D4.2
Overview of design requirements

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.

D4.2 Overview of design requirements

Work package	WP 4 – Design requirements for new 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) Bart Custers (Leiden University)
Internal review	Richard Stevens (IDC)
Due Date	M25 (January 2019)
Date	31 January 2019
Version	1.0
Type	Report
Dissemination level	Public

This document is Deliverable 4.2 of Work Package 4 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 deliverable discusses requirements for the design and use of big data solutions. The software solutions it focuses on are considered "privacy preserving" because they include privacy-preserving technologies or are used in an environment in which non-technical measures were taken to preserve privacy. The design of a solution can only lay the foundation; to achieve privacy-preserving data sharing and usage, it is important that the solution is used in an appropriate manner and environment. Four quite general requirements are proposed with the goal to provide guidance to developers and operators of big data solutions by highlighting aspects that need to be considered, outlining established principles, and describing good practices:

- **Embed security and privacy features** – The importance of privacy-preserving technologies grows continuously. Security and privacy features based on these technologies need to be embedded in big data solutions rather than provided as add-ons. The features need to be activated and, if possible, configured so that they provide a high level of privacy. A tight integration increases the probability that the technologies work effectively and efficiently. Apart from that, operators of big data solutions are unlikely to purchase and use add-ons.
- **Take preventive measures** – Privacy breaches need to be prevented because they may have negative implications for data subjects, may lead to significant fines and may negatively affect the reputation of the organisations involved. Recent data breaches demonstrate the limitations of reactive approaches that prescribe measures to be taken when privacy is violated or when certain rules are broken. Technologies are considered to be proactive in the sense that they prevent incidents or rule violations in the first place.
- **Connect people, processes and technology** – To ensure privacy, knowledgeable people as well as proper processes are essential complements to privacy-preserving technologies. In general, technologies have limitations and alone they are not sufficient. At some level, non-technical measures will always be necessary to make sure a given technology functions as expected. Moreover, data protection officials are needed that are aware and able to assess the impact on privacy or risks regarding personal data that is collected and used by organisations.
- **Comply with laws and corporate policies** – Big data solutions need to comply with laws and corporate policies. At the same time, they must be flexible enough to be adapted to changing demands, not only with respect to the business side but also with respect to the regulatory side. The heavy discussions that recently accompanied the entry into force of the EU General Data Protection Regulation (GDPR) show that achieving legal compliance is not always trivial, even if the legal text has been available for a long time.

If these requirements are met by organisations dealing with personal data, it is likely that trust in the data economy will increase in the long run. Privacy-preserving technologies are increasingly indispensable as information systems are more and more networked within organisations and across organisational boundaries, and datasets are becoming larger and more heterogeneous. Nevertheless, they need to be complemented by knowledgeable people, proper processes, and laws and corporate policies that not only exist but to which organizations are compliant.

The requirements discussed originate from the analysis of previous e-SIDES deliverables, a review of related previous work and an analysis of design challenges faced in the context of privacy-preserving technologies. The review of related work and the discussion of design challenges are provided as part of this deliverable. With respect to related work documents such as the OECD Privacy Guidelines, the international standard ISO/IEC 29100:2011 and the US Federal Trade Commission's fair information practices are introduced and discussed in the light of the e-SIDES project. Moreover, the EU General Data Protection Regulation as well as works concerning protection goals and various aspects of system design and privacy by design are addressed.

With respect to design challenges, this deliverable focuses on several classes of privacy-preserving technologies. For each class, the current state of development is described as well as key actors that are active in the respective field. For some of the classes, a more detailed example is given that indicates the direction in which the field is moving. The deliverable discusses aspects of the classes of privacy-preserving technologies in the light of the GDPR. Finally, we outline key design challenges that have to be overcome paying particular attention to the needs faced in the context of big data applications as well as promising ways to overcome them.

Contents

Grant Agreement number: 731873

Figures

Abbreviations

ABAC	Attribute-based access control
AI	Artificial intelligence
ASP	Anonymising service provider
DPA	Data protection authority
EU	European Union
FHE	Fully homomorphic encryption
FTC	Federal Trade Commission
GC	Garbled circuit
GDPR	General Data Protection Regulation
GPU	Graphics processing unit
MPC	Multi-party computation
PIA	Privacy impact assessment

1. Introduction

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

1.1. Background

This report is Deliverable 4.2 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).

This deliverable discusses requirements for the design and use of big data solutions. The solutions we focus on can be considered "privacy preserving" because they include privacy-preserving technologies or are used in an environment in which non-technical measures were taken to preserve privacy. In most cases, both points are likely to apply. An appropriate solution design can only lay the foundation. To achieve privacy-preserving data sharing and usage, it is not less important that the solutions are used in an appropriate manner and environment. The presented requirements are based on the results of the gap analyses provided in D4.1 and discussed in the light of related previous work.

Moreover, the report describes design challenges faced in the context of privacy-preserving technologies paying particular attention to the requirements faced in the context of big data. The focus is on the classes of technologies introduced in D3.1 and assessed in D3.2. For each class, both technological and legal aspects are discussed. It is essential to take both aspects into account, as the technologies are based on the possibilities and limitations of engineering as well as the relevant legal regimes in which they are introduced. The relevance of the classes depends not only on the application context but also on the stages of the data lifecycle:

- **Data collection:** Data that has been collected and is relevant for big data analytics is often sensitive. Moreover, ownership issues are not always clear. For instance, although service providers may own sensors or devices, data may pertain to the physical residents of smart homes or the drivers of connected cars. Therefore, these users should know what data is collected, stored and shared. Moreover, they should be able to interrupt the collection and ask for the data to be eliminated. Technologies for accountability and transparency, data provenance, and access, portability and user control are particularly relevant with respect to data collection.
- **Data transfer:** The transmission of data through unsecure networks must be protected. Confidentiality and integrity need to be ensured for any data transfer. Confidentiality is securing sensitive data against a malicious user and integrity is preserving the accuracy and consistency of the data. Encryption technologies are thus relevant for data transfer.
- **Data storage:** Data stored with personally identifiable information (or identifiers), particularly if third-party storage is used, is a serious threat to data privacy. Personal and quasi identifiers describe personally identifiable information. These attributes can directly or in-directly reveal personal information. Technologies for anonymisation and sanitisation as well as technologies for deletion are most relevant with respect to data storage.
- **Data processing:** The processing of data should, wherever possible, be independent of sensitive information. Storing the data used for analysis as mentioned above can achieve this. However, a

Grant Agreement number: 731873

key challenge is to find the right trade-off between the amount of privacy and information loss. Multi-party computation (MPC) technologies as well as certain encryption technologies have the potential to render finding a trade-off unnecessary.

- **Data usage:** The use of data should be regulated through proper authentication and authorisation. Big data solutions need to be configurable in a way that allows assigning rights to execute analysis jobs to appropriate users and access the generated results. Technologies for access control and policy enforcement are thus particularly relevant with respect to data usage.

An illustration of the lifecycle is provided in Figure 1. The lifecycle-related discussion takes ideas of Chakravorty et al.[1] and Chen & Zhao[2] into account.

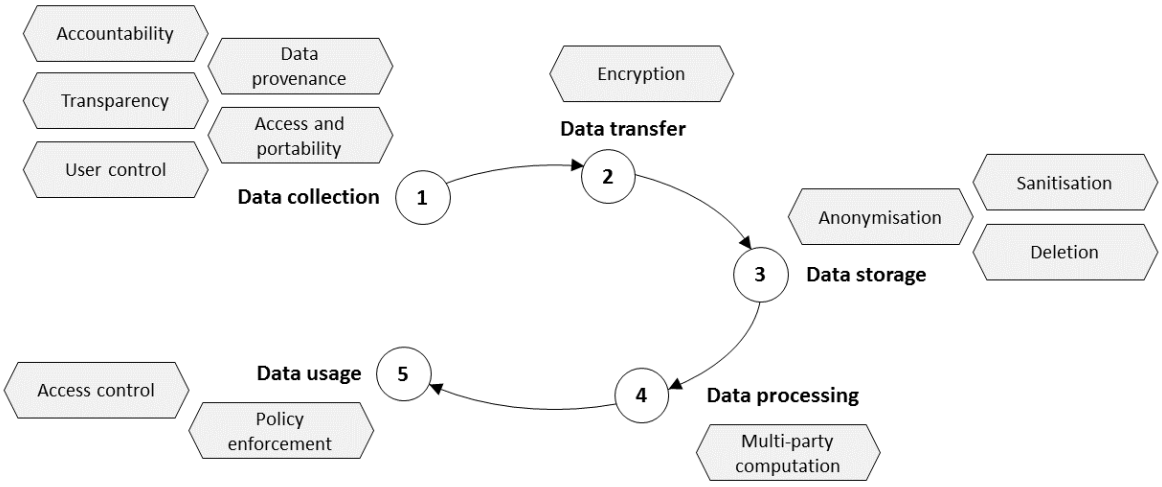

Figure 1 Relevance of technology classes along the data lifecycle

One natural commonality is that each of these technologies aims to uphold the principle of privacy by design, which is prescribed by the GDPR. The GDPR has introduced the (current) broadest existing definition for personal data within data protection law internationally, which is "any information related to an identified or identifiable natural person"[3]. Based on this broad definition, the extensive variety of available privacy-preserving technologies is certainly welcome as they aim to address, through differing means, how identified and identifiable information can be protected from undesired access. It is obvious, however, that not all approaches and techniques that may be relevant for addressing ethical and societal issues resulting from big data applications can be covered. e-SIDES covers technologies for anonymisation, sanitisation, encryption, deletion, MPC, access control, policy enforcement, accountability, data provenance, transparency, access and portability, and user control.

