social sciences

MDPI

Communication

Big Data: Some Ethical Concerns for the Social Sciences

Michael Weinhardt

Institute of Sociology, Technische Universität Berlin, Straße des 17. Juni 135, 10623 Berlin, Germany; michael.weinhardt@tu-berlin.de

Abstract: While big data (BD) has been around for a while now, the social sciences have been comparatively cautious in its adoption for research purposes. This article briefly discusses the scope and variety of BD, and its research potential and ethical implications for the social sciences and sociology, which derive from these characteristics. For example, BD allows for the analysis of actual (online) behavior and the analysis of networks on a grand scale. The sheer volume and variety of data allow for the detection of rare patterns and behaviors that would otherwise go unnoticed. However, there are also a range of ethical issues of BD that need consideration. These entail, amongst others, the imperative for documentation and dissemination of methods, data, and results, the problems of anonymization and re-identification, and the questions surrounding the ability of stakeholders in big data research and institutionalized bodies to handle ethical issues. There are also grave risks involved in the (mis)use of BD, as it holds great value for companies, criminals, and state actors alike. The article concludes that BD holds great potential for the social sciences, but that there are still a range of practical and ethical issues that need addressing.

Keywords: research ethics; online social research; digital trace data; data privacy; digital ethics

Citation: Weinhardt, Michael. 2021. Big Data: Some Ethical Concerns for the Social Sciences. *Social Sciences* 10: 36. https://doi.org/10.3390/socsci10020036

Academic Editor: Nigel Parton
Received: 27 October 2020
Accepted: 19 January 2021
Published: 24 January 2021

1. Introduction

Many aspects of everyday life are moving into the digital sphere and becoming more reliant on the digitalization of society (Marres 2017). The digital traces and footprints human beings leave behind allow for the recording of their activities in the constantly growing realm of big data (BD) (Lazer and Radford 2017). New analytic tools for large-scale data analysis such as artificial intelligence and machine learning allow us to extract information from data where previously nothing of value could be found. In combination with these new statistical modeling techniques, BD may enable advances in many areas as practically important such as the detection of cancer in patients from biometric data (Barrett et al. 2013). Large amounts of data are especially useful for making predictions, be it forecasting the weather, epidemics, or road traffic (Askitas and Zimmermann 2015; Lazer et al. 2009).

BD may be described along the three "classical" characteristics of volume, variety, and velocity (Salganik 2018; Dumbill 2012), but also has value as a fourth characteristic (Gantz and Reinsel 2011). Regarding *volume*, there has been a vast increase in the amount of data that has become available electronically over the past two decades due to the rise of personal use of information technology and, especially, the internet. In social media, for example, Facebook and Instagram each have more than two billion active users, and Twitter, a micro-blogging service, has over 300 million. All users produce content by posting text, pictures, and videos. By interacting with other users' content, connections and relationships are established, again stored as data. Search histories and browsing behavior leave digital traces in the online world, again becoming data of its own kind, potentially providing valuable insights into attitudes and behaviors of those online. The rise of smartphones has added to this development even further. This vast volume of data which reflects our lives online, but also offline, is accompanied by a rapid speed of how these data are stored, managed, and used for other purposes. Accordingly, the

high-*velocity* aspect of BD is a result of the online nature of the aforementioned services, providing instant access to people all around the world at any time. To be able to provide a smooth user experience, the processing of large amounts of data in real-time is necessary. In addition, for commercial agents, speed in transmitting information and analyzing patterns often is a key advantage over competitors (e.g., in stock market trading), making BD the basis of a new wave of business models (Hartmann et al. 2016). Managing such volumes of data while people have access to them constantly all the time requires hardware and software resources on a grand scale. Accordingly, "'Big Data' originally meant the volume of data that could not be processed (efficiently) by traditional database methods and tools" (Kaisler et al. 2013). This makes BD a relational concept whose understanding depends on the state of software and hardware development.

BD is not a singular entity. Rather, it comes in many different forms and flavors and encompasses data from a wide *variety* of sources such as internet connections, search engines, website usage, chat forums, email and messaging services, wearables such as fitness tracking devices, but also video portals, digital libraries, and the simple everyday usage of computers, smartphones, tablets, and internet browsers. As Lazer and Radford (2017) observe: "There are many discrete literatures around different BD sources, and even a complete list of those literatures would soon be obsolete." To conclude, the variety of BD is huge and may even be growing.

BD has become somewhat of a buzzword in many areas of society as well as in the natural sciences. Often it is called the "oil of the 21st century" (Rotella 2012) to describe its importance as a raw material and a basic element of the current economy. In contrast, social scientists and sociologists have been comparatively slow in using BD for their research: "The large majority of sociologically relevant analysis of big data is done by computer scientists, and there is relatively little reflection of the big data revolution in top sociology journals. For example, only 6 of 182 articles published between 2012 and 2016 in the American Journal of Sociology (AJS) and 9 of 240 in American Sociological Review (ASR) involve the use of big data" (Lazer and Radford 2017). This has changed somewhat in recent years, however, as there are increasing activities to support research in this area, such as special issues in journals, workshops, tutorials, summer schools, and conferences. The first textbooks on the matter have appeared, as well as specialized journals and even whole research institutes dedicated to BD in the social sciences (Salganik 2018; Veltri 2020; Marres 2017; Foster et al. 2017). All this activity will likely result in a steep upturn in research in this area and increase the usage of web-sourced data for social research.

While data from online sources share many aspects with traditional data sources such as survey data or social book-keeping data (Graeff and Baur 2020), they also exhibit features that are new, if not in principle, then at least in the scope of its occurrence. The variety of different data forms, such as text, pictures, and videos associated with tags, geo-codes, paradata, and metadata, are part of the reason why such data appear so appealing to social researchers (Evans and Foster 2019). They offer new opportunities to address research questions they have not been able to answer previously (Ruths and Pfeffer 2014). For example, BD often is the documentation of actual (online) behavior rather than just the (self-) representation of such behavior as collected through interviews or similar methods. Quite often, this is what social scientists strive to study and to explain: social interactions and behavior. For example, wearables such as fitness trackers allow the tracking of one's physical condition much more accurately and, most importantly, are much easier and more cost-effective than questionnaires or interviews. However, these new data forms pose new questions in the handling and analysis of these kinds of data (Shah et al. 2015). One of these questions concerns the ethical treatment of such data: whether we need new rules and guidelines for conducting research because of these differences to traditional forms of data (Weinhardt 2020). This article discusses some ethical concerns in the use of big data for social research informed by considerations related to sociology and neighboring disciplines.

