The GCAS Review Journal
Vol. I, Issue 2/2021

Available online at https://www.gcasreview.com/publications-technology-and-society

The GCAS
REVIEW

Big Data in Contemporary Science: Methodological and Ethical Implications for Everyday Life

Edna Alves de Souza PhD, José Artur Quilici-Gonzalez PhD, Maria Eunice Gonzalez PhD

Universidade Estadual Paulista and Universidade Federal do ABC

Abstract: *To what extent have techniques for recording and analyzing massive amounts of data, also known as Big Data, influenced scientific methodology? This question guides the present interdisciplinary reflection concerning the contemporary digital culture, and its ethical implications. We investigate the concepts of causality and correlation, arguing that data, organized by mining, analysis, and modeling techniques, may show correlation, but not necessarily causation. While acknowledging current controversy concerning the relevance of Big Data analysis in the scientific methodology, here we argue that in contrast to causal associations, correlation is unable to reveal the reason for the occurrence of events, only signaling what could be happening in specific locations and situations. Nonetheless, the study of correlations can be of great help for decision-making in many areas of science, politics, and economics, among others. Considering the analytical methodology of Big Data resources, which favors the study of correlations, rather than causal analysis, we also discuss some ethical implications of the use of Big Data analytics in scientific methodology, which might reverberate in everyday life.*

The GCAS Review Journal
Vol. I, Issue 2/2021

Available online at https://www.gcasreview.com/publications-technology-and-society

Keywords: Causality, Correlation, Data-mining, Scientific method, Digital culture, Ethics.

The GCAS Review-Journal is a Publication of GCAS College Dublin, Ltd.

The GCAS Review Journal
Vol. I, Issue 2/2021

Available online at https://www.gcasreview.com/publications-technology-and-society

I. Introduction

The development of information technologies with unprecedented application potential has given rise to exponential growth in the volume of data collected, stored, and organized in a variety of formats, as well as the speed of collection. At the same time, the most varied data and digital tracks serve as building blocks of what can be provisionally understood as Big Data.

Technically, Big Data comprises at least two steps. The first one includes mechanical processes of data collection, recording, and 'cleaning'. The second one involves analysis, interpretation, and modeling elaborations based on the first stage of organization of digital data. This minimalist classification is just a starting point for the understanding of Big Data as a process, which is not limited to a set of data, characterized by the traditional 3Vs: Volume, Variety, and Velocity. The context of the digital culture nourished by Big Data, and the socio-technological issues raised by it, are aspects to be considered in the present characterization of Big Data.

Considering the complexity of Big Data resources and their prominence in the contemporary world, ranging from everyday life to the scientific universe, the central problem that guides this investigation can be summarized as follows: to what extent could Big Data analysis influence the methodology of science, which might reverberate in everyday life? We do not have a definitive, or even consensual, answer to this problem, but we believe that investigations of possible implications of the use of Big Data Analytics for scientific and epistemological research are not only relevant, but also crucial for understanding of the possible future of the digital culture.

Since the concept of data is not obscured in advance by the utopian or dystopian bias of socio-technological issues, the analysis of this concept seems to have the potential to clarify the current state of science with regard to its growing use of Big Data. Thus, in this paper, the polysemy of meanings of the term 'data' is the object of analysis in the **first section**. In the **second section**, we consider Big Data as the result of an overlapping cultural, technological, and ethical-scientific relationship. Since the analytical methodology of Big Data resources is dependent on algorithms that might favor the detection of correlations, to the detriment of causal analysis, we investigate the concept of algorithm, and contrast, in the **third section**, the concepts of causality and correlation. In **the fourth section**, examples of algorithm applications are presented, ranging from recommendation systems to public policies and collective awareness, finally reflecting on the ethical implications of adopting, in science and everyday life, strictly correlational strategies, in contrast to causal ones or the combination of both.

The GCAS Review Journal
Vol. I, Issue 2/2021

Available online at https://www.gcasreview.com/publications-technology-and-society

1. The Concept of 'Data' Revisited: Scientific Method and Theories of Scientific Method

In what follows, we adopt Laudan's distinction between scientific method, understood as a set of techniques and procedures used by scientists in the production of theories and experiments, and theories of the scientific method, which consist of meta- scientific ideas used in the study of the logic of scientific inference.

The history of the scientific method, according to Laudan[1], can be understood as a descriptive and explanatory portrait of the art of experimenting, which does not require meta-scientific discernment. In contrast, the history of scientific method theories can be understood as a descriptive and explanatory analysis of concepts underlying scientific practice, such as those of induction, deduction, hypothesis, and scientific explanation. From this perspective, the present article will focus on theories of the scientific method, concentrating on the meta-scientific plan of Big Data analysis.

There is a tendency in computerized societies to overvalue and/or fear novelties expressed by digital sciences and technologies, without a consensual understanding of what is *data*, a key concept in the meta-scientific plan. We believe that confusing different meanings of 'data' can have relevant philosophical implications, and possible practical reverberations. Therefore, in this section, we discuss some conceptions of data. So, what is data, after all?

Furner[2] analyzes the concept of data, presenting a historical approach to its use since its mention in classical Latin, identifying nine 'main interpretations' of the meaning of 'data': Classical (Data as Gifts), Documentary (Data as Metadata), Ecclesiastical (Data as Gifts of God), Geometric (Data as Geometric Premises), Mathematical (Data as Mathematical Premises), Epistemic (Data as Evidence), Informational (Data as Attribute-Values), Computational (Data as Bits), and Diaphoric (Data as Differences). For our purposes, we will highlight the epistemic and computational interpretations of the concept of data.

According to Furner[3], until the end of the 19th century, at least in English, the main interpretation of 'data' was the epistemic one. According to this interpretation, data are "Things given, or admitted; quantities, principles or facts given, known, or admitted, by which to find things or results unknown"[4]. Epistemic interpretation, in this sense, already gives rise to two distinct approaches: those of data as known facts and as admitted premises.

In the first sense of the term, 'data' is considered as that which is the object of pure verification; it is presented, by immediate experience, before any theoretical elaboration. According to this perspective, in

[1] Laurens Laudan, "Theories of scientific method from Plato to Mach: a bibliographic review," in *History of Science 7* (1968), 1-63.

[2] Jonathan Furner, "Data": The data in *Information Cultures in the Digital Age*, ed. Kelly, M. & Bielby, J. (Wiesbaden: Springer Fachmedien, 2016).

[3] Jonathan Furner, "Data".

[4] Jonathan Furner, "Data", 295.

the data resides the evidence from the empirical experience, as it is captured immediately. The data, thus understood, can be opposed to that which is constructed, which is the object thought, elaborated as a function of a problem posed by the understanding or by a theory.

The second meaning of data as an admitted premise, in turn, detaches it from the need of a correspondence with immediate experience. What is admitted, as a premise, may not be subject to any empirical constraints. As an example, we can consider mathematical deductions created from certain theoretical premises admitted beforehand, constituting data from a mathematical demonstration.

