Ethical Issues in the Use of Big Data for Social Research

Weinhardt, Michael

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Empfohlene Zitierung / Suggested Citation:
Weinhardt, M. (2020). Ethical Issues in the Use of Big Data for Social Research. *Historical Social Research*, *45*(3), 342-368. https://doi.org/10.12759/hsr.45.2020.3.342-368

Ethical Issues in the Use of Big Data for Social Research

Michael Weinhardt *

Abstract: *»Ethische Fragen bei der Nutzung von Big Data in der Sozialforschung«*. With the advent of Big Data (BD) in the social sciences, vast amounts of data (and the tools to analyze them) have become available faster than ethical and legal standards could develop regarding the use of such data. At the same time, data collectors and analysts face new moral dilemmas as the proliferation of personal and impersonal data clearly poses new challenges to traditional assumptions about privacy and autonomy. The discussion of such ethical challenges seems to lag behind and the literature specifically dealing with the research ethics of BD is still scarce. This article asks which ethical and legal aspects need to be considered when collecting and analyzing data on individuals from the web and combining them to gain an enriched picture of human activities. It proceeds to provide a brief overview of existing research ethics regulations and outlines areas of particular relevance to the challenges that come with the use of BD, such as the delineation of human subject research, the (im)possibility of informed consent for these new kinds of data, the sources and public availability of data and questions of risk and risk assessment. It also formulates some generic recommendations in order to stimulate further debate, one of which posits that social scientists must address and discuss the challenges that emerge in research applications of BD more widely than it is currently the case.

Keywords: research ethics, digital research, human subject research, informed consent, computational social science, big data, data protection, social research.

1. Introduction

"Act in such a way that you treat humanity, whether in your own person or in the person of any other, never merely as a means to an end, but always at the same time as an end."
(Immanuel Kant, Groundwork for the Metaphysics of Morals)

"Art and science, research and teaching shall be free. […]"
(Basic Law for the Federal Republic of Germany*, Art. 5, para. 3*)

* Michael Weinhardt, Institut für Soziologie, Technische Universität Berlin, Fraunhoferstrasse 33-36, 10587 Berlin, Germany; michael.weinhardt@tu-berlin.de.

Historical Social Research 45 (2020) 3, 342-368 | published by GESIS
DOI: 10.12759/hsr.45.2020.3.342-368

Questions of privacy, data protection, and ethics have been discussed for decades in social research and related disciplines such as psychology and epidemiology, and solutions have been put in place to address these issues in practice. This is somewhat different in the area of Big Data (BD) in the social sciences. By now, BD is not a new phenomenon anymore and discussions about its potentials as well as its uses in actual research are increasingly widespread in the social sciences and emerging fields such as computational social sciences (Foster et al. 2016; Lazer and Radford 2017). However, the new availability of vast amounts of data and tools to analyze them have come faster than ethical and legal standards could develop regarding the use of such data. While the proliferation of personal and impersonal data clearly poses new challenges to traditional assumptions about privacy and autonomy (Bender et al. 2016), the discussion of such ethical challenges seems to lag behind.

Typically, the definition of BD is tied to the three dimensions of volume, variety, and velocity (Salganik 2018). Ganz and Reinsel (2011) define BD accordingly as "a new generation of technologies and architectures designed to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis." First, the sheer *volume* of information that has become available electronically over the past 20 years or so is astonishing. In social media, for example, Facebook and Instagram each have more than one billion users and the related mass of pictures, videos and text snippets uploaded every day seems unimaginable. 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).[1] Second, the high-*velocity* aspect is often a result of the online nature of BD. A smooth user experience necessitates the processing of large amounts of data in real-time. For commercial data use speed often is a key advantage (for example in the world of business analytics). The need for timely processing of data that does not interfere with users' online experience certainly has its own challenges, especially for database programming and computation power. Hence, the development of BD methods would not have been possible without an incredible growth in computing power that has roughly doubled every two years over at least five decades. Third, BD encompasses data from a wide *variety* of types and sources such as chat rooms, social media, communication tools such as

[1] Interestingly, according to this understanding, "big" is relative to context, especially historically. Every time a new medium for the storage of data was invented, the amount of data that could be easily accessed increased manifoldly. What was considered big some time ago was rather small a decade later. Also, compared to the amount of data that the tech giants around the globe have to manage every day, the data bases used for analyses in the social sciences usually are comparatively small, ranging from hundreds to maybe several millions of entries. Therefore, Riebling (2018) suggested the term *medium data* to emphasize the smaller scale of the data that social scientists are typically concerned with in contrast to the "really" BD computer scientists work on (and their problems of real-time processing).

messengers, wearables such as fitness tracking devices, but also video portals, digital libraries, and the simple everyday usage of internet browsers. In a way, we should rather speak of BD in the plural as it comes in so many different forms and flavors. The data itself may consist of different types of entities, such as plain numbers, text, audios, videos, or the combination of all of these.[2] In addition, as more and more archives digitize their collections (and projects such as Google Books and Project Gutenberg digitize millions of books from many different countries), BD begins to reach into (recent) history as well. Thus, the variety of BD is huge and may even be growing.

In *value*, Ganz and Reinsel (2011) find another characteristic that is central to the understanding of BD, one which is also deeply relevant to ethical issues in the use of BD. Many scientists agree that BD is very valuable in many respects, and not without reason is it often called the "oil of the 21st century" (e.g., Rotella 2012). In addition, new analytic tools for large-scale data analysis allow the extracting of information from data where previously nothing of value could be found. In combination with new statistical modeling techniques, BD may enable advances in many areas that are practically important, such as the detection of cancer in patients from biometric data. Other benefits are more benign. As BD is often a direct outcome of the digitalization of everyday life, one major advantage, for companies as well as consumers, is the increase in productivity and the reduction in transaction costs. Transnational communication, for example, even via video calls, is now available almost instantly and free of charge. However, this digitalization of life also involves huge risks. For example, in 2018 it was revealed that political data consulting firm had acquired access to the personal information of up to 87 million users of Facebook (cf. Hardcastle 2018). The company used the data to build personality profiles with the aim to most effectively target social media users with political advertising campaigns. While this practice has been discontinued since, at least officially, the example of Facebook – the Cambridge Analytica Scandal – exemplifies the dangers of the misuse of social media data.

The preceding example stresses the general risks involved in the use of BD. Consequently, data collectors and analysts face new moral dilemmas and today's scientists must address the challenges that emerge in applications of BD in both research and practice. This article addresses the ethical questions that arise particularly in social research settings. It asks which ethical and legal aspects need to be considered when collecting and analyzing data on individuals from the web and combining them to gain an enriched picture of human activities. While there is an emerging body of literature on the ethics of BD in general, literature specifically dealing with the *research* ethics of BD is still scarce (cf. Morena et al. 2013, Williams 2017). The article aims to help fill this

[2] 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."

gap by outlining some of the challenges that come with the use of BD in social research. While it seems impossible to give definitive answers to such challenges without a wider discussion of these issues with researchers and experts in the field, the article still formulates some generic recommendations and presents them for further debate.

2. Research Ethics: A Brief Overview

This section provides a brief overview of theoretical principles and existing regulations in the realm of research ethics, the legal requirements of data protection legislation, and existing codes that cover questions of BD. It lays the foundation for the later discussion that specifically addresses ethical issues in the use of BD for social research.

