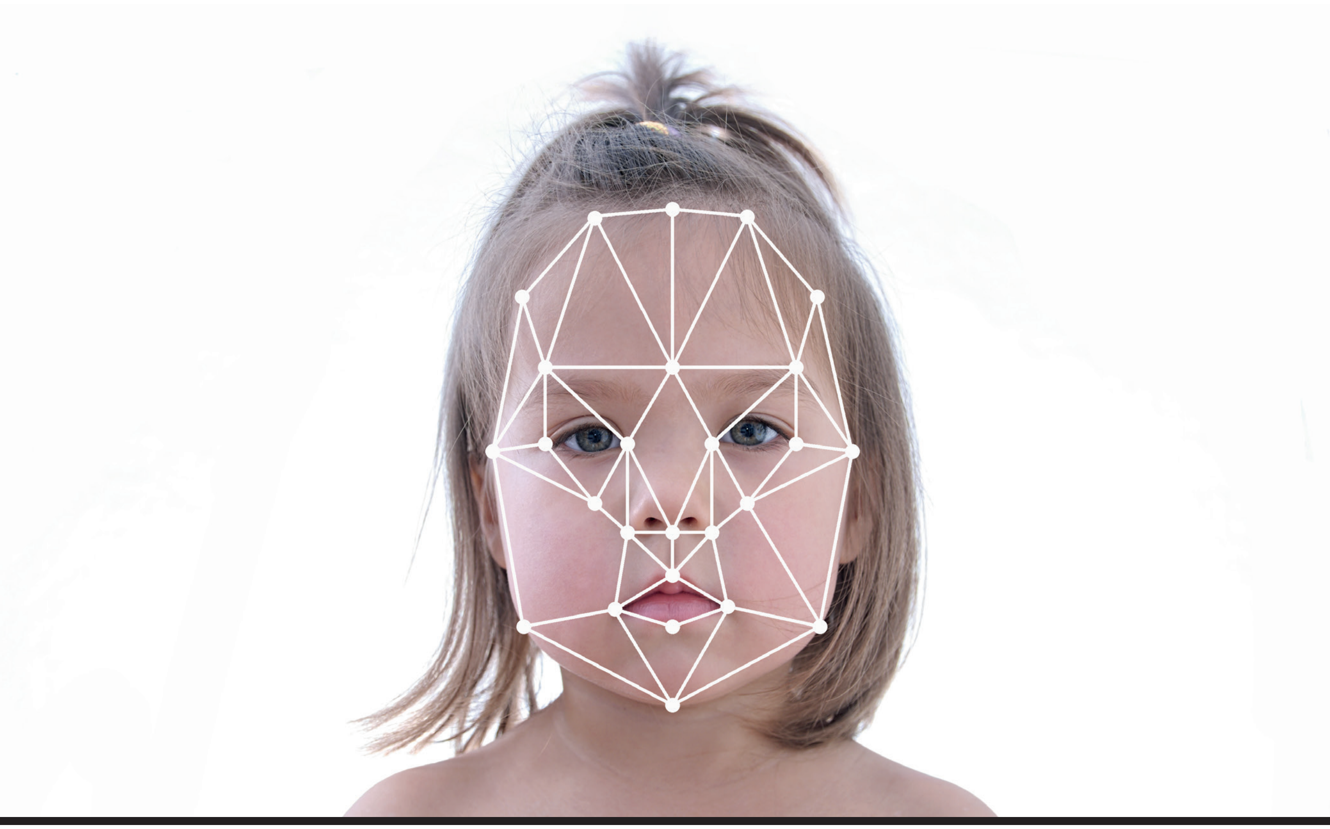
BIG DATA

in research: big opportunities, big challenges



The Norwegian National
RESEARCH ETHICS
COMMITTEES

1st edition, October 2020. English translation, September 2021.
Front page photo: Shutterstock
Design/layout: Ingrid S. Torp

www.researchethics.no

CONTENTS

PREFACE

The Norwegian National Research Ethics Committees (FEK) are independent bodies that address matters of research ethics in all subject areas. FEK seeks to ensure that research both public and private is conducted according to recognised ethical norms. Based on their own experience and cooperation with national and international partners, the committees believe an interdisciplinary review is needed to examine Big Data's implications for research ethics now and in future. In recent years, a number of FEK committees have expressed a variety of views on digitalisation and research ethics. Examples follow:

- The National Committee for Research Ethics in the Social Sciences and the Humanities (NESH) concluded in its "Guide to Internet Research Ethics" (2019) that many of the ethical aspects of using Big Data in research are interdisciplinary and so far-reaching that the committee is unable to report on them within the framework of the guide.

- In 2019, the National Committee for Research Ethics in Science and Technology (NENT) produced a report on artificial intelligence, focusing on the ethical implications of Big Data use. In NENT's view, studying those implications will require an interdisciplinary approach.

- The National Committee for Medical and Health Research Ethics (NEM) and the Regional Committees for Medical and Health Research Ethics (REK) have observed that Big Data poses a challenge to well-established principles such as the requirement of free and informed consent.

To initiate such a review, an interdisciplinary working group was appointed in 2019 to develop a report focused on three main questions: (1) In what ways is Big Data changing research? (2) Which ethical norms in research are coming under pressure? (3) How should researchers and other stakeholders address the emerging opportunities and challenges? The working group has consisted of representatives from each of the committees as well as the National Committee for Research Ethics on Human Remains and the FEK secretariat.

The working group has based its work on a comprehensive review of international and national guidelines and reports as well as relevant research literature. The issues have also been discussed within the FEK committees, commissions and secretariats and with research groups themselves. In addition, a reference group with representatives from a variety of disciplines has contributed along the way. This report summarises the key ethical issues that were identified and recommends appropriate responses in the management and use of Big Data in research.

SUMMARY

Big Data-research challenges the ethical frameworks and assumptions that traditionally apply to researchers and ethics review boards. Based on recent experience, and with input from international collaborators, the National Research Ethics Committees in Norway have identified the need for a review of the challenges posed by Big Data and their implications for research ethics. In 2019, an interdisciplinary committee was charged with exploring the topic and producing a report on Big Data-research and research ethics. The group focused on three main questions: In what ways is Big Data changing research? Which norms are tested? How should researchers and other stakeholders in the research system meet these challenges?

Based on discussions within and experiences from Norway's various research ethics committees, and dialogue with relevant research communities, the report presents the core ethical challenges identified in these conversations. Further, the report provides recommendations for ethically sound and responsible Big Data-research, and investigates how Big Data research tests three clusters of research ethics norms: (1) Norms that constitute good scientific practice, related to the quest for accurate, adequate and relevant knowledge, and norms that govern that the relationship to other researchers. For example, how do we ensure the fitness of data that have been collected for specific purposes and within specific contexts when the data are repurposed? (2) Norms regulating the use of individuals' and groups' personal data. How can researchers ensure the privacy of research subjects in a Big Data climate? Furthermore, how can informed consent requirements be met if the researcher collecting the information does not know what the data will be used for in the long term? (3) Norms regulating the consequences and use of research in society. For example, Big Data, in combination with machine learning, is often used to make decisions or guide decision makers. But even large data sets can be biased or incomplete, which may significantly impact results, fortify existing biases and cause discriminatory practices. How do we ensure that Big Data-based research directly or indirectly benefits society?

1. Introduction

1.1 Background

Big Data is affecting our lives, relationships and societies in new and unexpected ways, and expectations are high that Big Data can be of help in solving major societal challenges. The use of Big Data introduces new and more efficient ways of planning, executing, disseminating and sharing research. With Big Data, research can help to uncover connections that have been impossible to discern until now.

For example, research can help to predict the spread of epidemics and the risk of hereditary diseases. Weather data aligned with data on water conditions and rock and soil erosion can help in preventing natural disasters. Analysis of Big Data can also shed light on the effects of political measures and ensure a more equitable distribution of public goods.

Big Data is a wellspring of new research opportunities. Such opportunities arise from the enormous production of data and new ways of analysing and storing it. The digitalisation of society in general and the integration of sensors and computers into products lead to the continuous generation of large amounts of data. New technology makes it possible to collect, store and analyse such data faster and more cost-effectively than ever before. Artificial intelligence (AI) is a type of technology that expands analytical range and power while generating new data streams of its own.

The expectation that research will unleash the potential of Big Data is closely tied to political goals of open science and open data. The Research Council of Norway's Policy for Open Science, for example, declares that «research processes are to be as open as possible and as closed as necessary»[1], suggesting that the goal of transparency must be weighed against other considerations, such as privacy and data protection.

The mere existence of large amounts of data somewhere does not warrant use of the term Big Data. The data must also be put into service using specific technologies and new methodological approaches. In this report, therefore, Big Data may be understood as data sets which are:

- large
- involve rapidly evolving technologies
- often require new methodological approaches.