According to Furner's analysis, in the second half of the 19th century, there was a major change in the dominant interpretation of the concept of *data*. From then on, the term *data* started to be used to indicate the "content … about a referent", including values and (numeric) attributes contained in tables, systematically organized, in the areas of quantitative and qualitative studies. "The contents of these tables – the 'data' once collected and organized, became the raw materials for new, sophisticated forms of quantitative analysis – starting to be known as *data*"[5].

There is a significant change in the plan of analysis since the 19th century: data, considered in its ontological, epistemological, and semantic aspects, indicating elements for the unveiling of reality, starts to designate values and properties attributed to it. This interpretation of data as attributed values, measurements, and results of scientific research, with meaning established in a typically numerical way, is not abandoned; it still persists, above all, in Information Science. As Furner observes, in the computational scope, '*data*' came to be considered at a more abstract level than that of the attribute-values ascribed to it, being virtually synonymous with 'digitizable' or 'bits', that is, the binary 0 and 1 digits processed by computers[6].

Inspired by Furner[7], we propose the following classification of the term '*data*', outlined in the table below.

Table 1. Data typology.

Data$_1$	*Phenomenological data*: the basis from which knowledge of material objects can be experienced.
Data$_2$	*Informational data*: ordered patterns that assemble data$_1$ and also involve abstract data that can be expressed in numerical and propositional language.

[5] Jonathan Furner, "Data", 295.
[6] Jonathan Furner, "Data", 298.
[7] Jonathan Furner, "Data".

Data$_3$	*Digital data*: data$_1$ and data$_2$ mechanically organized in the form of binary digits.

We understand that the contemporary tendency to confuse these three types of data, namely data$_1$ (immediate), data$_2$ (ordered patterns), and data$_3$ (patterns organized in bits), can be a recurring source of misunderstandings in discussions about Data Science and Big Data[8].

According to Ibekwe-SanJuan & Bowker: "Although data are often presented as a natural phenomenon just waiting to be collected, nothing could be farther from reality"[9]. To support this statement, they rely on the discussion by Pushmann and Burgess[10] concerning the use of metaphors that associate the term Big Data with natural species such as 'gold', 'ocean', 'torrent', 'mineral', and 'oil', giving the impression that the former consists of a natural element. The usage of this metaphor indicates at least two contemporary tendencies: (a) to treat Big Data as a valuable commodity that can be not only stored, but also stolen; and (b) to treat it as a natural, ecological element, available as affordances for action in the environment.

We consider that the above assumption applies to the data$_2$ category, but not necessarily to the other data conceptions. When Ibekwe-SanJuan & Bowker emphasize that there is neither neutrality nor objectivity in the practice of data collection, they are referring to data$_2$. The collection exercise, as they emphasize, "[...] is governed by pragmatism (the goals of the study) and bound by technical constraints imposed by the data providers. This limits analysis possibilities in terms of data sources and content"[11].

In turn, the description of data$_3$, prepared by analysts, is not very different to what occurs when interpreting data$_2$, in that it seems to minimize pretensions regarding the naturalization of *phenomenological data* (data$_1$). In addition to the incompleteness of this type of data (existing in complex situations), changes in the data$_3$ set, selected for conducting a study, might also alter the 'discoveries' made from them.

In the Information Society, with the predominance of Big Data analysis, the analyzable data are mainly of the data$_3$ type. In this context, reality is not (nor is it intended to be) strictly described; what is mainly available, via data$_3$, is adaptable narratives suited to certain purposes. As stressed by Anderson[12], on the petabyte scale, data (for us, data$_3$) are first seen mathematically, and then a context is established in which they fit in the form of information. When properly contextualized, data become information!

[8] There is a long and fertile debate about data in philosophy, which we would not like to neglect, but it is not our task here to undertake a historiographical work; supported by Furner's survey (2016), we highlight only one aspect of this discussion: the possibilities of maintaining a conception of phenomenological data, while recognizing their interpretative character.
[9] Ibekwe-SanJuan & Bowker, "Implications of Big Data for Knowledge Organization," in *Knowl. Org. 44*, (2017), 193.
[10] Pushmann and Burgess, "Metaphors of Big Data," in *International Journal of Communication*, 8, (2014), 1690–1709.
[11] Ibekwe-SanJuan & Bowker, "Implications of Big Data for Knowledge Organization," (2017), 193.
[12] Anderson, "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete," in *Wired*, (June 23, 2008).

In summary, the trajectory outlined here aims to indicate that the selection and contextualization of data$_3$ leads to questions about the limits and scope of objectivity and knowledge of the reality provided by Big Data analysis resources. Emphasis should be given to the detail that 'data' often has different meanings, depending on the bias adopted by the researcher, or the area of research and its application. In this sense, the datification of science may represent, to a certain extent, an abandonment of the search for unveiling reality based on established scientific theories and practices.

The data interpretation, in traditional scientific practices, is obtained from the exploration of the causal nexus, while in scientific practices strongly anchored in Big Data strategies, data interpretation is usually extracted from correlational relationships. This computational research resource increases productivity and generates immediate practical results at a lower cost, although sometimes producing unreliable results.

In what follows, we describe the analysis of Big Data in depth, with emphasis on the concept of algorithm, indicating possible methodological and epistemological implications of its use in scientific, economic, and political research.

2 Big Data: from Data to Algorithms

Although there are several characterizations of Big Data, at least two conceptions deserve our attention. The first is predominant in the Humanities, conceiving Big Data as an area of investigation that focuses on a complex set of cultural, technological, and ethical-scientific relations; emphasis is given to datification in the study of ubiquitous information, available in the most diverse contexts. The second conception, predominant in the Exact Sciences, focuses on mining and 'cleaning' techniques for the massive amount of data that will be made available for analysis and modeling, employing computational resources. In this second approach, prominence is given to the concept of algorithm, though this concept is also present in the first meaning, especially in the area of Digital Humanities.

In line with the first conception of Big Data, Boyd and Crawford note that on the one hand, Big Data can be "[…] seen as a powerful tool to address various societal ills, offering the potential of new insights into areas as diverse as cancer research, terrorism, and climate change". But, on the other hand, it can also be seen "[...] as a troubling manifestation of Big Brother, enabling invasions of privacy, decreased civil freedoms, and increased state and corporate control"[13].

In the case of scientific research, the massive production of data, and the use of *datification* in scientific research, while investigating 'sociotechnical' questions and analyses, can lead to utopian or

[13] Boyd and Crawford, "Critical Questions for Big Data. Provocations for a cultural, technological, and scholarly phenomenon," in *Information, Communication & Society*, Vol. 15, No. 5, {June 2012), 662–679.

dystopian rhetorics, according to the propensities of those involved. As noted by Boyd and Crawford[14], these rhetorics are confused even with the very characterization of what is understood by Big Data: on the one hand, the exaggerated optimism of its enthusiasts, and on the other, the exacerbated pessimism of the radical critics.