1.1. BD for Social Research

BD has now been employed in the social sciences in various ways. Twitter is often used to investigate political discourse, for example to show how anti-immigration laws shape public sentiment (Flores 2017). Public discourse may also be studied by combining internet search data, social media postings, and digitalized newspaper articles (Vasi et al. 2015). Online platforms also serve as major source of data, such as in the investigation of the effect of employment histories on the chances of getting a job using data from an US freelancer website (Leung 2014). Online dating sites may be used to research the structural effects of race, gender, and education on personal interactions and mating behavior. Electronic communication may also be a key source of data. Goldberg et al. (2016) used a large corpus of emails among employees in a high-technology firm to look into the effects of cultural and structural embeddedness in personal networks. Digital administrative data may be used to investigate racial profiling (Legewie 2016). These examples show the variety of research questions that may be addressed using a wide range of different data sources of BD in the social sciences.

Important advantages for social research practices directly follow from the sheer amount of data that become available for analysis. Databases used for analyses in the social sciences are usually rather small compared to the amount of data that the tech giants around the globe have to manage on a daily basis (ranging from tens of thousands to maybe several billion entries). Still, the above-mentioned definition of "big" also holds for the social sciences, as datasets with millions of entries and thousands of variables typically were impossible to handle by standard software packages in earlier days (a problem that has mostly been dealt with in the meantime where new versions of software have become available). Growth in the *volume* of available data is so enormous that it may well be regarded as a change in quality. For example, almost by definition, sample size is not an issue when using BD (on the contrary, the meaning of p-value changes completely if datasets contain millions of entries). The "census-like" quality of BD allows us to draw connections where it was not possible before. Together with the still increasing capacities of modern computers and ever more sophisticated methods to analyze them, this allows for the detection of patterns among data that we were previously unable to identify (Foster et al. 2017). It holds huge benefits for social network analyses, for example, as it becomes possible to analyze whole network structures, something that was virtually impossible to do before (Stopczynski et al. 2014). Twitter and Wikipedia are examples often used in the analysis of many different kinds of networks (Miller 2011).

The *variety* of BD as discussed above is maybe the strongest bonus for social research. There are rich bodies of observational data, digital trace data, textual data, and of course pictures, audios, and videos that enable us to study social phenomena from different perspectives and potentially allow deep insights into many aspects of everyday life. From the observation of actual (online) behavior to the deduction of personality traits or political leanings and emotional states from Facebook postings it seems a wide range of social science questions can be tackled using BD. This is even truer as different data sources may be combined to capture a comprehensive view of human interactions in online settings and beyond. By combining many data points from many different sources and different types of data, it becomes possible to gain insights where it had not been possible before. Furthermore, BD is not only about present-day information from the internet age. As more and more archives (such as newspapers and libraries) digitize their collections, such as Google Books and Project Gutenberg who are digitizing millions of books, BD begins to reach into (recent) history as well. Hence, using BD, it is not only possible to study current development but it is also possible to look back in time to see things that have evolved over a long time.

Due to its variety, BD may be used to study wide-ranging social phenomena, from individual behavior and attitudes to groups and organizations. It may be particularly useful for the study of human interactions, as many of the associated data forms, such as call records and chat protocols, likes and posts on social media, ratings on websites,

and comments below newspaper articles, are exactly that, depictions of interactions in a digital format. Such interactions may be viewed as links and nodes within social networks. Thus, BD holds huge potential for the study of networks as well. In a similar vein, it may benefit international, trans-national, and cross-cultural research. Many online services are available internationally and generate data from different countries and cultures that may be used to study similarities and differences.

Velocity, it may be argued, is the BD feature with the least impact on social science research, as real-time processing is usually not needed to answer scientific questions per se. Still, it may be useful in terms of dissemination activities, policy evaluation and recommendations, and the realization of public awareness and societal impact of social research. While not the primary purpose of scientific inquiry, these aspects are valuable in their own right but also impact back on the sciences as they are important to legitimize its cost and importance.

1.2. Critical Perspectives on BD

Despite these potential benefits, it is not yet clear whether the new "data scape" available to social researchers really advances the possibilities of social research to a significant extent (Hamid Ekbia et al. 2015; Crawford 2013; Lazer et al. 2014). Social scientists have studied the quality of the data they use for a long time and know much about data sources that help them to understand that data often are not merely an objective representation of reality, but constructed artifacts which may be framed and biased in certain ways (Baur et al. 2020). These assertions also hold for BD which is never encountered as "raw" material, but rather as data shaped and formed by, among other things, the technological constraints and economic preferences of the providers of such data. By now, several frameworks for assessing the quality of BD based on such insights have been proposed. An obvious case in point is the fact that almost by definition, BD is a byproduct of some data-generating activity that is not specifically designed to address scientific questions (such as online *trace data*, cf. Lazer et al. 2009). Hence, they share the problem with other, more traditional types of large-scale data, as Merton already observed 50 years ago: "a circumstance which regularly confronts sociologists seeking to devise measures of theoretical concepts by drawing upon the array of social data which happen to be recorded in the statistical series established by agencies of the society—namely, the circumstance that these data of social bookkeeping which happen to be on hand are not necessarily the data which best measure the concept" (Merton 1968, p. 219). This problem certainly also applies to BD. In addition to these issues of data quality, there is a huge danger of selectivity in BD, because those people who are online and actively using a specific service are very different from those who are offline and do not. Those people who are online may also be very selective about the information they share online. While some may be very active in posting social and political commentary on Facebook, others solely like items of pop culture or impressions of everyday life. Hence, what is available about who is heavily skewed and the truthfulness of the information provided in such online settings is still another matter of debate. It is, therefore, safe to say that the early hopes about the end of the theory where "with enough data, the numbers speak for themselves" (Anderson 2008, p. 8) have not materialized yet and probably never will.

