The ethics of Big Data:

Balancing economic benefits and ethical questions of Big Data in the EU policy context

STUDY

European Economic and Social Committee

European Economic and Social Committee

The ethics of Big Data: Balancing economic benefits and ethical questions of Big Data in the EU policy context

Study

The information and views set out in this study are those of the authors and do not necessarily reflect the official opinion of the European Economic and Social Committee. The European Economic and Social Committee does not guarantee the accuracy of the data included in this study.
Neither the European Economic and Social Committee nor any person acting on the European Economic and Social Committee's behalf may be held responsible for the use which may be made of the information contained therein.

evodevo

This study has been prepared by Evodevo srl (Italy). The research team was coordinated by Loredana Ceccacci with the scientific supervision of Stefano De Luca. The research team was composed of Stefano De Luca, Loredana Ceccacci, Danilo Pecorino, Riccardo Troccoli, Claudia Corcione. The law firm E-Lex has provided advice on the legal and regulatory issues covered by the study.

Abstract

This study, carried out to support the activities of the EESC, explores the ethical dimensions of Big Data in an attempt to balance them with the need for economic growth within the EU. In the first part of the study an in-depth review of the available literature was carried out, to highlight ethical issues connected with Big Data. Five actions were devised as tools to strike the balance described above. The second phase of the study involved interviews with a number of key stakeholders and conducting a survey that acquired information on the general knowledge of the issues connected to the use of Big Data. Feedback on the proposed balancing actions was also sought and taken into consideration in the final analysis. Attitudes as emerged from interviews and survey most often ranged from concerned to worried, while benefits of Big Data were seldom discussed by the respondents. Benefits are, nevertheless, intrinsic to Big Data, as well as risks, and they are discussed more broadly throughout the study.

Table of contents

Executive summary

In June 2016, the European Social and Economic Committee (EESC) published a tender for a "Study on the ethics of Big Data: Balancing economic benefits and ethical questions of Big Data in the EU policy context". The EESC selected Evodevo srl to conduct the study.

The study, which was initiated in July 2016 and concluded in December 2016, mainly aims to ascertain how to balance human values that are fundamental to the European civil society, such as privacy, confidentiality, transparency, identity and free choice with the compelling uses of Big Data for economic gains.

The activities were divided into **two main phases**. During the first two months, intense information-gathering activities were performed, with **the collection and critical review** of articles, publications, reports, and studies. In parallel, the research team reviewed the **official positions** of EU institutions, organizations, and think tanks on the topic. The consideration of the material collected has supported the design of **five balancing actions** as policies that can be implemented to make the most of the Big Data while protecting fundamental human rights.

During the **second phase** of the study, concepts and themes emerged during the stocktaking process have been further discussed with relevant stakeholders, whose insights have been used to integrate and fine tune the balancing actions proposed.

A **legal framework** was also provided to understand how the topic of Big Data is reflected in the European legal framework. Several EU institutions reports argue that the existing legal framework – based, in particular, on Directive n. 46/95/EC and, today, on the General Regulation on the Protection of Personal Data - already offer adequate protection against any infringement of the fundamental rights of the citizens. However, a comprehensive and global strategy is required, taking into account the concentrations of power in private entities that collect and process personal data, as well as appropriate forms of protection in view of the changed perspective of privacy.

Over the years, **the protection has shifted from a right to exclude others to a right to the control of personal data**. We are now facing a third phase, which asks the question "who am I?" and the answer does not depend on the data subject but on the patterns selected by third parties that create an analytical profile and lead to rethinking the right to identity. The new frontier of data protection involves not exclusively personal data, but, more broadly, data. The passage is from data that is able to identify specific physical persons to data that can detect specific behaviours, consumptions, health data etc. of groups rather than of individuals.

The European debate that emerged from the desk research ranges from the multiple opportunities that Big Data can bring within the next decades to the concerns about the tangible impact that the massive analytics application can have on people's lives. The exposure to the influence of data analytics is a lifelong experience for individuals that, nonetheless, still have a low awareness of how their data are used to predict their behaviour and shape their virtual identity. This knowledge asymmetry makes individuals vulnerable, with limited resources to exercise fully their fundamental rights and freedoms. **Positive aspects** of Big Data, and their potential to bring improvement to everyday life in the near future, have been widely discussed in Europe. Examples span from **health services**, to **road safety**, **agriculture**, **retail**, **education** and **climate change mitigation** and are based on the direct use/collection of Big Data or inferences based on them.

The consideration of the material collected has supported the design of **five balancing actions** as policies that can be implemented to make the most of the Big Data while protecting fundamental human rights.

The **first** action, *EU privacy management platform*, comes from the idea, convincingly conveyed in the General Data Protection Regulation (GDPR), that natural persons should have control over their own personal data. The idea is to establish a pan-European web portal as a unique privacy management hub where the European citizens voluntarily register to get access to a personal dashboard to visualize the list of all the public and private entities that have gained and currently store, process, share and re-use their personal data.

The objective of the **second** balancing action, *Ethical Data Management Protocol*, is to increase transparency and making people aware of the level of compliance with the EU law of Big Data's holders, both public and private. The idea is to design a sound European **certification system** to identify the virtuous companies in the field of data protection. This measure is already outlined by the EU legislator in the GDPR (article 42) and during the study the research team reflected and discussed with the stakeholders to understand how it may be applied.

The **third** proposed action, *Data Management Statement*, involves companies on a voluntary level. It lies on the assumption that nowadays the success of an organisation increasingly depends upon the trust of shareholders, customers, employees and the general public. In order to boost the confidence of internal and external stakeholders, the organisations may submit declarations on how they collect, use or sell personal data coming from customers and general business activities.

The **fourth** action, *European e-Health Database*, involves the creation of a **European database which contains health related data of EU citizens, to be used for scientific research**. At the time an EU citizen is treated in a public hospital or a private institution that receives public funding, they are asked for consent for their data to be collected and stored in an EU managed database. Data collection and transfer would follow standard exchange protocols, such as those defined by the standard Health Level 7 in similar fields, defined beforehand at a European level. Concerning the use of health data for scientific studies, in order to comply with the GDPR, the person would have the right to give consent only to certain areas of research.

The **fifth** and final action, *Digital education on Big Data*, aims to create a broader digital culture in Europe, specifically aimed at the development of a much deeper understanding of Big Data, how it interacts with EU citizens throughout their life and how it affects each individual. In order to promote this, a series of educational programs are proposed aimed at the different age groups, administered via both compulsory school years and optional courses.

All the people interviewed during the **second phase of the study** expressed their own opinion on the topic and specifically on the actions proposed.

The research team contacted representatives of:

- Data protection authorities
- Research institutes and centres for statistics
- Consumers' associations
- Universities
- EU bodies
- Analytics big firms
- Companies making use of Big Data analytics.

The extraction of information from each interview followed four steps, the aim of which is the standardization of the issues expressed by each interviewee in order to compare the position of each stakeholder with that of the others.

To further deepen the understanding of the attitude towards Big Data, an online survey was conducted. After being published, it ran for one month, and it aimed at gathering the opinions of both big analytics firms and data-driven companies such as big retailers on the web, or utility providers.

Results of both interviews and surveys clearly show that the main stakeholders, and the European Data Protection Supervisor (EDPS) *in primis*, are looking towards concrete solutions to make the most of the Big Data value without sacrificing human fundamental rights.

The need to empower people and to raise the general understanding of the dynamics, interests and values affected in the use of personal data is something that has been clearly recognised in literature and in formal statements of relevant official entities and experts in Europe. This view has been confirmed by almost all the interviewees.

The discussions with the experts also made it clear that the **investment in education** and **awareness-raising** represents the core element that would enable the other policies because it generates a bottom-up demand for transparency and fairness emerging directly from citizens. This demand from citizens and consumers can't be disregarded by the data driven market.

Europe could work together to identify common contents to introduce in the education curricula. This can cover primary, secondary, post-secondary education, and even life-long learning initiatives. This policy could run in parallel and consistently with the actions that the European Commission are carrying on in the field of education, as the new Digital Skills and Jobs Coalition initiative that calls for concrete measures to bring digital skills to all levels of education and training to succeed in the digital world. Besides promoting the acquisition of technical skills in view of improving employability and competitiveness among European citizens, specific training could be designed to instruct people on privacy as a value and right, ethical issues of behaviour profiling, virtual identity related risks and digital reputation control, ownership of personal contents, digital footprints, intellectual property rights.

More specific integrations on ethics principles and requirements for quality and integrity in research could be foreseen for bachelor's and master's degree programmes in Statistics, Informatics, Data Science, Computer Science, Artificial Intelligence and correlated subjects.

Instead, the proposed idea of setting up a European Portal where collecting information on how the personal data of European citizens are stored and processed doesn't seem feasible or even very useful in the current reality.

The stakeholders expressed positive attitude toward the idea to promote the commitment of big companies and organizations. In this sense, the next step could be to open as soon as possible the procedure of designing a European certification system that can support the companies in complying with the GDPR and help people recognise the service providers that guarantee a fair, legal and transparent process of personal data. This procedure could take place in the context of a wider consultation, with the involvement of the private subjects that are expected to be most affected by the implementation of the Regulation itself, such as the analytics companies and big firms that make use of personal data to offer their services.

Governments could also promote among companies and organisations that make use of personal data ways to show and communicate to customers their commitment to acting in accordance with the new Regulation and their willingness to go beyond the rules, to guarantee an even higher level of data protection. A strategic approach for communicating such good practices could benefit the reputation of companies and increase the trust of customers toward their products and services. Connecting this solution with Corporate Social Responsibility strategies and Socially Responsible Investments could pave the way for this kind of approach.

Instead, the idea of creating a centralised datacentre at European level was perceived as too risky due to possible data breaches and misuses. The investment in this field should first move instead toward a standardisation of data collecting and storage at national level in each country, to create at least common and comparable databases. As Member States reach a good level of internal standardisation, the cooperation at European level can be simplified. This could then facilitate the sharing of mainly anonymised data among research focused entities such as scientific institutes, hospitals, public or private healthcare structures.

1. Introduction

Setting the scene

It is Sunday, and Alice wakes up early in the morning. She starts her day by running, which she monitors by wearing a device that tracks her timing, location and body conditions. Through a connection with her Smartphone, a software uses this data, matched with that of millions of other users, to offer advice as a virtual personal trainer. She can search for the most popular routes and share her achievements with her friends by publishing them on social networks. Back home, Alice has a good breakfast, as suggested by a mobile app that helps her count calories while recommending healthy foods.

She then opens her laptop and reads the news from a content aggregator of news websites and private blog posts. She receives personalized contents based on her interests and web history. Once on the web, she starts looking for a more gratifying job. Alice uses specialised websites to upload her curriculum vitae and receive recommendations from enthusiastic colleagues. Sometimes, she posts and shares interesting facts showing the world her skills and competences to build a stronger virtual image, a persona, and become a stronger candidate in the job market. Again, she receives tailored job offers based on her behaviour, skills and previous positions. This could be a good day to start a new career!

But wait! It is late. It is time to pick up her husband and daughter from the airport. They are arriving from a holiday in Pakistan where they went to enjoy the wonderful artistic heritage of that country. Alice is not in a hurry though because she has discovered that their flight is delayed after consulting the airline's app. Alice uses a collaborative navigation app that, by using real-time drivers' data, suggests the most suitable route, thus avoiding heavy traffic. Alice herself reports a car accident along the way. She then refuels her car from the cheapest gas station on her route based on community-shared fuel prices.

At the airport, Alice is tracked by cameras that scan her face to identify known terrorists. A few minutes' wait and her family arrives! Her daughter captures the moment with a selfie which she immediately posts on her favourite social network, where automatic face recognition is used to tag people in photos.

Alice's husband faints, due to jetlag and tiredness caused by the trip. In seconds, airport personnel recognise the problem with the support of a camera-based software that identifies the situation of "body to the ground", and take him to the emergency room. Health checks, including ECG and blood tests, show that there is no problem, although a software matches these tests against a large database, and warns about a 40% chance of having future problems with his blood pressure, considering that he is overweight. Finally, Alice can go back home with her husband and daughter.

Every day, we use and generate tons of data, feeding Big Data of government agencies, private companies and even private citizens. As shown in this brief story, we benefit in many ways from the existence and the use of Big Data, but we also need to remember that “there is no such thing as a free lunch”. There are risks in using Big Data, a sort of dark side. We can see some examples:

Alice is selected for a really interesting job position, potentially winning over more highly qualified candidates, simply because they did not use her same website and because they are not as skilled in setting up a digital persona able to attract companies’ attention. Another possibility is that a recruiter looking at Alice’s social network account, finds a photo of her husband with a long beard after his trip to Pakistan, discards her submission in order to avoid the risk of having a relative of a potential fundamentalist in the company.

By reading only news selected according to her recorded political, religious and ethical points of view, Alice risks having a limited, even though nicely customized world perception, building a gilded cage around her.

Also, Alice’s husband might have some problems: an insurance company may decline his application for life insurance or ask a higher price after assessing his risks of high blood pressure and obesity, as derived from accessing the electronic patient record of the emergency room, which is operated by an affiliated company.

Market changes can be produced by using Big Data: while they can be positive for someone, they may harm somebody else. For instance, a fuel pump with a higher price can be deserted by people using the same navigator app as Alice, but its owner cannot lower his retail price because his rent is higher due to being in a more expensive neighbourhood – so he is forced to shut down.

Potential risks also come from public agencies. For instance, Alice may be investigated by the police because she was recorded by the airport CCTV cameras while speaking to a person under surveillance, although she was simply asking a stranger for directions to the WC.

It is clear that Big Data can bring benefits to European citizens and companies, but they have to be balanced by an increased awareness of the “dark side” of Big Data and to exploit the new capabilities and opportunities that they are offering to all of us.

2. The Study on the Ethics of Big Data commissioned by the EESC

2.1 Origin and development of the Study

In June 2016, the European Social and Economic Committee – EESC - published the “Study on the ethics of Big Data: Balancing economic benefits and ethical questions of Big Data in an EU policy context” tender.

Evodevo submitted the tender and was selected by the EESC to conduct the Study.
The Study, which was initiated in July 2016 and concluded in December 2016, mainly aims to ascertain how to balance human values that are fundamental to the European civil society, such as privacy, confidentiality, transparency, identity and free choice with the compelling uses of Big Data for economic gains.

The activities were divided into two main phases. During the first two months, intense information-gathering activities were performed, with the collection and critical review of articles, publications, reports, and studies. In parallel, the research team reviewed the official positions of the main European Authorities in the EU, organizations, and think tanks on the topic.

The consideration of the material collected has supported the design of five balancing actions as policies that can be implemented to make the most of the Big Data while protecting fundamental human rights.

During the second phase of the study, concepts and themes emerged during the stocktaking process have been further discussed with relevant stakeholders, whose insights have been used to integrate and fine tune the balancing actions proposed in Chapter 9 of this study.

2.2 Structure of this study

The study is organized in eleven chapters:

- Chapter 1: Introduction, with a brief fantasy story that sets the scene of the study
- Chapter 2 This chapter describes the structure and development phases of the study. In this chapter the key definitions and concepts that are used throughout the rest of the study are provided, detailing in particular what is meant by Big Data.
- Chapter 3 Philosophical aspects. This chapter contains a philosophical perspective on the ethical dimensions identified in the study
- Chapter 4: European agenda on the digital society. This chapter contains a general overview of the main actions and programmes promoted by the European Institutions to boost the digital society and economy
- Chapter 5: General Review. This chapter examines the main dimensions connected to the topic as they are revealed from the literature review carried on. It also contains a description

of the relevant legislation, with a detailed discussion of the General Data Protection Regulation (Regulation EU 2016/679). Finally, a comprehensive analysis of the key stakeholders' positions about the topic is presented here.

- Chapter 6: A life of Big Data. This chapter presents actual scenarios in which European citizens find themselves dealing with Big Data. These scenarios are grouped by life phase in order to provide a perspective on how Big Data is presently interacting with individuals throughout their whole life.
- Chapter 7: Ethical issues. The scenarios described in chapter 6 give rise to ethical issues that are described here in this chapter. A summary table is provided at the end.
- Chapter 8: Real-world uses for Big Data: five case studies in Europe. The chapter presents five concrete experiences of using big data in innovative ways
- Chapter 9: Balancing actions and effectively balanced scenarios. In this chapter, five actions that are aimed at balancing the right to privacy of the EU citizens and the need for growth are proposed. In addition, their effects on the relevant ethical issues in producing effectively balanced scenarios are assessed.
- Chapter 10 Stakeholders' opinion landscape. The designed balancing actions have been broadly discussed with selected key persons heard via interviews and asked to fill in an online survey. In this chapter the main results of these consultations are summarized and reported with the conclusions and final comments about the balancing actions proposed.
- Chapter 11 General conclusions.

2.3 Definition of Big Data

As the amount of data keeps growing exponentially, compounded by the Internet, social media, cloud computing, mobile devices and governmental data, it poses both a threat and an opportunity for Europe in terms of how to manage and make use of this ever increasing amount of data for economic growth, while keeping EU citizens' rights safeguarded.

Since the production of user-generated data is expected to grow by 2000%[1] globally by 2020 and since it comes from a diverse range of sources[2], definitions of Big Data are varied[3] in terms of the focus they put on one aspect or another, but they all have in common the fact that they refer to a large amount of data, **much larger than what can be analysed on a single computer today**, coming from different sources and in different, often unstructured, formats. A definition was provided in 2001[4] and later modified to what was presented by IBM scientists[5] (see Figure 1), and it has become more and more accepted. It states that Big Data is characterised by:

- Volume, referring to the scale of data;
- Variety, since data is produced by different data sources in different formats;
- Velocity, which is connected to the analysis of streaming data;

[1] Tucker (2013). *Has Big Data made anonymity impossible?* In: Big Data gets personal – MIT Technology Review.

[2] Cumbley and Church (2013). *Is "Big Data" creepy?* Computer Law & Security Review, **29:** 601–609.

[3] Ward and Barker (2013). *Undefined by data: A survey of Big Data definitions*. arXiv preprint arXiv: 1309.5821

[4] Douglas (2001). *3d data management: Controlling data volume, velocity and variety*. Gartner.

[5] https://www-01.ibm.com/software/data/bigdata/what-is-big-data.html

- Veracity, as data is uncertain and needs to be verified before or during use;
- Value, which can be produced by analysing Big Data.

The volume of data produced and stored has been constantly increasing worldwide, and data generation has been estimated at 2.5 Exabytes of data per day in 2016, and it is expected to grow exponentially[6]. For instance, it is estimated that 90% of all data in the world today has been generated over the last two years. High volumes of data are closely connected to a constantly increasing need for quick analyses to generate fast insights, from a perspective of using Big Data for growth[7]. Both volume and velocity have a strong impact on veracity, since processing a very large amount of data in different formats coming in at high speed is worthless if that data is incorrect. Incorrect data, in fact, has the potential to generate issues when used in decision making processes by governments or companies, and ultimately affects citizens and consumers.

Therefore, the need to ensure that the data, as well as the analyses performed on this data, is correct is paramount when dealing with Big Data[8]. This is particularly relevant in automated decision-making, where no human is involved in the process.

Data types are especially varied in the case of Big Data, and they span from satellite imagery[9] to environmental data from sensors[10], use data of mobile devices[11], digital pictures and videos (e.g. videos uploaded on YouTube), health data collected by wearables[12], and data generated by web users or submitted during registration processes (i.e. registration forms). Moreover, a massive increase in the type of data, as well as volume, is expected to take place as soon as the use of items from an Internet-of-Things perspective becomes more widespread[13].

[6] Bello-Orgaz et al (2016). *Social Big Data: Recent achievements and new challenges*. Information Fusion, **28**: 45–59.
[7] SAS (2013). *Big Data Analytics – Adoption and Employment Trends, 2012–2017*
[8] Rockwell and Sinclair (2016). *False positives: Opportunities and dangers in Big-Data text analysis*. In Hermeneutica: Computer-assisted interpretation in the humanities. MIT Press, Cambridge (USA-MA)
[9] Vatsavai et al (2012). *Spatiotemporal Data Mining in the era of Big Spatial Data: Algorithms and applications.* Proceedings of the 1st ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data.
[10] Vitolo et al (2015). *Web technologies for environmental Big Data.* Environmental Modelling & Software, **63**: 185–198.
[11] Hull et al (2011). *Contextual gaps: Privacy issues on Facebook.* Ethics and Information Technology, **13(4)**: 289–302.
[12] Michael et al (2014). *Uberveillance and the Internet of Things and People.* Proceedings of the 1st International Conference of Contemporary Computing and Informatics, 1381–1386.
[13] Gudymenko et al (2011). *Privacy Implications of the Internet of Things.* Proceedings of the 2nd International Joint Conference on Ambient Intelligence, 280–286.

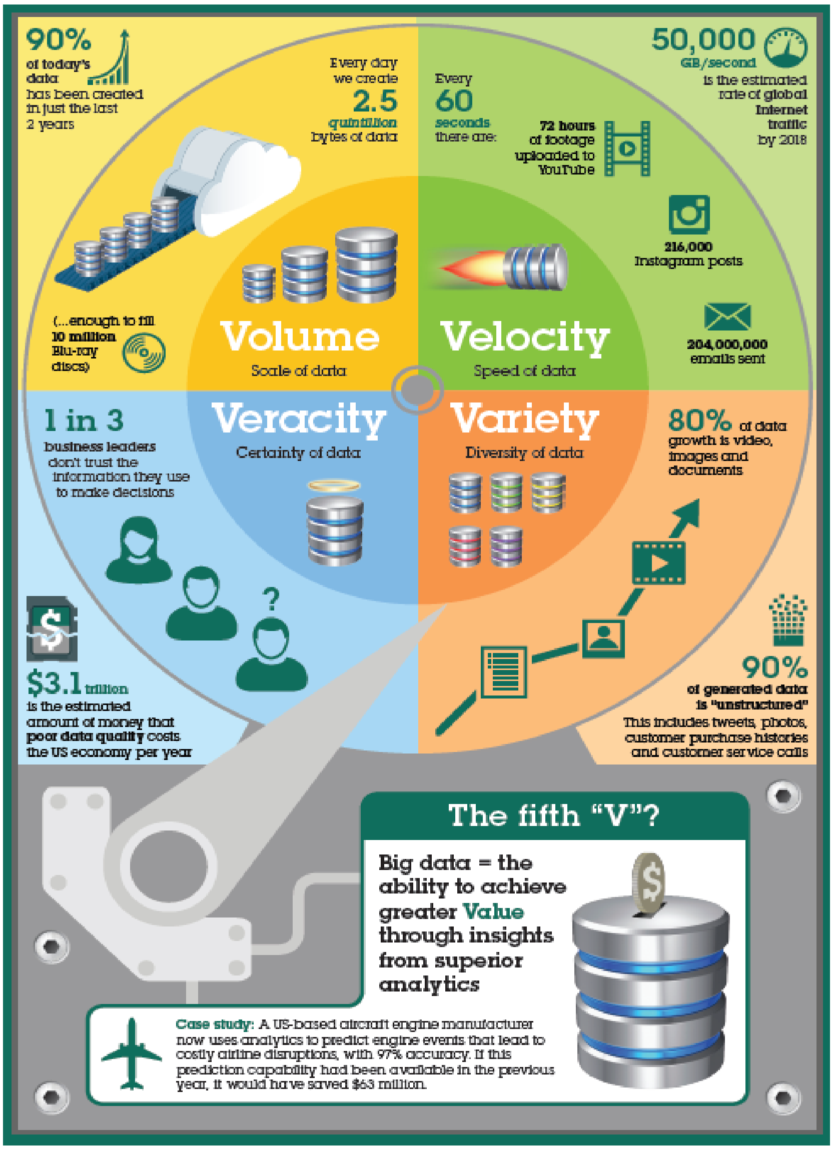

90% of today's data has been created in just the last 2 years
Every day we create 2.5 quintillion bytes of data
(...enough to fill 10 million Blu-ray discs)
Volume
Scale of data
Every 60 seconds there are:
72 hours of footage uploaded to YouTube
216,000 Instagram posts
204,000,000 emails sent
50,000 GB/second is the estimated rate of global Internet traffic by 2018
Velocity
Speed of data
1 in 3 business leaders don't trust the information they use to make decisions
$3.1 trillion is the estimated amount of money that poor data quality costs the US economy per year
Veracity
Certainty of data
Variety
Diversity of data
80% of data growth is video, images and documents
90% of generated data is "unstructured"
This includes tweets, photos, customer purchase histories and customer service calls
The fifth "V"?
Big data = the ability to achieve greater Value through insights from superior analytics
Case study: A US-based aircraft engine manufacturer now uses analytics to predict engine events that lead to costly airline disruptions, with 97% accuracy. If this prediction capability had been available in the previous year, it would have saved $63 million.

Figure 1. Infographic that illustrates the meaning of the five V's related to Big Data (http://www.ibmbigdatahub.com/infographic/four-vs-big-data)

Furthermore, it must be kept in mind that the variety of sources of Big Data can produce unexpected outputs when datasets, each bringing its own portion of information, are combined to produce, for instance, a more complete profile of a user.

This availability of data, which includes personal information of at times unaware citizens, is unprecedented and represents a unique opportunity for present day governments and companies to improve services and welfare with the added challenge of doing it while respecting privacy and human dignity, which are part of Europe's core values.

One of the seldom discussed topics is that Big Data simply does not exist without an Internet connection, regardless of the number of sensors, apps, CCTV cameras and satellites that are collecting data. When a mobile phone is measuring heart rate and walking speed, it still needs to send this data to a server where other data is stored in order to create compounded datasets that can be accessed and processed. If data were only stored locally on a device, the whole process of aggregation and processing of data, and joining of disparate datasets from various sources, could not take place. Sheer size, as in *volume* of data, is also one of the founding principles of the phenomenon of Big Data.

One of the reasons behind the rise of Big Data is, in fact, most likely found in the more widespread availability of Internet connection at affordable rates, due to the development of infrastructures and more efficient data transfer technologies that allow for a faster data transfer. If, in fact, Internet was at first a rare commodity only accessible through computers by people skilled in their use, it has progressively been integrated into other devices and Internet connection capabilities are now found in all computers, mobile devices, televisions, vehicles such as cars, and will soon be present in other everyday objects within an Internet-of-Things landscape.

The increasing availability of an Internet connection has also been quantified by adapting the Gini coefficient used in economics to assess the equality in distribution of income across a population. Results[14,15] show that there is a global trend which is leading to a more equally distributed access to ICT technologies in general, and the Internet more specifically, and this is in strong favour of supporting the idea that Internet access has been at the foundation of the rise of Big Data.

As an interesting note, a strong correlation between the GDP per capita and the ICT diffusion index is present, thus possibly indicating that a more developed economy is at the root of the development and adoption of innovative communication technologies.

2.3.1 Proposed classification dimensions of Big Data

Despite not being aware of it at times, Big Data is a phenomenon that human beings are dealing with. They are getting in touch with systems that store personal data in massive databases in many ways and they can be affected by the results of Big Data processing either as an individual or as part of a group. These different attributes of the interaction and use of Big Data are useful in classifying them.

The concept that data is shared and stored on servers through the use of Internet implies that this process can take place in two ways depending on how the interaction occurs between the subject that produces data and the storage system. According to the kind of interaction, we can identify:

- *Active Big Data*, when a user directly sends data to a storage system (e.g. data collected by the apps of mobile devices for which explicit consent was provided, data submitted during registration for the creation of a digital identity).
- *Passive Big Data*, when a citizen’s data is collected by another person and then input into an online storage system (e.g. personal details and results of analyses collected by hospital staff during health care).

[14] Howard et al (2010). *Comparing digital divides: Internet access and social inequality in Canada and the United States*. Canadian Journal of Communication, **35**: 109–128.

[15] United Nations (2006). *The digital divide report: ICT diffusion index 2005*. United Nations Publications.

Even in the case when data is directly transferred from a user’s device to a storage system, it should not be implied that proper notification, which is timely and explicit enough to be clearly understood, is given. Data transfer, in fact, regardless of how direct the connection between a subject and the storage system is, may take place in more or less explicit ways. It is, therefore, possible to distinguish between:

- *Consciously Transferred Data*, when a user is timely and clearly informed that data about them is being collected and stored, therefore awareness can be safely assumed.
- *Not-Consciously Transferred Data*, when a timely and clear notification has not been provided, therefore it cannot be assumed that a person is aware that data collection and storage is taking place.

After data collection following one or a combination of the possibilities just described, data is usually processed to generate insights. Processing of data follows different paths when it is used to identify and characterize groups within a dataset or find information about a single individual. Therefore, the three following dimensions of Big Data are devisable:

- *Individual dimension*, when an individual is the object of an analysis which is carried out to understand their behaviour, spending habits, etc. (e.g. purchase history of a user on an online marketplace can be used to generate recommendations).
- *Social dimension*, when analyses are aimed at identifying discrete groups within a population, and labelling them for future use (e.g. a group of students can be identified as fast learners, and this may lead to personalized learning actions).
- *Hybrid dimension*, when an individual is defined by the belonging to a group whose characteristics have already been defined and actions towards that individual are based on the features of this group (e.g. an organization might decide not to employ a candidate according to their religion due to a bias towards that group and the potential integration with the company’s policies).

3. Philosophical aspects

The analysis of Big Data from an ethical point of view involves two main intertwined aspects: a theoretical one, i.e. the philosophic description of the elements subject to ethic control, and a pragmatic view of the impact on the life of people and organizations.

The ethical impact of computing is hardly a new argument: it was already raised for computing in general, for instance in "Is there an Ethic of Computing?"[16], and for user interaction issues as in "Human Agency and responsible computing"[17]. An important debate on the ethical issues caused by artificial intelligence exists, and it ranges from problems derived by its use, as in the very deep analysis of the impact on job creation in "The future of employment"[18], to the creation of new types of moral actors as in "Why machine ethics?" or "Artificial Agents and Their Moral Nature"[19] about the emergence of increasingly complex autonomous software agents and robots. A strong relationship is present between Big Data and artificial intelligence, considering that data are useless without interpretation, and it is almost impossible to proceed from data to knowledge when the size of data exceeds the personal dimension. Hence, we are forced to use automatic tools such as artificial intelligence, or its derivate: machine learning, semantic analysis, data mining.

A largely used approach to ethics is moral agency[20] in which at least the three conditions of causality, knowledge and choice are described. Quoting Noorman:

- There should be a causal connection between the person and the outcome of actions. A person is usually only held responsible if she had some control over the outcome of events.
- The subject has to have knowledge of and be able to consider the possible consequences of their actions. We tend to excuse someone from blame if they could not have known that their actions would lead to a harmful event.
- The subject has to be able to freely choose to act in certain ways. That is, it does not make sense to hold someone responsible for a harmful event if their actions were completely determined by outside forces.

As already noted by Noorman, "computing can complicate the applicability of each of these conditions"[21] and, in our opinion, the emergence of Big Data complicates this matter even more.

Professor Floridi, in "The Fourth Revolution"[22], identifies the moral problem of Big Data with small pattern discovery: it represents a new frontier of innovation and competition, able to put out of

[16] Brown, Geoffrey (1991) *Is there an Ethics of Computing?* Journal of applied philosophy 8.1.: 19-26.

[17] Friedman, Batya, and Peter H. Kahn (1992). H*uman agency and responsible computing: Implications for computer system design.*, Journal of Systems and Software.7-14.

[18] Frey, Carl Benedikt, and Michael A. Osborne (2013). *The future of employment. How susceptible are jobs to computerisation*, Oxford Martin Programme on Technology and Employment.

[19] Allen, Colin, Wendell Wallach, and Iva Smit (2006) *Why machine ethics?* IEEE Intelligent Systems 21.4: 12-17.

[20] Eshleman, Andrew (2014). *Moral responsibility*. The stanford encyclopedia of philosophy.

[21] Noorman, Merel (2012) *Computing and moral responsibility* The Stanford encyclopedia of philosophy.

[22] Floridi, Luciano (2014) *The fourth revolution: How the infosphere is reshaping human reality*. OUP Oxford.

business or to create new companies, producing new important research insights or create problems to a country.

An associated problem is the risk of discovery of these patterns, because "they push the limit of what events or behaviours are predictable, and therefore may be anticipated." Floridi remarks how data must be properly aggregated, correlated, and integrated in order to become interesting[23].

