Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-020-00638-0 OPEN

How ethics combine with big data: a bibliometric analysis

Marta Kuc-Czarnecka [1]✉ & Magdalena Olczyk [1]

The term Big Data is becoming increasingly widespread throughout the world, and its use is no longer limited to the IT industry, quantitative scientific research, and entrepreneurship, but entered as well everyday media and conversations. The prevalence of Big Data is simply a result of its usefulness in searching, downloading, collecting and processing massive datasets. It is therefore not surprising that the number of scientific articles devoted to this issue is increasing. However, the vast majority of research papers deal with purely technical matters. Yet, large datasets coupled with complex analytical algorithms pose the risk of non-transparency, unfairness, e.g., racial or class bias, cherry-picking of data, or even intentional misleading of public opinion, including policymakers, for example by tampering with the electoral process in the context of 'cyberwars'. Thus, this work implements a bibliometric analysis to investigate the development of ethical concerns in the field of Big Data. The investigation covers articles obtained from the Web of Science Core Collection Database (WoS) published between 1900 and July 2020. A sample size of 892 research papers was evaluated using HistCite and VOSviewer software. The results of this investigation shed light on the evolution of the junction of two concepts: ethics and Big Data. In particular, the study revealed the following array of findings: the topic is relatively poorly represented in the scientific literature with the relatively slow growth of interest. In addition, ethical issues in Big Data are discussed mainly in the field of health and technology.

[1] Gdansk University of Technology, 80-233 Gdansk, Poland. ✉email: marta.kuc@zie.pg.edu.pl

Introduction

The concept of Big Data has emerged in recent years and has become an active field of research with great interest from academics and practitioners. An extensive body of literature exists concerning the technical potential and challenges of Big Data, as with their increased volume, the velocity, variety, and veracity of data analysis become more sophisticated (Díaz et al., 2012; Michael and Miller, 2013; Hashem et al., 2015). Despite the apparent interest in the use of Big Data tools in the scientific literature and the well-established field of the ethics of technology, the two themes are not often combined in scientific research.

Figure 1 presents the number of newly published scientific articles containing, respectively, "Big Data" (red bar), "ethic*" (green bar), and "Big data and ethics" (purple bar) in the topic search in the WoS database search engine. Figure 1 shows data from 1993 as publications in the field of Big Data began to appear in this year. It is unequivocal that the number of papers dealing simultaneously with big data and ethics is a minor fraction of the overall discussion around Big Data. Are there indeed only a few ethical doubts appearing in the context of Big Data systems?

In our paper, we decided to narrow down the analysis to Big Data as a result of the specificity of this phenomenon. The clarification of 'Big Data' meaning must be related to the notion of Data Science. Traditionally, Data Science is a broad notion, which encompasses mathematics, computer science and relevant expertise in the application domain (health, policing, insurance, etc.). Data Science applies scientific methods, processes, and systems to extract knowledge or insights from data in various forms, either structured or unstructured (Ley and Bordas, 2018). Three main fields in Data Science can be distinguished (Song and Zhu, 2016):

- Data analytics—data is extracted and categorised to obtain some useful patterns and behavioural data.
- Machine-learning—focuses on the development of computer programs that can access data and use it to learn for themselves.
- Big Data—concentrates on mining of useful information from large volumes of datasets.

Thus, 'Big Data' is a term that describes the large volume of data—both structured and unstructured—that inundates a business on a day-to-day basis. But it is not the amount of data that matters—what is truly important is what organisations are doing with the data. Big Data can be analysed for insights that lead to improved decisions and strategic business moves. By employing Big Data, companies and organisations have ample information about the products, services, buyers, suppliers, consumer preferences, etc. that can be captured and analysed. Therefore, the central question is what ethical issues are associated with the use and analysis of Big Data.

In 2012, Boyd and Crawford (Boyd and Crawdord, 2012) claimed that "very little is understood about the ethical implications underpinning the Big Data phenomenon." The literature maps out several ethical dilemmas that evolve in the Big Data context, which are well summarised in the book of O'Neil (2016), from discriminating behaviour toward minorities and people living in a poor neighbourhood to abusive labour practices to the exploitation of consumers. A discussion of these dimensions is offered in other works of the present special issue, e.g., (Sareen, Rommetveit and Saltelli, 2020). Perhaps the most visible form of alarm against algorithms is their use in various types of cyberwarfare, which is often militarily directed and deployed against industrial and military infrastructures (Halpern, 2019). Also, drawing alarm is the use of social media to disrupt elections (McNamee, 2019), and to operate forms of ethical sabotage such as amplifying discord in social conflicts. No aspect of public life is spared, from political figures to the use of vaccines (Broniatowski et al., 2018), from gun controls and mass shootings to migration, and—at the time of writing this work, even the COVID-19 is caught in the crossfire (Rankin, 2020). The dangers to liberty in the form of digital dictatorship are among the challenges identified by historian Yuval N. Harari (Harari, 2018), while many fear the Big Data contribution to the deployment of autonomous lethal weapons. A race seems to be taking place between the scripts of dystopian science fiction—e.g., the series Black Mirror, and what happens in reality. In the age of Jules Vernes, fiction limited itself to anticipating technology in the coming few decades; now it becomes a reality during the period of crafting and producing the script, in a process which has been called "rapidification" (Pope Francis, 2015).