The key to understanding of possible effects of the application of Big Data techniques in science, as well as in the autonomous action of moral agents, lies in a careful analysis of the data concepts, such as the one initiated in the previous section, and also of the concept of algorithm. We believe that the explanation of these concepts helps in the search for answers to the following question: What type of objectivity can algorithms, among other programming techniques, provide in the search for knowledge of empirical reality?

Instantiated in different contexts, $data_3$ constitutes the starting point of the computational processing performed by means of algorithms. However, the concept of algorithm (as well as that of data) is polysemic, ranging from structures organized by logical-mathematical rules to sets of procedures to be performed by machines using specific programs. According to Negnevitsky, an algorithm is "[...] a series of well-defined step-by-step operations"[15]. Nevertheless, confusions concerning the concept of algorithm are pointed out by Hill[16]:

> […] we see evidence that any procedure or decision process, however ill-defined, can be called an 'algorithm' in the press and in public discourse. We hear, in the news, of 'algorithms' that suggest potential mates for single people and algorithms that detect trends of financial benefit to marketers, with the implication that these algorithms may be right or wrong.

Following the trails of Hill[17] and Negnevitsky[18], Mittelstadt et al.[19] propose a conceptual analysis of the algorithm, distinguishing it from its instantiation, using the following definition:

> At the center, we follow Hill's[20] formal definition of an algorithm as a mathematical construct with "a finite, abstract, effective, compound control structure, imperatively given, accomplishing a given purpose under given provisions"[21].

The distinction between an algorithm and its instantiation is relevant for understanding the type of

[14] Boyd and Crawford, 2012.
[15] Negnevitsky, "Artificial Intelligence - A Guide to Intelligent Systems," (Harlow: Pearson Education Limited, Third Edition, 2011), 34.
[16] Hill, "What an Algorithm Is," in *Philosophy & Technology*, 56(6), (2015), 36.
[17] Hill, "What an Algorithm Is," 36.
[18] Negnevitsky, "Artificial Intelligence - A Guide to Intelligent Systems."
[19] Mittelstadt et al., "The Ethics of Algorithms: Mapping the Debate," in *Big Data & Society*, (July–December 2016), 1–21.
[20] Hill, "What an Algorithm Is," 47.
[21] Mittelstadt et al., "The Ethics of Algorithms: Mapping the Debate," (July–December 2016), 2.

information that instantiations of algorithms can provide in data manipulation. In contrast to specific instantiations, the algorithms, defined as abstract mathematical constructs, involve a finite number of components that characterize their logical structure, applicable in multiple tasks, in different contexts.

The Latent Dirichlet Allocation (LDA) algorithm, for example, can be used for a variety of functions, from clustering customers, based on their purchases, to automatic music creation. However, it is commonly applied to the detection of topics in a textual *corpus*. Behind the results of the application of this algorithm (whether they are considered successful or unsuccessful) are calculations that express: pre-established rules; parameters established *a priori* concerning the probability of occurrence of topics and the distribution of words in the topics; assigning weights to topics; training involving topic filtering; and, fundamentally, changes to the tool in order to adapt the model to the intended purpose. As it is implemented, this algorithm allows for several interpretations of the data.

In addition, the LDA is a 'bag-of-words' algorithm, in which the order of words does not matter. This characteristic of the LDA, in itself, already generates questions important to the philosophy of language. For example, if the meaning of a term depends on its use, as suggested by Wittgenstein[22], then the context is essential for obtaining a coherent textual interpretation that allows distinction of different meanings in an utterance that can have different interpretations (an example would be 'old friend'). In this case, bag-of-words is far from offering a natural context; the programmer has to contextualize the data to which the LDA is applied, in order to present a result that meets the motivations of the research or the user of this algorithm. As a direct consequence of methodological decisions, the meaning that is extracted from the set of data, grouped in different topics, is limited by the scope of the instantiation of the algorithm.

This line of reasoning favors the ontological aspects of concepts such as data and algorithms; they are inserted into the environment in which they are employed or created. In the case of problematic concepts, which in themselves provide little help in clarifying ambiguous propositions, it is necessary to clarify the conditions under which the instantiation of an algorithm is applied, in order to preserve minimal aspects of objectivity. But, what procedures should be adopted to obtain such a clarification? Within this context, John Searle suggests that:

> In the twentieth century, philosophers learned to be very cautious about asking questions of the form, 'What is . . .?', as in, for example, 'What is truth?', 'What is a number?', 'What is justice?'. The lessons of the twentieth century (though these lessons are rapidly being forgotten in the twenty-first century) suggest that the best way to approach such problems is to sneak up on them. Do not ask, 'What is truth?', but ask, 'Under what conditions do we say of a proposition that it is true?'. Do not ask, 'What is a number?', but ask, 'How do numerical expressions function in actual

[22] Wittgenstein, "Philosophical Investigations," Blackwell Publishing, (2011[1953]).

The GCAS Review Journal
Vol. I, Issue 2/2021

Available online at https://www.gcasreview.com/publications-technology-and-society

mathematical practice?'[23]

Following Searle's trail, we can ask: How do numerical expressions (articulated in the analysis of massive amounts of data, using algorithms) function in contemporary scientific practice? We investigate this issue, with an emphasis on scientific research, deepening the analysis of the debate 'correlation versus causality' based on Big Data resources.

3. Big Data: Correlation *versus* Causality?

Although causation involves correlation, in the sense that X is related to Y, there is an attempt in science to eliminate spurious correlations, in order to detect causation and demonstrate that X necessarily implies Y. However, this attempt is not very relevant in the context of the e-Science that is taking shape in the 21st century through the use of Big Data techniques.

Gray[24], in an enthusiastic spirit, considers that we are experiencing the emergence of a fourth paradigm, in which science is based on data. According to his analysis, a thousand years ago we had Experimental Science, characterized by the description of natural phenomena, of which Newton's laws and Maxwell's equations provide good examples. Later, sophisticated computational resources allowed the possibility of simulating complex phenomena. Today, we experience the development of a Data-centric Science: The e-Science.

In the e-Science, researchers deal primarily with data sets from different sources: data captured by instruments, extracted by sensor networks, generated by simulations at different scales, and so on. In this sense, as noted by Gray[25], e-Science is related to Big Data and the 'Internet of Things'. Thus, e-Science is based on the set of tools and technologies that involve the acquisition, recording, and 'cleaning' of data available for analysis, modeling, and interpretation. But, in what sense is the concept of data (of types 1, 2, or 3) being used? Possibly in the three senses indicated in Table 1, with predominance of the type data$_3$.