The baseline of the ethics of Big Data is the protection of privacy, freedom and the discretional power to autonomously decide. Although these three areas are conventionally applied to a person, there is a continuous tension between the individual needs and those of a community. For example, the right of people to maintain their flight information secret is contrasted by the community's (in this case, all the EU member states) right to access it for the prevention, detection, investigation and prosecution of terrorist offences and serious crimes (EU Passenger Name Record, PNR, directive 2/12/2015). The same holds true for the interception of communications, authorised by law when necessary in specific cases and for limited purposes (Article 8 of the European Convention of Human Rights).

It's possible to identify several ethical problems that derive from the exploitation of Big Data:

- Privacy
- Tailored reality and the filter bubble
- After death data management
- Algorithm bias
- Privacy vs. growing analysis power
- Purpose limitation
- User digital profile inertia and conformism
- User radicalization and sectarism
- Impact on personal capabilities and freedom
- Equal rights between data owner and data exploiter

Privacy. Privacy is a topic that encompasses most of the others. Its definition is really hard, and usually involves concepts such as liberty, autonomy, secrecy, and solitude.
Seclusion is the keyword that can define privacy, as in the definition of Alan F. Westin as "voluntary withdrawal of a person from the general society through physical [means] in a state of solitude"[24]; this definition can be extended to more modern themes that include data protection and data exposure.
More recently Moor and Tavani[25] defined a model of privacy named Restricted Access Control (RALC), based on the idea that an adequate theory of privacy needs to differentiate the concept of privacy itself from both the justification and the management of privacy. RALC has three components: an account of the concept of privacy, an account of the justification of privacy, an

[23] Floridi, Luciano (2014) *Artificial Agents and Their Moral Nature. The Moral Status of Technical Artefacts*. Springer Netherlands: 185-212.

[24] Westin, Alan F. "Privacy and freedom." *Washington and Lee Law Review* 25.1 (1968): 166.

[25] Tavani, Herman T. "Philosophical theories of privacy: Implications for an adequate online privacy policy." *Metaphilosophy* 38.1 (2007): 1-22.

account of the management of privacy. Privacy itself is divided in the *condition* of privacy and the *right* to privacy, related to the loss of privacy and invasion.

The definition is interesting since it has an operative side more natural to the data protection, and indeed Tavani in the quoted papers uses a data mining example, including data and artificial intelligence algorithms, to clarify RALC.

Privacy vs. growing analytical power. This problem is related to the emerging nature of information as a complex system: when data arising from different contexts are collated, the result is more than the simple sum of the parts. Looking at the data of a career related service such as Linkedin will give you a very controlled image of a person, but when all the comments of that person written on social networks, online newspapers, forums and so on are added, their image will not be under their direct control any longer: for instance, a potential employer might associate political opinions, sexual orientation, religious beliefs and even health related information on that person, so their decision to hire or not will be based on data that are indeed sensible data.

This problem will become prominent in the near future because it will become progressively cheaper and easier to analyse data (even not structured data as a post in a social network or a comment on a restaurant), widening the arena of who is able to exploit the data fusion.

Purpose limitation. Related to the previous point is that it is at present really hard, or, better, almost impossible, to limit the use of your data. You can give the right to publish your comment on a restaurant to a web site, but are you aware that it is possible that company is selling data to an employer that, from those comments, might assess if you are morally in line with his/her company policies?

Privacy is not a single block item; it is important to understand not only the invasion of privacy but also this subtle forms of loss of privacy.

Tailored reality and the filter bubble. When we interact with a server, we surrender a huge amount of information about us; this is the way, for instance, in which an online newspaper learns the kind of news that we like or dislike, and then it uses that information to build a model of our interests to suggest other news and articles that might be interesting for us. The same approach is used by online market places to recommend interesting products.

A problem arises when a system uses these models to filter information, rather than to provide recommendations. This way we might be induced to think that what we see is a complete vision of a specific context – in the example of the newspaper our view of the world – while we are limited by the “understanding” of an underlying algorithm of our goals.

The ethical impacts are multiple: a service can use this filtering approach to hide some pieces of information from us[26], imposing a bias of which we are unaware; our vision of the world might become progressively limited, producing even an effect of echo chamber[27] where there is a progressive reinforcement of a narrow view. In the long term this might generate momentum around a proposal or a point of view.

User digital profile inertia. This issue is related to the topic of tailored reality. In this case, the issue is that a model that infers a user's interests is usually based on their past behaviour and the collection of information given in the past. This way algorithms are not based on the current identity of a person, but on a previous version. For instance, if I am interested in children care in a moment of my life, I will receive news, recommendation, and web pages based on that interest. But if, in the meantime, I lost interest in that topic, the update time of the algorithms on which those recommendations are based might be very slow, and I will continue to receive items based on my previous, and now irrelevant, actions.

An inertial filter bubble, therefore, will impact on my real behaviour: I can be pushed to maintain my old interests, or I cannot discover other opportunities that might be more interesting for me. If a person were aware of the behaviour of recommendation or filtering systems, this would be a minor problem; but this awareness is mostly lacking, so there is a direct impact forcing, as a "strange attractor", to maintain old views.

User radicalization, conformism and sectarianism. This is another issue related to the topic of tailored reality. In this case, the issue is related to opinion formation. When a person holds an opinion, a trace of this interest will be left in data-driven applications such as web newspapers, online bookshops, online forums, social networks. Through a filtering/recommendation algorithm, the information, items, posts, friends and so on will be focused on this opinion. For instance, if I sustain a specific political position, it is more frequent for a social network to suggest I add people with the same position to my friends list. This process is reverberating: through the filtering/recommendation system I will be increasingly in contact with people, opinions and facts that will support my initial position. This is a well-known process typical of group formation; for instance, in "Membership Has Its (Epistemic) Rewards: Need for Closure Effects on In-Group Bias"[28] it is described how a function of in-groups is to define a "social reality" for their members based on a group consensus.

Again, the problem is that, while in an "physical" group this process is apparent, it is hidden from users of Big Data based systems, so a tendency to develop a bias, from a group conformism up to a radicalization of ideological positions, is a sort of data-driven unconscious process. We can even postulate the formation of a sort of technological subconscious that impacts on our personality development and ultimately on our social life. There are too few studies on this theme, that is clearly

[26] Pariser (2011). *The filter bubble – what the Internet is hiding from you.* Penguin Press, New York.

[27] Harris, Lisa, and Paul Harrigan. (2015) *Social Media in Politics: The Ultimate Voter Engagement Tool or Simply an Echo Chamber?*. Journal of Political Marketing 14.3: 251-283.

[28] Shah, James Y., Arie W. Kruglanski, and Erik P. Thompson, (1998). Membership Has Its (Epistemic) *Rewards: Need for Closure Effects on in-Group Bias.* Journal of Personality and Social Psychology 75, no. 2: 383.

underestimated; nevertheless, an interesting experiment conducted by Lev Muchnik[29] is present, “Social Influence Bias: A Randomized Experiment”, that concludes that social influence substantially biases rating dynamics in systems designed to harness collective intelligence, with an influence on the opinion impacting up to 32% of people. This problem is accentuated on criminal behaviours, ranging from teen bullying to terrorism. This theme is studied in particular with reference to social network impact, for instance in “Tweeting Propaganda, Radicalization and Recruitment”[30].

The emerging role of technology in criminal groups is explored in several studies, including “Gangs, Terrorism, and Radicalization”[31], that reports as “YouTube and related websites have eclipsed mainstream media sources—news, television, movies—as the source of new information that was not available to gang members two decades ago.” It is clear, for the authors and for us, that “how this information impacts gangs, extremist groups, and the transfer of radicalized beliefs and images across the globe should be a high priority for future research”.

It is important to realize that the distance between the “physical”, real, world and the Internet is strongly reducing. The study “Examining the Overlap in Internet Harassment and School Bullying”[32] reports that there is strong passage from Internet harassment to behavioural problems at school.

After death data management. We create lots of data and many of them live on Internet or in data-driven companies’ database. What happens the moment we die? Will our heirs inherit our data, too? Is it desirable that heirs can remove selected (or all) data from the digital world? This issue is a strong mixture of legal and technological problems: who owns the data? How to notify a company/government agency the death of a person in a trustworthy communication? How to remove all the data of a person from a database? How to remove all the duplicated data?

Impact on personal capabilities and freedom. Amartya Sen changed the definition of freedom by including the concept of capabilities, the open possibilities, as a factor of it[33]. This ethical issue is related to the digital divide, to the capability of using data and to benefit from giving away one’s own data.

The idea that data diffusion, even in the case of open data, is automatically connected to more freedom is challenged by several authors. An interesting study is “From Open Data to Information Justice”[34]; it refers to a seven-layer model for promoting an effective use of open data by Gurstein[35], that includes general digital-divide items and more data specific ones:

[29] Lev Muchnik, Sinan Aral, and Sean J. Taylor (2013). Social Influence Bias: A Randomized Experiment, Science 341, no. 6146: 647–651

[30] Takeoka Chatfield, Akemi, Christopher G. Reddick, and Uuf Brajawidagda, (2015). *Tweeting Propaganda, Radicalization and Recruitment: Islamic State Supporters Multi-Sided Twitter Networks* in Proceedings of the 16th Annual International Conference on Digital Government Research.

[31] Decker, Scott and Pyrooz, David (2011). Gangs, Terrorism, and Radicalization. Journal of Strategic Security, no. 4: 151.

[32] Ybarra Michele L., Diener-West, Marie and Philip Leaf (2007). Examining the Overlap in Internet Harassment and School Bullying: Implications for School Intervention, Journal of Adolescent Health 41, no. 6: S42–S50.

33 Sen, Amartya (2001) *Development as freedom*. Oxford Paperbacks.
Sen, Amartya (2004) *Rationality and freedom*. Harvard University Press.

[34] Johnson Jeffrey Alan, (2014). *From Open Data to Information Justice*, Ethics and Information Technology 16, no. 4: 263-274.

[35] Gurstein, Michael B. (2011). Open Data: Empowering the Empowered or Effective Data Use for Everyone? First Monday 16, no. 2 http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/viewArticle/3316.

- Interpretation and sense-making skills, including both data analysis knowledge and local knowledge that adds value and relevance.
- Advocacy in order to translate knowledge into concrete benefits.
- Governance that establishes a regime for the other characteristics

Johnson assumes that "in the absence of these conditions diverse individuals are not able to use information to act on or become something that they value" and that "this problem is likely to be exacerbated by the emergence of 'big' data."

Equal rights between data owner and data exploiter. An emerging ethical issue is the unbalance between who generates and who collects and manipulates data. For instance, when I use a wearable bracelet and a running app to monitor my physical conditions and exercise, I produce massive amounts of data: the tracing of runs, my speed, my cardiac measures and so on. These data are used, with my permission, by the hardware/software producer to give me a personalized service, so they rightfully own my data. But, strangely, I do not own them!
To balance this asymmetry, it is strongly ethically desirable that the company that collects and processes my data give me back the data I produced, allowing me to download them and totally or partially deleting them.

It is also desirable to have a common data format, so that, when I switch from a running app to another, I will receive my data in a compatible format that lets me (in accordance to the concept of "capability" as defined by the Nobel Prize economist and philosopher Amartya Sen) to easily analyse all the collected data.

Algorithm bias. Big Data are almost useless without interpretation; this is given by using algorithms. But are they neutral or do they have some form of bias? Algorithms are designed by humans, and the same is true for data selection and, lastly for data presentation, so it is likely that some form of bias is present. In addition, there is a possibility that an error in an algorithm might introduce forms of bias. The most ethically relevant biases are those that have the potential to harm minority groups, not only related to race or gender, but also poverty, joblessness, social marginalization.

A recent paper, "Battling Algorithmic Bias"[36], describes several real-world applications of algorithmic bias. In the USA, a software created by Northpoint is gaining momentum. It is able to determine the likelihood of committing future crimes; one non-profit investigative journalism organization, Pro Publica, challenged the output of this algorithm, asserting that it is strongly biased on race and house position. This algorithm is not just for study. A BBC article, "How maths can get you locked up"[37], reports a law case in Wisconsin, in which one of person involved in a crime was charged to a higher number of years to serve in jail because "the court noted that he had been identified as an "*individual who is at high risk to the community" by something called a Compas*

[36] Kirkpatrick, Keith, (2016) *Battling Algorithmic Bias: How Do We Ensure Algorithms Treat Us Fairly?* Communications of the ACM 59, no. 10: 16-17.

[37] Maybin Simon (2016). *How Maths Can Get You Locked up*, Available at http://www.bbc.com/news/magazine-37658374.

assessment. The acronym - which stands for Correctional Offender Management Profiling for Alternative Sanctions".

Julie Angwin, Pro Publica, attacked this software since it includes questions like: "Your criminal history, and whether anyone in your family has ever been arrested; whether you live in a crime-ridden neighbourhood; if you have friends who are in a gang; what your work history is; your school history. And then some questions about what is called criminal thinking, so if you agree or disagree with statements like 'it's okay for a hungry person to steal'." Letting an algorithm decide for a fate of a person is a really hazardous choice, given that it was developed by human beings, that the elements of the algorithms are opaque and not known to an external person (in this case, to the court, which furthermore most likely do not have the skills to understand its computer science background), and it is based on data that can already be biased in the moment of collection. For instance, if a dataset is characterized by an excess of crime committed by people of a certain race, machine learning algorithms tend to over classify crimes for that race and under classify for the other races.

There are even most subtle forms of bias. A paper by Kirkpatrick cites other examples with discrimination of gender (Google returns less pictures of women CEO than men, or display less payed jobs to women), race (ads of different universities or products are displayed in base of the race and geographical position).

An ethical issue to be deeply explored is our trust in algorithms. Most people think that "machines" are neutral by definition, but we see that this is absolutely not true, and risks for people might be very high.

4. The European agenda on the digital society

The European Union's "digital" opinions have sped up exponentially over the last 5 years. Since the present study is focused on the ethical aspects of Big Data, it is necessary to provide a general perspective of the strategy and the vision of the European authorities on the evolution of the digital world in order to produce a comprehensive point of view.

In September 2012, the European Commission published "**Unleashing the Potential of Cloud Computing in Europe**"[38], a document outlining a plan explaining how Cloud computing works and what benefits it can bring (e.g. the opportunity to store information in a simple standard way) with the aim of enhancing people's knowledge by building and spreading digital confidence. This plan will push member states to "embrace the potential of cloud computing" for the 2014-2020 growth-phase.

Five months later, in February 2013, the European Commission published "**An Open, Safe and Secure Cyberspace**"[39]. This document aimed to describe cyberspace as a 'place' where the behaviour of users should tend to conform to their behaviour offline. To this end, European authorities should pay attention to fundamental rights, democracy and the rule of law that need to be protected in cyberspace. In the same document, the Commission explicitly drew a boundary separating Europe from other countries where citizens are potentially controlled by their government in opposition to the ideal of freedom and respect of civil rights as meant within Europe. The European Union must promote "freedom online" and ensure the "respect of fundamental rights online". This issue could be solved by granting access to anyone, creating a democratic and efficient multi-stakeholders governance, and by sharing responsibility to ensure security. Strategic priorities and actions start from promoting:

- the achievement of cyber-resilience by raising awareness,
- a drastic reduction of cybercrime with a strong and effective legislation, enhancing operational capability to combat cybercrime and improving coordination at an EU level,
- the development of a cyber defence policy and the industrial resources for cyber security, also by promoting a Single Market for cyber security products and fostering R&D investments and innovation,
- the establishment of a coherent international cyberspace policy for the EU
- mainstream cyberspace issues into EU external relations and Common Foreign and Security Policy.

This document concluded by aiming at "protecting and promoting citizen rights, to make the EU's online environment the safest in the world".

[38] European Commission (2012). Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions – *Unleashing the Potential of cloud Computing in Europe*.

[39] European Commission (2013). Joint communication to the European Parliament, the Council, the European Economic and Social Committee and the committee of the Regions – Cybersecurity Strategy of the European Union: An Open, Safe and Secure Cyberspace.

After the launch of Horizon 2020, other initiatives of the EU on the theme of Big Data were included in several documents, such as the guide of ICT-related activities, the European Commission Communication “Towards a thriving data-driven economy” of 2014 or the interest shown in the improvement of connectivity for society.

In particular, the mentioned Communication clarifies the essential features of a data-driven economy, described as an ecosystem of different players interacting in a Digital Single Market. It states that such economy can grow if it is based on trusted and high quality data, enabling infrastructures, sufficient domain experts and a range of application areas where improved big data handling can make a difference.[42]

In 2015, the European Data Protection Supervisor wrote Opinion 7/2015[43], that discussed the important topic of Big Data and opened a discussion on the risks and challenges of Big Data itself, providing insights on the orientation of the General Data Protection Regulation, that was being written at the time, and that was approved and entered in force in May 2016, and will be applied by May 2018.

What emerges from this brief analysis is that the European Union and the whole European landscape appear to be particularly careful of themes such as the digital economy and cyberspace issues, also by practical means such as Horizon2020, which is particularly active on the topics of Big Data, the whole European Digital Economy and its development.

[42] European Commission (2014). Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions - *Towards a thriving data-driven economy.*

[43] 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.

5. General review

As first step of the Study, a revision of the European and International debate on the topic as expressed in literature has been carried out. Almost 100 publications on the topic has been reviewed consulting digital libraries, sources of peer-reviewed articles, studies, journals, technical magazines, newsletters and books. In parallel, an analysis on how the General Data Protection Regulation (Regulation EU 2016/679) adopted in May 2016 can impact on the actual scenario has been produced. Finally, a comprehensive stocktaking of the key stakeholders' positions about the topic has been realized.

5.1 Knowledge base review

At the moment of writing this study, the majority of the literature available on the topic of Big Data comes from the United States, with a minor participation of Asian countries and, so far, a small academic contribution from the European countries, which have expanded much more on the legal side of the matter. The debate ranges from the multiple opportunities that Big Data can bring within the next decades to the concerns about the tangible impact that the massive analytics application can have on people's lives.

5.1.1 Opportunities

Positive aspects of Big Data, and its potential to bring improvement to everyday life in the near future, have been widely discussed. Examples span from health services, to road safety, agriculture, retail, education and climate change mitigation and are based on the direct use/collection of Big Data or inferences based on them.

Simple *collection* of Big Data is discussed in a recent study[44] where health services delivered via Internet are discussed in-depth, especially services aimed at impacting the behaviour of an individual and change their habits to healthier ones, and the authors assess that these services will bloom in the near future improving the general health conditions of the population.

"*Nowcasting*" sits on the boundary between collection and inferences, and refers to the possibility of generating real-time reports and analyses with alerting purposes, as detailed in a study by the Pew Research Center[45] about environmental crises.

Inferences from Big Data generate new knowledge and insights, such as in the case of road safety that was cited[46] within a study also covering other topics, where authors talk about very practical implications that Big Data could have, such as the use of GPS positioning to infer car crashes and

[44] Lim and Thuemmler (2015). *Opportunities and challenges of Internet-based health interventions in the future Internet*. Proceedings of the 12th International Conference on Information Technology – New Generations, 567–573.

[45] Anderson and Rainie (2012*). Big Data: Experts say new forms of information analysis will help people be more nimble and adaptive, but worry over humans' capacity to understand and use these new tools well.* The Future of Internet. Pew Research Center.

[46] Bertot and Choi (2013). *Big Data and e-government: Issues, policies and recommendations*. Proceedings of the 14th Internation Conference on Digital Government Research, 1–10.

potential accidents, and send alert messages to drivers and requests for assistance. Inferences are the core of another application discussed in the same study, that involves the use of weather data and other environmental variables to produce useful outputs to determine the amount of weather insurance farmers need and pay-outs that they would need to make. The use of insights produced by the analyses of Big Data has become more and more widespread (and controversial) in the retail field, with one famous case[47] being that reported by the cofounder of the cosmetic company 100% Pure, who, after the use of a predictive algorithm by Freshplum, managed to increase the company's online sales by 13% in 3 months, highlighting the potential of Big Data for generating economic growth. This case became famous since it sparked a wider discussion on the use of similar algorithms by much larger companies, such as Amazon.

Collection of Big Data, nowcasting and inferences, as a source of both economic growth and generation of value, have been widely discussed from a business point of view[48,49], which is often that Big Data is a factor of production. This point of view stems from the consideration that a pragmatic use of Big Data boosts the productivity of companies and, therefore, growth at multiple levels, from the single company itself to communities and society in general. Big Data creates values in many ways, such as by making data more accessible and easier to share among compartments (transparency); by enabling experimentation to discover needs, expose variability and improve performances; by segmenting populations to customize actions; by replacing and supporting human decision making with automated algorithms (portrayed as positive in the author's intentions); by introducing innovative new business models, products and services. For example, the authors of this study "estimate that a retailer embracing Big Data has the potential to increase its operating margin by more than 60 percent".

This new paradigm that follows the use of Big Data in an economic context will potentially generate new employment opportunities that will be characterized by uncertainties over the working ethics and conditions of this new category of workers, now globally defined as data scientists. A major worry has come from the consideration that data scientists tend to work on data about subjects they do not know, and have never been in touch with, and are often estranged by the end product of their work (i.e. application of analyses). Both are risk factors in a perspective of worker alienation. A recent study[50], nevertheless, has tackled this matter by stating that digital alienation is very different from alienation in the Marxist sense, since Big Data workers are involved throughout the data analyses and use process and cover a creative role, where autonomous decisions and new approaches are constantly required.

Looking at Big Data, and at the "Digital Revolution" in general, from another point of view, some authors[51] point out that the effects of digital innovations may not be as strong as the Industrial Revolution ones. The idea is that the growth enhancing effect of the digitalization in our societies has

[47] Tanner (2014). *Different customers, different prices, thanks to Big Data*. Forbes 26 March 2014. Available at http://www.forbes.com/sites/adamtanner/2014/03/26/different-customers-different-prices-thanks-to-big-data/.

[48] Manyika et al (2011). *Big Data: The next frontier for innovation, competition and productivity*. McKinsey Global Institute.

[49] Villars et al (2011). *Big data: What it is and why you should care.* IDC White Paper.

[50] Dainow (2015). *Digital alienation as the foundation of online privacy concerns.* SIGCAS Computer & Society, 45: 109–117.

[51] Gordon (2016). *Computers and the Internet from the Mainframe to Facebook.* In *The Rise and Fall of American Growth: The U.S. Standard of Living Since The Civil War.* Princeton University Press, Princeton and Oxford.

already come, and the related economic growth has stopped in many western countries. This is surely a controversial opinion that should be taken into account.

Big Data could have (and has already had, in some instances) broader influences at a governmental level, thus positively affecting several aspects of the citizens' lives. Such a large amount of information might be efficiently disseminated by public institutions, and freely accessed by a growing number of citizens. This could be realised through the application of transparency and open government policies, such as Open Data[52]. National and international (e.g. European) laws, acts and regulations, indeed, adopt such policies as important aspects for the development of more democratic and participative societies[53, 54, 55, 56]. Additionally, there have been discussions [48] about how correct use of Big Data in sectors such as public health care and government administration in general could improve efficiency, with an estimated 100 billion € saving for the European developed economies.

A further example is connected to the use of Big Data in health care and disease prevention, particularly the detection of early outbreaks. It has been shown that it is possible to generate alerts on the appearance of disease that could potentially progress into epidemics by using predictive algorithms based on data other than health. For instance, keywords used in Google queries were found to be strongly correlated to the appearance of the 2014 Ebola outbreak in West Africa[57], and other authors have proposed to use a similar approach, dubbed Digital Disease Detection, in the future[58].

A clearly integrated approach to the collection of Big Data, the use of nowcasting and predictions is that reported by an American study[59] on education and how student data can be used to monitor their performance. After the production of predictive models on their behaviour, they can be used to generate early warnings of potential drop outs, who would subsequently be approached with recovery interventions. Authors report positive feedback between research and practice, with discoveries quickly taken up by practice, thus reducing the number of students dropping out of schools and improving general scholastic performance.

Integrated approaches are also reported in literature in the case of websites that, through self-provisioning of data by registered users, use algorithms to provide matching of various sorts. One important example is online dating, which can include the use of algorithms to help users to find potential partners. This phenomenon has become more and more accepted in the American society,

[52] Buhr and Kleiner (2012). *European Open Data policy: Challenges and opportunities.* Policy Advice and Political Consulting, 5(3):141–146.

[53] An Act to amend section 3 of the Administrative Procedure Act, chapter 324, of the Act of June 11, 1946 (60 Stat. 238), to clarify and protect the right of the public to information, and for other purposes (FOIA). United State Congress, 6 July 1967.

[54] Italian Legislative Decree 97 of 25 May 2016. Revision and simplification of the dispositions on prevention of corruption, publicity and transparency, correcting law number 190, 6 November 2012, and legislative decree 33, 14 March 2013, following art. 7 law 124, 7 August 2015 on reorganization of Public Administrations.

[55] Directive 2013/37/EU of the European Parliament and of the Council on the use of the public sector information.

[56] Regulation 1049/2001 of the European Parliament and the Council regarding public access to European Parliament, Council and Commission documents.

[57] Alicino et al (2015). *Assessing Ebola-related web search behaviour: insights and implications from an analytical study of Google Trends-based query volumes.* Infectious Diseases of Poverty, 4: 2–13.

[58] Vayena et al (2015). *Ethical challenges of Big Data in public health.* PLOS Computational Biology, 11: DOI:10.1371/journal.pcbi.1003904.

[59] Siemens and Baker (2016). *Educational data mining and learning analytics: towards communication and collaboration.* Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, 252–254.

and represents a side shoot of Big Data exploitation[60]. A further field in which Big Data might represent a potential source of value is political campaigns in a USA context, where citizens' personal data and their previous donations to parties are at the base of an efficient management of campaigns themselves[61].

5.1.2 Challenges

Some of the topics previously discussed under a positive light, are treated differently by other authors, that focus more closely on the potential ethical and social issues that might derive from the collection and use of Big Data[62].

The field of retail, which clearly benefits from Big Data as discussed above, and whose growth has positive repercussions on a country's economy, also generated some concerns. Details are provided in an Irish study[50] that explains how excessive use of algorithms on customer's details might lead to online business models where commercial surveillance is being used to commodify private life.

In some authors' opinions[63], the great benefits that could be derived from Big Data in health care and the healthcare field more in general might be offset by the impossibility to respect some of the ethical boundaries that are required today when dealing with sensitive data. Authors point out that consent to data collection and treatment, for example, might not be feasible to acquire before each processing instance when dealing with tens of thousands of individual's data. In addition, the epistemological shift that Big Data is bringing about might also prevent citizens from providing consent to data uses that are not foreseeable at the present time.

Also, education applications of Big Data are not exempt from potential issues, as clearly stated in a study[64] that aims to provide a framework for the development of the so-called "big data regulation" in Europe by looking at the lessons learned in the USA. This study voices sharp concerns about the fact that constantly collecting students' data might cause them to feel under surveillance at all times. This might lead, as known, to reduced creativity and higher levels of stress. Assessment of students only on the basis of collected data (and inferred results) might also lead to precocious labelling as under-performing, with related potential discrimination. Another study[65] on education carried out through MOOCs (Massive Open Online Courses), instead, explores the general issue of privacy acquired by means of anonymization of personal data, suggesting that a better approach to privacy protection should involve a stricter data access policy rather than newer and more complicated anonymization algorithms that can be more or less easily breached.

[60] Smith and Anderson (2016). *5 facts about online dating.* Pew Research Center. Available at http://www.pewresearch.org/fact-tank/2016/02/29/5-facts-about-online-dating/

[61] Nickerson and Rogers (2014). *Political campaigns and Big Data.* The Journal of Economic Perspectives, 28(2): 51–73.

[62] Boyd and Crawford (2012). *Critical questions for Big Data.* Information, Communication & Society, 15(5): 662–679.

[63] Mittelstadt and Floridi (2015). *The ethics of Big Data: current and foreseeable issues in biomedical context.* Science and Engineering Ethics, 22(2): 303–341.

[64] Carmel (2016). *Regulating "Big Data education" in Europe: lessons learned from the US.* Internet Policy Review, 5(1): DOI: 10.14763/2016.1.402.

[65] Daries at al (2014). *Privacy, anonymity and Big Data in the social sciences.* Communication of the ACM, 57(9): 56–63.

Other general issues have emerged from the analyses of the available knowledge base as described by articles and studies.

The topic of the **digital divide**, as defined by the OECD[66], has not yet been fully discussed by the available literature. A recent study[67] highlighted potential inclusion concerns, with the possibility for some individuals or portions of a society to not be considered during the process of defining new policies based on a data driven approach. In particular, these concerns revolve around the worries that, if we promote a new policy which was built on data collected via sensors, social media, etc., then the risk is that this policy will only account for the needs of people that have access to these technological means. Quantification of the digital divide by means of appropriately defined indices14’[68] has been attempted and has proven useful in the monitoring of the effects of policies aimed at bridging it, which can be found in literature in a North-American14, a European[69] and an Asian context[70].

A study[71] draws attention to how web services providers (such as Facebook and Google) offer Single Sign On services, in order to grant access to third party websites. This login type may cause a decrease of a users’ control and awareness of data types provided to these companies. This is likely due to the used consent dialogs, which lack clarity and efficiency in showing and explaining what information will be transferred to the third parties, who will store and potentially reuse this data. This alleged negative effect may also be fostered by a user’s excess of trust in some companies, similar to what happens with the use of mobile phones and apps, as reported by a German study[72].

Another ethical issue related to the use and analysis of Big Data is explored in a study of the American Economic Association[73]. The intent is to highlight the limits of common anonymization techniques[74]: the substitution of a user’s personal information in a dataset is not enough to guarantee privacy. In fact, simple joins between these anonymous datasets and other datasets such as web search history, and personal data, allow de-anonymization in a relatively short period of time. This may cause privacy and security issues, circumventing security practices.

[66] https://stats.oecd.org/glossary/detail.esp?ID=4719.

[67] Poel et al (2015). *Data for policy: A study of Big Data and other innovative data-driven approaches for evidence-informed policymaking.* Available at http://www.data4policy.eu/.

[68] Coria et al (2013). *Delta score: A novel and simplified measurement for digital divide of cities.* The Proceedings of the 14th Annual International Conference on Digital Government Research.

[69] Negreiro (2015). *Bridging the digital divide in EU*. European Parliamentary Research Service – Member’s Research Service

[70] Sung (2015). *A study on the effect of smartphones on the digital divide*. Proceedings of the 16th Annual International Conference on Digital Government Research.

[71] Bauer et al (2013). *A comparison of users’ perceptions of and willingness to use Google, Facebook, and Google+ Single-Sign-On functionality.* Proceedings of the 2013 ACM Workshop on Digital Identity Management.

[72] Benenson et al (2013). *Android and iOS users’ differences concerning security and privacy.* CHI 2013: Changing perspectives. Extended abstracts, 817–822.

[73] Heffetz and Ligett (2014). *Privacy and data-based research.* The Journal of Economic Perspectives, 28(2): 75–98.

[74] Machanavajjhala and Reiter (2012). *Big privacy: Protecting confidentiality and Big Data*. XRDS, 19(1): 20–23.

In March 2013, the Commission launched a **Grand Coalition for Digital Jobs**[40], a project developed over three years, and is now launching a multi-stakeholder partnership, to “tackle the lack of digital skills in Europe and the thousands of unfilled ICT-related vacancies across all industry sectors”. In October 2013, the official opinion of the European Council stated the need for a strong digital economy for European growth and competitiveness, a field which should be promoted by investing in it. Investing in the digital economy, as a matter of fact, would push innovation and better development of citizens’ digital skills, fostering the creation of a Digital Single Market, for which the Commission is setting up ICT standardization priorities. This initiative aims at re-energising the standard-setting system in Europe as a step towards industrial global leadership and digital innovation.

Moreover, the **Digital Single Market** strategy has the purpose of allowing better access for consumers and businesses to online goods and services across Europe. This will remove the key differences between online and offline worlds, thus breaking down existing barriers to cross-border online activity. These strategies should act synergistically, building the necessary confidence to enjoy all the different aspects of the digital world safely. All the parts that contribute to define the digital experience can be considered to broadly fall under the all-encompassing topic of Big Data, to which the world has opened its boundaries only recently. Several activities have been activated over the last two years on very different fronts, and a plethora of documents have been produced by governments and citizens to understand how Big Data works and how it can be best exploited.

The European Commission commitment on this topic led to the constitution of ISA (Interoperability solutions for public administrations, businesses and citizens) and subsequently to the Joinup platform, a community for exchanging information, experiences and best practices around open source solutions for use in public administrations. In 2015, Joinup published “**eGovernment in the European Union**”[41]. This document is pivotal in order to understand the path undertaken by the EU regarding Big Data and its use. One of the points that emerge from a first analysis of the ideas brought up and discussed in this document is that the European Information Society and e-Government there described appear a bit too far-fetched, given present-day ICT infrastructures available throughout the continent’s territory.