Due to the complex ethical concerns and high relevance of Big Data, it becomes increasingly difficult or even impossible to understand the overall structure and development of this field

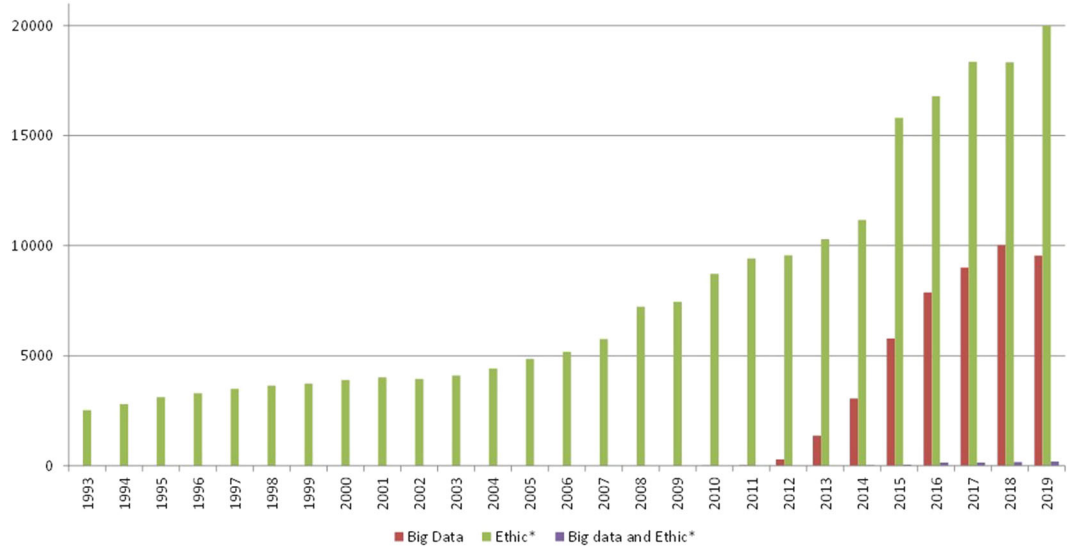

Fig. 1 Number of research papers on the topic of "Big Data", "Ethic*" and "Big Data and ethic*". Search on www.webofknowledge.com using the search string: TOPIC ("Big Data"); TOPIC (ethic*); TOPIC ("big data" AND ethic*) (July 6th, 2020).

without more in-depth analytical approaches. Not many papers have addressed the issue of the evolution of the concept of combining ethics with Big Data. That is the gap in the nascent literature that we aim to fill by providing extensive insights into publication patterns. According to the best knowledge of the authors, a study of ethics in Big Data using bibliometric methods has not yet been carried out. However, it was successfully used to analyse the relationship between ethics and entrepreneurship (Vallaster et al., 2019).

Methodology

Bibliometric analysis involving the application of mathematical and statistical methods to scholarly publications (Pritchard, 1969) is the cornerstone of modern literature research (Bornmann, 2017). It allows investigating knowledge structure, developing research fields, and capturing the interdisciplinarity of research topics (Reuters, 2008; Pauna et al., 2018; Zou et al., 2018). The goal of traditional citation analysis is to investigate two issues: (i) whether the two articles are connected through citations, (ii) and how many quotes an article has accrued. It is assumed that scientific impact is defined as the extent to which given research papers have been used by other researchers (Bornmann et al., 2008), so citation is taken as the main channel of communication between scientists. The number of quotations and average citations is often, though not always, (Osterloh and Frey, 2020), correlated with the quality and influence of scholars (Tang et al., 2018). However, it is also said that a high number of citations is a necessary, but not sufficient condition of 'being influential' (Small, 1978). Nevertheless, it should be noted that different scientific fields have different citation rates (Radicchi et al., 2008). Therefore measurements of performance based on citation count cannot be directly compared across various research fields.

In our research, we are focusing on a numerical feature of citation, i.e., we are assuming that research impact is not intangible, but measurable in a quantitative way (Zhang et al., 2013). We have decided to apply three different bibliometric methods to investigate the development of the relationship between ethics and Big Data: (i) descriptive analysis, (ii) network-citation analysis, and (iii) co-occurrence analysis.

The first approach concerns a descriptive analysis of fundamental indicators, such as the number of research papers over time, the number of global and local citations. The difference between local and global citations are expressed in the set from which quotes are counted. Local cited reference (LCR), shows the number of citations in a paper's reference list to other manuscripts within the created collection. In comparison, global citation score (GCL) presents the total number of citations to an article in the Web of Science Core Collection. Hence, in our study, we will focus on local citations, which should be understood as a contribution to the development of the field being analysed. To make an example, we are not interested in how many geographers mentioned in their study, research referring to the ethics of Big Data. In fact, we are interested in how many scientists writing about ethics in Big Data used a given article on this subject. Thus, a paper with a large number of global citations (GCL) that has reached many researchers from other fields, but has a low LCR indicates a small contribution to the development of the field related to the topic of the article.

In the second step, we employ network-citation analysis to disclose the relationship between the most-cited publications (Small, 1973). At this stage of our investigation, we are using HistCite software (2005) to generate a historiograph—which is a graphical representation of the network between the most-cited works (based on LCR indicator). In a historiograph, the vertical axis represents time, and the horizontal axis shows citation network nodes. Each node refers to a single research paper having its unique number, while the size of the node reflects the number of citations in the local database. The arrows express the relationship between cited publication—from the analysed manuscript to the previously published one. Moreover, this visualisation allows us to present the timeline of publications under consideration. According to Griffith (Griffith et al., 1974), the top forty research items with the highest number of citations are the optimal number to create the historiograph.

In the third step, we use VOSviewer software (Van Eck and Waltman, 2009) to conduct co-occurrence term analysis to ascertain trends and to identify "hotspots" domains (Cho and Khang, 2006; Williams and Plouffe, 2007). The co-occurrence method measures the distance between two terms. The more often two phrases co-occur in the same line of text, the smaller the distance between them. VOSviewer applies a natural language processing algorithm (NLP) (Van Eck and Waltman, 2011) to identify the strength of association among noun phrases[1]. The software creates a distribution function for each second-ordered phrase and compares it with the overall distribution function of co-occurrences over noun phrases (Van Eck and Waltman, 2010). The lower the distance between phrases in a semantic context, the higher association strength is expected. Based on the word count and association strength, VOSviewer creates a co-occurrence map, allowing us to distinguish main clusters characterised by strong association. To construct the map, VOSviewer uses the SMACOF algorithm (Borg and Groenen, 2005), which minimises the function:

$$V(X_1 \ldots \ldots \ldots X_n) = \sum_{i<j} S_{ij} ||\boldsymbol{X_i} - \boldsymbol{X_j}||^2$$

under the constraints:

$$\frac{2}{n(n-1)} \sum_{i<j} ||\boldsymbol{X_i} - \boldsymbol{X_j}|| = 1$$

where:

n–the number of nodes in a network,

X_i–the locations of node i in a two-dimensional space,

$||X_i - X_j||$–the Euclidean distance between nodes i and j.