[1] 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)
[2] Deyan Chen and Hong Zhao, "Data Security and Privacy Protection Issues in Cloud Computing," in *Proceedings of the 2012 International Conference on Computer Science and Electronics Engineering*, 647–51 (IEEE, 2012)
[3] Art. 4(1) of the GDPR

Grant Agreement number: 731873

We demonstrate in this deliverable that design challenges regarding each class of privacy-preserving technologies are predicated upon the big data context and the relevant legal regime in which they are introduced. For the introduction of any technology class within certain sectors, such as healthcare, additional legal regimes would be applicable. Distinctions relevant for the processing of special categories of data under the GDPR would be directly applicable for personal data processing within healthcare. But in other contexts, such as transportation, sensitive information can be inferred about travellers during profiling practices when, for instance, the data related to the disabilities of certain travellers can be revealed.[4] While there may be multiple legal regimes that touch upon the processing of personal data, the GDPR is applicable across all contexts except for law enforcement and as such, it will be relied upon and referenced when discussing design challenges pertaining to certain technologies. Further, many of the principles mentioned in related previous works such as the OECD Privacy Guidelines (see section 2.1) are built into its binding prescriptions.

Essentially, the goal of this deliverable is to provide guidance to developers and operators of big data solutions by

- highlighting aspects that need to be considered,
- outlining established principles, and
- describing good practices.

This deliverable lays the foundation for the assessment of big data solutions under development and the formulation of recommendations in WP5. The assessment will focus on issues including the extent to which developers and operators of big data solutions currently meet the requirements proposed in this deliverable, and what this means for the issues and values e-SIDES has identified. Additionally, we will study barriers that keep developers and users from meeting the requirements. The work on barriers will form the foundation for the recommendations.

1.2. Methodology

This deliverable is based on desk research.

The overview of work related to design requirements for big data solutions focuses on aspects such as privacy principles, protection goals, system design, and privacy by design and its implementation.[5] We investigate not only the particular relevance of previous work for e-SIDES but also how it is related to different technology classes. The main purpose of the review is to make sure the e-SIDES design requirements build upon and are well-aligned with previous work.

While the assessment presented in D3.2 focused on the suitability of the privacy-preserving technologies to address societal and ethical issues, we now focus on the state of development of the technologies. For each technology class, we analyse the current state of development, the main actors active in the

[4] Bart H. M. Custers, "Data Mining and Group Profiling on the Internet," in *Ethics and the internet*, ed. A.H Vedder, 87–104 (Antwerpen, Groningen: Intersentia, 2001)

[5] See also Jeroen van den Hoven, "Value Sensitive Design and Responsible Innovation," in *Responsible innovation: Managing the responsible emergence of science and innovation in society*, ed. Richard Owen, J. R. Bessant and Maggy Heintz, 75–83 (Chichester: Wiley, 2013)

respective field, key legal aspects as well as key design challenges and promising ways how they can be overcome. For some of the technology classes, we provide a concrete example showing in a bit more detail a direction in which the field goes.

The proposed requirements for the design and use of big data solutions, which can be considered privacy preserving because they include privacy-preserving technologies or are used in an environment in which non-technical measures were taken to preserve privacy, are based on previous e-SIDES deliverables (especially, the gap analysis documented in D4.1 and the assessment of classes of privacy-preserving technologies presented in D3.2), a review related previous work (presented in section 2) and an analysis of design challenges in the context of privacy-preserving technologies (presented in section 3).

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 work related to design requirements for big data solutions that can be considered privacy preserving.
- Section 3 describes design challenges faced in the context of privacy-preserving technologies paying particular attention to requirements faced in the context of big data.
- Section 4 concludes the deliverable with a description of requirements for the design and use of big data solutions that can be considered privacy preserving because they include privacy-preserving technologies or are used in an environment in which non-technical measures were taken to preserve privacy.

2. Related previous work

This section provides an overview of work related to design requirements for big data solutions that can be considered privacy preserving. The overview covers work that focuses on aspects such as privacy principles, protection goals, system design, and privacy by design and its implementation. We discuss not only the particular relevance of previous work for e-SIDES but also how it is related to the classes of technologies first described within the scope of D3.1. The main purpose of this section is to make sure the e-SIDES design requirements build upon and are aligned with previous work.

2.1. OECD principles

In July 2013, the OECD Council adopted a revised Recommendation Concerning Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data ("Privacy Guidelines"). This revision was the first since the original 1980 release of the guidelines, which were the first comprehensive set of legal principles enforcing the protection of the right to privacy. The new OECD guidelines arose out of a call by Ministers in the 2008 Seoul Declaration for the Future of the Internet Economy to assess the guidelines in the light of changing technologies, markets and user behaviour, and the growing importance of digital identities.

Within the scope of its updated guidelines, the OECD refers to the following principles:[6]

- **Collection limitation** – The collection of personal data is lawful, limited, and happens with the knowledge or consent of the data subject.
- **Data quality** – Personal data should be relevant to the purposes for which they are to be used, be accurate, complete and kept up-to-date.
- **Use limitation** – The purposes of the collection must be specified upfront (purpose specification), and the use of the data after collection is limited to that purpose.
- **Security safeguards** – Personal data must be adequately protected.
- **Openness** – The nature and extent of the data processing and the controller responsible must be readily available.
- **Individual participation** – Individuals have the right to view, erase, rectify, complete or amend personal data stored that relates to him.
- **Accountability** – A data controller must be accountable for complying with these principles.

The e-SIDES classes of privacy-preserving technologies are well suited to put these principles into practice. Of particular relevance for all principles but the one focusing on security safeguards are technologies for user control, access and portability, data provenance, transparency and accountability. With respect to security safeguards, technologies for encryption, deletion, access control and policy enforcement are most relevant.

Moreover, the OECD introduces a number of new concepts, including:

[6] "The OECD Privacy Framework," (OECD, 2013), https://www.oecd.org/sti/ieconomy/oecd_privacy_framework.pdf (accessed October 23, 2018)

- **National privacy strategies** – While effective laws are essential, the strategic importance of privacy today also requires a multifaceted national strategy co-ordinated at the highest levels of government.
- **Privacy management programmes** – These serve as the core operational mechanism through which organisations implement privacy protection.
- **Data security breach notification** – This provision covers both notice to an authority and notice to an individual affected by a security breach affecting personal data.

The new concepts are not directly related to the e-SIDES classes of privacy-preserving technologies. The use of the technologies, however, can be promoted by referring to them explicitly in national privacy strategies and privacy management programmes. Incident notification can be made more reliable if technologies for transparency and accountability are used.

2.2. ISO/IEC 29100:2011 principles

ISO/IEC 29100:2011 provides a privacy framework, which specifies a common privacy terminology, defines the actors and their roles in processing personal data, describes privacy safeguarding considerations and provides references to known privacy principles for information technology.[7] The framework is applicable to natural persons and organisations involved in specifying, procuring, architecting, designing, developing, testing, maintaining, administering, and operating information and communication technology systems or services where privacy controls are required for the processing of personal data. The framework was first published in 2011 and amended in 2018.

ISO/IEC describes the following privacy principles that form the basis of their framework:[8]

- **Consent and choice** – Provisions should be made to provide data subjects with the opportunity to choose how their personal data is handled and to allow them to withdraw consent easily and free of charge.
- **Purpose legitimacy and specification** – A purpose can require a legal basis or a specific authorisation by a data protection authority or a government authority. The purpose should be specified using language that is both clear and appropriately adapted to the circumstances.
- **Collection limitation** – Organisations should not collect personal data indiscriminately. Both the amount and the type of personal data collected should be limited to that which is necessary to fulfil the purpose specified by the data controller.
- **Data minimisation** – Data processing procedures and ICT systems should be designed and implemented in a way that minimises the personal data which is processed, ensures the adoption of the need-to-know principle, avoids the identification of data subjects where possible and deletes personal data as soon as possible.
- **Use, retention and disclosure limitation** – Use, retention and disclosure of personal data should be limited to that which is necessary in order to fulfil specific, explicit and legitimate purposes.

[7] https://www.iso.org/standard/45123.html
[8] "ISO/IEC 29100: Information technology - Security techniques - Privacy framework," (ISO/IEC, 2011)

Grant Agreement number: 731873

- **Accuracy and quality** – Personal data which is processed should be accurate, complete, up-to-date, adequate and relevant for the purpose of use. The reliability of personal data collected from a source other than the data subject should be ensured before it is processed.[9]
- **Openness, transparency and notice** – Data subjects should be provided with clear and easily accessible information about the data controller's policies, procedures and practices with respect to the processing of personal data.
- **Individual participation and access** – Data subjects should be given the ability to access and review their personal data. Moreover, data subjects should be allowed to challenge the accuracy and completeness of the personal data and to have it amended, corrected or removed.
- **Accountability** – The processing of personal data entails a duty of care and the adoption of concrete and practical measures for its protection.
- **Information security** – Personal data should be protected with appropriate controls at the operational, functional and strategic level to ensure the integrity, confidentiality and availability of the data, and protect it against risks such as unauthorised access, destruction, use, modification, disclosure or loss throughout the whole of its life cycle.
- **Privacy compliance** – It should be verified and demonstrated that the processing meets data protection and privacy safeguarding requirements by periodically conducting audits using internal auditors or trusted third-party auditors. Additionally, privacy risk assessments should be developed and maintained.

The privacy principles described by ISO/IEC were derived from existing principles developed by a number of states, countries and international organisations. The framework focuses on the implementation of the principles in ICT systems and the development of privacy management systems to be implemented within organisational ICT systems. According to ISO/IEC, the principles should be used to guide the design, development and implementation of privacy policies and privacy controls. Additionally, they may be used as a baseline in the monitoring and measurement of performance, benchmarking and auditing aspects of privacy management programs in an organisation.

The principles of ISO/IEC 29100:2011 go beyond the principles of the OECD guidelines. Therefore, all of the classes of technologies mentioned in the previous section are also relevant here. The principle focusing on purpose legitimacy and specification, however, clearly shows the limits of technical measures. For example, technologies are not able to assess if a purpose has a legal basis or if there is a specific authorisation by a data protection authority or a government authority. Similarly, there are no technologies capable of specifying a given purpose using language that is clear and adapted to specific circumstances.

