D3.2
Assessment of existing technologies

Ethical and Societal Implications of Data Sciences

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731873

e-SIDES – Ethical and Societal Implications of Data Sciences

Data-driven innovation is deeply transforming society and the economy. Although there are potentially enormous economic and social benefits this innovation also brings new challenges for individual and collective privacy, security, as well as democracy and participation. The main objective of the CSA e-SIDES is to complement the research on privacy-preserving big data technologies, by analyzing, mapping and clearly identifying the main societal and ethical challenges emerging from the adoption of big data technologies, conforming to the principles of responsible research and innovation; setting up and organizing a sustainable dialogue between industry, research and social actors, as well as networking with the main Research and Innovation Actions and Large Scale Pilots and other framework program projects interested in these issues. It will investigate stakeholders' concerns, and collect their input, framing these results in a clear conceptual framework showing the potential trade-offs between conflicting needs and providing a basis to validate privacy-preserving technologies. It will prepare and widely disseminate community shared conclusions and recommendations highlighting the best way to ultimately build confidence of citizens and businesses towards big data and the data economy.

D3.2 Assessment of existing technologies

Work package	WP 3 – Review of existing technologies
Lead author	Daniel Bachlechner (Fraunhofer ISI)
Contributing authors	Michael Friedewald (Fraunhofer ISI) Jana Weitkamp (Fraunhofer ISI) Melek Akca Prill (Fraunhofer ISI) Karolina La Fors (Leiden University) Alan M. Sears (Leiden University)
Internal review	Duncan Brown (IDC) Bart Custers (Leiden University) Karolina La Fors (Leiden University) Alan M. Sears (Leiden University)
Due Date	M16 (April 2018)
Date	22 May 2018
Version	0.99 (draft)
Type	Report
Dissemination level	Public

This document is Deliverable 3.2 of Work Package 3 of the e-SIDES project on Ethical and Societal Implications of Data Science. e-SIDES is an EU funded Coordination and Support Action (CSA) that complements Research and Innovation Actions (RIAs) on privacy-preserving big data technologies by exploring the societal and ethical implications of big data technologies and providing a broad basis and wider context to validate privacy-preserving technologies. All interested stakeholders are invited to look for further information about the e-SIDES results and initiatives at www.e-sides.eu.

Grant Agreement number: 731873

Executive Summary

This report assesses privacy-preserving technologies taking ethical and societal issues into account. The assessment consists of two parts: a technology-specific assessment of selected classes of technologies and a more general assessment of the technologies. Among the assessed technologies are technologies for anonymisation and sanitisation, encryption, multi-party computation, access control, policy enforcement, accountability and transparency, data provenance, and access, portability and user control. With respect to the issues, the focus is on privacy but issues and values such as self-determination, welfare interdependency, trustworthiness, accountability, fairness and legislation are also taken into account. The assessment is based on a series of interviews and desk research.

The assessment led to the following overarching observations:

- The predefined set of classes of technologies is **quite comprehensive**. Interviewees suggested additional technologies, but most of them were at least related to the predefined classes. Blockchains, quantum computing, data classification, homomorphic encryption and sketches were mentioned explicitly.
- The classes of technologies contribute to privacy preservation in different ways and are closely linked with each other. In practice, the technologies **need to be combined** to be effective and there is not one most important class of technologies.
- The technologies **pursue different aims**. While some aim at overcoming the need for trust in other parties (e.g., multi-party computation, homomorphic encryption), others aim at increasing trust in other parties (e.g., access control, policy enforcement, accountability, transparency).
- A **multidimensional measure** for privacy preservation is needed that covers relevant factors in a balanced way. Apart from the degree of protection and the cost of protection, it is important to also take the societal value of data into account.
- Some observe a **fundamental tension** between the objective of big data, which is collecting a lot of personal data and using it, and privacy, which is about protecting personal data. The two goals seem almost antagonistic. Yet, classes of technologies constitute attempts to bridge gaps between goals as well as differences that are, among others, regional, legal, organisational or procedural.

With respect to the specific classes of technologies, the focus was on the effectiveness of the technologies in addressing the issues and problems that may arise when addressing the issues with the technologies:

Anonymisation and sanitisation: It is important to differentiate between the anonymisation of datasets that include personal data and anonymisation to prevent the collection of data that is identifiable. With respect to the anonymisation of datasets, key problems are to determine the optimal balance between improved privacy protection through anonymisation and sanitisation, and the usefulness of data for decision making, the uniqueness of certain characteristics or behaviours, and the fact that it is usually unknown what other information is available to potential adversaries.

Encryption: It is often stressed that the battle between law enforcement and privacy is being fought over encryption. A solution that is considered to maximise privacy and query expressiveness, at least theoretically, is homomorphic encryption. While a key challenge is computation cost that makes it

relatively slow compared to other methods, a big advantage is that it does not lead to a loss in data quality. It has to be kept in mind that some of the current encryption methods may be useless in the era of quantum computing.

Multi-party computation: Multi-party computation can have a large impact on a number of areas. The technology is useful to unlock new possibilities in the context of joint data processing. There are some areas in which multi-party computation proved to be quite efficient. It is mostly practical implementation problems that stand in the way of wider adoption. While anonymisation allows using standard analytics tools, the value of the data may be decreased. In contrast, multi-party computation allows working with the data as it is, but restricts the efficiency of the analysis and the range of tools that can be used.

Access control: Access control gets increasingly difficult when higher numbers of devices are used. Access control in itself is considered inherently inadequate as a framework for addressing privacy. Those who obtain access to the data, legitimately or not, can use the data without restriction. What is needed is the provision of access in a highly granular and dynamic way. Technologies for access control, policy enforcement, accountability and transparency rely very much on a strong trust model as they have a single point of failure. There is usually only one entity that is designing the access control system.

Policy enforcement: To be able to enforce sensitive data security policies, a policy enforcement framework has to be flexible and support different data processing requirements. Policy enforcement gets increasingly difficult as the chain of responsibilities becomes longer and the roles more geographically dispersed. Policy enforcement is needed as it is very hard for data protection authorities to exercise their enforcement power, particularly, if actors outside the European Union (EU) are involved.

Accountability and transparency: More and more tools and methods are developed to build trust over accountability and transparency. However, the increasing use of machine learning algorithms makes things difficult. Companies should be fully transparent to customers. It is fundamental to find a balance between transparency and confidentiality. Instead of providing users with information on how and why their personal data is processed, accountability should concentrate on monitoring the use of data through mechanisms such as auditability, technical design of algorithms and software-designed regulation.

Data provenance: The challenges that are introduced by the volume, variety and velocity of big data, also pose challenges for provenance and quality of big data. Key challenges include the size of provenance data, the overhead linked to its collection, the integration of distributed provenance data and the reproduction of an execution from provenance data. What is needed are flexible provenance query tools and provenance visualisation tools.

Access, portability, user control: Empowering users, informing them and giving them access to their data is not only for the users' benefit but also for the benefit of the organisation using the data. However, even if empowered, users do not practice related opportunities very often. This may be because they are not informed about the opportunity, because the process is too complicated or because they do not care. It is difficult for users to understand what his or her data can be used for and how certain database transactions will end up being inefficient or unfair. To some extent, user control is eroded in the big data context.

With respect to the technologies in general, attention was paid to the integration of the technologies in today's big data solutions, the demand for big data solutions that include the technologies, regional and cultural differences with respect to the availability and use of the technologies, the need for non-technical measures as a complement to technologies, and the responsibility for addressing the issues:

Today's solutions: There is wide agreement that technologies are integrated only to a limited extent in today's big data solutions. There are strong solutions in research but there is a big gap when it comes to deployment. Incidents, however, may result in further reluctance of companies to share data. What is needed are technologies that protect the data but at the same time allow sharing it. Privacy by design can be effective since traditional legal instruments may not be implemented in big data settings because they are in conflict with business models and perceived as blocking; and it will be very hard to block technology development in the future. Companies increasingly seem to try to brand themselves as privacy protectors. However, although companies sometimes claim that data has been anonymised, it is important to check that carefully. Sometimes, only personally identifiable information is removed, which is not sufficient to avoid re-identification.

Customers and users: One would expect that the handling of personal data and privacy protection are very important for clients of companies that are highly networked, deal with big data and process their personal data. However, it seems that quite some of the clients are blinded by the benefits they get in return for their personal data. There seems to be broad consensus with respect to the rather low demand from the customer side for technologies to protect privacy. To change that, a change in culture is considered necessary. Features protecting users must be embedded in the products rather than provided as add-ons. Moreover, people shouldn't have to pay extra for privacy preservation. Policy makers and regulators could play an important role with respect to the demand. Currently, legislation and scandals taken up by the media seem to be the key driver of demand.

Regional differences: There is little doubt that there are considerable regional and cultural differences. In North America, it is common not to constrain things and see what happens and then generalise and apply case-based legal decisions. The European historical context is more rule-driven. North Americans do not seem to trust their government much. In Europe, governments are seen as important protectors of privacy. It is possible that the EU becomes an exporter of norms that have the potential to lead to technological changes globally. However, it is also possible that the EU is deprived of leading technologies. North America, and in particular the Silicon Valley, hold the utility of data in high regard. The things that can be learned from data weigh significantly more than possible concerns about privacy.

Organisational measures: There is consensus that the combination of technical and organisational measures is essential. The Cambridge Analytica and Facebook scandal shows vividly the limitations of reactive approaches that prescribe norms and actions that are taken if there is a breach of privacy or some types of rules are broken. Technical solutions that are proactive in the sense that they prevent breaches or rule violations in the first place are needed. Awareness and education are key aspects in modern organisations. Moreover, technologies need to be accompanied by the right processes as well as legislation, agreements and policies.

Responsibility: It is often stressed that consumers need to protect themselves because nobody else will do it for them. In general, organisations collecting, using and distributing data are responsible for data

management and anonymisation. There is wide agreement that the strongest party should have the biggest responsibilities. It is important that data protection is not considered as *"somebody else's problem"* as this point of view passes the responsibility from one hand to another. Supervisory authorities and governments should have a role because they have to shape the framework conditions.

Contents

Grant Agreement number: 731873

Figures

Tables

Abbreviations

ABAC	Attribute-based access control
ABE	Attribute-based encryption
CPBR	Consumer Privacy Bill of Rights
DPO	Data Protection Officer
ENISA	European Union Agency for Network and Information Security
EU	European Union
FCC	Federal Communications Commission
FHE	Fully homomorphic encryption
GAFA	Google, Apple, Facebook and Amazon
GDPR	General Data Protection Regulation
HDFS	Hadoop Distributed File System
IBE	Identity-based encryption
MPC	Multi-party computation
PDP	Provable data processing
POR	Poofs of retrievability
PRE	Proxy re-encryption
RAMP	Reduce and Map Provenance
RBAC	Role-based access control

Grant Agreement number: 731873

1. Introduction

This section outlines the background, the methodology and the structure of this document.

1.1. Background

This report is Deliverable 3.2 of the e-SIDES project. In this project, the ethical, legal, societal and economic implications of big data applications are examined in order to complement the research on privacy-preserving big data technologies (mainly carried out by ICT-18-2016 projects) and data-driven innovation (carried out, for instance, by ICT-14-2016-2017 and ICT-15-2016-2017 projects).

Within the scope of the report, the technologies described in Deliverable 3.1 are assessed taking the societal and ethical issues discussed in Deliverable 2.2 into account. The assessment consists of two parts: a technology-specific assessment of selected classes of technologies and a more general assessment of the technologies.

1.2. Methodology

The assessment of technologies is based on nine interviews and additional desk research.

The interviews consisted of two main parts:

The **first part** of the interviews focused on the assessment of the relevance and applicability of selected technologies for addressing ethical and societal issues faced in the context of big data applications.

The following questions were discussed:

- What societal and ethical issues can be addressed by the technologies?
- How effective are the technologies in addressing the issues?
- What problems may arise when addressing the issues with the technologies?

In the **second part** of the interviews, the technologies were assessed more generally by discussing the following questions:

- To what extent are the technologies integrated in today's big data solutions?
- Is there a significant demand for big data solutions that include the technologies?
- What role do regional/cultural differences play (mainly concentrating on North America and Europe)?
- To what extent do the technologies need to be complemented by non-technical measures?
- Who along the data value chain is or should be responsible for addressing the issues?

The interviews were conducted with renowned experts. Among them were researchers as well as company representatives and members of relevant organisations such as data protection authorities. The interviewees were based in Europe, North America or the Middle East at the time of the interview; many of the interviewees have had the chance to collect experience in multiple regions. The codes used to refer to the individual interviewees are explained in Table 1. The interviews lasted between 30 and 45 minutes.

Code	Primary activity at the time of the interview
I1	Technology advisor for a national data protection authority in Europe
I2	Research associate focusing on transparent computer systems at a European university
I3	Associate professor focusing on the design, analysis and application of technologies to protect privacy at a European university
I4	Professor focusing on machine learning, data and text mining, and privacy at a North American university
I5	Professor focusing on privacy at a North American university
I6	Professor focusing on privacy, cybersecurity, Internet policy and telecommunications law at a university in the Middle East
I7	Senior scientist at a multinational technology company headquartered in Europe
I8	Regional technical lead in Europe at a multinational technology company headquartered in North America
I9	Privacy and civil liberties engineer at a software and services company specialised on big data analysis headquartered in North America

Table 1 Overview of the interviewees

Despite the fact that the interviewees were male and female, only the male form is used in the text to improve readability and further protect the anonymity of the interviewees.

The interviews were completed with desk research. The desk research focused on the same set of questions as the interviews. The existing research that was taken into account had not only been published in journals or conference proceedings but also in pertinent magazines and newspapers.

Among the **search terms** used to identify relevant publications were: *"privacy-preserving big data solutions"*, *"privacy + big data"*, *"big data and societal challenges"* and *"big data + (all classes of technologies)"*.

The results of both approaches, the series of interviews and the desk research are presented in an integrated manner.

1.3. Structure

This deliverable is structured as follows:

- Section 1 outlines the background, the methodology and the structure of the deliverable.
- Section 2 provides an overview of the assessed technologies as well as the ethical and societal issues taken into account.
- Section 3 describes the results of the assessment of the relevance and applicability of selected technologies for addressing ethical and societal issues faced in the context of big data applications.
- Section 4 describes the results of a more general assessment of the technologies.
- Section 5 concludes the deliverable.

2. Relevant technologies and issues

This section provides an overview of the assessed technologies (see D3.1) as well as the ethical and societal issues (see D2.2) taken into account.

Technologies for addressing ethical and societal issues in the context of data-driven applications have already been studied for several decades. Privacy issues have clearly received most attention so far. Torra and Navarro-Arribas[1], for instance, state that there exists a solid and useful set of technologies for ensuring different levels of privacy. This does not mean however, that all issues are solved. Particularly, since big data poses new challenges to the field. Weathington, for instance, stresses that besides privacy, discrimination due to biased algorithms and inaccurate analysis due to fake data are also increasingly relevant risks in the context of data-driven business.[2]

Privacy-preserving technologies gain importance with the need of processing large volumes of data that include sensitive attributes. Datasets dealing with medical, financial or social topics, in particular, often contain sensitive attributes such as sex, race, colour, language, religion, political or other opinion, national or social origin, association with a national minority, property, birth or other status. The collection, use and distribution of such datasets requires great attention. Looking at technologies for support is therefore a logical step.

2.1. Technologies

Figure 1 provides an overview of the classes of technologies. D3.1 provides a detailed description of the classes. For the sake of simplicity and for continuity with D3.1, approaches and methods are also included when the term *technologies* is used.

Anonymisation is performed by encrypting or removing personally identifiable information from datasets. Traditional anonymisation techniques fail in the context of big data applications because there are hundreds of data points for a single individual. A full de-identification cannot be achieved. Privacy models that may be used when anonymizing data include k-anonymity, l-diversity, t-closeness and differential privacy.

Sanitisation is done by encrypting or removing sensitive information from datasets. Anonymisation is a type of sanitisation aiming at privacy protection. Sanitisation techniques other than encryption and removal of columns include masking data, substitution, shuffling and number variance. In the big data era, for instance, it can be difficult to find substitution data in large quantities.

[1] Vicenç Torra and Guillermo Navarro-Arribas, "Big Data Privacy and Anonymization," in *Privacy and Identity Management: Facing up to Next Steps*, ed. Anja Lehmann et al., 15–26 (Springer, 2016)

[2] John Weathington, "Big data privacy is a bigger issue than you think,", https://www.techrepublic.com/article/big-data-privacy-is-a-bigger-issue-than-you-think/ (accessed February 23, 2018)

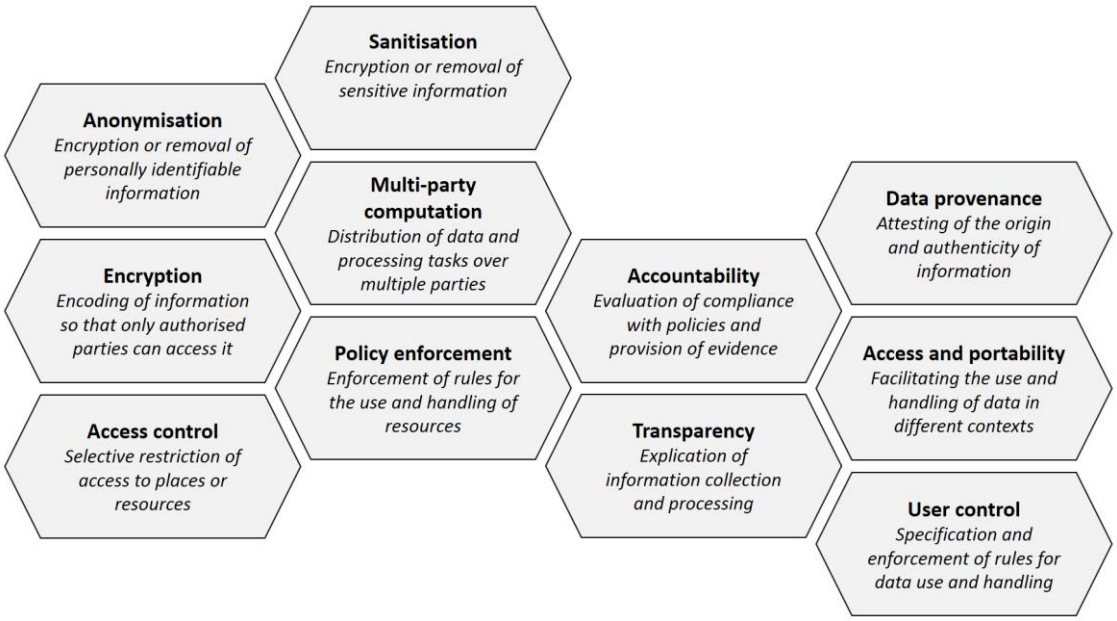

Figure 1 Overview of the classes of technologies

Encryption is the encoding of information so that only authorised parties can access it. In the context of big data applications, it is not enough to combine the advantages of public key encryption in scalability and key management with the speed and space advantages of symmetric encryption. Fine grade sharing policies are necessary that go beyond the *encrypt all or nothing* model. Relevant cryptographic primitives include ABE, IBE, PRE and functional encryption.