VOSviewer builds clusters of nodes by maximising the following function:

$$V(c_1 \ldots c_n) = \sum_{i<j} \delta(c_i, c_j)(s_{ij} - \gamma)$$

where:

c_i–the cluster to which node i is assigned,

$\delta(c_1, c_j)$–a function that equals one if $c_i = c_j$; and zero otherwise,

γ–a resolution parameter that determines the level of detail of the clustering (the higher γ is, the higher the number of clusters).

Although there is a significant overlap between the content of Scopus and WOS databases (Norris and Oppenheim, 2007), we have decided to use the Web of Science Core Collection Database (WoS) because it does not have the following disadvantages of Scopus databases. First, in Scopus, the citation matching algorithm seems to need improvement (Valderrama-Zurián et al., 2015). Second, duplicate publications in Scopus represent a vital data quality problem that requires serious attention (Van Eck and Waltman, 2017).

As fas as WoS is concerned, a general limitation is a fact that its coverage in the social sciences and humanities is still limited (Mingers and Leydesdorff, 2015). It is connected with the relatively small coverage of book publications, despite the fact that during the last five years, the number of indexed books has been increasing. Also, non-English language journals are underrepresented in the WoS database. Despite this, the Web of Science

Table 1 Principal bibliometric indicators in the WoS database.

Period: 1900–July 2020		
Number of records: 892	Number of authors: 2696	Number of journals: 591
Number of countries: 75	Number of institutions: 1257	Number of languages: 19
Total local citations (LCS): 798	Total global citations (GCS): 8621	

Source: Bibliometric data from the Web of Science Core Collection retrieved on July 6th, 2020.

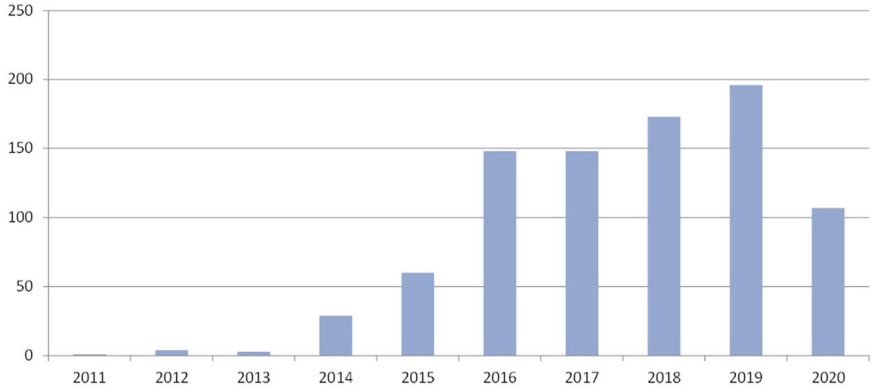

Fig. 2 Scientific productivity on ethics in Big Data over the period 2012–2019 based on data taken from the WoS database. Source: Authors' calculation based on the local database.

Core database is regarded by scholars—e.g., by (Byl et al., 2016), as a suitable tool for bibliometric evaluation.

Empirical results

In the initial stage of our research, we surveyed the literature on ethics in Big Data. Documents were collected on July 6th, 2020, by research on the web search engine Web of Science Core Collection. We searched for the topic: "Big Data" and "Ethic*" in all categories in the period, 1900–2020. In the WoS database, the fields mined to return results in a common 'topic search' are:

- The title of the article, review, proceedings, book, etc.
- The abstract—which is the work's summary containing the key points discussed, such as research question, methodology, discussion, and conclusion. This field is supplied by the author(s) of the paper.
- The keywords and keywords plus fields: The keywords field is the one supplied by the author(s) and "tags" the main and sub-topics of the paper's content. The keywords plus field is an algorithm that provides expanded terms stemming from the record's cited references or bibliography.

The total number of obtained documents was 892 (Table 1). Evaluation of data was conducted with the use of bibliometric software, HistCite, and VOSviewer. Based on the collected information, we aim to show where the topic of ethics in Big Data began and identify primordial papers and authors. Basic statistics referring to the created local bibliometric database are presented in Table 1.

The analysis of 892 records showed a substantial dispersion of publications measured as the ratio of the number of articles per one journal (i.e., on average each journal in the database was represented by 1.5 scientific papers about the topic under consideration). A moderate concentration was observed in the relationship of the average number of authors per journal (on average, 4.5 authors per journal). There is a substantial difference in the global and local number of citations, as selected publications were cited 8621 times in the whole WoS database, while only 798 times among the database created for the study. It can be assumed that researchers from fields other than data science were more willing to use work-related to ethics in Big Data. Thus the field itself was developing quite slowly (a relatively small number of connections in the local database).