In contrast to the optimistic view of the use of Big Data in science, Marcus & Davis[26] present a critical opinion in relation to this practice, as well as to what they consider its extravagant self-promotion. In line with these researchers, we understand that instruments for manipulating massive amounts of data could be a complement, but not an alternative or a substitute, for traditional scientific research. Although Big Data techniques are efficient for detecting correlations, they do not specify the criteria of relevance for

[23] Searle, "What is an institution?," in *Journal of Institutional Economics*, 1:1, (2005), 2.
[24] Gray, "On eScience: A Transformed Scientific Method," in *The Fourth Paradigm*, ed. Tony Hey, Stewart Tansley & Kristin Tolle, Data-Intensive Scientific Discovery, (Washington: Microsoft Research, 2009), (Based on the transcript of a talk given by Jim Gray to the NRC-CSTB in Mountain View, CA, on January 11, 2007).
[25] Gray, "On eScience: A Transformed Scientific Method," 2009.
[26] Marcus & Davis, "Eight (No, Nine!) Problems with Big Data," in *The New York Times*, (April 6, 2014).

selecting significant correlations; they also do not explain why the criteria are considered relevant[27].

Another important aspect to be considered, as mentioned by Ibekwe-SanJuan & Bowker[28], is that research guided by Big Data is difficult to replicate, due to the private character of several algorithms. One must also consider the transient, dynamic, and heterogeneous nature of Big Data, which hinders the well-established scientific test of reproducibility of experiments: since replicability has, so far, been one of the canons of science.

Despite the difficulty in replicating experiments and modeling, Big Data techniques can be of great help in assisting literary, bibliographic, and political research, by enabling the gathering and crossing of massive amounts of data, at different times and in different contexts. This is one of the reasons for the adoption of its use in the most diverse fields of knowledge. Especially in science, Big Data is bringing novelty considered revolutionary by some thinkers. Anderson[29], for example, states in the subtitle of his article that: "The data deluge makes the scientific method obsolete". He decrees the end of traditional scientific theory, concerned with the creation of new taxonomies or scientific models, by stating that in times of cloud computing and massive sets of data, the real challenge is to sift the data in order to find significant correlations. However, Anderson[30] does not clarify what could be the criteria of relevance involved in the selection of data for the detection of significant correlations.

According to Anderson: "With enough data, the numbers speak for themselves"[31]. The idea promulgated is that we are experiencing the era of large amounts of data, applied mathematics, and data mining techniques, which are replacing other techniques and analytical tools that have been used in science so far. From his point of view, theories of human behavior (from linguistics to sociology), taxonomy, ontology, and psychology do not matter, because "Who knows why people do what they do? The point is they do it". Anderson's assumption[32] is that with Big Data, we can track and evaluate these actions with unprecedented fidelity.

The above perspective counters a conception of science according to which the scientific method is built from testable hypotheses based on law-like causal relationships, that is, on strong regularities. Scientists, especially in the Exact and Biological Sciences, commonly conceive systems in which their

[27] Tyler Vigen (2015) was known for presenting funny spurious correlations on his website *Spurious Correlations*. He developed a computer program that searches for correlations in massive databases, providing results that are at least provocative, such as correlation between reduction in per capita consumption of margarine and reduction of divorces. However, Vigen's purpose is not to provide humor, but rather to denounce the irresponsible use of statistical correlations propagated, above all, on social networks.
[28] Ibekwe-SanJuan & Bowker, "Implications of Big Data for Knowledge Organization," (2017).
[29] Anderson, "The End of Theory," (2008).
[30] Anderson, "The End of Theory," (2008).
[31] Anderson, "The End of Theory," (2008), 2.
[32] Anderson, "The End of Theory," (2008).

hypotheses take place, serving them as models of causal explanation, which are then empirically tested. Experiments validate or invalidate theoretical models of the functioning of regularities observed in the world. To avoid hasty conclusions, based on mere correlations between X and Y, experiments are objects of discussion and testing among members of the scientific community.

Despite the various conceptualizations of *causality*, we assume here the current one according to which, in an efficient causal relationship, if X is the cause of Y, then X must precede Y ($X \rightarrow Y$), so that, logically speaking, "if not Y, then not X ($\neg Y \rightarrow \neg X$)", while the correlation is a statistical relationship of probabilistic dependence or association between two variables. In this sense, the causal relationship can be identified as a correlation, but the reverse does not follow, that is, not all correlational relationships can be identified as causal. In the correlation, there is no necessary relationship between X and Y, but only the co-occurrence of these variables. In this case, the occurrence of X commonly accompanies that of Y, but there may be situations in which X occurs despite the non-occurrence of Y. Thus, the correlation between the consumption of margarine and the divorce rate does not imply causality between these events, although, in practice, they may coincide.

There is not, necessarily, a dichotomy between the correlational and causal strategies, but they might be complementary. Current scientific practices that use Big Data and give prominence to correlations raise the problem of selecting those ones that are significant, not spurious. In efforts to understand the underlying mechanisms that connect X to Y, in addition to mere correlation, philosophers and scientists have for centuries been developing causal theories, with logical and empirical foundations. Contrary to the assertions of Anderson[33], here we understand that the structural basis of the established scientific method remains founded on causal relations, despite the development of instruments and resources to deal with correlations among massive amounts of data.

However, Anderson, in a dichotomous perspective, insists that the classic method of scientific investigation (of hypothesis, model, and test) is becoming obsolete in light of Data Science and Big Data techniques. The way forward, he envisions, is to find significant correlations in Big Data. In his words:

> Petabytes allow us to say: "Correlation is enough". We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot[34].

The passage above reinforces the assumption that the availability of massive amounts of data, with the statistical tools and algorithms suitable for processing these numbers, offers a new way of understanding

[33] Anderson, "The End of Theory," (2008).
[34] Anderson, "The End of Theory," (2008), 3.

the world. In this new form of scientific practice, correlation tends to replace causation, and science can advance without coherent models, unified theories or, in short, without a causal explanation.

In summary, we understand that Anderson's[35] reading of the power of Big Data underestimates the importance of traditional scientific methods. Contrary to what he claims, a heap of data alone has little, if any, relevance to scientific explanation. Unless we know what to look for, and have a criterion of relevance for developing a model of what can happen in certain situations, we might not even ask relevant questions about the data: a theoretical basis and a criterion of relevance are fundamental for the formulation of coherent questions. As Timmer pointed out, "Correlations are a way of catching a scientist's attention, but the models and mechanisms that explain them are how we make the predictions that not only advance science, but generate practical applications"[36]. Correlations can produce illusory results, even if large data sets are used.

Several authors argue that due to its remarkable specificities, Big Data constitutes a paradigm shift in scientific research. Thus, for example, Fan et al.[37] draw attention to the fact that while, in traditional datasets, the sample size is usually larger than the dimension (number of variables), Big Data is characterized by huge data sizes and high dimensionality. Also, Gandomi and Haider[38] brings together work by other authors demonstrating that the massive size of datasets tends to falsely correlate independent variables. Fan et al.[39] present a computer simulation in which the correlation coefficient of independent random variables increases as the size of the datasets grows. To understand why high dataset dimensionality favors the appearance of spurious correlations, consider the fact that the larger the population of a city, the greater the chances of finding two people of similar appearance but without any genetic relation.