2.1 Existing Regulations and Theoretical Principles

For the purpose of this article, I will understand ethics as being concerned with the question of whether my conduct is justifiable toward other human beings, especially if my actions have an impact on them (cf. Fuchs et al. 2010, 4).[3] In the *deontological* view of ethics, other persons need to be considered when we contemplate our actions as other persons are an end in themselves and bearers of some irrevocable basic rights. In the *consequential* or *teleological* view of ethics, we especially need to consider the consequences of our actions and how they affect others, whether these consequences are intended or not (cf. Fuchs et al. 2010, 15). Here, the notion of risk becomes important and the assessment of potential risks of one's own conduct becomes an ethical virtue in itself (see Salganik 2018 for a brief discussion of overarching ethical principles).

In the social sciences, it is important to acknowledge that the subjects of research are humans who have the right to be protected from harmful conduct. It may be argued that the main intention of research ethics is the protection of its research participants, i.e., the subjects under study, from being harmed through any element of scientific inquiry. While wider ethical consideration, e.g., about the use of scientific discoveries, are pertinent to all scientific disciplines (natural scientists have often struggled with the question of whether their discoveries may be used for the development of new arms for example), this issue is specific to disciplines which study human beings. The questions of the extent to

[3] Another issue concerns the ethical handling of other social entities such as companies and organizations or communities. While human subjects should be the main focus of concern, others argue that other social entities should also be included in our considerations. For example, some German courts have applied data privacy law to judicial persons such as companies and organizations, although the laws were written with natural persons in mind.

which humans are subjected to research practices, and of the risks of harming those who are involved, itself becomes an important ethical issue (Buchanan and Zimmer 2018).

While many ethical considerations have become enshrined into law, as is the case for example with data privacy legislation, this is not always the case. Quite often, societal developments and technological advancements are much faster than the legislative process and hence there typically is the need for ethical contemplation, which goes beyond the existing legal framework. Additionally, ethical considerations are not and cannot be the same as established codes of ethics, because otherwise, the implementation of new rules or the criticism of existing rules would be impossible. In this view, many important questions of ethical behavior in science only start beyond what the law requires in any event.[4] Ethical behavior becomes everything that should be done according to some ethical standards but that is not legally binding or already codified into standard practice (cf. RatSWD 2017b, 15). Still, this makes existing frameworks of data protection and ethical research a reasonable starting point for the discussion.

Research ethics now have a longstanding tradition in human subject research (cf. von Unger and Simon 2016). Generally, ethical guidelines are especially important and prominent in epidemiology and health research, a consequence of the cruel and deadly treatment that individuals have received historically in the name of the advancement of science. One major motivation for the development of ethical research codes were the atrocities committed by doctors in Nazi concentration camps in the name of science during World War II. A result of this was the establishment of the Nuremberg code, which laid out basic principles to protect human life especially in the medical sciences which received a further level of codification in the *declaration of Helsinki*,[5] binding medical researchers to the well-being of their research subjects. In the US, the infamous, racist Tuskegee Syphilis study –where subjects were left untreated in order to study its long-term effects – also led to a review of ethical practices in clinical research and the establishment of institutional review boards (IRBs) to oversee such research (cf. Salganik 2018, 326).[6] The investigation into this and

[4] Mostly, law only regulates questions of data protection and privacy. However, this may vary widely between countries.

[5] <https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/>.

[6] IRBs are more prevalent in Anglo-Saxon Countries and often mandatory for all kinds of research involving human beings. In Europe, the requirements regarding the necessity of an IRB review are much less strict and vary widely between disciplines. For example, they are common in psychology at the institutional level. In German sociology, there is one review board run by the DGS where researchers may submit their proposals and research ideas voluntarily. However, there is a debate about how suited IRBs are for qualitative research in principle (e.g., Hammersley 2008; Von Unger et al. 2016).

other studies led to the *Belmont Report*, that further established many of the rules which nowadays govern what the report called "Human subjects research." The report defined three main principles: respect for persons, beneficence, and justice (Hoyle et al. 2002). Respect for persons entails that people should be treated as autonomous agents and establishes the concept of informed consent (IC) according to which people should choose to participate in research voluntarily and freely, based on all relevant information especially about the potential risks involved. The principle of beneficence sets out that researchers should do no harm and protect their participants, providing potential benefits and minimizing risks and potentially negative outcomes, which often necessitates a risk-benefit analysis before the research is conducted. The third principle of justice states that the selection of participants should assure that risks and benefits are distributed fairly and do not unduly burden groups of people already disadvantaged in society.

One should keep in mind that these principles, updated and refined in what is called the Common Rule, widely govern the ethical practice of research and the review practice of IRBs, especially in Northern America, have their origins in guidelines for medical research. Compared to this, social sciences typically involve low-impact studies usually without risks of lasting physical or psychological harm (Kämper 2016; Metcalf 2016). It should be noted that these standards and their application through IRBs have not gone uncontested (Dingwall 2008). One argument put forward, especially by qualitative researchers, states that such a standardized approach is not suited to the flexible arrangements often encountered in (qualitative) social research (e.g., Haggerty 2004; von Unger 2014). It has also been questioned to what extent they are suited to guide internet research because here the human subjects research model may not be entirely appropriate (Bassett and K. O'Riordan 2002). Here, it is important to consider whether research deals with (active) participants which in turn depends on the level of involvement of those researched (Keller and Lee 2003).

Before I turn to the more generic questions of research ethics specifically in the context of BD, I briefly describe the legal requirements enshrined in data privacy law as a baseline for discussion. This will help to focus this article on issues that scientific communities want to address voluntarily rather than what is already obligatory given the legal context. Existing codes of research ethics will also be discussed to see whether key concepts from the "old days" of scientific research and the handling of individual and sensitive data are still valid and important in the times of "BD," even if only to a smaller degree. Such concepts include, among others, privacy, anonymity, public availability, and IC.

2.2 Data Protection and Privacy Regulation

This section attempts a brief overview of the rules set out in existing data protection regulations. Such regulations are ethical requirements enshrined into law and build the baseline for any ethical research that needs to be known and adhered to in any research setting. In this sense, it is clearly important to understand what they entail.[7] Traditionally, data protection holds a special place in the German legislature in which the right to personal data sovereignty (*informationelle Selbstbestimmung*) even has the status of a fundamental constitutional right (Mühlichen 2014). However, data privacy law in Europe is now mainly regulated by the new EU general data protection regulation (EU-GDPR). In May 2018, the EU-GDPR came into effect and set out the overarching principles for data protection in all EU countries. Its introduction has proved influential even beyond the EU, an instance of what has been called the "Brussels Effect" (Bradford 2012). Even American companies have implemented GDPR- principles into their services throughout the world and not just for users in Europe.[8]

There is the conceptual question of what actually constitutes personal information on the one hand and what is meant by the identification of a person on the other. Many internet services, for example, track users online when they surf the web and visit other websites and services, enabling them to gather information on these surfing habits and to show personalized advertisements. These techniques, often by the use of cookies,[9] make the person identifiable online even though not necessarily in the traditional sense of connecting names and addresses. While the use of cookies can be monitored and is controllable at least theoretically, other techniques such as "fingerprinting" fulfill the same purpose without users' consent or knowledge simply by using identifying information from any machines soft- and hardware components that cannot be hidden for technical reasons. Hence, people are always identifiable online, even if they are not necessarily connected to the offline world. The answers to these questions determine what kind of information needs to be protected and how it can be protected by anonymizing data. This is also important from a technical

[7] For a good introduction to the situation in Germany, including changes implemented through the EU-GDPR, see Schar (2016) and RatSWD (2017).

[8] While the GDPR implements overarching rules which now apply internationally, it grants national legislatures the authority to establish their own rules in certain areas (cf. Schaar 2016, 10). Hence, it will still be necessary to consult national data protection law, for example in the relationship between employers and their employees. This section will, therefore, provide a brief introduction into the core principles of the concept of personal data, IC, and the stated purpose for using personal data as stated in the EU-GDPR, complemented with information from the German data protection law where it seems informative.