The distribution over time of analysed publications is presented in Fig. 2. The pioneering work in the context of ethics in Big Data was an article published in 2011 by Helbring and Balietti in which one of the goals was to "elaborate ethical standards regarding the storage, processing, evaluation, and publication of social and economic data" (Helbring and Balietii, 2011). However, Danah Boyd and Kate Crawford are considered the mothers of the field. In their paper from 2012, they raised the issue of data privacy in social media and the issue of ignoring research ethics because "data is seemingly public" (Boyd and Crawdord, 2012). In the same year, three other publications that met the criteria of our search were published, but they did not achieve much success as measured by the number of citations. Within a year, three more articles were published that drew attention to the usage of online data for social research (Loader and Dutton, 2012; Wright, 2012; Nunan and Di Domenico, 2013). In subsequent years, a slow increase in interest in research on ethical issues in Big Data can be seen. So far, the peak of interest is in 2018 and 2019, during which years 173 and 196 scientific articles were published, respectively. As of July 6th 2020, 107 research papers on this subject have been published, which indicates that research on ethics in Big Data is still slowly entering the field of scientific research. Taking into account the current epidemic and voices about the questionable reliability of some publications on COVID-19, a sharp increase in research on ethics in Big Data can be expected.

The importance of individual authors for the development of research on ethics in Big Data can be assessed based on the number of citations of their publications in the created database

Table 2 Ranking of authors with the highest number of local citations (LCS).

Rank	Author	Number of local citations (LCS)	Number of publications (Q)	LCS/Q
1	Crawford K	181	7	25.9
2	Boyd D	160	2	80.0
3	Floridi L	90	5	18.0
4	Mittelstadt BD	89	5	17.8
5	Vayena E	25	11	2.7
6	Allo P	21	1	21.0
7	Di Domenico M	21	3	7.0
8	Nunan D	21	3	7.0
9	Taddeo M	21	1	21.0
10	Wachter S	21	2	10.5

Source: Authors' calculation based on the local database.

(LCS). Table 2 presents a list of the ten most frequently cited authors. Crawford, with 181 citations of her works in the local database, is the leader in the created ranking. The second place went to Boyd, whose work was cited 160 times in the analysed database. However, it is worth mentioning here that Boyd is co-author of Crawford's two works. Crowford, with her seven publications, is the second most productive author within the created database. This qualification of most prolific author belongs to Vayena, with her eleven manuscripts published in the field of ethics in Big Data. Analysing the data contained in Table 2, one can see, that the ratio of the number of local citations (LCS) to the number of publications (Q) for each of the authors is not particularly high, which may indicate that researches on this subject are so far at the initial stages. Interestingly, if one adds up the citation per-gender, one finds that female scholars total 429 citations against male scholars' 211, a rare partition in a usually male-dominated academy.

We also compiled a rank of the most frequently cited publications in the field of ethics in Big Data (Table 3). It is worth reminding that the generated database only contains scientific articles and conference publications, with only some books, book chapters and reports.

The most significant publication in the evolution of ethics in Big Data is the previously mentioned work of Boyd and Crawford (2012) entitled "Critical questions for big data. Provocations for a cultural, technological, and scholarly phenomenon". This research paper is focusing on ethical problems concerning data privacy in social media and the problem of a lack of understanding of ethical boards with respect to "the processes of mining and anonymising Big Data" (Boyd and Crawdord, 2012). The second most cited work is the paper prepared by Mittelstadt and Floridi, "The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts" (Mittelstadt and Floridi, 2015). In this paper, as the title suggests, the authors investigate a biomedical context, not ignoring such issues as privacy, ownership or epistemology, and objectivity. They are also noticing that "as is often the case with the cutting edge of scientific and technological progress, understanding of the ethical implications of Big Data lags behind" (Mittelstadt and Floridi, 2015). The third article with the highest LCS is "The ethics of algorithms: Mapping the debate" which clarifies the ethical importance of algorithmic mediation (Mittelstadt et al., 2016). A thought-provoking element of this article is a conceptual map of the ethics of algorithms, allowing more rigorous diagnosis of algorithms' ethical challenges. Among normative concerns authors distinguish unfair outcomes and transformative effects, while epistemic considerations consist of: inconclusive, inscrutable and misguided evidence. In turn, the

Table 3 The main papers related to ethics and Big Data.

Rank	Authors	Title of the publication	Year	LCS	Node no.
1	Boyd D, Crawford K	Critical questions for big data. provocations for a cultural, technological and scholarly phenomenon	2012	160	27
2	Mittelstadt BD, Floridi L	The ethics of big data: current and foreseeable issues in biomedical contexts	2015	61	201
3	Mittelstadt BD., Allo P, Taddeo M, Wachter S, Floridi L	The ethics of algorithms: mapping the debate	2016	21	257
4	Fairfield J, Shtein H	Big data, big problems: emerging issues in the ethics of data science and journalism	2014	18	45
5	Nunan D, Di Domenico M	Market research and the ethics of big data	2015	17	29
6	Crawford K, Miltner K, Gray ML	Critiquing big data: politics, ethics, epistemology special section introduction	2014	17	39
7	Lupton D	The commodification of patient opinion: the digital patient experience economy in the age of big data	2014	15	53
8	Martin KE	Ethical issues in the big data industry	2015	15	93
9	Cohen IG, Amarasingham R, Shah A, Xie B, Lo B	The legal and ethical concerns that arise from using complex predictive analytics in health care	2014	13	51
10	Markowetz A, Blaszkiewicz K, Montag C, Switala C, Schlaepfer TE	Psycho-informatics: big data shaping modern psychometrics	2014	10	48

Source: Authors' calculation based on the local database.