On a more pragmatic level, the establishment of the **Horizon 2020** funding programme gives the European Union the opportunity to foster the development of new sectors of interest, namely the ethical exploitation of Big Data, where innovation is most needed. Big Data is the central topic of the development of digital economy in a perspective of the creation of new sectors of interest to generate new jobs. To better clarify how the topic of Big Data is at the core of Horizon 2020, it should suffice to cite what is present on the Commission’s (who is responsible for the implementation of this funding scheme) web page about the Digital Single Market: “Data has become a key asset for the economy and our societies similar to the classic categories of human and financial resources. Whether it is geographical information, statistics, weather data, research data, transport data, energy consumption data, or health data, the need to make sense of "Big Data" is leading to innovations in technology, development of new tools and new skills”.

[40] https://ec.europa.eu/digital-single-market/en//digital-skills-jobs-coalition

[41] Joinup (2015) *eGovernment in the European Union*.

5.2 Legal Framework

5.2.1 Introduction

It is common opinion that the use of big data represents a fundamental step for the European economy. However, the use of big data poses significant legal problems, primarily if focused on the angle of perspective of data protection.

Several reports of the Community institutions argue that the existing legal framework – based, in particular, on Directive n. 46/95/EC and, today, on the General Regulation on the Protection of Personal Data - already offer adequate protection against any infringement of the fundamental rights of the citizens.

However, as remarked below, a comprehensive and global strategy, taking into account the concentrations of power in private entities that collect and process personal data, as well as appropriate forms of protection that take into account the changed perspective of privacy, is required. It is undeniable fact that the right to privacy arose as a right to be let alone and, thus, as a right to exclude others from personal facts (*ius excludendi alios*) following a proprietary schema that has its roots in Roman law. However, over the years, the protection has shifted from a right to exclude others to a right to the control of personal data. We are now facing a third phase, which answers the question "who am I?" and the answer does not depend on the data subject but on the patterns selected by third parties that create an analytical profile and lead to rethinking the right to identity.

The new frontier of data protection involves not exclusively only personal data, but, more broadly, data. The passage is from data that is able to identify specific physical persons to data that can detect specific behaviours, consumptions, health data, etc. of groups rather than of individuals.

The collection and aggregation of mass big data (even in the form of the so-called bulk data), as mentioned below, is not subject to the application of data protection regulations. Originally, privacy rules were conceived to protect intrusion to a person’s personal life and to avoid forms of discrimination based on the information collected. At the present moment, structured information of Big Data and quantitative analysis allow new insights that are able to stigmatize commercial choices and other personal information on groups. As the groups get smaller (identified by geographical, age, sex, etc. settings) forms of discrimination are more likely.

In 2014, the so-called Podesta Report similarly stated “that big data analytics have the potential to eclipse longstanding civil rights protections in how personal information is used in housing, credit, employment, health, education, and the marketplace”[75].

Thus, innovative ways of rethinking the protection of citizens are necessary as the legal framework, even if theoretically applicable to the new scenario, does not seem to offer adequate and full protection.

[75] Executive. Office of the President (2014). *Big Data: Seizing Opportunities, Preserving Values*, Available at: http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_5.1.14_final_print.pdf .

5.2.2 Territorial Scope of the GDPR

One of the main subjects introduced by the GDPR concerns the territorial scope. In fact, the application of the GDPR is not limited to the companies which are established in the EU territory (regardless of whether the processing takes place in the EU or not), but covers the data processing of data subjects who are in the EU by a controller or processor not established in the EU.

However, in this latter case, pursuant to article 3, par. 2 of the GDPR, the application of the GDPR may occur in two cases where the processing of personal data is related to:

a) the offering of goods or services, irrespective of whether a payment of the data subject is required, to such data subjects in the Union; or
b) the monitoring of their behaviour as far as their behaviour takes place within the Union.

The provision seems to be significant, since it allows to extend the new rules to all Internet service providers who are not established within the EU. Moreover, the payment of the service is irrelevant, as well as the place where, physically, the processing of personal data is made.

Therefore, the GDPR is applicable to large data aggregators (TLC companies as well as the so-called over-the-top IT companies), regardless of geographical or physical connections.

5.2.3 Definition of Big Data, personal data and anonymous data

European regulations do not provide a binding definition of Big Data. According to the Opinion 3/2013 of the European Working Group on data protection (Article 29 Group), the notion of Big Data may be the following:

“Big Data is a term that refers to the enormous increase in access to and automated use of information: It refers to the gigantic amounts of digital data controlled by companies, authorities and other large organisations which are subjected to extensive analysis based on the use of algorithms. Big Data may be used to identify general trends and correlations, but it can also be used such that it affects individuals directly”.

However, this definition, even if useful as a starting point, mostly focuses on the volume of the data, not accurately considering the reuse of the personal data and its secondary value.

Regulation no. 2016/679 defines (article 4, par. 1) personal data as "any information relating to an identified or identifiable natural person (‘data subject’); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person”. This data is, for instance: name, address, sex, occupation, date of birth, telephone number, e-mail address, town, country, license plates, username and password.

Among personal data, a specific category is sensitive data, which is subjected to specific legal rules and which includes personal information regarding racial or ethnic origins, political and philosophical opinions, religious beliefs, sexual behaviours and preferences and health data.

Anonymous information is not covered by the EU Regulation, but according to Whereas n. 26 of the GDPR, “The principles of data protection should apply to any information concerning an identified or identifiable natural person. Personal data which has undergone pseudonymisation, which could be attributed to a natural person by the use of additional information should be considered to be information on an identifiable natural person. To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly.

To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments. The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable. This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes”.

Furthermore, pursuant to article 4, n. 5) “‘pseudonymisation’ means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person”.

It is a crucial point as personal data, once anonymized (or pseudo-anonymized), may be freely processed, without any prior authorization by the data subject. On the other hand, one of the main issues with Big Data concerns the possibility of the re-identification of the data subject. It may occur in two different ways, at least: using technologies of de-anonymization which allow to be traced back to the original personal data and also through multiple and specific sets which may allow to identify specific physical persons or small groups.

5.2.4 Personal Data Processing Steps

The processing of personal data is defined, in a broad and all-encompassing way, by article 4, paragraph (2) as “any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction”.

The necessary processing for Big Data, in general, includes several processing activities of personal data and in particular:

a) data collection
b) data storage
c) data aggregation
d) analysis of the data and use of the analysis results

Each of these activities is subject to specific rules, which will be detailed below.

5.2.5 Data Processing Principles

Article 5 of the GDPR sets out the principles that must be followed in processing data. Many of these principles directly affect the processing of Big Data, and even before, the methods of collection and data retention.

The principles are the following ones:

a) lawfulness, fairness and transparency: transparency is one of the key points in relation to Big Data. Users are often not fully informed and in the position of properly understanding the privacy policy of the services (e.g. app) which collect their data. The data subject must be, among other things, in a position to easily access the information related to his data, to contact the data protection officer and to be aware of the methods and purposes of the processing. They must also be able to exercise the rights granted to them by the GDPR.
b) Purpose limitation: the person who collects the data needs to inform the data subject of the purposes for which the data is collected. Subsequently, personal data may be processed only for those purposes and not be used for different purposes.
c) Data minimization: The data must be relevant and limited to the purposes for which they were collected. Therefore, personal information cannot be collected if it is not closely related to the purposes of collection or, rather, only personal data which is necessary for those purposes can be collected. Thus, the amount of personal data processed must be limited to the minimal amount possible.
d) Accuracy and updating: the data must be constantly updated and rectified in case of request by the person concerned. In relation to Big Data, among these rights, the right of cancellation or erasure is specifically important.
e) Storage limitation: data can only be kept for the time necessary for processing and later destroyed. Personal data can also be stored for longer time periods, insofar as the personal data will be processed solely for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes.
f) Integrity and confidentiality: the data controller must ensure adequate security of personal data by appropriate technical and organizational measures, including protection against unauthorized or unlawful processing and against the loss, destruction or accidental damage.

5.2.6 Privacy Policy and Transparency

Before collecting personal data, the data controller is expected to provide the data subject with a complete privacy policy. The key information which may be included in the privacy policy is:
The first step is data collection. Personal data may be collected in several ways and for several reasons: to process a registration on a website or a service; to comply with legal, administrative or contractual obligations; to respond to users' requests for marketing or advertising activities; for business and market analysis; to subscribe to mailing lists and newsletters; to participate in discussion forums or to publish reviews on specialized websites; etc.

First of all, data may be necessary in order to provide the service (e.g. through completion of forms) or in order to give process to an agreement. In this case, the data minimization principle must be met, in the sense that the provider of the service is not entitled to require personal information which is not strictly needed for the service itself. For example, a social network platform (such as Facebook or Instagram) is not entitled to collect medical or political information. However, sometimes this information is voluntarily disclosed by users (e.g. through "likes" on Facebook or hashtags on Instagram) and stored and automatically collected by the owners of these platforms.

A different case is when personal data is collected unbeknownst to the data subject such as in the case of cookies, web tracking or geolocation data. In this case, the privacy policy should contain information regarding this aspect. The data subject may also submit personal data even if not expressly required (e.g. personal data published on social network platforms).

Furthermore, there are many other ways of collecting personal data: for example, from the participation in contests, sweepstakes and other commercial promotions; through market researches; through users' participation in discussion forums or writing comments and reviews on websites.

In any case, the data controller is due to cover all the relevant information concerning how all the information, including both personal data and aggregate information, will be used.

5.2.7 Data Processing Purposes

As already mentioned, the data collected can be used only for the purposes included in the privacy policy provided to the data subject.

An exception is included in article 89 of the GDPR, concerning statistical purposes. In this case, "Processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes, shall be subject to appropriate safeguards, in accordance with this Regulation, for the rights and freedoms of the data subject. Those safeguards shall ensure that technical and organisational measures are in place in particular in order to ensure respect for the principle of data minimisation. Those measures may include pseudonymisation provided that those purposes can be fulfilled in that manner. Where those purposes can be fulfilled by further processing which does not permit or no longer permits the identification of data subjects, those purposes shall be fulfilled in that manner”.

The definition held by article 89 does not include the use of data for commercial purposes as it is undoubtedly limited to non-commercial applications.

Therefore, Big Data collected for commercial purposes, which represent one of the main uses for profiling users' activities, falls outside the boundaries of the definition of statistical purposes. Nevertheless, anonymised and aggregated data can be processed for the purpose of identifying the commercial behaviour of specific categories of consumers, whose consumption habits can be divided for multiple criteria (e.g. location, age, gender, education level, etc.).

In this regard, however, we must distinguish two hypotheses:

a) the data controller collects data and performs data processing anonymization. In this case, since the subject who collected the data is the same that anonymizes them, then special precautions should be taken (e.g. encryption that does not allow to return to the original data, etc.). In other words, process of anonymization, as long as operated by the data controller, who will also use the data, should be mentioned in the privacy policy and also the uses of Big Data should be authorized by the data subjects or by the national Data Protection Authority;
b) the subject who anonymizes personal data transfers it, once anonymized, to a third party and that third party uses the anonymized data. In this case, there are no particular legal issues, as far as the subject who uses the anonymous data is not in a position to trace the identity of the subjects.

5.2.8 Privacy by design and privacy by default

One of the most interesting changes of the GDPR is the formalization of the privacy by design and of the privacy by default. The concepts were not included in the EU Data Protection Directive (Directive n. 96/45/CE), and the Directive only contained the obligation for data controllers to implement technical and organizational measures in order to fully protect personal data against unlawful conduct. Similarly, member States’ regulations did not contain any specific rules on these issues, even if some Data Protection Authorities (e.g. UK's ICO) have already issued specific guidelines for implementing such measures by default or by design.

According to Whereas n. 78 of the GDPR, “In order to be able to demonstrate compliance with this Regulation, the controller should adopt internal policies and implement measures which meet in particular the principles of data protection by design and data protection by default. Such measures could consist, inter alia, of minimising the processing of personal data, pseudonymising personal data as soon as possible, transparency with regard to the functions and processing of personal data, enabling the data subject to monitor the data processing, enabling the controller to create and improve security features. When developing, designing, selecting and using applications, services and products that are based on the processing of personal data or process personal data to fulfil their task, producers of the products, services and applications should be encouraged to take into account the right to data protection when developing and designing such products, services and applications and, with due regard to the state of the art, to make sure that controllers and processors are able to fulfil their data protection obligations”.

Pursuant to article 25, paragraph 2 of the GDPR “The controller shall implement appropriate technical and organisational measures for ensuring that, by default, only personal data which is necessary for each specific purpose of the processing are processed. That obligation applies to the amount of personal data collected, the extent of its processing, the period of its storage and its accessibility”.

Pressuring data controllers to implement privacy by design and privacy by default measures, granting them incentives in order to facilitate the adoption of these practises, should be strongly encouraged. In fact, through this approach companies would be compliant with legal regulation by automated settings, limiting expenses and avoiding legal consequences.

Though, these practises should be agreed or validated by Data Protection Authorities, preferably by European authorities, in order to achieve the harmonization effort which leads the GDPR.

5.2.9 The (Legal) Paradox of Big Data

Big Data and their use configure at least two paradoxes.

On the one hand, Big Data ensure maximum transparency, but, at the same time, there is no adequate transparency regarding the use of Big Data.

As stated above, the lack of clarity on the possibility of returning to the identity of the data subject after the process of anonymization of the data is one of the main issues. Business companies that are willing to use Big Data should adopt transparent procedures and ensure that these procedures are easily accessible and knowable by the public.

In this sense, the adoption of tools set of privacy by design or privacy by default, with settings also agreed with the Data Protection Authorities, seems to be a way forward, with practices that should be favoured with legislative incentives (e.g. by preventing sanctions on those who adopt certain practices shared by the regulatory authorities).

On the other hand, transparency is a crucial issue because it affects the ability of a data subject to allow disclosure of their information. It is shown that social network platforms have allowed (for example, in the case of the so called Arab Spring) the aggregation of people and the circulation of information. At the same time, however, they allowed totalitarian regimes to monitor the participants in these initiatives, with obvious repercussions on the civil rights of citizens.

So, an obvious corollary of transparency is that people are not frightened by the use of Big Data and not limited by a chilling effect in their freedom of expression until they are in a position to control access to these data by third parties.

5.3 Stakeholders’ opinions landscape

The ethical dimension of Big Data is becoming more and more central in EU debate. There are a variety of papers, opinions, reports, strategies, and statements by Authorities, associations, Agencies and of course the European Data Protection Supervisor (EDPS) that have given a strong contribution to this debate. Looking at this massive documentation, it is clear that the main stakeholders and the EDPS *in primis*, are looking toward concrete solutions to make the most of the Big Data value without sacrificing human fundamental rights.

The EDPS pointed out in many of its official positions that the right to privacy and the right to protection of personal data play a fundamental role in order to fully respect **human dignity**, which is an inviolable right of human beings, recognised in the first article of the EU Charter of fundamental rights. In the context of digital economy, taking into account the crucial importance of personal data, violations of dignity may take the form of **objectification**, where a person is treated as a tool serving someone else's purposes. Therefore, privacy is an integral part of human dignity.

One of the central ideas that arises from the documents collected is the need to **overcome the conceptual conflict** between privacy and Big Data and between privacy and innovation. Instead, as pointed out in the report of the European Union Agency for Network And Information Security (ENISA)[76], they can coexist and prosper together. And this is strongly reaffirmed despite some arguments that, in the era of Big Data, thinking about concepts such as data minimization and purpose limitation is not realistic.

This is something also strongly supported by the EDPS for whom there is no dichotomy between ethics and innovation but it is pivotal to identify the ways to include ethical assessment in the development of innovations.

Starting from the idea of the **Big Data value chain** which includes the consecutive phases of data collection, data analysis, data curation, data storage and data usage, many stakeholders, including the EDPS, affirm that the data protection principles should be taken into consideration at a very early stage. This is the principle of **privacy by design**, that now is officially included in the new EU Regulation 2016/679[77] According to that regulation, data controllers shall already implement privacy measures and privacy enhancing technologies (PETs) at the time of the determination of the means for processing and at the time of the processing itself.

Many *privacy by design* strategies have been already identified, in particular by ENISA[76], such as data minimization, hiding personal data and their interrelations, separate processing of personal data, choosing the highest level of aggregation once personal data is processed, informing the data subject in the most transparent way, monitoring by specific agencies, privacy policy compatible with legal requirements, compliance with any applicable legal requirements.

[76] ENISA (2015) *Privacy by design in Big Data. An overview of privacy enhancing technologies in the era of Big Data analytics*

[77] Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), (25)

These strategies need specific privacy enhancing technologies (**PETs**) to be adequately implemented. Some of these technologies are already applied in the Big Data industry but an adequate **investment in this sector is still missing**, as attested by the very low number of patents for 'PETs' compared to those granted for data analytics. Among these technologies, there is anonymization, encryption, security and accountability control, transparency and access, consent ownership and control mechanisms.

In particular, one of the main pillars on which the coexistence of Big Data exploitation and data protection can be founded, is **data subject control**, because control brings **transparency** and **trust** between users and digital service providers. As pointed out by the EDPS, *building trust in the online environment is key to economic development. The lack of confidence will continue to slow down the development of innovative uses of new technologies, to act as an obstacle to economic growth*[78]. For the sake of transparency and trust building, the EDPS invites the digital companies to "open their black boxes", to disclose the secret algorithms behind Big Data analytics where there is an effect (direct or indirect) on the individual.

This trusting relationship can benefit both. As underlined in the Impact Assessment of the GDPR, *"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 digitisation of its services"*[78].

Stakeholders agree on the idea that in order to effectively exercise control, **traditional consent** models are rather **insufficient** and **obsolete**. In fact, in the era of Big Data and massive secondary use of data, wide-ranging consent from the users is dangerous and useless in terms of transparency. Instead, consent should be granular enough to cover all the different processing and purposes of processing and reuse of personal data.

One of the most promising measures to support and enhance data subject control is the **data portability**. It is strongly advocated by the EDPS, that in the Opinion 7/2015[79] stresses the importance of guaranteeing citizens the right to access and correct one's personal data, already mentioned in the Charter of Fundamental Rights of the European Union. Very few users are aware about the right to access their data and even fewer exercise this right.

Data portability expands the idea of simple control and nurtures the concept of allowing people to share the benefits (also economic) of data and help to reduce the economic imbalance between Big Data industry on one hand and individuals on the other.

[78] Commission Staff Working Paper Impact Assessment Accompanying the document GDPR and Directive of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data by competent authorities for the purposes of prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and the free movement of such data.

[79] 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*.

The EDPS states that data portability can represent one of the most powerful ways to enhance consumers' awareness and control, since it *allows users to transfer between online services in a similar way that users of telephone services may change providers but keep their telephone numbers. In addition, data portability would allow users to give their data to third parties offering different value-added services* [80]. Users are not conscious about the value of their personal data, which is the real fuel of many Internet services. As the EDPS points out[80], '*Free' online services are 'paid for' using personal data which has been valued in total at over EUR 300 billion and has been forecast to treble by 2020*. In this data driven economy, instead of consumer, we should rather talk about **prosumers**, to refer to users that at the same time produce and consume information.

EDPS has often stated that personal data should be considered like other precious resources, like oil, that are traded by equally well informed parties to the transaction. The market of personal information is neither transparent nor fair, because between the personal data suppliers and Big Data industry the relationship is unfair, and customers are not compensated for their personal information which is traded.

Following this perspective, the EDPS argues that data portability can foster a more competitive market environment, by allowing customers to switch providers more easily, perhaps choosing the one that shows more respect toward personal data and invests in technical measures and internal procedures to fully comply with the European legislation.

If the data protection policy becomes a strategical economic asset, a virtuous circle can be triggered, where companies invest money and human resources to find the best ways to guarantee the privacy of their customers.

Another possible technology enabling user's empowerment is the **personal data store** (PDS). This is a technology that *enables individuals to gather, store, update, correct, analyse, and/or share personal data. Of particular importance is the ability to grant and withdraw consent to third parties for access to data about oneself.* [81]

The personal data store (or personal data space) is a concept, framework, and architectural implementation that shifts data acquisition and control from a distributed data model to a **user-centric model**.

Using PDS technology, the users could have full control over their personal data and take informed choices about what personal details can be stored, analysed and reused.
The key element to make these solutions be developed is the strong commitment of the Big Data industry.
The common position expressed by many stakeholders, and strongly supported by the EDPS is working towards **accountable** companies, reducing at the same time the bureaucracy in data

[80] European Data Protection Supervisor (2014) *Preliminary Opinion on Privacy and competitiveness in the age of Big Data: The interplay between data protection, competition law and consumer protection in the Digital Economy.*

[81] *Study on Personal Data Stores*, Cambridge University, 2015.

protection law and making personal data protection and privacy respect an economic asset for digital players.

Instead of red tape, the EDPS suggests giving more room for more responsible initiatives by businesses, supported by guidance from data protection authorities. For example, *codes of conduct, audits, certification and a new generation of contractual clauses and binding corporate rules can help build a robust trust in the digital market*[82].

But what does being an accountable company in the data driven business mean exactly?
Accountability requires first of all putting in place internal policies and control systems that ensure compliance with the EU law but this does not seem to be enough among the main stakeholders, if the concept of *accountability over mechanical compliance with the letter of the law*[83] is stressed in the EDPS Strategy. The approach is more focused on finding pragmatic solutions to guarantee the application of an ethical approach to data processing, avoiding making companies get stuck in bureaucracy.

This means finding smart and dynamic solutions to guarantee the compliance with the main principles of EU law such as “data minimization”, “purpose limitation”, “data quality”, “fair and transparent data processing”, “data protection by default and by design”, “storage limitation” and “integrity and confidentiality”.

There are other stakeholders that express close positions to this approach, like the **European Privacy Association**, that goes even further and proposes a new way to see data protection for digital companies, as part of Corporate Social Responsibility (CSR), so a way to make a positive contribution to society while doing business[84]. According to this position, personal data protection and privacy protection should no longer be considered as mere legal compliance obligations, instead as they should be seen as assets that can help companies to responsibly further their economic targets. But this rather ambitious scenario can be realized only with strong change of business models and a mind-set transformation in the Big Data industry, which can bring effective investments in PETs and stronger internal policies and procedures for the sake of a sustainable business.

To sum up, it is clear that the challenges posed by the Big Data industries can be tackled only by adopting different measures, from technological solutions to new business models and enforcement cooperation. And, as the EDPS pointed out, only with an interdependent ecosystem of legislators, corporations, IT developers and individuals is it possible to come up with a data driven industry that fully respects human dignity.

[82] European Data Protection Supervisor, Opinion 4/2015 Towards a new digital ethics Data, dignity and technology.
[83] See Action 4 of the EDPS Strategy 2015-2020, developing an ethical dimension to data protection.
[84] European Privacy Association, EPA and Corporate Social Responsibility.

6. A Life of Big Data

Nowadays people disseminate personal data in the digital world even before their birth and keep feeding Big Data in many different ways and at different levels of consciousness throughout their life cycle. The following paragraphs provide a description of how each person in the digital society is exposed to the influence of data analytics throughout the main life phases.

6.1 Birth

In our information society, data collection and processing on a human being starts even before they are born. Genetic data acquired during the prenatal period is becoming crucial to monitoring correct foetal development and for early identification of potential congenital anomalies.

Also, a variety of information about parents and especially the mother is usually stored. Data collected during the prenatal period may be also stored in registries used in epidemiological studies to monitor temporal or geographical patterns over many years.

Furthermore, shortly after birth, hospitals and clinics routinely collect a lot of data about each new-born in order to make a complete physical assessment of the baby's health within the very first minutes of life. For each new-born, medical teams report weight, heart and respiratory rates, muscle tone, reflexes, etc.

Babies are also screened for a list of disorders that are treatable, such as some metabolic disorders. Screening programs are often run by national or regional governing bodies and vary from Member State to Member State.

All this data is fed into vast historical national or regional databases and is often used for epidemiological investigations. People are usually not fully aware of the type of data stored, retention period and purposes behind the processing of their data. Indeed, as far as the individual health records are concerned, getting a full informed consent is almost impossible, since the long-term or short-term relevancy of certain data is not always clear for which uses or purposes.

Even if scientific research could make remarkable steps forward in tackling some of the current health challenges and improve healthcare by having access to an individual's health data, the collection and use of this personal data presents a number of relevant risks. In fact, an unexpected violation of security systems could lead to the revelation of details of diseases or inherited disorders about an identifiable person, rendering that person exposed to psychosocial and financial harm at any time in life.

The way that data is stored in public registries and how it is processed depends on the legislation of each Member State. The EU Countries have quite different standards to protect health data.

Anyway, the new EU Regulation explicitly states that it is often not possible to fully identify the purpose of personal data processing for scientific research at the time of data collection. This law identifies a possible trade-off between the data protection principle and the needs of scientific research, and acknowledges a person's right to give their consent to certain areas of scientific research while, at the same time, keeping within recognised ethical standards for scientific research.

6.2 Infancy

For the purposes of this study a person is considered "infant" when aged from 0 to 10 years.
From a very early age, children's data starts to be recorded in the context of education, leisure activities and even Internet and mobile communications.

6.2.1 Education

Schools collect and store a lot of data about their students, from basic student identifying information such as name, date of birth, gender, address, identifying details of the parents or holders of parental responsibility over the pupils, to more sensitive information such as children's disabilities, family status, and behavioural problems. During the whole school lifetime of each single pupil, schools also trace grades, qualitative assessments from teachers, and level of learning.

Furthermore, with the continuing growth of online learning, students use a wide range of learning resources, interact with a variety of applications, enhance their experience in virtual environments, and communicate with peers and teachers through different platforms, such as MOOCs (Massive Open Online Courses), learning management systems like Moodle, etc.

The interaction with these innovative tools generates a vast amount of granular learning-related data that make a huge amount of information on how students learn available, so the learning experience can also be assessed in depth in order to identify possible scope for improvement.

Therefore, on the one hand, Big Data analysis in education can bring concrete benefits for students, because learning tools can be more responsive and better reflect the learning attitudes and pace of single students, to even develop a form of personalized education. But it can also have unintended consequences, such as the potential discrimination that can arise from permanently labelling students as underperformers or slow learners.

6.2.2 Entertainment

Nowadays, children are more and more attracted to online game platforms, already when they are 5–6 years old. There is a variety of games out there, from boxed games, purchased from stores or online, to those that are downloaded directly to the console, PC or handheld device. Some of these games are played with subscription while some are free-to-play games supported by ads.
These games and consoles offer amusement and even learning experiences to children but at the same time often require a lot of personal data to allow users to fully enjoy the different social and gaming features. Moreover, the majority of the most popular games offer both social network and marketplace

features. For instance, users are invited to buy and sell digital collectables for platform specific credits, and to share scores and performances with friends or even unknown platform users. This way, game providers can not only collect data such as personal details, friends and contacts, credit card numbers, purchased goods, but also behavioural information about game preferences, time spent playing, etc.

The EU law, according to the new Regulation (2016/679), considers processing the personal data of a child lawful only when that child is at least 16 years old. If that child is below the age of 16 years, such processing is lawful only if and to the extent that consent is given or authorized by the holder of parental responsibility over the child.
But this provision risks being continuously disregarded, because it is hard to determine the real age of an online service user, hence children pretend to be older to access certain services.

6.3 Adolescence

For the purposes of this study a person is considered "adolescent" when aged from 11 to 19 years old.

6.3.1 Online services

Once a person grows and starts to join and explore social contexts more autonomously, his/her use of digital devices and services becomes more pervasive in the daily routines, especially nowadays, with the intense use of social networks. These Internet platforms are used to stay in touch with peers, meet new people, share contents, and find information without any kind of filter and often without control of parents or other adults. Social networks such as Facebook are designed to encourage the extension of one's networks and the sharing of as much personal and even sensitive information as possible.
For most of Facebook's youngest users, this social network represents an extension of their offline interactions and many of them publish a great amount of personal information that can include mobile phone numbers.
They are also invited to show preferences and interests to offer targeted ads. Facebook even analyses users' preferences to promote ads to their friends. It also uses plugins and scripts to tie browsing data to Facebook accounts.

Moreover, adolescents start to have their own money to spend and this fact turns them in very attractive consumers who can easily get access to marketplaces to buy goods. Their digital identities become even more accurate. Many digital services ask users to create a unique virtual identity to make use of many different free services. The use of these kinds of services imply often giving data such as: personal details (name, date of birth, place of birth etc.) home address, phone number, credit card number, GPS position, etc.

Search engines deserve to be mentioned separately, as they are pervasive and often make use of the information connected to a user's digital identity[85,86]. When searching by keywords in Google, Bing,

[85] Pariser (2011). *The filter bubble – what the Internet is hiding from you.* The Penguin Press, New York.

Yahoo or using services like Siri and Cortana (Apple's and Microsoft's digital assistants), search topics are saved as data connected to that user. In the case of digital assistants, a user's voice is even sent to servers in order to be analysed and translated into words and topics to be searched. The immediate effect is that results that are more related to that user's interests are shown on top of the results list, but this might come at a cost. In the mid-term, and progressively as a user performs more and more searches, their results list becomes more tailored to their profile as described by their previous search keywords, clicked result pages and digital identity.

6.3.2 Mobile communication

Adolescents start quite early to use mobile communication (in some cases even before adolescence) and to have their own mobile device. Through mobile applications on a smartphone, they can access all the digital services available on Internet, such as social networks, games, digital marketplaces, music applications, as well as instant messaging services such as WhatsApp, Instagram, Telegram, Snap Chat, etc. Smartphones also give very fast ways to pay for services, even to buy very cheap applications.

6.3.3 Streaming services

Access to contents, such as videos and music, has an important impact on personality development during adolescence. Furthermore, contents creators and sharers have the ever-increasing opportunity to reach a larger public, often made up of young and very young people. This also holds true in the case of authors at an amateur level.

This fact represents not only an important sociological issue, since it involves social dynamics like the development of personal opinions and interests, but it also gives the opportunity to service provider companies to profile and archive teen users' data.

Internet video streaming providers (e.g. YouTube, Daily Motion and Netflix) collect a large variety of data about their users, from personal data to visualization habits, preferences and tendencies. A specific case: when, as a Google account owner, a registered user watches YouTube videos, all of their navigation data are archived and analysed, including "likes" and comments related to those specific contents.

The potential young age of users may involve more critical privacy and security issues, even from a legal side. The same topics are significant when it comes to music streaming platforms (e.g. Spotify, Deezer etc.).

Particular attention should be paid to adult contents streaming and downloading sites: these are some of the most visited web sites in the world, and the average user age has decreased over the last decades. This kind of service provider (from free platforms like YouPorn and Xvideos, to subscription ones) collect very sensitive data regarding sexual preferences, that may provide insights to a user's gender orientation.

[86] Feigenbaum et al (2014). *Open vs Closed systems for accountability.* Proceedings of the 2014 Symposium and Bootcamp on the Science of Security.

Another critical question about web porn is users age: access to contents is generally prohibited to people under the age of 18, but control over real age of visitors can be easily circumvented, so that often the data of people who are too young is thrown into the Big Data vortex.

6.4 Young adulthood

Adulthood represents the phase of a person's life during which an individual's choices are more determinant over which kind of data and to whom they are provided. It is, in fact, during this time that a young person acquires independence and creates their own life according to their own preferences. This reflects on their digital representation.

For the sake of clarity, this section will be discussed in two sections that refer, respectively, to young adults and middle-aged persons. In the first case, we refer broadly to individuals between 20 and 39 years old, during which human beings tend to be at their healthiest and most productive stage and they are mostly concerned with achieving life goals. In the second case, we mostly refer to individuals aged 40 to 60 years old, when health status has gone over its peak and has generally started to deteriorate slightly, and personal goals (family, career, gratifying social network) have usually already been achieved, and are being enjoyed.

6.4.1 Job search

One of the unifying scenarios that is likely to occur to a young person in their 20's is job search, whose acquisition is necessary for most of the subsequent steps of an individual's life.