[9] A cookie is a small file that is placed on visitor's computer when she visits a website. With the information stored in the file, which are often used to continue web-sessions started previously, it is possible to re-identify returning visitors, and to track their activities.

point of view because if researchers are to combine large amounts of data from different sources for the same individuals they need to be sure that identification works. If they do not get the correct person for data linkage this will lead to faulty data and probably to incorrect conclusions. All this begs the question of how personal information is defined, which is itself tied to the possibility of identifying a person. In short, and similarly, across countries and continents, private data protected by judicial regulation is such data that allows for the identification of a natural human being as well as the personal information about those human beings identified in the research process. According to GDPR (Art. 4, 1), "'personal data' means any information relating to an identified or identifiable natural person ('data subject')." The GDPR specifically protects certain types of personal data "revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation" (GDPR, art. 9, para. 1). It should be noted that these kinds of data are often the focus of empirical social research.

One important provision of legal data protection obligations is the limitation of purpose for the usage of the data. That data must be used for specified, explicit and legitimate purposes only, with a prohibition of general data retention (*Vorratsdatenspeicherung*). Information on purpose and proceedings must be provided "in a concise, transparent, intelligible and easily accessible form, using clear and plain language" (GDPR, art. 12, para. 1), so that consent can be given freely and be withdrawn at any time (GDPR, art. 7). While the purpose limitation in the EU-GDPR is somewhat wider than it used to be in German legislation,[10] this typically still means that personal data must only be used for the research project they were collected for and not shared with third parties. It also follows that all personal data needs to be deleted after it was used for the scientific purpose specified and communicated to the research participants (commonly, this is the end date of the research project at the latest). If personal data is to be used for other purposes or to be shared with third parties, this necessitates further IC from the participants.

2.3 Existing Ethical Guidelines for the Use of Big Data

Discussions about the threats that BD poses to individuals and societies abound. Hence, the literature on the ethical issues of BD and how to handle them is growing. The main topic in ethical debates around BD is the intrusion of privacy. Often, it is argued that the sprawling exploitation of private data for profitability and state-surveillance threatens freedom and autonomy. However,

[10] Compare Schaar (2016, 5) for the changes in purpose limitation and the benefits for research it brings.

while general ethical considerations regarding the use of BD are broad and all-encompassing, issues of research ethics are more fine-grained and specific and, therefore, much different from general ethical considerations in this area. One reason is the privilege of science mentioned above. There is a general public interest in good science and properly grounded scientific knowledge that stands against the individuals' claims to complete privacy. However, the literature on the specific topic of research ethics and BD is much less prevalent and so is its codification in ethical guidelines for scientific research. BD is hardly mentioned in existing codes on the research ethics of most social science associations (at least the ones we checked from the US, the UK, and Germany).[11] I continue with a brief overview of the existing discussion of research ethics in BD. This may start with guidelines provided by the British Sociological Association (BSA), which is a positive exemption in this respect.[12] They not only have developed but also published a series of case studies exemplifying and discussing ethical dilemmas coming out of online and BD research.

Another positive example is the ethical guidelines from the Association of Internet Researchers (AoIR) who have just recently updated them (franzke et al. 2020). According to Lomborg (2013), they

> advocate a bottom-up, case-based approach to research ethics, one that emphasizes that ethical judgment must be based on a sensible examination of the unique object and circumstances of a study, its research questions, the data involved, and the expected analysis and reporting of results, along with the possible ethical dilemmas arising from the case.

However, this approach has also drawn criticism (Eynon and Schroeder 2016) and overall, there have been too few debates around these issues. As a result, the research ethics of BD are contested and underdeveloped. As Lazer and Radford (2017) put it:

> The problem, however, is that there is no consensus on what the rules should be, and the policies and recommendations set forth by scientific associations vary substantially, often contradicting one another. Rules will eventually become clear, but the risks to researchers, universities, and the public remain high until they do.

[11] This at least holds for the German research foundation (DFG), the German Political Science Association (GPSA), the German sociological association (DGS) as well as its counterpart after the recent schism, the Academy for Sociology (AS), but also even for the American Association for Public Opinion Research (AAPOR).

[12] It seems that researchers in the UK overall are quicker to react to the new challenges as the British Psychological Society (BPS) and the British Society of Criminology (BSC) have also updated their guidelines to include online research (<http://www.bps.org.uk/system/files/images/2012_ethics_committee_social_media.pdf> respectively <http://www.britsoccrim.org/documents/BSCEthics2015.pdf>)

3. Ethical Issues of BD in Social Research

This section is going to discuss the ethical challenges that follow from the innovations of BD for social research. These new challenges stem from the specific features of BD and beg important ethical questions, especially regarding the use of personal information of individuals. Kaisler et al. (2013) see, amongst others, the following compliance challenges when using BD: "What rules and regulations should exist for prohibiting the collection and storage of data about individuals – either centralized or distributed?" and "What rules and regulations should exist regarding combining data from multiple sources about individuals into a single repository?" As Lazer and Radford (2017) observe, "there are major ethical issues regarding the acquisition and use of BD for researchers, institutions, and society at large." However, they also point out that "some issues are new, but many are new versions of long-standing issues." This may warrant the assumption that standard ethical guidelines governing social research may need adapting in the face of the availability of BD, while at the same time some of the practices that have been proven useful in the past are suited to cover ethical problems involving BD as well. When discussing these issues we must also consider the peculiarities of textual, "qualitative" BD, as they may be different from more quantitative data, for example in the possibility to anonymize such data.

3.1 New Ethical Challenges for Social Science Research

It seems useful to identify the ethical challenges of BD in the light of the new opportunities discussed in the previous section and the four key features of BD: volume, variety, velocity, and value. For example, while the variety of BD sources that are potentially of interest for social scientists is vast, this variety of BD is a challenge in itself. Many of the ethical issues involving BD will have to deal with questions of personal data and the threats BD brings to anonymity and confidentiality through new possibilities of re-identification (cf. Barocas and Nissenbaum 2014; Bender at al. 2016,). However, there are many different kinds of data which vary greatly and hence may not be treated equally under a unified ethics framework but rather require special solutions for different kind of scenarios. It is even possible that the attempt to establish one overarching form of research ethics for BD as such may be futile. Where the nature and content of the data are so diverse, the ethical challenges related to this data may be similarly varied.

One example of crowd-sourced personal information from the web is personal information from social networks such as Facebook (or Myspace in earlier days) or the Russian or Chinese equivalents. This may encompass biographical information as well as personal interests, political leanings, and even activities collected via personally shared content or the clicks and likes of the

content provided by others – be they commercial distributors (advertisers, news agencies) or other private individuals. In addition to that, a simple but important piece of information is the network of people that an individual is connected to. Such data provide rich material for many different research questions for a range of disciplines from politics and economics to sociology and psychology. Precisely because this material is rich and detailed and may allow an intimate assessment of an individual's personality and values, it is also highly sensitive in nature.

Combining data points to gain insights creates another set of problems. For example, data linkage of different sources almost by definition requires the (re-)identification of individuals in order for the sources to be linked. Given the quality and depth of the data at hand, this not only poses a huge challenge to be done accurately, but also poses a great risk if linkage goes wrong and data is matched which do not pertain to the same individual. This may result in false conclusions about connections in the population and false claims about individuals with possibly damaging results. Another challenge relates to the question of how such datasets, in which information on individuals is integrated from a range of different sources, may be anonymized to allow for archiving, re-use, secondary analysis, and the confirmation of findings in replication studies.