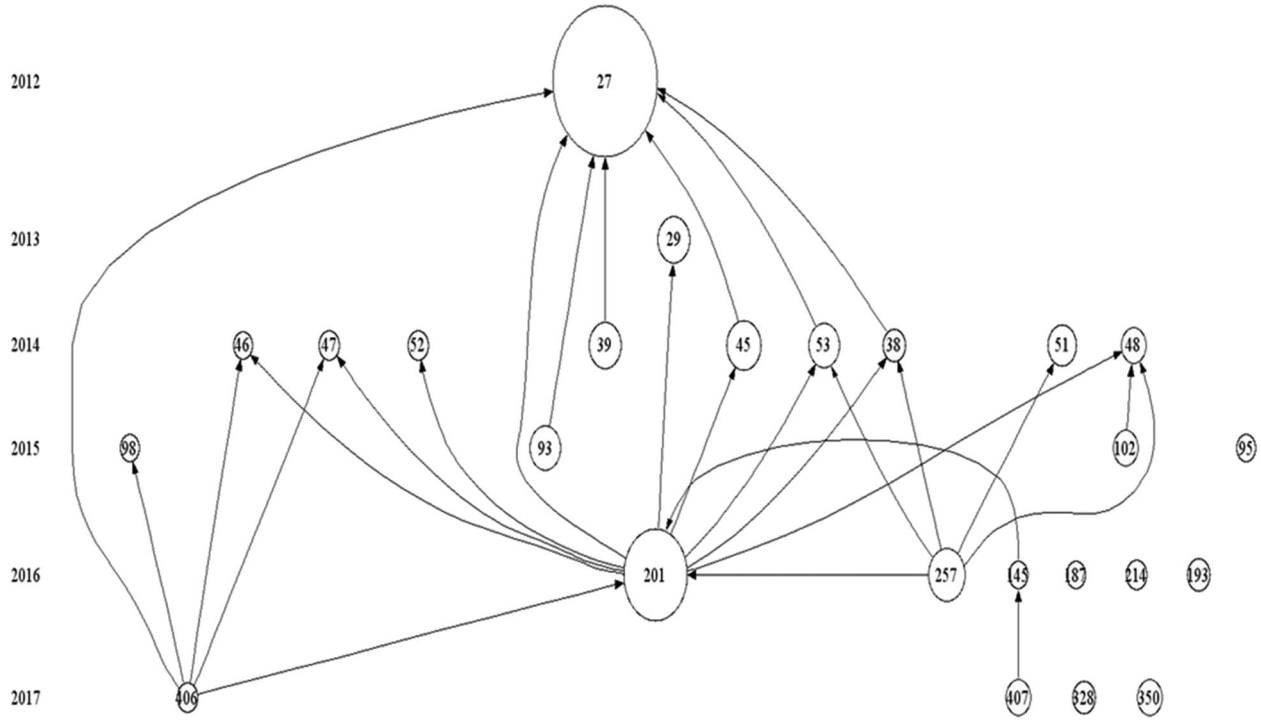

Fig. 3 Historiograph for the 25 most highly cited research papers in the local database. Source: Authors' investigation based on the local database.

fourth most cited publication, "Big Data, Big Problems: Emerging Issues in the Ethics of Data Science and Journalism," pointed out the uncertain status of data collected through telemetry or public submission (Fairfield and Shtein, 2014). Authors indicate the growing ethical problems of media and research using big data techniques, clearly observable now in the era of the COVID-19 pandemic and the spreading of dubiously ethical studies. The main conclusion from their paper focuses on the need to use the framework combining stability with flexibility, as the best way to achieve the original purpose of fundamental ethical principles.

The content of Tables 2 and 3 is, of course, correlated, i.e., the most cited researchers are the authors of the most important publications in the field of the discussed issue. It may be surprising that none of the works by Vayena, the most prolific author, is included in Table 3. However, it should be noted that despite publishing 11 research papers on the analysed topic, they were cited only 25 times, which is an average of 2.7 per manuscript (LCS/Q score). Therefore, none of them managed to impact significantly the development of the field being analysed.

The crucial part of this analysis is not to identify the most frequently cited publications but to establish a network of connections between them. Thus, using the HistCite software, we have prepared a historiograph (Fig. 3) involving, typically, around 5% of the publications that are the most-cited in the local database (Garfield et al., 2003). As can be seen in Fig. 3, the biggest node (no. 27) represents the very first publication of Boyd and Crawford (Boyd and Crawdord, 2012). Their research turned out to be innovative and groundbreaking enough to contribute to further ethical considerations in the context of Big Data. In principle, all subsequent publications relate directly or indirectly to this particular study. For example, the work of Vayena (node no. 145) refers to the work of Mittelstadt and Floridi (node no. 201) that was inspired by Boyd and Crawford (node no. 27).

It is not surprising that the second-largest node (no. 201) refers to the second work in terms of citability, which is Mittelstadt and Floridi (2015). They refer to the first publication and become an essential source of inspiration for papers issued after 2016. Interestingly, none of the research published between 2013 and 2015 gained as much popularity as the work of Mittelstadt and Floridi (2015). The success of this work probably results from the authors' explicit embedding of ethics in the context of biomedical research. The vast majority of later publications contain references to the two articles mentioned above. It is also worth noting that none of the papers published in 2018 or later is included in the top most-cited list. The hypothesis arises that ethical issues in Big Data relate more to the biological and medical sciences than other disciplines.

At the last stage of our analysis, we use the co-occurrence map, which helps to identify the various areas of research and understand the direction in which the ethics combine with Big Data. We used information included in the title, abstract, and keywords as term sources obtaining 156555 unique terms extracted from the local database. We applied the text mining functionality of the VOSviewer to identify the noun phrases in the text, and then to convert all plural noun phrases into singular ones. A minimum number of occurrences was assumed as 20, so 172 terms met the threshold. For those 172 words, a relevance score was calculated by VOSviewer, and then we selected the 60% most important phrases. Finally, we ended up with 103 terms, from which we excluded terms not germane to analysis goals such as specific place names, general statistical terms or measures reflecting such things as time, quantity, and rate. The same VOSviewer software was also used to construct a bibliometric diagram visualising the co-occurrence of the extracted texts. Figure 4 presents the co-occurrence term map. Each term is represented by a blurred circle, where the size of the label represents the term's frequency; the colour characterises the cluster to which it conceptually belongs, and proximity to another phrase indicates the degree of relatedness between them. The analysis of Fig. 4 showed science (the biggest font size) as the most frequently mentioned phrase followed by the words: health, medicine, governance, artificial intelligence, and knowledge.