As BD is not only quantitative but also qualitative in nature, this in theory allows for the study of subjective meanings as in qualitative research but on a much greater scale (Fuhse et al. 2020). However, the extraction of meaning from large bodies of text via algorithms still seems to be a challenging prospect (cf. Weichbold et al. 2020). It is still very difficult to use automated sentiment analyses to attach simple emotional inclinations from online text snippets. In addition, artificial intelligence still struggles with, for example, the moderation of online content and its screening for breaches of net-etiquette and hate speech in online forums, as human irony and similar patterns of ambiguity are very difficult to detect correctly via algorithms. Still, despite these difficulties, given the value BD may add as a source about social life, BD should find its place in the toolbox for social

researchers among other more traditional forms of data. Researchers should also be aware, however, of the knowledge that social scientists have accumulated over decades about the particularities of the social world when they use BD for their research (Mützel 2015).

2. Ethical Concerns in BD Research

The research ethics of BD are somewhat different from general ethical concerns about the use of BD in society. While there is a vivid debate about the ethics of BD in general, the literature specifically on the research ethics of BD is still scarce (Moreno et al. 2013). Historically, ethical guidelines for research have been especially important and prominent in epidemiology and the health sciences. In this context, it was defined that so-called human subject research needs to adhere to certain ethical principles (Metcalf and Crawford 2016). Two main principles are that research subjects must not be harmed and that participation in scientific studies must be voluntary (Hoyle et al. 2002). From this, it follows that research participants must give their consent to participate in a study and scientists must provide sufficient information about the study so that potential participants can make a reasonable decision on their participation (Keller and Lee 2003). Consequently, "informed consent" is a major topic in research ethics, especially surrounding social research online (Froomkin 2019), but also privacy concerns of individuals. Many other topics deserve attention and have not been touched upon here, such as the question of proper risk assessment, proper procedures to protect privacy (e.g., privacy-preserving record linkage, Vatsalan et al. 2013), data sharing and archiving (e.g., Zimmer 2010; Borgman 2012; Bishop and Gray 2017), and the issue of data ownership (Politou et al. 2018; Ruppert 2015). In the following, I briefly discuss four issues that overall feature less often in the discussion of research ethics for use of BD. I start by outlining the link between the value of BD and the risk involved for individual citizens.

2.1. Value and Risk

According to Gantz and Reinsel (2011)there is a fourth defining characteristic of BD: its value. For whom different BD applications are valuable varies widely, from private consumers to state actors and criminals. From an ethical point of view, it is important to recognize that value and risk are inherently linked. What makes BD applications so valuable to private companies, state actors, and criminals directly poses threats to privacy and self-determination of citizens. The following section discusses some real-world examples to highlight the risk to citizens by showing the value involved to other actors.

Many of the services that are now available online offer a great deal of value for their consumers, if only for entertainment and enjoyment. Otherwise, there would not be this massive new amount of data that may be used for analyses. People use social media platforms, online retailers, or streaming services because they reduce the burdens of everyday lives, ease communication with friends and family, and even bring joy and excitement. With the advent of electronic personal assistants in every smartphone or smart speaker driven by artificial intelligence and constantly improved by machine learning algorithms (such as Apple's Siri or Amazon's Alexa), early claims about the power of machines to know ourselves better than we do (Negroponte 1996) may finally become true. This way, BD helps to "enhance" the online experience of consumers, based on our stated preference as expressed online through likes or previous consumption histories. Netflix, for example, an online TV streaming service, is known for its data-driven approach to bind consumers to their never-ending stream of content. They use data mining into the viewing habits of their customers to reveal preferences that the consumers themselves may not even be aware of. This way, they can suggest additional programs and series which target specific groups, and quite critically, to produce content and whole series specifically tailored towards the taste of their customers, thereby tying them to their network with the obvious aim to stabilize and increase the stream of revenue. While this practice reduces customers' burden in dealing with the vast amount of content online, it also represents a not-so subtle influence on the cultural self-determination of individuals.

Maybe the most prominent example of extracting monetary value from BD and personal information is tailored advertising, also known as micro-targeting (Barbu 2014). The sales of online ad space specifically targeted to certain consumer groups given what is known about them regarding social demographics, online habits, and consumer preference based on social media profiles and search histories generates billions of dollars in revenue from this business model for Facebook, Google, and other companies alike. For this purpose, personal information is sold and traded, i.e., exploited commercially, often without the conscious awareness of those affected.

BD has other uses for commercial companies also. Employers, for example, not only use social media to find recruits via advertising and tailored targeting of potential candidates but also for screening the social media appearances of their candidates for potentially incriminating information. The misuse of such data in the employment process led the State of Maryland even "to prohibit by law employers asking for Facebook and other social media passwords during employment interviews and afterwards" (Kaisler et al. 2013). Assurance companies are another example for who BD may prove extremely valuable as it can help to predict risks of certain events from illness to car accidents from lifestyle preferences and online habits or even the actual way we drive our cars as these become connected to the internet and collect vast amounts of data. Credit companies may try to harvest personal online data to compute credit ratings and predict credit defaults. From these examples it can be seen that the (mis)use of data may interfere heavily in people's everyday lives, increasing loan and insurance costs or even impeding their chances to gain employment.

BD also opens up completely new possibilities for political interference, state control, and surveillance. In the political realm, the Facebook/Cambridge Analytica scandal showed how the harvesting of personal information on social networks may be used to develop data-driven political campaigns, even if they are designed to misinform or spread falsehoods to influence election results (Confessore 2018). The knowledge gained on the connection between social and political attitudes and preferences proved to be useful to influence people's political preferences. This helps political campaigns in micro-targeting their political ads, but also was exploited by Russian intelligence agencies in Western elections by using bots and false accounts to spread false and misleading information to undermine unwanted and promote more desired candidates and parties, respectively. This poses grave risks not only at the individual level, but also to whole political systems.

State actors have for some time now recognized the vast potential that comes with the digitalization of life and communication. Historically, official statistics have been developed by state actors to govern their populations (Diaz-Bone 2019) and administrations use BD for this purpose too (Thévenot 2020). For example, database matching has always been a desire of police forces and security agencies as they claim it helps them identifying and catching criminals and terrorists. With the advent of widespread security cameras in private as well as public spaces, the developments in face recognition software, computing power, and bandwidth, the police increasingly receive the power to track almost anyone anywhere for (hopefully) legal purposes. With such means, predictive policing goes way beyond crime prevention (which may also be aided by data analyses) and becomes a real possibility as data algorithms identify potential crime hotspots before they even occur (Williams et al. 2017a; Egbert and Krasmann 2019). Still, this is nothing compared to the width and depth of data penetration the NSA, the US National Security Agency, routinely undertake in their efforts to prevent terrorist attacks and other security threats (Landau 2013). Meanwhile, Chinese authorities seem to proceed even further into the realm of full state surveillance of citizen's private lives by introducing a social credit system that scores a wide variety of online and offline behavior, from tax fraud and parking tickets to social media tweets critical of the regime (Creemers 2018). Such an ability to link and search personal information on such a grand scale at the state level yields immense power and holds immense risks for individual freedom and liberty.