As early as the 1980s, researchers such as Clark Glymour and his graduate students were already studying causation in the social sciences, using structural equation models. The introduction of computational resources to resolve equations in the social sciences would have strong impacts in theoretical and applied research on the epistemology of causation, as well as an alignment with computer science research groups, such as the Judea Pearl group, which independently elaborated an algorithmic causal theory based on Bayesian networks.

As reported by Scheines[40], the Glymour group represented causal modeling with linear parametric

---

[35] Anderson, "The End of Theory," (2008).
[36] Timmer, "First the cloud, now AI takes on the scientific method. Cloud didn't make the scientific method irrelevant in '08 - AI won't do it in '17, either," in *Ars Technic*, (May 28, 2017), 3.
[37] Fan et al., "Challenges of Big Data Analysis," in *National Science Review*, 1, doi: 10.1093/nsr/nwt032, (2014), 293-314.
[38] Gandomi & Haider, "Beyond the Hype: Big Data Concepts, Methods and Analytics," in *International Journal of Information Management 35*, (Elsevier, 2015), 137-144.
[39] Fan et al., "Challenges of Big Data Analysis," (2008).
[40] Scheines, "Computation and Causation," in *Cyber Philosophy: the Intersection of Computing and Philosophy*, edited by James H. Moor and Terrell W. Bynum. (Oxford: Blackwell Publishing Ltd., 2002).

systems, while the Pearl group used qualitative directed graphs. As probabilistic causal models represent the dependencies between variables by means of directed acyclic graphs, these models have been used to simulate inferences and learning in structures without feedbacks, while structures with feedbacks are represented by graphical causal models.

With the availability of massive amounts of data, probabilistic causal models have placed research in the areas of exact and biological sciences in a new stage of scientific advances. Thus, for example, with the help of Bayesian networks, data from thousands of patients affected by a certain disease are being used in fully computational research to discover probabilistic relationships with their possible causes. However, their use in human sciences should be viewed more cautiously, because the complex subtleties involved in moral, legal, cultural, and political issues might not be properly modeled by linear and direct causalities.

Mistaken recommendations grounded on spurious correlations can also be considered in the dynamics of opinion and action, as presented in the next section, which indicates ethical implications of the use of Big Data analytics that might reverberate in everyday life.

4. Big Data: From Recommendation Systems to Collective Awareness

In addition to spurious correlations, possibly generated by misuse of Big Data in e-Science, examples of unreliable recommendations can be considered in the dynamics of opinion and action, supported by statistical correlations from Amazon, Netflix, and others.

In the Music Genome Project[41], for example, each song is classified by an algorithm, using approximately four hundred and fifty descriptive attributes, such as the musical instruments used, the singer's genre, theme, rhythm, and so on. These attributes can be interpreted as components of a vector, and different songs can be represented in the same vector space. The shorter the distance between two vectors, the more similarities there are between the songs they represent. Thus, the algorithm organizes music according to the genres: Pop/Rock, Hip-Hop/Electronic, Jazz, World Music, and Classical. Similarly, dating sites ask the person interested in finding a compatible partner hundreds of questions, such as: What is your age? Do you smoke? Are you liberal or conservative? Are you vegetarian? Are you monogamous? Do you want children? The answers provided can be represented in a vector space, so that the distance between two vectors measures the possible affinity between people.

Considering the recommendations of songs or films, the internet users behave as economic agents (ones that can be passively induced to consumption), but for the choice of a relationship, the users behave as moral agents, with rooted beliefs and ethical principles, with a certain degree of political awareness and sometimes with very strong emotional motivation and autonomy. Recommendation systems have shown

[41] Cf. Patent of Music Genome Project: https://patents.google.com/patent/US7003515.

success for economic agents, and a little success for moral agents with a certain degree of awareness and autonomy against external pressures, because confirmation biases (the tendency to select, remember, interpret, research, or deliberate from information that, in general, reaffirms beliefs or initial hypotheses) filter the incoming data.

Although causality (understood as an efficient cause) can explain the relationship among several physical events, such as the grouping of music in a vector space so that the distance between two vectors measures the similarity between them, it is not sufficient to explain the action of a moral agent, which presupposes a complex directionality. The directionality of the action differentiates it from pure movement, which is typical of physical bodies. Regarding this matter, Juarrero[42] proposes an explanation of the intentional action as follows: a stimulus received (efficient cause) acts in the organism (material cause), which, by means of its internal dispositions (formal cause), updates its purposes in a given action.

In recommendation systems, serendipity is an important measure of how surprising or unusual a recommendation can be. Obvious recommendations have no economic value to customers, even though they may seem very pertinent. In scientific research, the high serendipity of a correlation between the cause and effect of a disease may be the key to a discovery in medicine. However, the originality of an association or correlation depends on further analysis to find a causal link and validate its application. Suppose, for example, that a search in a large database associates allergy to certain foods with previous consumption of some medication, such as an antacid. How can we know if we are facing yet another spurious correlation, or an important scientific discovery with high serendipity? Currently, the validation of such correlations is accomplished by studies based on the traditional scientific research methodology, strongly supported by causality.

In contrast, in the methodology based on Big Data techniques, data mining tools detect patterns of association in large databases, generating several hypotheses to be selected, ideally among the most reasonable ones. New tests are carried out to refute, reformulate, or corroborate these hypotheses, with the difference (compared to traditional methodologies) that the process is now entirely virtual, using computational resources structured mainly on the basis of correlational associations. In this case, testing of hypotheses is performed in the same computational environment that produced them, potentially creating inbreeding scenarios.

Search engines, recommendations for reading, and suggestions for purchases, among other services, are being improved, including ones based on disastrous or positive practical experiences. It is an open question as to whether these mechanisms will be able to achieve, or simulate, the degrees of complexity, personalization, and subtlety that human beings present in dynamic social contexts. However, the resources

[42] Juarrero, "Dynamics in Action: Intentional Behavior as a Complex System," (Cambridge: MIT, 1999).

we now have suggest the need for caution regarding the generalized scope of the datification project.

The use of correlations as a predictive basis can be disastrous, especially in the dynamics of public opinion and action. Search and recommendation systems can inadvertently induce people to change their behaviour, instigating undesirable patterns of fear or euphoria, as can occur, for example, in pandemic situations. In these complex circumstances, no mechanical and autonomous search and recommendation systems are highly desirable. In the same way, in the era of Big Data, the shift towards the use of datification and correlation resources in science and everyday life deserves attention from researchers, considering both their performance and the associated ethical implications.

5. Possible Ethical Consequences of the use of Big Data to Model Social Events

An example worth mentioning in the context of Big Data modeling in the domain of social events is the controversial algorithm COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), used in the criminal justice system of the United States to determine the degree of danger of possible criminals, together with the likelihood of them repeating the offense/crime. The use of this algorithm, which is decisive for the attribution of sentences, leads to the question of whether, in fact, it helps with the imposition of sentences that are more just and objective, compared to those imposed by a regular court influenced by subjectivity, errors, bias, and preconceptions.