2.3. EU GDPR principles

The EU General Data Protection Regulation (GDPR) can be considered as currently the most developed legal framework for the protection of data and privacy. Since 25 May 2018, the GDPR has been in force,

[9] Bart H. M. Custers, "Effects of Unreliable Group Profiling by Means of Data Mining," in *Discovery science: 6th international conference, DS 2003 Sapporo, Japan, October 17-19, 2003 : proceedings*, ed. Gunter Grieser, Y. Tanaka and Akihiro Yamamoto, 291–6, Lecture Notes in Computer Science 2843. Lecture notes in artificial intelligence (Berlin, New York: Springer, 2003)

offering a comprehensive set of rights for data subjects in order to empower them in exercising more control over data controllers and the whereabouts of their data. The GDPR is the first legislation of its kind that puts an explicit obligation upon the shoulders of all data controllers – who handle the data of EU residents even beyond the territory of the EU – to demonstrate compliance with all GDPR rules and principles. This provision fosters the capabilities of data subjects to hold data controllers accountable for how they process data subjects' data. In case of non-compliance, the GDPR also assigns significant powers to data protection authorities in order to carry out punishments, such as in the case of data breaches (Art. 33). Furthermore, the regulations requires data controllers to take security measures in order to ensure the lawful, fair and transparent processing of data (Art. 32). This is evident in the privacy by design and privacy by default principles (Art. 25), which were expounded upon in an ENISA report[10] on privacy-by design.

AS the GDPR currently offers the most comprehensive set of principles and rules for the protection of data and privacy, we focus on this legal framework when discussing legal aspects related to design challenges faced in the context of privacy-preserving technologies in section 3.

The GDPR describes six data protection principles that organisations need to follow when processing personal data. The following principles are described in Art. 5 of the GDPR:[11]

- **Lawfulness, fairness and transparency** – Personal data must be processed lawfully, fairly and in a transparent manner in relation to the data subject.
- **Purpose limitation** – Personal data must be collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes.
- **Data minimisation** – Personal data must be adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed.
- **Accuracy** – Personal data must be accurate and, where necessary, kept up to date.
- **Storage limitation** – Personal data must be kept in a form, which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed.
- **Integrity and confidentiality** – Personal data must be processed in a manner that ensures appropriate security of the personal data, including protection against unauthorised or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organisational measures.

According to the GDPR, the data controller is responsible for complying with the principles and must be able to demonstrate the organisation's compliance with the principles (**Accountability**).

The GDPR principles are also mostly in line with those of the OECD guidelines. Therefore, the classes of technologies mentioned in the section on the OECD principles are also relevant here. The e-SIDES classes

[10] 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)

[11] "Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC," (European Parliament, Council, 2016)

of technologies are relevant to put the principles into practice but there are limits, for instance, with respect to the assessment of lawfulness or legitimacy.

2.4. US FTC fair information practices

In 1998, the US Federal Trade Commission (FTC) summarised widely-accepted principles regarding the collection, use and dissemination of personal data. The FTC identified five core principles of privacy protection:[12]

- **Notice/awareness** – Data collectors must disclose their information practices before collecting personal information from consumers.
- **Choice/consent** – Consumers must be given options with respect to whether and how personal information collected from them may be used for purposes beyond those for which the information was provided.
- **Access/participation** – Consumers should be able to view and contest the accuracy and completeness of data collected about them.
- **Integrity/security** – Data collectors must take reasonable steps to assure that information collected from consumers is accurate and secure from unauthorised use.
- **Enforcement/redress** – A reliable mechanism to impose sanctions for noncompliance with these fair information practices is critical.

The US FTC fair information practices served as basis for many later works on privacy. Technologies for transparency and user control are most relevant to put the first three principles into practice. Technologies for encryption, deletion, access control and policy enforcement are the technologies of choice for the principle focusing on integrity and security. Technologies are less relevant for the principle focusing on enforcement and redress. Technologies for accountability may be useful but the limits of technical measures become obvious here. Imposing sanctions for noncompliance usually does still go beyond current technical capabilities.

2.5. Protection goals

Hansen et al.[13] describe six protection goals for privacy engineering that provide a scheme for addressing the legal, technical, economic, and societal dimensions of privacy and data protection in complex IT systems:

- **Confidentiality** – One technical realisation of the confidentiality protection goal is commonly found in the domain of cryptography. Furthermore, access control enforcement contributes to the confidentiality protection goal as well.
- **Integrity** – The realisation of integrity is prevalently based on a cryptographic scheme that allows for detection of modifications of data. Other means of realisation of the integrity protection goal are based on redundancy and comparison, data verification, and access control enforcement.

[12] https://web.archive.org/web/20090331134113/http://www.ftc.gov/reports/privacy3/fairinfo.shtm

[13] Marit Hansen, Meiko Jensen and Martin Rost, "Protection Goals for Privacy Engineering," in *Proceedings of the 2015 IEEE Security and Privacy Workshops*, 159–66 (San Jose, CA, USA: IEEE, 2015)

- **Availability** – Availability can only be realised by adding redundancy to the system. This way, if one of the copies is destroyed or altered, the other copies remain intact, and the original data can be restored.
- **Unlinkability** – Under the umbrella of this protection goal, a lot of commonly known properties are subsumed. For instance, unlinkability refers to the property of anonymity and its different aspects of realisation. Another typical technique for enhancing unlinkability consists in the implementation of access restrictions for data and processes. However, the idea of the unlinkability protection goal goes way beyond these technologies. According to Hansen et al., one of the core challenges here is the external processing paradox: in order to use the services an external entity provides, it becomes necessary to provide the required input data for these services to that entity. Here, manifold experimental and early cryptographical approaches have emerged during the last decades. Techniques of secure computation and homomorphic encryption support the processing of data without learning the data.
- **Transparency** – The common techniques for fostering the protection goal of transparency are centred around storing and delivering information. Beyond system documentation, dedicated and complete logging mechanisms play an essential role in the provisioning of effective transparency. The most promising technologies regarding transparency improvements are not based on the use of data modification techniques, but require dedicated transparency services to be implemented alongside the core services of the particular IT systems. These services then take care of storing, linking, aggregating, and providing all information required for achieving a sufficient level of transparency regarding personal data. Moreover, these services then need to be made accessible to the entitled entities in a usable and understandable way.
- **Intervenability** – The protection goal of intervenability does not have many technologies and techniques elaborated to the degree of daily use. The idea of intervenability is to enable direct action by entitled entities, such as the data processor, a supervisory authority or the data subject. What typically can be found in terms of intervenability is a configuration menu for the user's core personal data and a clerk-operated help desk. Interference with ongoing processes, however, is rarely implemented.

Hansen et al. emphasise that there is no possibility to ensure 100% of each of the goals simultaneously. The three explicit conflict axes are shown in Figure 2.

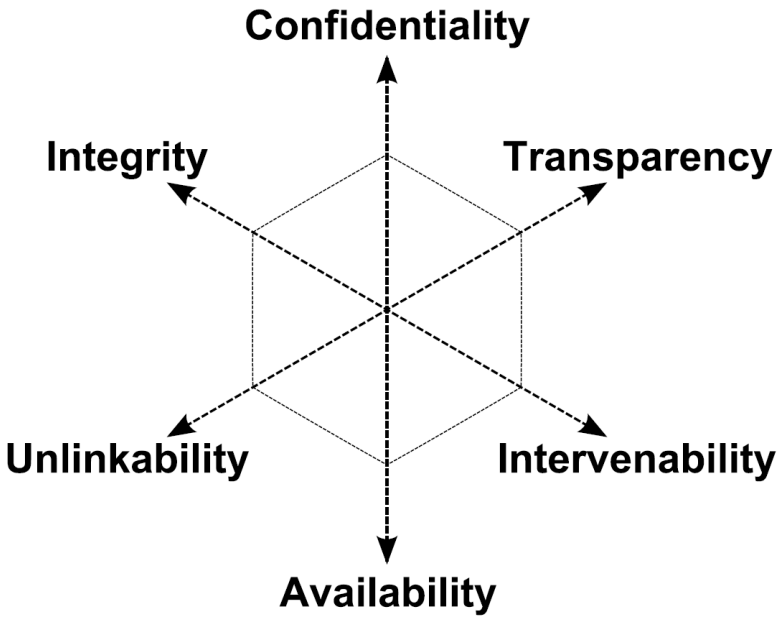

Figure 2 The six protection goals for privacy engineering

Danezis et al.[14] stress that it is essential to properly balance the requirements derived from the six protection goals. Considerations on lawfulness, fairness and accountability may provide guidance for balancing the requirements and deciding on design choices and appropriate safeguards.

As mentioned above, confidentiality and integrity can best be achieved by technologies for encryption and access control. As adding redundancy is a means that plays a roles with respect to integrity and availability, technologies for deletion are relevant too. Unlinkability can best be achieved my means of technologies for anonymisation, sanitisation, access control and MPC. Technologies for transparency and accountability are most relevant to achieve the protection goal of transparency. Concerning intervenability, technologies for user control and policy enforcement appear to be most relevant. Although at least two classes of privacy-preserving technologies can be mentioned for each of the six protection goals, it becomes obvious that the technologies have clear limits as well as that there are strong interactions and interdependencies between the technologies. For example, technologies for anonymisation, sanitisation and deletion on the one side and technologies for accountability and transparency on the other side may be contradictory.

2.6. System design

With respect to system design, we introduce common design strategies and mistakes, describe a design framework and outline the idea of privacy patterns.

[14] Danezis et al., "Privacy and Data Protection by Design - from policy to engineering"

Grant Agreement number: 731873

2.6.1. Design strategies

Hoepman[15] identified eight privacy design strategies. Most of them were not only described generally but also put in relation with specific design patterns. The design patterns allow connection the design strategies with the classes of privacy-preserving technologies studied in e-SIDES.