Multi-party computation (MPC) relies on the distribution of data and processing tasks over multiple parties. MPC is a field of cryptography with the aim to allow securely computing the result of any function without revealing the input data. Although MPC was proven to be theoretically plausible, there are still no practical solutions. Key challenges in the big data context are utility, performance and ease of use.

Access control describes the selective restriction of access to places or resources. Big data applications typically require fine-grained access control. Traditional approaches such as RBAC and user-based access control are becoming less and less manageable. ABAC in an example for a set of approaches that can conceptually support fine grained access control policies in big data based on attributes that are evaluated at run-time.

Policy enforcement focuses on the enforcement of rules for the use and handling of resources. Automated policy enforcement mechanisms are particularly important in the big data era as policies get easily lost or neglected in the course of data being transferred between different systems. Data expiration policies, for instance, are already enforced by some big data solutions.

Accountability requires the evaluation of compliance with policies and the provision of evidence. A cornerstone of accountability in the context of big data applications is the provision of automated and scalable control and auditing processes that can evaluate the level of compliance with policies. PDP and POR are among the main approaches to cloud data-integrity verification without retrieval.

Grant Agreement number: 731873

Data provenance relies on being able to attest the origin and authenticity of information. The aim is to provide a record of the processing history of pieces of data. Fine-grained provenance is difficult to achieve because big data is typically highly heterogeneous. Additionally, the use of many different analytics and storage solutions may result in prohibitively large amount of provenance information to be transferred between systems.

Transparency calls for the explication of information collection and processing. In the context of big data applications, transparency may be achieved by purely textual information, multichannel and layered approaches, or standardized icons and pictograms. Transparency is considered critical to allow data subjects informed choices.

Access and portability facilitates the use and handling of data in different contexts. Having access to data means that data subjects can look through and check the data stored. Portability gives data subjects the possibility to change service providers without losing their data.

User control refers to the specification and enforcement of rules for data use and handling. Consent mechanisms are one means that allows reaching user control, others are privacy preferences, sticky policies and personal data stores.

Some of the interviewees mentioned additional technologies that are relevant in the context of trust in big data applications. Most of the technologies mentioned, however, are at least related to the classes listed above.

For instance, blockchains, quantum computing and data classification were mentioned by a company representative who participated in the series of interviews (I8). With respect to blockchains, the interviewee stated that the technology will have significant impact on privacy and accountability as well as that the technology increases transparency as all parties in the chain know when something is changed. A key problem of blockchains is, according to the interviewee, that it is difficult to technically enforce, for instance, the right to be forgotten. Moreover, it is admitted that the use of one blockchain for millions of people is simply not possible at the moment. Currently, blockchains work well in sectoral settings only. The interviewee considers quantum computing highly relevant because some of the current encryption methods will be useless in the era of quantum computing. Finally, the interviewee stated that data classification is highly relevant for privacy protection. Different classes of data need to be addressed in different ways. It was stressed, for instance, that data provided by job applicants can only be stored by organisations for a limited time. It was further pointed out that, while detailed data has to be disposed, aggregated data can be kept for future use. The high level outcome of an analysis is not as sensitive as the actual detailed records.

A professor focusing on machine learning, data and text mining as well as privacy (I4) mentioned blockchains, homomorphic encryption and sketches. With respect to blockchains, the interviewee pointed out that they allow building solutions that are trusted by a number of parties. A key challenge in the context of blockchains is computational cost, not only in terms of CPU time but also in terms of transmission time as the underlying cryptographic schemes involve numerous key exchanges, which rely on Internet communications. Concerning homomorphic encryption, it was stated that it allows data owners to encrypt data in a way that other parties can perform certain types of operations without seeing the actual data. The party that performed the operations can then return the encrypted results to the

data owner. This means that data owners can benefit from advanced data analysis done by competent data scientists without having to disclose the data. Sketches are high-level data summaries that can be produced for the purpose of performing a specific computation. The data, according to the interviewee, is aggregated in a way that individuals are not identifiable but that selected advanced statistical or data analytics operations are possible. There is some data quality loss but it can be controlled. A key challenge in the context of sketches is to build summaries with minimal data quality loss.

Additionally, the interviewee suggested a multidimensional measure for privacy and related technologies. The measure should not only represent the degree of protection of the data against breaking privacy but also the cost of the protection in terms of how much harder it makes the data to be used, if the data protection scheme is used, and the societal value of the data. The interviewee emphasised the role of the societal value of data as many data protection mechanisms and governance rules seem to be built under an assumption of zero tolerance. The value of data is not taken into account and, for instance, a person's shoe size is dealt with just as the person's genetic data. This is closely related to the relevance of data classification mentioned by another interviewee (I8). The value of data is often not taken into account and very pessimistic assumptions are made, according to the interviewee who is professor at a North American university.

Another interviewee (I3) stated that all classes of technologies are very important. He explained that it is important to acknowledge that they are used in different contexts and often need to be combined. All classes contribute to privacy preservation in different ways. The interviewee stressed that the fact that an organisation uses anonymisation does not mean it doesn't need accountability and transparency. Moreover, the technologies correspond to different trust models. Some of the technologies, such as MPC, aim at overcoming the need for trust in other parties. Others, such as access control, policy enforcement, accountability and transparency aim at increasing trust in other parties. The interviewee, who is associate professor focusing on technologies to protect privacy, explained that the conceptions differ a lot: in *data protection*, the data controller is not only a trusted party but also responsible for protecting people's privacy, while in the context of *privacy technologies*, systems are designed in a way that no trusted parties are needed.

Another professor focusing on privacy (I5) also pointed out within the scope of an interview that it is very useful to combine the technologies. Explicitly, he suggested to combine technologies for transparency, data provenance, portability and user control to achieve an architecture or a tool that allows data subjects to trace where his or her data has gone and for what purposes it has been used. The interviewee, however, admitted that it would be very hard to develop such an architecture or tool. A key challenge is that erasure becomes very difficult when there are many copies of the data at several locations, particularly if the copies are in different countries with different jurisdictions and held by different organisations. Another key challenge the interviewee referred to is addressing the fundamental tension between the objective of big data, which is, according to the interviewee, collecting a lot of personal data and using it, and privacy, which is about protecting personal data. The interviewee stressed that for him big data is almost solely composed of personal data. Phrased that way, the two goals are almost like antagonists, according to the interviewee.

2.2. Issues

Figure 2 provides an overview of the ethical and societal issues and values taken into account. D2.2 provides a detailed description of the issues and values.

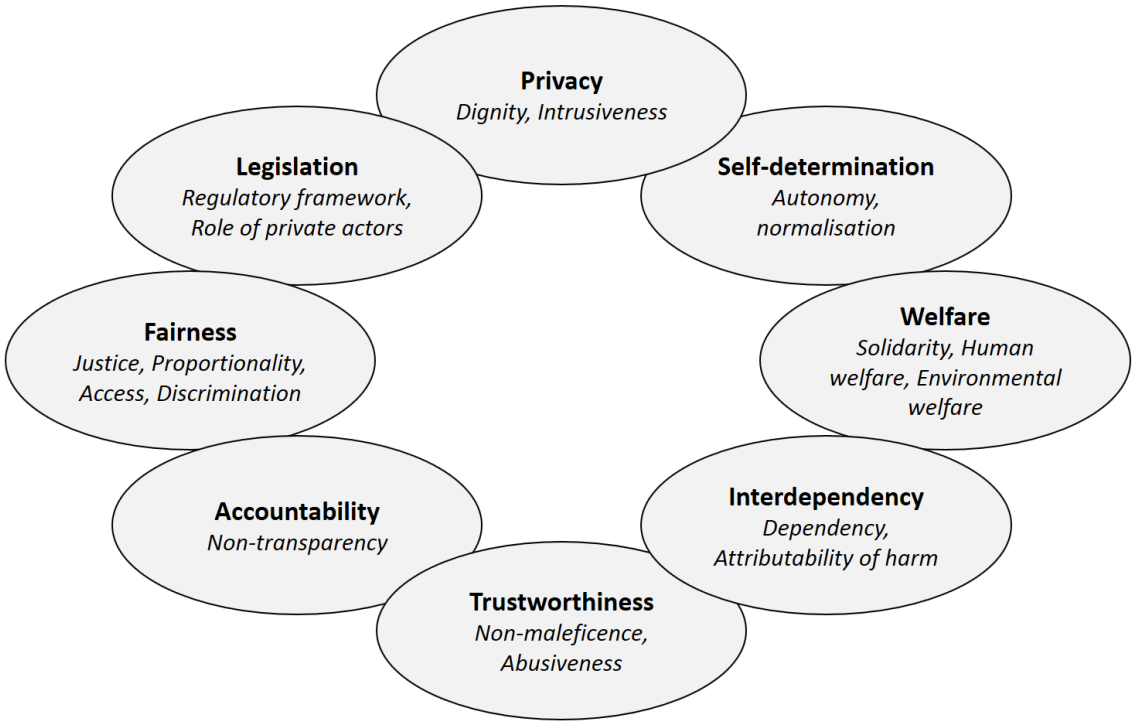

Figure 2 Overview of the issues and values

Privacy refers to a claim, an entitlement or a right of an individual to determine what information about himself or herself can be communicated to others. Human dignity includes both self-respect and respect towards all humans by humans without any interests. The intrusion into peoples' privacy and organisations' business practices is perceived as problematic. Big data has integrated itself into nearly every part of people's online life and to some extent also in their offline experience.

Self-determination describes the free choice of one's own acts without external compulsion. Big data-driven profiling practices can limit free will, free choice and be manipulative in raising awareness about, for instance, news, culture, politics and consumption. Thereby, they impede autonomy. The pressure towards conformity is referred to as normalisation. This restricts the breadth of choices, and pushed back pluralism and individuality.

Welfare is concerned with the general state of health or the degree of success of a person. Big data-based calculations in which commercial interests are prioritised rather than non-profit-led interests, are examples of situations in which solidarity is under pressure. Detrimental implications can emerge in the contexts of employment, schooling or travelling by various forms of big data-mediated unfair treatment of citizens and adversely affect human welfare. Big data has indirect effects on the environment.

Interdependency refers to the dependence of two or more people or organisations on each other. The dependency of people and organisations on organisations and technology leads to a limitation of

Grant Agreement number: 731873

flexibility. Organisations are strongly dependent on the data as well as the big data technologies they use. Due to the fact that the application of big data technology is not a single linear process, but consists of different stages, with different actors involved, the harms connected to it can have an incremental character that is not only difficult to articulate but also difficult to attribute to any given stage or actor.

Trustworthiness describes the ability to be relied on as honest or truthful. Data reuse in the world of big data can have diverse detrimental effects for citizens. This puts non-maleficence as a value under pressure. The risk of abuse is not limited to unauthorised actors alone but also to an overexpansion of the purposes of data use by authorised actors.

Reports based on the data from trust barometers show that levels of customer trust in companies have been declining for years.[3] One of the reasons for lower customer trust is the intense use of personal data by companies. Trustworthiness is a substantial determinant for consumers' willingness to share their data. Figure 3 shows that primary care doctors and payment or credit card companies are widely considered as "completely trustworthy" or "trustworthy" by consumers.[4] Internet giants such as Google and Yahoo, governments, media and entertainment companies, and social media firms are considered as trustworthy by a considerably smaller share of consumers.

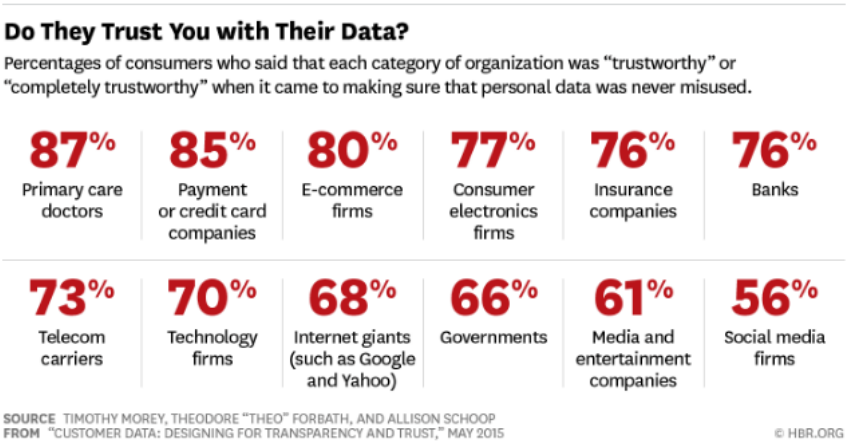

Figure 3 Trustworthiness of different types of organisations

Accountability refers to the properties that ensure that the actions of a person or organisation can be traced uniquely to the person or organisation. For instance, consumers often do not know who to turn to when their data is shared via surveys for research and marketing purposes. There is a lack of transparency with respect to organisational algorithms and business practices. Algorithms, for instance, are often like black boxes to average citizens, they are not only opaque but also mostly unregulated and thus perceived as incontestable.

[3] Neil Davey, "Customer data collection: How to be trustworthy and transparent,", https://www.mycustomer.com/marketing/data/customer-data-collection-how-to-be-trustworthy-and-transparent (accessed March 17, 2018)
[4] Timothy Morey, Theodore Forbath and Allison Schoop, "Customer Data: Designing for Transparency and Trust," *Harvard Business Review* 93, no. 5 (2015)

Fairness is characterised by impartial and just treatment or behaviour without favouritism or discrimination. Unfairness puts constant pressure on the value of justice. As none of the relevant rights is of absolute character, in cases of conflict with another right or interest, a right can be limited pursuant to the principle of proportionality. Not everybody or every organisation is in the same starting position with respect to big data. Discrimination is understood as the unfair treatment of people and organisations based on certain characteristics.

As Figure 4 shows, when the data is used for improving a product or a service, consumers generally perceive it as a fair trade for their data. But consumers expect more value in return for data used to facilitate targeted marketing, and the most value for data that is sold to third parties. The findings come from a survey by Morey et al.[5], which was published in 2015. For consumers, it is an important factor to receive value in return from the companies, which use their data.

Figure 4 How people swap value for data

Legislation is a law or set of laws suggested by a government and made official by a parliament. The current legal framework for protection of human rights is considered to display a number of vulnerabilities in the context of big data applications. Legislation provides several provisions which shift the obligation of balancing different fundamental rights and interests on private actors, and it does so without providing

[5] ibid.

any guidance. Opting for such a solution is problematic, as it raises the question of whether such decisions are legitimate, especially where the obligations of private actors are coupled with lack of transparency.

The interviewees did not mention any additional issues. Within the scope of the interviews, there was a clear focus on privacy.

3. Assessment of selected technologies

This section describes the results of a technology-specific assessment. Particular attention is paid to the societal and ethical issues that can be addressed by the technologies, the effectiveness of the technologies in addressing the issues, and the problems that may arise when addressing the issues with the technologies.

3.1. Anonymisation and sanitisation

This section focuses on technologies that help removing personally identifiable or other sensitive information from datasets.

The notion of *personal data* is very wide. Personal data is any information that relates to an identified or identifiable natural person.[6] Developments in data technologies increasingly allow natural persons to be re-identified from anonymised data. The emergence of artificial intelligence in general and machine learning in particular is particularly relevant in this context. *Anonymous data* is defined as any information from which the person to whom the data relates cannot be identified, neither by the organisation processing the data nor by anybody else. Under the EU data protection law, data can only be considered anonymous if re-identification is impossible. Common anonymisation techniques are:[7]

- **Noise addition** means that imprecision is added to the original data. For example, a doctor measures a person's weight correctly but after noise addition it shows a weight bandwidth of +/-10 kg.
- **Substitution** means that information values of the original data are replaced with other values. For example, instead of indicating a patient's height with 1.65 m, the value is substituted by the word "blue". If patient's height is 1.70 m, it is registered as "yellow".
- **Aggregation** focuses on making sure that an individual cannot be singled out. Therefore, individuals are grouped with several others sharing, for instance, their residency or their age. For example, a dataset does not cover the inhabitants of Munich with certain characteristics, but the inhabitants of Bavaria. *K-anonymity*, for instance, is a form of aggregation that, on the one hand, impedes re-identification by removing some information but, on the other hand, ensures that the data can still be used. Methods of K-anonymity are data suppression and data generalisation.[8]
- **Differential privacy** is relevant if an organisation gives a third party access to an anonymised dataset. While a copy of the original data remains with the organisation, the third-party only receives an anonymous dataset.[9]

Pseudonymous data is also mentioned in EU data protection law. It is characterised by the reduced linkability of the dataset with the original identities of individuals. Unlike anonymous data, where re-

[6] Cédric Burton and Sára Hoffman, "Personal Data, Anonymization, and Pseudonymization in the EU,", https://www.wsgrdataadvisor.com/2015/09/personal-data-anonymization-and-pseudonymization-in-the-eu/ (accessed March 27, 2018)
[7] ibid.
[8] ibid.
[9] ibid.

identification is impossible, pseudonymous data allows some form of indirect or remote re-identification.[10] Common pseudonymisation techniques are hashing and tokenisation.