Another huge problem consists of personal information that is not provided by subjects themselves but by third parties such as family, friends, employers, or even state agencies. A simple example are smartphone apps who ask for access to one's contact details in order to check whether friends or family already use the same service. While possibly helpful, it constitutes unauthorized sharing of personal information (the contact details) to a third party that may easily be used to identify persons and link them to other online profiles. Without one's knowledge, the circulation of one's contact details may be widespread.

Social networking sites are a particular challenge for privacy, not only because of the often dubious practices of using personal data by the platform providers, but also because of the rich amount of information on personal profiles and the high connectedness with other people who often share similar traits. For example, even though an individual may not want to disclose their last name or general family affiliation on a social networking site, one might deduce it from close relatives who they are connected to on the platform and who have not anonymized their personal information. A related challenge is the willingness of some individuals to disclose information about third parties in comments and statements, but also by tagging friends in a photo for example. Even for people who are not even members of a platform, by using social network theory and some probabilistic theorizing, one may deduce their personal network and connections if at least some of their friends and families are on the platform. In addition, analysts can make very educated guesses, simply based

on statistical models, about cultural and political preferences for people who share the more easily observable traits that correlate with the hidden traits. Finally, many mobile phones collect data and provide it to Facebook, for example, which in this way collects personal information on hundreds of millions of people without them even being on Facebook (Hoppensedt 2018).

One of the new challenges for research ethics is the fact that information is bound to stay accessible for anyone at any time for the foreseeable future once it has been made public. This increases the risk of de-anonymization but also the scope of potential analyses that could not have been foreseen at the time of the creation of the data. Sources and data types that previously were anonymous now become searchable, linkable, and personalized. This is problematic because it was probably impossible for individuals to know back then what the information they had provided on the internet is now likely or possible to be used for, that it is still available after a long time and that it even may be linked to a range of other data that can be found on the web. In another vein, face recognition software turns anonymous snapshots into very specific personalized information that may be easily linked to other personal information. This may pose an even greater threat when historical archives become digitalized and old pictures can be searched for the persons who are depicted.[13]

Clearly, before using data for research, it is important to clarify which type of data we are concerned with and where the data comes from. This is meant geographically, as this will govern the rules and laws that have to be followed. This also concerns the ultimate source of the data. At times, this may be difficult to establish because the online site where we as researchers encounter the data may not be the original source of the data. The latter will, however, determine whether the original data was meant to be shared and to be distributed on the web (through user agreements, for example) or whether it simply appeared in other places erroneously or even maliciously. Pictures are an obvious example that people can copy easily to appear on other websites (in an extreme form, to fake a personal account) but which never have been meant to be shared or copied. Another level of complexity is added by the temporal dimension, as data is increasingly available from previous years or even decades as archives, newspapers, books, and other types of data typically stored in paper format become digitized through either scanning and ORC (or even manual keying). This way, information becomes available to data analysts and may be linked to other sources which may have been inconceivable at the time those texts were produced. An interesting example are newspaper articles. In those, individuals may reveal substantial personal information but not specifically for

[13] This threat has become much greater in recent years as the amount of pictures taken is now by far greater than it used to be previously thanks to digital cameras, smart phones, and increased storage capacity. Also, the resolution of pictures is much greater than it used to be.

the purpose of the data being collected, compiled, and used for research purposes. Such instances beg the question of whether IC can be assumed or whether it is even necessary given that people freely provided this information to a news outlet where it must have been clear that the information would be made available to the public. For individuals, at the time it would have been impossible to know how far this information could be available electronically and could be linked to other sources and hence a realistic assessment of impact and risk of collecting and providing information was probably impossible. This is somewhat mitigated by the very fact that some of this information may be old by now. However, while some information loses its sensitivity over time, other information, such as a criminal conviction, for example, may not.

Data and services easily cross national boundaries, which makes it harder to establish legal rules and to police them. While many principles are valid universally in all social research around the globe, there will always be specific topics to address given the country the research is conducted in. Established ethical traditions will differ between cultures and nations states. Germany, for example, has a unique tradition of protecting personal information, partly through their historical experience with two different dictatorships in the past century and the unique conception of human dignity as the anchor and most important principle of its constitution. However, what is also important for questions of scientific ethics, scientific freedom also ranks as a constitutional principle with spill-over effects into the private interactions of citizens. These constellations, where both the right for informational self-determination and academic freedom enjoy constitutional status, may be unique and hence not generalizable across countries; the backing of ethical principles in ethical traditions may not be the same everywhere. This, however, becomes more and more difficult when using web-sourced data as these travel easily and are hardly bound by the boundaries of nation-states. However, rules, especially legal ones, may still be bound by territory. For example, in some countries, data protection laws only cover natural persons, while in others they protect judicial persons as well. This distinction may be very consequential for researchers especially when they are interested in firms and organizations. This raises the question of which regulations you need to comply with when using BD and whether researchers actually have a real chance of knowing which rules from which jurisdictions are all invoked and need consideration in their research.

3.2 The Delineation of Human Subject Research

The question of when BD actually deals with human subjects is central in determining whether and when ethical considerations are necessary. In BD research, humans primarily appear and are represented as data, in their digital correlates. However, there are many BD applications that do not involve data on humans and with some forms of data, the level of human involvement is

marginal. Many of those capture social processes at another level than the individual. Examples may involve the development of stock prices around the world, the tracking of trucks in automated toll systems for real-time forecasting of GDP developments, or the extraction of rental housing market information from websites and dedicated portals to estimate the development of rents over time. Also, individual-level data, such as terms searched for in online search engines, may be used for aggregated level results, such as forecasting the outbreak of influenza or measles. Hence, BD is present and may be accessed for many different social units and at many different levels of aggregation, all of which may have different ethical implications. The clearest and strongest implications, however, pertain to the individual as a bearer of rights and an end in themselves that needs to be considered when doing research. How can human subject research be defined under such circumstances? Currently, according to 45 CFR 46,[14] a human subject is "a living individual about whom an investigator (whether professional or student) conducting research:

- Obtains information or biospecimens through intervention or interaction with the individual, and uses, studies, or analyzes the information or biospecimens; or
- Obtains, uses, studies, analyzes, or generates identifiable private information or identifiable biospecimens."[15]

Thus, human subject research is tied to either intervention or interaction with participants or the use of identifiable private information. Hence, anonymized data or data on an aggregate level that was obtained without such intervention or interaction with an individual would not fall under the rules of human subject research. One may argue that this is still too restrictive and that the involvement of human subjects or the potential risk involved needs to pass a certain threshold for the rules of human subject research to kick in (i.e., there needs to be a certain severity of intrusion into the dealings of the research participants and their state of mind and affairs – compare the discussion around the Facebook experiment on emotional contagion; Salganik 2018). Whatever the final definition of human subject research will finally look like in the area of BD research, it is clear that the first consideration of ethical issues needs to be to answer the question of whether and to what extent humans are involved, as this will govern the following steps and the application of ethical rules and guidelines.

[14] <https://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/index.html>.

[15] Until 2018, the Common Rule defined human subjects as follows: "(f) Human subject means a living individual about whom an investigator (whether professional or student) conducting research obtains (1) Data through intervention or interaction with the individual, or (2) Identifiable private information" (Metcalf 2016).

3.3 The Conundrum of Informed Consent

The question of IC for doing social research likely needs to be answered afresh with BD in its various forms (cf. Barocas and Nissenbaum 2014). As discussed above, participation in scientific research should happen voluntarily without force. IC has been established as a principle to assure exactly that: a person consents to take part in a study based on his or her free will, which requires an adequate understanding of the procedures and risks that this involves. Asking participants in a study for IC has, therefore, become a key requirement in ethical research and, in fact, a legal requirement for the use of personal data. IC typically covers two aspects of research practices separately: First, participation in a scientific study, which is rooted in the assumption that participation in a scientific study should be voluntary and must not be forced; second, the permission to use personal data for research purposes. While closely related (it is hardly possible to participate in a study without providing any kind of personal information), these two points are not identical. In recent years, it has also been argued that researchers may need additional consent for the storage of research data in scientific data archives, and that other researchers may use the data for secondary analysis. This third point is based on data privacy regulations and expresses the idea that personal data should only be collected for predefined purposes and for a limited amount of time, i.e., until the data was used for this purpose.