Maybe unsurprisingly, BD is very valuable for criminal activities also. There are many different ways to illegally extract money from people, through spam emails and phishing attempts but also things like ransomware attacks used for blackmailing whole institutions, where criminal hackers maliciously encrypt computers to collect a ransom. For example, in a data breach on a US adultery platform, private information of its customers was stolen and used for blackmailing campaigns (Zetter 2015). Identity theft is another danger where people steal complete personal profiles from social media platforms to pretend they are this person, for example to apply for loans or other benefits. This way, criminals may not only extract money from those who are directly involved in a data breach of their personal information, but also from third parties where false identities are used to fraudulently extract money.

From these examples it becomes clear that there are severe risks from (mis)use of BD by private companies, state actors, and criminals. Those features that render BD valuable for many parties at the same time poses such a threat to the privacy and everyday lives of citizens around the globe (e.g., Jackson and Orebaugh 2018). Hence, assessing such risks and preserving privacy for individuals are major issues in the ethics of BD, although not the only issues.

2.2. Anonymization and Re-Identification

To protect data privacy and the confidentiality of personal information, and to preclude the risks of re-identification of research subjects, research data are typically anonymized at various stages of the data-handling process. Where data are fully anonymized, they lose their character as personal data, and therefore informed consent is no longer necessary to handle the data. There are, however, differences in the definitions of privacy and personal information, depending on the national context of data sources and their usage. While the definition of what constitutes confidential data is somewhat open to discussion and may not be determined objectively, most data privacy law in Europe is now regulated by the new EU general data protection regulation (EU-GDPR 2016). European data protection law provides an explicit list of certain information that the legislature considers to be sensitive and which therefore must be considered as a binding minimum for researchers working in Europe. However, full anonymization is typically not possible, at least not without huge losses of information important to the research question. This is already a relevant issue when archiving and sharing quantitative survey data. With BD, the risk of re-identification vastly increases, as the amount of data that is available on the internet becomes a big challenge for the practice of data anonymization (Bender et al. 2017). This will become even more relevant as the possibility of linking different data sources increases (some even argue it becomes increasingly impossible). That *de-anonymization* of research data is possible in this way with data that is already available online has been proven numerous times (Cecaj et al. 2014; Lubarsky 2017; Archie et al. 2018). An early project using Facebook data released their dataset anonymized for scientific purposes only to find that the information it contained quite easily allowed for the identification of research subjects in the dataset on Facebook itself. The combination of data together with the knowledge that all subjects were students in one particular university actually made this task relatively easy (Zimmer 2010). While this example mainly proves that researchers had not thought carefully enough about the risks of de-anonymization, other examples show that the task is increasingly difficult to fulfill. Netflix, an online service for streaming television, had publicly released a dataset of their users' viewing habits as part of a public challenge to increase their content and service provision. A team of researchers was able to identify users by matching viewers' preferences to ratings people provide on IMDB, an online platform for film and television reviews, which in many cases included peoples' names (Narayanan and Shmatikov 2008). Other researchers claimed to have identified Banksy, a street artist notorious for shielding his personal identity, by comparing publicly available records on local housing and voting with the known locations of the street artist's work (Hauge et al. 2016). These examples make it clear that anonymization in the field of

BD poses huge challenges. The problem is somewhat mitigated by the fact that researchers must not use research data for anything other than scientific purposes. For researchers who want to provide BD for other researchers to use, this can be assured by the appropriate formulation of user agreements. Hence, any uses to marketize archival data or to transgress into someone's privacy are per se improper, if not outright illegal. This will not prevent people with criminal intent, of course, and proper assurance of anonymity in the data is still an ethical demand. However, the output of scientific research and accompanying datasets is likely not of any interest for companies using data sourcing as their business model. As the data are potentially already out there, the actual risks involved in providing BD in archives specifically for researchers might be comparatively small.

2.3. Documentation and Dissemination

Documentation of research practices as well as the dissemination of source materials and results are key principles of what has been called Open Science (Fecher and Friesike 2013; Vicente-Saez and Martinez-Fuentes 2018). This umbrella term describes the general idea to make all stages of scientific inquiry accessible to a broader audience, including both professional scientists and non-scientists. The general goal is to facilitate the publication and communication of scientific knowledge and scientific practices. Important aspects are the open accessibility of publications and research data, as they allow for the replication of findings by other researchers. Therefore, the preregistration of research questions and hypotheses is another principle that has grown in acceptance recently and which is intended to strengthen the reproducibility of results. To make Open Science workable, general principles of data management should apply during all phases of the research process with the two imperatives of keeping stored data not only safe and secure but also retrievable and accessible. This includes planning for archiving and secondary usage of the data, ensuring that all efforts meet the FAIR principles of research data management: findable, accessible, interoperable, and re-usable (Wilkinson et al. 2016).

One might argue that the question of open data and research reproducibility is a more salient issue in BD than other types of social research because the workflows and research results typically involved in such projects easily lend themselves for this purpose. Proper data management practices in data-driven research, which involves the heavy use of software and code, already demand the documentation of procedures and transformation to allow for collaboration among project members. From this internal process of project documentation, if implemented correctly, it is a comparatively small step to share code, data, and documentation on online platforms (such as github.com, gitlab.com, or others) for the benefit of other researchers and the wider public.

While the sharing of code and other research materials seems rather less problematic from an ethical point of view, sharing the actual data with a wider audience may be problematic. A sometimes neglected distinction in this regard concerns the collection and usage of data by primary researchers vs. the secondary usage of research data by other researchers (van Deth 2003). Secondary research, on the other hand, is based on the dissemination of research data to a wider (scientific) audience in order to replicate earlier analyses or to answer completely new research questions. While there are examples for primary research that meet the criteria of BD, such as the personality experiment worldwide (Stillwell and Kosinski 2012), in the context of BD, we typically think of using data that has been collected by others already, i.e., instances of secondary research. In case personal data is to be provided for the benefit of other researchers, efforts should be undertaken to anonymize the data, as is standard practice in current quantitative social research. Currently, the archiving and dissemination of BD used for research to other scientists is an issue that deserves more attention as there are no standardized solutions or archival practices established yet, and a discussion around this issue has also started to emerge (Williams et al. 2017b). As documentation and dissemination become a cornerstone of scientific endeavor, the question of *anonymization* becomes an even more pressing issue.