---

[1] This principle underpins The Research Council of Norway's «The Research Council Policy for Open Science», 2020.

Anything that creates new opportunities inevitably has an ethical dimension. One purpose of research ethics is to strengthen society's confidence that researchers, research institutions and other research stakeholders will manage and employ data in an ethically sound and responsible manner. The growing reliance on Big Data and its spread to more and more academic disciplines heighten the need to discuss what it means to conduct research in an ethically sound and responsible manner while employing such data.

1.2 Purpose

The purpose of this report is twofold: (1) to point out the key ethical issues that arise when generating and using Big Data in research, and (2) to provide recommendations for generating and using Big Data ethically and responsibly in research.

The report is intended first and foremost for researchers across academic disciplines, but also for research consumers, research funding bodies and public administration. It is intended for anyone who uses or generates Big Data in research. The ethics of Big Data research extends beyond individual researchers and research institutions; the stakeholders providing funding or engaging in other aspects of research also bear responsibility for ensuring that the research is conducted and used in accordance with ethical norms in research. With Big Data-research there is more need for coordination among researchers, the business community and public agencies. In Norway, oversight of research ethics differs by subject area. In medicine and health, for example, advance approval of research projects is required under the Health Research Act. This is not the case for other disciplines. However, all researchers and research institutions in Norway are required to conduct their research in accordance with recognised ethical norms as defined in national research ethics guidelines.[2] I In this report, we maintain that ethical assessments are necessary regardless of the type of organisation responsible for a research project, and that research funders and the stakeholders that issue research guidelines also have an obligation to uphold research ethics.

Many of the issues discussed here are regulated by law. Still, it is important to keep in mind that ethics cannot be reduced to legal affairs, and that ethical questions and challenges also arise within the limits established by legislation. National and international guidelines on research ethics form the framework of this report. Research ethics are defined here as the diverse range of values, norms and

---

[2] Research Ethics Act, Lovdata.no, section 5, https://lovdata.no/dokument/NL/lov/2017-04-28-23.

institutional arrangements that constitute and regulate scientific activity.[3]

Ethical norms in research can be divided into three categories.

(1) Norms related to the pursuit of knowledge and norms that regulate the relationships to other researchers

The way large amounts of data are managed influences and alters research practices. Such changes are consequential for the norms that regulate sound scientific practice, quality assurance and respect for the work of others. The question then becomes: How can researchers control and protect data quality when they are managing vast amounts of data, including data that may have been collected for other purposes?

(2) Norms that regulate the relationship to individuals (including animals) and groups participating in the research

Aligning data from multiple sources may reduce the ability to obtain truly informed consent and could make genuine anonymisation impossible. This raises questions such as: How can researchers ensure that individuals and groups are accorded the respect they are entitled to?

(3) Norms that regulate the relationship to society and the environment

The use of Big Data changes how research affects individuals, society and the environment in the long term. Researchers employing Big Data run the risk of having data used in a way that threatens the democratic rights of individuals. The question is then: How can researchers and those who employ their results prevent or minimise the risk to human dignity and democratic rights?

The report closely examines how the three categories of norms cited above may come under pressure when researchers employ Big Data, and how researchers often have to balance different ethical considerations. Quality may come at the expense of transparency, and transparency may sometimes conflict with matters of privacy and data protection. The benefit of research must be weighed against any risk of harm to the affected parties. Many of the issues raised here are also relevant to research that does not employ Big Data, but may be more acute with Big Data.

The report's recommendations are spelled out in varying levels of detail. Some of

---

[3] The National Committee for Research Ethics in Science and Technology (NENT), «Guidelines for Research Ethics in Science and Technology» (Oslo: The National Research Ethics Committees, 2016), https://www.forskningsetikk.no/retningslinjer/nat-tek/forskningsetiske-retningslinjer-for-naturvitenskap-og-teknologi/.

the subject areas we address have been aired enough previously that it is possible to provide substantive ethical guidance, while other subjects remain at an early stage of discussion and therefore receive more general treatment in this report. In those cases, we hope to raise awareness and move the debate a few steps forward.

In Part 2 we take a closer look at what Big Data is and show how various types of Big Data raise different ethical questions for researchers. Part 3 addresses the ways that Big Data alters research, challenges fundamental research ethical norms and demands new approaches.

2 Big Data in research

2.1 What is Big Data?

There is no established, universal definition of «Big Data». The term is used in a variety of different contexts related to the processing of large data sets. Such data sets exist in a number of forms such as genetic information, electronic health data registries, meteorological data and data from social media. Often there are data from multiple sources to be aligned. The term «Big Data» is often used to refer to data that pose three typical analytical challenges:[4]

- *Volume*: Large amounts of data place new demands on data processing, storage or sharing;
- *Velocity*: The source data can change rapidly when produced in real time or with high frequency;
- *Variety*: The data may be compiled from different sources and can be classified as structured, semi-structured or unstructured.[5]

This definition, known as «the 3 Vs», is often used in the literature, but as a practical matter it has its limitations. One question, for example, is the size threshold for data volume to be considered an analytical challenge. While «Big Data» once referred to data sets so large that processing them required supercomputers, today's very ordinary computers can carry out those same operations. Whether data are considered high-velocity or high-variation will also depend on the available technology.

---

[4] Doug Laney introduced the 3 Vs in 2001 in a MetaGroup report on e-commerce. Doug Laney, «3D data management: Controlling data volume, velocity, and variety», 2001.

[5] Structured data are the easiest to organise and use since their formatting is predefined. Numbers in an Excel sheet are examples of structured data. Unstructured data, such as images or text documents, cannot be organised into such systems, making it more difficult to organise and proces.

«Big Data» is used somewhat more loosely to describe data of significant volume that typically are also heterogenous and produced at a rapid rate. In this report, however, the term is meant to encompass more than isolated and abstract data sets. The actions performed on and with data sets are themselves a form of Big Data and give rise to at least as many ethical challenges.[6] This means that the methodologies and technologies that create a basis for such actions must be included in our understanding of what Big Data is. As a result, our assessment of what constitutes Big Data may well change as technology advances. The phenomenon of Big Data cannot be understood independent of the technology that enables collection, storage and analysis of large data sets.

2.2 Different types of Big Data used in research

The research community distinguishes between different types of Big Data that raise ethical issues in research to varying degrees and in different ways. One distinction is between primary vs. secondary use of data. Primary data have been collected for a specific research project or research purpose – for example, the Trøndelag Health Study (HUNT), a research project encompassing health data and biological samples from over 140 000 inhabitants of Nord-Trøndelag county. Secondary data refers to data collected for other purposes, and often by people other than those who intend to study them/it. This may include data from social media, search engines or commercial websites, sensors embedded in products, or public registries. It may also include data collected under a different research project than the one using them/it. Many of the research ethics issues discussed in connection with Big Data revolve around secondary use of data. Quality assurance is one such issue: how does a researcher ensure that the data – which were collected for other purposes and in a different context than the current project will use them – are of adequate quality and suitable for the current project?

Another distinction is made between *personal data*, which contains identifying information about individuals, and *anonymous data*. The difference here can be particularly difficult to determine in research that uses Big Data. Aligning large data sets containing anonymised data carries the risk of re-identification of individuals. Researchers must therefore address fundamental issues of informed consent and risk of harming those who participate or volunteer their data. Can requirements for informed consent be met at all if there is uncertainty as to how the data might be used later in completely different contexts? What potential consequences could Big Data research have on people's control over their own information? Anonymised data can also lead to other ethical pitfalls, such as the creation of unfair biases. For example, algorithms that are trained or developed on an unrepresentative selection

---

[6] Jonathan Stuart Ward and Adam Barker, "Undefined By Data: A survey of Big Data definitions", 2013, http://arxiv.org/abs/1309.5821.

of images or other data can become a source of discrimination. A facial recognition system not trained to recognise black people would be a case in point, possibly resulting in a disproportionate level of scrutiny for them at passport control posts. A third distinction that applies to secondary use of data is between *data collected by public stakeholders*, such as the Norwegian Labour and Welfare Administration and Statistics Norway, and *data collected by commercial stakeholders*, such as Google and Facebook. Here, a lack of clarity can arise around data ownership, access and sharing. This weakens the principle of fairness, a cornerstone of research ethics. It is therefore vital to address issues such as how to facilitate fair access to data and avoid discriminating against certain researchers, and how to ensure that research results benefit all of society and not only certain privileged groups.

A fourth distinction is between *data collected by Norwegian researchers or in Norway* and *data collected by foreign researchers outside Norway*. Norwegian researchers are now expected to share their data more than previously. But it can be hard to know what laws and guidelines have been followed for data collected abroad and whether ethical standards will be upheld when data, including personal data, are made available to researchers in other countries or to commercial stakeholders. The research community must figure out how to responsibly share data across national borders and regulatory systems.

3 Ethical challenges in research

3.1 The research community

The use of Big Data affects how knowledge is produced, and in the longer term may even alter our understanding of what is considered research. Big Data-based research is data-driven research whereby knowledge is primarily developed by accumulating large volumes of data and using methods to detect meaningful patterns within. Society's expectations as to the potential of Big Data's contributions may lead researchers to turn their focus to issues and disciplines that already have access to large volumes of data. If this limits or interferes with the exploration of other issues, it could threaten academic freedom. The value accorded to Big Data could also lead to disparities between nations, institutions, disciplines, and research groups stemming from their unequal access to large volumes of data and the necessary infrastructure. Section 3.3 explores this in greater detail.

The relationship between causal and predictive knowledge is a key point of discussion in the literature regarding how Big Data is changing knowledge production. The strength of Big Data lies not in its capacity to understand or explain causal relationships, but rather in predicting behaviour by detecting patterns within large data sets.[7] Emphasising the predictive power of Big Data is quite different from hailing Big Data as the end to all theory .[8] To be meaningful, data requires a certain amount of interpretation and a framework. Such frameworks, however, are not necessarily predefined in Big Data research, and are more likely to occur via inductive analysis of volumes of data. It is the tool itself that picks out what is of interest. This makes it essential to discuss the responsibility of researchers to ensure scientific quality and openness, as well as researchers' actual possibilities of carrying out this duty. In this report we do not answer all the questions raised, but we assert that research environments should be aware of the issues involved and highlight them in future discussions about Big Data-based research.

The first set of challenges that we discuss in Section 3.1.2 concerns how researchers should go about ensuring that the data they use are of sufficient quality when the volumes are large and complex, and how to facilitate the further use and dissemination of research results based on Big Data.

The second set of challenges discussed in section 3.1.2 involves how researchers should credit research contributions that are particularly relevant when Big Data is used, such as the production of data sets. This is an important point since the potential of Big Data can only be fully realised through cooperation across disciplines and sectors of society, a process that demands collaboration between disciplines, institutions and nations with differing standards, such as rules for authorship.

3.1.1 Quality assurance

Quality is pivotal in the scientific search for truth. Large data sets offer a basis for more reliable analysis, but large data volumes do not necessarily mean better data volumes; large data sets can contain unreliable data, and even in large data sets, small biases or errors can affect results. This holds true for all research, but since Big Data is used for many different purposes, the consequences of low-quality data containing errors may be more far-reaching than for other types of research.[9]

A number of factors make it especially challenging to determine the quality of Big Data – factors relating to attributes that are characteristic of Big Data.

---

[7] Domenico Napoletani, Marco Panza, and Daniele Struppa, «Is Big Data Enough? A Reflection on the Changing Role of Mathematics in Applications», *Notices of the American Mathematical Society* 61, no. 485–490 (2014).

[8] Chris Anderson, «The end of theory: The data deluge makes the scientific method obsolete», *Wired*, juni 2008, https://www.wired.com/2008/06/pb-theory/.

[9] Dhana Rao, Venkat N. Gudivada, og Vijay V. Raghavan, «Data quality issues in big data», *Proceedings - 2015 IEEE International Conference on Big Data, IEEE Big Data 2015*, 2015, 2654, https://doi.org/10.1109/BigData.2015.7364065.

Firstly, data sets are often unstructured – an estimated 80–90 % of all data worldwide are thought to be unstructured. Large variation makes data difficult to analyse, and it can be costly and demanding to structure the data for further processing. Great progress has been made in recent years in analysing unstructured data using methods borrowed from text analysis and image and speech recognition, but it is difficult to develop standards for determining the quality of this type of data. Secondly, the very speed with which data is generated poses a challenge. Data that are generated continually at rapid rates can be of great value by providing detailed insight and precise predictions. Such data must be curated, however – which is to say organised and put into context. That demands the development of new expertise.

If there is insufficient information about the circumstances and conditions under which data were collected, the secondary-use challenges addressed above are further compounded. Unbeknownst to the researcher, collection methods and algorithms may also have changed during the collection period or processing period, making it difficult to compare findings if a data set is employed at different points in time. «Data noise»– the existence of irregular or incorrect data in a data set – can be another problem with secondary use. Noise can be due to errors during data collection, as in the case of defective sensors, or errors arising during data processing, such as when digitalising a physical document. Data noise is common in very large data sets, but even a low noise level can have a large impact when processing the data set. This is a risk with machine learning, i.e., artificial intelligence systems, if the algorithm learns from noise and makes predictions based on it.

In recent years, a number of technical advances have been made that may help to solve these problems. Data cleansing, for example, is a technology that corrects or removes data that does not belong in a data set, either because it is erroneous, not relevant for the purpose at hand, or it contains information that can lead to biases. When cleansing data, one must decide which properties and variables should count and which should be ignored. But no technology can filter out all noise.

A special challenge in using Big Data is ensuring that the data are representative. Data volume alone is not sufficient to ensure a representative sample. Data collected from the internet and social media may be from a limited user sample, for example, and thus be unsuited for studying patterns within an entire population. A survey conducted in China a few years ago showed that 70 % of social media users were under 30 years of age. This skewed user population makes social media data a poor basis for generalising to the population as a whole, although such data may be suitable for several other purposes.[10] When researchers make use of such data sets, it is important that they understand and explain not only the limitations of the data set, but also what kinds of questions the data are suitable for, and which interpretations may be possible.[11]

In general, making the leap from causal relationships to predictions in Big Data research entails a higher risk of making unfounded interpretations of findings. Correlations between variables, for example, may be misperceived as causal relationships. When researchers make use of large data sets to interpret social structures and predict human behaviour and events, sources of error in data sets and algorithms can have serious consequences. Researchers have a duty to provide information about the uncertainty and limitations involved in any further use of the data, and to clarify this when disseminating research results.

Recommendations

- Researchers should assume independent responsibility for ensuring that data is of adequate quality for the research purpose at hand.

- Researchers should strive for openness in their own approaches and facilitate the further use of data sets, so that researchers who use the data later can familiarise themselves with how the data was collected and any limitations that may apply to further use.

3.1.2 Respect for others' research contributions

A key principle of research ethics is that researchers are to respect one another's contributions by, among other things, adhering to recognised standards for co-authorship and collaboration. This helps to ensure that research is responsible and verifiable. This section of the report addresses proper attribution of credit in interdisciplinary research projects and research contributions in the form of producing data sets.

Big Data research is often interdisciplinary, requiring researchers from different fields to take each other's discipline-specific standards into account. Research questions and data may be rooted in biology, for example, while IT disciplines

---

[10] Jianzheng Liu et al., «Rethinking big data: A review on the data quality and usage issues», IS-PRS Journal of Photogrammetry and Remote Sensing 115 (2016): 138, https://doi.org/10.1016/j.isprsj-prs.2015.11.006.

[11] Kate Crawford, «Critical questions for big data. Provocations for a cultural, technological, and scholarly phenomenon», *Information, Communication & Society* 15, nr. 5 (2012): 669–70, https://doi.org/10.1080/1369118X.2012.678878.

contribute to managing and refining the data to be used. One potential issue in such interdisciplinary projects is how to order the list of authors. In the Vancouver recommendations, authorship credit assumes substantial contributions to the following: (1) the conception or design of the work, or the acquisition, analysis or interpretation of data; (2) drafting the work or revising the manuscript critically; (3) final approval of the version to be published[12] and (4, a criterion added in 2013) agreement to be accountable for all aspects of the work.

Publication traditions, however, may vary from one discipline to another. Academic articles within medicine, science and technology typically list more authors than those in the humanities and social sciences. The disciplines have their own guidelines for assigning authorship credit and determining the order of authors. If machine learning is used in analysing Big Data, with contributions from IT experts and mathematicians, but the research question is rooted in, say, philosophy, it may be complicated to satisfy a discipline's traditional notion of who deserves authorship credit. Thus, in Big Data research, it may be necessary to agree upon the list of authors at an early stage of the project.

It can also be difficult to determine the types of contribution that warrant authorship. Examples include contributions by technical personnel or statistical and paraclinical experts in clinical research. Managing and analysing Big Data may comprise a substantial component of the work, but it may be unclear whether that contribution qualifies someone for authorship. Research contributions can also be made in ways other than article-writing. Section 3.1.1 examined some of the difficulties that can arise when researchers compile, align and quality-assure data sets in order to use them for research purposes. If quality assurance is to be recognised, it is important to approve the production or preparation of data sets, such as curating and standardising, as genuine research contributions. One way of achieving this is to facilitate the citation of data sets. This acknowledges the value of the data while giving recognition to the research efforts of those who produced it/them. In recent years, several initiatives and technical solutions have been developed that make this possible. Data sets are assigned DOI numbers or other identifiers, including those used by DataCite. Specialised peer-reviewed journals have been established for publishing data sets, such as Scientific Data, Data in Brief and GigaScience.

Recommendations

- Project managers of interdisciplinary collaborations should encourage any authorship issues to be clarified and agreed upon at an early stage.

---

[12] International Committee of Medical Journal Editors (ICMJE), «Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals», 2019, http://www.icmje.org/recommendations/.

- When using others' data sets, researchers should state each data set's origin. Any metadata such as DOIs should be included in the list of references.

3.2 Consideration for individuals and groups

Trust between researchers and participants is critical in all research involving humans. For example, participants must have confidence that the information given about the project is truthful, and researchers must trust participants to report honestly.[13] Big data changes the relationship between researchers and those to whom the data pertains, and it challenges and expands the use of terms such as «harm» and «research participant» in research ethics.[14] The data are usually not collected through direct contact between researchers and participants. In many cases, the researchers use digital trace data from social media or apps that are then aligned and employed in new ways. As the distance between researcher and participant grows, the risk of harm can seem increasingly hypothetical and less specific, especially if future use of the data is unclear. When the relationship is abstract or limited and communication is not via direct contact, it can be particularly difficult to establish trust.

The next set of norms discussed in this report can be tied to two fundamental, absolute principles of research ethics: (1) ensuring respect for individuals and groups that participate or that may be directly affected by the research, and (2) minimising any adverse effects, meaning risk of harm, discomfort or other repercussions for the subjects being studied. In practice these principles are administered as *informed consent requirements* and *the confidentiality requirements*. Dissemination of personal information about individuals is, in other words, to be confined to certain persons.

3.2.1 Personal data, confidentiality and anonymisation

In a legal sense, privacy and data protection concerns the processing of personal data and is regulated by data protection regulations. Research ethics employs a broader sense of privacy and data protection that extends to protecting individuals' human dignity. In the realm of research ethics, privacy and data protection encompasses not only the protection of personal data but also the principle that each individual has interests and integrity that must not casually be set aside in

---

[13] Helene Ingierd og Hallvard Fossheim, «Innledning», i *Forskeres taushetsplikt og meldeplikt* " [Researchers' duty of confidentiality and duty of notification] (Oslo: The Norwegian National Research Ethics Committees, 2013), 7–8.

[14] Jacob Metcalf og Kate Crawford, «Where are human subjects in big data research? The emerging ethics divide», *Big Data & Society* 3, nr. 1 (2017): 2, https://doi.org/10.1177/2053951716650211.

the interest of research.[15] Researchers are to treat the processing of any and all information about personal matters confidentially.[16] The research material usually must be anonymised to protect privacy and avoid harm and undue strain on the persons being studied. In Big Data-based research this is a particularly challenging requirement to meet. With further use of Big Data, human dignity issues arise. This is discussed in section 3.3.

A distinction is to be made between *re-identification of de-identified data*, and *re-identification of anonymised data*. Data that are linked to an individual can be de-identified to make it more difficult to pinpoint the individual – e.g., by replacing a name with a code – while retaining the ability to re-link it to the individual. Data can be considered anonymised when individuals are no longer identifiable. To prevent re-identification of anonymised data, any identifiers available in the data must be eliminated and the data set must be processed in a way that prevents someone re-identifying individuals through variables which individually are not sufficient to identify someone but in combination may well do so. While de-identified data are considered personal data and thus regulated under the Personal Data Act, anonymised data are not. Additionally, data that are currently considered anonymised may not necessarily remain so as technology advances.

As the volumes of data increase, it will be possible to align and realign data and study it in perpetuity. With the advent of new technology, for example, genetic data considered anonymous today could be linked back to the identities of individuals. There will be more and more examples of researchers aligning data sets to one another and uncovering sensitive information through access to anonymised data. A 2019 US study showed that 15 demographic attributes are enough to re-identify individuals in a large data set with 99.98 % accuracy.[17]

This means that in practice it may be difficult or nearly impossible for researchers to guarantee that data are truly anonymised. Researchers should be aware that collecting and processing anonymised data carries an ethical responsibility towards the individual participants. When a researcher is obtaining consent, he or she should therefore inform participants of the risk of identification.

It is worth noting that the perception of what constitutes personal data may vary between individuals and cultures. Data which in principle is not encompassed

---

[15] The National Committee for Research Ethics in the Social Sciences and Humanities (NESH), *A Guide to Internet Research Ethics* (Oslo: The Norwegian National Research Ethics Committees, 2018).

[16] The National Committee for Research Ethics in the Social Sciences and Humanities (NESH), *Guidelines for Research Ethics in the Social Sciences, Humanities, Law and Theology*, (Oslo: The Norwegian National Research Ethics Committees, 2016).

[17] Luc Rocher, Julien M. Hendrickx, og Yves-Alexandre de Montjoye, «Estimating the success of re-identifications in incomplete datasets using generative models», *Nature Communications* 10, nr. 1 (2019), https://doi.org/10.1038/s41467-019-10933-3.

by the legal definition of personal data may indeed be perceived as personal data by those to whom the data pertain. Researchers thus have a duty to consider the context within which the information is collected. Moreover, researchers may gain access to personal data not originally requested, but which nevertheless must be processed. This may occur when conducting research indirectly on or with people, as in the case of environmental monitoring and the use of drone technology or other measuring devices. In general, researchers should be aware of the risk their research poses to individuals, and not only the type of data being collected.

Over time, a number of methods have been developed to help safeguard personal and sensitive information. In recent years, machine learning has been used to generate synthetic data sets that reflect the statistical macro-characteristics of sensitive data sets. All the individuals in such data sets are fictitious and none needs to be an exact copy of a real person, but in the aggregate this artificial, fictitious population has the same statistical traits as the population being modelled. The use of such synthetic data removes all risk of re-identification (since there are no actual persons to be re-identified), but of course it may still be possible to reveal the traits of real people by studying the fictitious population.

Another alternative is to allow only machines to process the data. Machine processing of personal data is processing nonetheless, and therefore constitutes an encroachment on personal privacy. Machine processing locks out human access to data, but statistical methods can still detect and adjust any biases in the data sets.[18] In this way, researchers can make use of data without having access to it. Data refined in this way is called microdata.

In Norway, the Norwegian Centre for Research Data (NSD) and Statistics Norway have collaborated to develop microdata.no. The idea behind microdata.no is that researchers can use non-anonymised data through a platform that protects confidentiality, where the researcher only has indirect contact with personal data, where data remains in Statistics Norway's possession, and where the platform ensures that all output is anonymous.[19] Such services can help researchers safeguard personal privacy while allowing them access to registry data without lengthy delays and high cost. The microdata solution has certain limitations since researchers must use the embedded tool for statistical analysis.

Recommendations

- Researchers should be aware of the complexity involved in ensuring anonymity

---

[18] Teknologirådet, «Kunstig intelligens og norske helsedata» [Artificial intelligence and Norwegian health data], 2019, 2.

[19] «Om microdata.no», microdata.no, undated, https://microdata.no/about/.

and confidentiality in Big Data research and should acknowledge both the legal and ethical dimensions of privacy and data protection. For any research project that uses personal data, or where there is a genuine risk of re-identification, researchers should draw up a plan for safeguarding privacy.

- Research institutions should ensure that infrastructure includes adequate data security measures, and that researchers who will be employing data sets receive training in maintaining data security.

3.2.2 Informed consent

Informed consent is a basic requirement in research on humans when the research involves data registration or any kind of discomfort or risk to research subjects. Obtaining informed consent for research participation is a key rule in research ethics, regardless of the legal basis for processing personal data. Under privacy and data protection regulations, consent is only one of multiple legal bases.[20]

Consent is an essential part of ensuring respect for the freedom and self-determination of individuals participating in research; such consent must be free, informed and explicit.[21] In sum this means that information must be provided about the research project's purpose, funding, expected risks and possible benefits for research participants. It also means that participation is to occur without external pressure or restrictions on personal freedom of action; that individuals may withdraw at any time; and that their consent is to be expressed explicitly by an affirmative action.

In large-database research, meeting the information requirement can be difficult. The information someone agrees to share may be harmless but processing it and aligning it with other data can lead to new information. Big Data thus allows us to develop new and more detailed information about individuals – information that can be misused. As a result, it is sometimes difficult to inform prospective research participants about all the expected risks and possible benefits. Part of what makes Big Data so valuable is its usefulness in repurposing previously collected data. When Big Data-based research reuses existing research data, the purpose limitation principle cited in the Personal Data Act comes into play. This principle states that personal data is not to be processed without the existence of a legitimate, specific

---

[20] The National Committee for Research Ethics in the Social Sciences and Humanities (NESH), *Guidelines for Research Ethics in the Social Sciences, Humanities, Law and Theology*, sections 5–8

[21] «Lov om behandling av personopplysninger (personopplysningsloven)» [Act of 14 April 2000 no. 31 relating to the processing of personal data (Personal Data Act)], Lovdata.no, 2018, art. 6a, https://lovdata.no/lov/2018-06-15-38/gdpr/a6.

and explicitly stated purpose.[22] Adhering to this principle requires identifying and describing the purpose of the processing that will be carried out and ensuring that all persons concerned know how their personal data will be used.[23]

However, it is not necessary under the informed consent requirement to give a prospective provider of information full details about all the conceivable benefits and risks of handing over their data. Lengthy written explanations may in fact go unread, leaving research participants unaware of the content. What is crucial is that research participants gain enough knowledge about the research project to realistically assess its purpose and risks. To achieve this, the researcher must have some control over who will have access to the information contributed by the participant, and this knowledge must be communicated to prospective participants. Research institutions should facilitate the use of technical solutions that let research participants see who has access to their data. Researchers heedful of research ethics should inform participants that their data may be reused for other purposes, if that is the case. If the research participant cannot be informed of the purpose in advance, or the participant is unable to evaluate the possible benefits or risks of providing data, then there is little basis for describing his or her consent as informed under research ethics guidelines.

Big Data-based research can make it difficult to fulfil the requirements that consent is to be expressed explicitly by means of an affirmative action, that no external pressure or freedom restrictions are to be applied, and that participants are to be permitted to withdraw at any time. Researchers using Big Data can have trouble retracting information once it has been shared with others, so participants must be told clearly when they can withdraw their consent and when withdrawal will no longer be possible.

The problem of defining what constitutes informed consent is heightened with *secondary use*. This is when researchers obtain data from sources such as social media, mobile phone apps or public registries that contain personal and sensitive information or obtain data that can lead to personal identification when aligned with other information. In some cases, consent is not relevant, as when data originates with public agencies such as the tax authority or when user data has been uploaded voluntarily to social media. Sometimes individuals agree in advance

---

[22] Personal Data Act, Lovdata.no, 2018, https://lovdata.no/dokument/NL/lov/2018-06-15-38.

[23] Data protection legislation recognises that it is not always possible in a research context to fully identify the purpose for processing personal data at the time data is collected. Data subjects should therefore be allowed to give their consent «to certain areas of scientific research when in keeping with recognised ethical standards for scientific research». In addition, data subjects should have the opportunity to consent only to certain areas of research or parts of research projects to the extent allowed by the intended purpose. Therefore, stating the purpose as «developing artificial intelligence» is unlikely to fulfil the duty of purpose limitation; but that duty may be fulfilled if one specifies the type of artificial intelligence that is to be developed and the expected operations these systems will perform.

that the data they provide – to a mobile phone app, for example – may be used for research later. In other cases, it can be hard to know for sure where data is coming from and whether valid consent has been given. In general, informed consent should be obtained anew when the researcher intends to reuse personal and sensitive information for a new purpose, or when the information can lead to identification if aligned with other material. This will often be difficult or impossible in practice – as with data pertaining to dead persons. If the data in question is personal or sensitive, or if there is a risk of identification, it is ethically problematic to conduct research using material from a deceased person when informed consent has not been given.

The norm of informed consent and norms relating to transparency and utility may conflict. In registry research and biobank research, consent models have been developed to maintain respect for the individual while making data more widely available. *Broad consent* is one widely used model. Such declarations are designed to allow future use of data in a set of broadly defined areas of research without the need to reacquire consent. Although data thus obtained may be used in addressing new research questions, such general forms of consent do not authorise use of the data for any and all research projects. Another model is *tiered consent*, in which research participants tick off certain areas their data can be used for. For example, a participant may consent to his or her data being used in cancer research but disallow its use in genomic research.[24] This approach also allows research participants to decide for themselves whether to be informed if the research in question turns up information of relevance to their personal health, such as the likelihood of developing certain diseases. The *dynamic consent model* goes a step further. It employs communications technology to inform and engage research participants over time and thereby allow them to decide if and when their personal data can be used again, and for what purposes. The Norwegian Mother, Father and Child Cohort Survey (MoBa), one of the world's largest health surveys, employs this model. Another option is to give every resident of the country a right of reservation. This allows citizens to exclude all or parts of their data. In the UK, a centralised *national data opt-out service* has been established, giving residents the ability to exclude their confidential patient information from use by specific organisations. If you do not opt out, your data may be used.[25]

A risk associated with all these approaches is that more is demanded of the research participants than with a traditional consent model. Participants must keep up to date as the research progresses and make active choices over a long period. To

---

[24] Brent Daniel Mittelstadt og Luciano Floridi, «The ethics of big data: Current and foreseeable issues in biomedical contexts», *Science and Engineering Ethics* 22, nr. 2 (2016): 312, https://doi.org/10.1007/s11948-015-9652-2.

[25] Norwegian Board of Technology, «Kunstig intelligens og norsk helsedata» [Artificial intelligence and Norwegian health data], 2020, 2.

address this challenge, research leaders must/should ensure an accessible flow of communication concerning their research and how the data is being managed. Providing substantive information to those who may be affected by the research is essential to building trust and goodwill, which in turn encourages participants to share their data, including in the long run. In time, the technological advances that have brought about the challenges discussed here may also provide a practical means of resolving them.

In certain types of Big Data-based research, there may be legitimate reasons to make an exception to the requirement of free, informed consent. In assessing whether such an exception can be made, it is important to distinguish between public and private information.[26] In simple terms, the processing of private data will often require informed consent, while information obtained through observation of «open arenas, on streets and public squares» will not.[27] Nevertheless, it is not the case that all information that is openly available, and thus apparently public, can be made the subject of research without informing the parties concerned or asking their consent. In determining what is private and what is public in terms of research ethics, it is helpful to employ the concepts of *reasonable expectation of publicity* and *contextual integrity*. When there are expectations of publicity, it is incumbent on the researcher to assess whether the informants understand and expect that their actions and statements will in fact be public, and whether they understand and expect that this information may be used for purposes other than what they originally thought (NESH 2019). Maintaining contextual integrity requires the researcher to consider the context in which information is found or where communication occurs.[28] For example, people in closed Facebook groups may differ in their understanding of how public the information they share is. It is important to clarify not only whether the data are private or public, but how much protection of individuals is sufficient in specific contexts.

Even where there are legitimate exceptions to the informed consent requirement, researchers are responsible for safeguarding the integrity of those who have provided information for research purposes. This responsibility requires arguing for the utility value of one's research and informing those directly concerned and the public in general of the project's purpose and findings. It also requires processing of data in a responsible and transparent way and exerting control over who can access it and how it is used.

---

[26] The National Committee for Research Ethics in the Social Sciences and Humanities (NESH), *A Guide to Internet Research Ethics*.

[27] The National Committee for Research Ethics in the Social Sciences and Humanities (NESH), *Guidelines for Research Ethics in the Social Sciences, Humanities, Law and Theology*, section 7.

[28] Helen Nissenbaum, *Privacy in context: Technology, policy, and the integrity of social life* (Stanford: Stanford University Press, 2011).

Recommendations

- Researchers should assess possible risks and benefits for participants and communicate them in a clear way. When managing personal data, researchers should carefully consider whether the data can be reused for new and different research purposes than those they were provided for. When a clear description of research purpose cannot be given, researchers should consider whether obtaining valid informed consent is even possible. If it is possible, they should inform participants as to when consent can be withdrawn and at what point withdrawal will no longer be possible.

- With secondary use of data, researchers have a separate responsibility to safeguard the integrity of research participants. As a general rule this means informed consent is to be obtained again when the researcher plans to reuse personal data for a new purpose or when there is a risk of identification in connection with data alignment. When a researcher or ethics committee deems it ethically sound to carry out research without informed consent, the reason for not obtaining it should be clearly communicated to those directly concerned and to the general public. The integrity of individual participants should be protected and reducing the risk of identification is one way of doing so.

- Research institutions should facilitate the use of technical solutions that let research participants see who has access to their data.

3.2.3 Safeguarding group interests

Legislation and research ethics guidelines are more oriented towards the individual than towards groups.[29] For example, data protection guidelines are mainly about safeguarding the interests, dignity and autonomy of individuals as well as the right to control what kinds of data about them are collected and how they are used. The legal framework, too, aims primarily at protecting the personal data of individuals.

With Big Data, however, analysis often takes place at group level. Artificial intelligence and algorithms can be used to create groups, and the information collected on such groups can be used to prevent crime, predict the spread of diseases or facilitate urban planning. Identifying individuals is not the goal of such analysis. Some experts have espoused introducing the concept of «group protection», but views differ on how to define group interests in the context of Big

---

[29] Mittelstadt og Floridi, «The ethics of big data: Current and foreseeable issues in biomedical contexts», 25.

Data.[30]

As the volume of data increases, more data points emerge that can be used to identify common denominators between people. Patterns established within such datasets make it possible to identify and group individuals by behaviour, preferences and other characteristics, such as place of residence, gender or age. Groups are delineated by the parameters selected, whether by a data analyst or an algorithm that searches for patterns independent of predetermined criteria. Such groups exist as clusters within data sets, rather than as collections of individuals who share an identity or experience. Knowledge of the group's existence may be limited to the data analyst, and in fully automated processes – to the system itself. A group's membership may also change in the course of analysis, since the group itself can be continually redefined, existing only as long as it takes to carry out a specific analysis. While a group is traditionally understood as consisting of people who share an identity or an experience (such as ethnic minorities or members of a club), in this case we are dealing with algorithmically determined groups whose members do not necessarily identify themselves as group members.

Existing data protection practices focus largely on whether or not individuals are identifiable. Even if individuals are anonymised, the use of their data can still affect groups that are identified or profiled using Big Data technology. Examples would include those whose consumption habits identify them as ripe for targeted marketing and those with a profile suggesting susceptibility to certain diseases.

Group affiliation can also be a basis for discrimination, as in the case of people subjected to targeted influence based on selected behavioural traits. Legally protected groups, such as minority groups, may fall into this category.

The fact that individual group members may experience discrimination while not understand why it is happening is itself problematic.

Another risk is re-identification. As the number of variables used to define a group increases, the group itself becomes more narrowly drawn.[31] It also becomes easier to identify which particular individuals are encompassed. For example, a group may be large enough that the individuals in it are relatively anonymous if they are residents of a city. But that changes if the group is narrowed further by gender, age, residence in a particular street and residence in a particular street while also owning a dog.

---

[30] Linnet Taylor, Luciano Floridi, og Bart van der Sloot, «Introduction: A new perspective on privacy», i *Group Privacy: New Challenges of Data Technologies* (Springer, 2017), 6.

[31] Lanah Kammourieh et al., «Group privacy in the age of big data», in Group privacy: *New challenges of data technologies*, red. Luciano Floridi, Linnet Taylor, og Bart van der Sloot (Springer International Publishing, 2017), 37–66.

To some extent, existing data protection practices are able to strengthen protection for members of certain groups. The Personal Data Act, for example, increases the protection of special categories of personal data, such as information on racial or ethnic origin, political views or genetic and biometric data.[32] Nevertheless, new grey zones have appeared surrounding groups and individuals whose interests are inadequately protected, a development which heightens the need for an approach to group protection and a discussion of emerging challenges.

Recommendations

- Researchers should reflect on whether data and algorithms can create a risk of re-identification or discriminatory outcomes for groups. Researchers should prepare a plan to address group interests in research projects where such a risk exists.

- Research institutions should take responsibility for addressing group protection by regulating the processing of data and use of algorithms.

3.3 Social responsibility in research

This portion of the report deals with research norms related to the ethics of social responsibility and how Big Data research can put such norms under pressure.

By analysing Big Data, public authorities and business interests are able to gather information on people to predict patterns of action. Such information can benefit individuals and society but can also be misused. A recent example was the introduction of a smart-phone app called Smittestopp («Infection stop») in Norway in the spring of 2020. The idea was to collect location data using GPS and Bluetooth. One purpose of the app was to make corona virus infection tracing more efficient and thereby slow the virus's spread. However, the app raised concerns that the collection and alignment of such data could be misused. As a result of the objections raised, the Norwegian Institute of Public Health no longer recommends use of the original app, and a new app for infection tracing is under development.

A basic principle in research ethics is that researchers have a responsibility to ensure that their research benefits society directly or indirectly and to minimise any adverse effects from the use of research results.[33] However, the responsibility

---

[32] Regulation on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation), Article 9.

[33] The National Committee for Research Ethics in Science and Technology (NENT), *Guidelines for Research Ethics in Science and Technology.*

for how research results are put to use is shared by multiple parties. Powerful public and private stakeholders control large amounts of data and infrastructure, and the researcher is one of many nodes in a system whose data may be generated outside of the research project and whose research results may be employed by other actors. Consideration must be given to how other parties in Big Data-based research can be made to assume responsibility for methods of data generation, management and use, and to what responsibility researchers and research institutions have in relation to those parties.

3.3.1 Human dignity

In research, ethical norms concerning individuals and groups rest on fundamental principles of self-determination, human dignity and democratic rights. Such principles are put to the test when Big Data-based research is used as a tool for manipulation, such as influencing individuals in a certain way without their knowledge.

Big Data, in combination with machine learning, can be used to analyse, predict and influence future preferences and behaviour patterns. This technique is called profiling. Profiling can help Big Data to provide services of higher quality and greater benefit. For example, Amazon has for many years recommended books on the basis of previous purchases and preferences. Yet as various services assemble more and more information about each of us, they go a long way towards identifying and defining who we are. Today, algorithms are better than our friends and family – and perhaps even ourselves – at predicting our preferences. Such data can be used for *nudging*, in which methods are used, sometimes imperceptibly, to modify our behaviour in ways beneficial to ourselves or society – perhaps prompting us to eat healthier food or sort our recyclables. Such influencing techniques can also be used in ways that are more problematic, nudging us to buy specific products, say, or creating reply forms and bureaucratic processes so cumbersome as to prevent users from accessing benefits they are entitled to.[34] Whether exerting positive or negative influences, nudging may eventually threaten our self-determination and the ability to shape our own identities.[35] The more information algorithms receive about us, the more our choices can be swayed.

Respecting the dignity of human beings involves placing limits on the categorisation of individuals using Big Data and algorithms. If people are treated as mere aggregations of data when they apply for social security benefits or a home mortgage, it is questionable whether due regard is being paid to their integrity as

---

[34] Richard H. Thaler, «Nudge, not sludge», *Science* 361, nr. 6401 (2018): 431, https://science.sciencemag.org/content/361/6401/431.

[35] Robindra Prabhu, «Big data - big trouble?», in *Internet research ethics*, eds. Hallvard Fossheim and Helene Ingierd (Oslo: Cappelen Damm, 2015), 162.

human beings.[36]

Big Data can also put strain on democratic rights. Efforts by both state and non-state actors to influence democratic processes have received considerable international attention in the aftermath of the 2016 US presidential election as well as the UK referendum on EU membership the same year. In both election run-ups, the consulting firm Cambridge Analytica was able to harvest large amounts of data – more than 50 million raw profiles – and employ Big Data analysis to create profiles and send targeted information to individuals. Such targeted influence operations are increasingly undermining confidence in democratic processes and institutions. Activities include spreading false news and disinformation, often with intent to sow confusion and doubt about what is true.

The potential consequences of Big Data-based research and the many ways of misusing personal data exceed the grasp of the individuals providing the data. Research funding bodies, research institutions and researchers seeking to counter the threat to individual self-determination, human dignity and democratic rights should reflect critically on the purpose of their research as well as how it will be used and how it could be misused. To minimise the risk of misuse, funding providers should support broad interdisciplinary research on the new capabilities engendered by Big Data to manipulate organisations, decision-making processes and individuals. Researchers and research institutions should work towards an «embedded ethics» in which broad-based ethical assessments of research projects, including their effect on individuals and society, are conducted right from the start. Researchers in different academic disciplines should develop the ability to identify and critically explore any biases or manipulative potential in the algorithms they employ. Based on what they find, researchers and other research stakeholders should try to minimise the potential risks, perhaps by building in restrictions on use. Researchers should also inform users about the potential for misuse.

Recommendations

- Researchers and research institutions should work to develop «ethics by design», and maintain critical awareness of the purpose of their research, how it is to be used and how it could be misused.

- Research institutions should encourage academic communities to develop the ability to identify and critically explore the manipulation potential of the algorithms they employ.

- Research funding bodies should support broad interdisciplinary research on

---

[36] European Data Protector Supervisor, «Opinion 4/2015. Towards a new digital ethics: Data, dignity and technology», 2015.

the potential use of Big Data analysis in manipulating organisations, decision-making processes, and individuals.

3.3.2 Fairness

Fairness is a fundamental principle in research ethics. Its scope extends to the issue of data access, including the degree to which all research environments enjoy equal access. As shown earlier in this report, Big Data-based research also raises the question of whom the research benefits. Its beneficiaries are in fact hard to predict. Both aspects of fairness are discussed here.

Discrimination in access to data

Unequal access to data can help to create a new digital divide between the “Big Data rich» and the “Big Data poor».[37] The largest global data sets are owned by international companies such as Facebook, Google and Amazon. The wider research community in many cases lacks access to this data, either because the companies do not want to release data for competitive reasons or because acquiring access would cost too much. In recent years, a number of research policy measures have been instituted nationally and internationally to promote open research, with a focus on open, sharable data.[38] An important argument for such measures is that knowledge should be available to all, and of potential benefit to all. However, the relevant policy documents distinguish between publicly funded research data and other data. Publicly funded data refers to data that are collected through publicly funded research or that form the basis of published articles resulting from publicly funded research. Private and public stakeholders thus face different transparency requirements.

This can affect the issues that are explored and the production of knowledge in society. Evaluations or research carried out by private companies have commercial purposes, and topics unlikely to produce short-term value may be deprioritised. The

---

[37] Crawford, «Critical questions for big data. Provocations for a cultural, technological, and scholarly phenomenon».

[38] Since 2017, the EU’s Horizon 2020 research and innovation programme has had open access requirements for research data as well as data management planning requirements. In 2018, the European Open Science Cloud (EOSC) was launched. The cloud will encourage open research and innovation across national borders in Europe by offering storage, management, analysis and reuse of research data. The international FAIR principles (Findability, Accessibility, Interoperability and Reuse) were developed to facilitate extended use of research data. Since 2017, the FAIR principles have been incorporated into the Research Council of Norway’s policy on making research data available. In 2017 the Ministry of Education and Research issued its «National strategy on access to and sharing of research data». The strategy establishes basic principles for the management and curation of publicly funded research data and presents measures to improve accessibility and sharing of research data.

motivation and interests of researchers with close ties to private companies may be swayed in ways that affect research priorities and results. In the long run, this could prevent research from benefiting the whole of society. In addition, other researchers may have trouble verifying the findings and evaluating the work of such colleagues, with consequences for the integrity and credibility of the research.

Frameworks and infrastructure for storing and managing research data should be adapted to allow data value to be exploited in ways that produce an overall benefit to society at large. Neither commercial factors nor data protection considerations should be misused by powerful actors whose intent is to avoid transparency and accountability.[39] Public and private data stakeholders involved in research have a duty to conduct research ethics assessments. It is essential that such parties clarify whether their activities do or do not constitute research. If their activities qualify as research, they must develop the procedures and expertise needed to safeguard privacy and uphold research ethical standards.

Recommendations

- Frameworks and infrastructure for storing and managing research data should be adapted to allow the value of data to be exploited in ways that benefit society at large.

- Public and private data stakeholders involved in research have a duty to conduct research ethics assessments.

Potential discriminatory outcomes

Big Data is often used in combination with machine learning to make decisions or guide decision-makers. However, one often hears a concern that data used in research gives rise to biases that can exaggerate unwelcome differences between social groups and lead, in some cases, to discrimination. When the algorithm learns from experiences in a data set – often referred to as guided learning – it means machines are learning from data that has been collected. Such data can reflect underlying distortions and result in prejudicial decision-making.

Profiling – an array of techniques used to analyse, predict, and possibly influence future preferences and patterns of behaviour – is used by the police in the United States to estimate the likelihood of where and when criminal acts are committed

---

[39] European Data Protector Supervisor, «A preliminary opinion on data protection and scientific research», 2020, 2, https://edps.europa.eu/sites/edp/files/publication/20-01-06_opinion_research_en.pdf.

and who commits them.[40] In China, facial recognition is combined with artificial intelligence to develop better systems of social control and sanctioning of the population, especially minority groups.[41] These are examples of how profiling can reinforce prejudices and stereotypes and pave the way for exclusionary and discriminatory practices like subjecting selected people to differential treatment according to ethnicity, gender or place of residence.

In some contexts, it can be difficult to know where a particular solution will lead and whether it will have a discriminatory effect. Algorithms, however, can also be used to expose discriminatory processes, such as by making it easier to examine an entire decision-making process and determine whether discrimination has taken place.[42] With oversight and openness of research data and algorithms, biases can be rectified, and discriminatory outcomes prevented. Within Big Data research, however, the ethical requirement to maintain transparency is a challenge when artificial intelligence is employed in processing and analysis. Researchers and outsiders alike may find it hard to determine how decisions have been reached. A predicament called the «black box problem» arises when artificial intelligence systems and algorithms become so complex that no one understands how they arrive at the answers they provide. If decisions are increasingly taken by algorithms, a «data dictatorship» could arise.[43] This is part of why data protection regulations require automated decisions to be explainable. It is helpful to distinguish between two types of «black boxes»: an *involuntary black box* where the lack of transparency is due to the nature of the model, and a *voluntary black box* where the lack of transparency is due to security or commercial concerns. In both cases the problem is that it is impossible to verify the machine's line of reasoning, so the responsible parties cannot explain the conclusions. On the other hand, a system containing an involuntary black box often performs better than a more open system. A balance must therefore be struck between quality and openness.

Researchers should be frank about the need for such trade-offs, and one goal of their research should be to produce «glass boxes», or systems that are transparent.

Researchers should be aware of and ready to assess how bias from the data or algorithms they employ may affect end users. Based on their findings, they should also consider fair ways of resolving any such bias. If there is a risk of discriminatory

---

[40] Hilde Lovett, «Kan vi forutsi det neste ranet I Oslo?» [Can we predict the next robbery in Oslo?], Norwegian Board of Technology, 2013, https://teknologiradet.no/kan-vi-forutsi-det-neste-ranet-i-oslo/.

[41] Paul Mozur, «One month, 500,000 face scans: How China is using A.I. to profile a minority», *New York Times*, 14 April 2019.

[42] Jon Kleinberg mfl., «Discrimination in the age of algorithms», 2019, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3329669.

[43] Viktor Mayer-Schönberger og Kenneth Cukier, «Big Data: A Revolution That Will Transform How We Live, Work, and Think» (London: Joh Murray, 2013).

outcomes, fairness should be incorporated as a goal for the algorithm to fulfil. The potential consequences of biases should be communicated to users.

Recommendations

- Research institutions should encourage the development of methods that can detect discrimination.

- Researchers should strive for transparency in data and algorithms and explain any lack of transparency.

- Researchers should reflect on the risk of bias from algorithms and incorporate fairness as a criterion when designing them. Researchers are also responsible for communicating to users any risk of discrimination from employing the research.

4 Conclusion

Big Data-based research opens a multitude of opportunities and can help us to solve major societal challenges, but it also raises a variety of new ethical challenges. Research ethics are absolutely key to ensuring that Big Data-based research serves us well, and that society retains its confidence in research. An overarching challenge in Big Data-based research is the complexity of the technology and the diversity of stakeholders who are often involved in collecting the data, developing methods of analysis and storing the data. Source data are often obtained outside of the research project, and research data is often transferred to large companies with storage and analytical resources sufficient to accommodate Big Data. This makes it hard for individual researchers to maintain a clear overview of the data's quality, the treatment of participants and what the research could mean for society. In this report we have identified research ethical norms that Big Data-based research may threaten and that we believe require special attention. We have also formulated a variety of recommendations for researchers, research institutions and other research stakeholders. FEK seeks dialogue on how to move forward with the issues and recommendations formulated here.

5 List of terms

Algorithm
A computer program with instructions to automatically employ data to produce a desired result or make a decision as efficiently as possible.

Anonymisation
A process that removes the risk of someone being identified in data material.

Artificial intelligence
Techniques developed to integrate computers into technological systems in ways that empower them to behave intelligently – that is, to solve cognitive and physical tasks previously reserved for humans. Some systems learn by themselves (see machine learning), while others are governed by rules defined by humans.

Big Data
Big Data is best understood in three interconnected and mutually reinforcing aspects: size, technology and methodological approach. First, Big Data generally comprises large data sets. Second, Big Data embraces relatively new and rapidly evolving technologies. Third, the methods developed to manage such large and often unstructured data sets are crucial; to exploit the potential of large amounts of data that have been produced or may be produced using new technology, new methodological approaches are often required.

Big Data-based research
Research that employs Big Data or generates Big Data.

Black box
A system or machine unit where the data entered and the result that emerges are visible, but not how the data was processed to produce the result.

Consent
In research ethics, obtaining informed consent obliges the researcher to give prospective participants enough information about a planned research project to allow them to make an informed and free choice about participation.

Data cleansing
The process of detecting and correcting corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.

Data curation
All activities that enable meaningful access to and use of data in a longer-term perspective.

Data minimisation
A basic principle of privacy and data protection. It asserts that the amount of personal data collected should be limited to what is necessary to achieve the purpose it was originally collected for.

Data mining
A process of reviewing existing data sets to generate new information.

Data noise
Incorrect or distracting information within a data set, including corrupted data. The terms «data noise» and «corrupted data» are often used interchangeably.

Data protection and privacy
From a legal point of view, data protection concerns the right to decide over one's own personal data and the right to privacy. Research ethics are more broad in scope, extending to matters of human dignity and integrity.

De-identification
Data associated with an individual, like a name, is replaced by a code. Identification therefore becomes more complicated, but re-identification is possible and the data can be linked back to the person.

Machine learning
Artificial intelligence systems that learn on their own.

Personal data
Any information that can be linked to an identifiable individual, whether directly or indirectly.

Profiling
Techniques used to analyse, predict and possibly influence future preferences and behaviour patterns.

Purpose limitation
A basic data protection principle which calls for identifying and precisely describing the purpose of the processing of personal data, and for the data to be processed strictly in accordance with the stated purposes.

Research
Research is a systematic activity, including data collection and analysis, that leads

to new insights and is carried out in accordance with scientific methods and norms. Ethical assessments are necessary regardless of the type of organisation or enterprise that conducts the research.

Research ethics
Values, principles and norms that help to compose and regulate scientific activity

Structured data
Data sets that conform to a formal structure, such as a customer registry where the data are arranged in rows and columns.

Unstructured data
Data sets that contain various types of data that cannot be sorted easily. Examples include images, video files or text files.

REFERENCES

Anderson, Chris. «The end of theory: The data deluge makes the scientific method obsolete». *Wired*, juni 2008. https://www.wired.com/2008/06/pb-theory/.

Crawford, Kate. «Critical questions for big data. Provocations for a cultural, technological, and scholarly phenomenon». *Information, Communication & Society* 15, nr. 5 (2012): 662–79. https://doi.org/10.1080/1369118X.2012.678878.

The National Committee for Research Ethics in Science and Technology (NENT), *Guidelines for Research Ethics in Science and Technology*. Oslo: The National Research Ethics Committees, 2016), https://www.forskningsetikk.no/retningslinjer/nat-tek/forskningsetiske-retningslinjer-for-naturvitenskap-og-teknologi/.

The National Committee for Research Ethics in the Social Sciences and Humanities (NESH), *A Guide to Internet Research Ethics.* Oslo: The Norwegian National Research Ethics Committees, 2018.

———. *Guidelines for Research Ethics in the Social Sciences, Humanities, Law and Theolog.* Oslo: The Norwegian National Research Ethics Committees, 2016.

European Data Protector Supervisor. «A preliminary opinion on data protection and scientific research», 2020. https://edps.europa.eu/sites/edp/files/publication/20-01-06_opinion_research_en.pdf.

———. «Opinion 4/2015. Towards a new digital ethics: Data, dignity and technology», 2015.

The Research Council of Norway. «The Research Council Policy for Open Science», 2020.

Regulation on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation).

Helene Ingierd og Hallvard Fossheim, «Innledning», in *Forskeres taushetsplikt og meldeplikt* [Researchers' duty of confidentiality and duty of notification], 7-12. Oslo: The Norwegian National Research Ethics Committees, 2013.

International Committee of Medical Journal Editors (ICMJE). «Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals», 2019. http://www.icmje.org/recommendations/.

Kammourieh, Lanah, Thomas Baar, Jos Berens, Emmanuel Letouzé, John Palmer, og Patrick Vinck. «Group privacy in the age of big data». I *Group privacy: New challenges of data technologies*, redigert av Luciano Floridi, Linnet Taylor, og Bart van der Sloot, 37–66. Springer International Publishing, 2017.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, og Cass R. Sunstein. «Discrimination in the age of algorithms», 2019. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3329669.

Laney, Doug. «3D data management: Controlling data volume, velocity, and variety», 2001.

Liu, Jianzheng, Jie Li, Weifeng Li, og Jiansheng Wu. «Rethinking big data: A review on the data quality and usage issues». *ISPRS Journal of Photogrammetry and Remote Sensing* 115 (2016): 134–42. https://doi.org/10.1016/j.isprsjprs.2015.11.006.

Lovdata.no. «Lov om behandling av personopplysninger (personopplysningsloven)» [Act of 14 April 2000 no. 31 relating to the processing of personal data (Personal Data Act)], Lovdata.no, 2018, art. 6a, https://lovdata.no/lov/2018-06-15-38/gdpr/a6.

Lovdata.no. «Lov om organisering av forskningsetisk arbeid (forskningsetikkloven)», 2017. https://lovdata.no/dokument/NL/lov/2017-04-28-23.

Lovett, Hilde. «Kan vi forutsi det neste ranet i Oslo?» [Can we predict the next robbery in Oslo?], Norwegian Board of Technology, 2013. https://teknologiradet.no/kan-vi-forutsi-det-neste-ranet-i-oslo/.

Mayer-Schönberger, Viktor, og Kenneth Cukier. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. London: Joh Murray, 2013.

Metcalf, Jacob, og Kate Crawford. «Where are human subjects in big data research? The emerging ethics divide». *Big Data & Society* 3, nr. 1 (2017): 205395171665021. https://doi.org/10.1177/2053951716650211.

Mittelstadt, Brent. «From individual to group privacy in big data analytics». *Philosophy and Technology* 30, nr. 5 (2017): 475–94.

Mittelstadt, Brent Daniel, og Luciano Floridi. «The ethics of big data: Current and foreseeable issues in biomedical contexts». *Science and Engineering Ethics* 22, nr. 2 (2016): 303–41. https://doi.org/10.1007/s11948-015-9652-2.

Mozur, Paul. «One month, 500,000 face scans: How China is using A.I. to profile a minority». *New York Times*, 14. april 2019.

Napoletani, Domenico, Marco Panza, og Daniele Struppa. «Is big data big enough? A reflection on the changing role of mathematics in applications». *Notices of the American Mathematical Society* 61, nr. 485–490 (2014).

Nissenbaum, Helen. *Privacy in context: Technology, policy, and the integrity of social life*. Stanford: Stanford University Press, 2011.

microdata.no. «Om microdata.no», undated. https://microdata.no/about/.

Prabhu, Robindra. «Big data - big trouble?» In *Internet research ethics*, edited by Hallvard Fossheim and Helene, 157–72. Oslo: Cappelen Damm, 2015.

Rao, Dhana, Venkat N. Gudivada, og Vijay V. Raghavan. «Data quality issues in big data». *Proceedings - 2015 IEEE International Conference on Big Data, IEEE Big Data 2015*, 2015, 2654–60. https://doi.org/10.1109/BigData.2015.7364065.

Rocher, Luc, Julien M. Hendrickx, og Yves-Alexandre de Montjoye. «Estimating the success of re-identifications in incomplete datasets using generative models». *Nature Communications* 10, nr. 1 (2019). https://doi.org/10.1038/s41467-019-10933-3.

Taylor, Linnet, Luciano Floridi, og Bart van der Sloot. «Introduction: A new perspective on privacy». I *Group Privacy: New Challenges of Data Technologies*, 1–12. Springer, 2017.

Teknologirådet. «Kunstig intelligens og norske helsedata» [Artificial intelligence and Norwegian health data], 2019.

Thaler, Richard H. «Nudge, not sludge». *Science* 361, nr. 6401 (2018): 431. https://science.sciencemag.org/content/361/6401/431.

Ward, Jonathan Stuart, og Adam Barker. «Undefined by data: A survey of big data definitions», 2013. http://arxiv.org/abs/1309.5821.

APPENDIX 1: MANDATE, WORKING GROUP, REFERENCE GROUP

The project has had a working group and a reference group. The working group has been responsible for the planning, execution and dissemination of the report. The reference group has been an advisory body for the working group, and has contributed in all phases of the project as needed.

Mandate

The working group will investigate research ethical dimensions of big data and thereby contribute to increased awareness, guidance and discussion of research ethics issues that arise with the use of such data. With this, FEK wants to contribute to good and responsible big data research in Norway.

The working group will especially consider:

What (national and international) resources already exist in this field, and which may be relevant for a research ethics study?
What is big data?
What are the key developments in the field?
What opportunities arise with big data research?
Which research ethical norms are affected and which challenges are raised?
What good examples of the use of big data exist?
What recommendations will the research communities and other relevant actors give on good and responsible use of big data?

Members of the working group

Sean Dexter Denham, The National Committee for Research Ethics on Human Remains

Trine B. Haugen, NEM

Helene Ingierd, FEK (Head of the working group)

Ivar Kolstad, NESH

Rune Nydal, NENT

Hannah Monsrud Sandvik, project employee in FEK 1/1/2019-1/13/2020, member of the working group since 1/13/2020

Ingrid Torp, FEK

Members of the reference group

Øyvind Kongstun Arnesen, The National Commission for the Investigation of Research Misconduct

Lene Bomann Larsen, NESH

Hallvard Fossheim

Christine Gjerdevik

Jostein Holmen

Hilde Lovett

Robindra Prabhu

Jim Tørresen, NENT

Lise Øvreås, NENT

The Norwegian National Research Ethics Committees
Kongens gate 14, N-0153 Oslo
Phone.: +47 23 31 83 00
www.researchethics.no