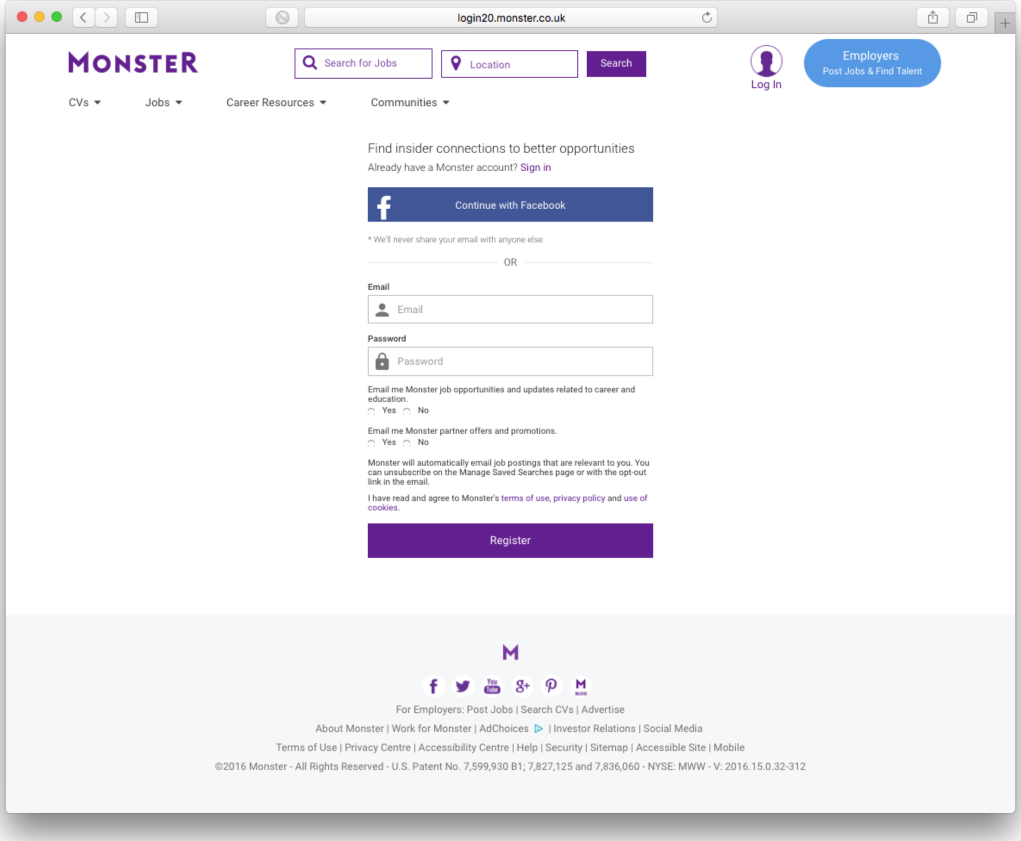
This field has undergone a rapid change over the last decade, with most job applications now submitted online, either via a company's web portal or, much more likely, via global job portals such as Monster[87], InfoJobs[88] and LinkedIn[89]. Typically, young EU citizens will subscribe to all the available services in order to maximize their likelihood of succeeding in their job hunt, and they will deal with different approaches to data collection and services provided.

For instance, in the case of Monster (Figure 2), during registration an initial choice is given to use one's own Facebook account to join the portal ('Continue with Facebook'). While this grants the user access to Monster's services without the need of an additional digital identity and a new password thanks to the single-sign-on mechanism, profile details are then shared between Facebook and Monster, producing potential privacy issues, notwithstanding the potential lack of awareness that this process is taking place. Registration, in the case of Monster, InfoJobs and other similar portals then follows a usual pattern where **personal details** are provided, in most cases down to the level of the **postal code**, **career** and **education level**. A *curriculum vitae* is then created online or updated as a pdf or text document.

[87] http://www.monster.com
[88] http://www.infojobs.com
[89] https://www.linkedin.com/

Even though almost none of the **finer grained personal details** are compulsory, a young adult is surely pushed to provide as much information as possible in order to find relevant job offers e.g. near the place where they live or with a profile similar to theirs. The creation of a digital identity is not required in these cases.

The actual job search on these portals then follows a pattern where the user searches available positions on the basis of key words, level of the position, geographical area, etc. and available positions are provided, together with the possibility to apply directly or through the company offering the position itself.



Find insider connections to better opportunities

Already have a Monster account? Sign in

* We'll never share your email with anyone else.

Email

Email

Password

Password

Email me Monster job opportunities and updates related to career and education.

○ Yes ○ No

Email me Monster partner offers and promotions.

○ Yes ○ No

Monster will automatically email job postings that are relevant to you. You can unsubscribe on the Manage Saved Searches page or with the opt-out link in the email.

I have read and agree to Monster's terms of use, privacy policy and use of cookies.

For Employers: Post Jobs | Search CVs | Advertise

About Monster | Work for Monster | AdChoices | Investor Relations | Social Media

Terms of Use | Privacy Centre | Accessibility Centre | Help | Security | Sitemap | Accessible Site | Mobile

©2016 Monster - All Rights Reserved - U.S. Patent No. 7,599,930 B1; 7,827,125 and 7,836,060 - NYSE: MWW - V: 2016.15.0.32-312

Figure 2. A screenshot of Monster's registration page where the initial option of using one's own Facebook profile to access services is given.

LinkedIn works slightly differently and, after the usual input of **name**, **surname**, **e-mail** and the choice of a password, a young adult is then prompted to create a personal profile, which is a new *digital identity*, by providing **personal details**, **geographical position**, **company** and **job position** (if already employed), and **areas of professional interests**, to name a few. The user interface is similar to a social network, with a personal page including the user's data and CV, and a page containing a news feed on the contact's activity and available positions. It is possible to connect to other users, follow companies and share private messages. Some of these features are only available to paid accounts, and marketing e-mails with commercial offers are regularly sent to users.

This kind of professional social network is similar to other social networks in different fields (e.g. Facebook, Twitter), and is, therefore, readily usable by a young adult who is familiar with this approach. Furthermore, it allows for a quick job search, ease of application, and the possibility to check the company that posted the advertisement.

Big Data in this case has a huge potential for matching applicants with the most suitable vacant positions, and it has a positive influence on both companies and citizens by favouring this process, especially during the present day economic crisis. This use of Big Data empowers a young adult in allowing for an easier application process and a broader view of the available positions and allows companies to improve the selection process by choosing more quickly from a larger candidate pool.

On the other hand, some young adults might not be versed enough in the use of the tools Internet provides or they might have a limited access to Internet due to the place they live in and so they might have a more difficult experience in finding a job. This *digital divide* generates the potential for discrimination since individuals might not be selected for a position, even though their skills match those required, because they never entered the candidate pool.

6.4.2 Banking services and insurance

Following the start of a new job, *banking services* become essential for a young adult, who needs a bank account number for their salary, rent and house utilities.

Procedures can change quite dramatically from one bank to the other within Europe (e.g. Deutsche Bank, BNP Paribas, Unicredit, ING), but the initial part of setting up a bank account usually takes place “on paper” and involves filling in forms that reiterate the request for common **personal details**. A working bank account with, most often, an online banking service is then provided. While this process does not usually involve surrendering a large amount of private data, additional services such as credit cards and ATM cards generate an impressive amount of data about an individual’s **behaviour**. **Spending habits** and, when a purchase is made from a physical shop, **geographical position** are recorded for approval of every purchase, on top of the basic information relative to the bank **account balance**. Since accounts are often used for direct payments of utilities (electrical power, gas, water, mobile communication, landline and Internet connection, cable or satellite television), a complete depiction of the management of an individual’s life is basically projected in data form.

Banking services are often linked to other financial services, such as *insurance*, following a so-called Bank Insurance Model (BIM), which is particularly common in Europe and offered by major banking corporations such as Unicredit, Deutsche Bank, BNP Paribas and ING. These banks have an insurance department that offers services such as life or health insurance, car insurance, and others. Insurance companies also operate independently, and some of the largest in Europe are AXA, Allianz, and Generali Assicurazioni.

Dealing with the insurance of a wide array of items, such as vehicles, health, life, house, etc., these companies possess an impressive amount of data about an **individual’s lifestyle**, **living conditions** and **health status** in general. One of the primary uses of this data is classifying a new applicant, such as a young person who has just bought their first car and is in need of car insurance in order to provide a precise and fair risk assessment and compute a premium which is proportional to this risk.

6.4.3 Utilities

Having the job and bank account situation all sorted, young adults often move out and start an independent life in their own home, often with more responsibilities than those associated with house sharing while studying at university. This usually represents a major step forward and comes with the need for utilities contracts to be signed and supply set up.

A new home requires an electrical power, gas and water supply, often a landline and an Internet connection, and at times satellite or cable television. Some of these utilities might be supplied by the same provider, as is often the case of gas and electrical power (ENI, Engie, E.ON are some examples of joined providers), or separately, and they always require that a contract is signed and **common personal details** provided. Data is then generated with everyday consumption and the relative provider is then granted a supply of usage data (that reflects behavioural data) in exchange for e.g. electrical power. It is not uncommon to receive commercial offers from these providers also after the contract has expired or has been cancelled, and these offers are specifically tailored according to the customer's habits as inferred by the **consumption data**, **number of calls to the customer service**, **time of day** when these calls are made, etc. These variables are, in fact, part of the predictive models that are often employed by utilities providers to make forecasts on the possibility of a customer leaving for another provider, the so-called "churning".

6.4.4 Online dating and Pornography

One of the areas that has been constantly gaining momentum for the last decade, especially among young adults, is online dating, with the majority of users between 18 and 44 years old[90]. Among the most popular worldwide, and in Europe, are Meetic, Tinder, be2, eHarmony and Meet me, but others are present and are addressed to specific groups of people according to their gender identity, age class or marital status. Extreme cases of online dating providers that use genetic matchmaking also exist (DatingDNA.com, GenePartner.com, ScientificMatch.com), for now confined to the USA market.

Once a young adult has settled into their independent life, a natural desire to have a relationship often follows, and online dating sites might provide a solution for meeting new people. This is where Big Data comes into play, as this kind of site blends the capabilities of a social network with chats, geolocation based searches and matching algorithms. In addition to the usual **name**, **last name**, **date and place of birth**, **place of residence** and **e-mail address**, during the subscription a plethora of personal information is asked, some of which sensitive, including **interests**, **photographs**, a **physical description**, **sexual orientation**, **personal aspirations**, **smoking and drinking habits**, **social life habits**, etc. While almost none of them is compulsory, a young person looking for a potential good

[90] *5 facts about online dating.* Smith & Anderson, 29 February 2016. Pew Research Center. Available at http://www.pewresearch.org/fact-tank/2016/02/29/5-facts-about-online-dating/

match will probably input most of them, thus generating a very detailed virtual identity, at times even more detailed than that present on other social networks such as Facebook, since it includes very personal details about one's own private life.

While some of these services are based on matching algorithms that process subscribers' data to produce a good match, thus generating privacy and awareness concerns, as well as a feeling of loss of control, others choose a different approach. It is the case of the mobile app Tinder, which has rapidly become popular in the USA and then spread over Europe. Tinder is an app that **connects to a user's Facebook profile** and uses a geolocation algorithm that uses **GPS data** to match people within a certain radius, and allows users to quickly swipe left or right with their thumb to discard or select a potential match.

One of the features that has the potential to generate serious concerns about a user's privacy is the fact that some of these services, while not directly collecting data that are outside of the privacy statement accepted by users, might redirect them to third party websites that operate out of the notified privacy statement and to which approval has not been previously given. These websites occasionally collect data that is then sent back to the original online dating website or app[91].

6.5 Middle age

6.5.1 Health services

Although not unique to the life of a middle-aged person, health services appear in this section because they tend to get more frequent and necessary during this phase, and tend to increase as a person ages. They may be provided within a private or a public health context, and collected data tend to be the same in both instances.

When a middle-aged person receives health care, minimum basic **personal details** are provided, often including **a national identification number**, which is part of the information on the European Health Insurance Card.

When the actual health care starts, the kind of data collected can vary wildly depending on the illness or condition that a person is being treated for. In general, it is **physiological** and **behavioural data**, often accompanied by a **description**. All this is then used for direct diagnosis and treatment of the person seeking health care, and represents an invaluable source for research to produce cures and improve life conditions and expectancy for the whole population.

6.5.2 Wearables and health & fitness apps

[91] The following is an extract of the Third Party Section of Tinder's privacy statement: "There are a number of places on our Service where you may click on a link to access other websites that do not operate under this Privacy Policy. For example, if you click on an advertisement on our Service, you may be taken to a website that we do not control. These third- party websites may independently solicit and collect information, including personal information, from you and, in some instances, provide us with information about your activities on those websites."

The potential range of physiological data collected has been recently dramatically expanded by the introduction of wearable devices equipped with a wide array of sensors. The use of these devices appears almost equally distributed across young to middle aged people, especially men[92]. Even though retail price for these items is at present expensive, it will likely decrease as their use becomes more widespread.

A middle-aged adult, therefore, is likely, over the course of their life, to purchase and use one of these devices, such as an Apple Watch, designed for the broader public, or one of the wearables specifically designed for health and fitness such as UP by Jawbone or the newer Suunto Spartan. These devices can measure position using GPS, track your sleep pattern, heart rate, body temperature, speed, etc. and this list is very likely to increase in the near future. Apart from the wide array of variables measured, one impressive feature is that they are measured constantly while the devices are worn, providing exceptional insights to fine grained temporal patterns in an individual's health state.

These devices are often paired to mobile apps with which data is exchanged and capabilities extended. One example is RunKeeper, an app for iOS and Android mobile phones, which can also be installed on an Apple Watch and Android Wear. This app adds a social side to the monitoring of one's physical activity and physiological parameters by allowing users to connect and check on each other's activity and running routes.

6.5.3 Job searching

The transformation of the job market that has led to a higher mobility of employees, and the rising unemployment rate have the potential to impact a middle-aged person's life. Differently from when they had to look for a job a couple of decades (or even a few years) ago, a middle-aged adult who loses their job might now be in the situation of starting with application processes.

As described in section 6.4.1 above, the job search has dramatically shifted towards Internet based means (job portals or professional networks such as LinkedIn), and this might be in stark contrast to how an adult looked for a position 20 years ago, sending applications via regular mail or taking them in person to the hiring company. While the modern approach is easier in that it allows for quicker job searches and applications, it is new and might require a fair amount of learning on the part of a senior applicant, who might benefit from the new scenario or be hindered by the presence of a digital divide in the use of technology.

6.5.4 Online dating

A situation similar to that detailed in 6.5.3 aboveis the one that a middle-aged person dealing with a break up or a divorce might have to go through while looking for a new partner. Since online dating

[92] https://www.npd.com/wps/portal/npd/us/news/press-releases/2015/the-demographic-divide-fitness-trackers-and-smartwatches-attracting-very-different-segments-of-the-market-according-to-the-npd-group/

has increasingly become an accepted practice, they might be willing to try it, and will have to surrender a large amount of **personal details** (see 6.4.4 abovefor a non-exhaustive list) while, potentially, not being fully aware of the data collection taking place and the potential use and misuse of their data.

6.6 Old age

6.6.1 Health care

Old age is a time when health care has the most important role in the life of an individual, if we are to grant healthy ageing to European citizens. At present, health care for old people is mostly carried out through local general practitioners who rely heavily on personal visits to their patients to conduct proper health checks, and hospitals or rest homes, when needed.

While not being fully exploited now in this field, wearables equipped with sensors to check on **heart rate**, **blood pressure**, **body temperature**, **position** (upright or horizontal) and **acceleration** (to detect falls) are going to be the main protagonists of the health care of the elderly in the near future. Smartphone apps to monitor the state of health of people suffering from **Parkinson's disease**[93] or **asthma**[94] already exist, and the spread of other similar apps is going to give rise to an unprecedented ease of surveillance and care for the elderly, although potentially to the expense of a reduced privacy.

6.7 Death

Although seldom included, death is a pivotal part of the life of a human being. Data collection does not exactly end with the life of a person, as data about the **date and place of death**, **place of burial** and **personal details of heirs** is produced and all without the consent of the deceased. Furthermore, while bureaucratic formalities eventually finish and data collection potentially as well, digital identities persist and are still part of the datasets used by companies to produce analyses results, forecasts and commercial offers.

No unified and globally accepted way of dealing with the passing away of a human being and his digital identities is presently available, although some providers such as Yahoo! include a 'death clause' in their terms of agreement that is activated by the receipt of a death certificate. Another mechanism involves the deletion of an account and all connected data after a prolonged period of inactivity, but what 'prolonged' means is not always clear and may vary according to the situation. An overview of the different approaches is provided in a recent study[95].

[93] https://www.michaeljfox.org/foundation/publication-detail.html?id=562&category=7

[94] http://apps.icahn.mssm.edu/asthma/

[95] Locasto et al (2011). *Security and privacy considerations in digital death.* Proceedings of the 2011 New Security Paradigms Workshop, 1–10.

7. Ethical issues

This chapter illustrates the issues that arise when the scenarios detailed above are looked at from an ethical perspective, taking into account an individual's need for privacy and self-determination. Each ethical issue will be detailed in a paragraph bearing the same name of the issue it deals with.

7.1 Awareness

Since registration to online services, such as the creation of a digital identity or the subscription to an identity provider, are often dealt with quickly and information is only partially provided to a subscriber during this process, awareness of the kind of data that is being provided is often scarce.
This issue is heightened by the possibility that users are given the opportunity to use a digital identity (e.g. Facebook profile, Google identity) to access third party services more and more often. This leads to a quicker and easier use of online resources, but, at the same time, creates opacity about the data that is shared between the identity provider and the service used. This becomes even more concerning if we think that identity providers, on top of the details surrendered at the moment of subscription, also hold data collected during a user's web surfing while logged in. These data is extremely detailed and personal.

This use of Big Data diminishes an individual's power and freedom by removing the necessary condition of knowing which data is collected and how it is processed. An increased awareness might produce a higher participation level of citizens, thus resulting in more data available for governments and private companies within an ethical context of respect of human dignity.

7.2 Control

As already pointed out in literature[96] "individuals can feel powerless in the relation to data. This is because the human-data relation is asymmetric; there seems to be an asymmetry of control with data having the upper hand". This refers to the situation often faced by users when they decide they want part or all of the data they provided to a service deleted. Even in the case when the service provider decided to follow up on the user's request and actually delete all their data, this might not affect that user's data that has already been sold to other companies or has been heavily processed, thus leading to a loss of control over the access to one's own personal data[97].
The so-called right to be forgotten is considered in the EU as one of the pillars of an individual's control over their personal data, and, after its introduction in principle in 1995[98], it has sparked discussions[99], also due to the many practical details required to apply this principle.

[96] Swan (2015). Philosophy of Big Data: Expanding the human-data relation with Big Data science services. *IEEE 1st International Conference on Big Data Computing Service and Applications*, 468–477.

[97] Bauer et al (2013). A comparison of users' perceptions of and willingness to use Google, Facebook, and Google+ single-sign-on functionality. *Proceedings of the 2013 ACM Workshop on Digital Identity Management*, 25–36.

[98] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. *Official Journal* L 281, 23/11/1995, 31–50.

7.3 Trust

Trust has emerged as a complex topic within the present day Big Data context due to its many interdependencies with the broader issue of privacy in general and awareness. One of the most prominent features of trust is that, so far, it has mostly been addressed from a strictly technological perspective. As pointed out in the relevant literature[100], we still do not have a full understanding of how the instauration of trust bonds take place between human beings and in the context of desktop computing. The creation of hardware and software architectures that might be conducive to the creation of trust bonds between human beings and objects, for instance in an IoT context, therefore still seems far off in the future.

Even when the use of online services simply mediates the interaction between humans, as in the case of online health interventions[101], trust issues are at the base of a user's acceptance to provide personal data ("Patients need to know who is behind an avatar before exchanging private information.").

7.4 Ownership

A more complicated matter revolves around how to consider a user's data that was produced after processing the original dataset: are they still a user's data, or do they belong to the company that carried out the analyses? Or to the company that collected the original data?

This issue has been tackled by limiting the place where physical storage of data takes place, namely the countries where servers are based, and the EU approach has been to progressively restrict the possibility for EU citizens' data to be stored out of the so-called "Euro cloud". This approach still leaves the problem of where already processed data is stored, 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.

7.5 Surveillance and security

Since more and more data sources are available, and the advancement of technology has made it both easier and faster to analyse data to generate insights, the idea that our position might be in some way known at most times has become commonplace. The ubiquitous use of CCTV circuits, coupled with the GPS positioning capabilities built in mobile devices, and the use of credit cards and ATM cards for payments and withdrawals represent only some of the available means to track one's position over time.

[99] Mantelero (2013). The EU Proposal for a General Data Protection Regulation and the roots of the 'right to be forgotten'. *Computer Law & Security Review*, **29**: 229–235.

[100] Schrammel et al (2011). Privacy, trust and interaction in the Internet of things. *International Joint Conference on Ambient Intelligence*, 378–379. *12th International Conference of Information Technology – New Generations*, 567–573.

[101] Lim and Thuemmler (2015). Opportunities and challenges of Internet-based health interventions in the future Internet.

This ease of tracking has surely increased safety (or the perception of it) across Europe and allows for a more effective and focused workflow of police forces during investigations, but this might come at a cost. Active surveillance is an extremely effective mean of limiting citizens' liberties, and has already been used as such by totalitarian government over the course of EU's history. When surveillance is carried out by employers (e.g. by controlling or restricting access to websites, by encouraging the use of a company's devices, by installing cameras in the workplace) this might lead to an increased stress level of workers, which often translates into lower productivity. The awareness of the possibility of being watched at any moment, furthermore, creates an ideal panopticon in which an individual's actions tend to conform to the expected norm, as shown by field experiments[102].

Surveillance is diagonal to the whole society and it is non-directional, as it can take place across levels of a society or within a level itself, and it can be from the higher to the lower level or vice versa. These features have led to the creation of a taxonomy of surveillance in which surveillance *sensu strictu* (top down, such as an employer on their employees), surveillance (bottom up, such as citizens controlling a government's initiatives), peer surveillance (horizontal, such as Facebook users checking on each other's profile and status updates) and self-surveillance (an individual recording their own every action) have been identified[103].

7.6 Digital Identity

The creation of digital identities has the obvious advantage of generating the possibility of accessing online contents and all related services through them. The widespread use of digital identities has created fertile ground for the practice of retrieving publicly available information on a person (for example following a job application) from the web, in order to generate insights before actually meeting them.

While this process, within boundaries, is accepted as legal, it has the potential to generate discrimination based on the representation of a person as portrayed by their data, as opposed to their real self, in a process known as *dictatorship of data* where "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"[104], and personal interaction is placed in a later step after analysis of digital identities.

7.7 Tailored Reality

Every time we e.g. search by keywords using a search engine, when we purchase an item from an online marketplace, or provide a personal detail to an identity provider, the data we produce is

[102] Panagopoulos (2011). *Social Pressure, Surveillance and Community Size: Evidence from Field Experiments on Voter Turnout*. Special Symposium: Electoral Forecasting Symposium, 30: 353–357.

[103] Briggs et al (2016). *Everyday surveillance*. *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 3566–3573.

[104] Norwegian Data Protection Authority's (2013). *Big data-privacy principles under pressure*.

potentially stored. Data processing and analysis are known to be used in providing, during subsequent accesses, personalised results in terms of order of the result pages shown by search engines, marketing offers that are received in our e-mail boxes, advertisements that appear on social networks and other services pages, thus generating a narrower and more personalized version of a user's online experience (the so-called "filter bubble"[105]).

One of the advantages of this extreme personalization is that a user will most likely find what they need in a matter of a few clicks. Nevertheless, this lack of exposure to different items, perspectives and, ultimately, ideas in the long run could represent a strong hindrance factor for creativity and the development of a tolerant attitude by fracturing the reference points necessary for a shared political and social life[106].

7.8 De-Anonymization

This issue has become prominent only recently due to the introduction of de-anonymization techniques allowed by the increased computational power of modern day personal computers. Traditional anonymization techniques focus on the data and aim to make each data entry non-identifiable by removing (or substituting) uniquely identifiable information (e.g. fiscal code, NHS number). While apparently effective, this method fails to account for the powerful insights that are produced when datasets from different sources (e.g. voting lists and social network profiles) are joined[107]. When doing this, perfectly anonymous information such as the date of birth, might be combined with the city of residence and marital status to uniquely identify an individual with varying degrees of certainty.

When pushed to the extreme, profiling can become a powerful, although discomforting, tool as shown in the famous study where authors presented the adaption of a geographical profiling technique used in criminology to the identification of UK street art artist Banksy[108]. This topic raised concerns over the degree to which the intrusion in an individual's life is allowable, as well as discussion over security and surveillance.

7.9 Digital divide

Digital divide refers to the difficulty in accessing services delivered through the use of new technologies, such as the Internet, and difficulties in actually understanding how these processes work due to not being familiar with them. One of the major issues produced by the digital divide is that referring to job hunting, since it is now conducted mainly online in ways at times difficult to understand by senior citizens that might find themselves uneasy in approaching this new way of

[105] Pariser (2011). *The filter bubble – what the Internet is hiding from you.* Penguin Press, New York.

[106] Crawford et al (2014). *Critiquing Big Data: politics, ethics, epistemology.* International Journal of Communication, 8: 1663–1672.

[107] Narayanan and Shmatikov (2010). *Myths and fallacies of "Personally identifiable information".* Communications of the ACM, 53: 24–26.

[108] Hauge et al (2016). *Tagging Banksy: using geographic profiling to investigate a modern art mystery*. Journal of Spatial Science, 61: 185–190.

dealing with, for instance, job loss. This is an issue that has the potential to affect large portions of a society, especially considered the higher mobility that is now part of the job market.

A similar situation takes place with senior citizens that approach online dating services as a means to finding a new partner, after the loss of a life companion. Frustration might arise from a new perspective on social interactions, that might ultimately lead to giving up and cause social withdrawal.

7.10 Privacy

Privacy is an all-encompassing topic overarching the issues discussed above and their ethical implications, since it broadly refers to the possibility to withhold personal information and prevent its use without consent, considering that all these issues impact in some way on a citizen’s right to privacy.

While it may be argued that a citizen might be willing to give up some of their privacy in exchange for a higher level of security and personal safety, it cannot be safely assumed for everyone and, furthermore, the use of privacy as a bargaining chip might be considered unethical to some extent. Trust also comes into play when we talk about privacy, especially in the medical field, where personal and behavioural details of the utmost importance, and the existence of a trust bond between a patient and the data collector (e.g. the medical doctor and the staff of a hospital) is essential for the well-being of that specific citizen. Differently from the situation considered before, in a case like this it is the decision not to provide a full disclosure of the requested information that might be considered unethical, as it may directly impact the health of the citizen directly involved and indirectly hinder progress in research in that field with repercussion on other individuals.

Personal discretion in the use of privacy, at this point, might appear to be a sensible choice, but case-based applications of a principle are, nevertheless, risky as they leave much to the interpretation of the people involved and an ordered society requires clear guidelines to abide to if we are to give equal opportunities and rights to everyone.

In the table below, a schematic view of where the ethical issues just discussed come into play within each life phase of a human being is provided.

Table 1. Identification of the actual scenarios for each life phase, with an indication of the ethical issues arising in each.

LIFE PHASE	ACTUAL SCENARIO	ETHICAL ISSUE
Birth	Pre-natal screenings	Awareness
		Consent
		Ownership
		Privacy
Infancy	Education	Control
		Privacy
	Entertainment	Awareness
		Consent
		Digital identity
		Ownership
		Privacy
		Trust
Adolescence	Online services	Awareness
		Consent
		Control
		Digital identity
		Privacy
		Surveillance
		Tailored reality
		Trust
	Mobile communication	Awareness
		Consent
		Privacy
		Surveillance
		Security
Adulthood	Job search	Awareness
		Consent
		De-anonymization
		Digital divide
		Digital identity
		Ownership
		Privacy
		Trust
	Banking services and insurances	Awareness
		Control
		De-anonymization
		Ownership

LIFE PHASE	ACTUAL SCENARIO	ETHICAL ISSUE
		Privacy
		Trust
	Utilities	Consent
		Privacy
	Online dating	Awareness
		Consent
		Control
		De-anonymization
		Digital divide
		Digital identity
		Ownership
		Privacy
		Trust
	Health care	De-anonymization
		Ownership
		Privacy
		Trust
	Wearables and health fitness & apps	Awareness
		Consent
		Control
		Ownership
		Privacy
		Surveillance
		Security
Old age	Health care	De-anonymization
		Ownership
		Privacy
		Trust
Death	Post-mortem	Awareness
		Consent
		Digital identity

8. Real-world uses for Big Data: five case studies in Europe

When Big Data are implemented in policies by governments at any level, benefits tend to be immediate for both citizens, institutions and companies. Benefits for citizens mostly comprise an increased ease of access to services, as well as a wider array of offers, while institutions can mostly reduce costs (after an initial investment in ICT infrastructures and training of professionals) of their routine operations and optimize the use of their resources. Companies mostly benefit in terms of the effectiveness of their marketing campaigns and productive processes, thus producing a higher return of investment.

In order to better communicate benefits at all levels that are subsequent to the introduction of Big Data technologies in planning and management, five case studies are described in the section that follows. These case studies refer to actual implementations of the use of Big Data within the EU member states and include an explanation of the positive impacts that followed (and are still following, in most case) their adoption.

Furthermore, these case studies, that span fields from e.g. the collection and use of sanitary data to provide more efficient services to citizens, to the use of Big Data for evidence-informed policy-making, offer a clear hint to what the future aspect of the use of Big Data will be within the EU, since they can be interpreted as showing the path that the future development will undertake.

8.1 Managing health data – the "FedERa" project

The first case study regards the management of citizens' health data at regional level. The Italian Region Emilia Romagna in 2007 started the developing of the project "FedERa"[109], to allow the citizens to access all online services of the Public Regional Entities via a single digital profile. One of the services made available through this system is the Electronic Health Record, that collects health data and information of each citizen and keep trace of the individual medical history. The record represents also the gateway to access the e-health services provided by the Emilia-Romagna Region.

The Electronic Health Record is a self-feeding system, and the documents collected are produced by a network of doctors and paediatricians, health care facilities, hospitals and increasing private health structures that have received the accreditation.

The System complies with the Italian Personal Data Protection Code and the national Regulation on electronic health records entered into force in 2015.

Through this system, the citizens can see their own health data, medical reports, prescriptions, and make use of useful services such book a medical examination or change the family doctor.

The data are not collected in a unique data warehouse but is a federation of data centres at regional level.

[109] https://federa.lepida.it/

8.2 Smart cities – the cutting-edge application of big data

In Rimini, an Italian city situated in Emilia Romagna region, an innovative enterprise has reproduced on a 1:1 scale a small district of the city, called EasyLand[110], where the company has installed all the latest technologies in the field of the Internet of Things (I.o.T.) dedicated to the Smart City in order to make visible a prototype of future cities.

The aim is to make accessible to administrators and citizens a concrete picture of how a digital city may appear and how it can change and improve the everyday life of citizens.

EasyLand contains sensors that measure environmental pollution, cameras for video surveillance, traffic monitoring, vehicle number plate recognition systems, parking sensors, smart bins, interactive multimedia information points and much more, all for the sake of energy saving, giving more information to the citizens and improving the security of urban environment.

8.3 Big Data in statistics – an international project

Big Data offer unprecedented opportunities for official statistics in terms of additional data sources, faster ways to produce analysis and timely statistics.

In April 2013, the United Nations Economic Commission for Europe (UNECE) Expert Group on the Management of Statistical Information Systems (MSIS) identified Big Data as a key challenge for official statistics, and called on the High-Level Group for the Modernisation of Statistical Production and Services (HLG) to focus on this topic in its plans for future work. In this context, the HLG sponsored a series of international collaboration projects to better understand how to exploit the potential of Big Data and other new data sources, to support the production of official statistics. In 2014 the international project" The Role of Big Data in the Modernisation of Statistical Production"[111] started, bringing together 75 experts from 20 different entities specifically National Statistical Offices and International Organizations national and international statistical organisations around the world. The researchers collected data from different sources, such as page view data of Wikipedia and social media data from Twitter, with the aim of exploring their feasibility as sources for official statistics.

The main output of this project is the “Sandbox”, a shared computation environment for the storage and the analysis of large-scale datasets. The Sandbox is the first example of shared international statistical Big Data research capability. The Sandbox is now available for developing and evaluating new software programmes, methodologies or exploring the potential of new data sources, and to share non-confidential data sets that cover multiple countries, as well as public-use micro-data sets.

8.4 Big data for evidence-informed policy-making

Big Data can be an inexhaustible source of information for policy makers, to increase the variables to consider and predict the evolution of social and economic patterns.