2.4. Stakeholders in Social Research Practices

Another important ethical aspect involves the question of which groups of persons might be affected through BD research and hence, who holds a stake in the scientific use of BD. Here, one needs to clarify how the research might affect them and how it can be assured that their interests are respected and their rights protected. The list of potential stakeholders obviously includes research subjects (i.e., the persons the data is about) who should be protected from harm. Often, research ethics considers certain groups as particularly vulnerable, such as children but also seniors. The particular risk a group faces must be considered in the decision whether they should be considered vulnerable. In the case of BD, the question of vulnerability, is among other things, closely linked to digital literacy and tech-savviness. This is due to their possible inexperience and ignorance towards existing perils and possibilities of BD research.

While it seems clear that we should care about the subjects of our research, there are other stakeholders to consider when we think about the ethical impact of our studies. The data providers from which we retrieve BD e.g., through online databases and websites, may also be considered stakeholders. While many of them are likely to be powerful commercial companies or state agencies, this is not always the case. Web-scraping data from websites and online databases for example may interfere with their performance and limit their functionality for other users if the scraping is done carelessly. It is therefore necessary to include data providers in our ethical considerations also.

Other researchers and scientists also have a stake in our proper scientific conduct. For example, research is seldom done alone, but in the context of cooperation with other scientists and research workers. However, project members and associates are an often-overlooked party in research ethics but still deserve attention. Researchers as a community share a stake in the ethical discussion of research issues as well. As colleagues and fellow researchers, we have an interest that everyone acts properly so that data sources remain accessible for future research and are not rendered inaccessible due to mishaps and mistakes in research practices. If other stakeholders perceive some scientists' conduct as improper, fellow scientists may be prevented from using the same or similar data in the future. It is, therefore, in the self-interest of scientists to act ethically. While professional conduct between scientists is often covered in the general rules of good scientific practice by the various national science foundations, there are likely to be specific ethical issues in relation to fellow researchers.

Finally, yet importantly, science itself is an important societal endeavor and scientists should have certain privileges in doing their research (something that is also recognized in the EU-GDPR). The importance of scientific research and the possibility to conduct it freely without undue inference by other actors is something that should and must be weighed against the claims of other stakeholders, and also against the rightful claims of research subjects. This is where the scientific usage of BD should be privileged over other private and commercial uses.

2.5. Ethical Regulation and Institutional Review Board (IRBs)

A discussion on research ethics should acknowledge that there are also critical voices regarding the role of ethics and the "human subject research model" (Bassett and O'Riordan 2002), especially in qualitative research (Haggerty 2004; Hammersley 2009), but also regarding the practical implementation of research ethics through review boards (von Unger et al. 2016). It is argued that ethical reviews before the research is conducted in the field requires pre-fixation of the research design that is incompatible with the principle openness of qualitative inquiry, where research questions, topics of the inquiry, and sample selection are constantly adjusted in the light of experiences in the field. The analysis of some kinds of BD, for example the analysis of textual data, resembles qualitative approaches to some extent (e.g., the inductive exploration of data to find patterns not previously envisaged). For example, the collection of online data may be observational data similar to ethnographic research where acquiring informed consent may be difficult. Thus, the

question arises whether the same concerns against ethical reviews can also be raised in the realm of BD.

However, overall, the implementation and deeper integration of regulatory bodies and ethics committees in the conduct of social science studies seems desirable, not least because of the advent of BD and the challenges it poses for privacy and anonymity. As this is still a relatively new area for research in the social sciences, practices and protocols still have to develop. It would be naive to simply and only trust in the self-guidance and morality of individual researchers. First, knowledge and acceptance of ethical standards and principles may vary widely between countries and disciplines and a unitary approach is therefore desirable for all stakeholders involved. Second, the transformation of international research into something akin to a capitalistic competition of scientific reputation with rankings and impact factors deciding on the standing and status of researchers (Münch 2014), the pressure on the individual researcher in this system is too high to expect everyone to adhere simply to their conscience everywhere and all the time during every phase of research. Third, as one of the advantages of BD is the possibility to conduct international research and country comparisons more easily, this complicates the question of research ethics, as researchers have to keep their research in compliance with regulations in different countries and regions. Hence, some kind of institutionalized standard-bearer of ethical research practices, which provides oversight, but also guidance and assistance, seems desirable. While the ideal form of such a format still needs to be developed, this may take, for example, the form of an ombudsman that all stakeholders in the research process may turn to if they have questions or concerns.

3. Discussion and Conclusions

The usage of BD is more and more common in social research. While data from online sources share many aspects with traditional data sources such as survey data or archival data, they also exhibit some features that are new, if not in principle, then at least in the magnitude of its occurrence. This is the reason they offer new opportunities for social researchers to address research questions they have not been able to answer previously. Yet, BD poses new questions for the handling and ethical treatment of such data, simply because of the differences to traditional forms of data.

Ethical concerns around BD often involve issues of data sharing, privacy, and security. Data privacy scandals such as the one which included Facebook and Cambridge Analytica for influencing political elections (Zuiderveen Borgesius et al. 2018) show the risk involved in the use of BD for research purposes. However, these and other scandals also led to a global public that is more sensitive to such large-scale breaches of privacy and the consequences they may have on a grand scale. The tech and social media giants have come under heightened pressure to be more open about their data-sharing practices with third parties as well as to be more restrictive in the way they provide access to their data. Still, the misuse of online data by various actors is likely to continue, as the potential gains are huge. It may therefore be argued that compared to the risks users of digital communication face in their daily routines, those risks from scientific inquiries are arguably relatively small (Kämper 2016). However, researchers should not be left alone in addressing these issues. Rather, they should receive support from institutionalized agencies trained and tasked in the handling of ethical issues. These do not need to take the form of established IRBs and should provide guidance as much as restrictions.