Technologies for anonymisation and sanitisation are certainly essential to deal with **privacy**. However, anonymisation and sanitisation technologies also play a key role with respect to **legality**. Data protection law, above all the new EU General Data Protection Regulation (GDPR), clearly states what measures have to be taken by the affected organisations.

There is a substantial amount of literature on the effectiveness of the technologies as well as problems that may arise when using them. The key problem is to determine the optimal balance between improved privacy protection through anonymisation and sanitisation, and the usefulness of data for decision making. Acquisti and College[11], for instance, state that technologies can be used to protect, anonymise or aggregate data in ways that are effective and efficient. Effective means, in this context, that re-identifying individual information becomes either impossible or costly enough to be unprofitable. Efficient means that the desired transaction can be completed with no additional costs for the involved parties. Acquisti[12] points out that identity management systems allow sharing information that, for instance, is needed to provide personalised and targeted services, while personal data in general remains protected.

One of the interviewees (I3) stressed that it is important to differentiate between the anonymisation of datasets that include personal data to make sure that it is not possible to identify the individuals the data related to, and anonymisation to prevent the collection of data that is identifiable. According to the interviewee, there are technologies that provide quite strong guarantees concerning the prevention of personal data collection. Tor was mentioned as a quite robust example. The interviewee emphasised that the situation is more difficult when it comes to data anonymisation. He explained that it is not just about providing a functionality for privacy preservation that is similar to one for enhancing security. The difficulty is caused by the fact that, on the one hand, the data should be utilised for data mining and extracting value, and, on the other hand, the re-identification of the data should be at least very hard, if not impossible. Key problems are, for instance, the uniqueness of certain characteristics or behaviours, and the fact that it is usually unknown what other information is available to a potential adversary. Sometimes it might be that there is no satisfying trade-off: either some utility and very weak privacy, or some privacy and hardly any utility. In contrast, in anonymous communication systems, the data is only needed to connect the source and destination. According to the interviewee, an associate professor focusing on technologies to protect privacy, technologies based on differential privacy are most promising at the moment. Concentrating on sanitisation only, is considered very weak. Sanitisation is good only to prevent accidental disclosure, not to provide protection from a motivated adversary.

An interviewee who advises a national data protection authority (I1) stated that anonymisation and sanitisation technologies represent the most important class of technologies in the context of big data

[10] ibid.

[11] Alessandro Acquisti and Heinz College, "The Economics of Personal Data and the Economics of Privacy: 30 Years after the OECD Privacy Guidelines,", Background Paper 3 (OECD, 2010), https://www.oecd.org/sti/ieconomy/46968784.pdf (accessed April 23, 2018)

[12] Alessandro Acquisti, "Identity Management, Privacy, and Price Discrimination," *IEEE Security & Privacy* 6, no. 2 (2008)

and privacy. The interviewee explained that the mathematical background for anonymisation is established and very stable. This made him conclude that it is the right moment to foster the implementation of these technologies. The results of researchers must now be transformed into software, according to the interviewee. Similarly, another interviewee, a professor focusing on machine learning, data and text mining, and privacy (I4) pointed out that anonymisation technologies are the most mature and most widely used class of technologies, and that the respective field is the most understood. Anonymisation technologies are followed by encryption technologies with respect to the mentioned attributes, according to the interviewee.

A company representative who participated in the series of interviews (I8) stated that the company he works for assists its clients with respect to GDPR compliance, for instance, if a test dataset for a new product or service based on real-life data is created. The company helps to make real-life datasets unidentifiable in a way that the datasets are still useful for the intended business purposes. The interviewee confirmed that achieving this objective is a challenge in most cases. An interviewee representing another company (I7) reported that he has experienced that anonymisation is particularly difficult with respect to medical data. The difficulties are caused, for instance, by free text in data that includes names, indirect descriptions of things such as diseases or treatments, or dates that can easily be cross-related with other data sources. Therefore, purpose limitation has particular relevance in this context.

A professor focusing on privacy who participated in an interview (I5) stated that anonymisation and sanitisation technologies are very relevant in the big data context. As data cannot be controlled anymore as soon as it has been released, it is important that measures are taken to reduce the risk that individuals are re-identified or that sensitive attributes are inferred. A key drawback is, according to the interviewee, that there are many examples of failure of anonymisation. It is considerably easier to re-identify data than to propose an effective approach to anonymisation. The interviewee stressed that there is still a lot of work to do in this field. Moreover, he pointed out that in some cases so much noise needs to be added to dataset to remove the privacy risk that the data's utility is gone. According to the interviewee, there is quite some attention on the topic in the research community. Another professor, who focuses on privacy, cybersecurity, Internet policy and telecommunications law (I6), highlighted that it is not unlikely that data that was anonymised at a specific point in time can be re-identified in the future, given various technological and social changes. Consequently, it would be necessary to re-examine the situation over time. This also shows that technological measures have to be complemented by organisational ones.

3.2. Encryption

Technologies that allow encoding information, so that only authorised parties can access it, are in the focus of this section.

Encryption is fundamental for the protection of data and privacy-preserving data analysis. It transforms data in a way that only authorised parties can read it, which is ultimately a strong protection measure for

personal data.[13] Encryption technologies play an important role in protecting and advancing the freedom of opinion and expression. The Special Rapporteur of the United Nations Human Rights Office of the High Commissioner (OHCHR) highlights the relevance of encryption to protect the right to privacy.[14]

Encryption technologies are particularly relevant with respect to the issues **privacy** and **accountability** but also with respect to **legality**. Technical measures to ensure accountability are typically based on encryption. Some laws require organisations to take reasonable measures to avoid data breaches. As stated above, encryption is fundamental in this regard.

There is a vast amount of literature on the effectiveness of encryption technologies as well as problems that may arise when using them. According to Roth, for instance, the battle between law enforcement and privacy is being fought over encryption.[15] Many private organisations have taken measures to enhance security and privacy. These measures include end-to-end encryption for digital communications, disk encryption and software updates to fill the security gaps.[16] But even where end-to-end encryption is used, the exchange of information can already be subject to judicially-ordered surveillance. As network metadata such as source and destination addresses cannot be properly encrypted, it remains available to government monitoring by appropriate judicial order.

According to UNESCO's Series on Internet Freedom, state actors should put more efforts in encouraging the use of encryption and related techniques, including financial subsidies for software and hardware development.[17] The encryption of user data through organisations reduces the probability of identity theft even if data is breached.[18] Searchable encryption, homomorphic encryption and secure MPC are promising technologies with a high interest for the research community.

In the era of big data, encryption is crucial to ensure the confidentiality and integrity of sensitive information. As the size of data is growing on massive scale and as cloud storage is increasingly used, encryption even for large data is a modern requirement.[19] According to Roth, encryption and key management should be considered the cornerstone of any data security strategy. ENISA (European Union

[13] Giuseppe D'Acquisto et al., "Privacy by design in big data: An overview of privacy enhancing technologies in the era of big data analytics," (ENISA, 2015), https://www.enisa.europa.eu/publications/big-data-protection/at_download/fullReport (accessed September 26, 2017)

[14] "Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression," (OHCHR, 2016), http://www2.ohchr.org/english/bodies/hrcouncil/docs/17session/A.HRC.17.27_en.pdf (accessed March 27, 2018)

[15] Kenneth Roth, "The battle over encryption and what it means for our privacy,", https://www.hrw.org/news/2017/06/28/battle-over-encryption-and-what-it-means-our-privacy (accessed April 23, 2018)

[16] "Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression"

[17] Joseph A. Cannataci et al., "Privacy, free expression and transparency: Redefining their new boundaries in the digital age," (UNESCO, 2016), http://unesdoc.unesco.org/images/0024/002466/246610E.pdf (accessed March 27, 2018)

[18] Acquisti and College, "The Economics of Personal Data and the Economics of Privacy"

[19] Abdullah Al Mamun et al., "BigCrypt for big data encryption," in *Proceedings of the 4th International Conference on Software Defined Systems*, 93–9 (Valencia, Spain: IEEE, 2017)

Agency for Network and Information Security) set up the following essentials for protecting data in big data environments:[20]

- Encryption of data in transit and at rest to ensure data confidentiality and integrity.
- Deployment of a proper encryption key management solution, considering all relevant devices.
- Consideration of the timeframe for which data must be kept; data protection regulations might require the disposal of some data due to its nature after a certain time period.
- Design of databases with confidentiality in mind; for example, any sensitive data could be contained in separate fields so that they can be easily filtered out or encrypted.

The main drivers for using encryption technologies, according to a study conducted by the Ponemon Institute for Thales e-Security[21], are shown in Figure 5. The primary driver is compliance with regulations and requirements.

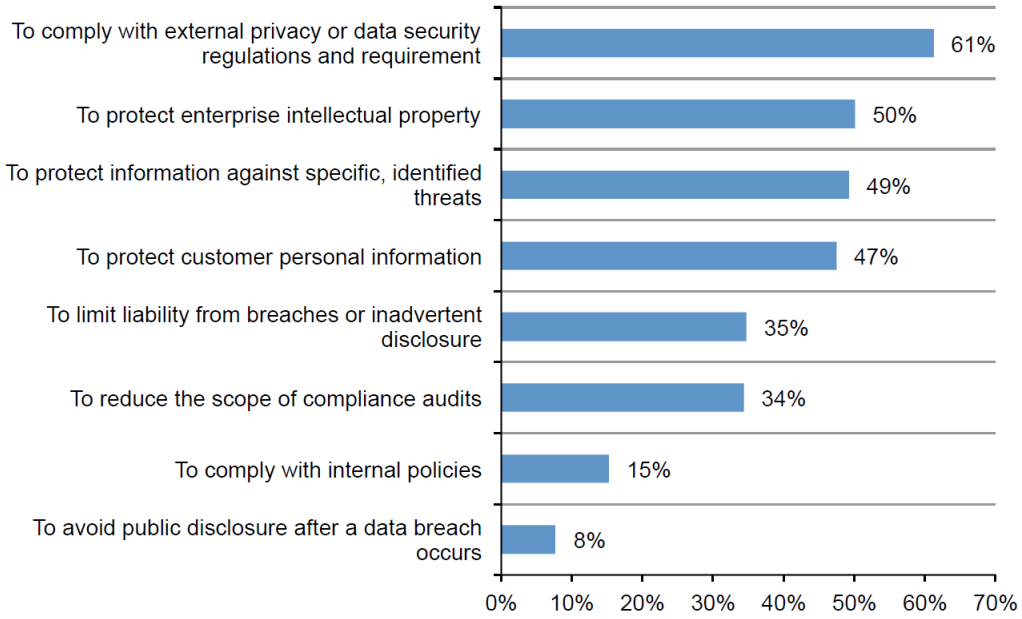

Figure 5 Main drivers for using encryption technology solutions[22]

However, while encryption is a cornerstone of data security, it is not sufficient in isolation. Rather, it is suggested to be tightly integrated with other security controls, including endpoint security, network security, application security and physical security systems, which are increasingly being run over IP-based networks.

[20] Rossen Naydenov et al., "Big Data Security: Good Practices and Recommendations on the Security of Big Data Systems," (ENISA, 2015), https://www.enisa.europa.eu/publications/big-data-security/at_download/fullReport (accessed April 24, 2018)

[21] "2016 Global Encryption Trends Study," (Thales e-Security; Ponemon Institute, 2016), https://www.ciosummits.com/Ponemon_Global_Encryption_Trends_Report_2016.pdf (accessed April 24, 2018)

[22] ibid.

A solution that is considered to maximise privacy and query expressiveness, at least theoretically, is Fully Homomorphic Encryption (FHE). It has the potential to substantially contribute to solving security and privacy problems in big data and to facilitate the adoption of cloud computing services. FHE is an emerging research field. Currently, the performance of FHE is not yet convincing. While some predict that feasible FHE solutions will be available already in the next decade, others assume that it will take at least 40 years until fully-implemented FHE technology is applied.[23]

A professor who participated in the series of interviews (I4) stated clearly that a key challenge with respect to homomorphic encryption is computation cost that makes it relatively slow compared to other methods. Therefore, further research must still be performed. The technology is particularly important for medical applications. The reason is that it does not lead to a loss in data quality and that is a large advantage. If zero quality loss is not the number one requirement, there are, according to the interviewee, other more interesting solutions. A technology advisor working for a national data protection authority (I1) stated that not FHE but semi-homomorphic encryption is more or less mature enough to be integrated into products. According to the interviewee, age verification, for instance, is mature.

An associate professor focusing on technologies to protect privacy (I3) explained more generally that encryption is generally very strong. However, it is important to keep the trust model in mind. As long as fully trusted parties are exchanging encrypted data and related keys, everything is fine. If the parties do not fully trust each other, encryption does not provide any protection. A representative of a technology company (I8) added that some of the current encryption methods will be useless in the era of quantum computing. The interviewee mentioned lattice-based cryptography as an interesting solution in this regard and pointed out that the US National Institute of Standards and Technology (NIST) is working on this topic.

3.3. Multi-party computation

Technologies supporting the distribution of data and processing tasks over multiple parties are discussed in this section.

The main goal of secure MPC is to provide an opportunity to perform distributed tasks in a secure way. Havron explains secure MPC as allowing "*two or more parties to compute a function on sensitive input data provided by both parties, without revealing anything about the inputs*".[24] MPC is often considered as the counterpart of sending encrypted data to a trustworthy third-party who would return the intended result. To avoid having to trust third-parties, MPC can be a key technology providing access to data while ensuring strong privacy protection.

According to Sadler[25], MPC can have a large impact on a number of areas. For instance, it could help unlocking the large quantity of health data that is currently inaccessible for study due to privacy concerns. MPC has already been used to evaluate gender pay disparities, detect tax fraud and prevent satellite

[23] D'Acquisto et al., "Privacy by design in big data"
[24] Samuel Havron, "Poster: Secure Multi-Party Computation as a Tool for Privacy-Preserving Data Analysis," in *Proceedings of the 37th IEEE Symposium on Security and Privacy* (IEEE, 2016)
[25] Christopher Sadler, "Protecting Privacy with Secure Multi-Party Computation,", https://www.newamerica.org/oti/blog/protecting-privacy-secure-multi-party-computation/ (accessed April 24, 2018)

Grant Agreement number: 731873

collisions. According to D'Acquisto et al.[26], often cited examples for practical implementations in which MPC has proved to be quite efficient are trading sugar beet by Danish farmers[27] and smart metering deployment in the Netherlands.[28] Nevertheless, depending on the MPC design and the number of involved parties, computing power and bandwidth use can still be potential obstacles. Additionally, the use of MPC typically requires the involvement of experts.[29]

Accordingly, MPC technologies are certainly most relevant for ensuring **privacy**. Moreover, MPC technologies are also relevant with respect to **trustworthiness**, as the misuse of data through one organisation is largely excluded.

Secure MPC is less secure than FHE, especially when many untrusted parties are involved but more efficient in certain types of implementations. There are many approaches to MPC including different cryptographic tools such as zero-knowledge proof, oblivious transfer and Yao's millionaire protocol.[30]

One of the interviewees, a senior scientist at a technology company (I7), stressed that MPC has a long history but that practical implementation problems stand in the way of wide adoption. Research projects can help by bringing the technology closer to a large audience by addressing the practical limitations, for instance, by providing algorithms that speed up common analytics methods. In terms of audience, the interviewee stated that it is important to go beyond the crypto community and also address companies, regulators and the popular press. The interviewee further pointed out that MPC is a technology that is useful to address a certain class of problems and that can unlock new possibilities in the context of joint data processing in domains such as medical research where legal and procedural hurdles are prevalent. Moreover, the interviewee compared the advantages and disadvantages of MPC and anonymisation. While anonymisation allows using standard analytics tools, the value of the data may be decreased by the anonymisation process. MPC allows working with the data as it is but restricts the efficiency of the analysis and the range of tools that can be used.

Moreover, the interviewee pointed out that the application domain is not relevant with respect to the effect of MPC on privacy. However, if the focus is on other objectives such as fairness, for instance, MPC can be particularly relevant in the financial context. The interviewee mentioned the role that MPC has been playing in the context of auctions. Finally, a reference was made by the interviewee to a practical implementation in Estonia in which MPC was used to check if there is a relation between Estonia's quickly growing IT industry and students failing at school. The Sharemind Platform, which is based on MPC and

[26] D'Acquisto et al., "Privacy by design in big data"

[27] Peter Bogetoft et al., "Secure Multiparty Computation Goes Live," in *Financial Cryptography and Data Security: 13th International Conference, FC 2009, Accra Beach, Barbados, February 23-26, 2009. Revised Selected Papers*, vol. 5628, ed. Roger Dingledine and Philippe Golle, 325–43, Lecture Notes in Computer Science 5628 (Berlin, Heidelberg: Springer Berlin Heidelberg, 2009)

[28] Benessa Defend and Klaus Kursawe, "Implementation of privacy-friendly aggregation for the smart grid," in *Proceedings of the 1st ACM Workshop on Smart Energy Grid Security*, 65–74 (ACM, 2013)

[29] Christopher Sadler, "Protecting Privacy with Secure Multi-Party Computation"

[30] George Danezis et al., "Privacy and Data Protection by Design - from policy to engineering," (ENISA), https://www.enisa.europa.eu/publications/privacy-and-data-protection-by-design/at_download/fullReport (accessed December 14, 2017)

offered by the Estonia-based company Cybernetica[31], was used in this context to perform a privacy-preserving statistical analysis on linked databases.

A professor focusing on privacy (I5) mentioned MPC as a key technology in his field of research. According to the interviewee, the technology is particularly relevant if a number of parties have a dataset but the parties do not trust each other to the extent that they would centralise the data at one of the parties or at a third party. The interviewee considers MPC technology as quite mature but also stated that, if it is used on a large scale or not, is a different issue, because this may also depend on political decisions. He stressed that some quite promising work has been done over the last decade to make the step from protocols that were very expensive in terms of computation and communication to very efficient protocols that can even be used in big data contexts.