There are three clusters in Fig. 4. The red cluster can be called the 'legal cluster.' This cluster groups terms associated with

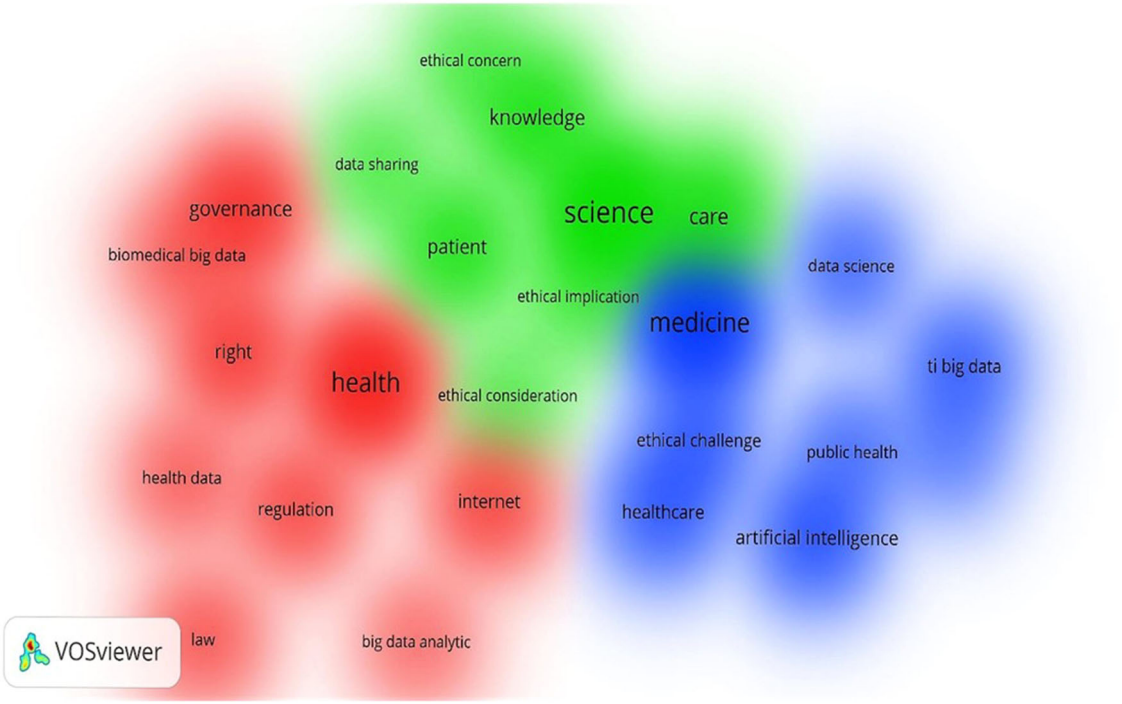

Fig. 4 Clusters in the ethics in Big Data literature by term co-occurrence analysis. Source: Authors' investigation based on the local database.

governance, regulation, law, and rights concerning gathering health and biomedical data, but also the ethical issues of obtaining private data on the Internet. The green cluster, which can be called the 'scientific cluster,' shows the ethical concerns and implications in data sharing and access to knowledge and research results. The blue cluster, named the 'medical cluster,' points to the importance of ethics in medicine, healthcare, and artificial intelligence. As was shown in the previous part of the analysis, ethics in the biomedical context are one of the biggest worries in the implementation of Big Data analysis. All clusters are located close to each other, proving a strong relationship between the topics covered within each group. One may even be tempted to say that the main phrases in clusters are located on their borderlands, demonstrating the interpenetration of the discussed phenomena. In fact, the subject of obtaining and processing medical data is the most pressing ethical issue related to Big Data, and references to this topic are undoubtedly visible in each of the clusters.

Conclusion

Big Data is a rapidly developing research area that attracts a lot of interdisciplinary attention, including on the ethical issues which arise in the course of the implementation of this new technology. The results of this study reveal that the current studies about ethics and Big Data are dominated by Boyd, Crawford, Mittelstadt, and Floridi, and that the thematic scope itself mainly relates to health and medical issues. It seems that these trends will also be maintained in the time of the COVID-19 outbreak as many ethical questions related to tracking the spread of the virus are raised (Jamrozik and Selgelid, 2020; Robert et al., 2020; WHO, 2020). In this particularly difficult period, attention is being paid to the issue of individual freedom, both in terms of traceability of movement and social networks, but also in terms of voluntariness of vaccination. Though not covered by the present investigation, growing ethical attention is focusing on so-called "challenge study," in which healthy subjects are given a prospective vaccine and then infected with the coronavirus (Elliot, 2020), and on the fact that participants in medical research studies such as these are often minorities or ex-detainees. Closer to the topic of this work, the issue of the ethics of Big Data also comes into play in the issue of contact-tracing applications for fighting the pandemic. While these were apparently a success in some countries (Holmes, 2020), they were less so in others, while the concerns about the privacy and security risks of the technologies let to an intense ethical debate (Singer, 2020). Ethical assessments of the potential benefits and risks of each action should be made in light of the best available empirical data and models. The expected harms and benefits of different proposed research programmes concerning not only COVID-19 but also all other areas, should be taken into consideration, and a nascent debate has sprung out about what numbers are being used to decide what policies to fight the pandemic (Caduff, 2020; Didier, 2020), a topic which relates to the ethics of quantification (Saltelli, 2020) and to the present special issue (Sareen et al., 2020; Saltelli and Di Fiore, 2020).