Four of the design strategies are data-oriented strategies:

- **Minimise** – The amount of data that is processed should be restricted to the minimum possible. Common design patterns that implement this strategy are *Select Before You Collect*, *Pseudonymisation* and *Anonymisation*.
- **Hide** – Any personal data, and their interrelationships, should be hidden from plain view. The design patterns that belong to this strategy are manifold. One of them is *Encryption* of data. Another one is *Mix Networks*[16] to hide traffic patterns.
- **Separate** – Personal data should be processed in a distributed fashion whenever possible.
- **Aggregate** – Personal data should be processed at the highest level of aggregation and with the least possible detail in which it is (still) useful. Examples of design patterns that belong to this strategy are *Aggregation Over Time* (used in smart metering) and *Dynamic Location Granularity* (used in location-based services).

With respect to the classes of privacy-preserving technologies, technologies for anonymisation, sanitisation and encryption are most relevant for the first two data-oriented strategies. The other two strategies are concerned with data storage and thus go beyond the scope of privacy-preserving technologies. Nevertheless, it makes a lot of sense to consider their implementation in the context of big data solutions.

The other four are process-oriented strategies:

- **Inform** – Data subjects should be adequately informed whenever personal data is processed. Possible design patterns for this strategy are the *Platform for Privacy Preferences* (P3P) and *Data Breach Notifications*.
- **Control** – Data subjects should be provided agency over the processing of their personal data.
- **Enforce** – A privacy policy compatible with legal requirements should be in place and should be enforced. *Access Control* and *Sticky Policies* are examples of design patterns that implement this strategy. Another example, according to Hoepman, is *Privacy Rights Management*, which is a form of digital rights management involving licenses to personal data.

[15] Jaap-Henk Hoepman, "Privacy Design Strategies," in *ICT systems security and privacy protection: 29th IFIP TC 11 International Conference, SEC 2014, Marrakech, Morocco, June 2-4, 2014. Proceedings*, vol. 428, ed. Nora Cuppens-Boulahia et al., 446–59, IFIP Advances in Information and Communication Technology 428 (Heidelberg: Springer, 2014)

[16] Mix networks is a technique to provide anonymous communications. To make tracing messages through a network as difficult as possible, they are relayed by a sequence of trusted intermediaries.

Grant Agreement number: 731873

- **Demonstrate** – Data controllers must be able to demonstrate compliance with the privacy policy and any applicable legal requirements. Design patterns that implement this strategy are, for example, the use of *Privacy Management Systems*, *Logging* and *Auditing*.

Technologies for transparency, user control, access control, policy enforcement and accountability are most relevant with respect to the process-oriented strategies.

2.6.2. Design mistakes

Hansen[17] describes ten common mistakes in system design from a privacy perspective. Hansen notes that the list is neither complete nor the order of the errors should be overestimated, but draws on her experience working for a data protection authority where the errors have all occurred several times and in different circumstances:

- **Storage as default** – Storing data is a precondition for all kinds of data processing, which has to be considered when assessing privacy risks. The data may reside on a plenitude of IT systems with individual providers while there is no guarantee that the data will be effectively erased as soon as they are not necessary any more. In particular, temporary files and log files are regularly neglected when assessing privacy risks. This also refers to the right to be forgotten and erasure (GDPR).[18]
- **Linkability as default** – For data processing, it is easier to address objects by specific identifiers, and generally this means to assign unique identifiers that enable the linkage "identifier – object" and "identifier – same identifier". Pushing the idea further would lead to a central global database of all subjects and objects where different parties would get different access rights, which would be a nightmare from a privacy perspective because of the mass of linkable data. On the contrary, data minimisation is based on unlinkability as far as possible, and linkage control for data subjects are key for their privacy.
- **Real name as default** – From a privacy perspective, the real name policy of many online services is problematic, rather one or many pseudonyms should be used. With technologies such as private credentials, anonymity and accountability can even be achieved at the same time.
- **Function creep as feature** – Widening of the data processing beyond the original purpose or context violates the principle of purpose binding and can pose risks to privacy that have to be considered when assessing the system. Instead, the objective of contextual integrity should be taken seriously.
- **Fuzzy or incomplete information as default** – From a privacy perspective, accurate and complete information on the planned and performed data processing is a necessity, as data controllers and data processors have to know how their IT systems and organisational procedures work, and this information is required when asking data subjects for consent or being asked by supervisory authorities.

---

[17] Marit Hansen, "Top 10 Mistakes in System Design from a Privacy Perspective and Privacy Protection Goals," in *Privacy and identity management for life: 7th IFIP WG 9.2, 9.6/11.7, 11.4, 11.6 international summer school, Trento, Italy, September 5-9, 2011; revised selected papers*, vol. 375, ed. Jan Camenisch, 14–31, IFIP Advances in Information and Communication Technology 375 ([Berlin], [Heidelberg]: Springer, 2012)

[18] Hans Graux, Jef Ausloos and Peggy Valcke, "The Right to Be Forgotten in the Internet Era,", ICRI Research Paper 11 (KU Leuven, 2012)

- **Location does not matter** – Since there is no common globally valid and accepted law, the location of data processing is relevant for all kinds of access to the data that is stored or transmitted in a country.
- **No lifecycle assessment** – Many problems occur because the system design did not consider the full lifecycle of the data, the organisation or the system itself. Related are lock-in effects where data portability is not offered: If a user has not foreseen an exit strategy, it may be hard to change a provider because there are already established dependencies. If the provider can be changed, there is no guarantee that the data are erased on the provider's side.
- **Changing assumptions or surplus functionality** – Surplus functionality may water down or contradict the initially intended privacy guarantees. In particular, a surplus payment method, a business model based on profiling and advertising, or obligations from the police or homeland security could render all privacy efforts useless.
- **No intervenability foreseen** – The possibility to intervene (change and shut off the system) is relevant for the entities processing the data, for the supervisory authorities that may inspect the data processing system, and – at least partially – for data subjects whose data is being processed, simply because of their data subject rights.
- **Consent not providing a valid legal ground** – All processing of personal data is only lawful if a statutory provision permits it or if the data subject has consented. A consent that provides a valid legal ground has to meet various requirements: it has to be freely given, specific and informed, otherwise it is not valid.

Technical measures are not sufficient to avoid the design mistakes described by Hansen. The privacy-preserving technologies e-SIDES focuses upon are relevant with respect to some of the mistakes including those related to storage, information completeness, lifecycle assessment and intervenability. Technologies for deletion, transparency, access and portability, and user control are most relevant to deal with these common mistakes. Form a privacy perspective, however, it is essential to avoid all mistakes when developing big data solutions. Apart from integrating appropriate privacy-preserving technologies into solutions, it is clearly necessary to take further technical measures. Moreover, it is important that the operators of the solutions use them responsibly and take the necessary organisational measures.

2.6.3. Design framework

Spiekermann & Cranor[19] provide a framework for privacy-friendly system design. Figure 3 provides an overview of the framework.

[19] S. Spiekermann and L. F. Cranor, "Engineering Privacy," *IEEE Transactions on Software Engineering* 35, no. 1 (2009)

Privacy stages	identifiability	Approach to privacy protection	Linkability of data to personal identifiers	System Characteristics
0	identified	privacy by policy (notice and choice)	linked	• unique identifiers across databases • contact information stored with profile information
1	pseudonymous		linkable with reasonable & automatable effort	• no unique identifies across databases • common attributes across databases • contact information stored separately from profile or transaction information
2		privacy by architecture	not linkable with reasonable effort	• no unique identifiers across databases • no common attributes across databases • random identifiers • contact information stored separately from profile or transaction information • collection of long term person characteristics on a low level of granularity • technically enforced deletion of profile details at regular intervals
3	anonymous		unlinkable	• no collection of contact information • no collection of long term person characteristics • *k*-anonymity with large value of *k*

Figure 3 Framework for privacy-friendly system design

Spiekermann & Cranor discuss two approaches for systematically engineering privacy friendliness: privacy by architecture and privacy by policy. The privacy by policy approach focuses on the implementation of the notice and choice principles of fair information practices, whereas the privacy by architecture approach minimises the collection of identifiable personal data and emphasises anonymisation and client-side data storage and processing.

As the framework focuses on linkability, with respect to privacy-preserving technologies, technologies for anonymisation, sanitisation, deletion and policy enforcement are most relevant to achieve the desired system characteristics. Technical measures beyond the use of privacy-preserving technologies, however, are important too. The achievements of certain characteristics is, for instance, linked to question of data storage.

2.6.4. Privacy patterns

It was found to be difficult to translate concerns about the flow of personal data into technical artefacts such as software or hardware that address them. The aim of privacy patterns is to standardise the language for privacy-preserving technologies, document common solutions to privacy problems and help designers identify and address privacy concerns. A comprehensive list of such patterns is provided by privacypatterns.org[20]. The creation of the list was supported by the US Department of Homeland Security,

[20] https://privacypatterns.org/patterns/

the US National Institute of Standards and Technology, the Berkeley Center for Law and Technology and the EU via the PRIPARE project[21].

Examples for privacy patterns are:

- **Data breach notification** – This pattern assures that a certain minimum data breach notification delay is not exceeded.
- **Privacy icons** – A privacy policy, which is hard to understand by general audience, is summarised and translated into commonly agreed visual icons. A privacy icon is considered worth a thousand-word policy.
- **Onion routing** – This pattern provides unlinkability between senders and receivers by encapsulating the data in different layers of encryption, limiting the knowledge of each node along the delivery path.
- **Sticky policies** – Machine-readable policies are attached to data to define allowed usage and obligations as it travels across multiple parties, enabling users to improve control over their personal information.
- **Anonymity set** – This pattern aggregates multiple entities into a set, such that they cannot be distinguished anymore.
- **Personal data store** – Subjects keep control on their personal data that are stored on a personal device.

Drozd[22] provides an overview of privacy patterns and describes a catalogue that was developed specifically to be used by software architects. He defined privacy patterns for the principles specified in ISO/IEC 29100. The results of Drozd for the ISO/IEC 29100 principle *Collection limitation*, for instance, are in line with the patterns identified by Hoepman[23] concerning the *Minimise* strategy.