This understanding of IC brings up the question of whether all usages of human-related BD falls under the necessity of acquiring IC. One may argue that IC is not necessary if participants are not affected by the collection of data or the usage of the data. In this perspective, IC would not be necessary as long as the research is only passive and no active steps or efforts are required on behalf of the human subjects themselves. In a way, there would be nothing to consent to because the individual person is not actively involved in the research and does not have to fear any risks from the usage of their data. This directly leads to the precondition for the second aspect of IC, whether the data at hand entails personal information that invokes data privacy issues. Even if data somehow stems from human subjects, it is important to assess whether personal data is involved or not, as many research questions do not involve such data. For example, analyses at the country level may use data that originated from individual-level data but may no longer be personal when aggregated to a higher level. Many other social entities (e.g., organizations or networks) may be interesting to study for social scientists, but do not directly involve person-level information. However, if they do, IC in the sense of data privacy protection is necessary.

Besides the question of necessity, the practical questions involved in obtaining IC in BD are quite challenging. If the data is collected newly by researchers themselves, they have the opportunity to ask for IC properly and definitely

should do so (if deemed necessary). Data collection via specialized tools on social networks, for example, may reach millions of respondents without the need for subsampling and the possibility to collect IC beforehand. However, most instances of the use of BD will be cases where data has already been collected by other actors. Here, it seems impractical and even impossible, given the scale of the datasets, to ask for IC retroactively (in many cases there will be no way to contact the persons involved anyway). Therefore, the question becomes whether the original data providers (be they a social network or a movie rating website) have obtained IC properly and whether these ICs cover the case of third parties using the data, maybe specifically mentioning the case of scientific research. Here we encounter the well-known problem that the formulations of ICs, as part of user-agreements in general, are regularly hard to understand or even outright misleading (cf. Wehofsits 2016). Especially in the past, commercial data often exhibited a laissez-fair approach to data protection issues, which is in part understandable as many business models depend on gathering as much personal information as possible. For example, on Facebook, for a long time, the status of personal information was set to "public" as default and could, therefore, be found freely on the internet through search engines or other means. This needed users to become active and change the settings if they were not wanted, but fortunately Facebook has changed this default since. Thus, whether people understand the extent to which their personal data is publicly available, collected, and shared with third parties is quite doubtful. Here, one is confronted with the very difficult problem that we know user agreements are typically ignored by people using online services who are therefore not aware of the potential sharing of their personal information. Whether such data should still be used for scientific purposes under such circumstances becomes doubtful too. The situation seems to have changed for the better, however, as new legislation (notably the EU-GDPR) has been put in place and public scrutiny regarding the data practices of big tech companies has increased due to a wide range of data breach scandals. However, it is still an open question as to whether ICs can ensure an informed decision to freely participate in research when BD may be used for a wide range of scientific "secondary" analyses with research questions that will only be formed in the future (cf. Wagner 2017, 4).

3.4 The Question of Sources and Public Availability

The discussion of IC directly leads to the question of the sources of the data and whether it has been obtained properly and lawfully. Just because something is available does not mean that it should be used in research. Here, one needs to consider whether researchers collect data themselves or whether they use data that has been collected by others, e.g., social networks, already. However, the benefits of BD are likely to arise with data that already has been collected because researchers typically do not possess the means to collect BD them-

selves. Sometimes, they still do - the myPersonality project, for example, was able to collect data on the personality traits of six million Facebook users via an app that would run a personality test between 2007 and 2012 (Stilwell and Kosinski 2012). This would be an instance of data collection by researchers on a scale that amounts to BD in our sense, but such instances do not occur often. Rather, typically, when we think of BD in the social sciences we think of the usage of data collected by others, which then amounts to some form of secondary data analysis. This is important because ethical duties are likely to differ between the collectors of data and its users for secondary purposes, and typically the former will be confronted with higher ethical expectations than the latter (cf. RatSWD 2017b, 25). However, I will argue that researchers who use BD collected by others have an ethical obligation concerning the question of whether the data has been collected ethically and lawfully in the first place. If this is not the case, or if the proper sources of such data cannot be obtained, it should not be used for research purposes.

When considering data sources, it is also important to establish whether they are strictly private or not. For data that is already public, the question of IC becomes irrelevant. It is therefore essential to be clear about what can be safely assumed to be public information and therefore accessible for research. For example, tweets on Twitter, a microblogging platform, are meant to be public and accessible by anyone and therefore can be used for research purposes.[16] While you need to have an account to post something, everyone can read what is tweeted even without having an account. However, it should be clear that, whenever some information is *only* accessible after logging on to a certain platform or service, it should not be considered public anymore. The matter quickly becomes complex, as the level of publicness of a datum can vary widely and change quickly, even within the same social network (see Williams et al. 2017 for a thorough discussion of privacy issues regarding using Twitter data, including research on users' views on these issues). Users' perception of what is meant to be public and private varies as well (Sugiura et al. 2017). For example, some content may be accessible by anyone, some only by members of the same platform, some others only by people marked as friends or acquaintances. A glimpse into the Facebook Data Policy shows how complex it can get:

> Public information is any information you share with a public audience, as well as information in your Public Profile, or content you share on a Facebook Page or another public forum. Public information is available to anyone on or off our Services and can be seen or accessed through online search engines, APIs, and offline media, such as on TV.

[16] "Terms of service specifically state users' posts that are public will be made available to third parties, and by accepting these terms users legally consent to this" (Williams et al. 2017, see also Twitter 2015, 2016).

> In some cases, people you share and communicate with may download or re-share this content with others on and off our Services. When you comment on another person's post or like their content on Facebook, that person decides the audience who can see your comment or like. If their audience is public, your comment will also be public.[17]

Thus, it is very easy to inadvertently share some personal preferences publicly by liking something else that is public without you really noticing. Also, it acknowledges the danger of other private actors involuntarily sharing some of your private content to a wider audience without realizing their mistake. Thus, there is an inherent risk in sharing material digitally as you may lose control over who will be able to see it and who will not. It seems that only recently people have become more aware of these issues and they are now more careful about what they share online.

However, not only contemporary online sources are relevant here. A case in point are old newspaper articles where it is clear that the information they contain was intended for a public audience. However, people cannot have known that the information they provided, maybe in an interview 20 or 30 years ago, would still be available at some point and would be even easier to find due to the digitalization of newspaper archives. Thus, the consent to provide information was not "informed" as the full scope of risks and consequences was not known at the time. However, this has changed – one may argue that nowadays people can be expected to assume that all information they provide to a newspaper will become searchable online eventually. Thus, this concern is mostly relevant for information that was provided to analog and offline media before BD and the Internet became an essential part of everybody's life.

As the discussion makes clear, it becomes essential for researchers to clarify which data is accessible publicly as this may be used for research without the need to obtain consent. However, this necessitates a clear understanding of what constitutes the "public" realm, which again differs between countries. While existing concepts such as the "public domain" may provide guidance, they originate in other contexts, such as copyright legislation. Hence, there is a pressing need for further clarification of the question of what information constitutes public data and what information does not.