Another interviewee, a technology advisor (I1), stressed that MPC is a promising field since it allows a better framing of the controllership responsibilities. According to the interviewee, a big problem in data protection is finding the right balance of responsibilities between controllers and processors. It needs to be defined who does what and who is responsible of what. The interviewee expects the introduction of MPC to make the relationship between controllers and processors smoother. He explained that MPC technology is not yet as mature as anonymisation technologies but that faster networks and increasing computation capacity increasingly allow transforming MPC concepts into practical implementations.

Finally, an associate professor focusing on technologies to protect privacy (I3) summarised that MPC provides very strong security guarantees but sometimes lacks performance. Guarantees are strong as only the result of the computation is revealed. The interviewee explained that there are solutions for specific problems that are efficient but there are no general solutions that can deal with everything. He also confirmed that there is quite an active research community focusing on the field of MPC.

3.4. Access control

This section focuses on technologies allowing the selective restriction of access to places or resources.

Big data applications typically require fine-grained access control. Access control consists of two main components: authentication and authorisation. Authentication is a technique used to verify that someone is who he or she claims to be. It must be combined with an additional layer of authorisation, which determines whether a user should be allowed to access the data.

It is each organisation's responsibility to ensure effective mechanisms for access control. This is especially important in non-traditional work environments, where employees do not only work from the office but also from home and while travelling. Access control gets even more difficult when multiple devices are used, such as (multiple) computers and laptops, but also tablets, smartphones and other IoT devices. Access control is a key component when it comes to prevent data breaches and it is in the hand of security

[31] https://cyber.ee/en/

professionals to ensure access control mechanisms are in place to prevent high-profile data breaches such as stolen passwords.[32]

Technologies for access control are certainly essential for **accountability**. However, the technologies also play a key role with respect to **legality**. Just as encryption, access control is fundamental when measures are taken to avoid data breaches.

There is a substantial amount of literature on the effectiveness of technologies related to access control as well as problems that may arise when using them. Martin[33], for instance, mentions five key challenges for enforcing access control:

- **Persistent policies**: Valid in hybrid environments with a large range of devices.
- Appropriate **control model**: Based on the type and sensitivity of data that is processed and operational requirements for data access.
- Consider **multiple solutions** for access control: Multiple technologies may need to be combined to achieve the desired level of access control.
- Authorisation as Achilles' heel: While the need for multi-factor authentication is taken seriously by most organisations, **authorisation** still is sometimes **neglected**. In addition, it needs constant monitoring with regards to the corporate security policy as well as operationally.
- Access control policies in **dynamic environments**: A sophisticated access control policy can be adapted dynamically to respond to evolving risk factors, in case of a breach enabling a company to link back to the relevant employees and data resources to minimise damage.

However, some scientists doubt that privacy can be adequately protected by means of access control. Kagal and Abelson[34] stress that access control in itself is inherently inadequate as a framework for addressing privacy on the Internet as in a pure access restriction system. Those who obtain access to the data, legitimately or not, can use the data without restriction. Instead of enforcing privacy policies through restricted access, the authors suggest focusing on helping users conform to policies by making them aware of the usage restrictions associated with the data and helping them understand the implications of their actions and of violating the policy, while simultaneously encouraging transparency and accountability in how user data is collected and used.

The authors suggest building the following principles into information systems:

- Users should be given due notice both in the case of collection and usage of their data, so that they can either take action to protect their privacy or voluntarily give it up in exchange for better service.
- Accountability including provenance, machine-understandable policies and policy tools.

---

[32] James A. Martin, "What is access control? 5 enforcement challenges security professionals need to know,", https://www.csoonline.com/article/3251714/authentication/what-is-access-control-5-enforcement-challenges-security-professionals-need-to-know.html (accessed April 10, 2018)

[33] ibid.

[34] Lalana Kagal and Hal Abelson, "Access control is an inadequate framework for privacy protection," in *Proceedings of the W3C Workshop on Privacy for Advanced Web APIs* (2010)

A company representative (I9) reported that the software and services company he works for has put a lot of engineering effort into developing an access control model that allows organisations to provision access in a highly granular way and in a highly dynamic way. As most of the company's clients have a legal basis for processing the data they have, anonymisation and sanitisation do not play a major role for them. They rather need support around access control, policy enforcement and accountability. Accordingly, the company the interviewee works for allows its clients to enforce their own policies effectively. With respect to granularity, the company allows granting access to data all the way down to the sub-cell level for data in tabular format. With respect to dynamics, access is granted for each session based on the user's role as well as the needs of the specific task he or she is performing. For instance, a case management system used to do policing work grants access not only based on the user's general status but also taking the severity of the crime into account he or she is focusing on. This allows restricting access to potentially privacy-invasive datasets in case of minor crimes.

Another interviewee, an associate professor focusing on technologies to protect privacy (I3), made a more general comment that is, among others, related to access control. According to him, some classes of technologies such as access control, policy enforcement, accountability and transparency rely very much on a strong trust model and thus have a single point of failure. There is usually one entity that is designing the access control system or specific policies, and responding to accountability and transparency requirements. Moreover, the opacity is often intensified by business secrets or certain business practices that are not completely open. According to the interviewee, this is typically not the kind of trust one appreciates when it comes to guarantees. In such cases, although sometimes cumbersome, non-technical things such as the legal system become relevant as complement or substitute for technologies. This is closely related to the discussion in section 4.4.

3.5. Policy enforcement

Technologies supporting the enforcement of rules for the use and handling of resources are discussed in this section. Automated policy enforcement mechanisms are particularly important in the big data era as policies get easily lost or neglected in the course of data being transferred between different systems. Data expiration policies, for instance, are already enforced by some big data solutions.

There is quite some literature discussing the effectiveness of policy enforcement technologies as well as problems that may arise when using them. According to Inukollu, for instance, big data security and privacy are challenges for both users and service providers.[35] In order to enforce sensitive data security policies, a policy enforcement framework has to be flexible and support different data processing requirements.

Policy enforcement technologies are particularly relevant not only in the **privacy** context but also in the context of **trustworthiness** and **accountability**.

One of the interviewees (I1), stressed that alongside anonymisation technologies, technologies for policy enforcement play a key role in the context of big data. Enforcement, according to the interviewee, who is an advisor for a European national data protection authority, is very relevant because the chain of

[35] Venkata N. Inukollu, Sailaja Arsi and Srinivasa Rao Ravuri, "Security Issues Associated with Big Data in Cloud Computing," *International Journal of Network Security & Its Applications* 6, no. 3 (2014)

responsibilities becomes longer and the roles more geographically dispersed. Moreover, it is very hard for data protection authorities to exercise their enforcement power, particularly if actors outside the EU are involved. Technical instruments, which assure proper functioning of things and in such a way offer the possibility of self-enforcement, are thus considered to be highly relevant. According to the interviewee, policy enforcement is not yet mature.

3.6. Accountability and transparency

This section focuses on technologies that help evaluating the compliance with policies, and explicating information collection and processing. Both accountability and transparency are often considered key prerequisites for trust.

Technologies for accountability and transparency are certainly particularly relevant with respect to the issues **trustworthiness** and **accountability**. Moreover, accountability technologies are also relevant concerning **legislation**.

According to Boujemaa[36], it is often assumed that big data techniques are unbiased because of the volume of the data and because the techniques are implemented through algorithmic systems. Boujemaa states that it is a mistake to assume that they are objective simply because they are data-driven – a view often referred to as data fundamentalism. Bias-related problems can impact the accuracy of big data applications and thus people's lives. Data inputs can become a problem because of:

- poorly selected data
- incomplete, incorrect or outdated data
- data that disproportionately represents certain populations
- malicious attacks

The design of algorithmic systems and machine learning can become a problem because of:

- poorly designed matching systems
- unintentional perpetuation and promotion of historical biases
- decision-making systems that assume correlation implies causation

Therefore, more and more tools and methods are developed to build trust, particularly over accountability and transparency for data and algorithms.[37] Implementing the responsible-by-design principle, means that issues such as fairness, equity, loyalty and neutrality are taken seriously and are addressed.

With decision responsibility in mind, Boujemaa lists several criteria related to accountability and transparency that big data applications should meet:

- decision explanation and tractability
- robustness to bias, diversion and corruption
- careful software reuse

[36] Nozha Boujemaa, "Algorithmic Systems Transparency and Accountability in the Big Data Era,", https://de.slideshare.net/NozhaBoujemaa/nboujemaa-datadrivenparis (accessed March 17, 2018)
[37] ibid.

Transparency is one of the most important issues in order to minimise problems associated with privacy. Companies should be upfront with the subjects that it is working with and be fully transparent to customers, by letting them know what the company knows and what information it receives about them. Being transparent to customers does not mean giving away strategic secrets, otherwise company's competitiveness will be abraded. Therefore, it is important to carry out proper methods to find a balance between transparency and confidentiality. For example, the company's IT department can launch pro-active communication campaigns, which could include PR, speaking, social media, and outreach programs by explaining more about what and how the company performs data-related activities. The company can also share its prediction accuracy, which will help with reasoning in the absence of methodology. For example, if the company is merging their customers' Facebook data with Twitter data to better understand their interests, this information should be shared with the customers. This level of transparency will not erode a company's competitiveness, rather it will build trust with the community.

In addition to arguments regarding the limitations of transparency, there are some opinions that the problems and any potential harms of big data analytics do not rise from how data is collected but rather from how data is used. For example, some commentators pondered the emerged harm to an individual by the collection of their data, for instance through tight surveillance by governments. A growing focus on the use of data has led to the preponderance of accountability as an answer to big data issues, as contrary to transparency. Instead of providing users with the information "how" and "why" their personal data is processed, accountability concentrates on monitoring its use through mechanisms such as auditability[38], technical design of algorithms[39] and software-designed regulation[40].

Trust and transparency are interrelated elements. The more transparency about the use and protection of their customers' data a company provides, the more it reinforces trust.

Accountability and transparency were considered particularly important by a research associate focusing on transparent computer systems (I2). The interviewee emphasised that it is essential to understand what exactly is done with the data. Transparency, according to the interviewee, is currently a big issue in the IoT context, and the concept should also include machine learning interpretability. Many machine learning modules that drive IoT devices are black boxes. Such devices may perform in an undesired way and it's almost impossible to find out why. The performance is not interpretable by non-experts. The interviewee highlighted, however, that there is a growing field of research that aims to explain why certain decisions are made. Companies such as Google, Apple, Facebook and Amazon (GAFA) are considered to be particularly interested. A key problem is that most IoT devices do typically not have a screen. This makes it difficult to let them explain what they do in a visual manner. Moreover, records of the interactions between devices need to be stored somewhere, ideally at an independent and trusted party. Finally, a link to the GDPR was established as it requires organisations to explain what they are doing with data. The

[38] Nicholas Diakopoulos and Sorelle Friedler, "How to hold algorithms accountable," *MIT Technology Review* (2016)
[39] Freek Bomhof, "In Order to Trust Big Data, Transparency Is Not Enough,", https://datafloq.com/read/transparency-in-big-data-is-not-enough/138 (accessed December 21, 2016)
[40] Hemant Taneja, "The need for algorithmic accountability,", https://techcrunch.com/2016/09/08/the-need-for-algorithmic-accountability/ (accessed December 21, 2016)

interviewee concludes that corresponding accountability and transparency technologies should already be quite mature.

Similarly, a privacy and civil liberties engineer at a software and services company (I9) stated that it is fundamental to be able to audit code that transforms data. Of particular importance are the weights that are being given to certain features. Without doubt, reviewing them is much more difficult if the algorithms used are based on machine learning. However, according to the interviewee, the hype around machine learning in particular and artificial intelligence in general has outpaced the reality in terms of what can be implemented in a production environment. While it is relatively easy to train a model in an experimental setting, it is very hard to get a machine learning model working in an industrial setting, and especially to maintain it over time.

A professor focusing on privacy who participated in the series of interviews (I5) stated that accountability and transparency technologies are particularly relevant in the big data context. For instance, if a classifier is trained using a large dataset, it may be necessary to ensure that the classifier does not discriminate against a particular group in the population. Doing this would require a measure for fairness as well as a way to ensure that the learning is fair by design. Moreover, explanations of certain decisions may be needed. For example, if the classifier is used to decide based on profiles whether somebody receives a loan or not, then it would be good to have an explanation on the respective decision. The interviewee does not consider accountability and transparency technologies mature but sees that efforts are making progress in this regard.

Finally, an associate professor focusing on technologies to protect privacy (I3), stated that transparency is not only very important but also very challenging to achieve. Particularly in the big data context, explaining what algorithms do with data or previewing what would be the outcome of providing data is extremely difficult.

3.7. Data provenance

Technologies helping to attest the origin and authenticity of information are in the focus of this section. The importance of data provenance has been increasingly recognized by both users and publishers of data. For users of data, the basis of their analysis relies largely on the credibility and trustworthiness of their input data. For publishers of data, the provision of provenance as part of their published data is an important factor that determines the value of the data. In today's Internet era, where complex ecosystems of data are even more prevalent, data provenance has become a key topic of research.

Technologies for data provenance are particularly relevant with respect to the issues of **trustworthiness** and **accountability**.

The effectiveness of data provenance technologies as well as problems related to the technologies have been addressed in literature to some extent. Glavic[41] states, for instance, that the information that data provenance provides is useful for debugging data and transformations, auditing, evaluating the quality of and trust in data, modelling authenticity, and implementing access control for derived data. Provenance

[41] Boris Glavic, "Big Data Provenance: Challenges and Implications for Benchmarking," in *Proceedings of the 2nd Workshop on Big Data Benchmarking*, 72–80 (2012)

in the context of big data is also referred to as big provenance. Its investigation has started around the beginning of the decade.

According to Wang et al.[42] , the challenges that are introduced by the volume, variety and velocity of big data, also pose related challenges for provenance and quality of big data, defined as veracity. The increasing size and variety of distributed big data provenance information bring new technical challenges and opportunities throughout the provenance lifecycle. Wang et al. mention the following challenges:

- The provenance data from big data workflows is **too large**. To get a fine-grain provenance tracking of a workflow execution, the recorded provenance could easily be several times larger than the original data to be processed.
- Provenance collection **overhead** during workflow execution is too much. There is always an execution overhead when recording provenance on top of the computation cost related to the analysis. This overhead problem often gets worse for big data workflows due to their distributed nature. A challenge is to minimise the provenance collection overhead.
- It is hard to **integrate** distributed provenance. The provenance of user-defined Map/Reduce functions running on big data systems is often initially saved on distributed non-permanent nodes. The information collected needs to be either communicated as the analysis is happening or stitched together in the end. The first choice generates a lot of communication overhead but is useful to monitor the application progress. The second choice is more efficient but requires an additional step to upload the information before freeing the computation nodes. The stitching of the data to be centralised in both choices requires additional integration steps.
- It is hard to **reproduce** an execution from provenance for big data applications. Many existing provenance systems only record intermediate data generated during execution and their dependencies. Execution environment information, which is also important for reproducibility, is often neglected. Execution environment information includes the hardware information and parameter configurations of big data engines. We have found this information is not only critical to execution performance but also could affect the final results.

Apart from Wang et al., Cuzzocrea[43] discusses big provenance challenges. He briefly outlines 14 challenges from accessing big data to minimising computational overhead to secure and privacy-preserving provenance. Two of the challenges mentioned focus on tools. Cuzzocrea states that flexible provenance query tools and provenance visualisation tools are needed. Among the concrete solutions mentioned by both Wang et al. and Cuzzocrea are the Hadoop extensions Reduce and Map Provenance (RAMP) and Hadoop-Prov as well as the hybrid big provenance systems Pig Lipstick.

A company representative (I9) pointed out that the company he works for offers products that allow not only integrating data from different sources but also preserving the sources of all components in the unified model. Each property that is related to an object can come from a different source. The sources

[42] Jianwu Wang et al., "Big data provenance: Challenges, state of the art and opportunities," in *Proceedings of the 2015 IEEE International Conference on Big Data*, 2509–16 (IEEE, 2015), http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7364047 (accessed December 14, 2017)

[43] Alfredo Cuzzocrea, "Big Data Provenance: State-Of-The-Art Analysis and Emerging Research Challenges,", http://ceur-ws.org/Vol-1558/paper37.pdf (accessed December 14, 2017)

can also be taken into account in the context of access control. The interviewee stressed that the company he works for puts a lot of development effort into notions of data provenance. The company does not only offer products that maintain relationships and dependencies between states or transformations of the data but also support versioning. Versioning is done not only of the data itself but also of the code used to transform the data. The company also includes accountability at this point by attributing the writer of the code. The interviewee underlined the tight connection between data provenance, accountability, access control and policy enforcement. The measures taken by the company with focus on data provenance allow users to comprehensively investigate certain values that seem wrong. They cannot only check if, how and when errors were introduced but also implement a fix and, based on the provenance tree, rebuild the dataset with the corrected values. An important issue, according to another interviewee (I3), is to have assurance that data is not fake.

3.8. Access, portability and user control

This section focuses on technologies that facilitate the use and handling of data in different contexts. Moreover, it discusses technologies for the specification and enforcement of rules for data use and handling.

Empowering users to access their data is fundamental for privacy protection as well as an obligation for data controllers. The scope of users' access right is specified in Article 15 of the GDPR. Accordingly, related technologies are particularly relevant for **privacy** and **legality**. Technologies for portability and access control are highly relevant for the issues **interdependency**, **self-determination** and **trustworthiness**. Apart from access, portability is a prerequisite for interdependency. Technologies for user control have the potential to increase self-determination and trust.

There is quite some literature on access, portability and user control. Empowering users, informing them and giving them access to their data is not only for the users' benefit but also for the benefit of the organisation using the data. Nevertheless, users do not practice this opportunity very often. This may be because they are not informed about the opportunity, because the process is too complicated or because they do not care. According to the European Data Protection Supervisor (EDPS)[44], the situation might change if individuals get the chance to benefit from accessing their data in a tangible way. Such a benefit could be accomplished by providing access as a service feature to users, instead of as an administrative burden. The format of the system should be portable, user-friendly and machine-readable.[45] In this context, user access to online banking information is considered as a good example.

Data portability, which is closely related to data access, is also highly relevant for users as it enables them to change their service providers without losing their data. In this regard, there are some remarkable initiatives described in the literature, such as the Midata initiative[46] in the United Kingdom (UK), providing

[44] "Meeting the challenges of big data: A call for transparency, user control, data protection by design and accountability,", Opinion 7/2015 (European Data Protection Supervisor (EDPS), 2015), https://edps.europa.eu/sites/edp/files/publication/15-11-19_big_data_en.pdf (accessed March 28, 2018)

[45] "Big data, artificial intelligence, machine learning and data protection," (Information Commissioner's Office (ICO), 2017), https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf (accessed March 28, 2018)

[46] https://www.gov.uk/government/news/the-midata-vision-of-consumer-empowerment

access to transactions and consumption for energy, finance, telecommunications and retail sectors. Another example is the French MesInfos platform[47], designed for the access to financial, communication, health, insurance and energy data.[48]

Data portability can be supported through personal data stores (or personal information management services), which hold the re-use of individuals' personal data under their control. These are third-party services holding users' data on their behalf and making it available to organisations by the permission of individuals.[49] The EDPS considers these stores as an alternative to deal with users' concerns regarding the loss of data control.[50] Obar does not consider the control by individuals over their personal data reasonable. Instead, he suggests a *"representative data management"*, which is a system of intermediaries that manages user data on behalf of the users.[51]

There are some examples of public acceptance for personal data stores in the UK. For instance, survey results of the Digital Catapult Centre show that 30% of people would be ready to use a service to assist them by collecting, managing and preserving their personal data. In this regard, Mydex is a good example in the UK, which provides free encrypted personal data stores. By applying for a service of Mydex, people can share data from their personal data store and also verify their digital identity.[52]

According to D'Acquisto et al., the "*big data analytics industry and the data controllers need to work on new transparency and control measures, putting the individuals in charge of the processing of their data*".[53] Data protection authorities should support these efforts.

A university professor who participated in the series of interviews (I6) referred to basic problems with the information economy when talking about user control. The interviewee considered the informational asymmetry between data subjects and data collectors as a central problem. According to the interviewee, the fact that an individual gets control over his or her data does not resolve the asymmetry. The data subject cannot understand what his or her data can be used for and how certain database transactions will end up being inefficient or unfair. The fact that the individual has greater control does not change this. The interviewee also thinks that providing the data subject with more information to make the consent more informed does not really change much.

A technology advisor (I1) underlined that portability is an opportunity for the market that fosters competition. However, the interviewee also mentioned that portability can create problems as it leads to the duplication of data. Data may be brought from one domain where there are safeguards to another domain that is more risky. Moreover, the interviewee states that portability needs trust between the involved parties. A user needs to be sure that whenever he or she brings data from one place to another, the data is processed according to his or her expectations. This cannot be done without a prior agreement

[47] http://mesinfos.fing.org/english/
[48] D'Acquisto et al., "Privacy by design in big data"
[49] "Big data, artificial intelligence, machine learning and data protection"
[50] "Meeting the challenges of big data"
[51] Jonathan A. Obar, "Big Data and The Phantom Public: Walter Lippmann and the fallacy of data privacy self-management," *Big Data & Society* 2, no. 2 (2015)
[52] "Meeting the challenges of big data"
[53] D'Acquisto et al., "Privacy by design in big data"

between the parties. As service providers do not only face a higher risk of losing clients but also see the opportunity to gain clients, they will most likely accept the paradigm of portability in the end.

Similarly, another interviewee (I3) considers portability good for competition. Access is seen by the interviewee as closely related to accountability. The interviewee has concerns with respect to user control. It is certainly great to empower people but to some extent it also means that responsibility is pushed to them. Therefore, the interviewee is in favour of designing systems in ways that make it difficult for users to endanger themselves rather than just giving them controls and expecting them to know how to use them. Even for experts it is sometimes difficult, according to the interviewee, to understand what the outcome of certain decisions will be, and thus non-expert users should not be expected to understand them and make meaningful decisions. This is also closely related to the discussion of responsibility in section 4.5.

One of the interviewees (I4) considered the work on tools that allow users to exercise better control over the uses of their data as important. He thinks that this is difficult, not only technically but also at the cultural level because many people are reluctant to take time and learn about the complexities and the language used to work with even a simple interface. A solution might be to create some models of individuals that have certain privacy rules decided and other could use these models as theirs. User control technologies are currently neither ripe nor user-friendly enough to be used by millions of people. The interviewee emphasised that it is critical that the research community comes up with ways of giving people the power over their data.

Another interviewee, a professor focusing on privacy (I5), had doubts with respect to technologies for user control. He argued that once data is released, asking for the consent of the data subject may be difficult. The interviewee stated that user control is eroded in the big data context, especially if big data is understood as trying to do analyses without knowing what the results will be used for in advance. In general, the interviewee emphasised that the right to erasure or the right to portability are very important. However, in the big data context, if, for instance, data was used to train a classifier, it is very hard to erase data and disregard the impact it had on computations or analyses.

4. General assessment of the technologies

This section describes the results of a more general assessment of the technologies. Particular attention is paid to the integration of the technologies in today's big data solutions, the demand for big data solutions that include the technologies, regional and cultural differences with respect to the availability and use of the technologies, the need for non-technical measures as a complement to technologies, and the responsibility for addressing the issues.

4.1. Today's solutions

This section discusses to what extent privacy-preserving technologies are integrated into today's big data solutions or what privacy-preserving add-ons are available. The section deals with the supply side and addresses commercial as well as non-commercial products. Ethical and societal issues other than privacy are also taken into account.

Hadoop and Spark are two fundamental big data technologies. While Hadoop provides processing data on disk, Spark processes data in memory. Both are open-source projects developed by Apache to handle big data. Each of them provides a distributed computing platform that allows fast, efficient, fault-tolerant and scalable processing of large datasets. In 2017, a big data anonymisation model based on Spark was proposed.[54] In the context of Hadoop, security and privacy issues were already discussed earlier.[55]

Most of the existing scientific literature that deals with privacy-preserving technologies focuses on technical details. The integration of technologies in today's big data solutions is typically not addressed at all. There are some publications, however, that discuss the pros and cons of technologies as well as the challenges faced and the ways to overcome them in specific scenarios. Such publications contribute to the understanding of the role of the technologies in practice. Among the scenarios addressed are, for instance, Smart Homes[56], healthcare[57] and outsourcing[58].

Literature on existing big data solutions that include specific technologies to preserve privacy is difficult to find. One of the rare exceptions was authored by AlMahmoud et al.[59] and proposes a privacy-preserving platform for collaborative spam detection named Spamdoop. Collaborative spam detection solutions can typically deal with large-scale e-mail data contributed by multiple sources. However, they have the well-known problem of requiring the disclosure of e-mail content. Spamdoop is a privacy-preserving big data

[54] Yavuz Canbay and Seref Sagiroglu, "Big data anonymization with Spark," in *Proceedings of the 2017 International Conference on Computer Science and Engineering (UBMK)*, 833–8 (IEEE, 2017)
[55] Karim Abouelmehdi et al., "Big data emerging issues: Hadoop security and privacy," in *Proceedings of the 2016 5th International Conference on Multimedia Computing and Systems (ICMCS)*, 731–6 (IEEE, 2016)
[56] Antorweep Chakravorty, Tomasz Wlodarczyk and Chunming Rong, "Privacy Preserving Data Analytics for Smart Homes," in *Proceedings of the 2013 IEEE Security and Privacy Workshops*, 23–7 (Piscataway, NJ: IEEE, 2013)
[57] Priyank Jain, Manasi Gyanchandani and Nilay Khare, "Big data privacy: A technological perspective and review," *Journal of Big Data* 3, no. 1 (2016)
[58] Nikolaos P. Karvelas et al., "Efficient Privacy-Preserving Big Data Processing through Proxy-Assisted ORAM," (IACR Cryptology ePrint Archive, 2014), https://eprint.iacr.org/2014/072.pdf
[59] Abdelrahman AlMahmoud et al., "Spamdoop: A privacy-preserving Big Data platform for collaborative spam detection," *IEEE Transactions on Big Data* (2017)

Grant Agreement number: 731873

platform built on top of a standard MapReduce facility. The platform uses a highly parallel encoding technique that enables the detection of spam campaigns in competitive times. Spamdoop, however, is still rather a research prototype than a solution to be used in practice.

The websites of relevant companies are more informative concerning the integration of privacy-preserving technologies in today's big data solutions than scientific literature. Thus, we carried out an analysis of the websites of companies that offer big data solutions. To be more precise, we especially looked at where the company resides, what it offers regarding privacy-preserving big data solutions and how the solutions are described. The results of this analysis are described below.

Examples for commercial products from **Privitar**[60], a UK-based privacy engineering company that offers privacy-by-design software products, are:

- Privitar Publisher: a software application, which anonymises sensitive data and creates a safe copy suitable for system development and testing, analytics, data science and machine learning, sharing with third parties and processing in cloud environments. It removes the identifying information while preserving the valuable patterns and relationships in large-scale data.
- Privitar Lens: a privacy-preserving query interface for the statistical analysis of sensitive datasets. It prevents direct access to the underlying sensitive data by dynamically applied privacy controls to each query submitted and thereby allows analysis of high-dimensional or longitudinal datasets.
- Privitar SecureLink: a data de-identification system that can be used by organisations to join data from many contributing organisations. The data is de-identified and the central organisation cannot recover the identifiers.

Relevant keywords and phrases on Privitar's website:

- engineering privacy
- privacy-by-design
- compliance
- privacy empowers innovation, businesses
- protection against data breaches, regulatory penalties, misuse of data

Another example of a commercial product is the platform developed by **trust-hub**[61]. The company is based in the UK. This platform offers tools to manage privacy and consent, encryption and secure data storage to support compliance and simplify data management. It offers secure storage, processing and rights management for personal data on a case-by-case basis and is designed to support data protection and privacy regulations including the GDPR. According to the developers, trust-hub works by incorporating data-protection-by-design into the platform and with multi-layered security built into the architecture to ensure data is protected across all components.

[60] https://www.privitar.com/privacy-engineering-products
[61] https://www.trust-hub.com/

Relevant keywords and phrases on trust-hub's website:

- trust-hub's dynamic mapping process means you always know what data you hold and what you are doing with it
- advanced features to ensure the GDPR compliance processes that underpin your business are fair and lawful
- demonstrate your GDPR compliance; collect only the personal data that is required and store it securely for only as long as it is needed
- regulatory compliance

Further relevant companies that are based in the EU are DPO Consulting[62], Dawex[63], SynerScope[64], SAP[65], Pyramid Analytics[66] and BOARD International[67]. DPO Consulting and Dawex are based in France. While DPO Consulting offers consulting services and supports cross-border compliance with EU requirements on data privacy, Dawex offers a trusted third-party platform and allows to exchange data containing personal information in full compliance with regulations. SynerScope and Pyramid Analytics are based in the Netherlands. SynerScope offers a platform that finds structured and unstructured personally identifiable data in an organisation's systems with the aim to open it up for full control and compliance (including the GDPR). Pyramid Analytics offers business intelligence software for data analytics. Privacy and other ethical and societal issues are not explicitly mentioned on their website. Germany-based SAP offers comprehensive solutions for all business processes as well as various big data solutions. Privacy and other ethical and societal issues, however, are not explicitly mentioned. BOARD International is based in Switzerland. The company offers business intelligence and corporate performance management software. Again, privacy and other ethical and societal issues are not explicitly mentioned.

Another company that focuses on privacy, compliance and risk management solutions is US-based **Trust Arc**[68]. It offers a data privacy management platform with various components (Data Flow Manager, Assessment Manager, Consent Manager, Website Monitoring Manager, Individual Rights Manager, Dispute Resolution Manager, Ads Compliance Manager) paired with consulting and certification services on privacy compliance.

Relevant keywords and phrases on Trust Arc's website:

- privacy-compliance (also GDPR)
- risk management
- trust

[62] https://www.dpo-consulting.com/
[63] https://www.dawex.com/
[64] http://www.synerscope.com/
[65] https://www.sap.com/
[66] https://www.pyramidanalytics.com/
[67] https://www.board.com/
[68] https://www.trustarc.com/

A company with a similar focus that is also based in the US is **OneTrust**[69]. It provides a privacy management software platform to comply with privacy regulations including GDPR and the EU-US Privacy Shield. According to the company, its solutions include readiness and privacy impact assessments, data inventory and mapping automation, website scanning and consent management, subject rights requests, incident reporting, and vendor risk management.

Relevant keywords and phrases on OneTrust's website:

- GDPR compliance
- privacy-by-design
- consent management
- breach management
- accountability

Other companies that offer privacy-preserving solutions are PHEMI[70], Nymity[71], Anonos[72], Integris Software[73], Privacy Shield Framework[74], PeerNova[75], Zettaset[76], HyTrust[77], Code42[78] and BlueTalon[79]. While PHEMI is based in Canada, all of the other companies are based in the US.

PHEMI offers big data technologies to handle any volume and variety of data, while providing advanced features for data management, privacy and governance.

Relevant keywords and phrases on PHEMI's website:

- privacy, security & governance
- consent, data sharing agreements, de-identification, compliance

Nymity offers accountability, risk and compliance privacy solutions.

Relevant keywords and phrases on Nymity's website:

- privacy compliance (GDPR + world)
- accountability

Anonos offers solutions on data risk management, security and privacy.

[69] https://onetrust.com/
[70] https://phemi.com/
[71] https://www.nymity.com/
[72] https://www.anonos.com/
[73] https://integris.io/
[74] https://www.privacyshield.gov/
[75] http://peernova.com/
[76] https://www.zettaset.com/
[77] https://www.hytrust.com/
[78] https://www.code42.com/
[79] http://bluetalon.com/

Relevant keywords and phrases on Anonos' website:

- GDPR compliance
- data protection
- sharing to maximize value
- unstructured data
- risk-reduction

Integris Software offers an automated system for privacy management (including the creation, validation and maintaining of a company's data map, responding to subject rights requests, and managing corporate risks).

Relevant keywords and phrases on Integris Software's website:

- privacy intelligence
- regulatory compliance (including GDPR)
- risk reduction

Zettaset offers a data encryption platform for sensitive regulated and business information.

Relevant keywords and phrases on Zettaset's website:

- data protection
- encryption
- prevention from data breaches

Hytrust offers security, compliance and control software for virtualization of information technology infrastructure.

Relevant keywords and phrases on Hytrust's website:

- encryption
- policy and access control

Code42 offers solution that backs up distributed end-user data on a single, secure platform.

Relevant keywords and phrases on Code42's website:

- limit risks
- meet data privacy regulations
- recover from data loss

Blue Talon provides data-centric security, user access control, data masking and auditing solutions for complex, hybrid data environments.

Relevant keywords and phrases on Blue Talon's website:

- authorisation
- access control

As the above-mentioned examples show, selected existing commercial solutions already emphasise privacy and regulatory compliance. With the EU GDPR about to come into force, GDPR compliance is also prominently highlighted. In this context, the term *"risk reduction"* is also used quite frequently on the respective companies' websites, as well as the solutions' potential to *"maximise value"*. Details on the featured technologies are rarely mentioned on the websites. Exceptions are, for instance, the websites of trust-hub (transparency and accountability), One Trust (accountability) and Nymity (accountability).

Schonschek and Litzel[80] provide an overview of products focusing explicitly on big data encryption to protect confidentiality in a communication channel. They mention the family of Ethernet encryptors R&C SITline ETH provided by the German company **Rohde & Schwarz**. The encryptors were designed specifically for protecting the data centre and site-to-site connections against eavesdropping and manipulation. Another product mentioned is **Gemalto's** SafeNet ProtectFile, which provides transparent and automated file system-level encryption of server data at rest. The product allows, for instance, protecting sensitive data in Hadoop clusters. The US-based company SafeNet was acquired by Gemalto in 2014. Gemalto is based in the Netherlands and still uses SafeNet as a product brand. SafeNet offerings focus on authentication, encryption and key management. **Protegrity** is a US-based company dealing with enterprise security. The company's Big Data Protector provides data-centric security for Hadoop-based platforms. The product secures sensitive data utilising tokenisation and encryption at rest in the Hadoop Distributed File System (HDFS), in use during MapReduce, Hive and Pig processing, and in transit to and from other data systems. Further big data encryption products mentioned by Schonschek and Litzel are offered by Voltage Security, Zettaset and Gazzang, which was acquired by Cloudera in 2014. All of them are based in the US.

An example for a non-commercial product, which is still under development though, is **OPAL**[81]. It focuses on building a secured infrastructure to allow the use of data, while giving people strong guarantees that the data is used in a privacy-conscientious manner.[82] OPAL's core will consist of an open suite of software and open algorithms providing access to statistical information extracted from anonymised, secured and formatted data.[83]

A professor who participated in the series of interviews (I4) stated that the technologies are integrated minimally in today's big data solutions. The interviewee stressed that he had not seen them being used widely but that the Facebook situation was a shocking proof that indeed not much is done. He added that the lack of such technologies is a big impediment for the companies referred to as GAFA. Unfortunately, the Cambridge Analytica and Facebook incident may result in further reluctance of the GAFA and similar companies to share data. What is needed are privacy-preserving technologies that make sharing data safe. The GAFA companies as well as mobile phone companies have incredibly valuable data, which may in fact be the key to some medical breakthroughs. The North American interviewee concluded that he does not

[80] Oliver Schonschek and Nico Litzel, "Big Data Encryption – das große Verschlüsseln,", https://www.bigdata-insider.de/big-data-encryption-das-grosse-verschluesseln-a-484370/ (accessed March 16, 2018)
[81] http://www.opalproject.org/
[82] Mekhala Roy, "Data anonymization techniques less reliable in era of big data,", http://searchcompliance.techtarget.com/feature/High-dimensional-info-complicates-data-anonymization-techniques (accessed February 23, 2018)
[83] http://www.opalproject.org/

have the feeling that enough is done. Currently, all kinds of legal agreements are signed, but these agreements have, as the most recent incident showed, limited value. What is needed are technologies that protect the data but at the same time allow sharing it.