While some privacy patterns are related to the e-SIDES classes of privacy-preserving technologies others are not. The large number of patterns and differing descriptions make it difficult to discuss the relationships in detail. In any case, the patterns can help to translate concerns about the flow of personal data into technical artefacts and should therefore be checked for their relevance when developing systems.

2.7. Privacy by design

Concerning privacy by design, we address features and criteria that have to be taken into account, provide an overview of recommendations and discuss risk mitigation strategies.

[21] http://pripareproject.eu/
[22] Olha Drozd, "Privacy Pattern Catalogue: A Tool for Integrating Privacy Principles of ISO/IEC 29100 into the Software Development Process," in *Privacy and Identity Management. Time for a Revolution? 10th IFIP WG 9.2, 9.5, 9.6/11.7, 11.4, 11.6/SIG 9.2.2 International Summer School, Edinburgh, UK, August 16-21, 2015, Revised Selected Papers*, ed. David Aspinall et al., 129–40, IFIP Advances in Information and Communication Technology 476 (Cham: Springer International Publishing, 2016)
[23] Jaap-Henk Hoepman, "Privacy Design Strategies" in *ICT systems security and privacy protection*

2.7.1. Features

In line with the idea of privacy by design, Cavoukian & Jonas[24] argue that privacy is best proactively interwoven into business processes and practices. They describe seven privacy by design features:

- **Full attribution** – Every record needs to know from where it came and when. There cannot be merge/purge processing whereby entire records or fields are discarded.
- **Data tethering** – Adds, changes and deletes occurring in systems of record must be accounted for immediately.
- **Analytics on anonymised data** – The ability to perform advanced analytics over cryptographically altered data means organisations can anonymise more data before information sharing.
- **Tamper-resistant audit logs** – Every user search should be logged in a tamper-resistant manner. Even the database administrator should not be able to alter the evidence contained in this audit log.
- **False negative favouring methods** – The capability to more strongly favour false negatives is of critical importance in systems that could be used to affect someone's civil liberties.
- **Self-correcting false positives** – With every new data point presented, prior assertions are re-evaluated to ensure they are still correct, and if no longer correct, these earlier assertions can often be repaired.
- **Information transfer accounting** – Every secondary transfer of data, whether to human eyeball or a tertiary system, can be recorded to allow stakeholders to understand how their data is flowing.

With respect to the e-SIDES classes of privacy-preserving technologies, technologies for accountability are clearly most relevant for implementing the privacy by design features. However, they are not equally relevant for all features. Full attribution rather calls for data provenance technologies. Analytics on anonymised data require technologies for anonymisation, sanitisation and encryption. Information transfer accounting would benefit from transparency technologies. While anonymisation and sanitisation technologies can be useful to make sure false negatives are favoured, self-correcting false positives require technical measures that go beyond the e-SDIES classes of privacy-preserving technologies.

2.7.2. Criteria

Danezis et al.[25] describe key criteria to be considered in a privacy by design methodology:

- **Trust assumptions** – The choice of the trust relationships is a driving factor in the selection of architectural options and privacy-preserving technologies. Any disclosure of personal data is conditional upon a form of trust between the discloser and the recipient. There are different types of trust. Blind trust is the strongest form of trust; from a technical point of view it could lead to the weakest solutions, the ones most vulnerable to misplaced trust. Verifiable trust is granted by

[24] Ann Cavoukian and Jeff Jonas, "Privacy by Design in the Age of Big Data," (Information and Privacy Commissioner, Ontario, Canada, 2012), https://jeffjonas.typepad.com/Privacy-by-Design-in-the-Era-of-Big-Data.pdf (accessed December 14, 2017)
[25] Danezis et al., "Privacy and Data Protection by Design - from policy to engineering"

Grant Agreement number: 731873

default but verifications can be carried out a posteriori (for example using commitments and spot checks) to check that the trusted party has not cheated. In contrast, verified trust amounts, technically speaking, to a "no trust" option; it relies on cryptographic algorithms and protocols (such as zero knowledge proofs, secure MPC or homomorphic encryption) to guarantee, by construction, the desired property. Furthermore the amount of trust necessary can be reduced by the use of cryptographic techniques or distribution of data.

- **Involvement of the user** – For some systems (smart metering, electronic traffic pricing, etc.), it may be the case that no interaction with the user is necessary to get his consent (which is supposed to have been delivered through other, non-technical means, or which is not required because it is not the legal ground for the collection of the data) or to allow him to exercise his rights. In other cases, these interactions have to be implemented. Great care should be taken that the interface of the system allows subjects to exercise all their rights (informed consent, access, correction, deletion, etc.) without undue constraints.
- **Technical constraints** – Some constraints on the environment usually have to be taken into account such as, for example, the fact that a given input data is located in a specific area, provided by a sensor that may have limited capacities, or the existence (or lack of) communication channel between two components.
- **Architecture** – The last stage is then the definition of the architecture, including the type of components used, the stakeholders controlling them, the localisation of the computations, the communication links and information flows between the components.

Technologies for encryption and MPC are obviously the most relevant pricy preserving technologies to build verified trust relationships. With respect to user involvement in cases where user interactions have to be implemented, technologies for user control, access and portability are particularly relevant. Technical constraints cannot be taken into account and requirements concerning the architecture cannot be achieved by integrating privacy-preserving technologies into big data solutions.

2.7.3. Recommendations

D'Acquisto et al.[26] compiled a set of recommendations related to privacy by design in the context of big data for ENISA. The recommendations are:

- **Privacy by design applied** – Data Protection Authorities (DPAs), data controllers and the big data analytics industry need to actively interact in order to define how privacy by design can be practically implemented (and demonstrated) in the area of big data analytics, including relevant support processes and tools.
- **Decentralised versus centralised data analytics** – The research community and the big data analytics industry need to continue and combine their efforts towards decentralised privacy-preserving analytics models. Policy makers need to encourage and promote such efforts, both at the research and the implementation level.

[26] 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)

- **Support and automation of policy enforcement** – The research community and the big data analytics industry need to explore the area of policy definition and to embody relevant mechanisms for automated enforcement of privacy requirements and preferences. The policy makers need to facilitate the dialogue between research, industry and DPAs for effective policy development and automation models.
- **Transparency and control** – 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. DPAs need to support these efforts, encouraging the implementation of practical use cases and effective examples of transparency and control mechanisms that are compatible with legal obligations.
- **User awareness and promotion of privacy-preserving technologies** – The research community needs to adequately address aspects related to the reliability and usability of privacy-preserving technologies. The role of the DPAs is central in user awareness and promotion of privacy-preserving processes and tools in online and mobile applications.
- **A coherent approach towards privacy and big data** – Policy makers need to approach privacy and data protection principles (and technologies) as a core aspect of big data projects and relevant decision-making processes.

Privacy-preserving technologies are useful for implementing two of the recommendations made by D'Acquisto et al. for ENISA. Technologies for policy enforcement, transparency, user control are relevant for implementing the third and the fourth recommendation. The other recommendations require technical measures that go beyond the core capabilities of the e-SIDES classes of privacy-preserving technologies.

Danezis et al.[27], for example, also give recommendations related to privacy by design.

2.7.4. Risk mitigation strategies

In addition to general recommendations, D'Acquisto et al.[28] also focused on risk mitigation strategies. The authors suggest the following strategies:

- **Transparency and awareness** – The different stakeholders should agree on the allocation of data protection responsibilities and on designating who plays the role of single point of contact (PoC) for privacy-related issues, as well as for any technical assistance the user may need. This will facilitate the exercise of users' rights and increase the effectiveness of the information. The creation of a web resource where the user can retrieve his/her own data may also be a valuable tool in this sense.
- **User control** – User control needs to be reinforced, enabling users to effectively select between a number of choices (after having been informed on their impact on the user experience). Tools such as privacy preferences and personal data stores can be very interesting in this regard, providing also for smoothness and ease of use (no need to "exit" or disrupt the service).

[27] Danezis et al., "Privacy and Data Protection by Design - from policy to engineering"
[28] D'Acquisto et al., "Privacy by design in big data"

Grant Agreement number: 731873

- **Data minimisation** – In order to strengthen the data minimisation principles, as much as possible the information handed over between the parties should be engineered in order to be Boolean rather than fully analytical. This is part of the data collection process and should be explicitly addressed in each different application (e.g., smart metering).
- **Access control and encryption** – In order to avoid crime related risks, access to data should be based on the respect of the principle of separation of duties, where each actor is enabled to have access to the data relating to the portion of the service that he provides, on a strict need-to-know basis. Solutions within this context are encryption based access controls and search mechanisms. Other technical and organisational safeguards to minimise any risk of data misuse include log management, aggregation of data whenever individual level data are not required, audit and policy enforcement.
- **Retention, deletion and anonymisation** – In general, data should be kept only for the period that are absolutely necessary for the processing. Also, upon a data subject's specific request, and if no other legitimate interests or legally binding constraints exist, personal data should be deleted. If data need to be stored beyond the data retention periods, appropriate anonymisation methods need to be applied.
- **Privacy impact assessment** – A practical tool in all big data applications where many stakeholders are involved is the use of a Privacy Impact Assessment (PIA) prior to the deployment of the service. In the specific context of smart grids, the European Commission has developed a PIA template, which can be applied by network and smart grid system operators.[29]

The classes of privacy-preserving technologies studied in e-SIDES are highly relevant for the implementation of the proposed risk mitigation strategies. Technologies for transparency, access and portability, user control, access control, encryption, anonymisation, sanitisation and deletion may be useful to implement the challenges. Addressing the strategy focusing on data minimisation requires technical measures that go beyond privacy-preserving technologies. The use of privacy-preserving technologies may be the consequence of conducting a privacy impact assessment but typically, technical measures do not play a key role with respect to the actual assessment.