As described by Maybin[43], COMPAS is based on the use of a questionnaire aimed at identifying the general traits of a person's behavior, employing a system of points with ratings from one to ten. The COMPAS algorithm is a commercial trade secret, so there are no published details of how it transforms responses into points and performs the calculation. All that we know at the present stage is that this algorithm deals with information about where the person lives, the crime rate in that region, family crime/prison history, the number of times the individual has been arrested, links to people belonging to gangs, and school and professional history, among other aspects.

Maybin explains that COMPAS assigns the highest scores to offenders from ethnic minorities: analysis of the data for two accused with the same profile (age, sex, and criminal record), one being black and the other white, showed that the former was 45% more likely to receive a high score. Although there are no race-specific questions in COMPAS, the results obtained can be informative in terms of racial issues. It may be that the algorithm itself does not, in fact, have a racially prejudiced bias, but that it exposes racial preconceptions existent in the penal system and in a specific society, reinforcing and even maximizing such tendencies. If applied in another society, it might signal other differences, such as greater criminalization of

[43] Maybin, "How maths can get you locked up," in *BBC News*, (2016), available at: https://www.bbc.com/news/magazine-37658374. Accessed: 25/09/2020.

the poor.

The above example illustrates the popularization of social networks, with detailed profiles of users, which has generated extraordinary amounts of opinion that also envisions the possibility of creating mental clones, as a kind of individual's double, by means of advanced Artificial Intelligence simulation techniques. Insurance companies, government agencies, and private employers have already used personal data to approve or reject individual or collective solicitations.

In studies of collective awareness using the voluntary publication of detailed personal data on the Internet and social networks, Big Data resources (with the creation of mental clones) allow modeling of the ways in which individual and collective reality can be explained. In the mental experiment known as Newcomb's paradox4, Nozick[44] illustrates this possibility by considering a virtual situation involving two participants: one of them, participant 1, has the function of making future predictions about the behavior of participant 2, who has to decide between choosing box B, or both box A and box B. Participant 2 knows that inside box A, there is $1,000, while the content of box B is uncertain and may be $1,000,000 or nothing.

Assuming that participant 1's predictions are correct (and participant 2 knows that all of participant 1's previous predictions were correct), if participant 2 chooses only box B, it will contain $1,000,000, but if participant 2 decides to take boxes A and B, in this case box B will be empty. The experiment is paradoxical, because the analysis of two distinct and equally plausible strategies for maximizing the gain of participant 2 produces conflicting results.

In strategy 1 (the maximizing of gain is based on the expected-utility principle), participant 2 decides that participant 1 has excellent forecasting capacity, and therefore the best strategy is to take only box B with $1,000,000. But in maximization strategy 2, based on the dominance principle, participant 2 decides that taking boxes A and B will earn him/her $1,000 more than taking just box B, regardless of the prediction of participant 1, eliminating the possibility of taking only box B and that it is empty.

Considering the hypothesis that participant 1's prediction is incorrect, the experiment could yield either $0 (choosing only the empty B box) or $1,001,000 for participant 2. Therefore, taking boxes A and B would ensure either $1,000 or $1,001,000 for participant 2. Strategy 1 is strongly influenced by a deterministic, non-fallibilist approach, while strategy 2 is anchored in the idea of rationality and degrees of free will.

To make this mental experiment closer to simulated reality, consider the hypothesis that participant 1 has excellent ability to make correct predictions, because he/she is a mental clone of participant 2 with

[44] Nozick, "Newcomb's Problem and Two Principles of Choice," in *Paradoxes of Rationality and Cooperation, Prisoner's Dilemma and Newcomb's Problem*, edited by Richmond Campbell and Lanning Sowden. (The University of British Columbia Press, 1985).

personal data extracted from that part of Big Data related to participant 2's past life. If the probability of participant 1 making correct predictions is p, then the probability of incorrect predictions is 1-p.

In short, the above illustration of a simulation environment illustrates the fact that the simple adjustment of a single parameter can produce diametrically opposite results. But, what is the relevance of this illustration for practical matters of everyday life?

To appreciate the relevance of Newcomb's paradox4 in everyday life, one might consider a situation in which a researcher has to choose between two candidates for a scholarship. As their performance in the entrance exam was equivalent, the researcher makes the decision to use, as a tiebreaker, the possibility that the student might cheat in exams in the future. In a scenario where Big Data allows the creation of mental clones, based on unstructured data from social networks, the university technicians can simulate the behavior of each of the candidates in tempting situations, and thus choose which one deserves to receive the scholarship, even before either of them committed the offense of cheating in exams. It seems that the ethical consequences of this kind of inference are worthy of attention.

Some people are aware that all their data present on the Internet may be used by Big Data analytics to decide on their requests. In a negative scenario, this situation can be compared to Bentham's Panopticon, designed to obtain power of mind over mind, in order to control prisoners by means of mental force, rather than using forced labor[45]. But, as Foucault[46] observed, the guilty person is only one of the targets of punishment, which in practice ends up being extended to all others potentially guilty. For better or worse, most people change their behavior as soon as they realize they are being watched.

In this scenario, what could happen with the attempt to replace, for example, judges with computers assisted by Big Data? Our provisional hypothesis is that serious methodological deficiencies might occur if Hume's principle, according to which morality constitutes an independent domain of thought, is held to be valid.

In his Is-Ought problem, Hume[47] argues that when we are describing or studying the world, we are dealing with facts and descriptive propositions (about what is), whereas when we are judging someone's conduct, we are dealing with values and normative propositions (about what ought to be). According to Hume, it is not possible to rationally justify the passage from the domain of "what is" to the domain of "what ought to be". The main reason is that prescriptive or normative statements are mainly justified by emotional intelligence or cultural customs, which are transformed into habits of mind, generic norms, or

---

[45] Bentham, "The Panopticon Writing," (London: Verso, 1995), available at: https://www.fcsh.unl.pt/docentes/rmonteiro/pdf/panopticon_jeremybentham.pdf. Accessed 13-01-2022.
[46] Foucault, "Discipline and Punish: The Birth of the Prison," (New York, NY: Random House, Inc., 1995).

[47] Hume, "A Treatise of Human Nature," (The Gutemberg Project, Release Date: March 4, 2002 [1739]), available at: https://www.gutenberg.org/files/4705/4705-h/4705-h.htm.

laws[48]. As stressed by Dworkin:

> "Hume's principle - itself a moral principle - is sound: any argument that either supports or undermines a moral claim must include or presuppose further moral claim or assumptions."[49]

It is true that the scientific methodology employed by the natural or the social sciences uses the mode of induction, also put under suspicion by Hume in the case that they operate from the particular to the universal. However, particular data obtained in laboratories or collected from real life are used to test hypotheses and establish universal causal laws, valid in the "what is" domain.

In the humanities or law, established laws have validity in the "what ought to be" domain. When we use data extracted from the "what is" domain and by means of Big Data predictions we make decisions that reach into the "what ought to be" domain, we are putting into practice a methodology denounced as suspicion by Hume.