This paper also touched upon the topic of stakeholders who may be affected by social research using BD and who hence should be part of the ethical considerations leading up to BD research projects. However, this was only a preliminary exercise and a thorough mapping of all different kinds of stakeholders, as is often done in other contexts (Aaltonen and Kujala 2016; Brugha and Varvasovszky 2000), is still lacking. Conducting such a stakeholder mapping may in itself be seen as an ethical requirement. When considering the stakeholders of BD research, one must consider the special role that science as a social enterprise should hold in these considerations. As the institution tasked with

providing grounded knowledge for society, scientists should hold a privileged position when conducting research and, therefore, when using BD for their research.

While the article made these and other points and observations, they should not be seen as definitive conclusions, but rather as invitations for further discussion and investigation. Indeed, there should be a wide and open debate on the potential ethical pitfalls of BD in the social sciences and social scientists must take part more actively and visibly in these discussions. The peculiarities of the social world such as its pre-structuration as a symbolic realm to be interpreted by self-aware actors need to be kept in mind when discussing the potential benefits for social research. So far, the field of BD research is too often dominated by natural and computer scientists and their take on the social world, who lack knowledge about the specificity of the social objects under study (Lazer et al. 2009). Likewise, they are often unaware of the social and ethical dilemmas such research may actually involve. As we need new rules and guidelines for conducting research in the digital realm, it, therefore, becomes an ethical requirement in itself for social scientists to share their knowledge on studying the social world, on substantial as well as methodological, and ethical topics, with scientists from other fields and disciplines. However, as this article is mostly informed by considerations related to sociology and neighboring disciplines, it may well be that in other disciplines the potential for research as well as the ethical questions differ and, therefore, need to be addressed separately.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

Aaltonen, Kirsi, and Jaakko Kujala. 2016. Towards an improved understanding of project stakeholder landscapes. *International Journal of Project Management* 34: 1537–52. [CrossRef]

Anderson, Chris. 2008. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired Magazine*. p. 16. Available online: http://www.uvm.edu/~{}pdodds/files/papers/others/2008/anderson2008a.pdf (accessed on 27 October 2020).

Archie, Maryam, Sophie Gershon, Abigail Katcoff, and Aaron Zeng. 2018. Who' s Watching? De-Anonymization of Netflix Reviews Using Amazon Reviews. Available online: https://www.semanticscholar.org/paper/Who-%E2%80%99-s-Watching-De-anonymization-of-Netflix-using-Archie-Gershon/d2183ba370dd77e3f7f4847a017567619d72a85d?p2df (accessed on 27 October 2020).

Askitas, Nikolaos, and Klaus F. Zimmermann. 2015. The internet as a data source for advancement in social sciences. *International Journal of Manpower* 36: 2–12. [CrossRef]

Barbu, Oana. 2014. Advertising, Microtargeting and Social Media. *Procedia Social and Behavioral Sciences* 163: 44–49. [CrossRef]

Barrett, Meredith A., Olivier Humblet, Robert A. Hiatt, and Nancy E. Adler. 2013. Big Data and Disease Prevention: From Quantified Self to Quantified Communities. *Big Data* 1: 168–75. [CrossRef]

Bassett, Elizabeth H., and Kate O'Riordan. 2002. Ethics of Internet Research: Contesting the Human Subjects Research Model. *Ethics and Information Technology* 4: 233–47. [CrossRef]

Baur, Nina, Peter Graeff, Lilli Braunisch, and Malte Schweia. 2020. The Quality of Big Data. Development, Problems, and Possibilities of Use of Process-Generated Data in the Digital Age. *Historical Social Research Historische Sozialforschung* 45: 209–43. [CrossRef]

Bender, Stefan, Ron Jarmin, Frauke Kreuter, and Julia Lane. 2017. Privacy and Confidentiality. In *Big Data and Social Science: A Practical Guide to Methods and Tools*. Edited by Ian Foster, Rayid Ghani, Ron S. Jarmin, Frauke Kreuter and Julia Lane. Chapman & Hall/CRC Statistics in the Social and Behavioral Sciences Series. Boca Raton: CRC Press, pp. 299–311.

Bishop, Libby, and Daniel Gray. 2017. Chapter 7: Ethical Challenges of Publishing and Sharing Social Media Research Data. In *The Ethics of Online Research*. Edited by Kandy Woodfield. Advances in research ethics and integrity. Emerald Publishing Limited: vol. 2, pp. 159–87.

Borgman, Christine L. 2012. The conundrum of sharing research data. *Acta Anaesthesiologica Scandinavica* 63: 1059–78. [CrossRef]

Brugha, Ruairí, and Zsuzsa Varvasovszky. 2000. Stakeholder Analysis: A Review. *Health Policy and Planning* 15: 239–46. [CrossRef]

Cecaj, Alket, Marco Mamei, and Nicola Bicocchi. 2014. Re-identification of anonymized CDR datasets using social network data. Paper presented at 2014 IEEE International Conference on Pervasive Computing and Communication Workshops (Percom Workshops), Budapest, Hungary, March 24–28; New York: IEEE, pp. 237–42.

Confessore, Nicholas. 2018. Cambridge Analytica and Facebook: The Scandal and the Fallout so Far. *The New York Times*. Available online: https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html (accessed on 27 October 2020).