[29] David Wright, "The state of the art in privacy impact assessment," *Computer Law & Security Review* 28, no. 1 (2012)

3. Design challenges in the context of privacy-preserving technologies

This section describes design challenges faced in the context of privacy-preserving technologies paying particular attention to requirements faced in the context of big data. The design challenges focus on a comprehensive set of classes of privacy-preserving technologies.

While the focus of the assessment presented in D3.2 was on the suitability of the technologies to address societal and ethical issues, we now focus on the state of development of the technologies. For each class, we first describe the current state of development, mention key actors active in the respective field and may give an example that indicates the direction in which the field goes in a bit more detail. Afterwards, we discuss key legal aspects. Finally, we outline key design challenges that have to be overcome paying particular attention to the needs faced in the context of big data applications. Similar to Martin[30], who derives guidelines for a sustainable big data industry from an analysis of selected ethical issues and possible solutions, we analyse design challenges and ways to overcome them to back up the requirements for the design and use of big data solutions we provide.

The purpose of this section is to identify research needs and to outline avenues for future research and development in the field of privacy-preserving technologies. The classes of technologies introduced in D3.1 and assessed in D3.2 are discussed. Each of the following subsections will discuss both technological and legal aspects that are characteristic to the design of the respective class of privacy-preserving technologies.

3.1. Anonymisation and sanitisation

Anonymisation is performed by encrypting or removing personally identifiable information from datasets. Examples for privacy models that may be used include k-anonymity and differential privacy. Sanitisation is done by encrypting or removing sensitive information from datasets. Anonymisation is a type of sanitisation aiming at privacy protection. Sensitive information is removed from datasets by techniques such as masking data, substitution, shuffling or number variance. Anonymisation and sanitisation technologies are very relevant in the big data context since data cannot be controlled anymore as soon as it has been released. It is important that measures are taken to reduce the risk that people are re-identified or that sensitive attributes are inferred. There is a substantial amount of literature on technologies for anonymisation and sanitisation. Technologies can be used to protect, anonymise or aggregate data in ways that are effective and efficient.[31]

The key challenge is to determine the optimal balance between improved privacy protection and the usefulness of the data for decision-making. 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,

[30] Kirsten E. Martin, "Ethical Issues in the Big Data Industry," *MIS Quarterly Executive* 14, no. 2 (2015); see also Shannon Vallor, *Technology and the virtues: A philosophical guide to a future worth wanting*, First issued as an Oxford University Press paperback (New York, NY, United States of America: Oxford University Press, 2016)

[31] 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)

if not impossible. Sometimes it may be that there is no satisfying trade-off: either some utility and very weak privacy, or some privacy and hardly any utility. Further challenges 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. Technologies based on differential privacy are most promising. Sanitisation is good only to prevent accidental disclosure, not to provide protection from a motivated adversary. A key drawback is that there are many examples where anonymisation has failed. Anonymisation is particularly difficult with respect to medical data. The difficulties are caused, for instance, by free text 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. The mathematical background for anonymisation is established and very stable. However, guidelines for the choice of parameters such as ϵ in differential privacy are missing.[32]

European institutions that are particularly active in the research field are the Cardiff University and the University of Southampton. The Alessandro Faedo Institute of Information Science and Technologies is an institute of the Italian National Research Council (CNR) that is also very active in this field.[33]

There are also a number of legal concerns and requirements concerning anonymisation. The GDPR does not apply to the processing of truly anonymous data as it no longer relates to an identifiable natural person;[34] yet, where anonymised data is used in algorithmic correlations of other data, such correlations can still have detrimental consequences for individual persons despite the fact that their anonymised data fell beyond the definition of personal data and thus lies potentially beyond the GDPR. In most cases, such as where pseudonymisation is used and a person may be re-identified, the data controller will however be subject to the GDPR. The GDPR requires that where processing is conducted for "a purpose other than that for which the personal data have been collected is not based on the data subject's consent", the controller must ascertain whether another purpose is compatible with the purpose for which the data were initially collected, and one of the factors that must be taken into account is the use of appropriate safeguards, such as pseudonymisation.[35] In Art. 25(1) of the GDPR on privacy by design, measures such as pseudonymisation are considered to be designed to implement data protection principles like data minimisation, and are considered necessary to comply with the legislation and to protect the rights of data subjects. The GDPR also requires that organisations implement security measures that are appropriate to the risk, which may include pseudonymisation, as per Art. 32(1)(a). While the GDPR does not apply to processing of truly anonymous data as it no longer relates to an identifiable natural person.[36]

Design challenges faced in the context of anonymisation and sanitisation include:

---

[32] Stamatis Karnouskos and Florian Kerschbaum, "Privacy and Integrity Considerations in Hyperconnected Autonomous Vehicles," *Proceedings of the IEEE* 106, no. 1 (2018)
[33] Based on a Microsoft Academic query using the search terms "anonymisation" or "anonymization"
[34] Recital 26 of the GDPR
[35] Art. 6(4)(e) of the GDPR. Pseudonymisation is defined in Art. 4(5) of the GDPR as "the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person".
[36] Recital 26 of the GDPR

- **Balance between utility and privacy** – Ensuring the usefulness of electronic data sources while providing necessary privacy guarantees is considered an important unsolved problem.[37] Anonymisation and sanitisation, which are cornerstones of privacy preservation, change datasets and thus often lead to a loss of information and utility. Finding an adequate balance between the protection of privacy and the utility of data (i.e., the usefulness of data for decision making) is difficult.[38]
- **Failure of anonymisation** – Complete de-identification can usually not be guaranteed.[39] Traditional anonymisation techniques fail in the context of big data applications because there are high numbers of data points for single individuals. The uniqueness of certain characteristics or behaviours is another aspect that makes re-identifying individuals relatively easy. Furthermore, it is usually unknown what other information that may be useful for re-identifying individuals is available to potential malicious actors. Finally, the effective de-identification of certain types of personal data such as genomic data is particularly very difficult.
- **Uniqueness of certain characteristics or behaviours** – An indirect identifier can be any attribute (or set of attributes) that, whilst not structurally unique, is likely to be unique for at least some individuals in a dataset and in the world.[40] The important point is that rare combinations can crop up and create a risk of someone spontaneously re-identifying individuals. Uniqueness – and particularly data uniqueness – does not in itself re-identify anybody. Uniqueness does indicate vulnerability, though.
- **Balance between personalisation and privacy** – Personalisation is well known in online stores and web-based information systems, and is used in a wide range of applications and services. Many big data applications rely on the identifiability of individuals to be able to provide personalised services.[41]

Another challenge that came up during a series of interviews conducted in an earlier phase of the project[42] is the creation of large quantities of substitutions data that may be needed for certain approaches to data sanitisation in the big data era. In the literature, we could not find much evidence referring to this challenge. The use of synthetic data[43] generated with the help of artificial intelligence (AI) may be part of a solution approach here. Other than that, it was noted that it may be a problem in certain situations that the background knowledge that a potential adversary has in unknown.

---

[37] Lalitha Sankar, S. R. Rajagopalan and H. V. Poor, "Utility-Privacy Tradeoffs in Databases: An Information-Theoretic Approach," *IEEE Transactions on Information Forensics and Security* 8, no. 6 (2013)

[38] Arpita Ghosh, Tim Roughgarden and Mukund Sundararajan, "Universally Utility-maximizing Privacy Mechanisms," *SIAM Journal on Computing* 41, no. 6 (2012)

[39] Paul Ohm, "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization," *UCLA Law Review* 57 (2010)

[40] Mark Elliot et al., "The Anonymisation Decision-Making Framework," (University of Manchester, 2016), http://ukanon.net/wp-content/uploads/2015/05/The-Anonymisation-Decision-making-Framework.pdf (accessed January 20, 2019)

[41] Esma Aïmeur, "Personalisation and Privacy Issues in the Age of Exposure," in *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*, 375–6 (ACM, 2018)

[42] See e-SIDES D3.2 - Assessment of existing technologies

[43] Jordi Soria-Comas and Josep Domingo-Ferrer, "Big Data Privacy: Challenges to Privacy Principles and Models," *Data Science and Engineering* 1, no. 1 (2016)

Promising ways to overcome challenges faced in the context of anonymisation and sanitisation include:

- **Differential privacy** – Publishing fully accurate information maximises utility while minimising privacy, while publishing random noise accomplishes the opposite. Privacy can be rigorously quantified using the framework of differential privacy, which requires that a mechanism's output distribution is nearly the same whether a given database row is included.[44] Geng & Viswanath[45], for instance, study noise-adding mechanisms for different types of query functions under a utility-maximisation/cost-minimisation framework.
- **Maximum-knowledge intruder model** – Domingo-Ferrer & Muralidhar[46] propose an approach that makes any assumptions about background knowledge unnecessary. They describe how an intruder can best guess the correspondence between anonymised and original records and how he or she can assess the accuracy of the guess. Further, the authors show how to protect against such a powerful intruder by using anonymisation methods that provide an adequate level of protection.
- **Statistical properties** – Privacy-preserving technologies have tried to ensure that de-identified data, although anonymised, still maintains its statistical properties.[47] This even allows analyses to be performed by third parties keeping privacy risks small.