In the negative scenario under consideration, as long as we do not have (yet?) computers with emotional intelligence or the cultural sensitivity necessary to interpret laws according to the particularities of each case or historical epoch, the use of machines to make moral decisions based only on the application of algorithms for legal rules may represent a serious methodological deficiency. Differently, in an embodied embedded environment, a judge, when analyzing the criminal facts imputed to a defendant, incorporates from the very beginning an approach that considers the moral values of his/her time. In this way, he/she can provide a fair sentence, using valid scientific methodology, without clashing with Hume's principle that morality constitutes an independent domain of human thought.

From the point of view of the ethical implications that the use of induction by human and machines can raise, in the domain of law enforcement, for example, it is more reasonable to accept a sentence written by a judge, based on limited empirical data, but with statistical significance acquired with legal wisdom, than to accept a generalization based on simulations using Big Data and hypothetical parameters.

For centuries, jurists have been discussing whether someone can be considered both guilty and mad, since the two concepts seem to be legally exclusive. It is hard to imagine how induction based on machine learning will ever be able to handle this kind of subtlety with the same legal complexity exercised

---

[48] In his *A Treatise of the Human Nature*, Hume argues that: "In every system of morality, which I have hitherto met with, I have always remarked, that the author proceeds for some time in the ordinary way of reasoning […] when of a sudden I am surprized to find, that instead of the usual copulations of propositions, is, and is not, I meet with no proposition that is not connected with an ought, or an ought not. This change is imperceptible; but is, however, of the last consequence. For as this ought, or ought not, expresses some new relation or affirmation, it is necessary that it should be observed and explained; and at the same time that a reason should be given, for what seems altogether inconceivable, how this new relation can be a deduction from others, which are entirely different from it." (Hume, 2002, Book 3, Section 1, Part 1).
[49] Dworkin, "Justice for Hedgehogs," (London: The Belknap Press of Harvard University Press, 2011), 99.

The GCAS Review Journal
Vol. I, Issue 2/2021

Available online at https://www.gcasreview.com/publications-technology-and-society

by an experienced judge. Although the inductive reasoning of judges is subject to criticisms, it predominantly takes place within the independent domain of morality, while the inductive algorithmic method of computers starts from syntactic premises and arrives at semantic conclusions.

The point defended here is that in a positive scenario, the autonomy that computers acquire might be admirable in certain circumstances, for example when driving a spacecraft without direct inputs from astronauts, or in automated research using genetic databases. However, the same autonomy provided to computers in the realm of human sciences, without the supervision of a human being, could bring enormous damage to the moral values most crucial for the emancipation of the human being.

Similar situations can be envisaged where criteria need to be established for public policies in areas such as the distribution of social benefits, home financing, and rates of taxation for rich and poor, among others. Some situations, given their dynamics, can be understood from a perspective of circular causality (retroactive and recursive), rather than from a linear one. Linear causality occurs when event A causes event B, without the latter having any effect on the former. In circular causality, event B can affect event A, even if the latter was the original cause of that the former, such that the effect of a cause affects its own cause, altering it and therefore subsequently being altered by it.

In a positive scenario, an application of scientific interest for methodologies based on Big Data is the modeling of (emergent) collective behaviour under psychological stress that can contribute to disease pandemics with lethality distributed throughout all social classes. Circular causality is also present in Marx's principle (discussed in *The Eighteenth Brumaire of Louis Bonaparte*) according to which human beings make their own history, but in conditions inherited from their predecessors and not chosen by them. If this principle is valid, the dynamics of opinion and action in societies can be studied, using Big Data analytics, from the perspective of complex systems with circular causality present in the form of emergent properties.

To provisionally conclude this essay, we will consider ethical consequences, sometimes unexpected, from the use of circular causality not only in scientific methodology, but also in the manipulation of massive amounts of data, with effects on public opinion and action.

Provisional Conclusions

In this article, we discussed ways to answer the central question that guided the present investigation: To what extent have Big Data analysis techniques influenced scientific methodology? We emphasized that the relevance attributed to data, characterized as constructs or biased appropriations, is a strong contemporary trend, due to the prominence of Big Data analysis resources. This trend, known as Dataism, suggests that not even the traditional sciences, from physics to sociology, would be immune to datification, given the central relevance of digital data and correlational methodologies largely guided by algorithms. We highlighted that the correlationist strategy, in contrast to the causalist strategy, does not require knowledge about the cause of a phenomenon; immediate practical results are expected, rather than a deep understanding of reality as sought in the epistemological tradition.

Although it seems commonly accepted in e-Science that Big Data resources can reveal significant useful correlations, the importance of established scientific methods should not be underestimated. As we have suggested, a pile of data alone has little, if any, relevance to scientific explanation: If *what* -present in correlations - can be sufficient for an 'instrumental' rationality, the criticality of *why* - present in causal explanations - still seems to be fundamental to human reason.

We also point out problems with datification in the field of collective opinion and action. We recognize the historical contingency of knowledge (a reflection of healthy scientific fallibility), as well as the postmodern insistence on the relevance of the multiplicity of local knowledge. However, we believe that the prevalence of $data_1$ (phenomenological data that constitutes the basis from which knowledge of material objects can be experienced) should be a mainstay in efforts to preserve commitment to a project of science adjusted to humanist aspirations, seeking to understand reality and respecting individual freedom, social justice, and autonomy of action. In times of Big Data, when practical results are being sought in the most diverse areas of knowledge, such humanistic desires, instead of being abandoned, ought to be carefully cultivated.

From this standpoint, private corporate interests or corporate copyright should not restrict access to Big Data. Although personal data should be protected by the right to privacy, this same data should not become secret, that is, with due respect to legal process, any part of Big Data should be accessible to legitimate stakeholders and be protected against unauthorized exploitation.

We hold that a methodology based on the information gathered from people's data needs to consider the ethical implications of the use of such data. Currently, based on ethical principles, members of the scientific community may contest laboratory experiments involving the use of animals or humans in degrading or painful circumstances. In the same vein, the scientific methodology that justifies or challenges such principles must also consider the ethical and legal implications of the use of personal data in the Big Data era. Although the data collected about people might not be private, the distribution and use of these resources have ethical implications.

Two scientific methods were discussed here: the traditional method and e-Science, based on different theories, with one being causal, based on $data_1$, and the other being correlational, based on $data_3$. If the correlation is not spurious, the results of different scientific methods must be equally valid. If the correlation used as a scientific inference is spurious or accidental, the researchers will face difficulties in detecting this problem in the virtual Big Data environment itself. They will have to resort to the established scientific method of causation. The root of this problem lies in the distinct nature of the data used and the corresponding laws that govern it.

The aim of the reflection initiated here is to open a dialogue envisaging feedback from the scientific and philosophical community about the influence of Big Data analytics in scientific methodology and ethics. We understand that a scientific methodology that does not consider the historicity and the complex idiosyncrasy of a society, treating it mainly as a physical system with mathematical laws and Boolean logic, may subject the logic of societies to that of physical systems. However, if it is not recognized that the logic that underlies physical events is not always the same as that which dominates political or social events, then there is the possibility of unexpected and ethically undesirable consequences.