[110] http://www.easylumen.it/it/easyland/

[111] http://www1.unece.org/stat/platform/display/bigdata/Big+Data+Projects

Two main types of data are being used in modern policymaking[112]: public datasets from administrative sources, in particular official statistics about population, economic indicators, etc. and data from social media, sensors and mobile phones that can be processed using innovative processing methodologies, such as sentiment analysis, location mapping or advanced social network analysis.

As for the policy areas covered, analytics are being used to implement policies in heterogeneous fields, from transport and mobility to environment, from government transparency to education.

In the debate about the risks connected to the massive use of data science in policy making, worries are expressed about the risk that the public sector could have a too close eye over the people life, as a "big brother" with unprecedented powers. Moreover, other experts pointed out the skills shortage in public administration to adequately manage the full value chain of data collection, analytics, interpretation and getting value for policies design. Furthermore, when the policy making process follows a data-driven approach, it can be perceived as less transparent and understandable by stakeholders and even more by the general public.

So far, across Europe quite a number of projects explored the use of analytics in policy design, many of which having been funded with the EU Research and Innovation funding programmes.

An interesting example is the FP7 funded project called "Insight"[113] that aims at exploring how ICT, with particular focus on data science and complexity theory, can help European cities to formulate and evaluate policies to stimulate a balanced economy. The project has been carried out by a European consortium of six partners coordinated by Universidad Politécnica de Madrid.

8.5 Big data in education: the new frontier of Learning analytics

Technologies have transformed the way people learn and is trained, offering a variety of different new opportunities such as massive online courses, online video-based learning, educational apps, webinars, virtual classes, and many other instruments. Moving education into technological era is producing big data around the learners and the contexts in which learning takes place.

The quite recent term "Learning analytics" refers to the measurement, collection, analysis and reporting of data about the progress of learners and the learning process itself. Today an educational institution collects already a lot of information about its students such as grades, credit hours, rates of participation in specific lessons, that can be turned into insights to improve the quality of the learning experience and the engagement of the students.

In this context, the Nottingham Trent University (NTU) has carried on an innovative project aimed at enhancing the academic experience for its 28,000 students, particularly their engagement with their course and identifying through a predictive algorithm, the students most at risk of early dropout[114]. The project is considered one of the most prominent learning analytics initiatives in the UK.

[112] Data for Policy: A study of big data and other innovative data-driven approaches for evidence-informed policymaking, 2015
[113] www.insight-fp7.eu
[114] Case Study I: Predictive analytics at Nottingham Trent University, Jisc

The university uses an analytics model to process data from several sources collected in an NTU data warehouse, such as the virtual learning environment, the library system, the student card and the student information system. A dashboard was also designed both for students to sense check their engagement against their peers and for tutors to discuss the students' performance with them. The results from the implementation of this system showed a strong association between engagement measured in the Dashboard and both progression and academic attainment of the students and consequently a strong predictive power of the algorithm in use.

9. Balancing actions and effectively balanced scenarios

The first phase of the study ended with the design of five "balancing actions", aimed at reaching a trade-off between economic growth in the EU and the respect for citizens' privacy and dignity in a context of exploitation of Big Data's potential.

During the second and last phase of the study, these actions were discussed extensively with selected stakeholders coming from different European countries and expressing various interests, positions and visions.

In the following chapter, these balancing actions are described to explain their possible impact on the actual scenario, the extent of potential benefits and the ethical concerns that each can contribute to address. Chapter 7 reports the reactions of the stakeholders and their suggestions on how integrate, change or enhance these actions.

For each balancing action a description is provided, followed by an identification of benefits and potential risks, together with measures to mitigate those risks.

9.1 Balancing action n. 1 – EU privacy management platform

9.1.1 Description

The first action comes from the idea, convincingly pointed out in the GDPR that natural persons should have control of their own personal data. Moreover, the new Directive 679/2016 forcefully points out how to apply the principle of transparency in the current scenario, where *the proliferation of actors and the technological complexity of practice make it difficult for the data subject to know and understand whether, by whom and for what purpose personal data relating to him*[115]. It is therefore essential to directly empower citizens and provide them with a concrete instrument to take control over their data and virtual identity details.

The idea is to establish a pan-European web portal as a unique privacy management hub where the European citizens voluntarily register to get access to a personal dashboard to visualize the list of all the public and private entities that have gained and currently store, process, share and re-use their personal data. For each company, services provider, public body, etc., the platform visualizes:

- The kind personal data each entity collects (e.g. name, date of birth, purchased goods, credit card number), but not the actual data
- How the entity manages personal data/how it complies with the EU law (in particular with the GDPR)

[115] Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), (58).

- What kind of services are provided in exchange for which data, and which service is currently active
- Whether or not the data is shared with third parties
- The logic behind the automatic personal data processing and, at least when based on profiling, the consequences of such processing
- Information on how to withdraw consent and ask for data deletion and/or service deactivation
- Identification of a data controller, data protection officer.

Since people are often not familiar with how to manage the privacy settings in web or mobile environments, the platform gives the possibility to easily find the pages to opt out certain services, denying consent of personal data processing. Ideally, the users should also be supported in understanding what happens in case they decide to withdraw their consent, for instance, in terms of giving up some services. The right to be informed about the possible consequences of consent denial is clearly affirmed in the Regulation (EU) 2016/679[116].

Comparing the data collected with the services provided, and more generally, the purposes behind the choose of gaining data, can also help to understand if the company is applying the principle of data minimization, which provide that personal data collected must be adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed[117].

From the entities' point of view, two sub scenarios are conceivable: one with mandatory registration and the other where the companies and institutions can choose voluntarily if they want to share the information requested.

9.1.2 Benefits

This action is a way to give back to people the direct control of what is happening in the digital communication with their personal data, so it is a strong measure to enhance awareness and informed consent. Also, it is a system to tackle the quite common attitude of digital services' users of not paying attention to the data they daily spread and to not reflect about possible misuses of them. The platform would also be used as a support to exercise the right to access, ask for rectification or erasure of personal data.

This scenario can have two different configurations, depending on the choice of making it compulsory or not for entities to register and indicate for all their users the personal data collected and processed.
In both cases, but especially if the registration is not mandatory, the companies that decide to join the platform can have a great benefit in terms of reputation and the trust of people.

Once a company makes available its data policy and personal data management, each user can directly check how it deals with personal data and a new relationship based on transparency and trust can result.

[116] Regulation (EU) 2016/679, (60)
[117] Regulation (EU) 2016/679, (39)

9.1.3 Risks and mitigation measures

The most apparent risk is that companies will not subscribe to the data management portal, which is addressed as explained in the section above by opting for a compulsory subscription. Heightened security issues are not devisable at the moment, since only information on the kind of data collected (and not the actual data; e.g. 'data of birth' would appear on the portal, and not, for instance, '18 August 1984') is present.

Another potential risk is the low reliability of the information provided by each company, who would hardly perceive any immediate benefit in communicating openly. A control system operated by a selected EU authority could be put in place in order to clarify any reported discrepancy between the information communicated by a company through the portal and the actual data collected during the provision of a service.

9.2 Balancing action n. 2 "Ethical Data Management Protocol (ED-MaP)"

9.2.1 Description

The objective of the second balancing action is similar to the first one's: increasing transparency and making people aware of the level of compliance with the EU law of Big Data's holders, both public and private. However, this measure does not imply a direct action of the users, but rather a commitment of companies, institutions or any other subject that for business, scientific research or other reasons keep and process large amounts of (personal) data.

The idea is to design a sound European certification system to identify the virtuous companies in the field of data protection. This measure is already outlined by the EU legislator in the GDPR (article 42) and during the study the research team reflected and discussed with the stakeholders to understand how it may be applied.

The certification, to do on a voluntary basis, must be based on the main principles pointed out by the GDPR, with particular reference to:

- Data minimization
- Special care of sensitive data and health data
- The respect of the right to be forgotten
- Data portability
- Data protection by design and by default

In order to create a general system that can work for all kinds of companies and services provided and to identify the quality criteria for complying with the data minimization principle, specific "sector studies" must be carried on defining the kind of data needed to supply each service, also clearly stating the temporal span data should reasonably be stored on a company's servers (data detention latency), and the scope as per the already defined purpose.

In other words, the sector study is needed to give indications on how to apply the EU standards and principles.

The standard may be designed by the International Organization for Standardization along the lines of the new EU Regulation.

9.2.1.1 Certification process

The proposed certification addresses the six following issues:

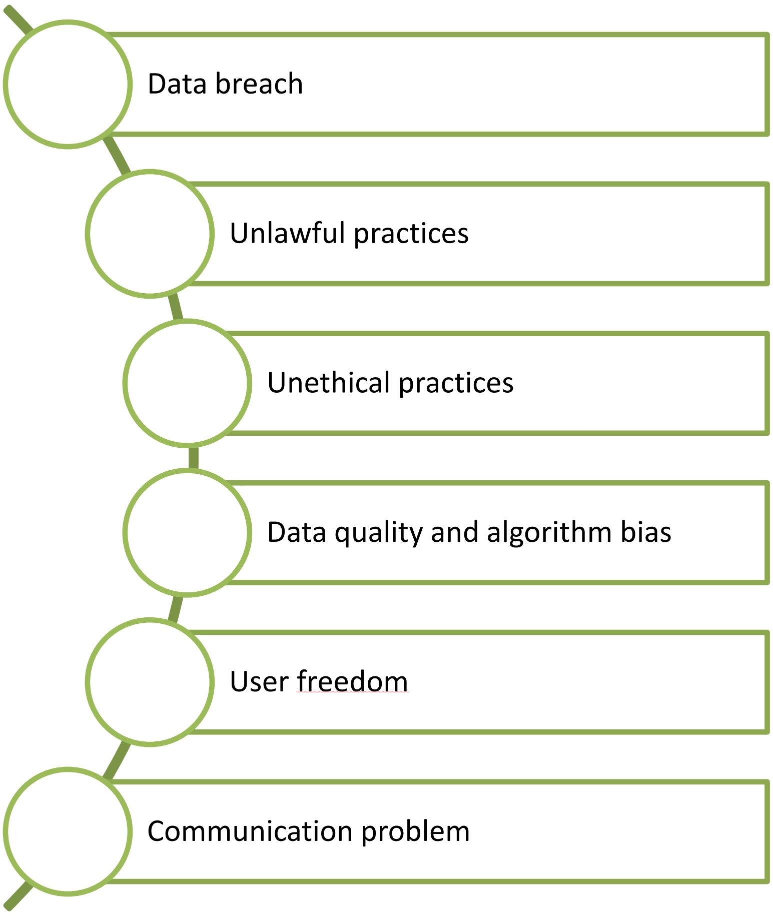


Figure 3 Issues covered by the certification

Data breach and loss: refers to the risk of violation of database containing personal and potentially sensitive data of customers and employees. It also refers to the physical integrity of the infrastructures where data are stored, including server replication, high availability and disaster recovery, plus general risks such as the physical theft of machines.

Unlawful practices: this issue takes into account potential and/or actual violations to present day laws, from a national/local level to an international level (comprising all the countries where a subject company operates or is present).

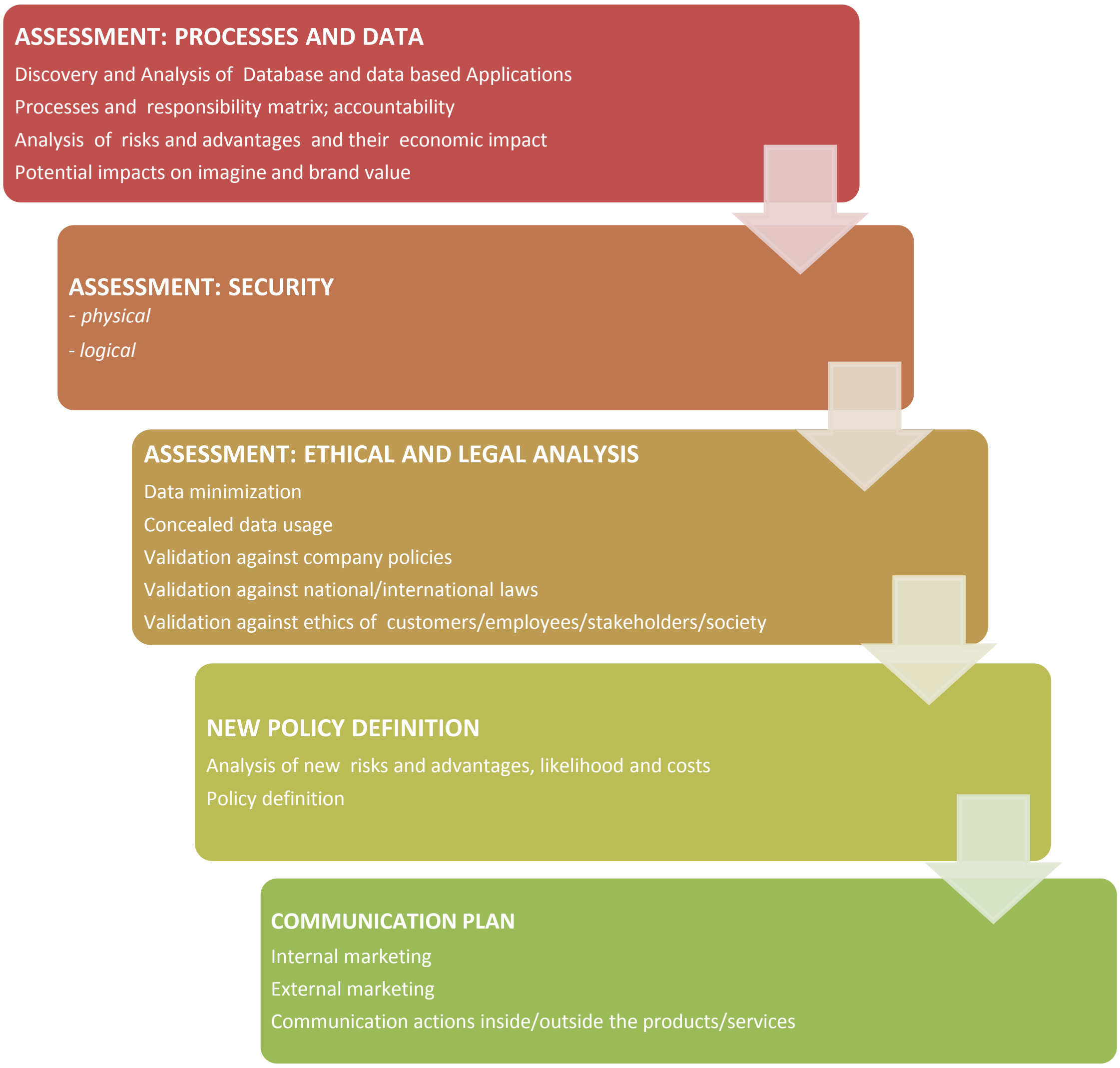
Unethical practices: this topic involves practices such as the use of sensitive data for purposes other than what stated for a specific product or service, or selling data to third parties, or the introduction of unethical biases to e.g. marketing strategies.

Data quality and algorithm bias: refers to potential issues connected with poor quality of input data, or low quality of results of the application of algorithms, potentially due to a bad choice of the algorithm itself or a narrower-than-ideal training dataset.

User freedom: it makes explicit reference to the possibility for a user to acquire and transfer their own data with ease, in a standard manner, up to exploiting the concept of data portability.

Communication problems: this is a common issue connected to the lack of effective external communication of the high ethical standards used within a company, as well as a lack of proper communication to customers regarding the use of their data.

The envisaged certification process would follow the illustrated steps:



ASSESSMENT: PROCESSES AND DATA

Discovery and Analysis of Database and data based Applications

Processes and responsibility matrix; accountability

Analysis of risks and advantages and their economic impact

Potential impacts on imagine and brand value

ASSESSMENT: SECURITY

- physical

- logical

ASSESSMENT: ETHICAL AND LEGAL ANALYSIS

Data minimization

Concealed data usage

Validation against company policies

Validation against national/international laws

Validation against ethics of customers/employees/stakeholders/society

NEW POLICY DEFINITION

Analysis of new risks and advantages, likelihood and costs

Policy definition

COMMUNICATION PLAN

Internal marketing

External marketing

Communication actions inside/outside the products/services

Figure 4 Certification process

An initial assessment, as detailed in the picture above, has the purpose to capture the situation as is and to provide a quantification of the likelihood of both risks and advantages and their economic impacts. ISO 27000 (*Information technology – Security techniques – Information security management systems*) and ISO 31000 (*Risk management*) will form the basis for this part of the process.

The following step is the definition of a new policy based on the results of the previous steps, in order to minimize risks and maximize the ethical data management of the company. This policy will act towards the definition of the company's **data management sustainability**.

The last part of the process is the definition of a communication plan, for both internal and external purposes, which will effectively promote the image of the company by providing clear information, as well as stimulate feedback from customers and stakeholders in a continuous and proactive ever-improving system.

9.2.2 Benefits

From a customer's perspective, a logo acquired after the certification of standardized procedures for the ethical management of personal data would represent a guarantee in terms of trust.

Why should a private company be interested in getting the certification on personal data management? First of all, to show and communicate its interest in respecting the law and citizens' rights while doing their business. Secondly, the company can start to consider privacy and data protection as assets that can help them to responsibly further their economic targets. An immediate benefit would be an increased reputation, with positive effects on both relationships with customers and other companies. Furthermore, the certification can represent a way for the company to periodically verify their compliance with the EU regulation.

9.2.3 Risks and mitigation measures

The main risk is a low uptake of this certification by companies, or a low renewal rate. A potential incentive may be to make it a mandatory criterion accessing specific European or national tenders. Another risk, which is relative to certifications in general, and not specific of the ethical certification here discussed, is that they are often focused on certifying processes rather than final products (ethics, in this case). This issue is often resolved by defining strict standards for the processes certified in order not to leave room for variation in the final product.

9.3 Balancing action n. 3 “Data Management Statement”

9.3.1 Description

The third proposed action, involves companies on a voluntary level. It lies on the assumption that nowadays the success of an organization increasingly depends upon the trust of shareholders, customers, employees and the general public. In order to boost the confidence of internal and external stakeholders, the organizations may submit declarations on how they collect, use, eventually sell personal data coming from customers and, in general, business activities.

The goal of this balancing action is the creation of a “Data Management Statement”, a periodical balance document describing:

- Adopted policies
- Qualitative description of data collected
- Previsions for next period

In the Data Management Statement, companies will describe, on a periodical basis, what the adopted policies are (both one-off and permanent ones), the specific actions performed in order to grant rights like data security and privacy control. An example of a one-off policy is a media campaign communicating the high standard data protection policies adopted by the enterprise. Permanent policies, on the other hand, are those affecting the core procedures and structures of the company itself, such as the creation of a specific department following data privacy matters.

All the elements emerging from this part of the document will **attest the effort to prevent, reduce or assess issues** like:

- Excessive data collection
- Risks for users’ privacy
- Third parties’ harmful/unethical use of data
- Security holes

The subscribing company will be able to put in evidence every single data management choice, increasing users’ trust and awareness.

According to this schema, this section will contain a summary of the specific **measures taken during the last period**, in addition to a **data privacy and protection policies review**: for example, companies will indicate percentage of revenue to be reinvested on privacy, transparency and users’ awareness increasing matters; explaining, in addition, what are the services devoted to supporting and informing customers and users about privacy and ethical data management issues.

This section of the statement should contain (if needed) a specific part regarding any difficulties (legal, technical, infrastructural etc.) encountered by the applicant in performing or making the chosen actions effective, with advice or suggestions in terms of European Community policies.

In this part of the statement, companies will specify **what kind of data they collect and how, in terms of what kind of personal information they take, from which source this data is taken and at what frequency**.

Another important aspect will include **data processing strategies**, such as data aggregation level or whether their datasets are joined to other datasets before/after processing or not. In this case, the secondary data sources will be indicated. In addition, a precise temporal span after which data is deleted from their servers must be specified, and the use of cryptography and anonymization techniques clearly stated. All of these will be at a detailed enough level to provide reassurance about privacy concerns (reinforcing trust in the company), but not detailed enough to unveil technical details that might undermine data safety and business secrets.

In order to encourage a programmatic approach, this Statement model will include a section dedicated to **plans for the continuous improvement of the ethical data management framework** with clear indications for the coming period. This will grant the opportunity to users, as well as to the institutions, to evaluate and properly consider the progress obtained by the subscribing company over the years. Forecasts will focus on the aspects detailed above, and will provide quantitative (and measurable) estimates on how each indicator will be improved.

In order to ensure that statements from different subjects are comparable, and thus their ethical performances, this statement will follow specific guidelines. To this end, a **standard will be defined following a discussion among different stakeholders, actors and regulating bodies at a European level**, such as:

- European institutions and committees;
- Data protection authorities;
- Employers' associations and chambers of commerce;
- Citizens' associations.

Citizens' participation in the definition of this standard may be accomplished by means of popular consultation e.g. a website where potential aspects to be included in the standard can be put forward and voted.

9.3.2 Benefits

From a citizen's perspective, the main benefit would be a higher degree of transparency regarding the use, storage and processing of their personal data, which would produce a better awareness on these topics. Companies, on the other hand, could use this statement as a mean to increase their reputation and, therefore, as a marketing tool, provided that the statement quality standards on the information included are met.

9.3.3 Risks and mitigation measures

The main risk of a statement of this kind is a limited impact, in the sense that a risk exists that it would not really improve a customer's trust in a company's practices, potentially due to a limited diffusion of the statement itself. In order to prevent it, part of the standards for the production of the statement could include the obligation to post this document on the company's website, to send it via e-mail to registered customers and to make it available in print at information points within the company's premises.

9.4 Balancing action n. 4 "European E-health Database"

9.4.1 Database description

One of the concerns about the impact of GDPR is the use of big amount of data for scientific research purposes, especially in the field of human health. This action aims at balancing the need of exploiting the huge potential of Big Data in terms of accuracy and quality of information for scientific research with the need to guarantee the highest possible security level from risks of data breach or dangerous misuse of these information.

This action involves the creation of a **European database which contains health related data of EU citizens**. At the time an EU citizen is treated in a public hospital or a private institution that receives public funding, they are asked for consent for their data to be collected and stored in an EU managed database. This consent also includes a part that authorizes that hospital to use the data then collected for the patient's treatment. Data collection and transfer will follow standard exchange protocols, such as those defined by Health Level 7[118] in similar fields, defined beforehand at a European level. Concerning the use of health data for scientific studies, in order to comply with the EU Regulation 679/2016, the person has the right to give consent only to certain areas of research.

Every precaution will be taken, using all modern anonymization techniques as new and more efficient ones are developed, to protect and effectively encrypt personal data. Encryption techniques will be standard and agreed upon by all participating EU member states.

Data access is managed differently according to the subject asking for it. Four cases are envisaged:

1. Public hospitals and private institutions that receive public funding
2. Scientists and research institutions
3. Private EU citizens (after authentication)
4. Public

The way data is accessed, and the kind of data provided, differs for each case. When an EU citizen is treated, they are asked for informed consent, and authorize that **hospital** (case 1) to access their health history directly from the database. The kind of data that is shared with that specific hospital is defined by fields that must be indicated clearly in the consent statement, so that only data that is necessary for a certain treatment can be accessed (or all, if a global authorization is given by the treated citizen).

[118] www.hl7.org

Scientist and research institutions (case 2) are required to submit an application to the EU body that manages the database. This application must include specific details on the scientist and the research team, the institution they work for, etc. in order to be identifiable. The research project this data is required for (or the application for the research project), and the financing body or institution, must be stated, together with a detailed list of the required information, a justification for each data field asked for and expected outcome. This application is assessed by the managing EU body, which will grant permission to access data or not, and the data will be provided in an appropriately anonymized way and aggregated to a level detailed enough to carry out the research.

Private EU citizens (case 3) can use their own official digital identity (e.g. the Italian Public System for Digital Identities – SPID) to access the database through a web portal to visualize, download, manage and change authorization to the use of their personal healthcare data. A system to authorize a specific hospital online beforehand for the use of certain data can be envisaged.

The **general public** (case 4) will be able to access a web portal that allows the visualization and download of aggregated Open Data. This data must be anonymous and aggregated at such a level in order to prevent de-anonymization and privacy threats. Data download will be granted through the portal itself and through web services in order to encourage public use of this data in a perspective, for instance, of identifying institutions that excel in the treatment of a condition.

Data thus collected will find **application** in many contexts:

- Healthcare data of a citizen will be easily available to hospitals that will provide a higher quality service based on a broader information base on that specific person. This will improve time to gather data and time to treat condition, thus freeing resources for other patients as well.
- E-health data acquisition time will be dramatically cut down for research. Therefore, the availability of a huge amount of data will foster an improvement of health conditions and an improved quality of life as a consequence of new findings and treatments for diseases.
- EU citizens will have a clearer picture of their own healthcare data, and they will be able to manage and share is as needed without requiring unnecessary bureaucratic steps.
- Healthcare Open Data will find a general application in the everyday decision making processes of private citizens, citizens associations and policy makers, as well as companies and firms, since specific pieces of information will be freely available. It is possible to foresee the development of services based on Healthcare Open Data.

9.4.2 Benefits

The most important benefits that follow directly from the implementation of the e-health data management system here described can be summarized as follows:

- EU citizens will have a greater control on their personal health data.
- They will be much more aware of how and when their health data is collected, and how and where it is stored and used, and by whom.

- An increased trust in health care services and research will result from the increased transparency.
- Digital identities of EU citizens will be enriched by this system and their use will be encouraged.

These aspects will coexist with an improved quality of life that is associated with lower healthcare expenses that are closely connected to a healthier population. Through the use of Open Data to describe the performances of various hospitals, furthermore, citizens will have a clearer picture of the excellence fields of each and will be able to make informed decisions relative to their healthcare.

9.4.3 Risks and mitigation measures

The main devisable risk of a centralized EU database mainly concerns matters of security and potential data breaches. More specifically, in case of leakage of information there stored, a potential for an illegal surveillance of citizens' habits and might exist, as well as a risk of sales of personal health data to e.g. insurance companies for unlawful purposes that might produce distortions in the free market. These issues are generic and relate to any centralized source of information, and they are often dealt with encryption (to make data unusable in case of leakage) and with the use of federated databases, in place of physically centralized databases, to prevent the physical aggregation of servers containing data themselves. By using these two means, the risk of data breaches can be minimized and, in case a leakage did occur, the impact following this illicit could be sensibly reduced.

As for research, one potential risk is related to the fact that health data, once an access authorization has been granted to a scientist, would reside on their computers. A mechanism to prevent the unlawful use of those data, or a use out of the purpose initially specified, could involve the liability of the scientist itself in case of an illegal event by making them responsible for the correct use and storage of data. A similar risk, and connected mitigation measures, might be envisaged in the case of the use of data by hospitals personnel.

9.5 Balancing action n. 5 "Digital education on Big Data"

9.5.1 Description

This action aims to create a broader digital culture in Europe, specifically aimed at the development of a much deeper understanding of Big Data, how it interacts with EU citizens throughout their life and how it affects each individual.

In order to promote this increased awareness, we hypothesize that a series of educational programmes aimed at the different age groups will exist and will be administered via both compulsory school years and optional courses.

Following this balancing action, the following initiatives have been adopted:

- Digital curricula for **primary school children**, and **junior high** and **high school students** that explain Big Data, what it is, how it is collected, its risks and how to prevent an excessive exposure of personal information, together with clear explanations of the pros and cons of digital identities.
- Integration of an ethical approach to the creation and use of Big Data in **university** degrees that train data scientists and similar professional figures, with the optional/compulsory sitting of ethics papers.
- Massive Open Online Courses (MOOCs) aimed at the **general population** and sponsored by the EU to provide EU citizens with the necessary tools to understand Big Data and help the development of a more comprehensive digital education. Courses will be diversified according to the required level of confidence with Big Data, and modular in order to fit with and integrate any kind of previous knowledge level (e.g. senior citizens have very different learning needs compared to middle aged citizens).
- Theme based face-to-face seminars and courses administered in person **down to the level of single municipalities**. Themes are practical and tailored to the needs of each age group (with particular attention to middle aged and senior citizens that are more vulnerable in this respect). They include topics such as how to set up an account to communicate with friends and family (e.g. VoIP services), how to use job hunting portals and professional social networks effectively, how to use Internet banking, e-healthcare services, and access to welfare system services. General concepts such as the risks connected to a lack of control over one's own personal data, the potential for scams via e-mail, websites and apps, and potential real life risks deriving from an improper use of online services are covered.

9.5.2 Benefits

A broader understanding of Big Data will enable citizens to make active use of the opportunities that this paradigm shift brings with it. More specifically, EU citizens will:

- Develop a more aware attitude to the use of Internet (and IoT) avoiding pitfalls related to the uncontrolled propagation of their data.
- An increased understanding of how Big Data and offered services are intertwined might lead to the acceptance of the necessity to surrender some personal data in order to use those services, thus leading to an increase of trust in new technologies.
- Ethics trained data scientists will concede that personal data is treated ethically, diminishing privacy breaches in the mid to long term.
- The provision of digital skills to middle aged and senior citizens, as well as people with limited access to online services, will empower them by teaching them how to e.g. effectively look for jobs in case of premature unemployment, and use social welfare online services.

9.5.3 Risks and mitigation measures

One of the risks, which is mostly limited to the case of face-to-face seminars and course, is a non-effective use of the funding resources, meaning that some of the funded education projects might not lead to an actual improvement of the digital knowledge of the target population, or a wider uptake and use of new ICT technologies. In order to prevent it, funded educational projects should only be funded in municipalities where the basics infrastructures for Internet connectivity are available.

9.6 Balancing actions impacts summary

Considered that each proposed action affects many life phases and its impacts are manifold, a table is presented below to provide a global overview.

Table 2. Identification of the life phases and the ethical issues impacted by each balancing action.

Balancing action	Life phase	Ethical issue
1	Adolescence, adulthood, middle age, old age	Awareness, control, ownership, digital identity, privacy, trust
2	All	Control, privacy, trust
3	All	Trust, awareness, privacy
4	All	Control, awareness, trust, digital identity
5	All (higher focus on infancy, adolescence, middle age and old age)	Awareness, trust, privacy, digital divide

10. Second phase - stakeholders' opinions landscape

During the second phase of the study a set of interviews were conducted and an online survey was published, with the aim of discussing the ethical issues and the balancing actions with some of the main stakeholders in Europe, acting in the public and private sectors. These discussions helped the research team to fine tune the proposed actions and to arrive at the final outcomes of the study.

10.1 Interviews

10.1.1 Stakeholders selection

The stakeholders were selected by taking into account the complexity of the debate about the ethical issues of Big Data exploitation, which involves a variety of social, economic, even philosophical aspects and fields of expertise.

After having designed the five balancing actions, the research team was very interested in discussing those actions in detail with people concretely involved in this debate, covering different points of view, and together trying to identify the most hidden implications of those actions, both as good opportunities and risks for citizens and for the protection of human rights.

In order to get a wide panorama of opinions, data protection Authorities, big data and analytics companies, SMEs, consumers' associations, representatives from EU bodies, academia, data scientists were involved.

All the people interviewed expressed their own opinion, which do not necessarily reflect the position of the public or private organization/company they work for.

According to the Study first design, and after further reflection, the research team started to contact representatives of:

- EU data protection Authorities
- Research Institutes and Centres for Statistics
- Consumers' Associations
- Universities
- EU bodies
- Analytics Big Firms
- Companies making use of Big Data Analytics

In the end, 16 stakeholders were contacted and expressed their views:

- The European Data Protection Supervisor
- The National Supervisory Authority for Personal Data Processing in Romania
- The Italian Data Protection Authority
- A Research Supervisor at the Italian Institute of Statistics (ISTAT)
- An expert at CBS, The Netherlands.
- One expert of the Eurostat Task Force Big Data
- A data scientist of ESA-ESRIN (European Space Research Institute)

- One legal expert of BEUC The European Consumer Organization
- A member of the EESC
- An expert in Network and Information Security of the European Agency for Network and Information Security (ENISA)
- A member of the Ethics Advisory Group set up by the EDPS
- A full professor of Legal Philosophy at Federico II University of Naples
- An associate professor of Legal Informatics at Bologna University.
- A sale manager at Oracle Italy
- A regional manager of Booking
- A big data independent expert

10.1.2 Topic guide

All the interviews conducted as open discussions about the main topics of the Study. However, in order to cover all the central dimensions emerged during the desk research and to collect comparable information from the respondents, the interviewers followed a common topic guide. Compared to a conversational interview free from any structure, the guided approach allows a better control of the discussion, but still leaves a degree of freedom and adaptability in getting the information from the interviewee.

Each interview consisted in two main phases: during the first one, the discussion aimed at allowing topics to emerge spontaneously, to understand which dimensions arise more often (for instance trust, virtual identities, transparency and so on) and the negative or positive connotations related to the dimensions. During the second phase, each balancing action was discussed in depth in order to understand the position of the stakeholders on the action, specifically in terms of efficiency and feasibility.

10.1.3 Information retrieval

The extraction of information from each interview followed four steps, the aim of which is the standardization of the issues expressed by each interviewee in order to compare the position of each stakeholder with that of the others.

Each interview was divided in six parts, one referring to the general discussion on the ethical aspects of Big Data, and one for each proposed balancing action. For each part, the first step involved listening to each interview and taking notes on the issues expressed by a stakeholder while discussing each point of the topic guide. An example of an issue is “Need for clearer consent statements” or, in the case of balancing actions, “Centralization poses too many risks”.

Once issues were identified, a second step involving a second round of listening was carried out for each part of each interview in order to count the number of times each issue was brought up and discussed by the interviewee. Issues were counted only when a stakeholder expressed a complete point of view about it.

Issues were then standardized across stakeholders during a third step. Issues belonging to different stakeholders that identified similar points of view (e.g. "Centralization poses too many risks" for stakeholder 1 and "Central data storage is not safe" for stakeholder 2) were merged in a new issue with a new name that is unique across all stakeholders. The final point of this step was the production of a matrix for each part of the interviews where each row is a stakeholder, columns are the issues and cells contain the frequency at which a specific issue was discussed by a specific stakeholder.

During the fourth step, a mathematical technique called Principal Component Analysis (PCA) was applied. This technique allows, among others, to produce a chart for the general discussion and a chart for each balancing action where stakeholders are represented by symbols associated with the interest they express towards Big Data. The chosen stakeholder categories are **Academic**, **Company manager**, **Consumer representative**, **Data specialist**, **National data protection authority** and **EU body expert**. PCA provides a numerical quantification of the importance of each issue along the negative and the positive halves of the axes of the chart. For each of the halves of both the horizontal and the vertical axes, the issue that had the highest importance score was chosen as a starting point to provide a name to that half axis. When more than one issue scored similarly high, they were all used in characterizing the half axis they referred to.

As an example, for the chart on the general attitude towards the ethical aspects of Big Data, for the negative half of the vertical axis, "Privacy as asset" was the most important topic (with a score of 0.32), and "GDPR impacts privacy" and "Surveillance" followed (with a score of 0.25 and 0.24 respectively). All the other issues lagged behind with much lower scores. The chosen name for that half axis, therefore, was "Privacy concerns", as it sums up the more detailed positions expressed by the issues and provides an explanation of the attitude towards the overarching topic of Big Data and ethics. The same approach was used to name each half axis for each chart.

For the interpretation of the charts, the closer two stakeholders are in a chart, the more similar their attitude is. Similarly, the farther away from the origin of the axes they stand, the more extreme their attitude is. The overall attitude of each stakeholder is inferred by their position relative to the axes. The attitude emerging from each chart was interpreted together with the qualitative knowledge on the position expressed by each stakeholder during the interview, and charts are provided in this study as a graphical support to convey quantitative information more effectively.

10.2 Survey

10.2.1 Questionnaire

The online survey published, and kept open for one month, aimed at gathering the opinions of both big analytics firms and data-driven companies such as big retailers on the web, or utility providers. The survey questions have been formulated to follow the same themes of the interview topic guide, in view of getting comparable results.

The questionnaire was distributed among about 200 companies, and 21 filled-out forms were returned. The respondents stayed anonymous but they were asked to provide the country where they live, their company name and their role as compulsory information.

The questionnaire contained additional resources for those respondents who were not familiar with the topic, such as a Glossary for the most technical terms and a brief video to contextualize the topic.

The first block of questions was related to Big Data ethical issues and the respondents were called on to select what they considered to be the three most crucial matters. Then the questionnaire asked the respondents to select from a list of options, the main risk and the main benefit for people in the data-driven society, the main risk and opportunity for companies in business.

The second block of questions were about the GDPR, aimed at finding out the familiarity of the respondents with the new Regulation and exploring their opinions of it.

The last block was related to the five balancing actions proposed in the study. Respondents were asked to give feedback about these possible solutions in terms of efficacy and feasibility.

10.2.2 Information retrieval

Answers to the online survey were collated in a matrix where respondents were rows and questions were columns. At the intersection of each respondent and each question, their reply to that specific question was present. The frequency at which each answer was given was then expressed as a percentage and results reported as bar graphs or pie charts.

A discussion of the results is provided within each paragraph relative to the general attitude towards ethics and Big Data and each balancing action (sections from 7.3.1 to 7.3.6). Charts were collated in the form of a focus section at the end of section 7.3.

10.3 Results and conclusions of interviews and survey

10.3.1 General attitude

As a general remark, almost all of the people interviewed expressed more worries than optimism about the current Big Data scenario.

One of the most urgent issues identified by almost all the stakeholders interviewed is the lack of awareness of the users. At the moment, the situation seems to be like a “perfect storm” that makes people increasingly vulnerable: from one side, there is a strong social pressure to use the new technologies to manage their own social relationship, to take decisions in everyday life. Deciding to take a step back in using these services does not seem to be a real alternative in the digital society.

The services are already there, all available apparently for free, or at least this is what people generally think. And this is another aspect that was clearly pointed out during the interviews: data subjects are not aware about the value of their data, there is not enough reflection on the fact that personal data is the currency to pay for the "free of charge" services that they use on a daily basis.
Closely linked to awareness is also the issue of control over both the personal data people give in exchange for a service and the self-generated data (for instance by using a sport tracker or an app that monitors health). It is not only a problem of lack of control but also of not paying attention, not even having the feeling of losing control.

The transparency was also a very important issue: digital services are perceived by the stakeholders as "closed boxes" where the data subjects cannot see anything. And the actual system of getting users' consent is not helping the data subjects to understand the purposes of the processing of their personal data.

There is also a clear concern about the security of data and risk of breach, which has already occurred to some big digital players.

The technologies already used to guarantee data protection (privacy enabling technology) like anonymization techniques, do not seem strong enough to guarantee the safety of people's more personal and even sensitive data; almost all the stakeholders mention the concept of de-anonymization risks, as results of datasets combinations.

Another often mentioned risk is the so called "filter bubble" and the development of a "tailored reality", even though in this debate two different positions emerge. One stakeholder from academia thinks that this is a risk that pre-existed digital devices, because it is strongly connected to the need of human beings to select the information according to their values and certainties. Most of the other stakeholders clearly disagree with this point of view and are instead convinced that digital services like internet search engines or online shopping services, massively increased this natural attitude, which has become much stronger than in the past. Moreover, while in previous times people at least decided autonomously which information source to use, today people still have the illusion of having free choice.

Connected to this issue is the virtual identity topic. The data that digital aggregators get from the person, by detecting choices, preferences, the behaviour on the net, and in general collecting data from a potentially huge number of sources (wearables, sensors, etc.), builds a virtual identity that can produce effects in the real world in terms of job opportunities, credit rating, risk profile for insurance companies, etc. Moreover, the virtual identity can have an impact on the self-perception of an individual, especially in case of adolescents or youngsters, and the risk is that the two identities become too narrow and indistinguishable. And if a person wants to change their digital identity because they feel the distance from the real personality, he or she does not have the tools to do that.
The digital divide did not emerge often from the interviews. It was rather pointed out by the interviewers. When the discussion came to this topic, stakeholders seemed more concerned about the "digital illiteracy" than the lack of connection.

When it comes to data portability, some stakeholders are rather sceptical about these solutions and many of them express concerns that this can be only one single thing that can help only informed people. The idea is that you cannot give people these kinds of tools if they are not even aware that there is a problem of privacy violation. Besides, for data portability many people see difficulties in the technical application.

The interviews were also a good opportunity to see the reactions of some key people to the GDPR adopted in May 2016. Most of the opinions are in favour of the Regulation, but stress the fact that this is only a first step and many implementation issues will certainly occur because in some parts the text is rather vague and leaves the Member States the freedom to decide how to apply certain rules, as for instance regards the role of the Data Protection Officer.

Some stakeholders from academia and EU entities expressed worries about the fact that consumers are at the centre of the debate because of what they know, what they do not know and want, however their voice is absolutely not heard in this debate. It is a closed dialogue between “insiders”. But in order to play an active role the stakeholders stressed the fact that massive educational programs must be organized, starting from the very early stages of education systems.

As mentioned before, the benefits were less discussed by the interviewees. However, they all recognized the great value of Big Data in many different fields, such as better medical treatments and earlier diagnosis, socio-economic, energy and environmental modelling, more comprehensive and better-informed policy making, unprecedented contribution to official statistics, economic growth boosted by innovative business models.

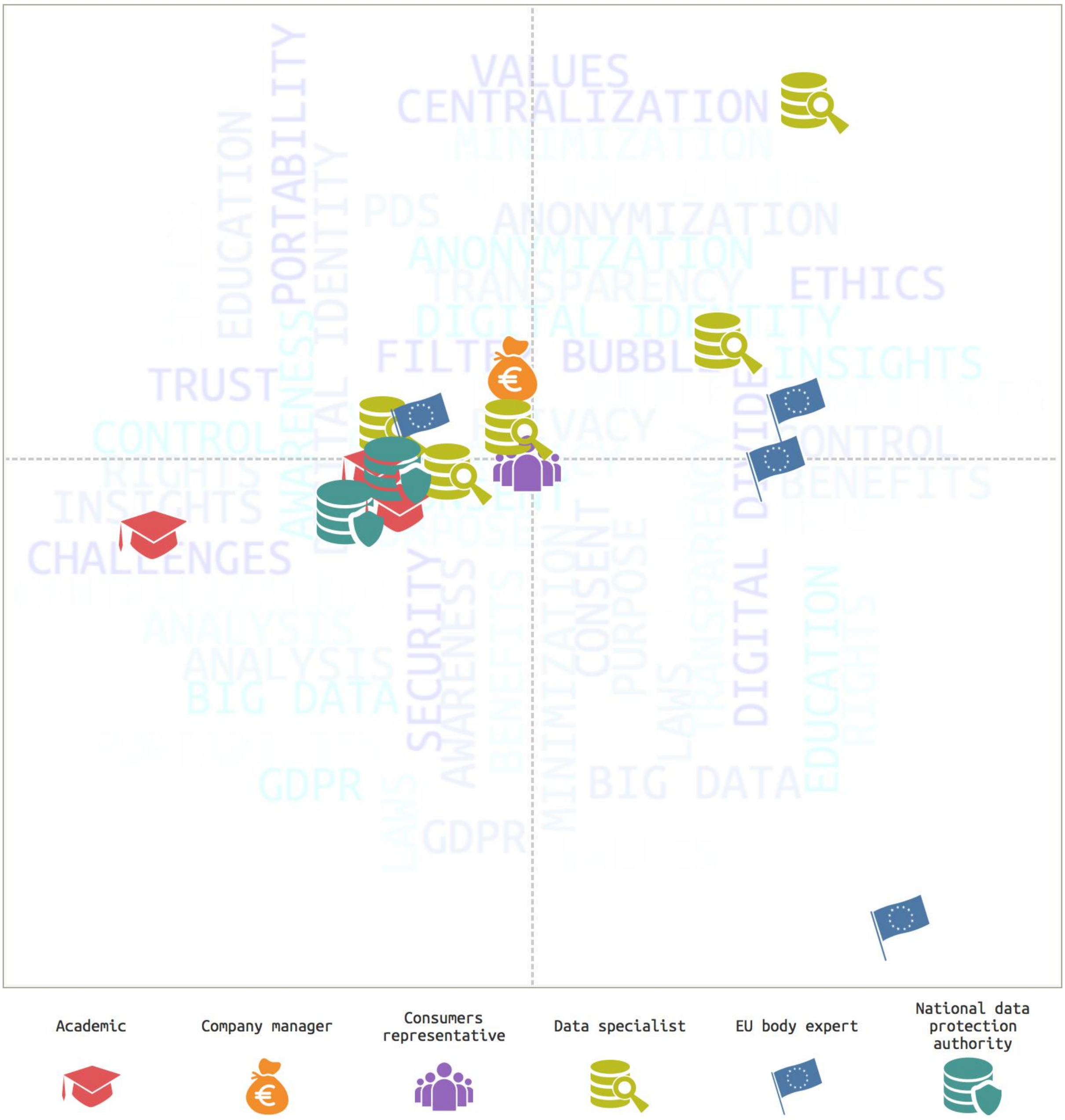


Figure 5. Stakeholders' general attitude towards the ethical aspects of Big Data

The analysis of topic frequency, as detailed in section 7.1.3 above, highlighted that most stakeholders cannot be grouped according to the category they belong to. They expressed opinions that were more on a personal level, rather than aiming to put forward the interests of their respective group. For instance, only academics and national data protection authorities fall closely clustered within the same small portion of the chart above, while EU body experts, data specialists, consumers' representatives and company managers are roughly homogenously scattered throughout the chart itself. This means that academics and national data protection authorities tend to describe their position towards Big Data and ethics more similarly compared to what the other stakeholders do.

Academics and **national data protection authorities** are mainly concerned with topics such as a general need for an **ethical use of data** and **transparency**, they give voice to the issues connected with an **excessive personalization of the digital experience** (the so-called filter bubble), and its interplay with the digital representation of human beings, which may be biased. They also indicated the **loss of control** over one's own personal data as a **threat to privacy**.

Data specialists did not provide a unique point of view on the debate, but expressed two distinct attitudes. In fact, they do not form a single group, but are arranged in two smaller clusters, one in the upper right side of the chart itself, and one which is part of a large group on the immediate left of the origin of the axes. The group of two in the upper right hand side mostly indicated the **potential of Big Data** to generate new insights, although **analytical techniques** are presently lagging behind, and they were aware of the fact that a need for increased digital education is present, and that companies should not be too strictly burdened with requirements when processing Big Data. They have an optimistic outlook on the future, as they expressed the idea that **general awareness of EU citizens on this topic is increasing**. The group of three expressed general ethical concerns, shared by one EU body expert, a consumers' representative and a company manager.

Of the remaining three stakeholders, who are all **EU body experts**, two grouped together on the right-hand side of the chart, close to the horizontal axis. They agreed on the fact, and were mostly concerned with the idea, that **consent mechanisms** are presently not up to the task of dealing with a Big Data scenario, and that new paradigms must be developed to grant a full exploitation of the potential of Big Data in a context of respect for **privacy** and protection from **de-anonymization. The EU body expert** in the lower right hand side of the chart was deeply concerned with stressing how **privacy is a value** and should be considered as an asset by companies, how **surveillance** is increasing with the ever-increasing diffusion of Big Data technologies. This stakeholder also voiced a positive opinion on how the GDPR will increase privacy protection.

Survey results indicate that a large majority of respondents were aware of the current debate on the challenges and opportunities of Big Data, and their foremost ethical issues were privacy, data protection and transparency. At a personal level, they were mostly worried that a data-driven society may pose a risk of loss of control over their own personal data, although they appreciated smart solutions as the main benefit for people. As for the benefits for companies, respondents thought most of the issues were connected to the quality of data collected, and a slight majority identified a potential benefit in the possibility to offer new and personalized services to customers. As it emerged, less than a quarter of the respondents were familiar with the new General Data Protection Regulation, and, among them, the vast majority tended to agree with statements that offered a positive interpretation of the impact of the GDPR on both companies and people.

10.3.2 Balancing action 1

The attitude towards balancing action number 1 is generally **rather positive**. It is seen as an action that can make service providers' data policies more transparent and increase people's awareness of which company holds their data in exchange for which "free" services.

On the other hand, this centralized solution raises concerns about **security and accessibility**. The preferred option is to provide a database containing only **metadata** and not the real data of people, to avoid duplications of datasets and further risks of data disclosure. Most of the people heard are convinced that only if this European platform is joined on a **voluntary basis** by private companies and institutions does it have a chance of success and of becoming a useful tool.

Another mentioned issue was how to be sure that the data holders publish correct and complete information on the data they collect. One interesting suggestion was to integrate machine-readable **icons** in the platform, as mentioned by the GDPR (art. 12), in order to guarantee a transparent way to inform people about the processing of their personal data. In the long term, the platform can become an informative hub of privacy policies in Europe, to promote and disseminate new initiatives, best practices of European companies, and so on.

The concrete realization of this measure is not conceivable without a strong commitment from European Institutions, which, through ad hoc communication campaigns and networking, can promote this opportunity among the citizens and recognize the commitment of the public and private actors that accept to share information about the personal data they manage. Only the joint attention of EU Institutions and citizens can raise the interest of the big players in joining the platform, and trigger a virtuous competition in the field of transparency.

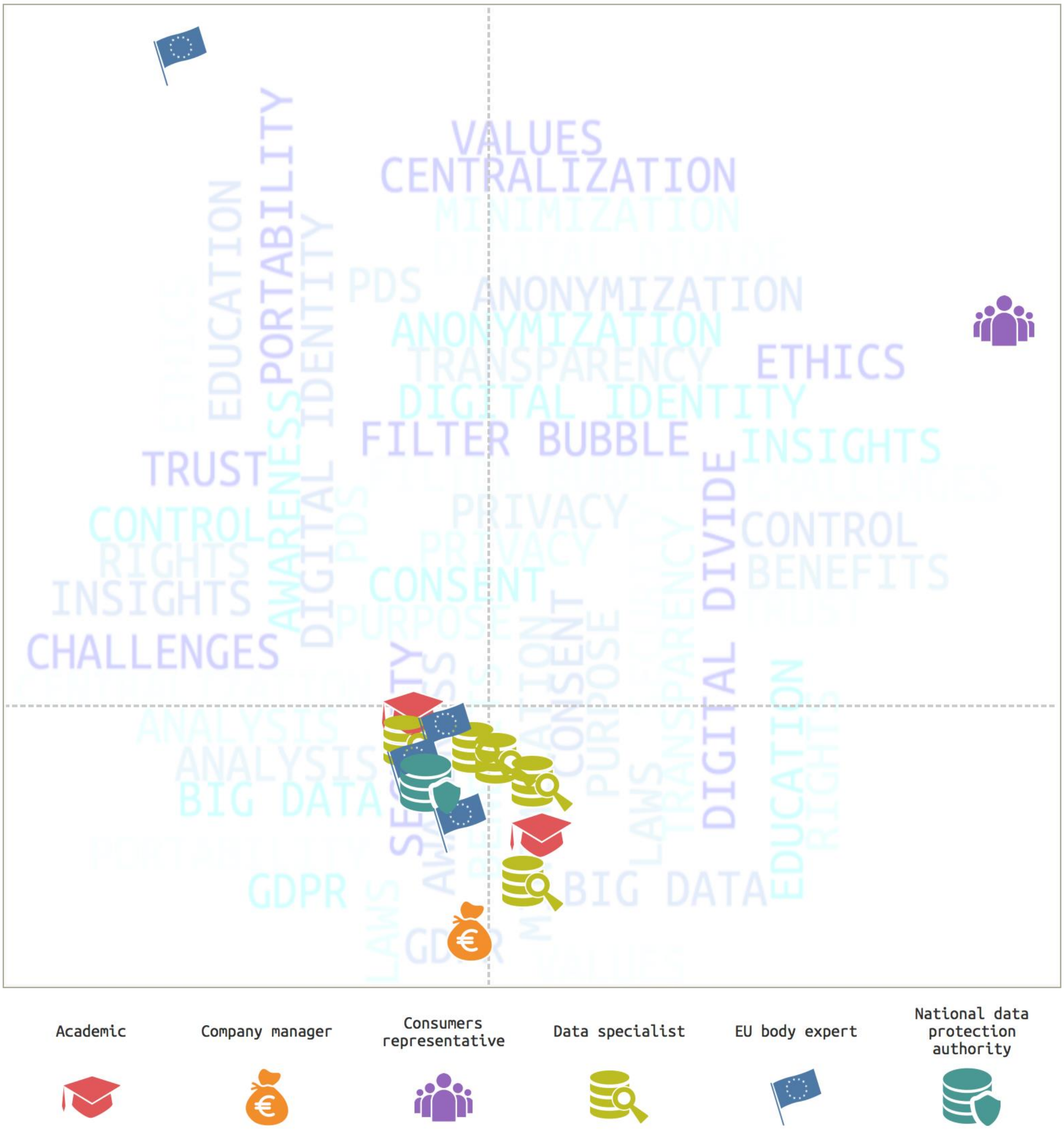


Figure 6. Stakeholders' attitude towards Balancing Action 1 – EU Data Management Portal

The attitude of stakeholders towards Balancing Action 1 was pretty much the same for all the interviewees, as shown by the clustering of almost all of them in one group at the lower side of the chart itself. Regardless of the group each stakeholder belongs to, they all agreed that an EU Data Management Portal could **give control over one's own data** back to citizens to a certain extent. Nevertheless, they all stated that it would be really hard to control the **reliability of the information provided** by companies, thus potentially causing this portal to fail its purpose.

One of the **EU body experts** had a very different opinion on Balancing Action 1, as they positively noted it could act as an **information hub for EU bodies** to monitor companies' behaviour, although they thought it would be **hard to ensure that both companies and citizens would subscribe**, and **security issues** would be a hindrance to encourage uptake of the portal itself by the population.

The **consumers' representative** too, had an opinion which separated from the main group, and perceived a EU Data Management Portal as a good chance to **increase transparency**, provided it also gave users the opportunity to **change their own personal data** through it, and provided that it was integrated with other initiatives, such as the use of embedded **machine readable icons**.

The survey indicated that respondents thought that an EU Data Management portal should be mandatory in order to be effective, and they thought that this action is feasible in the current scenario and could increase awareness.

10.3.3 Balancing action 2

This action gave rise to quite divergent opinions. Many stakeholders interviewed think that action can represent a good starting point to improve transparency and support people in identifying the most virtuous companies in this matter. However, it would only work if implemented together with an effective sound certification process designed by a central Authority and if independent bodies can guarantee the achievement of the certification only by companies that are fully compliant with the data protection law.

Some concerns were expressed by the persons interviewed, which underlined how ISO certifications seem more focused on processes and less on results and that a variety of trademarks already exist that do not necessarily guarantee the commitment of the entities and the full compliance to a given standard.

During the interviews, the research team presented the idea that the certification can also be an opportunity to support controllers and processors in concretely and correctly applying some of the provisions of the GDPR. The legal text contains quite general indications on how to implement the provisions and the Member States are rather free in the implementation.

Considering that different interpretations of the regulation might lead to different levels of privacy, a comprehensive certification process would act as a guideline for companies, public authorities and bodies to standardize the way such requirements are addressed. This seems particularly useful in determining the communication means and protocols to the data subjects, and in particular to children, that the controllers must put in place to guarantee transparency in personal data gathering and processing and to allow data subjects to fully exercise their rights. Furthermore, the certification may contain a guidance for controllers to identify the appropriate measures to implement data-protection principles both at the time of the determination of the means for processing and at the time of the processing itself (principle of "privacy by design") and the measures to ensure that, by default, only personal data which is necessary for each specific purpose of the processing is processed (principle of "privacy by default").

Finally, the certification may support the controllers when the data protection impact assessment is required to measure and mitigate the high risk to the rights and freedoms of individuals.

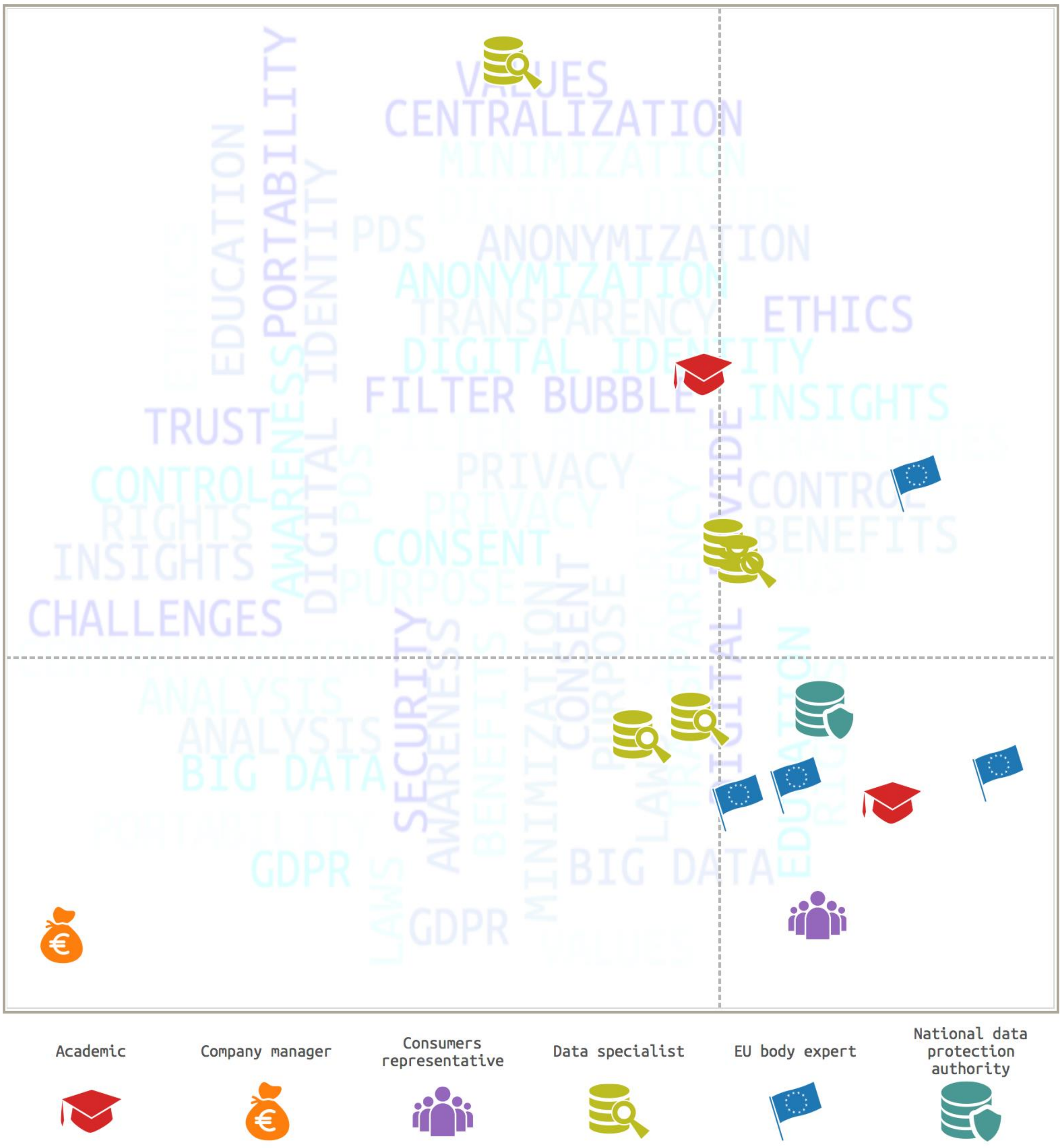


Figure 7. Stakeholders' attitude towards Balancing Action 2 – EU Ethical Certification

The attitude of stakeholders towards Balancing Action 2 as reported in the chart above seemed to be correlated with the group they belong to in the case of data specialists, who cluster close to the origin of the axes, and in the case of EU body experts, who are, except for one, located in the lower right corner.

Data specialists had a very mild interest in the certification and thought it could be an **added burden for companie**s, for a **slight gain in increased awareness and control** over one's own data. These topics were only mentioned and not elaborated in depth. As for **EU body experts** and a **national data protection authority**, they insisted that it would be crucial to **assess costs and benefits** of this certification in order for it to succeed and that a **higher degree of standardization** would be needed to function properly.

One **data specialist** who is set apart from the main group and is located at the upper end of the chart repeated several times that this kind of certification should be **totally optional** because it is a **burden** for companies to undergo certification processes. This topic was also prominent in the opinion expressed by the **academic** who sits alone in the upper half of the chart.

Differently from the other stakeholders interviewed, the only **company manager** present did not express the opinion that the idea of this certification as an added burden and his attitude is best described as correlated to the idea that it would be **a good tool to increase customers' awareness** on the use of their data.

The majority of the respondents agreed with statements concerning the effectiveness of an ethical certification and regarded it as feasible in the present-day scenario.

10.3.4 Balancing action 3

The idea of promoting ethical statements for companies is considered interesting but with a limited efficacy, especially if not accompanied by other kinds of measures. Many stakeholders though express the idea that showing a clear commitment in the field of data protection would be a big reputational benefit for companies that process big amounts of data to provide their services and products.

This statement would also be part of the CSR strategies of private companies, especially in terms of social accountability and the protection of fundamental human rights. However, stakeholders indicated that this should not be used to show the compliance with the legal requirement, which should be taken for granted, especially once the new Regulation is fully applied from May 2018. Instead, it should reveal the efforts in going beyond the legal requirements, through innovative approaches and solutions.

One more important aspect was raised while the interviews were being conducted. It was suggested that the personal data policies of private companies as an ethical indicator in Socially Responsible Investments be included. The term Socially Responsible Investment (SRI) refers to the incorporation of Environmental, Social and Governance (ESG) issues as well as criteria linked to a values-based approach into the investment strategies. Without compromising the profitability of the investment, the responsible investor seeks to create long-term value, supporting businesses that can bring concrete benefits to society.

In particular, social criteria examine how a company manages relationships with its employees, suppliers, customers and the communities where it operates.

Both institutional and individual investors can become socially responsible, even though SRI fund growth is driven by major institutional investors so far. SRI strategies are growing in the European asset management market and there is room to explore further indicators of ethical and responsible business.

Among the different kinds of SRI approaches, there is the norms-based screening which allows investors to assess the degree to which each company in their portfolios respects issues that impact Environmental, Social and Governance criteria by adhering to global norms on environmental protection, human rights, labour standards and anti-corruption. In this context, the compliance with the GDPR main requirements may represent a crucial criterion to include companies in ethical portfolios.

Possible stakeholders to involve in this debate should be the companies that manage ethical investment funds and ESG rating agencies, that may be interested in enriching the set of standards against which to evaluate companies.

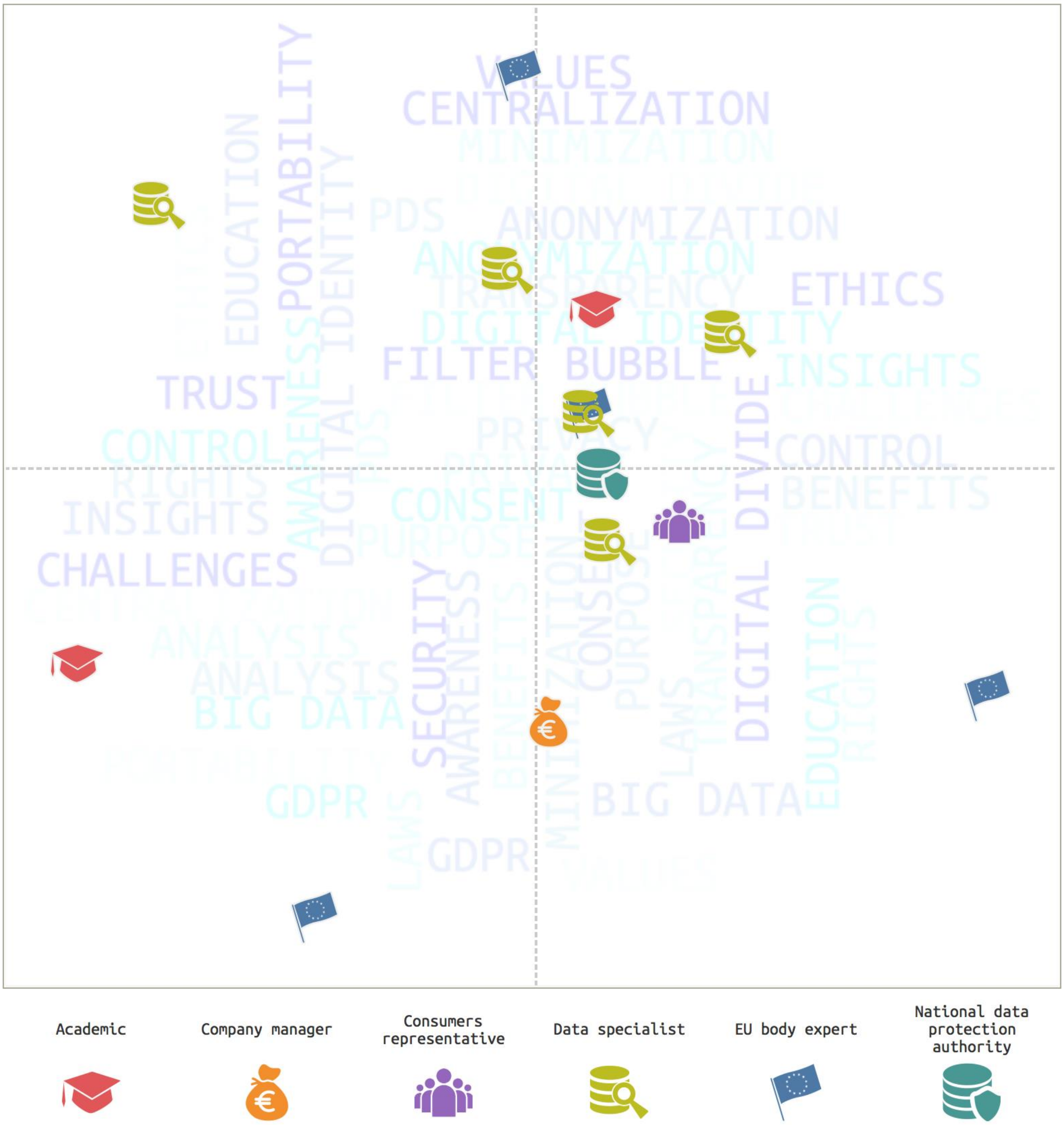


Figure 8. Stakeholders' attitude towards Balancing Action 3 – Ethical Statement

Attitudes towards Balancing Action 3, as emerged from the Principal Components Analysis and are reported in the chart above, varied greatly among stakeholders and no clustering related to the group of each stakeholder is clearly evidenced in the chart.

One of the EU bodies experts (lower right side of the chart) insisted that a statement such as that proposed as part of this action should be closely **linked to Corporate Social Responsibility**. This idea is shared, although to a lesser extent, by a **data specialist**, by a **national data protection authority** and by **the EU consumers' representative** (all in the lower right quadrant close to the origin of the axes), but did not appear foremost and relevant in the opinions expressed by **another EU body expert** (lower left side of the chart), who was more interested in the use of the ethical statement as a tool to **increase transparency and awareness**, and in the opinion of an **academic** (left side of the chart), who was concerned with **security issues** and conceded that an **increase in awareness** might be a consequence of the implementation of this statement. Another EU bodies expert (upper side of the chart) noted that customers might not rely on the information provided on the statement itself due to a lack of trust towards large companies.

Balancing Action 3 was perceived optimistically among those who replied to the survey and they thought that it could promote data protection policies among companies and it was regarded as feasible.

10.3.5 Balancing action 4

This is the action that gives rise to a number of concerns. Even recognizing the huge potential benefits in terms of better analytics at the disposal of scientific research and optimization of therapies, the majority of the stakeholders heard expressed worries on centralized solutions like this one that would imply a duplication of personal data and would expose sensitive information to further risks of data breach.

The general attitude expressed is toward a stronger investment of the European countries in a stronger standardization of their own data centres to guarantee the possibility of comparing data at least at a national level. At a European level, the research institutes could share anonymized results and information-, with limited possibilities of access to personal data only if this is indispensable and only under strict security protocols.

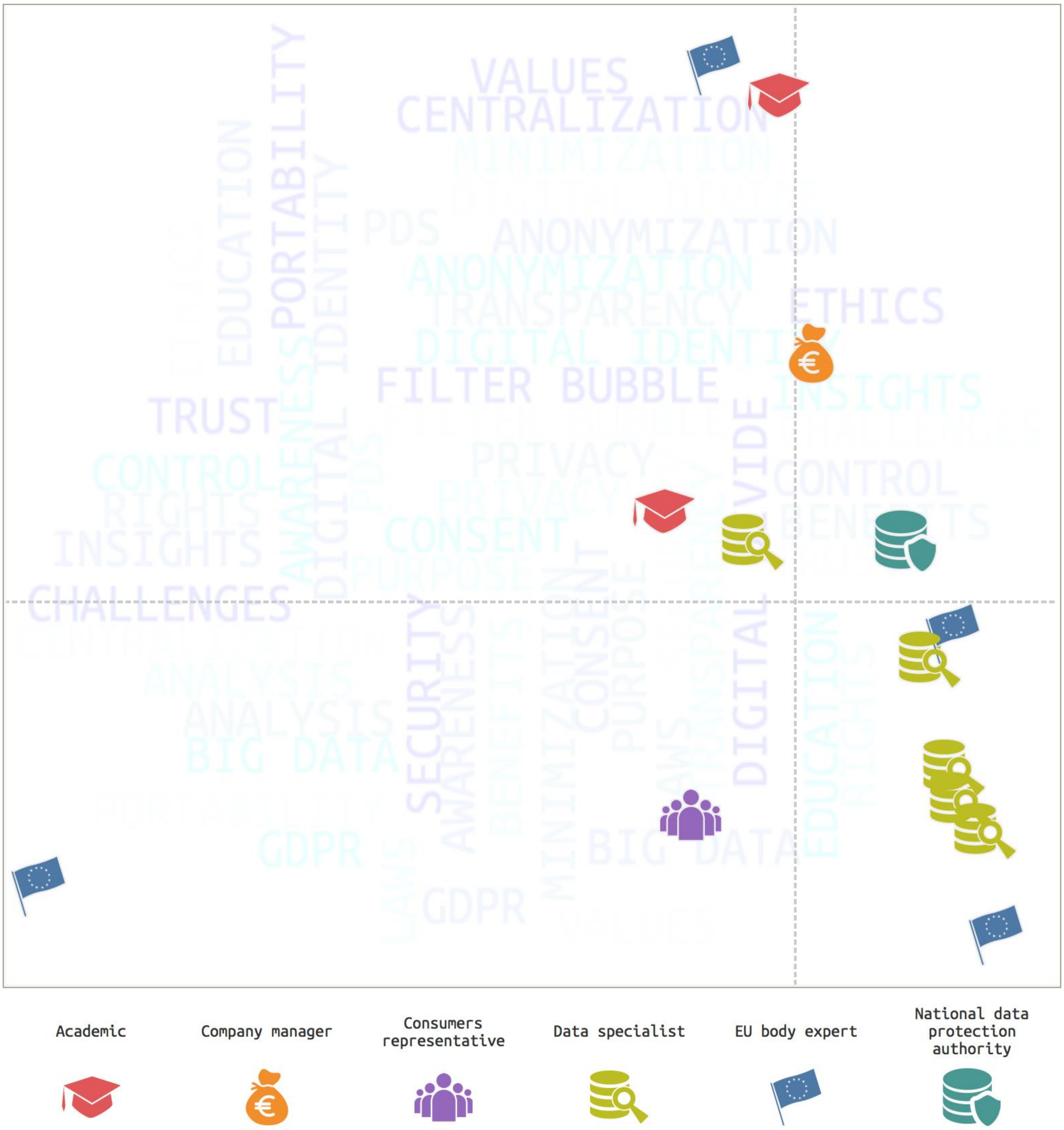


Figure 9. Stakeholders' attitude towards Balancing Action 4 – EU e-Health Database

The stakeholders' attitude towards Balancing Action 4, as emerged from the Principal Component Analysis and is shown in the chart above, was mostly characterized for all by a strong **objection to the idea of a centralized system** for the collection and storage of healthcare data. The stakeholders that mostly expressed this concern are those located in the lower right corner of the chart and include **four data specialists** and **two EU body experts**. A **national data protection authority** expressed a similar concern with the centralization of data, and was even more worried by the potential **security issues** connected with this action. **Five other stakeholders** (two academics, one data specialist, one company manager and one EU bodies expert located in the upper half of the chart) also expressed the idea that **security issues**, specifically the risk of **de-anonymization** of sensitive data, would be closely linked to a centralized system as that described as part of Balancing Action 5.

Only one stakeholder, namely **one EU bodies expert** (lower left side of the chart), expressed a positive attitude towards the idea of a European health database, suggesting that it could represent an application of the **principle of data portability** and greatly improve the **quality of research**.

The survey indicated that respondents thought a European e-Health Database could help in taking advantage of Big Data in healthcare and scientific research, although they did not regard it as particularly feasible at present.

10.3.6 Balancing action 5

The digital education programme is the action that convinces the stakeholders interviewed the most. It has been indicated that the introduction of digital education in curricula of all educational levels, from primary schools to universities, can play an essential role in empowering the future generations in the information society.

It has been highlighted, in particular, that there is a need to raise the awareness of younger generations to the intrinsic value of personal data, the need to protect one's privacy, as the other fundamental human rights, the benefits and risks of virtual identities and how they can interfere with the real life of a person.

In this context, the experts heard expressed more worries on the poor digital literacy of connected users, especially youngsters, than on the digital divide affecting people that cannot access the digital services because of their age, financial resources or because they live in remote areas. In other words, the risks connected to the unfair use of Big Data and the arising ethical matters are perceived as much more urgent and dangerous than the risk of isolation and discrimination that can be experienced unconnected people.

Particular attention is devoted to the education of future statisticians, data scientists and Artificial Intelligence Engineers that should be trained to apply an ethical approach when collecting and processing personal data. This means, for instance, the concrete implementation of privacy by design and by default principles and the systematic use of the more advanced Privacy-Enhancing Technologies. Furthermore, taking into account the increasing importance of algorithms in everyday life, the future Big Data scientists should be aware of the values, biases and potential discriminations and prejudices embedded in the algorithms they design, especially when they are used to predict human behaviours and have a relevant impact on people's lives.

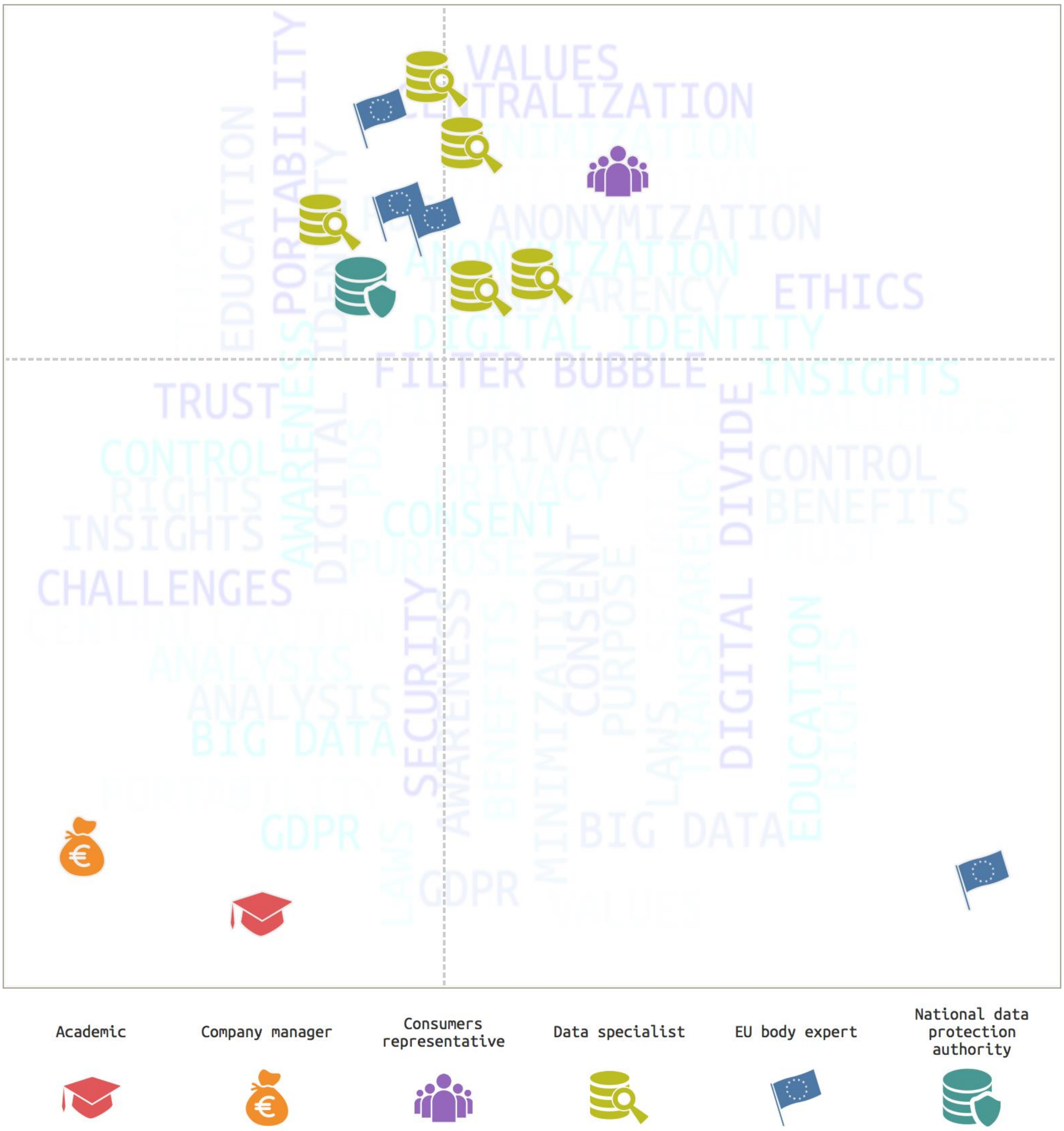


Figure 10. Stakeholders' attitude towards Balancing Action 5 – Digital Education

Almost all of the stakeholders interviewed (grouped in the upper half of the chart) expressed a **general positive attitude** towards the idea of a comprehensive scheme for the development of digital education in Europe, and they all agreed on the idea that general benefits for the population would follow this action. Two stakeholders (one **company manager** and one **academic**, lower left side of the chart) also noted that this action could specifically generate an **increase of awareness**, while one **EU bodies expert** felt strongly that this action should **start as early as possible** (e.g. primary school) and that it could help in **bridging the digital divide**.

Balancing Action 5 emerged in the survey as being globally regarded as positive, with the vast majority of the respondents thinking that digital education would be effective in raising awareness among both the general public and practitioners, and thought it was feasible in the current scenario.

10.3.7 Cross comparison on feasibility criteria

The following table shows an assessment of three feasibility criteria (namely magnitude of impact, costs and degree of support) across the five proposed balancing actions, as emerging from the interviews with the relevant stakeholders and the survey results.

The three criteria are:

- Magnitude of impact, referring to the breadth and depth of the expected outcomes;
- Costs, as a general hypothesis of the public economic investments needed to set up the action;
- Degree of support, as expressed by the stakeholders.

Balancing Action	Magnitude of impact	Costs	Degree of support
EU privacy management platform	Medium	High	Low
Ethical Data Management Protocol (ED-MaP)	Medium to High	Medium	High
Data Management Statement	Low	Low	Medium
European E-health Database	Very high	Very high	Very low
Digital education on Big Data	Very high	High	Very high

11. General conclusions

Taking stock of this study is not an easy task, because the general framework emerged is characterized by a lot of interlinked issues, and potential scenarios designed from different perspectives, interests and experiences.

As the European Data Protection Supervisor stated during the interview, it’s worth considering that the debate is anticipating the technological development and we’re still not experiencing all the possible effects of the big data since this scenario is still in its infancy. Therefore, we are still speculating on the ultimate impact at social, economic and environmental level.

The main objective of this Study was to find out how to balance the human values that are fundamental for the European civil society, like privacy, confidentiality, transparency, identity and free choice with the compelling uses of big data for economic gains. In order to reach this ambitious goal and to figure out feasible balancing solutions, the research team started an exploration in different disciplines that had already addressed this topic, examining the technical, legal, philosophical, moral, sociological and even psychological perspective toward Big Data.

It was immediately clear that this topic and its related concerns have their roots in the complex and ambiguous relationship between human beings and artificial intelligence and the overwhelming use of computing in vital sectors of human organization such as health, education, security and social networking.

This fascinating exploration sheds a new light on traditional human traits such as trust, ownership, identity, reality, self-realization that in the digital environment appear more nuanced and complex than how they are experienced in the “real” world.

The study also made evident how being exposed to the influence of data analytics can be a lifelong experience for individuals that, nonetheless, still have little awareness of how their data are used to predict their behaviour and shape their virtual identity. This knowledge asymmetry makes individuals vulnerable, with limited resources to fully exercise their fundamental rights and freedoms.

The elements gathered at the end of this study provide a framework of the policies that have the best chances of being implemented in the short and medium term and that can have the most relevant impact on society.

The need to empower people and to raise the general understanding of the dynamics, interests and values affected in the use of personal data is something that has been clearly recognised in literature and in formal statements of relevant official entities and experts in Europe. This view has been confirmed by almost all the interviewees.

The discussions with the experts made also clear that the investment on education and awareness raising represents the core element that would enable the other policies because it generates a bottom-up demand for transparency and fairness emerging directly from citizens. This demand from citizens and consumers, can’t be disregarded by the data driven market.

Moving the persons at the centre of the debate seems to be the most urgent and necessary task, taking also into account that Europe is investing in the innovative landscape of the data portability solutions, such as the Personal Information Management Systems to transform the current provider centric system into a human centric system[119].

Europe could work together to identify common contents that could be introduced in the education curricula. This can cover primary, secondary, post-secondary education, and even life-long learning initiatives. This policy could run in parallel and consistently with the actions that the European Commission are carrying out in the field of education, as the new Digital Skills and Jobs Coalition initiative that calls for concrete measures to bring digital skills to all levels of education and training to succeed in the digital world. Besides promoting the acquisition of technical skills in view of improving employability and competitiveness among European citizens, specific training could be designed to instruct people on privacy as a value and right, ethical issues of behaviour profiling, virtual identity related risks and digital reputation control, ownership of personal contents, digital footprints, intellectual property rights.

More specific integration of ethics principles and requirements for quality and integrity in research, could be foreseen for bachelor’s and master’s degree programmes in Statistics, Informatics, Data Science, Computer Science, Artificial Intelligence and correlated subjects.

On the other hand, the proposed idea (Balancing Action n.1) of setting up a European Portal where collecting information on how the personal data of European citizens are stored and processed doesn’t seem feasible and even that useful in the current reality.

As for the two actions to promote the commitment of large companies and organizations, the EU and its Members States could explore new ways of recognising the efforts and commitment of those companies which invest in ethical relationship with customers.

In this sense, the next step could be to open as soon as possible the procedure of designing a European certification system that can support the companies in complying with the GDPR and help people recognise the service providers that guarantee a fair, legal and transparent process of personal data. This procedure could take place in the context of a wider consultation, with the involvement of stakeholders that are expected to be most affected by the implementation of the Regulation itself, such as the analytics companies and big firms that make use of personal data to offer their services.

Together with this action, governments could promote among companies and organizations that make use of personal data, ways to show and communicate to customers their commitment in acting along the line of what the new Regulation states, and their willingness to go beyond the rules to guarantee

[119]EDPS Opinion 9/2016 on Personal Information Management Systems

an even higher level of data protection. A strategic approach for communicating such good practices could benefit the reputation of companies and increase the trust of customers toward their products and services. Connecting this solution with Corporate Social Responsibility strategies and Socially Responsible Investments could pave the way for this kind of approach.

In order to support the scientific research that makes use of personal data, especially research on health and diseases prevention, the idea of creating a centralized data centre at European level could pose too many risks of data breach and misuses. The investment in this field should first move instead toward a standardization of data collecting and storage at national level in each country, to create at least common and comparable databases. Once the Member States reach a good level of internal standardisation, the cooperation at European level can be simplified. This could then facilitate the sharing of mainly anonymised data among research focused entities such as scientific institutes, hospitals, public or private healthcare structures. A disclosure of personal data would be conceivable only in exceptional cases, under strict protocols that make use of adequate privacy enhancing technologies, such as encryption.

12. Bibliography

- Alicino et al (2015). *Assessing Ebola-related web search behaviour: insights and implications from an analytical study of Google Trends-based query volumes. Infectious Diseases of Poverty*.
- Allen, Colin, Wendell Wallach, and Iva Smit (2006) *Why machine ethics?* IEEE Intelligent Systems 21.4.
- Anderson and Rainie (2012). *Big Data: Experts say new forms of information analysis will help people be more nimble and adaptive, but worry over humans' capacity to understand and use these new tools well. The Future of Internet*. Pew Research Center.
- Bauer et al (2013). *A comparison of users' perceptions of and willingness to use Google, Facebook, and Google+ single-sign-on functionality*. Proceedings of the 2013 ACM Workshop on Digital Identity Management.
- Bauer et al (2013). *A comparison of users' perceptions of and willingness to use Google, Facebook, and Google+ Single-Sign-On functionality*. Proceedings of the 2013 ACM Workshop on Digital Identity Management.
- Bello-Orgaz et al (2016). *Social Big Data: Recent achievements and new challenges*. Information Fusion.
- Benenson et al (2013). *Android and iOS users' differences concerning security and privacy*. CHI 2013: Changing perspectives. Extended abstracts.
- Bertot and Choi (2013). *Big Data and e-government: Issues, policies and recommendations*. Proceedings of the 14th International Conference on Digital Government Research.
- Boyd and Crawford (2012). *Critical questions for Big Data. Information, Communication & Society*.
- Briggs et al (2016). *Everyday surveillance*. Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems.
- Brown, Geoffrey (1991) *Is there an Ethics of Computing?* Journal of applied philosophy 8.1.
- Buhr and Kleiner (2012). *European Open Data policy: Challenges and opportunities*. Policy Advice and Political Consulting, 5(3).
- Carmel (2016). *Regulating "Big Data education" in Europe: lessons learned from the US*. Internet Policy Review, 5(1): DOI: 10.14763/2016.1.402.
- Coria et al (2013). *Delta score: A novel and simplified measurement for digital divide of cities*. The Proceedings of the 14th Annual International Conference on Digital Government Research.
- Crawford et al (2014). Critiquing Big Data: politics, ethics, epistemology. International Journal of Communication.
- Cumbley and Church (2013). *Is "Big Data" creepy?* Computer Law & Security Review.
- Dainow (2015). *Digital alienation as the foundation of online privacy concerns*. SIGCAS Computer & Society.
- Daries at al (2014). *Privacy, anonymity and Big Data in the social science*s. Communication of the ACM, 57(9).
- Decker, Scott and Pyrooz, David (2011). *Gangs, Terrorism, and Radicalization*. Journal of Strategic Security, no. 4: 151.

- Directive 2013/37/EU of the European Parliament and of the Council on the use of the public sector information.
- Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. Official Journal L 281, 23/11/1995.
- Douglas (2001). *3d data management: Controlling data volume, velocity and variety*. Gartner.
- ENISA (2015). *Privacy by design in Big Data. An overview of privacy enhancing technologies in the era of Big Data analytics*.
- Eshleman, Andrew (2014). *Moral responsibility*. The Stanford encyclopedia of philosophy.
- European Commission (2012). Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions – *Unleashing the Potential of cloud Computing in Europe.*
- European Commission (2014). Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions - *Towards a thriving data-driven economy.*
- European Commission (2013). *Digital Single Market portal – Digital Skills and Jobs Coalition*.
- European Commission (2013). Joint communication to the European Parliament, the Council, the European Economic and Social Committee and the committee of the Regions – *Cybersecurity Strategy of the European Union: An Open, Safe and Secure Cyberspace*.
- European Commission (2016). *A guide to ICT-related activities* in WP2016-2017.
- European Commission (2016). Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions - *Connectivity for a Competitive Digital Single Market* - *Towards a European Gigabit Society.*
- European Commission (2016). *Digital Single Market portal–Big Data.*
- European Commission (2016). *Digital Single Market portal – Towards a thriving data-driven economy.*
- European Commission Staff Working Paper Impact Assessment Accompanying the document GDPR and Directive of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data by competent authorities for the purposes of prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and the free movement of such data.
- European Data Protection Supervisor, Opinion 9/2016 on *Personal Information Management Systems*.
- European Data Protection Supervisor Preliminary (2014). *Opinion on Privacy and competitiveness in the age of Big Data: The interplay between data protection, competition law and consumer protection in the Digital Economy*.
- European Data Protection Supervisor, Opinion 4/2015 *Towards a new digital ethics Data, dignity and technology*.
- 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*.

- European Data Protection Supervisor *Strategy 2015-2020, developing an ethical dimension to data protection*.
- European Privacy Association, *EPA and Corporate Social Responsibility*.
- Executive Office of the President (2014). *Big Data: Seizing Opportunities, Preserving Values*,
- Feigenbaum et al (2014). *Open vs Closed systems for accountability*. Proceedings of the 2014 Symposium and Bootcamp on the Science of Security.
- Floridi, Luciano (2014). Artif*icial Agents and Their Moral Nature. The Moral Status of Technical Artefacts*. Springer Netherlands.
- Floridi, Luciano (2014). *The fourth revolution: How the infosphere is reshaping human reality*. OUP Oxford.
- Frey, Carl Benedikt, and Michael A. Osborne. (2013). The future of employment. How susceptible are jobs to computerisation", Oxford Martin Programme on Technology and Employment
- Friedman, Batya, and Peter H. Kahn (1992). *Human agency and responsible computing: Implications for computer system design*. Journal of Systems and Software.
- Gordon (2016). *Computers and the Internet from the Mainframe to Facebook*. In The Rise and Fall of American Growth: The U.S. Standard of Living Since The Civil War. Princeton University Press, Princeton and Oxford.
- Gudymenko et al (2011). *Privacy Implications of the Internet of Things*. Proceedings of the 2nd International Joint Conference on Ambient Intelligence.
- Gurstein, Michael B. (2011). *Open Data: Empowering the Empowered or Effective Data Use for Everyone?* First Monday 16, no. 2.
- Harris, Lisa, Paul Harrigan (2015). *Social Media in Politics: The Ultimate Voter Engagement Tool or Simply an Echo Chamber?* Journal of Political Marketing 14.3.
- Hauge et al (2016). *Tagging Banksy: using geographic profiling to investigate a modern art mystery*. Journal of Spatial Science, 61.
- Heffetz and Ligett (2014). *Privacy and data-based research*. The Journal of Economic Perspectives, 28(2).
- Howard et al (2010). *Comparing digital divides: Internet access and social inequality in Canada and the United States*. Canadian Journal of Communication, 35.
- Hull et al (2011). *Contextual gaps: Privacy issues on Facebook*. Ethics and Information Technology, 13(4).
- Italian Legislative Decree 97 of 25 May 2016. Revision and simplification of the dispositions on prevention of corruption, publicity and transparency, correcting law number 190, 6 November 2012, and legislative decree 33, 14 March 2013, following art. 7 law 124, 7 August 2015 on reorganization of Public Administrations.
- Johnson Jeffrey Alan, (2014). *From Open Data to Information Justice*, Ethics and Information Technology 16, no. 4.
- Joinup (2015) eGovernment in the European Union.
- Ybarra Michele L., Diener-West, Marie and Philip Leaf (2007). *Examining the Overlap in Internet Harassment and School Bullying: Implications for School Intervention*, Journal of Adolescent Health 41, no. 6.

- Kirkpatrick, Keith, (2016) Bat*tling Algorithmic Bias: How Do We Ensure Algorithms Treat Us Fairly?* Communications of the ACM 59, no. 10.
- Lev Muchnik, Sinan Aral, and Sean J. Taylor (2013). *Social Influence Bias: A Randomized Experiment*, Science 341, no. 6146.
- Lim and Thuemmler (2015). *Opportunities and challenges of Internet-based health interventions in the future Internet.* Proceedings of the 12th International Conference on Information Technology – New Generations.
- Locasto et al (2011). *Security and privacy considerations in digital death.* Proceedings of the 2011 New Security Paradigms Workshop.
- Machanavajjhala and Reiter (2012). *Big privacy: Protecting confidentiality and Big Data.* XRDS, 19(1).
- Maybin Simon (2016). *How Maths Can Get You Locked up.*
- Mantelero (2013). *The EU Proposal for a General Data Protection Regulation and the roots of the 'right to be forgotten'.* Computer Law & Security Review, 29.
- Manyika et al (2011). *Big Data: The next frontier for innovation, competition and productivity*. McKinsey Global Institute.
- Michael et al (2014). *Uberveillance and the Internet of Things and People.* Proceedings of the 1st International Conference of Contemporary Computing and Informatics.
- Mittelstadt and Floridi (2015). *The ethics of Big Data: current and foreseeable issues in biomedical context.* Science and Engineering Ethics, 22(2).
- Narayanan and Shmatikov (2010). *Myths and fallacies of "Personally identifiable information*". Communications of the ACM, 53.
- Negreiro (2015). *Bridging the digital divide in EU.* European Parliamentary Research Service – Member's Research Service.
- Nickerson and Rogers (2014). *Political campaigns and Big Data.* The Journal of Economic Perspectives, 28(2).
- Noorman, Merel (2012). *Computing and moral responsibility*. The Stanford encyclopedia of philosophy.
- Norwegian Data Protection Authority's (2013). *Big data-privacy principles under pressure*.
- Panagopoulos (2011). *Social Pressure, Surveillance and Community Size: Evidence from Field Experiments on Voter Turnout*. Special Symposium: Electoral Forecasting Symposium, 30.
- Pariser (2011). *The filter bubble – what the Internet is hiding from you.* Penguin Press, New York.
- Poel et al (2015). *Data for policy: A study of Big Data and other innovative data-driven approaches for evidence-informed policymaking*.
- Regulation 1049/2001 of the European Parliament and the Council regarding public access to European Parliament, Council and Commission documents.
- Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

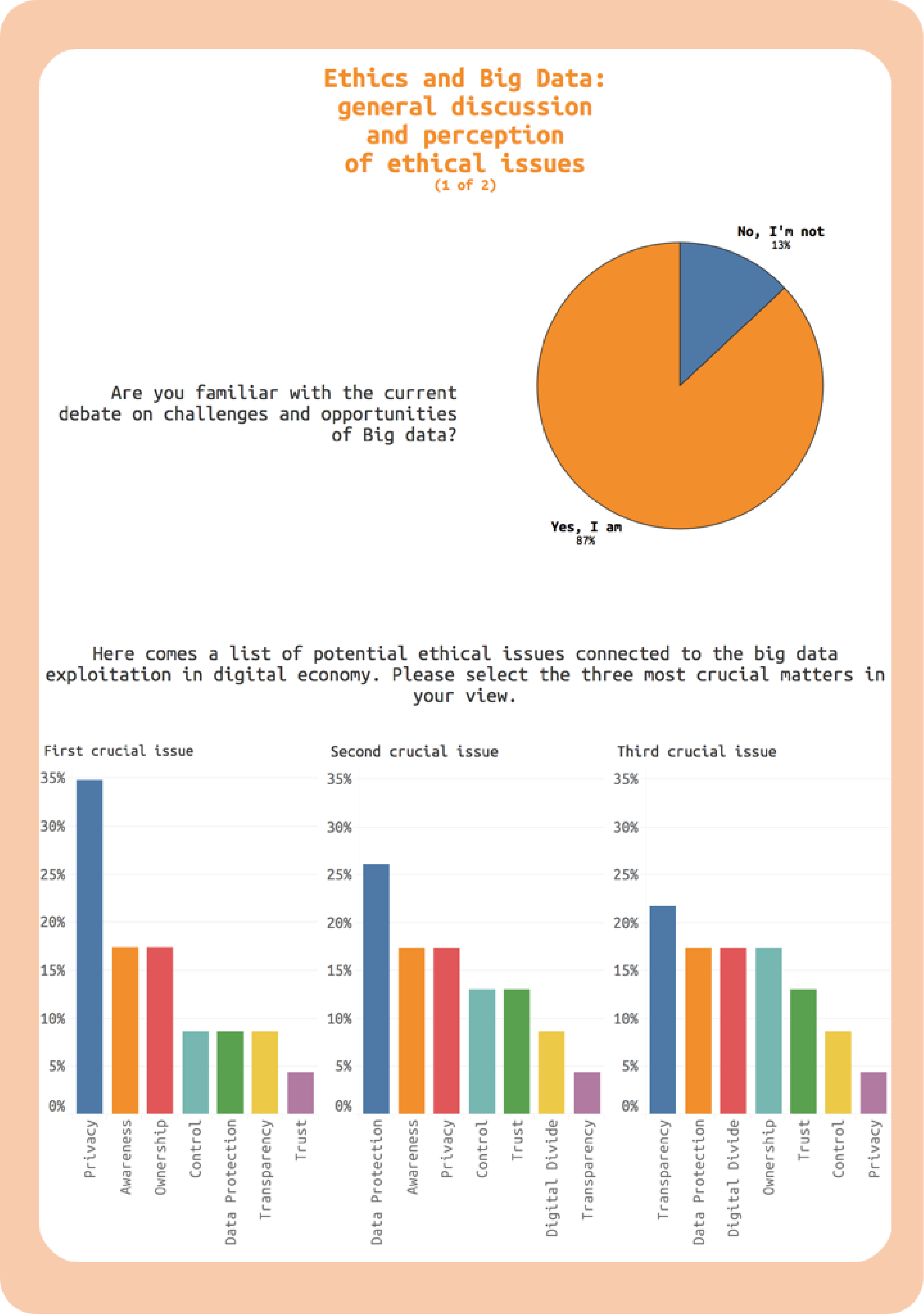
- Rockwell and Sinclair (2016). *False positives: Opportunities and dangers in Big¬-Data text analysis*. In Hermeneutica: Computer-assisted interpretation in the humanities. MIT Press, Cambridge (USA-MA).
- SAS (2013). *Big Data Analytics – Adoption and Employment Trends, 2012–2017.*
- Schrammel et al (2011). *Privacy, trust and interaction in the Internet of things. International Joint Conference on Ambient Intelligence. 12th* International Conference of Information Technology – New Generations.
- Sen, Amartya (2001) *Development as freedom*. Oxford Paperbacks.
- Sen, Amartya (2004) *Rationality and freedom*. Harvard University Press.
- Shah, James Y., Arie W. Kruglanski, and Erik P. Thompson, (1998). *Membership Has Its (Epistemic) Rewards: Need for Closure Effects on in-Group Bias.* Journal of Personality and Social Psychology 75, no. 2.
- Siemens and Baker (2016). *Educational data mining and learning analytics: towards communication and collaboration.* Proceedings of the 2nd International Conference on Learning Analytics and Knowledge.
- Smith & Anderson *5 facts about online dating* (2016). Pew Research Center.
- *Study on Personal Data Stores*, Cambridge University, 2015.
- Sung (2015). *A study on the effect of smartphones on the digital divide*. Proceedings of the 16th Annual International Conference on Digital Government Research.
- Swan (2015). *Philosophy of Big Data: Expanding the human-data relation with Big Data science services*. IEEE 1st International Conference on Big Data Computing Service and Applications.
- Takeoka Chatfield, Akemi, Christopher G. Reddick, and Uuf Brajawidagda, *Tweeting Propaganda, Radicalization and Recruitment: Islamic State Supporters Multi-Sided Twitter Networks*. Proceedings of the 16th Annual International Conference on Digital Government Research (ACM, 2015).
- Tanner (2014) *Different customers, different prices, thanks to Big Data*. Forbes 26 March 2014.
- Tavani, Herman T. (2007). *Philosophical theories of privacy: Implications for an adequate online privacy policy*. Metaphilosophy 38.1.
- Tucker (2013). *Has Big Data made anonymity impossible?* In: Big Data gets personal – MIT Technology Review.
- United Nations (2006). *The digital divide report: ICT diffusion index 2005*. United Nations Publications.
- Vatsavai et al (2012). *Spatiotemporal Data Mining in the era of Big Spatial Data: Algorithms and applications.* Proceedings of the 1st ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data.
- Vayena et al (2015). *Ethical challenges of Big Data in public health*. PLOS Computational Biology, 11.
- Villars et al (2011). *Big data: What it is and why you should care*. IDC White Paper.
- Vitolo et al (2015). *Web technologies for environmental Big Data*. Environmental Modelling & Software, 63.
- Ward and Barker (2013). *Undefined by data: A survey of Big Data definitions.*

- Westin, Alan F. (1968) *Privacy and freedom*. Washington and Lee Law Review 25.1: 166.

13. Web sites

- www.apps.icahn.mssm.edu/asthma/
- www.bbc.com/news/magazine-37658374
- www.data4policy.eu/
- www.dl.acm.org/citation.cfm?id=2757408
- www.easylumen.it/it/easyland
- www.ec.europa.eu/digital-single-market/en//digital-skills-jobs-coalition
- www.ec.europa.eu/digital-single-market/en/big-data
- www.ec.europa.eu/digital-single-market/en//digital-skills-jobs-coalition
- www.ec.europa.eu/digital-single-market/en/towards-thriving-data-driven-economy
- www.federa.lepida.it
- www.firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/viewArticle/3316
- www.forbes.com/sites/adamtanner/2014/03/26/different-customers-different-prices-thanks-to-big-data
- www.hl7.org
- www.infojobs.com
- www.insight-fp7.eu
- www.linkedin.com
- www.michaeljfox.org/foundation/publication-detail.html?id=562&category=7
- www.monster.com
- www.npd.com/wps/portal/npd/us/news/press-releases/2015/the-demographic-divide-fitness-trackers-and-smartwatches-attracting-very-different-segments-of-the-market-according-to-the-npd-group
- www.pewresearch.org/fact-tank/2016/02/29/5-facts-about-online-dating
- www.stats.oecd.org/glossary/detail.esp?ID=4719
- www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_5.1.14_final_print.pdf
- http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html
- http://www1.unece.org/stat/platform/display/bigdata/Big+Data+Projects

14. Appendix 1 – Focus section on the results of the survey “The Ethics of Big Data in Europe. Balancing the economic benefits of Big Data with the need to protect fundamental human rights”



Ethics and Big Data:
general discussion
and perception
of ethical issues
(1 of 2)

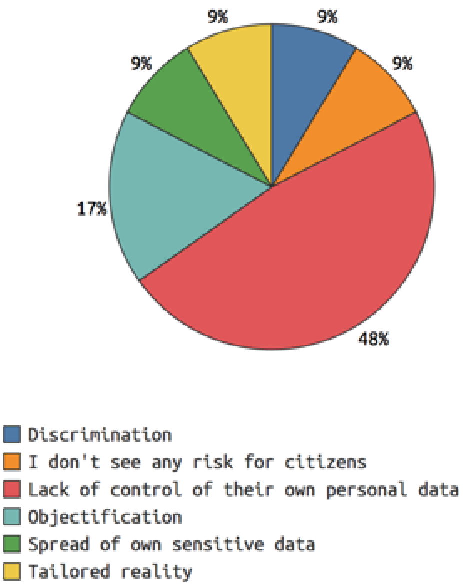
Are you familiar with the current debate on challenges and opportunities of Big data?

Here comes a list of potential ethical issues connected to the big data exploitation in digital economy. Please select the three most crucial matters in your view.

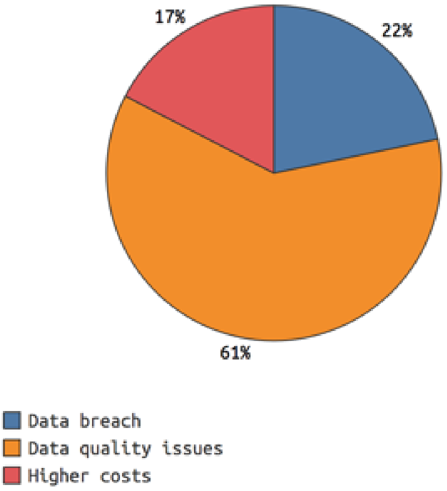
Ethics and Big Data: general discussion and perception of ethical issues

(2 of 2)

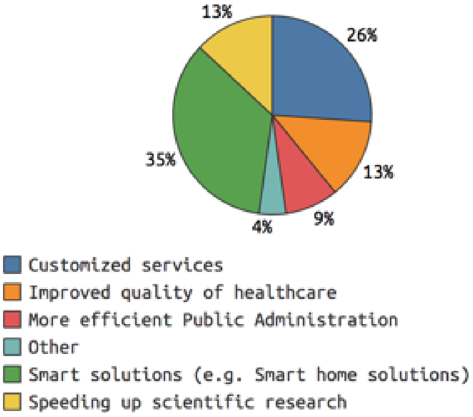
Which do you think is the **main risk** for **people** in a data-driven society?



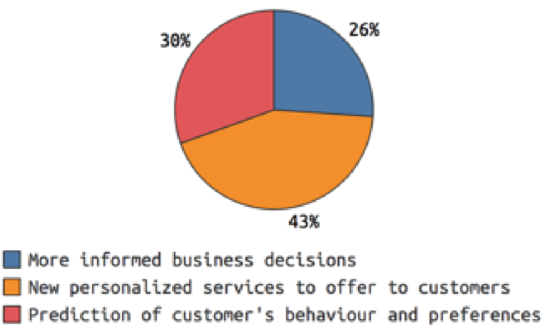
Which do you think is the **main risk** for **companies** in a data-driven society?



Which do you think is the **main benefit** that Big Data can bring to **people**?

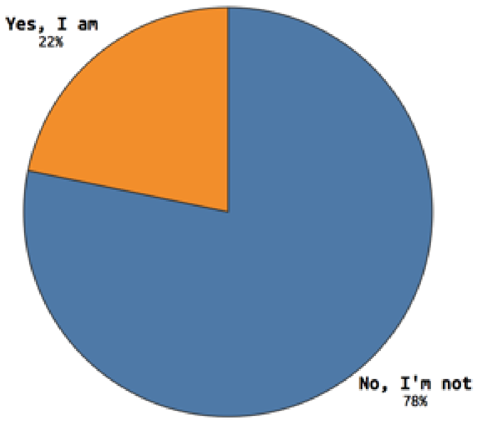


Which do you think is the **main opportunity** for data-driven **businesses**?

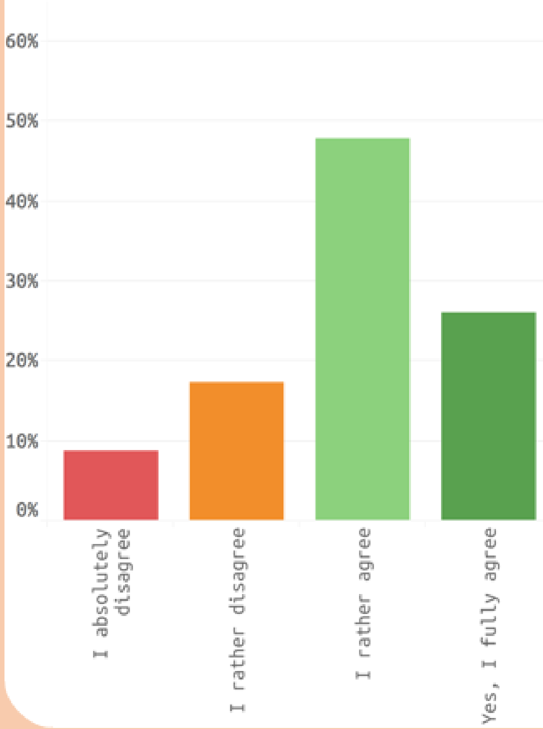


Ethics and Big Data:
The General Data Protection Regulation

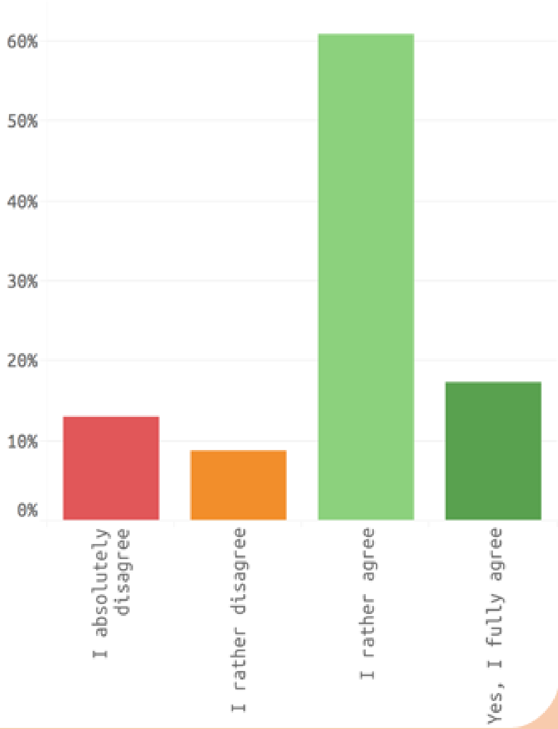
Are you familiar with the new European Regulation on Data Protection (GDPR), entered into force on May 2016?

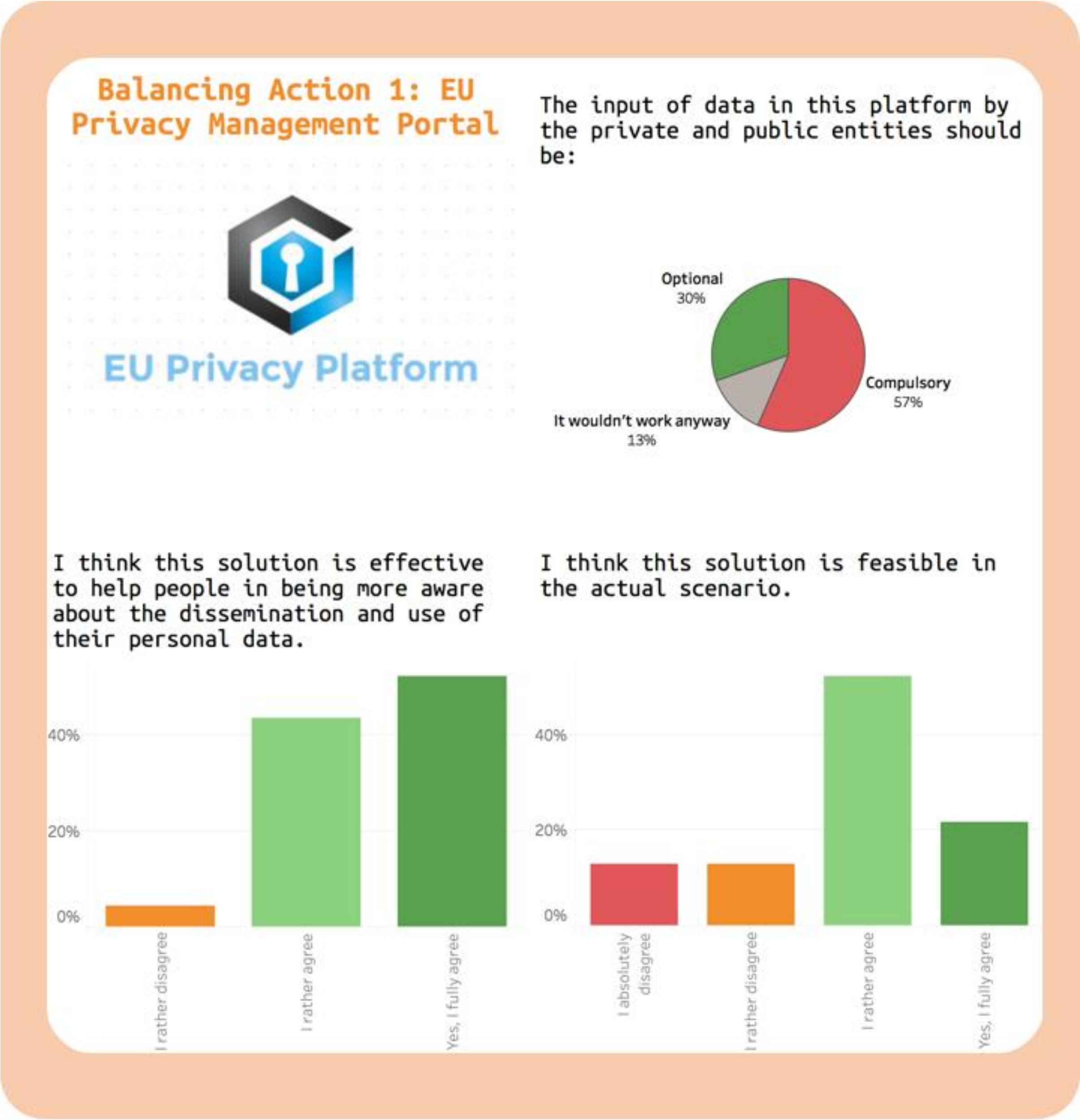


The GDPR clarifies the procedures for companies to deal with personal data of their customers.



The GDPR will guarantee a better protection of people's personal data.



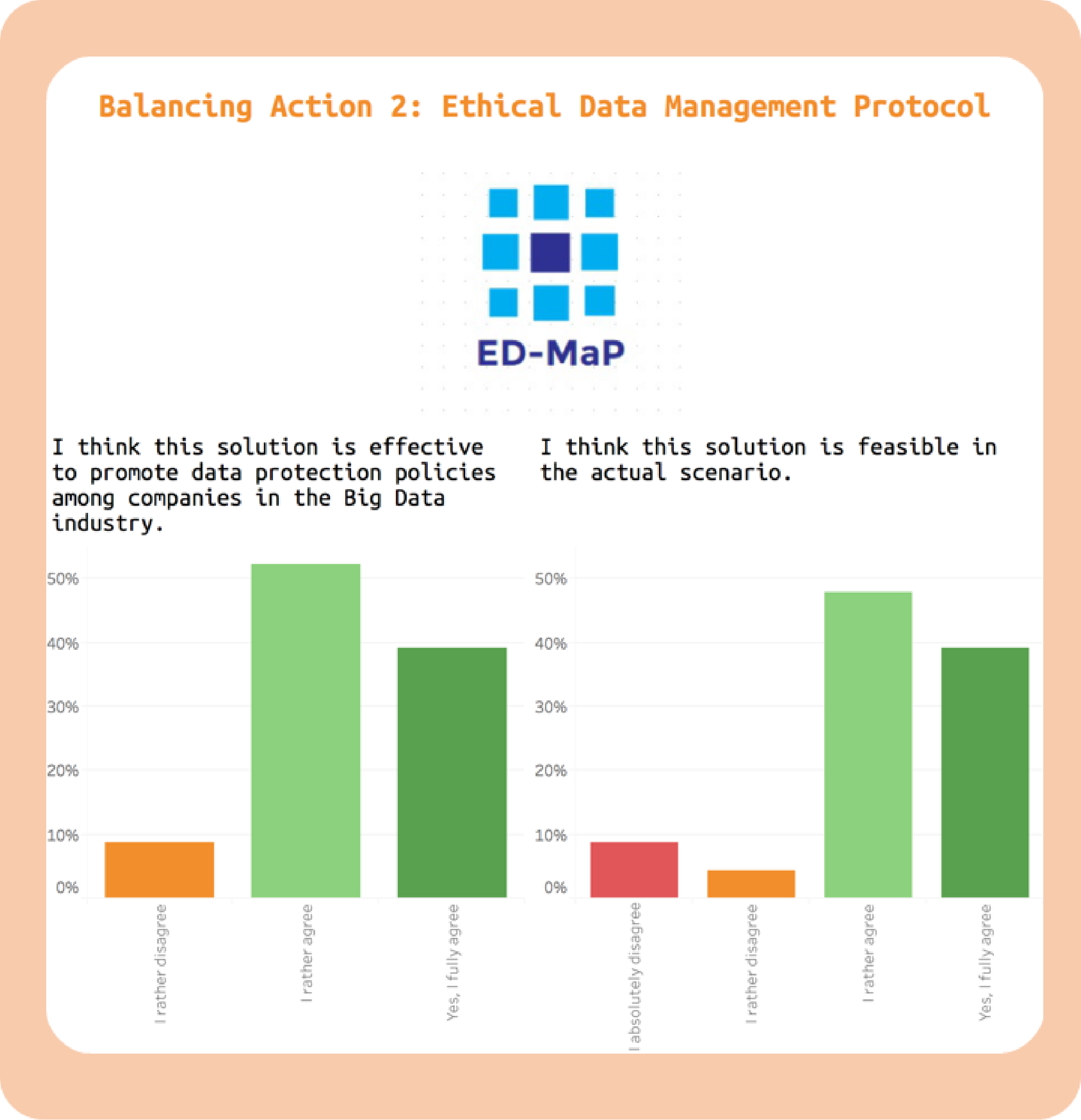


Balancing Action 1: EU Privacy Management Portal

The input of data in this platform by the private and public entities should be:

I think this solution is effective to help people in being more aware about the dissemination and use of their personal data.

I think this solution is feasible in the actual scenario.

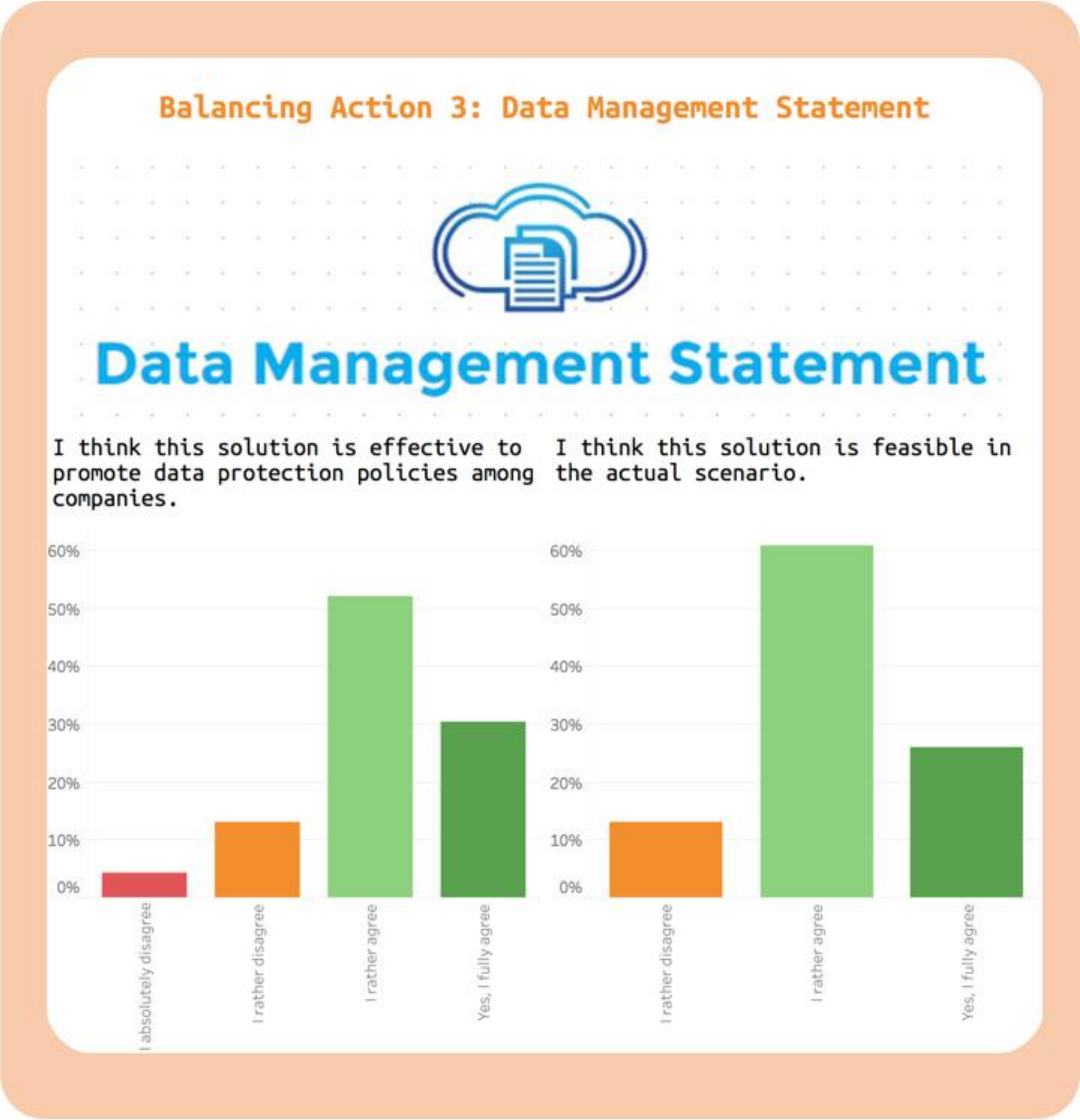


Balancing Action 2: Ethical Data Management Protocol

ED-MaP

I think this solution is effective to promote data protection policies among companies in the Big Data industry.

I think this solution is feasible in the actual scenario.

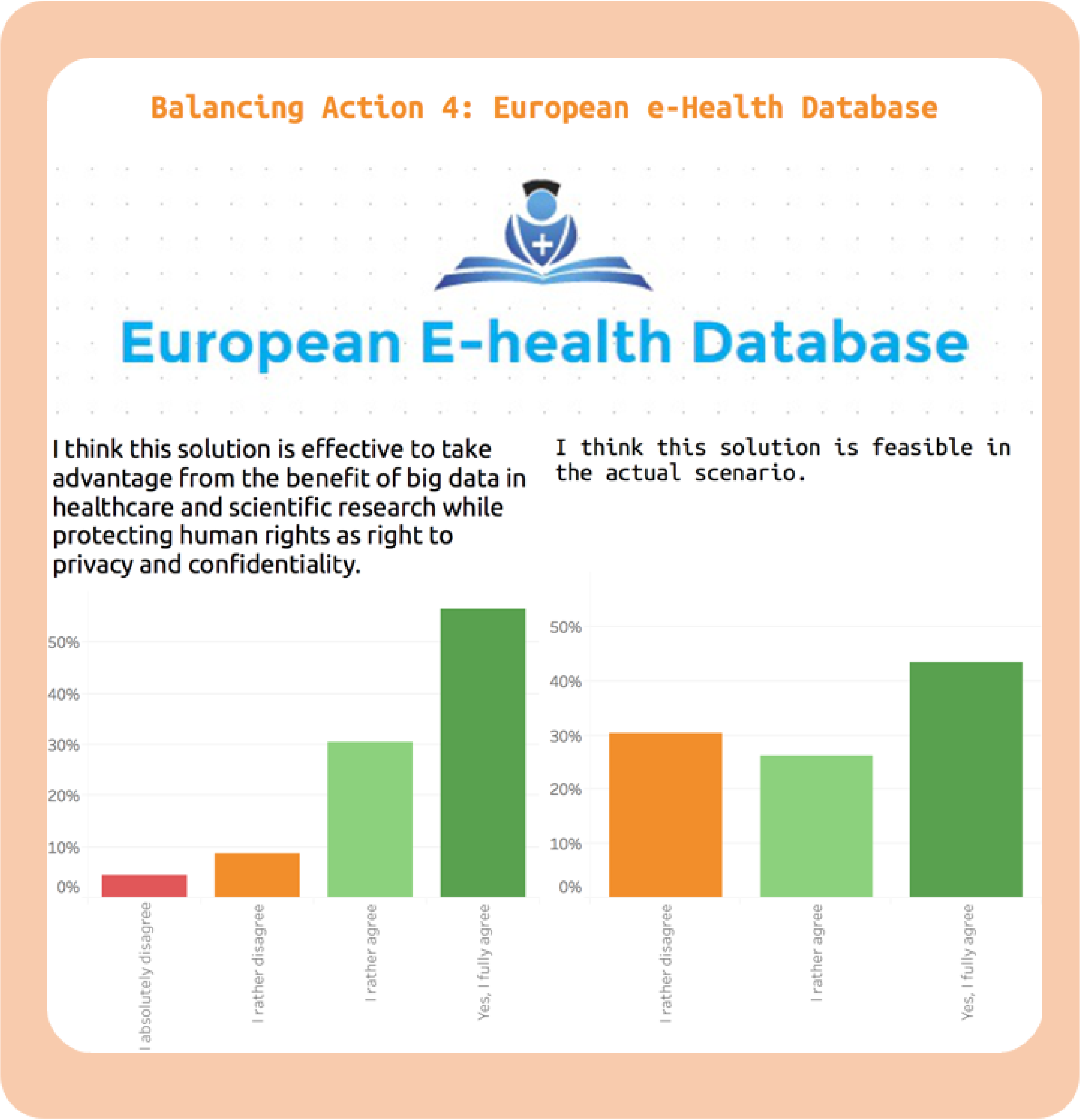


Balancing Action 3: Data Management Statement

Data Management Statement

I think this solution is effective to promote data protection policies among companies.

I think this solution is feasible in the actual scenario.

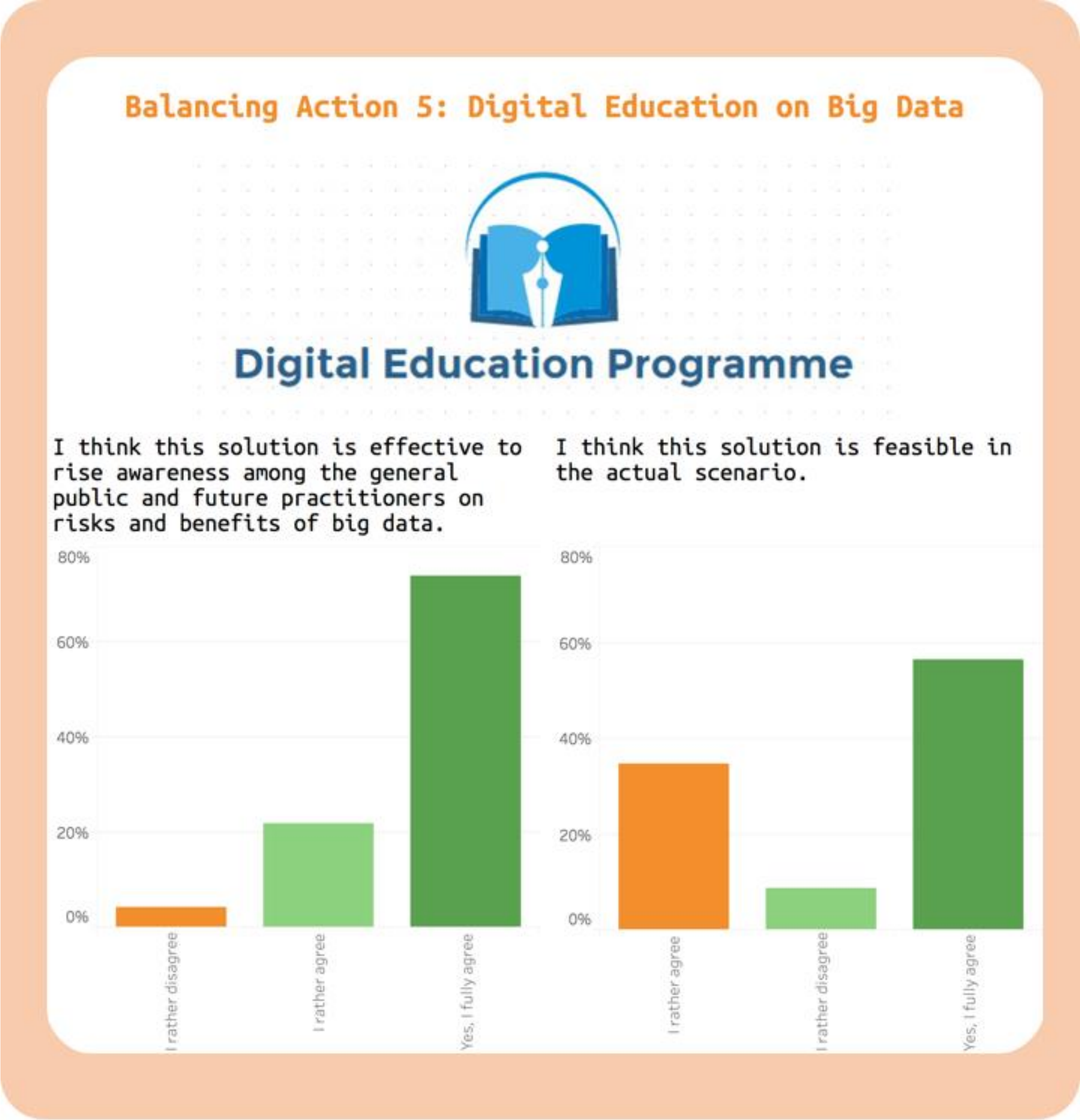


Balancing Action 4: European e-Health Database

European E-health Database

I think this solution is effective to take advantage from the benefit of big data in healthcare and scientific research while protecting human rights as right to privacy and confidentiality.

I think this solution is feasible in the actual scenario.



Balancing Action 5: Digital Education on Big Data

Digital Education Programme

I think this solution is effective to rise awareness among the general public and future practitioners on risks and benefits of big data.

I think this solution is feasible in the actual scenario.

TEN
SECTION

For further information please visit our website:
www.eesc.europa.eu/ten
@EESC_TEN
or contact us: **ten@eesc.europa.eu**

European Economic and Social Committee

Rue Belliard/Belliardstraat 99
1040 Bruxelles/Brussel
BELGIQUE/BELGIË

Published by: "Visits and Publications" Unit
EESC-2017-41-EN
www.eesc.europa.eu

evodevo

Print
QE-04-17-306-EN-C
ISBN 978-92-830-3532-9
doi:10.2864/446901

Online
QE-04-17-306-EN-N
ISBN 978-92-830-3534-3
doi:10.2864/97426

EN