We figured out that there is a lack of well-recognised literature on ethical issues in Big Data related to micro and macroeconomic, political and sociological analyses. To make an example, in 892 papers reviewed, only 12 are from Economics, 42 from Management, 38 form Business, 27 from Sociology and 20 from Political Science. The small share of papers related to ethics in Big Data in the total number of published scientific research is striking. The individual works appearing in those topics are still somewhat limited, fringe research area. However, we realise that the contextual and multi-level phenomenon of ethic and Big Data is a demanding research area, requiring extensive knowledge, both philosophical and purely technical. These factors may contribute to the relatively low popularity of the issue raised in the scientific literature. In the case of medical research, the subject is also industry-specific. The enormous emphasis on ethical issues in medical sciences results mostly from working on sensitive data, but also the perilous consequences of unreliable studies, e.g., linking autism with vaccinations.

Despite the restrictions arising from the very nature of bibliometric research and the database used (including only some books and book chapters), the analysis allowed us to reconstruct

the effects of scientific productivity in terms of concreteness in historical terms. Our main contributions in this work are the analysis of statistical patterns and the provision of an informative overview of the different contexts and intersections between ethics and Big Data—at a moment where the field is likely to experience transformation and accelerations.

Data availability

The datasets analysed during this study are available in the Harvard Dataverse repository: https://doi.org/10.7910/DVN/RU8KTN Ethics and Big Data-bibliometric analysis.

Received: 7 April 2020; Accepted: 20 October 2020;
Published online: 04 November 2020

Note

1 Noun phrase consists of a head, which is typically a noun, and of elements which (either obligatorily or optionally) determine the head and (optionally) modify the head, or complement another element in the phrase. Noun phrase 'consists of a noun and all the words and word groups that belong with the noun and cluster around it' Stagaberg (1979).

References

Broniatowski DA, Jamison AM, Qi S, AlKulaib L, Chen T, Benton A, Quinn SC, Dredze M (2018) Weaponized health communication: twitter bots and russian trolls amplify the vaccine debate. Am J Public Health 108 (10):1378–1384. https://doi.org/10.2105/AJPH.2018.304567

Borg I, Groenen JP (2005) Modern multidimensional scaling, 2nd edn. Springer, New York

Boyd D, Crawford K (2012) Critical questions for Big Data Provocations for a cultural, technological, and scholarly phenomenon. Inform Commun Soc 15 (5):662–679. https://doi.org/10.1080/1369118X.2012.678878

Bornmann L, Mutz R, Neuhaus C, Daniel H-D (2008) Citation counts for research evaluation: standards of good practice for analysing bibliometric data and presenting and interpreting results. Ethics Sci Environ Politics 8:93–102. https://doi.org/10.3354/esep00084

Bornmann L (2017) Measuring impact in research evaluations: a thorough discussion of methods for, effects of, and problems with impact measurements. Higher Educ 73(5):775–787. https://doi.org/10.1007/s10734-016-9995-x

Byl L, Carson J, Feltracco A, Gooch S, Gordon S, Kenyon T, Muirhead B, Seskar-Hencic D, MacDonald K, Tamer Özsu M, Stirling P (2016) White Paper: Measuring Research Outputs Through Bibliometrics. UWSpace. http://hdl.handle.net/10012/10323

Caduff C (2020) What went wrong: Corona and the world after the full stop. Med Anthropol Q

Cho CH, Khang HK (2006) The state of internet-related research in communications, marketing, and advertising: 1994-2003. J Advert 35(3):143–163. https://doi.org/10.2753/JOA0091-3367350309

Díaz M, Juan G, Lucas O, Ryuga A (2012) Big data on the internet of things: an example for the E-health. Sixth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, Palermo, pp. 898–900

Didier E (2020) Politique du nombre de morts. AOC, Analyse Opinion Critique

Elliott C (2020) An ethical path to a covid vaccine. The New York Review of Books

Fairfield J, Shtein H (2014) Big data, big problems: emerging issues in the ethics of data science and journalism. J Mass Media Ethics 29(1):38–51. https://doi.org/10.1080/08900523.2014.863126

Garfield E, Pudovkin AI, Istomin VS (2003) Why do we need algorithmic historiography? J Am Soc Inform Sci Technol 54(5):400–412. https://doi.org/10.1002/asi.10226.

Griffith BC, Small HG, Stonehill JA, Dey S (1974) The structure of scientific literatures II: toward a macro- and microstructure for science. Sci Stud 4 (4):339–365. https://doi.org/10.1177/030631277400400402

Halpern S (2019) The drums of cyberwar. The New York Review of Books

Harari YN (2018) 21 lessons for the 21st century. Spiegel & Grau

Hashem IAT, Yaqoob I, Badrul Anuar N, Mokhtar S, Gani A, Khan SU (2015) The rise of "big data" on cloud computing: review and open research issues. Information Systems 47:98–115. https://doi.org/10.1016/j.is.2014.07.006

Helbring D, Balietii S (2011) From social data mining to forecasting socio-economic crises. Eur Phys J Special Topic 195(1):3–68. https://doi.org/10.1140/epjst/e2011-01401-8

HistCite (2005) Bibliographic Analysis and Visualization Software. http://garfield.library.upenn.edu/histcomp/

Holmes A (2020) How South Korea has used tech to successfully contain COVID-19 Business Insider

Jamrozik E, Selgelid MJ (2020) COVID-19 human challenge studies: ethical issues, www.thelancet.com/infection Published online May 29, 2020, 016/S1473-3099(20)30438-2

Ley C, Bordas (2018) What makes Data Science different? A discussion involving Statistics2.0 and Computational Sciences. Int J Data Sci Anal 6:167–175

Loader BD, Dutton WH (2012) A decade in internet time. Inform Commun Soc 15 (5):609–615. https://doi.org/10.1080/1369118X.2012.677053

McNamee R (2019) Zucked: waking up to the Facebook catastrophe. Penguin Press

Michael K, Miller KW (2013) Big data: new opportunities and new challenges [Guest editors' introduction]. Computer 46(6):22–24. https://doi.org/10.1109/MC.2013.196

Mingers J, Leydesdorff L (2015) A review of theory and practice in scientometrics. Eur J Operat Res https://doi.org/10.1016/j.ejor.2015.04.002.