Crawford, Kate. 2013. The Hidden Biases in Big Data. *Harvard Business Review*. p. 1. Available online: https://hbr.org/2013/04/the-hidden-biases-in-big-data (accessed on 27 October 2020).
Creemers, Rogier. 2018. China's Social Credit System: An Evolving Practice of Control. *SSRN*. p. 3175792. Available online: https://ssrn.com/abstract=3175792 (accessed on 27 October 2020).
Diaz-Bone, Rainer. 2019. Statistical Panopticism and Its Critique. *Historical Social Research* 44: 77–102. Available online: https://www.ssoar.info/ssoar/bitstream/document/61929/1/ssoar-hsr-2019-2-diaz-bone-Statistical_Panopticism_and_Its_Critique.pdf (accessed on 27 October 2020).
Dumbill, Edd. 2012. *Planning for Big Data: A CIO's Handbook to the Changing Data Landscape*. Beijing: O'Reilly Media, O'Reilly Media: Sebastopol.
Egbert, Simon, and Susanne Krasmann. 2019. Predictive policing: Not yet, but soon preemptive? *Policing and Society*, 1–15. [CrossRef]
EU General Data Protection Regulation. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union L*, 119/1.
Evans, James, and Jacob G. Foster. 2019. Computation and the Sociological Imagination. *Contexts* 18: 10–15. [CrossRef]
Fecher, Benedikt, and Sascha Friesike. 2013. Open Science: One Term, Five Schools of Thought. In *Opening Science: The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing*. Edited by Bartling Sönke and Friesike Sascha. Heidelberg, New York, Dordrecht and London: Springer, Cham, pp. 17–47.
Flores, René D. 2017. Do Anti-Immigrant Laws Shape Public Sentiment? A Study of Arizona's SB 1070 Using Twitter Data. *American Journal of Sociology* 123: 333–84. [CrossRef]
Foster, Ian, Rayid Ghani, Ron S. Jarmin, Frauke Kreuter, and Julia Lane, eds. 2017. *Big Data and Social Science: A Practical Guide to Methods and Tools*. Chapman & Hall/CRC Statistics in the Social and Behavioral Sciences Series. Boca Raton: CRC Press, Available online: http://lib.myilibrary.com/detail.asp?ID=950162 (accessed on 27 October 2020).
Froomkin, A. Michael. 2019. Big Data: Destroyer of Informed Consent. *Yale Journal of Health Policy Law and Ethics* 18: 27–54.
Fuhse, Jan, Oscar Stuhler, Jan Riebling, and John Levi Martin. 2020. Relating social and symbolic relations in quantitative text analysis. A study of parliamentary discourse in the Weimar Republic. *Poetics* 78: 101363. [CrossRef]
Gantz, John, and David Reinsel. 2011. Extracting Value from Chaos. *IDC iView* 1142: 1–12.
Goldberg, Amir, Sameer B. Srivastava, V. Govind Manian, William Monroe, and Christopher Potts. 2016. Fitting in or Standing Out? The Tradeoffs of Structural and Cultural Embeddedness. *American Sociological Review* 81: 1190–1222. [CrossRef]
Graeff, Peter, and Nina Baur. 2020. Digital Data, Administrative Data, and Survey Compared: Updating the Classical Toolbox for Assessing Data Quality of Big Data, Exemplified by the Generation of Corruption Data. *Historical Social Research* 45: 244–69.
Haggerty, Kevin D. 2004. Ethics Creep: Governing Social Science Research in the Name of Ethics. *Qualitative Sociology* 27: 391–414. [CrossRef]
Hamid Ekbia, Michael Mattioli, Inna Kouper G. Arave, Ali Ghazinejad, Timothy Bowman, Venkata Ratandeep Suri, Andrew Tsou, Scott Weingart, and Cassidy R. Sugimoto. 2015. Big Data, Bigger Dilemmas: A Critical Review. *Journal of the Association for Information Science and Technology* 66: 1523–45. [CrossRef]
Hammersley, Martyn. 2009. Against the ethicists: On the evils of ethical regulation. *International Journal of Social Research Methodology* 12: 211–25. [CrossRef]
Hartmann, Philipp Max, Mohamed Zaki, Niels Feldmann, and Andy Neely. 2016. Capturing value from big data—A taxonomy of data-driven business models used by start-up firms. *International Journal of Operations & Production Management* 36: 1382–1406. [CrossRef]
Hauge, Michelle V., Mark D. Stevenson, D. Kim Rossmo, and Steven C. Le Comber. 2016. Tagging Banksy: Using geographic profiling to investigate a modern art mystery. *Journal of Spatial Science* 61: 185–90. [CrossRef]
Hoyle, Rick, Monica J. Harris, and Judd Charles. 2002. *Research Methods in Social Relations*, 7th ed. Fort Worth: Wadsworth.
Jackson, Catherine, and Angela Orebaugh. 2018. A study of security and privacy issues associated with the Amazon Echo. *International Journal of Information Technology, Control and Automation* 1: 91. [CrossRef]
Kaisler, Stephen, Frank Armour, J. Alberto Espinosa, and William Money. 2013. Big Data: Issues and Challenges Moving Forward. Paper presented at 46th Annual Hawaii International Conference on System Sciences, Wailea, Maui, Hawaii, January 7–10; Edited by Ralph H. Sprague. Los Alamitos: IEEE Computer Society, pp. 995–1004.
Kämper, Eckard. 2016. Risiken sozialwissenschaftlicher Forschung? Forschungsethik, Datenschutz und Schutz von Persönlichkeitsrechten in den Sozial- und Verhaltenswissenschaften. RatSWD Working. Available online: https://www.econstor.eu/handle/10419/129793255 (accessed on 27 October 2020).
Keller, Heidi E., and Sandra Lee. 2003. Ethical Issues Surrounding Human Participants Research Using the Internet. *Ethics and Behavior* 13: 211–19. [CrossRef] [PubMed]
Landau, Susan. 2013. Making Sense from Snowden: What's Significant in the NSA Surveillance Revelations. *IEEE Security Privacy* 11: 54–63. [CrossRef]
Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. Big Data. the Parable of Google Flu: Traps in Big Data Analysis. *Science* 343: 1203–5. [CrossRef]
Lazer, David, Alex Pentland, Lada Adamic, Sinan Aral, Albert-Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, and et al. 2009. Social Science. Computational Social Science. *Science* 323: 721–23. [CrossRef]