3.5 The Role of Risks and Risk Assessment

Currently, it is common practice (at least in the US) "that IRB's are thus currently tasked with reviewing any research that risks harm to an individual person which the researcher is interacting or intervening with in order to collect data" (Metcalf 2016). Therefore, assessing the potential risk for individuals subjected to the research is common practice (and could be considered a cate-

[17] <https://www.facebook.com/full_data_use_policy>.

gorical, general principle). I argue that the risk involved with BD is a function of its value which is why I discuss risk in relation to the outline of the value of BD given above. In addition, the level of risk may be thought of as a combination of severity and likelihood. Some risks may occur more frequently but are very minor in their consequence. Other risks involve severe threats to an individual, but at the same they are not very likely. Accordingly, risk levels are especially high for risks with severe consequences and a high likelihood of occurrence (when no measures are undertaken to prevent this). Thus, any risk analysis will need to consider these two elements.

Typically, the potential harm (and its severity) involving web-sourced data is associated with the risk of (unintentionally) disclosing personal information. This may be harmful in different ways without being legally threatening. As Kaisler et al. (2013) put it:

> Perhaps the biggest threat to personal security is the unregulated accumulation of data by numerous social media companies. This data represents a severe security concern, especially when many individuals so willingly surrender such information. Questions of accuracy, dissemination, expiration, and access abound.

One example is the aforementioned data dump of the US adultery website Ashley Madison where the data was used to threaten and blackmail people by exposing them to their friends and families (Zetter 2015). Others include sensors and cameras of self-driving cars used to scan the environment – in combination with face recognition software this may allow the generation of mobility profiles of individuals without them even knowing. This opens the doors wide to the misuse of such data. That such things happen has been reported repeatedly. Amazon's Alexa, a smart speaker combined with artificial intelligence to act as an AI-driven personal assistant, has already been decried as an espionage device in the private home. Through its built-in microphones that record constantly the inner surroundings, it gathers everything that is being said in a room and stores it online. While Amazon has pledged to respect privacy, it was reported that employees were listening in on private conversations during their work (improving the AI software driving the smart speaker) just for fun (Hern 2019). Such examples show that misuse of such data happens easily, is possibly widespread, and difficult to prevent.

Overall, the threat from scientific research is likely to be smaller than general threats by other actors who likely have access to the same amount of information. To put it differently, risks involving the uses of BD in social research are comparatively low when compared with those risks people face online through their daily activities otherwise (and it is likely smaller than the

risks involved in life science or bio-medical science, cf. Kämper 2016).[18] Scientists are typically not malicious actors with criminal intent and are bound, or should be bound, by legal agreements to prevent the privacy of their subjects. If harm is done, this is likely to be unintentional rather than intentional due to carelessness and/or ignorance. This, however, does not mean that the researcher should not care about the risks involved for the subjects of their research. It does put this risk into a certain perspective, however. In addition, as long as research only uses information that is openly accessible and usable by others also, scientists actually do not add any risk to that which already exists. This may change if the data is processed and sources and combined. In such instances, however, questions of data distribution and dissemination arise and should be handled accordingly so as to not increase existing threats to privacy.

4. Discussion and Initial Recommendations

This article set out to identify and discuss some of the key ethical challenges that social researchers are confronted with in their use of BD. For this purpose, it reviewed existing literature on BD ethics and ethical guidelines on the use of BD in social research and briefly assessed the nature of BD to identify the potential benefits for social research. It discussed ethical pitfalls which follow from these research potentials and identified issues that warrant further discussion in the near future and the outline of what needs to be discussed, and what can be handled by established codes and practices has become clearer. For example, the major problems regarding researching human subjects and their data in this area derive from the fact that the data is already collected by some third parties like social networks. Also, it appears that the level of ethical precaution needs to be correlated with the level of impact and risks that the research practices bear on the subjects under study.

In the following, I present a short list of suggestions, proposals, and guidelines for social researchers using BD that are not necessarily covered in existing codes of ethics.[19] Most of them are simple and straightforward, invoke common sense, and are relatively easy to follow. Some of them are simple extensions of existing best practice examples to this new realm of data and may appear obvious. Others are more specific to the peculiarities of web-generated, mass data.

[18] The ethics code of the German Sociological Association (DGS), that those risks which go beyond what is common in everyday life require specific attention (DGS & BDS 2014, § 2 (5)).

[19] The order of the list is quite arbitrary and not necessarily indicative of impact or importance.

- Existing formal requirements need to be followed. Besides the professional codes of research ethics, there are obvious legal frameworks that may overlap with ethical questions (such as data protection law or copyright law) that researchers need to abide by. In addition, there are user agreements and terms of services of websites and platforms that researchers need to acknowledge if they want to use those services.
- The extent to which any given research is about human subjects should be clearly stated and defined. This is important as the outcome will govern the set of rules that will need to be considered and followed, and, potentially, whether an IRB review is necessary or not.
- Only data from "proper" sources should be used (if researchers do not collect data themselves) and the sources should be documented. To determine whether sources are "proper," researchers need to make sure that the data is legally available for research purposes. They also need to check that the original data collectors behaved properly and ethically when they collected the data. Finally, they should check the data privacy regulations and consent procedures used by the data providers. Researchers should only use the data if both seem satisfactory.
- The risks involved for research subjects should be assessed and a risk-benefit analysis should be conducted. While a "data protection impact assessment" has become obligatory under EU-law if new technologies are applied or certain types of sensitive data are to be processed (GDPR, art. 35), risk assessment should be done in any case. Such an assessment should consider who may be affected (not just persons subjected to research, but also providers of infrastructures and fellow scientists) as well as the severity and the likelihood of each risk involved. Rules and guidelines should be developed, and tools should be put in place that help researchers to conduct a standardized risk assessment.
- Territoriality needs consideration in BD research as well. While guidelines and codes are mostly bound locally, BD research often involves the use of online data that easily travels across borders and jurisdictions. Consequently, researchers need to check carefully which jurisdiction(s) they fall under so that it is always clear which rules apply and which do not. Also, researchers must avoid conducting their research in countries with lower ethical standards simply to circumvent those rules. To prevent such cases and ease the application of ethics rules, international cooperation for the definition of ethical standards is required.
- The development of ethical guidelines and tools for research using BD in the social sciences is in itself an ethical requirement. Given the dynamic of the field of computational social sciences and the ever-changing landscape of social networks, platforms, and other possible data sources, this is likely to be an ongoing task for the foreseeable future.

- There should be some form of institutionalization of research ethics, as self-contemplation of individual researchers about the ethical implications of their research is laudable, but not sufficient. As in other areas, self-governance at the level of the individual researcher or research team should be viewed critically (cf. RatSWD 2017b, 9). While institutionalization does not need to take the form of US-style IRBs, some form of ethics review of social research projects seems desirable (cf. Kämper 2016, 7; Wagner 2017). However, one should also think about an institutionalized body that researchers can turn to and that helps them to assess ethical challenges in their research. This aiding body would provide guidance and would keep the needs and interests of researchers in mind when they navigate existing institutional pitfalls in research ethics.

The recommendations above clearly do not present a definitive list of everything that needs to be considered in the area of BD ethics. Rather, they present discussion points on these matters. The question of further ethical guidelines for this emerging area of research is especially pressing, as there is a lack of ethics committees in the social sciences. Ethical approval procedures are less standardized (and required) here than in other disciplines such as psychology or health research and epidemiology. While this may reflect the fact that in general, there are fewer risks of personal harm from studies in sociology compared to psychology or medicine, it also puts a higher burden on the individual researchers as they have to establish for themselves whether their research is ethically proper or not. For example, rather than being an expert on research ethics in general, I am a researcher who is interested in using BD for his own research agendas but has found too little existing guidance in the field of BD research and research ethics. As there is still a gap regarding ethical guidelines around the use of BD in the social sciences, it becomes an ethical necessity to study such questions before one conducts such research. The development of ethical guidelines and frameworks becomes an ethical requirement in its own right as long as this is not achieved.