As the widespread generation and collection of human activity data becomes ubiquitous, Big Data may come to represent the collective awareness of a society at a given historical moment. In this way, it could be analogous to a nation's mother tongue, which is a form of collective awareness, constituting a unifying force with a property shared by the people. If this is a feasible possibility, then the proposal of ethical principles for scientific research, as well as for the use of Big Data resources in society, may be of great value.

We understand that the inadvertent incorporation of the presumption of continuity of circular causality applied to moral agents in legitimizing theories of scientific methodology, and in computer simulations to predict future decisions, can generate methodological mistakes with unpredictable consequences. However, the relevance of information of the type $data_1$ seems to speak for itself, especially in problematic situations where life is put at risk (such as during pandemics), and reminds us that the real is still more fundamental than the digital.

Bibliography

ANDERSON, C. (2008). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. **Wired**, June 23, 2008. Available at: https://www.wired.com/2008/06/pb-theory/. Accessed: 30-09-2019.

BENTHAM, J. (). **The Panopticon Writing**. London, England: Verso, 1995. Available at:

https://www.fcsh.unl.pt/docentes/rmonteiro/pdf/panopticon_jeremybentham.pdf. Accessed 13-01-2022.

BOYD, D. & CRAWFORD, K. (2012). Critical Questions for Big Data. Provocations for a cultural, technological, and scholarly phenomenon. **Information, Communication & Society**. Vol. 15, No. 5, June 2012, pp. 662–679.

DWORKIN, R (2011). **Justice for Hedgehogs**. London, England: The Belknap Press of Harvard University Press.

FAN, J. & LV, J. (2008). **Sure Independence Screening for Ultrahigh Dimensional Feature Space**. Journal of the Royal Statistical Society, Volume 70, Issue 5, pp. 849-911, November 2008. Available at:

https://rss.onlinelibrary.wiley.com/doi/10.1111/j.1467-9868.2008.00674.x. Accessed 15-01-2022.

FAN, J. et al. (2014). **Challenges of Big Data Analysis**. National Science Review, 1: pp. 293-314, doi: 10.1093/nsr/nwt032, 2014.

FOUCAULT, M. (1995). **Discipline and Punish: The Birth of the Prison**. New York, NY: Random House, Inc., 1995.

FURNER, J. (2016). "Data": The data. *In*: KELLY, M. & BIELBY, J. (Eds.)
Information Cultures in the Digital Age. Springer Fachmedien Wiesbaden.

GANDOMI, A & HAIDER, M. (2015). **Beyond the Hype: Big Data Concepts, Methods and Analytics**. International Journal of Information Management 35 (2015), pp. 137-144, Elsevier, 2015.

GLASER, W. T. et al. United States Patent No. US 7,003,515 B1. Date of Patent: Feb. 21, 2006. Available at: https://patents.google.com/patent/US7003515. Accessed: 04-04-2020.

GRAY, J. (2009). On eScience: A Transformed Scientific Method. *In*: Tony Hey, Stewart Tansley & Kristin Tolle (Eds.) **The Fourth Paradigm**. Data-Intensive Scientific Discovery. Washington: Microsoft Research. (Based on the transcript of a talk given by Jim Gray to the NRC-CSTB in Mountain View, CA, on January 11, 2007).

HILL, R. (2015). What an Algorithm Is. **Philosophy & Technology** 56(6), 2015.

HUME, D. (1739). A Treatise of Human Nature. The Gutemberg Project. Release Date: March 4, 2002. Available at:

https://www.gutenberg.org/files/4705/4705-h/4705-h.htm.

IBEKWE-SANJUAN, F. & BOWKER, G. (2017). Implications of Big Data for Knowledge Organization. **Knowl. Org**. 44, 2017.

JUARRERO, A. (1999). **Dynamics in Action**: Intentional Behavior as a Complex System. Cambridge: MIT.

LAUDAN, L. (1968). Theories of scientific method from Plato to Mach: a bibliographic review. **History of Science** 7 (1968), 1-63.

LOCKWOOD, D. (2015). **A New Approach to Rule-Consequentialism**, pdf unpublished paper version.

MARCUS, G. & DAVIS, E. (2014). Eight (No, Nine!) Problems with Big Data. **The New York Times**. April 6, 2014.

MARX, K. (1963 [1852]). **The Eighteenth Brumaire of Louis Bonaparte**. New York: International Publishers.

MAYBIN, S. (2016). How math can get you locked up. In: **BBC News**. Available at: https://www.bbc.com/news/magazine-37658374. Accessed: 25/09/2020.

MITTELSTADT, B. D. et al. (2016). The Ethics of Algorithms: Mapping the Debate. Big Data & Society July–December 2016: 1–21.

NEGNEVITSKY, M. (2011). **Artificial Intelligence** - A Guide to Intelligent Systems. Pearson Education Limited, Harlow, England. Third Edition.

NOZICK, R. (1985). Newcomb's Problem and Two Principles of Choice. *In*: **Paradoxes of Rationality and Cooperation**. Prisoner's Dilemma and Newcomb's Problem. Edited by Richmond Campbell and Lanning Sowden. The University of British Columbia Press.

PAULEEN, D. J.; ROONEY, D. & INTEZARI, A. (2015). Big Data, Little Wisdom: Trouble Brewing? Ethical Implications for the Information Systems Discipline. **Social Epistemology Review and Reply Collective** 4, no. 8, 9-33.

PUSHMANN, C. & BURGESS, J. (2014). Metaphors of Big Data. **International Journal of Communication**, 8, 2014, 1690–1709.

SCHEINES, R. (2002). Computation and Causation. In: Cyber Philosophy: The Intersection of Computing and Philosophy. Edited by James H. Moor and Terrell W. Bynum. Oxford, UK: Blackwell Publishing Ltd.

SEARLE, J. R. (2005). What is an institution? **Journal of Institutional Economics**, 1: 1, 2005, 1-22.

TIMMER, J. (2017). First the cloud, now AI takes on the scientific method. Cloud didn't make the scientific method irrelevant in '08 - AI won't do it in '17, either. **Ars Technic**, May 28, 2017. Available at: https://arstechnica.com/science/2017/05/why-the- cloud-cannot-obscure-the-scientific-method/. Accessed: 30-09-2018.

VIGEN, T. (2015). **Spurious Correlations**. New York: Hachette Books, 2015. Available at: http://tylervigen.com/spurious-correlations. Accessed: 03-09-2018.

WITTGENSTEIN, L. (2001 [1953]). **Philosophical Investigations**. Blackwell Publishing.

Acknowledgments

Thanks are due to Mariana Claudia Broens (UNESP), Guiou Kobayashi (UFABC), and Mariana Vitti Rodrigues (UNESP), who helped with the development of the ideas presented in this paper, and to colleagues from GAEC (UNESP) and CLE (UNICAMP), for their inspiring discussions. This research was supported by the Brazilian agencies FAPESP, CAPES, and CNPq.