Duan & Canny[48] argue that the practice of "selling privacy" created a mismatch between what the market offers and what customers need. According to them, customers need privacy and are willing to pay for it. As was discussed at length in D4.1, individuals often act irrationally in the economic sense when facing privacy sensitive decisions and sometimes underestimate their privacy risk. Instead of treating privacy as a product or commodity that is to be sold for a price, Duan & Canny follow the perspective of viewing it as a value and construct schemes that allow an Anonymising Service Provider (ASP) to add it to other existing services. The acceptance of an ASP is not dependent on creating a new "market for privacy". Rather, according to the authors, it rides on the popularity of other, more tangible services. By adding values to other services or goods, an ASP can make them more attractive thus the other business will be willing to participate. Using cryptographic tools, the approach provides provable unlinkablity and anonymity for user transactions even to the vendor who can see the contents of the transactions. With minimal change to existing delivery services, the scheme can anonymise transactions involving delivery of physical items, which require real world identity and physical address. The scheme features a simple architecture involving only the ASP, the vendor and the user. The ASP's role in the scheme is twofold. It performs user authentication and anonymises the communication. The vendor itself is in charge of processing user transactions. The ASP will know users' identities, which is necessary for authentication,

[44] Ghosh, Roughgarden and Sundararajan, "Universally Utility-maximizing Privacy Mechanisms"
[45] Quan Geng and Pramod Viswanath, "Optimal Noise Adding Mechanisms for Approximate Differential Privacy," *IEEE Transactions on Information Theory* 62, no. 2 (2016)
[46] Josep Domingo-Ferrer and Krishnamurty Muralidhar, "New directions in anonymization: Permutation paradigm, verifiability by subjects and intruders, transparency to users," *Information Sciences* 337-338 (2016)
[47] Antorweep Chakravorty, Tomasz Wlodarczyk and Chunming Rong, "Privacy Preserving Data Analytics for Smart Homes" in *Proceedings of the 2013 IEEE Security and Privacy Workshops*
[48] Yitao Duan and John Canny, "From Commodity to Value: A Privacy-Preserving e-Business Architecture," in *Proceeding of the 2006 IEEE International Conference on e-Business Engineering*, ed. Wei-Tek Tsai, 488–95 (Los Alamitos, CA, USA: IEEE, 2006)

Grant Agreement number: 731873

but not their transactions. The vendor knows the transactions, which are the product of service provision, but not their identities. Duan & Canny stress that although the ASP is involved in user transactions, it is not treated as a trusted party. The ASP acts as an intermediate router and cannot gain access to the actual transactions. The harm a malicious ASP can cause is minimised.

3.2. Encryption

Encryption is the encoding of information so that only authorised parties can access it. Examples for cryptographic primitives that are relevant in the context of big data include attribute-based encryption, functional encryption and homomorphic encryption. Encryption is generally strong and fundamental for the protection of data.[49] As cloud storage is increasingly used, encryption is crucial to ensure the confidentiality and integrity of sensitive information.[50] In the context of big data, it is necessary to go beyond the "encrypt all or nothing" model.

With respect to encryption, 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. An approach that is considered to maximise privacy and query expressiveness, at least theoretically, is fully-homomorphic encryption (FHE). A key challenge with respect to FHE is computation cost that makes it relatively slow compared to other methods. Therefore, further research must be performed. If zero quality loss is not the number one requirement, there are other more interesting approaches. Semi-homomorphic encryption, for instance, is mature enough to be integrated into products.

The number of research institutions active in the field of cryptography is very large. Among the European ones that are particularly active in this field are the University of Cambridge, the Catholic University of Leuven, the École Polytechnique Fédérale de Lausanne (EPFL), the University of Louvain and the Ruhr University Bochum. The Israel Institute of Technology (Technion) is also among the most active institutions in the field.[51]

Encryption is mentioned in many of the same provisions of the GDPR as pseudonymisation. For instance, when determining whether another purpose is compatible with the purpose it was initially collected, one factor to take into account is the use of appropriate safeguards, which also includes encryption.[52] Similarly, encryption is mentioned as one of the possible measures to use regarding the requirement that organisations must implement security measures that are appropriate to the risk in Art. 32(1)(a). While Art. 34 requires that persons be notified without undue delay when a "personal data breach is likely to result in a high risk to the rights and freedoms of natural persons", the data controller is exempt from this requirement if appropriate protection measures – such as encryption – were applied to the affected personal data that render them unintelligible.[53] Finally, and more generally, Recital 83 states that "[i]n

[49] D'Acquisto et al., "Privacy by design in big data"
[50] 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)
[51] Based on a Microsoft Academic query using the search term "cryptography"
[52] Art. 6(4)(e) of the GDPR
[53] Art. 34(3)(a) of the GDPR. It should be noted that in any case the controller must notify the relevant supervisory authority within 72 hours after becoming aware of the breach according to Art. 33(1).

order to maintain security and to prevent processing in infringement of this Regulation, the controller or processor should evaluate the risks inherent in the processing and implement measures to mitigate those risks, such as encryption."[54]

Design challenges faced in the context of encryption include:

- **Need for fine-grain sharing policies** – In the context of big data, encryption has to go beyond the *encrypt-all-or-nothing* model. 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. Reducing unnecessary encryption makes great sense for raising efficiency in big data processing.[55] The challenge of fine-grain sharing policies is closely related to the one of fine-grain access control discussed in section 3.5.
- **Computation cost/lack of performance** – The computation of encrypted data still leads to high computation cost. Moreover, it is relatively slow. FHE is considered nowhere near ready for real-life deployment due to serious efficiency impediments.[56] In 2015, FHE computation was 5 to 10 orders of magnitude slower than regular computation.[57] Nevertheless, FHE is ideally suited to protect sensitive data on untrusted servers. Somewhat and partially homomorphic encryption is usually more performant, but comes with limitations in the supported operations.
- **Emergence of quantum computing** – The security of cryptographic functionalities such as public key encryption depends on the difficulty of certain number theoretic problems such as integer factorisation or the discrete log problem over various groups. Shor[58] shows that quantum computing can efficiently solve each of these problems, thereby rendering all public key cryptosystems based on such assumptions impotent. This would seriously compromise the confidentiality and integrity of electronic communications.[59]

Further challenges that came up in an earlier phase of the project[60] are, very generally, that parties exchanging data in an encrypted form need to trust each other and, more specifically, that certain approaches are characterised by high implementation complexity. While the former cannot really be considered a technical design challenge, the later typically lies in the nature of the approach. The fact that

[54] Recital 83 of the GDPR
[55] Changli Zhou, Chunguang Ma and Songtao Yang, "An Improved Fine-Grained Encryption Method for Unstructured Big Data," in *Intelligent computation in big data era: International Conference of Young Computer Scientists, Engineers and Educators, ICYCSEE 2015, Harbin, China, January 10-12, 2015. Proceedings*, vol. 503, ed. Hongzhi Wang, 361–9, Communications in computer and information science 503 (Heidelberg: Springer, 2015)
[56] Wei Wang et al., "Exploring the Feasibility of Fully Homomorphic Encryption," *IEEE Transactions on Computers* 64, no. 3 (2015)
[57] David W. Archer et al., "Maturity and Performance of Programmable Secure Computation," *IEEE Security & Privacy* 14, no. 5 (2016)
[58] Peter W. Shor, "Polynomial-Time Algorithms for Prime Factorization and Discrete Logarithms on a Quantum Computer," *SIAM Review* 41, no. 2 (1999)
[59] Lily Chen et al., "Report on Post-Quantum Cryptography,", Internal Report 8105 (NIST, 2016), https://nvlpubs.nist.gov/nistpubs/ir/2016/NIST.IR.8105.pdf (accessed January 20, 2019)
[60] See e-SIDES D3.2 - Assessment of existing technologies

Grant Agreement number: 731873

a system is possible does not necessarily mean that it is practical.[61] There is limited literature regarding these challenges.

Promising ways to overcome challenges faced in the context of encryption include:

- **Somewhat or partially homomorphic encryption** – Somewhat homomorphic encryption schemes, which support a limited number of homomorphic operations, can be much faster, and more compact than FHE schemes.[62] The same usually holds for partially homomorphic cryptosystems. Moreover, methods such as pre-computations of fixed-basis powers, computing with reduced moduli and parallelisation.[63]
- **Graphics processing unit (GPU) acceleration** – Apart from algorithmic optimisations, the use of GPUs can accelerate computation of encrypted data[64] as well as the encryption and decryption of data. With continuous architectural improvements in recent years, GPUs have evolved into a massively parallel, multithreaded, many-core processor system with tremendous computational power.

According to Sharma et al.[65], there are two well-known generic approaches for privacy-preserving computation on untrusted platforms: FHE and Yao's *garbled circuit* (GC) together with *oblivious transfer* (OT). FHE allows an arbitrary number of additions and multiplications on encrypted data without decryption. GC provides basic logic gates such as AND, XOR, and OR, with which more complex circuits can be implemented. Conceptually, these techniques can be used to construct any data-mining algorithm. However, the schemes are, according to Sharma et al., very expensive in storage, communication, and computation. The ciphertext resulting from FHE schemes becomes several magnitudes larger than plaintext data, placing an excessive burden on storage and communication. Likewise, even the most optimised GC implementation incurs impractical communication costs between the participating parties. Moreover, building and implementing complex algorithms with GCs is a challenging task. Another example of a GC is provided in section 3.4.

3.3. Deletion

Deletion describes the permanent erasure of data from a physical medium. Choosing the right layer for secure deletion is a trade-off between abstracted access (user layer) and richer information (physical layer). There are user-level approaches as well as file-system focused approaches to deletion.[66] With

[61] Liam Morris, "Analysis of Partially and Fully Homomorphic Encryption," (Rochester Institute of Technology, 2013), http://gauss.ececs.uc.edu/Courses/c6056/pdf/homo-outline.pdf (accessed January 20, 2019)
[62] Michael Naehrig, Kristin Lauter and Vinod Vaikuntanathan, "Can homomorphic encryption be practical?," in *Proceedings of the 3rd ACM Workshop on Cloud Computing Security*, 113–24 (ACM, 2011)
[63] Christine Jost et al., "Encryption Performance Improvements of the Paillier Cryptosystem,", IACR Cryptology ePrint Archive (IACR, 2015), https://www.iacr.org/cryptodb/data/paper.php?pubkey=26497 (accessed January 20, 2019)
[64] Wang et al., "Exploring the Feasibility of Fully Homomorphic Encryption"
[65] Sagar Sharma, Keke Chen and Amit Sheth, "Toward Practical Privacy-Preserving Analytics for IoT and Cloud-Based Healthcare Systems," *IEEE internet computing* 22, no. 2 (2018)
[66] Joel Reardon, David Basin and Srdjan Capkun, "On Secure Data Deletion," *IEEE Security & Privacy* 12, no. 3 (2014)

Grant Agreement number: 731873

respect to the latter class, proposals have been made that address file systems that are frequently used in the context of big data applications.