One should keep in mind that a discussion about research ethics is not about research *quality* and the development of ethical guidelines is not intended to produce better research as such (whatever the yardstick for that may be). If this was the case, they would not be ethical guidelines, but rather best practice guidelines. Rather, ethical guidelines probably narrow the scope of what researchers can and should do; guidelines should prevent researchers from doing something they would have done otherwise because they have convinced themselves it would be unethical to do so. Thus, following ethical principles in research may prevent you from realizing the ideal study design to accommodate the interests of your subjects and they may thus limit what you like to do. If this was not the case, no one would need ethical guidelines in the first

place.[20] Still, it must also be emphasized that ethical considerations should not prevent social researchers from using BD altogether. There is an intrinsic value to scientific research that needs to be weighed against the interest of the subjects of research in the search for scientific truth and insight (cf. RatSWD 2017, 6). BD is a promising and valuable tool for answering scientific questions and should remain that way even after ethical issues have been addressed.

5. Conclusion

One major conclusion from this discussion is that the topic of data privacy and the threats to personal data are maybe a lesser concern to research ethics than, for example, the commercial uses of BD. In addition, for the majority of data, social research will use BD in terms of secondary analysis and hence may only use data where subjects have provided consent for scientific uses already. This leaves the burden of assuring IC with the original data collectors. However, scientists need to check whether the original IC achieved was proper and included the right to share personal data with third parties, especially scientists. This, however, has consequences. It seems that from the vast amount of data that is out there, only a small fraction actually qualifies for its use in scientific research if one requires sources to be proper and the provision of consent to be acceptable. It also already greatly minimizes the risk of harm for subjects whose data is analyzed for research.

Finally, there is a pressing need for more discussion of research ethics in the field of BD and social research, as institutional codes of conduct should offer more and better guidelines in this area than they currently do. Overall, it is clear that we need further research, ranging from issues such as de-anonymization and re-identification (Sweeney 2002; Daries et al. 2014), data-sharing (Zimmer 2010; Borgman 2012; Bishop 2017), and data-ownership (Ruppert 2015) to the question of the appropriateness and manageability of IC procedures and the vulnerability of specific groups in large-scale online settings (Stopczynski et al. 2014) and institutionalized ethical review boards (von Unger et al., 2016). As long as these issues are not settled, the discussion of ethical issues involving BD and the development of guidelines for that matter remains an ethical requirement in itself. As the landscape of data that becomes available is still expanding and the tools for analyzing them are still developing so quickly, this may remain a continuous task for some time to come.

[20] Thinking about ethical issues in your research may even lead to better and more thoughtful studies because an extra level of careful considerations is needed about what you want to do beforehand. You may also help the wider research community by keeping or increasing the level of trust people have in social research.

References

Anderson, Chris 2008. The end of theory: The data deluge makes the scientific method obsolete. *Wired magazine* 16 (7): 16-07.

Barocas, Solon, and Nissenbaum, Hellen 2014. Big data's end run around anonymity and consent. In *Privacy, Big Data and the Public Good: Frameworks for Engagement,* ed. Lane, J., Stodden, V., Bender, S. and Nissenbaum, H., 44-75. New York: Cambridge University Press.

Basic Law for the Federal Republic of Germany in the revised version published in the Federal Law Gazette Part III, classification number 100-1, as last amended by Article 1 of the Act of 28 March 2019 (Federal Law Gazette I p. 404).

Bassett, Elisabeth H., and Kate O'Riordan. 2002. Ethics of Internet research: Contesting the human subject research model. *Ethics and Information Technology* (4) 3: 233-247.

Bender, Stefan, Jarmin, Ron S., Kreuter, Frauke, and Lane, Julia 2016. Privacy and Confidentiality. In *Big data and social science: A practical guide to methods and tools*, ed. Foster, Ian, Ghani, R., Jarmin, Ron S., Kreuter, Frauke, and Lane, Julia, Boca Raton, FL; London; New York: Chapman and Hall/CRC.

Bishop, Libby 2017. Big data and data sharing: Ethical issues. UK Data Service, UK Data Archive.

Borgman, Christine L. 2012. The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*. doi: 10.1002/asi.22634.

Bradford, Anu 2012. The Brussels Effect. *Northwestern University Law Review* 107 (1), *Columbia Law and Economics Working Paper* 533.

Buchanan, Elizabeth A. and Zimmer, Michael. 2012. Internet Research Ethics. *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. <https://plato.stanford.edu/archives/win2018/entries/ethics-internet-research/> (Accessed 21 February 2020).

HardcastleConfessore, Nicholas. 2018. Cambridge Analytica and Facebook: The Scandal and the Fallout So Far. *The New York Times,* April 4. <https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html> (Accessed February 21, 2020)

Crawford, Kate 2013. The hidden biases in big data. *Harvard Business Review* 1.

Daries, John P., Reich, Justin, Waldo, Jim, Young, Elise M., Whittinghill, Jonathan, Ho, Andrew D., Seaton, Daniel T., and Chuang, Isaac 2014. Privacy, anonymity, and big data in the social sciences. *Communications of the ACM* 57 (9): 56-63.

Dingwall, Robert 2008. The ethical case against ethical regulation in humanities and social science research. *21st Century Society* 3 (1).

Eynon, Rebecca, Fry, Jenny and Schroeder, Ralph 2016. The ethics of Internet research. In *SAGE Handbook of Online Research Methods*, eds. Fielding N, Lee RM and Black G. London: SAGE.

Franzke, Aline. s., Bechmann, Anja, Zimmer, Michael, Ess, Charles and the Association of Internet Researchers. 2020. *Internet Research: Ethical Guidelines 3.0*. <https://aoir.org/reports/ethics3.pdf> (Accessed 21 February 2020)

Foster, Ian, Ghani, Rayid, Jarmin, Ron S., Kreuter, Frauke, and Lane, Julia 2016. *Big data and social science: A practical guide to methods and tools*. Boca Raton, FL; London; New York. Chapman and Hall/CRC.

Fuchs, Michael, Heinemann, Thomas, Heinrichs, Bert, Hübner, Dietmar, Kipper, Jens, Rottländer, Kathrin, Runkel, Thomas, Spranger, Tade Matthias, Vermeulen, Verena and Völker-Albert, Moritz 2016. *Forschungsethik: Eine Einführung*. Stuttgart, J.B. Metzler.

Gantz, John and *Reinsel*, David 2011. *Extracting Value from Chaos*. IDC iView, International Data Corporation (IDC).

Haggerty, Kevin D. 2004. Ethics Creep: Governing Social Science Research in the Name of Ethics. *Qualitative Sociology* 27 (4).

Hammersley, Martin 2008. Against the ethicists: on the evils of ethical regulation. *International Journal of Social Research Methodology* 12 (3): 211-225. doi: 10.1080/13645570802170288.

Hern, Alex 2018. Amazon staff listen to customers' Alexa recordings, report says. *The Guardian,* April, 11. <https://www.theguardian.com/technology/2019/apr/11/amazon-staff-listen-to-customers-alexa-recordings-report-says> (Accessed February 21, 2020).

Vice, December 29. <https://www.vice.com/de/article/439ebb/facebook-sdk-datenweitergabe-auch-ohne-facebook-account-tracking-kayak-shazam-spotify>.

Hoyle, Rick H., Harris, Monica J., and Judd, Charles M. 2002. *Research methods in social relations*. Fort Worth, TX: Thomson Learning.

Kant, Immanuel 1981. *Grounding for the Metaphysics of Morals*, trans. J. Ellington. Indianapolis: Hackett.