Another professor (I5) also stated that he does not see many privacy-preserving technologies being integrated in big data solutions. Encryption technologies, according to the interviewee, are quite mature and could be deployed. With respect to advances made in the field of homomorphic encryption in the last few years, the interviewee mentioned IBM as an example. Apart from encryption technologies, technologies for access control have also been used for a long time but they are not widely implemented, according to the interviewee. Concerning anonymisation, the interviewee mentioned Privacy Analytics[84] as an example. The Canadian company focuses on the anonymisation of medical data. The interviewee stated however that he considers anonymisation technologies to be not as mature as technologies for encryption, MPC and access control. He explained that encryption and access control have long been used in the security context. Although companies sometimes claim that data has been anonymised, it is important to check that carefully. Sometimes, only personally identifiable information is removed, which is not sufficient to avoid re-identification.

An interesting observation that one of the interviewees (I6) made is that companies increasingly try to brand themselves as privacy protectors. As examples, the interviewee, a professor focusing on privacy, cybersecurity, Internet policy and telecommunications law, named Apple and Microsoft. The question is whether the companies actually consider protecting their clients' privacy as important and the interviewee speculates that if they pretend to care there will be less governmental intervention. Additionally, the interviewee pointed out that companies might refrain from putting a lot of emphasis on privacy protection because they fear that they could be losing out on benefits that could come from data analysis.

A research associate focusing on transparent computer systems (I2) stated that technologies related to accountability and transparency as well as data provenance technologies are not yet well integrated in big data solutions. Although there is quite some research, the results have not been integrated yet in mainline products or popular libraries. The interviewee stressed that he is convinced that the technologies are quite mature and could be implemented. Therefore, it is mainly a question of time until some company understands the value of being able to explain where data comes from and what is actually done with the data. The interviewee believes that new regulations as well as trust issues taken up by the press can speed up the process.

Another interviewee, a technology advisor (I1), emphasised that he is a strong believer of privacy by design. The interviewee stated that he thinks that privacy by design can be effective since traditional legal instruments may not be implemented in big data settings because they are blocking; and it will be very difficult to block technology development in the future. Privacy by design, according to the interviewee, accompanies the processing with safeguards and does not prevent development. In general, the interviewee considers the level of integration of privacy-preserving technologies in today's solutions as not comprehensive. He stated that the big players, which of course pose the biggest risks, are also those which are most aware of the risks. As privacy preservation is sometimes in conflict with business

[84] https://privacy-analytics.com/

objectives, they do not integrate everything in their products although they would be able to do so. The interviewee emphasised that the researchers at Google are the best in their class in fields such as pseudonymisation, anonymisation and minimisation but he does not think that they implement all of their capabilities in the products. Nevertheless, Google does much more than others do.

In line with this, another interviewee, an associate professor focusing on technologies to protect privacy (I3), explained that he thinks that integrating privacy-preserving technologies in today's solutions would often be in conflict with the business models of the solutions' users. Therefore, big data analytics companies restrict certain functionalities. They are not interested in slowing down their business. They want solutions that give them the flexibility to do what they want. Another reason that might hamper the integration of the technologies in big data solutions is complexity. Most developers do not have the training to properly integrate the technologies. The interviewee concluded that there are very strong solutions in research but there is a big gap when it comes to deployment, particularly regarding encryption technologies. Encryption, according to the interviewee, is very widespread and advanced. However, companies rather implement them to protect their own datasets than to protect their clients from themselves. The interviewee concluded that technologies for policy enforcement, accountability and transparency are probably not only quite advanced but also, as compared to encryption technologies, preferred by companies to address privacy issues. Additionally, companies seem to prefer organisational over technical measures as they interfere less with their need for flexibility.

Finally, one of the interviewees (I8) highlighted the important role of data lakes in the context of big data. According to the representative of a technology company, having all data that is relevant for an organisation in one place allows making sure that governance is in place, that role-based access is in place, and that data is accessed and used in line with the class of data it belongs to. The data in an organisation's data lake, according to the interviewee, can range from ledger data, client data and financial data to marketing data, historical records and text records.

4.2. Customers and users

This section discusses whether there is a significant interest in big data solutions that include privacy-preserving technologies or in privacy-preserving add-ons. This means that the section deals with the demand side. Again, ethical and societal issues other than privacy are also taken into account.

Data analytics and data sharing are huge businesses. Both private companies and research organisations want to use the existing amount of data, however, privacy must be ensured. A lot of companies therefore turn to data anonymisation techniques. According to Gartner analyst Ramon Krikken, data anonymisation allows companies to make the most out of the existing data, as analyses are only possible when the privacy of individuals is ensured.[85] In her article on data anonymisation techniques, Roy[86] cites Yves-Alexandre de Montjoye, research scientist at MIT Media Lab, who stated that data anonymisation consists of pseudonymisation and de-identification, basically taking sensitive data like mobile phone and medical data and then removing any information that can link it back to an individual. However, within big datasets this principle is at risk as there are numerous data points for each individual that can easily link back to an

[85] ibid.
[86] ibid.

individual when combined. In fact, his research showed that it requires just four pieces of information to identify 90% of the people in a dataset containing credit card transactions of over a million users.

Other studies confirm this. There have been several known cases of re-identification, the majority performed by researchers on real datasets. In the majority of these cases, the data had been poorly anonymised in the first place. This was, for example, provoked if the organisations had retained too many identifying elements in the dataset, or, in connection with the publication of the data, had not analysed which other accessible datasets existed that could be used to deduce information from their own dataset. Furthermore, some types of data are more difficult to render anonymous than others. This applies, among others, to localisation as people's patterns of movement are so unique that the semantic part of the localisation data (the places where the data subject has been at a certain point in time) can reveal a lot about the data subject, even without other known attribute values. This was demonstrated in many representative academic studies.[87] Another example of personal data that can relatively easily be re-identified is genetic data. If the only anonymisation technique used is to remove the subject's identity, the combination of publicly available genetic resources (such as genealogy databases, obituaries, search engine results) and metadata about DNA donors (donation time, age, address) can reveal the initially-removed identity of certain people. This again is because genes are inherently unique.[88]

Re-identification might not only lead to legal consequences but could also affect the companies' image.[89] This argument is underlined by Weathington.[90] In his article on big data privacy and the included risks, he states that companies have to put emphasis on privacy if they want to prevent the embarrassment they would suffer if there is a data breach. Other risks mentioned are the risk of discrimination, inaccurate analysis caused by fake news, and identity reverse engineering by undoing anonymisation.[91]

It also is a company's responsibility to ensure effective mechanisms for access control. This is especially important in non-traditional work environments. Access control gets even more difficult when multiple devices are used. While the need for access control mechanisms is thus based in the modern employee's working habits, it is in the hand of security professionals to ensure access control mechanisms to prevent data breaches.[92]

The prevalence of data breaches along with the growth of mobile and public cloud services have also led to an increasing demand for encryption technologies. Data encryption can severely hinder attackers in their goal to steal confidential user and customer data and trade secrets. In addition, the increasing

[87] Yves-Alexandre de Montjoye et al., "Unique in the Crowd: The privacy bounds of human mobility," *Scientific reports* 3 (2013)
[88] Datatilsynet, "Challenges linked to anonymisation,", https://www.datatilsynet.no/en/regulations-and-tools/guidelines/anonymisation/challenges-linked-to-anonymisation/
[89] Mekhala Roy, "Data anonymization techniques less reliable in era of big data"
[90] John Weathington, "Big data privacy is a bigger issue than you think"
[91] ibid.
[92] James A. Martin, "What is access control? 5 enforcement challenges security professionals need to know"

adoption of new technologies such as mobility, cloud and virtualisation have pushed the need for encryption more than ever before.[93]

Personal data, furthermore, is a huge money-making business. Conn[94] cites Kay Firth-Butterfield, adjunct professor at the University of Texas in Austin that *"data is worth money"* and that individuals should have the right to choose if money should be made from their own data. While Francesca Rossi, research scientist for IBM, in general agrees with this statement, she also emphasises that the more data a system analyses, the more useful the outcomes potentially are. Therefore, it is even more important that companies protect their clients' data, so that trust is gained through a transparent handling of data.[95]

One would expect that the handling of personal data and privacy protection are very important for clients of companies that are highly networked, deal with big data and process their personal data. Personal data can affect everything from relationships to getting a job, and from qualifying for a loan to even getting on a plane.[96] However, it seems that quite some of the clients are blinded by the benefits of providing their personal data, which include, for instance, advertisements focused on what they actually want to buy.

In a 2014 study, a research team looked at consumers' awareness of how their data was collected and used, how they valued different types of data, their feelings about privacy, and what they expected in return for their data.[97] Five countries were analysed, and while the value assigned to personal data varied among individuals, a median by country for each type of data was determined. As Figure 6 shows, the responses revealed significant differences from country to country and from one type of data to another. The differences between countries are discussed in section 4.3.

[93] Mark Hickman, "The encryption challenge,", https://www.itproportal.com/features/the-encryption-challenge/ (accessed April 10, 2018)
[94] Ariel Conn, "Can We Ensure Privacy in the Era of Big Data?,", https://futureoflife.org/2017/02/10/can-ensure-privacy-era-big-data/ (accessed February 23, 2018)
[95] ibid.
[96] Taylor Armerding, "The 5 worst big data privacy risks (and how to guard against them),", https://www.csoonline.com/article/2855641/privacy/the-5-worst-big-data-privacy-risks-and-how-to-guard-against-them.html (accessed February 23, 2018)
[97] Morey, Forbath and Schoop, "Customer Data: Designing for Transparency and Trust"

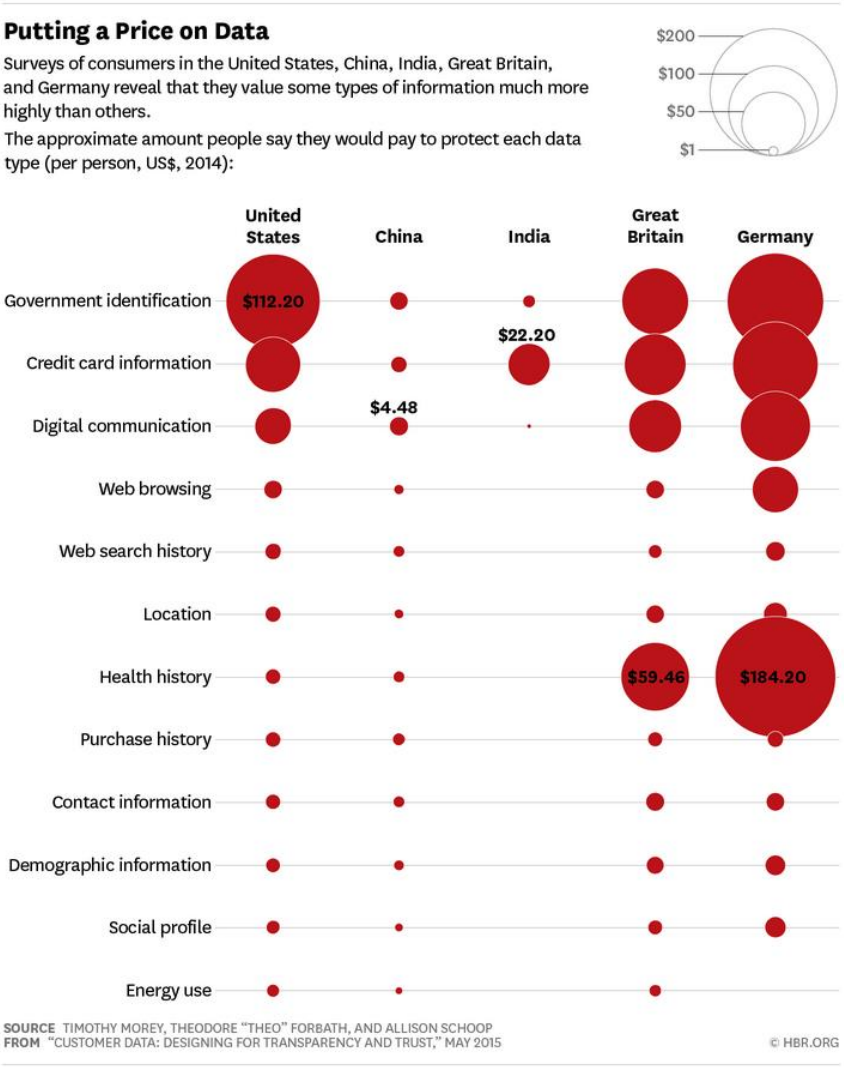

Figure 6 How people value different types of data

With respect to types of data, the analysis looked at three categories:

- *Self-reported data*, or information people volunteer about themselves, such as their e-mail addresses, work and educational history, and age and gender.
- *Digital exhaust*, such as location data and browsing history, which is created when using mobile devices, web services, or other connected technologies.
- *Profiling data*, or personal profiles used to make predictions about individuals' interests and behaviours, which are derived by combining self-reported, digital exhaust and other data.

The study shows that people value self-reported data the least, digital exhaust more, and profiling data the most.[98]

[98] ibid.

One of the interviewees, a technology advisor of a national data protection authority (I1), stated that there is rather low demand from the customer side for technologies to protect privacy. With respect to data, quantity prevails over quality and quick adoption over awareness. A change in culture is necessary, according to the interviewee, which will push the providers to implement technology safeguards. Policy makers could play an important role with respect to the demand. They could set priorities accordingly in education or put emphasis on privacy in public procurement. The related features must be embedded in the products rather than as add-ons. To prevail, privacy preservation has to lead to a win-win situation in the end.

A research associate focusing on transparent computer systems (I2) reported that people should care about their privacy but that he has the feeling that they often only do for a short time when something goes wrong and the press picks it up. He mentioned the recent Cambridge Analytica and Facebook scandal as an example. The interviewee concluded that there is at least some demand from the public. Regulators should step in and make sure that organisations are transparent in what they are doing; organisations cannot do any obviously bad things in that scenario and thus will maintain some standard. The interviewee also argued that people should not have to pay extra for privacy preservation; this should be something that can be taken for granted. Being able to understand what is done with data should be a prerequisite, not something to pay extra. The interviewee stressed that he thinks that what organisations processing personal data should do must go beyond telling individuals what they do; they should enable individuals to read what is done. The right to retract consent would be very artificial if individuals are not able to understand how their data is processed.

Another interviewee (I3) went in a similar direction. The interviewee stated that he has the feeling that people are worried but at the same time do not know what to do. Technologies and concepts are often complex and counter-intuitive. Moreover, people are not used to the adversarial thinking required to understand threats. The interviewee revealed that he thinks that it is also a question of age. Younger people might have less concerns, according to the interviewee, than people that are in their thirties or older. Something that shows that there is a demand is that companies begin using privacy protection as a selling point. Apple was mentioned as an example. Companies are distinguishing themselves from others by emphasising that they do not monetise user data. The interviewee, who is an associate professor focusing on technologies to protect privacy, mentioned two factors that are driving recent advances in the deployment of technologies: legislation, in particular the GDPR, and scandals taken up by the media. Both make companies offer something to their clients to make them feel more confident.

Finally, a professor focusing on privacy (I5) stated that demand is closely related to maturity. If the technologies are mature, according to the interviewee, it is the regulatory context that can push the use of the technologies. The interviewee emphasised that the Cambridge Analytica case and other privacy breaches might lead to regulation in this regard, especially in Europe. He thinks and hopes that there will be significant developments regarding privacy-preserving technologies over the next years as they are what companies need to implement privacy-by-design.

4.3. Regional differences

This section discusses the role that regional and cultural differences play with respect to the availability and use of privacy-preserving technologies. The focus lies on differences between Europe and North America.

Various studies suggest a difference in the perception of ethical and societal issues between European countries and North America. The survey of Morey et al.[99] mentioned in section 4.2 shows some countries value some types of information more highly than others. While Germans place the most value on their personal data, the Chinese and Indians value personal data the least, with UK and US respondents falling in the middle. Government identification, health and credit card information tended to be the most highly valued across countries, and location and demographic information among the least. For instance, while people in the US, the UK and Germany all value government identification, credit card information and digital communication the most, US citizens value these types of information significantly less than Germans (with UK citizens in between). Moreover, while information on their personal health history is of great value to Germans (in fact the study revealed that this type of information is the one of greatest importance there) and still important to UK citizens, it seems to be of no importance to US citizens. The same applies to the web browsing history, although to a much lesser extent.[100] The cultures of India and China, for example, are considered more hierarchical and collectivist, while Germany, the US, and the UK are more individualistic, which may account for their citizens' stronger feelings about personal information.

In the **EU**, as of 25 May 2018, all processing of personal data will have to be made in observance of the legal provisions of the EU GDPR. The addressees of this EU regulation are required to meet certain goals and objectives at their own discretion and adopting their own business decisions to keep them in compliance with the following principles:

- lawfulness, fairness and transparency
- purpose limitation
- data minimisation
- accuracy
- storage limitation
- integrity and confidentiality

Therefore, all big data projects have been or should be assessed bearing all these legal principles in mind.[101] Under the GDPR, organisations can face sanctions of up to 4% of their annual gross global revenue in the event of a breach or data mismanagement.[102]

[99] ibid.
[100] ibid.
[101] Rafael García del Poyo, Roger Segarra and Martínez Samuel, "Big Data analysis and anonymisation techniques under the EU General Data Protection Regulation,", https://www.financierworldwide.com/big-data-analysis-and-anonymisation-techniques-under-the-eu-general-data-protection-regulation/ (accessed May 9, 2018)
[102] Mekhala Roy, "Data anonymization techniques less reliable in era of big data"

The GDPR applies when a data controller or a data processor is based in the **EU**, but it also addresses the transfer of data outside the EU if the data collected or processed is from individuals based in the EU. Personal data in this regard refers to any information connected to an individual, but there are exceptions for data processed in an employment context or in national security that might be subject to country regulations. To be compliant with the GDPR, data controllers need to design data protection measures into their business processes for products and services.[103]

In the **US**, there is no single comprehensive federal law on the collection and use of personal data. Instead, there are various federal and state laws, regulations and guidelines, which with each congressional term are proposed to be standardised at a federal level.[104] Whereas the right to privacy and the right to data protection in Europe are fundamental rights of all citizens, the right to data protection and the right to privacy in the US are constitutional rights. This is an important distinction as it requires different forms of enforcement, but privacy by design is uniquely beneficial for both legal systems. In February 2012, the Obama administration published a blueprint for what it termed a Consumer Privacy Bill of Rights (CPBR), which became proposed legislation but was immediately attacked both by industry groups that stated that it would impose *"burdensome"* regulations and by privacy advocates who complained that it was riddled with loopholes. The blueprint thus never made it to a vote. Beyond that, the US government has not been able to agree on other privacy initiatives. In 2016, the Federal Communications Commission (FCC) issued the so-called broadband privacy rules right before the election. These rules would have limited data collection by Internet service providers but were repealed by Congress before they took effect.[105]

A common approach by the US Department of Commerce, the European Commission and the Swiss administration are the so-called Privacy Shield Frameworks. The frameworks are aimed to provide US, EU and Swiss companies with a mechanism to comply with data protection requirements when transferring personal data from the EU and Switzerland to the US in support of transatlantic commerce. The program is administered by the International Trade Administration (ITA) within the US Department of Commerce and enables US-based organisations to join one or both of the Privacy Shield Frameworks in order to benefit from the adequacy determinations. To join either Privacy Shield Framework, a US-based organisation needs to publicly commit to comply with the framework's requirements. While joining the Privacy Shield is voluntary, once an organisation makes the public commitment to comply with the framework's requirements, the commitment will become enforceable under US law.[106]

As already mentioned, there are several companies that develop and sell products with the aim of enhancing privacy protection, managing personal data and/or becoming compliant relevant laws (see section 4.1). Most of these companies are located in the US.

A professor who was interviewed (I4) stated that Europe is very strong in legal and governance solutions, while most of the technological innovations do not come from a European context. According to the interviewee, in North America, it is more of a framework not to constrain things and see what happens

[103] https://www.eugdpr.org/
[104] https://uk.practicallaw.thomsonreuters.com/6-502-0467
[105] Taylor Armerding, "The 5 worst big data privacy risks (and how to guard against them)"
[106] https://www.privacyshield.gov/

and then generalise and apply case-based legal decisions. The European historical context is more rule-driven. The interviewee, who works at a North American university, pointed out that he indeed thinks that Europe has to catch up with respect to technological solutions. The culture, according to the interviewee, is ripe for responsible data science solutions.

A representative of a company headquartered in North America (I9) emphasised that the company he works for has been consulting with its clients in the EU for many months already on how to get prepared for the GDPR. The interviewee stressed that there is definitely some regional variation in terms of how people think about privacy, but there is also variation within different domains. The company, for instance, is preparing itself for data protection impact assessments. The focus, according to the interviewee, is clearly not only on technical but also on organisational measures. Another interviewee (I1) explained that North American companies are certainly interested in European markets. The companies try to strike a balance between investments and safeguards. For instance, Google had to change its privacy policy and their mechanisms for getting consent from its users. Google had discovered that there was no way out.

Another professor who was interviewed (I6) also commented on the possible impact of the GDPR. According to the interviewee from the Middle East, it is possible that the EU becomes an exporter of norms that have the potential to lead to technological changes globally. However, it is also possible that the EU is deprived of leading technologies. The question with respect to such a scenario is if EU citizens would want to stay with the technologies they are getting or if they would opt for the technologies other people in the world have. Yet another professor (I5) pointed out that he is convinced that European legislation will have a global impact. It will also change, to some extent, the technologies developed and used in North America. Canadian policy makers, according to the interviewee working in North America, for instance, have even suggested to consider making national privacy regulations follow the ideas of the GDPR.

An interviewee who lives in Europe and North America (I2) stated that there clearly is a big divide on how people see things. According to him, for instance, North Americans do not seem to trust their government much. In Europe, governments are seen as important protectors of privacy. Furthermore, North Americans do not seem to have a problem with a successful private company making decisions about their lives. In Europe, people want the legal framework to decide what can be done with data and what cannot. Similarly, another interviewee (I3) stated that he has the impression that the European law and the European approach to privacy and personal data is not really in line with Silicon Valley and their view on data. North America in general, and the Silicon Valley in particular place a premium on the utility of data. The things that can be learned from data, according to the interviewee, weight significantly more than possible concerns about privacy. In Europe, the balance is different. With respect to the GDPR, the interviewee stressed that it seems that European legislation is going to have global impact, particularly because it is too difficult to make products that are differentiated by region. The interviewee referred to Apple and Facebook as examples. Apple, according to the interviewee, stated that it is modifying its products to comply with the GDPR, and the modification will be worldwide for everyone. Facebook, in contrast, stated that they might implement extra protections for Europeans to comply with the GDPR,

which will not be rolled out to people in other jurisdictions. Lomas[107], for instance, provides further details about Facebooks GDPR changes.

4.4. Organisational measures

This section discusses to what extent privacy-preserving technologies need to be complemented by non-technical (mostly organisational) measures. This means it looks at the relevance of complimentary organisational measures. Ethical and societal issues other than privacy are also taken into account.

As already mentioned, privacy is at risk in the era of big data. A MIT study shows that it requires just four pieces of information to identify 90% of the people in a dataset containing credit-card transactions of over a million users.[108] As most solutions work well only with a critical number of individuals participating, and with customer data being a growing source of competitive advantage, gaining consumers' confidence will be key. Numerous studies have found that transparency about the use and protection of consumers' data reinforces trust. Organisations that are transparent about the information they gather, give clients control over their personal data, and offer fair value in return for it will be trusted and earn ongoing and even expanded access. Those that conceal how they use personal data and fail to provide value for it stand to lose clients' goodwill and eventually their business. Companies may earn access to client data by offering value in return, but trust is an essential facilitator, research shows.[109] The more trusted a brand is, the more willing clients are to share their data. An organisation that is considered untrustworthy will find it difficult or impossible to collect certain types of data, regardless of the value offered in exchange. Highly trusted organisations may be able to collect it simply by asking, because clients are satisfied with past benefits received and confident the organisation will guard their data. This means that if two organisations offer the same value in exchange for certain data, the one with the higher trust will find clients more willing to share their data. Creating trust through the use of specific technology alone is hardly possible; it is at least necessary to communicate the use of the technology properly.

Chris Combemale, CEO of the industry body DMA Group, states that *"[a]s an industry, we need to reconnect with consumers on the issue of data and build new relationships based on transparency and trust"*.[110] Several suggestions to ensure transparency and win back the customer's trust focus on non-technical measures. According to the DMA Group[111], it is considered important to

- be clear about how data will be used (explain why user data is collected, what is done with the data, whether the data is sold or shared, and whether data is deleted when asked to do so),
- explain the benefits of sharing data (even if implementing a top down approach by placing focus on customer engagement),
- demonstrate explicitly how data is kept secure (defining security procedures, ensuring that there are comprehensive regulatory programmes in place and that legal requirements are respected),

[107] Natasha Lomas, "Data experts on Facebook's GDPR changes: Expect lawsuits,", https://techcrunch.com/2018/04/18/data-experts-on-facebooks-gdpr-changes-expect-lawsuits/ (accessed May 8, 2018)
[108] Mekhala Roy, "Data anonymization techniques less reliable in era of big data"
[109] Morey, Forbath and Schoop, "Customer Data: Designing for Transparency and Trust"
[110] https://dma.org.uk/press-release/dma-challenges-brands-to-re-connect-with-consumers-on-data
[111] https://dma.org.uk/press-release/dma-challenges-brands-to-re-connect-with-consumers-on-data

- give customers control over their data (especially social media companies can ensure customers' control over their data; giving control over data can be supported by technologies), and
- simplify the collection procedure (making terms and conditions easy to understand for consumers).

D'hulst and Kengen point out that data protection compliance as enforced by the GDPR does not only bring obligations and risks, but also various opportunities:[112]

- It can be a useful tool in increasing employee and consumer confidence in an organisation.
- In addition to providing an enhanced brand image, data protection compliance can also help in the management of organisational information.
- Data protection compliance acts as a reminder to companies that they should also act to protect company data and business secrets generally.
- Strong awareness of personal data can facilitate future projects using such data.
- For many organisations, personal data (including contacts and profiles) is a key asset. Documented compliance adds to the value of this information and the organisation.

Furthermore, the authors stress that full compliance needs long-term organisational and operational investments. The organisation's leadership must therefore allocate resources (time, personnel and budget) and ensure that all staff place importance on their tasks in the compliance process. A clear and explicit commitment and support of the company's leadership is deemed crucial in this regard as well as an awareness that the rules on data protection will change. The authors suggest mapping the areas within the organisation that are likely to be affected by the GDPR and set up a plan towards compliance that addresses both the organisational structure and the operational compliance strategy while taking into account the activities, structure and size of the organisation. Under the GDPR, some organisations will even be obliged to appoint a Data Protection Officer (DPO); an independent position within an organisation that is assisted by a network of contact persons in each department to effectively ensure the implementation of compliance measures and provide information about data protection activities when required. As a last organisational measure, the staff that gets into contact with personal data needs to be aware of the basic obligations under data protection laws and how to translate these to daily tasks.[113]

A professor focusing on machine learning, data and text mining, and privacy (I4) stated that a combination of technical and organisational measures is essential. So far, according to the interviewee, the focus was more on governance and legal solutions, but he thinks that the Facebook situation and others show very vividly the limitations of reactive approaches that usually prescribe norms and actions that are taken if there is a breach of privacy or some types of rules are broken. This, the interviewee expects, will create more interest in technical solutions that are proactive in the sense that they prevent breaches or rule violations in the first place. The interviewee pointed out that apart from a growing importance of

[112] Thibaut D'hulst and Lily Kengen, "Data protection compliance strategy,", Practical Law (2017), https://uk.practicallaw.thomsonreuters.com/Link/Document/Blob/I6509c58bcfce11e79bef99c0ee06c731.pdf?targetType=PLC-multimedia&originationContext=document&transitionType=DocumentImage&uniqueId=fe172116-b02e-451a-a2f9-cf91f68d27aa&contextData=%28sc.Default%29&comp=pluk (accessed April 10, 2018)
[113] ibid.

technologies he also sees a particular need for education targeting, for instance, young people and users of social networks. They need to know that things in fact are never erased.

Similarly, a technology advisor (I1) stressed that both are important for privacy preservation technology and commitment. He further explained that technologies are not the key challenge. In order to make them effective, it is not sufficient if just a single person in the organisation has the required expertise, the entire environment must be aware of the technologies and the related opportunities and threats. Another issue the interviewees mentioned is usability. Most of the technologies can only be used by experts. Awareness and education are thus key aspects according to the interviewee.

Another interviewee (I8) stressed that proper processes and governance are needed to ensure privacy. The representative of a technology company pointed out that having access to data does not mean that it can be used for every purpose. It was emphasised that on example is how marketing data from social media is used. However, not all data shared on social media is shared with the intention to be profiled. Technologies, particularly encryption and access control, are necessary in this context but they are not sufficient. Organisations also need to have the right processes in place, for instance, to check the legal compliance of own products and services before they are rolled out, and to deal with aspects such as consent of individuals. According to the interviewee, legislation such as the new GDPR is a key safeguard concerning privacy. Similarly, another interviewee (I6), stated that privacy preservation is not only an issue of technology but also an issue of, for instance, building processes and specific administrations. As an example, the interviewee, a professor focusing on privacy, cybersecurity, Internet policy and telecommunications law, mentioned that if medical data is about to be shared, this has to be approved by a committee that consists not only of lawyers and doctors but also of statisticians and computer scientists. Another interviewee (I2) also sees things similarly. The interviewee stated that a regulatory framework is essential to complement technologies. Within organisations, incentives are considered to play a key role.

One of the interviewees (I3) stated with respect to anonymisation that it is not possible to fully automate such a feature. The problem is not the technology but rather that two contradictory things are expected from the same dataset. Getting a little bit more anonymity and to make the data a bit more difficult to re-identify requires processing the data in a way that it loses utility. It is a trade-off and the technology cannot identify the sweet spot all by itself. In general, the interviewee, an associate professor, stated that technical solutions alone are not sufficient. At some point, organisational measures will always be necessary to make sure the technology functions as expected. Legislation, for instance, plays a key role according to the interviewee. Moreover, agreements and policies may be relevant. Another interviewee (I5) also pointed out that he does not think that technology can solve all problems related to privacy that are faced in big data contexts. Technical measures must be complemented by organisational measures and regulations. Regulations need to backup technologies and people in organisations need to be aware of possible privacy issues and means to address them. Moreover, data protection officials are needed that are aware and able to assess the privacy impact or risk of personal data that is collected and used by companies. The interviewee stressed that the risk depends a lot on the data that is used.

4.5. Responsibility

This section discusses who along the data value chain[114] is or should be responsible for addressing ethical and societal issues. Who in the data value chain needs to take action?[115] Although the focus is once again on technologies, non-technical measures are also addressed.

Big data analytics, according to Weathington[116], has the power to provide insights about people that are far and above what they know about themselves. With respect to the responsibility for addressing societal and ethical issues related to big data, he cites Stan Lee, an American writer, editor, and memoirist, who argues that *"with great power there must also come – great responsibility".* Armerding[117] cites Rebecca Herold, CEO of The Privacy Professor, with her statement that *"consumers need to protect themselves because nobody else will be doing it for them".*

Weathington emphasises the responsibility of the data holders to fully inform their subjects about the use of the data.[118] While he is careful to clarify not to offer too much information or even give away strategic secrets, this somehow passes on at least part of the responsibility to the consumer and thus underlines that consumers need to protect themselves.

Taking this concern a step further, Conn cites Roman Yampolskiy, an associate professor at the University of Louisville, arguing that *"the world's dictatorships are looking forward to opportunities to target their citizenry with extreme levels of precision"*. In his opinion, it is an essential part of the technology development to ensure that privacy becomes a fundamental cornerstone of any data analysis.[119]

In general, organisations collecting, using and distributing data are responsible for data management and anonymisation, unless somehow the liability is passed onto the anonymisation provider. But even if the latter is the case, the other organisation would be the one to face public relations exposure in case of a data breach, which can be more costly than the costs of a breach.[120]

In a future in which customer data is a growing source of competitive advantage, gaining clients' confidence will be key. Organisations that are transparent about the information they gather, give clients control over their personal data, and offer fair value in return for it will be trusted and will earn ongoing and even expanded access. Those that conceal how they use personal data and fail to provide value for it stand to lose clients' goodwill and eventually their business.[121]

---

[114] Among the ones that come into question are data suppliers, technology providers, data end users, data marketplaces, data subjects as well as regulators.
[115] Anton Vedder and Bart Custers, "Whose Responsibility Is It Anyway? Dealing with the Consequences of New Technologies," in *Evaluating new technologies: Methodological problems for the ethical assessment of technology developments*, vol. 3, ed. Anthony M. Cutter, Marcus Düwell and Paul Sollie, 1st ed., 21–34, International library of philosophy and scientific method 3 (New York: Springer, 2009)
[116] John Weathington, "Big data privacy is a bigger issue than you think"
[117] Taylor Armerding, "The 5 worst big data privacy risks (and how to guard against them)"
[118] John Weathington, "Big data privacy is a bigger issue than you think"
[119] Ariel Conn, "Can We Ensure Privacy in the Era of Big Data?"
[120] Mekhala Roy, "Data anonymization techniques less reliable in era of big data"
[121] ibid.

An important prerequisite for clients' trust is secure anonymisation. The data controller has various options during the process of anonymisation. At first, a decision must be made whether the personal data to be processed should be anonymised, pseudonymised or left identifiable. This choice will affect the organisation's responsibilities in relation to the data protection law during data processing. If anonymisation is chosen, any further processing will fall outside the scope of the data protection law.[122]

Another prerequisite is encryption. Organisations need to implement automatic encryption in their security policies, as encrypting at the source may not stop a hacker from gaining access to data, but it will prevent the data itself from being disclosed. According to Hickmann, many organisations think they are well protected from attack, but every organisation can easily suffer a data breach not only because of hacktivists or cyber criminals, but also if an employee accidentally miss-sends an email or loses his or her device. He emphasises that it is crucial for every organisation to take this attitude to ensure effective protection of sensitive data.[123]

It is important that data protection is not considered as *"somebody else's problem"* as this point of view passes the responsibility from one hand to another. In an organisation, this problem occurs when everyone thinks somebody else is possibly responsible for data protection; for instance, the legal department, the IT department, human resources, the employees themselves.

According to Murphy[124], the legal team does not hold the responsibility for data protection all alone. They make sure to have a legally sound basis to collect information, ensure that contracts are sound, make sure data processing to clients is explained well and requirements for disclosure on websites, in communications and on organisational property are met. All of these are important requirements, but they are not sufficient. For instance, it is not the legal department's responsibility to create content, marketing lists and multi-channel campaigns or assess a new CRM system; all of which are sensitive to misuse of personal data. The same applies to the IT department: While they surely are responsible for data security, meaning they implement appropriate security measures depending on the volume and sensitivity of the data that is stored, assess the potential threats and determine the appropriate security measures against these threats, investigate the available solutions and balance costs and risks as well as monitor the organisation's systems for suspicious activity, they are not responsible for the data's origin nor for client contacts. The conclusion from that is that everyone in an organization who handles personal data is responsible for data protection and that it is the responsibility of an organisation's management to ensure that every single employee knows about data protection policies and concrete strategies to ensure it in everyday business.

One of the interviewees (I9) went into detail with respect to the question of responsibility within the organisation. The company that the interviewee works for provides a software platform for data analytics and offers professional services to help its clients get use out of the platform. The company observed that parts of its client organisations are not sufficiently operationally involved. The interviewee reported that

[122] Datatilsynet, "Guide: The anonymisation of personal data", https://www.datatilsynet.no/en/regulations-and-tools/guidelines/anonymisation/?id=7636 (accessed May 22, 2018)
[123] Mark Hickman, "The encryption challenge"
[124] Marie Murphy, "How data protection is everyone's responsibility.", http://safedatamatters.com/everyones-responsibility/ (accessed April 10, 2018)

specifically people who are in data governance roles that sometimes are called data officers or data custodians often have responsibilities around maintaining data quality, and provisioning and revoking access to users. However, they are neither trained to nor feel empowered to perform anything like an audit of a machine learning algorithm or to make judgements about the proportionality of certain data for a given use case. Data protection officers, which are, according to the interviewee, sometimes mentioned in this regard, are traditionally not accustomed to using data or technology themselves. The interviewee thus suggests to get people with privacy and data protection responsibilities more involved with how the data is actually being used, and have them use the same technologies that the people are using who are actually processing or using the data. In his opinion, this could dramatically improve the overall privacy outcomes.

With respect to the responsibility along the value chain, the interviewee underlined that, as a software and services company, the company the interviewee works for wants to make sure that its clients can enforce their policies. For this purpose, the company puts effort into making its controls as granular (since rules are getting more and more complex), as flexible (since its clients are in very different domains) and as consistent (since its products are used at different layers of the technology stack) as possible. If a client wants to integrate data into the platform, it is forced to make decisions about who's going to have access and at what level, and how that is organised, but the company does not force its clients to adopt any single scheme about how to organise their roles and users.

An interviewee, a technology advisor, (I1) argued that the strongest party should have the biggest responsibilities. Consequently, the controller should play a very important role in this regard. Users have a great role in raising the awareness and exercising their rights properly because that results in pressure on controllers, which in turn gives them more safeguards. Moreover, the interviewee explains, the supervisory authority and governments should have a role because they have to shape the framework conditions. A research associate focusing on transparent computer systems (I2) explained that he does not see the responsibility for addressing the challenges with the users. According to him, it should be the responsibility of the provider of the devices and the processor of the data to make sure user expectations are respected and legal requirements met. A third party that the interviewee considers relevant with respect to responsibility are those organisations that provide services to the provider of the device or the processor of the data. If their infrastructure allows compromising the data, they should be responsible.

As discussed in section 3.8, one interviewee (I3) is afraid that responsibility is to some extent pushed to the users who may not fully understand what they are doing. According to the interviewee, the responsivity placed on the user should be as small as possible. Everybody involved should be responsible in line with their competency. In contrast, a professor who participated in the series of interviews (I4) stated that the data provider must provide the tools for the individual data owners to control what happens with their data. The research community, according to the interviewee, must develop these tools and they must be available to data providers cost-free or at a minimal cost.

Finally, an interviewee (I5) stated that in terms of responsibility, everyone should be working together. For the data subject, it would be good to have a way to sanitise or anonymise the data before it leaves his or her sphere of control. The data controller, the party that collects the data and uses big data solutions, should have a lot of responsibility. The creators of the solutions should, according to the interviewee, be aware of the privacy risk and integrate some of the discussed technologies. Similarly to others, the

interviewee pointed out that he does not think that much responsibility can be put on the data subject. He likes the idea of giving more control to the data subject but in the context of big data that seems very difficult as the data leaves the data subject's sphere of control.

5. Conclusion

Within the scope of this report, privacy-preserving technologies were assessed taking societal and ethical issues into account. The assessment consisted of two parts: a technology-specific assessment of selected classes of technologies and a more general assessment of the technologies. Among the assessed technologies were technologies for anonymisation and sanitisation, encryption, MPC, access control, policy enforcement, accountability and transparency, data provenance, and access, portability and user control. With respect to the issues, the focus was clearly on privacy but issues and values such as self-determination, welfare interdependency, trustworthiness, accountability, fairness and legislation were also taken into account. The assessment was based on a series of interviews and desk research.

The assessment led to the following overarching observations:

- The set of classes of technologies is **quite comprehensive**. Additional technologies were suggested but most of them were at least related to the existing classes. Blockchains, quantum computing, data classification, homomorphic encryption and sketches were mentioned explicitly.
- The classes of technologies contribute to privacy preservation in different ways and are closely linked with each other. In practice, the technologies **need to be combined** to be effective and there is no single most important class of technologies. There is no hierarchy between the technologies.
- The technologies **pursue different aims**. While some aim at overcoming the need for trust in other parties (e.g., MPC, homomorphic encryption), others aim at increasing trust in other parties (e.g., access control, policy enforcement, accountability, transparency).
- A **multidimensional measure** for privacy preservation is needed that covers relevant factors in a balanced way. Apart from the degree of protection and the cost of protection, it is important to also take the societal value of data into account.
- Some observe a **fundamental tension** between the objective of big data, which is collecting a lot of personal data and using it, and privacy, which is about protecting personal data. The two goals are almost antagonistic. Yet, classes of technologies constitute attempts to bridge gaps between goals as well as differences that are, among others, regional, legal, organisational or procedural.

Moreover, it was found that there is wide agreement that technologies are **integrated very little** in today's big data solutions. There are strong solutions in research but there is a big gap when it comes to deployment. There is broad consensus with respect to the **rather low demand** from the customer side for technologies to protect privacy. It seems that quite some of the data subjects are blinded by the benefits of providing their personal data. There is little doubt that there are **considerable regional and cultural differences**. North Americans do not seem to trust their government much. In Europe, governments are seen as important protectors of privacy. Perhaps this relates in part to the fact that privacy is protected as a constitutional right in the US; therefore, US governments have a rather indirect role in mingling into privacy protection. Whereas in Europe privacy rights are protected as fundamental rights and governments have a rather direct role in enforcing the protection of such rights and remedying any form of violation. There is consensus that the **combination of technical and organisational measures** is essential. Awareness and education are key aspects in modern organisations just as processes, legislation, agreements and policies. It is often stressed that consumers need to protect themselves because nobody

else will do it for them. In general, organisations collecting, using and distributing data are responsible for data management and anonymisation. There is wide agreement that **the strongest party** in the data value chain should have the biggest responsibilities. Although this should in principle be an acceptable starting point, yet in the era of big data agreeing upon the influence of stakeholders upfront might be problematic. Notably, the strength and influence of a certain stakeholder cannot always be predetermined because, due to the dynamics of the field, the strength of a certain stakeholder can become revealed entirely only after the implications of big data analytics are turned into economic or other form of value (see, for instance, the development of Facebook into a multibillion dollar company).

Bibliography

"2016 Global Encryption Trends Study.". https://www.ciosummits.com/Ponemon_Global_Encryption_Trends_Report_2016.pdf (accessed April 24, 2018).

Abouelmehdi, Karim, Abderrahim Beni-Hssane, Hayat Khaloufi, and Mostafa Saadi. "Big data emerging issues: Hadoop security and privacy." In *Proceedings of the 2016 5th International Conference on Multimedia Computing and Systems (ICMCS)*, 731–6. IEEE, 2016.

Acquisti, Alessandro. "Identity Management, Privacy, and Price Discrimination." *IEEE Security & Privacy* 6, no. 2 (2008): 46–50.

Acquisti, Alessandro, and Heinz College. "The Economics of Personal Data and the Economics of Privacy: 30 Years after the OECD Privacy Guidelines.". Background Paper 3. https://www.oecd.org/sti/ieconomy/46968784.pdf (accessed April 23, 2018).

Al Mamun, Abdullah, Khaled Salah, Somaya Al-maadeed, and Tarek R. Sheltami. "BigCrypt for big data encryption." In *Proceedings of the 4th International Conference on Software Defined Systems*, 93–9. Valencia, Spain: IEEE, 2017.

AlMahmoud, Abdelrahman, Ernesto Damiani, Hadi Otrok, and Yousof Al-Hammadi. "Spamdoop: A privacy-preserving Big Data platform for collaborative spam detection." *IEEE Transactions on Big Data* (2017): 1.

Armerding, Taylor. "The 5 worst big data privacy risks (and how to guard against them).". https://www.csoonline.com/article/2855641/privacy/the-5-worst-big-data-privacy-risks-and-how-to-guard-against-them.html (accessed February 23, 2018).

"Big data, artificial intelligence, machine learning and data protection.". https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf (accessed March 28, 2018).

Bogetoft, Peter, Dan L. Christensen, Ivan Damgård, Martin Geisler, Thomas Jakobsen, Mikkel Krøigaard, and Janus D. Nielsen et al. "Secure Multiparty Computation Goes Live." In *Financial Cryptography and Data Security: 13th International Conference, FC 2009, Accra Beach, Barbados, February 23-26, 2009. Revised Selected Papers.* vol. 5628. Edited by Roger Dingledine and Philippe Golle, 325–43. Lecture Notes in Computer Science 5628. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009.

Bomhof, Freek. "In Order to Trust Big Data, Transparency Is Not Enough.". https://datafloq.com/read/transparency-in-big-data-is-not-enough/138 (accessed December 21, 2016).

Boujemaa, Nozha. "Algorithmic Systems Transparency and Accountability in the Big Data Era.". https://de.slideshare.net/NozhaBoujemaa/nboujemaa-datadrivenparis (accessed March 17, 2018).

Burton, Cédric, and Sára Hoffman. "Personal Data, Anonymization, and Pseudonymization in the EU.". https://www.wsgrdataadvisor.com/2015/09/personal-data-anonymization-and-pseudonymization-in-the-eu/ (accessed March 27, 2018).

Canbay, Yavuz, and Seref Sagiroglu. "Big data anonymization with Spark." In *Proceedings of the 2017 International Conference on Computer Science and Engineering (UBMK)*, 833–8. IEEE, 2017.

Cannataci, Joseph A., Bo Zhao, Gemma Torres Vives, Shara Monteleone, Jeanne Mifsud Bonnici, and Evgeni Moyakine. "Privacy, free expression and transparency: Redefining their new boundaries in the digital age.". http://unesdoc.unesco.org/images/0024/002466/246610E.pdf (accessed March 27, 2018).

Chakravorty, Antorweep, Tomasz Wlodarczyk, and Chunming Rong. “Privacy Preserving Data Analytics for Smart Homes.” In *Proceedings of the 2013 IEEE Security and Privacy Workshops*, 23–7. Piscataway, NJ: IEEE, 2013.

Conn, Ariel. “Can We Ensure Privacy in the Era of Big Data?”. https://futureoflife.org/2017/02/10/can-ensure-privacy-era-big-data/ (accessed February 23, 2018).

Cuzzocrea, Alfredo. “Big Data Provenance: State-Of-The-Art Analysis and Emerging Research Challenges.”. http://ceur-ws.org/Vol-1558/paper37.pdf (accessed December 14, 2017).

D'Acquisto, Giuseppe, Josep Domingo-Ferrer, Panayiotis Kikiras, Vicenç Torra, Yves-Alexandre de Montjoye, and Athena Bourka. “Privacy by design in big data: An overview of privacy enhancing technologies in the era of big data analytics.”. https://www.enisa.europa.eu/publications/big-data-protection/at_download/fullReport (accessed September 26, 2017).

Danezis, George, Josep Domingo-Ferrer, Marit Hansen, Jaap-Henk Hoepman, Daniel Le Métayer, Rodica Tirtea, and Stefan Schiffner. “Privacy and Data Protection by Design - from policy to engineering.”. https://www.enisa.europa.eu/publications/privacy-and-data-protection-by-design/at_download/fullReport (accessed December 14, 2017).

Datatilsynet. “Challenges linked to anonymisation.”. https://www.datatilsynet.no/en/regulations-and-tools/guidelines/anonymisation/challenges-linked-to-anonymisation/.

Datatilsynet. “Guide: The anonymisation of personal data”. https://www.datatilsynet.no/en/regulations-and-tools/guidelines/anonymisation/?id=7636 (accessed May 22, 2018).

Davey, Neil. “Customer data collection: How to be trustworthy and transparent.”. https://www.mycustomer.com/marketing/data/customer-data-collection-how-to-be-trustworthy-and-transparent (accessed March 17, 2018).

Defend, Benessa, and Klaus Kursawe. “Implementation of privacy-friendly aggregation for the smart grid.” In *Proceedings of the 1st ACM Workshop on Smart Energy Grid Security*, 65–74. ACM, 2013.

D'hulst, Thibaut, and Lily Kengen. “Data protection compliance strategy.”. Practical Law. https://uk.practicallaw.thomsonreuters.com/Link/Document/Blob/I6509c58bcfce11e79bef99c0ee06c731.pdf?targetType=PLC-multimedia&originationContext=document&transitionType=DocumentImage&uniqueId=fe172116-b02e-451a-a2f9-cf91f68d27aa&contextData=%28sc.Default%29&comp=pluk (accessed April 10, 2018).

Diakopoulos, Nicholas, and Sorelle Friedler. “How to hold algorithms accountable.” *MIT Technology Review* (2016).

García del Poyo, Rafael, Roger Segarra, and Martínez Samuel. “Big Data analysis and anonymisation techniques under the EU General Data Protection Regulation.”. https://www.financierworldwide.com/big-data-analysis-and-anonymisation-techniques-under-the-eu-general-data-protection-regulation/ (accessed May 9, 2018).

Glavic, Boris. “Big Data Provenance: Challenges and Implications for Benchmarking.” In *Proceedings of the 2nd Workshop on Big Data Benchmarking*, 72–80. 2012.

Havron, Samuel. “Poster: Secure Multi-Party Computation as a Tool for Privacy-Preserving Data Analysis.” In *Proceedings of the 37th IEEE Symposium on Security and Privacy.* IEEE, 2016.

Hickman, Mark. “The encryption challenge.”. https://www.itproportal.com/features/the-encryption-challenge/ (accessed April 10, 2018).

Inukollu, Venkata N., Sailaja Arsi, and Srinivasa Rao Ravuri. “Security Issues Associated with Big Data in Cloud Computing.” *International Journal of Network Security & Its Applications* 6, no. 3 (2014): 45–56.

Jain, Priyank, Manasi Gyanchandani, and Nilay Khare. “Big data privacy: A technological perspective and review.” *Journal of Big Data* 3, no. 1 (2016): 120.

Kagal, Lalana, and Hal Abelson. “Access control is an inadequate framework for privacy protection.” In *Proceedings of the W3C Workshop on Privacy for Advanced Web APIs.* 2010.

Karvelas, Nikolaos P., Andreas Peter, Stefan Katzenbeisser, and Sebastian Biedermann. “Efficient Privacy-Preserving Big Data Processing through Proxy-Assisted ORAM.”. https://eprint.iacr.org/2014/072.pdf.

Lomas, Natasha. “Data experts on Facebook’s GDPR changes: Expect lawsuits.”. https://techcrunch.com/2018/04/18/data-experts-on-facebooks-gdpr-changes-expect-lawsuits/ (accessed May 8, 2018).

Martin, James A. “What is access control? 5 enforcement challenges security professionals need to know.”. https://www.csoonline.com/article/3251714/authentication/what-is-access-control-5-enforcement-challenges-security-professionals-need-to-know.html (accessed April 10, 2018).

“Meeting the challenges of big data: A call for transparency, user control, data protection by design and accountability.”. Opinion 7/2015. https://edps.europa.eu/sites/edp/files/publication/15-11-19_big_data_en.pdf (accessed March 28, 2018).

Montjoye, Yves-Alexandre de, César A. Hidalgo, Michel Verleysen, and Vincent D. Blondel. “Unique in the Crowd: The privacy bounds of human mobility.” *Scientific reports* 3 (2013): 1376.

Morey, Timothy, Theodore Forbath, and Allison Schoop. “Customer Data: Designing for Transparency and Trust.” *Harvard Business Review* 93, no. 5 (2015): 96–105.

Murphy, Marie. “How data protection is everyone’s responsibility.”. http://safedatamatters.com/everyones-responsibility/ (accessed April 10, 2018).

Naydenov, Rossen, Dimitra Liveri, Lionel Dupre, Eftychia Chalvatzi, and Christina Skouloudi. “Big Data Security: Good Practices and Recommendations on the Security of Big Data Systems.”. https://www.enisa.europa.eu/publications/big-data-security/at_download/fullReport (accessed April 24, 2018).

Obar, Jonathan A. “Big Data and The Phantom Public: Walter Lippmann and the fallacy of data privacy self-management.” *Big Data & Society* 2, no. 2 (2015): 205395171560887.

“Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression.”. http://www2.ohchr.org/english/bodies/hrcouncil/docs/17session/A.HRC.17.27_en.pdf (accessed March 27, 2018).

Roth, Kenneth. “The battle over encryption and what it means for our privacy.”. https://www.hrw.org/news/2017/06/28/battle-over-encryption-and-what-it-means-our-privacy (accessed April 23, 2018).

Roy, Mekhala. “Data anonymization techniques less reliable in era of big data.”. http://searchcompliance.techtarget.com/feature/High-dimensional-info-complicates-data-anonymization-techniques (accessed February 23, 2018).

Sadler, Christopher. “Protecting Privacy with Secure Multi-Party Computation.”. https://www.newamerica.org/oti/blog/protecting-privacy-secure-multi-party-computation/ (accessed April 24, 2018).

Schonschek, Oliver, and Nico Litzel. “Big Data Encryption – das große Verschlüsseln.”. https://www.bigdata-insider.de/big-data-encryption-das-grosse-verschluesseln-a-484370/ (accessed March 16, 2018).

Taneja, Hemant. “The need for algorithmic accountability.”. https://techcrunch.com/2016/09/08/the-need-for-algorithmic-accountability/ (accessed December 21, 2016).

Torra, Vicenç, and Guillermo Navarro-Arribas. “Big Data Privacy and Anonymization.” In *Privacy and Identity Management: Facing up to Next Steps*. Edited by Anja Lehmann et al., 15–26. Springer, 2016.

Vedder, Anton, and Bart Custers. “Whose Responsibility Is It Anyway? Dealing with the Consequences of New Technologies.” In *Evaluating new technologies: Methodological problems for the ethical assessment of technology developments.* vol. 3. Edited by Anthony M. Cutter, Marcus Düwell and Paul Sollie. 1st ed., 21–34. International library of philosophy and scientific method 3. New York: Springer, 2009.

Wang, Jianwu, Daniel Crawl, Shweta Purawat, Mai Nguyen, and Ilkay Altintas. “Big data provenance: Challenges, state of the art and opportunities.” In *Proceedings of the 2015 IEEE International Conference on Big Data*, 2509–16. IEEE, 2015. http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7364047 (accessed December 14, 2017).

Weathington, John. “Big data privacy is a bigger issue than you think.”. https://www.techrepublic.com/article/big-data-privacy-is-a-bigger-issue-than-you-think/ (accessed February 23, 2018).