Lazer, David, and Jason Radford. 2017. Data ex Machina: Introduction to Big Data. *Annual Review of Sociology* 43: 19–39. [CrossRef]
Legewie, Joscha. 2016. Racial Profiling and Use of Force in Police Stops: How Local Events Trigger Periods of Increased Discrimination. *American Journal of Sociology* 122: 379–424. [CrossRef]
Leung, Ming D. 2014. Dilettante or Renaissance Person? How the Order of Job Experiences Affects Hiring in an External Labor Market. *American Sociological Review* 79: 136–58. [CrossRef]
Lubarsky, Boris. 2017. Re-Identification of "Anonymized Data". 1 GEO. L. TECH. REV. 202. Available online: https://perma.cc/86RR-JUFT (accessed on 27 October 2020).
Marres, Noortje. 2017. *Digital Sociology: The Reinvention of Social Research*. Cambridge: Polity Press.
Merton, Robert K. 1968. *Social Theory and Social Structure*. New York: Free Press.
Metcalf, Jacob, and Kate Crawford. 2016. Where are human subjects in Big Data research? The emerging ethics divide. *Big Data and Society*, 3. [CrossRef]
Miller, Greg. 2011. Sociology. Social Scientists Wade into the Tweet Stream. *Science* 333: 1814–15. [CrossRef] [PubMed]
Moreno, Megan A., Natalie Goniu, Peter S. Moreno, and Douglas Diekema. 2013. Ethics of Social Media Research: Common Concerns and Practical Considerations. *Cyberpsychology, Behavior and Social Networking* 16: 708–13. [CrossRef] [PubMed]
Münch, Richard. 2014. *Academic Capitalism*. Abingdon: Routledge.
Mützel, Sophie. 2015. Facing Big Data: Making sociology relevant. *Big Data and Society*, 2. [CrossRef]
Narayanan, Arvind, and Vitaly Shmatikov. 2008. Robust De-anonymization of Large Sparse Datasets. Paper presented at 2008 IEEE Symposium on Security and Privacy (sp 2008), Oakland, CA, USA, May 18–21; New York: IEEE, pp. 111–25.
Negroponte, Nicholas. 1996. *Being Digital*, 1st ed. New York: Vintage Books.
Politou, Eugenia, Efthimios Alepis, and Constantinos Patsakis. 2018. Forgetting personal data and revoking consent under the GDPR: Challenges and proposed solutions. *Journal of Cybersecurity*, 4. [CrossRef]
Rotella, Perry. 2012. Is Data the New Oil? *Forbes*. April 2. Available online: https://www.forbes.com/sites/perryrotella/2012/04/02/is-data-the-new-oil/#5710586c7db3 (accessed on 15 October 2020).
Ruppert, Evelyn. 2015. Who Owns Big Data. *Discover Society*. p. 23. Available online: http://research.gold.ac.uk/12494/ (accessed on 27 October 2020).
Ruths, Derek, and Jürgen Pfeffer. 2014. Social Sciences. Social Media for Large Studies of Behavior. *Science* 346: 1063–64. [CrossRef]
Salganik, Matthew J. 2018. *Bit by Bit: Social Research in the Digital Age*. Princeton: Princeton University Press.
Shah, Dhavan V., Joseph N. Cappella, and W. Russell Neuman. 2015. Big Data, Digital Media, and Computational Social Science. *The ANNALS of the American Academy of Political and Social Science* 659: 6–13. [CrossRef]
Stillwell, David J., and Michal Kosinski. 2012. myPersonality project: Example of successful utilization of online social networks for large-scale social research. *American Psychologist*. 59, pp. 93–104. Available online: http://www.davidstillwell.co.uk/articles/Stillwell_and_Kosinski_(2012)_myPersonality_Introduction.pdf (accessed on 27 October 2020).
Stopczynski, Arkadiusz, Vedran Sekara, Piotr Sapiezynski, Andrea Cuttone, Mette My Madsen, Jakob Eg Larsen, and Sune Lehmann. 2014. Measuring Large-Scale Social Networks with High Resolution. *PLOS ONE* 9: e95978. [CrossRef]
Thévenot, Laurent. 2020. Measure for Measure: Politics of Quantifying Individuals to Govern Them. *Historical Social Research Historische Sozialforschung* 44: 44–76. [CrossRef]
von Unger, Hella, Hansjörg Dilger, and Michael Schönhuth. 2016. Ethikbegutachtung in der sozial- und kulturwissenschaftlichen Forschung? Ein Debattenbeitrag aus soziologischer und ethnologischer Sicht. *Forum Qualitative Sozialforschung Forum Qualitative Social Research* 17: 13. Available online: https://www.ssoar.info/ssoar/bitstream/document/57126/1/ssoar-fqs-2016-3-unger_et_al-Ethikbegutachtung_in_der_sozial-_und.pdf (accessed on 27 October 2020).
van Deth, Jan. 2003. Using published survey data. In *Cross-Cultural Survey Methods*. Edited by Janet A. Harkness, Fons J. R. van de Vijver and Peter Ph. Mohler. Hoboken: Wiley-Interscience, pp. 291–307.
Vasi, Ion Bogdan, Edward T. Walker, John S. Johnson, and Hui Fen Tan. 2015. "No Fracking Way!" Documentary Film, Discursive Opportunity, and Local Opposition against Hydraulic Fracturing in the United States, 2010 to 2013. *American Sociological Review* 80: 934–59. [CrossRef]
Vatsalan, Dinusha, Peter Christen, and Vassilios S. Verykios. 2013. A taxonomy of privacy-preserving record linkage techniques. *Information Systems* 38: 946–69. [CrossRef]
Veltri, Giuseppe A. 2020. *Digital Social Research*. Cambridge: Polity Press.
Vicente-Saez, Ruben, and Clara Martinez-Fuentes. 2018. Open Science now: A systematic literature review for an integrated definition. *Journal of Business Research* 88: 428–36. [CrossRef]
Weichbold, Martin, Alexander Seymer, Wolfgang Aschauer, and Thomas Herdin. 2020. Potential and Limits of Automated Classification of Big Data—A Case Study. *Historical Social Research Historische Sozialforschung* 45: 288–313. [CrossRef]
Weinhardt, Michael. 2020. Ethical Issues in the Use of Big Data for Social Research. *Historical Social Research Historische Sozialforschung* 45: 342–68. [CrossRef]
Wilkinson, Mark D., Michel Dumontier, I. Jsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, and et al. 2016. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data* 3: 160018. [CrossRef]
Williams, Matthew L., Pete Burnap, and Luke Sloan. 2017a. Crime Sensing with Big Data: The Affordances and Limitations of Using Open-source Communications to Estimate Crime Patterns. *The British Journal of Criminology* 57: 320–40. [CrossRef]

Williams, Matthew L., Pete Burnap, and Luke Sloan. 2017b. Towards an Ethical Framework for Publishing Twitter Data in Social Research: Taking into Account Users' Views, Online Context and Algorithmic Estimation. *Sociology* 51: 1149–68. [CrossRef]

Zetter, Kim. 2015. Hackers finally post stolen Ashley Madison data. *Wired*. Available online: https://www.wired.com/2015/08/happened-hackers-posted-stolen-ashley-madison-data/ (accessed on 26 October 2020).

Zimmer, Michael. 2010. "But the Data Is Already Public": On the Ethics of Research in Facebook. *Ethics and Information Technology* 12: 313–25. [CrossRef]

Zuiderveen Borgesius, Frederik, Judith Moeller, Sanne Kruikemeier, Ronan Ó Fathaigh, Kristina Irion, Tom Dobber, Balázs Bodó, and Claes H. de Vreese. 2018. Online Political Microtargeting: Promises and Threats for Democracy. *Utrecht Law Review* 14: 82–96.