Mittelstadt BD, Allo P, Taddeo P, Wachter S, Floridi L (2016) The ethics of algorithms: Mapping the debate, Big Data Soc 1–21, https://doi.org/10.1177/205395171667967

Mittelstadt BD, Floridi L (2015) The ethics of big data: current and foreseeable issues in biomedical contexts. Science and Engineering Ethics 22(2):303–41. https://doi.org/10.1007/s11948-015-9652-2

Norris M, Oppenheim C (2007) Comparing alternatives to the Web of Science for coverage of the social sciences' literature. J Informetr 1(2):161–169. https://doi.org/10.1016/j.joi.2006.12.001

Nunan D, Di Domenico M (2013) Market research & the ethics of big data. Int J Market Res 55(4):505–520. https://doi.org/10.2501/IJMR-2013-015

O'Neil C (2016) Weapons of math destruction: how big data increases inequality and threatens democracy. Random House Publishing Group

Osterloh M, Frey BS (2020) How to avoid borrowed plumes in academia. Res Policy 49(1):103831. https://doi.org/10.1016/j.respol.2019.103831

Pauna VH, Picone F, Le Guyader G, Buonocore E, Franseze PP (2018) The scientific research on ecosystem services: a bibliometric analysis. Ecol Quest 29 (3):53–62. https://doi.org/10.12775/EQ.2018.022

Pope Francis, "Laudato si" (2015). [Online] http://w2.vatican.va/content/francesco/en/encyclicals/documents/papa-francesco_20150524_enciclica-laudato-si.html. [Accessed: 11 May 2018]

Pritchard A (1969) Statistical bibliography or bibliometrics? J Document 25 (4):348–349

Radicchi F, Fortunato S, Castellano C (2008) Universality of citation distributions: toward an objective measure of scientific impact. Proc Natl Acad Sci USA 105 (45):17268–17272

Rankin J (18 Mar, 2020) "Russian media' spreading Covid-19 disinformation." *The Guardian*

Reuters T (2008) Whitepaper Using Bibliometrics: Thomson Reuters, 12

Robert R, Kentish-Barnes N, Boyer A, Laurent A. Azoluay E, Reignier J (2020) Ethical dilemmas due to the Covid-19 pandemic. Ann Intens Care 10(84) https://doi.org/10.1186/s13613-020-00702-7

Saltelli A (2020) Ethics of quantification or quantification of ethics?, FUTURES Vol. 116, February 2020, 102509

Saltelli A, Di Fiore M (2020) From sociology of quantification to ethics of quantification. Humanit Soc Sci Commun 7:1–8

Sareen S, Rommetveit K, Saltelli A (2020) Ethics of quantification: illumination, obfuscation and performative legitimation. Humanit Soc Sci Commun 6:1–5

Singer N (8 Jul, 2020) "Virus-Tracing Apps Are Rife With Problems. Governments Are Rushing to Fix Them." *The New York Times*

Small H (1973) Co-citation in the scientific literature: a new measure of the relationship between two documents. Journal of the Am So Inform Sci 2 (4):265–269. https://doi.org/10.1002/asi.4630240406

Small H (1978) Cited documents as concept symbols. Soc Stud Sci 8:327–340. https://doi.org/10.1177/030631277800800305

Tang M, Liao H, Wan Z, Herrera-Viedma E, Rosen MA (2018) Ten years of sustainability (2009 to 2018): a bibliometric overview. Sustainability 10 (5):1655. https://doi.org/10.3390/su10051655

World Health Organisation (2020) Ethical considerations to guide the use of digital proximity tracking technologies for COVID-19 contact tracing. http://www.WHO/2019-nCoV/Ethics_Contact_tracing_apps/2020.1

Williams BC, Plouffe CR (2007) Assessing the evolution of sales knowledge: a 20-year content analysis. Industr Market Manag 36(4):408–419. https://doi.org/10.1016/j.indmarman.2005.11.003

Wright DJ (2012) Theory and application in a post-GISystems world. Int J Geogr Inform Sci 26(12 Dec):2197–2209. https://doi.org/10.1080/13658816.2012.713957

Valderrama-Zurián JC, Aguilar-Moya R, Melero-Fuentes D, Aleixandre-Benavent R (2015) A systematic analysis of duplicate records in Scopus. J Informetr 9 (3):570–576. https://doi.org/10.1016/j.joi.2015.05.002

Vallaster C, Kraus S, Merigó Lindahl JM, Nielsen A (2019) Ethics and entrepreneurship: a bibliometric study and literature review. J Business Res 99:226–237. https://doi.org/10.1016/j.jbusres.2019.02.050

Van Eck NJ, Waltman L (2009) Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics 84:523–538. https://doi.org/10.1007/s11192-009-0146-3

Van Eck NJ, Waltman L (2010) Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics 84(2):523–538. https://doi.org/10.1007/s11192-009-0146-3

Van Eck NJ, Waltman L (2011) Text mining and visualisation using VOSviewer. ISSI Newsletter 7(3):50–54

Van Eck NJ, Waltman L (2017) Accuracy of citation data in Web of Science and Scopus. In: Proceedings of the 16th International Conference of the International Society for Scientometrics and Informetrics. ISSI, Wuhan Univeristy, China, pp. 1087–1092

Song I, Zhu Y (2016) Big data and data science: what should we teach. Expert Syst 33(4):364–373

Zhang G, Ding Y, Milojević S (2013) Citation content analysis (CCA): a framework for syntactic and semantic analysis of citation content. J Am Soc Inform Sci Technol 64(7):1490–1503. https://doi.org/10.1002/asi.22850

Zou X, Long W, Le H (2018) Visualisation and analysis of mapping knowledge domain of road safety studies. Accident Anal Prevent 118:131–145. https://doi.org/10.1016/j.aap.2018.06.010

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1057/s41599-020-00638-0.

Correspondence and requests for materials should be addressed to M.K.-C.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.