Kämper, Eckard 2016. Risiken sozialwissenschaftlicher Forschung? Forschungsethik, Datenschutz und Schutz von Persönlichkeitsrechten in den Sozial-und Verhaltenswissenschaften. *RatSWD Working Paper* 225.

Kaisler, Stephen, Armour, Frank, Espinosa, J. Alberto, and Money, William. 2013. *Big data: Issues and challenges moving forward*. Paper presented at 46th Hawaii International Conference on System Sciences, January, Hawaii, USA.

Keller, Heidi E., and Sandra Lee. 2003. Ethical issues surrounding human participants research using the Internet. *Ethics & Behavior* 13 (3): 211-9.

Lazer, David, Kennedy, Ryan, King, Gary, and Vespignani, Alessandro 2014. The parable of Google Flu: traps in big data analysis. *Science* 343 (6176): 1203-1205.

Lazer, David, and Radford, Jason 2017. Data ex machina: introduction to big data. *Annual Review of Sociology* 43: 19-39.

Lomborg, Stine 2013. Personal internet archives and ethics. *Research Ethics* 9(1): 20-31.

Merton, Robert K. 1968. *Social theory and social structure*. New York: Simon and Schuster.

Metcalf, Jacob. 2016. Human-Subjects Protections and Big Data: Open Questions and Changing Landscapes. *Council for Big Data, Ethics, and Society*. <http://bdes.datasociety.net/council-output/human-subjects-protections-and-big-data-open-questions-and-changing-landscapes/> (Accessed: 21 February 2020).

Moreno, Megan A., Goniu, Natalie, Moreno, Peter S., and Diekema, Douglas 2013. Ethics of social media research: common concerns and practical considerations. *Cyberpsychology, behavior, and social networking* 16 (9): 708-713.

Mühlichen, Andreas 2014. Informationelle Selbstbestimmung. In *Handbuch Methoden der empirischen Sozialforschung*, 93-102. Wiesbaden: Springer.

Negroponte, Nicholas 1996. *Being digital*. New York: Vintage.

RatSWD (Rat für Sozial- und Wirtschaftsdaten). 2017. Output 5, Handreichung Datenschutz. *RatSWD Output Series*. Berlin: Rat für Sozial und Wirtschaftsdaten (RatSWD). <https://www.ratswd.de/dl/RatSWD_Output5_HandreichungDatenschutz.pdf> (Accessed February 21, 2020).

RatSWD (Rat für Sozial- und Wirtschaftsdaten) (2017): Output 9, Forschungsethische Grundsätze und Prüfverfahren in den Sozial- und Wirtschaftswissenschaften. *RatSWD Output Series*. Berlin: Rat für Sozial und Wirtschaftsdaten (RatSWD). (Accessed February, 21 2020) <https://www.ratswd.de/dl/RatSWD_Output9_Forschungsethik.pdf>

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*. European Parliament and European Council. May 04, 2016.

Riebling, Jan R. 2018. The medium data problem in social science. *Computational Social Science in the Age of Big Data: Concepts, Methodologies, Tools, and Applications,* 77-103. Köln: Herbert von Halem.

Rotella, Perry 2012. Is data the new oil? *Forbes,* April 2. <http://www.forbes.com/sites/perryrotella/2012/04/02/is-data-the-new-oil/> (Accessed February, 21 2020).

Ruppert, Evelyn S. 2015. Who Owns Big Data? *Discover Society* 23.

Salganik, Matthew 2018. *Bit by bit: Social research in the digital age*. Princeton University Press.

Schar, Katrin 2016. Was hat die Wissenschaft beim Datenschutz künftig zu beachten? Allgemeine und spezifische Änderungen beim Datenschutz im Wissenschaftsbereich durch die neue Europäische Datenschutzgrundverordnung. RatSWD *Working Paper* 257. <https://www.ratswd.de/dl/RatSWD_WP_257.pdf>

Shah, Dhavan V., Cappella, Joseph N., and Neuman, W. Russell 2015. Big data, digital media, and computational social science: Possibilities and perils. *The ANNALS of the American Academy of Political and Social* Science 659(1): 6-13.

Stopczynski, Arkadiusz, Sekara, Vedran, Sapiezynski, Piotr, Cuttone, Andrea, Madsen, Mette My, Larsen, Jakob Eg, Lehman, Sune 2014. Measuring large-scale social networks with high resolution. *PLOS ONE* 9 (4): e95978.

Sugiura, Lisa, Wiles, Rosemary, and Pope, Catherine 2017. Ethical challenges in online research: Public/private perceptions. *Research Ethics* 13(3-4): 184-199.

Stillwell, David J. and Kosinski, Michael 2012. myPersonality project: Example of successful utilization of online social networks for large-scale social research. Proceedings of MobiSys (2012).

Sweeney, Latanya 2002. K-anonymity: a model for protecting privacy. *International Journey on Uncertainty; Fuzziness Knowledge Based Systems* 10 (5): 557-70.

Townsend, Leanne and Wallace, Claire 2016. Social media research: A guide to ethics. University of Aberdeen. <www.gla.ac.uk/media/media_487729_en.pdf>.

Twitter. 2015. *Developer agreement.* San Francisco, CA: Twitter. <https://dev.twitter.com/overview/terms/agreement-and-policy> (Accessed: 21 February 2020).
Twitter 2016. *Broadcast guidelines*. San Francisco, CA: Twitter. <https://about.twitter.com/en-gb/company/broadcast> (Accessed February 21, 2020).
von Unger, Hella 2014. Forschungsethik in der qualitativen Forschung: Grundsätze, Debatten und offene Fragen. In *Forschungsethik in der qualitativen Forschung*, 15-39. Wiesbaden: Springer VS.
von Unger, Hella and Simon, Dagmar 2016. Ethikkommissionen in den Sozialwissenschaften. Historische Entwicklungen und internationale Kontroversen. *RatSWD Working Paper* 253/2016. Berlin: Rat für Sozial- und Wirtschaftsdaten (RatSWD). <http://www.ratswd.de/dl/RatSWD_WP_253.pdf> (Accessed February, 21 2020).
von Unger, Hella, Dilger, Hansjörg, and Schönhuth, Michael 2016. Ethikbegutachtung in der sozial- und kulturwissenschaftlichen Forschung? Ein Debattenbeitrag aus soziologischer und ethnologischer Sicht. *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research* 17 (3).
Wagner, Gert G. 2017. Anmerkungen zu den vielfältigen Dimensionen einer Forschungsethik in den Sozial-, Verhaltens-und Wirtschaftswissenschaften. *RatSWD Working Paper* 265.
Wehofsits, Anna 2016. *Big Data. Ethische Fragen. Ein Report*. Berlin: Vodafone Institut für Gesellschaft und Kommunikation.
Williams, Matthew L., Burnap, Pete, and Sloan, Luke 2017. Towards an ethical framework for publishing Twitter data in social research: taking into account users' views, online context, and algorithmic estimation. *Sociology* 51(6). 1149-1168. doi: 10.1177/0038038517708140.
Williams, Mattew L., Burnap, Pete 2017. Using Twitter for Criminology Research. *Digital Research Ethics Case Study* 2. British Sociological Association.
Zetter, Kim 2015. Hackers Finally Post Stolen Ashley Madison Data. *Wired,* August 8. <https://www.wired.com/2015/08/happened-hackers-posted-stolen-ashley-madison-data/> (Accessed February, 21 2020).
Zimmer, Michael 2010. "But the data is already public": on the ethics of research in Facebook. *Ethics and Information Technology* 12 (4): 313-325.

Historical Social Research
Historische Sozialforschung

All articles published in this Forum: