Big Data Ethics

Nicolae Sfetcu

08.09.2019

Sfetcu, Nicolae, "Big Data Ethics", SetThings (8 septembrie 2019), URL =
https://www.setthings.com/en/big-data-ethics/

Email: nicolae@sfetcu.com

A partial translation of:
Sfetcu, Nicolae, "Etica Big Data în cercetare", SetThings (6 iulie 2019), DOI:
10.13140/RG.2.2.27629.33761, MultiMedia Publishing (ed.), ISBN: 978-606-033-228-2, URL =
https://www.setthings.com/ro/e-books/etica-big-data-in-cercetare/

Ethical issues

Big Data ethics involves adherence to the concepts of right and wrong behavior regarding data, especially personal data. Big Data ethics focuses on structured or unstructured data collectors and disseminators.

Big Data ethics is supported, at EU level, by extensive documentation, which seeks to find concrete solutions to maximize the value of Big Data without sacrificing fundamental human rights. The European Data Protection Supervisor (EDPS) supports the right to privacy and the right to the protection of personal data in the respect of human dignity. According to these documents, the conceptual conflict between privacy and Big Data, and between intimacy and innovation, must be overcome. It is essential to identify the ways of including the ethical dimension in the development of innovations. (European Economic and Social Committee 2017)

According to the new EU Regulation 2016/679, data operators must implement the confidentiality measures and technologies to improve the confidentiality when determining the processing modalities and the processing itself. Through ENISA75 many privacy strategies have been identified by design (data minimization, hiding personal data and their interconnections, separate processing of personal data, choosing the highest level of aggregation, transparency, monitoring, privacy policy, legal issues).

A basic way for peaceful coexistence between Big Data exploitation and data protection is user *control* of personal data, which leads to transparency and trust between users and digital service providers. As outlined in the GDPR impact assessment,

> "Building trust in the online environment is key to economic development. Lack of trust makes consumers hesitate to buy online and adopt new services, including public e-government services. If not addressed, this lack of confidence will continue to slow down the development of innovative uses of new technologies, to act as an obstacle to economic growth and to block the public sector from reaping the potential benefits of digitization of its services." (European Data Protection Supervisor, Opinion 7/2015 *Meeting the challenges of Big Data A call for transparency, user control, data protection by design and accountability*.)

In the case of Big Data, traditional *consent* models are insufficient and outdated. The "consent should be granular enough to cover all the different processing and purposes of processing and reuse of personal data." (European Economic and Social Committee 2017)

A special problem is data *portability*, supported at EU level by the EDPS in Opinion 7/2015, (MORO 2016) where it is necessary to guarantee the right of citizens to access and correct personal data through an expanded control. Data portability can help increase consumer awareness and control by transferring online services.

The EDPS considers that personal data should be treated just like other important resources, such as oil, where the trading takes place between equally well-informed parties (informational symmetry). In fact, the market for personal information has a character of informational asymmetry, being neither transparent nor fair, customers are not compensated for the personal information they provide. Thus, the portability of the data would encourage a more competitive environment among the beneficiaries of this data, the users having the possibility to choose who offers the personal data.

Another approach involves the *storage* of personal data, with the possibility for the user to grant or withdraw consent for his personal data. (MORO 2016) (DG Connect 2015) The storage of personal data involves a "concept, framework, and architectural implementation that shifts data acquisition and control from a distributed data model to a *user-centric model*." (European Economic and Social Committee 2017) Data portability could ensure this.

The EDPS supports promoting responsible beneficiaries and reducing bureaucracy in data protection, through codes of conduct, audits, certifications, and a new generation of contractual clauses and mandatory corporate rules. The responsibility of Big Data beneficiaries involves the establishment of internal policies and control systems in accordance with the legislation in force, through intelligent and dynamic solutions that guarantee the respect of fundamental principles (data minimization, purpose limitation, data quality, correct and transparent data processing, design, storage limitation, integrity and confidentiality).

Data ethics is based on the following principles: *ownership* (individuals own their data), *transparency* of transactions (users must have transparent access to the algorithm design), *consent* (the user must be informed and expressly consent to the use of personal data), *privacy* (user privacy must be protected), *financial* (the user should know the financial transactions resulting from the use of his personal data), and *openness* (aggregated data sets should be freely available).

Ethics in research

The term critical data studies (CDS) implies that researchers are investigating Big Data from critical perspectives. The study of data in this context involves, in addition to their analysis, the incorporation of data into practices (knowledge), political and economic institutions and systems, through the complex interaction between data and the entities that produce, own and use them.

An OECD report (2013) underlines that, unlike the ethical norms applied to common research data, in the case of Big Data: (OECD 2013)

- Data collection was not subject to a formal ethical review process.
- Common ethical rules will not be implemented in the case of Big Data
- The use of research data may differ from the initial purpose.
- Data is no longer held as discrete sets.

The relationship between those who provide the data and those who use it is often indirect and variable. A more recent OECD report (2016) argues that this relationship is weaker or non-existent, with Big Data limiting common capabilities. (OECD 2016)

Data storage is important for research integrity. The data must have a clear provenance, with known, identified and documented sources and processing.

Many data that are not specifically collected for research have different standards in data research.

For some data, often of commercial value (e.g., data collected on Twitter), there are legal restrictions on their reproduction. (UK Data Service 2017)

Data storage must comply with standards of transparency and reproducibility.

Awareness

Awareness of the type of data that is provided during an online registration (for creating an account, or a subscription, for example) is a rare fact, especially since there is the possibility of using an existing digital identity (Facebook profile, for example) instead of a separate registration for faster access. Such situations create an opacity regarding the data shared between the identity provider and the service used.

Consent

In order to use the personal data of a person, his or her informed and explicit consent is required regarding who, when, how and for what purpose they are used. When data needs to be shared, these uses must be made known to the person. It should always be possible to withdraw consent for future use.

In Big Data analytics, very little can be known about the intended future uses of data, and about the benefits and involved risks. Here, there are procedures for "broad" and "generic" consent to share genomic data, for example, and for different purposes. Even when done correctly, there are some specific practical challenges: obtaining informed consent can be impossible or very costly, and the validity of consent is disputed when the agreement is required to access a service.

Control

In today's world, personal data can be traded just like any currency in Big Data implementation. There are different opinions to what extent this situation is ethical, including who to participate in the profit obtained from these transactions.

In the trading model of personal data, the transmission of personal data is a framework that offers people the opportunity to control their digital identity and create granular agreements of data sharing.

The idea of open data, centered around the argument that data should be freely available, is now emerging. Willingness to share data varies by person.

In the case of children, parents or tutors have responsibility for their data, which cannot be traded for financial benefits.

At national level, a government is sovereign over the generated and collected data. On October 26, 2001, the Patriotic Act entered into force in the US, and on May 25, 2018, the General Data

Protection Regulation 2016/679 (GDPR) at the European Union level, for the issues related to the protection of personal data.

In Big Data, the human-data relationship is asymmetrical, based on data control. The "right to be forgotten", adopted at EU level, is one of the basic elements of an individual's control over his personal data.

Transparency

Anticipatory governance involves Big Data-based predictive analytics to evaluate potential behaviors, with ethical implications that can encourage prejudice and discrimination.

A person who accepts the inclusion of his personal data in Big Data has the right to know why the data is collected, how it will be used, how long it will be stored, and how it can be modified.

Trust

Confidence in Big Data systems is linked to interdependence with confidentiality and awareness. So far, trust has been considered from a strictly technological perspective. It is hoped that hardware and software architectures will be developed that could increase trust between human beings and objects, and thus a greater acceptance of the use of personal data.

Ownership

A fundamental question in the ethics of Big Data research is, who owns the data? This involves the subject of property rights and obligations. In European law, the GDPR indicates that people have own their own personal data.

The sum of an individual's personal data forms a digital identity.

The protection of the moral rights (the right to be identified as a source of data, and to control them) of an individual is based on the opinion that personal data are a direct expression of his personality, and can only be transferred to another person, possibly, by succession when the individual dies.

The property implies exclusivity, i.e. the implicit restriction of others regarding access to the property. An efficient ownership of personal data involves portability, the ability to use alternatives without losing data. Standardization would also help to clean up your personal data.

At present, the data is owned by the owner of the sensors, the one who makes the recording or the entity that owns the sensor.

In the EU, the possibility of EU citizens' data being stored outside the so-called "Euro cloud" has been progressively reduced, but the problem of data already stored and processed elsewhere has not been resolved, and "does not resolve the ethical dilemma of how data ownership is defined philosophically, before passing to a more down-to-earth approach of law and policy making." (European Economic and Social Committee 2017)

Surveillance and security

More and more data sources are available with the help of advanced technologies such as CCTV, GPS, mobile devices, credit cards, ATMs. Also, active surveillance is a method of collecting data, but at the same time limiting the freedoms of citizens. Such permanent surveillance determines the increase of people's stress and creates their tendency to behave in a certain way that conforms to the expected norms.

Digital identity

Digital identity has the advantage of quick access to online content and related services. The use of digital identity has the potential to generate discrimination based on the representation of a person according to their online data, which may often not correspond to the real situation, in a process called "data dictatorship" in which "we are no longer judged on the basis of our actions, but on the basis of what all the data about us indicates our probable actions may be", (Norwegian Data Protection Authority 2013) personal interaction not being placed in a secondary plan.

Tailored reality

Any interaction we have with the Internet implies the possibility of storing our personal data. The processing and analysis of this data determines the personalized results that appear later on the Internet, through our search results, the display of products in online stores, the display of advertisements, etc. This generates a narrower and more personalized version of a user's previous online experience (the so-called "filter bubble." (Pariser 2011) An advantage is that the user will quickly find what he or she usually looks for, but excluding certain aspects, perspectives and ideas can lead to a restriction of creativity and the development of a tolerant attitude through the political and social isolation of the other aspects, by the lack of pluralistic views. (Crawford, Gray, and Miltner 2014)

De-identification

De-identification involves deleting or hiding elements that could immediately identify a person or organization. Legislation in different countries on data protection defines different treatments for identifiable data. Identifiability is increasingly seen as a continuum, not a binary aspect. Disclosure risks increase simultaneously with the number of variables, data sources and the power of data analysis. Disclosure risks may be mitigated but not eliminated. De-identification remains a vital tool for ensuring the safe use of data. (UK Data Service 2017)

Perfectly anonymous information taken separately can be combined with other data to uniquely identify a person with varying degrees of certainty. Profiling can become a powerful tool, raising concerns about the degree to which intrusion into an individual's life is allowed, the possibility of ensuring security, and surveillance.

Digital inequality

The advantages of Big Data size are clear, but there are also opinions that the accumulation of data on a huge scale presents specific risks. Because of this, there are few entities that have access, through infrastructure and skills, to Big Data systems. In this context, the costs and skills needed for access lead to certain specific digital inequalities addressed by ethics.

Privacy

In data transactions it is very important to ensure confidentiality:

> "No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honour and reputation. Everyone has the right to the protection of the law against such interference or attacks." - United Nations Declaration of Human Rights Article 12.

In many countries, public monitoring of the data by the government to observe citizens requires explicit authorization through an appropriate judicial process. Privacy is not about keeping secrets, but about choice, human rights, and freedom.

Often privacy is wrongly viewed as a binary choice between isolation and scientific progress. Identity protection in data is technologically possible, for example using homomorphic encryption and algorithmic design.

Privacy as a limitation of the use of data can also be considered unethical, (Kostkova et al. 2016) especially in healthcare, but it should be kept in mind that it is possible to extract the value of the data without compromising privacy.

Privacy is recognized as a human right by numerous national and international regulations. Privacy in research is achieved through a combination of approaches: limiting the collected data, anonymizing them; and regulating access to data. In the case of Big Data research, specific problems arise: the ambiguity between the terms "privacy" and "confidentiality; the declaration of social spaces as public or private; the ignorance of the risks of privacy by users; the blurred distinction between public and private users. Currently there are disputes whether data science it should be classified as a research of human subjects, and therefore not subject to the usual rules of privacy.

Big Data research

Through the new concepts of "algorithmic damage", "predictive analysis", etc., the algorithms currently used in Big Data operations go beyond the traditional view of privacy. According to the US National Science and Technology Council,

> "“Analytical algorithms” as algorithms for prioritizing, classifying, filtering, and predicting. Their use can create privacy issues when the information used by algorithms is inappropriate or inaccurate, when incorrect decisions occur, when there is no reasonable means of redress, when an individual’s autonomy is directly related to algorithmic scoring, or when the use of predictive algorithms chills desirable behavior or encourages other privacy harms.” (NSTC (National Science and Technology Council) 2016, 18)

Big Data research is what the ethicist James Moor would call a "conceptual muddles" due to the "inability to properly conceptualize the ethical values and dilemmas at play in a new technological context." (Buchanan and Zimmer 2018) In this situation privacy is ensured through a combination of different tactics and practices (controlled or anonymous environments, limitation of personal information, anonymization of data, access restrictions, data security, etc.). In general, all related concepts become confusing in the case of Big Data. Thus, social posts are considered public on social networks in case of an appropriate setting. But social networks are complex environments of socio-technical interactions where users do not always understand the functionality of the settings and terms of use. Thus, there is uncertainty about users' intentions and expectations, and these conceptual deficiencies in the context of Big Data research lead to uncertainties regarding the need for informed consent.

Bibliography

Buchanan, Elizabeth A., and Michael Zimmer. 2018. “Internet Research Ethics.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2018. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2018/entries/ethics-internet-research/.

Crawford, Kate, Mary L. Gray, and Kate Miltner. 2014. "Big Data| Critiquing Big Data: Politics, Ethics, Epistemology | Special Section Introduction." *International Journal of Communication* 8 (0): 10. https://ijoc.org/index.php/ijoc/article/view/2167.

DG Connect. 2015. "Study on Personal Data Stores Conducted at the Cambridge University Judge Business School." Text. Digital Single Market - European Commission. August 7, 2015. https://ec.europa.eu/digital-single-market/en/news/study-personal-data-stores-conducted-cambridge-university-judge-business-school.

European Economic and Social Committee. 2017. "The Ethics of Big Data: Balancing Economic Benefits and Ethical Questions of Big Data in the EU Policy Context." European Economic and Social Committee. February 22, 2017. https://www.eesc.europa.eu/en/our-work/publications-other-work/publications/ethics-big-data.

Kostkova, Patty, Helen Brewer, Simon de Lusignan, Edward Fottrell, Ben Goldacre, Graham Hart, Phil Koczan, et al. 2016. "Who Owns the Data? Open Data for Healthcare." *Frontiers in Public Health* 4. https://doi.org/10.3389/fpubh.2016.00007.

MORO, Veronica. 2016. "Meeting the Challenges of Big Data." Text. European Data Protection Supervisor - European Data Protection Supervisor. November 16, 2016. https://edps.europa.eu/data-protection/our-work/publications/opinions/meeting-challenges-big-data_en.

Norwegian Data Protection Authority. 2013. "Big Data – Privacy Principles under Pressure." https://www.datatilsynet.no/globalassets/global/english/big-data-engelsk-web.pdf.

NSTC (National Science and Technology Council). 2016. "National Privacy Research Strategy." https://obamawhitehouse.archives.gov/sites/default/files/nprs_nstc_review_final.pdf.

OECD. 2013. "New Data for Understanding the Human Condition: International Perspectives." http://www.oecd.org/sti/inno/new-data-for-understanding-the-human-condition.pdf.

———. 2016. "Research Ethics and New Forms of Data for Social and Economic Research," November. https://doi.org/10.1787/5jln7vnpxs32-en.

Pariser, Eli. 2011. *The Filter Bubble: What The Internet Is Hiding From You*. Penguin Books Limited.

UK Data Service. 2017. "Big Data and Data Sharing: Ethical Issues." https://www.ukdataservice.ac.uk/media/604711/big-data-and-data-sharing_ethical-issues.pdf.