In Europe, the ETH Zurich has been one of the most active research institutions in the field of secure deletion.[67]

Under Art. 17 of the GDPR, data subjects have the right to erasure of their personal data where:

- the data is no longer necessary for the purposes for which they were collected or processed,
- the data subject withdraws consent and where there are no other legal grounds for processing,
- the data subject objects to the processing where there are no overriding legitimate grounds for the processing or where the data is used for marketing purposes, or
- the data was processed illegally.[68]

In addition to the rights of the data subject to have their personal data erased, the storage limitation principle of Art. 5(1)(e) of the GDPR holds that personal data must not be held longer than is necessary for the purposes for which they are processed.[69] This principle is related to the privacy by design principle of data minimisation,[70] which entails that personal data are "adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed".[71] Recital 39 states that this in particular requires "ensuring that the period for which the personal data are stored is limited to a strict minimum" and that "time limits should be established by the controller for erasure or for a periodic review". Personal data that are inaccurate should be rectified or deleted.[72]

Further, the competent supervisory authority must set binding corporate rules that specify, among other measures, the application of privacy by design principles including limited storage periods and data minimisation.[73] Finally, where processing is conducted by a processor on behalf of a controller it shall be governed by a contract that stipulates that the personal data must be deleted or returned to the controller at the end of the provision of processing services, unless they are required to retain the data by law.[74]

The key design challenge in the context of secure deletion is that it can only be ensured on the physical layer that a data object is truly irrecoverable. At the same time, however, the data object can only be identified easily on the user layer. Choosing the right layer for secure deletion is a **trade-off between abstracted access and richer information**. Consequently, secure deletion is particularly difficult in distributed cloud storage.

---

[67] Based on a Microsoft Academic query using the search term "secure deletion"
[68] This list is not comprehensive. Recitals 59, 65 and 66 of the GDPR give more context to this right.
[69] The exception to this rule is if it the data are being kept for public interest archiving, scientific or historical research, or statistical purposes. Where data is held for one of these reasons, appropriate safeguards must be utilised so that the principle of data minimisation is respected. See Art. 89(1) and Recital 156 of the GDPR. Where processing is carried out in the public interest or in the exercise of official authority, storage periods may be specified by law. See Recital 45 of the GDPR.
[70] Art. 25 of the GDPR
[71] Art. 5(1)(c) of the GDPR
[72] Recital 39 of the GDPR
[73] Art. 47(1) and (2)(d) of the GDPR
[74] Art. 28(3)(g) and Recital 81 of the GDPR. The choice of deleting or returning the data is the controller's.

Promising ways to overcome challenges faced in the context of secure deletion include:

- **Application of blockchains** – When data is stored in the cloud, data deletion becomes a security challenge because the cloud server may not delete the data. In most cases, data owners cannot verify if their data has actually been deleted. Yang et al.[75] propose a blockchain-based data deletion scheme, which can make the deletion operation more transparent. In this scheme, the data owner can verify the deletion result no matter how malevolently the cloud server behaves. Moreover, with the application of blockchain, the proposed scheme can achieve public verification without any trusted third party.
- **Provable data deletion** – The correctness and availability of data becomes a primary concern of data owners when data is transferred from one cloud to another. To address this issue, Xue et al.[76] propose a provable data transfer protocol based on provable data possession and deletion for secure cloud storage. The data owner can transfer the outsourced data from one cloud provider to another without worrying about the deletion of the removed data by the previous cloud provider. The scheme can generate the deletion evidence on the cloud data regardless of whether the data is encrypted or not.

3.4. Multi-party computation

Multi-party computation (MPC) relies on the distribution of data and processing tasks over multiple parties. MPC is a field of cryptography that aims to allow securely computing the result of functions without revealing the input data.

Secure MPC is less secure than FHE, especially when many untrusted parties are involved but more efficient in certain types of implementations.[77] MPC has a long history, but practical implementation problems stand in the way of wide adoption. Research can help by addressing the problems, for instance, by providing algorithms that speed up common analytics methods. MPC can unlock new possibilities in the context of joint data processing in domains where legal and procedural hurdles are prevalent. While anonymisation allows for the use of standard analytical tools, the value of the data may be decreased in relation to the scope of the anonymisation process. MPC enables working with the data as it is, but it restricts the efficiency of the analysis and the range of tools that can be used.

MPC currently receives a lot of attention in research. For instance, the EU-funded projects SODA and SPECIAL pay particular attention to this class of privacy-preserving technologies. In Europe, the Aarhus University has been particularly active in the research field.[78] With the Technion, the Bar-Ilan University, the Hebrew University of Jerusalem, the Tel Aviv University and the University of Haifa, there are several institutions from Israel among the most active institutions in the field. Further relevant institutions from Europe include the ETH Zurich, the Darmstadt University of Technology and the EPFL. The Estonian

[75] Changsong Yang, Xiaofeng Chen and Yang Xiang, "Blockchain-based publicly verifiable data deletion scheme for cloud storage," *Journal of Network and Computer Applications* 103 (2018)
[76] Liang Xue, Ni, Jianbing, Li, Yannan and Jian Shen, "Provable data transfer from provable data possession and deletion in cloud storage," *Computer Standards & Interfaces* 54, no. 1 (2017)
[77] Danezis et al., "Privacy and Data Protection by Design - from policy to engineering"
[78] Based on a Microsoft Academic query using the search term "secure multi party computation"

company Cybernetica offers a promising solution called Sharemind MPC[79]. The Danish company Partisia[80] work on solutions based on MPC.

While the GDPR does not specifically mention MPC, as a privacy-preserving technology it embodies many of the principles the GDPR was created to uphold. For instance, privacy by design principles, such as data minimisation and limitations on accessibility,[81] are embedded in MPC, and its use can aid in protecting the fundamental right to data protection.[82] Further, MPC can also provide a better framing of the controllership responsibilities,[83] as it can help in distinguishing between controllers and processors and therefore also in determining their respective obligations under the GDPR.[84]

The key design challenge faced in the context of MPC include is **lack of performance**. Although it has been greatly improved over time[85], the efficiency of analyses in MPC contexts is still rather low. Just as FHE, MPC is a technology for programmable secure computing.[86] GC protocols are mostly used when two parties want to compute a function over their inputs. Protocols based on linear secret sharing (LSS) are usually deployed when more than two parties collaborate. FHE implementations tend to be slower than MPC implementations.

Another challenge that came up during a series of interviews conducted in an earlier phase of the project[87] is that setting up working solutions is difficult. Although the number of MPC implementations is increasing, it is still quite low. We could not find much evidence referring to this challenge in literature.

With respect to promising ways to overcome challenges faced in the context of MPC, it is important to note that while the literature generally focuses on protocols that enable the highest number of parties with the highest security, applications in practice often make use of a smaller numbers of parties and can often rely on weaker forms of security.[88] This does not necessarily imply the overall system is less secure, it is more a reflection of the trust relationship that the parties have.

GC is a cryptographic technique that enables the computation of a function where the inputs are held by two different parties in a completely secure manner. It is a general technique that can be used to compute any function representable in the form of a circuit. Although restricted to two parties, it supports a rich

[79] https://sharemind.cyber.ee/
[80] https://partisia.com/
[81] Art. 25(1) and (2) of the GDPR
[82] Recital 1 and 2 of the GDPR, = Art. 8(1) of the Charter of Fundamental Rights of the European Union, and Art. 16(1) of the Treaty on the Functioning of the European Union (TFEU).
[83] e-SIDES D3.2 section 3.3
[84] See Art. 4, 5, 24 and 28 of the GDPR for more details regarding the differences between controllers and processors and their respective obligations therein.
[85] Chuan Zhao et al., "Secure Multi-Party Computation: Theory, practice and applications," *Information Sciences* 476 (2019)
[86] Archer et al., "Maturity and Performance of Programmable Secure Computation"
[87] See e-SIDES D3.2 - Assessment of existing technologies
[88] David W. Archer et al., "From Keys to Databases—Real-World Applications of Secure Multi-Party Computation," *The Computer Journal* 9 (2018)

Grant Agreement number: 731873

class of operations. Chen et al.[89] state that existing record linkage solutions assume that data are centralised with no privacy or security concerns restricting sharing. However, according to the authors, that is often untrue. Therefore, they designed and implemented a portable method for privacy-preserving record linkage based on GCs to accurately and securely match records. Moreover, they developed an approximate matching mechanism that significantly improves efficiency. Chen et al. propose a completely secure deterministic record linkage method for two parties. The method guarantees that no extra information is revealed to the two parties beyond the results of the matching, and no information is revealed to any other party. The authors provide a basic approach based on a classical implementation of GCs and a computationally more efficient approach using a filtering strategy. Unlike methods requiring honest-broker intermediaries, the approach does not require an external third party. Chen et al. provide a proof of concept for secure two-party record linkage using state-of-the-art GC methods that can be readily adopted in small to medium-scale linkage tasks with a strong security guarantee.

3.5. Access control

Access control describes the selective restriction of access to resources. Attribute-based access control (ABAC) refers to a set of approaches that support fine-grained access control policies based on attributes that are evaluated at run-time. Big data applications typically require fine-grained access control.

One of the experts we interviewed in an earlier phase of the project[90] 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 and dynamic way. The expert's statement makes the requirements clear that have to be met in the big data context. With respect to granularity, according to the expert, 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 the user is performing. The expert stated that a case management system used to do policing work, for instance, grants access not only based on the user's general status but also taking the severity of the crime into account. This allows restricting access to privacy-invasive datasets in case of minor crimes.

Access control technologies have a long history. The University of Milan is one of the most active European institutions doing research in the field of access control. [91] Apart from the University of Milan, Royal Holloway, a constituent college of the University of London, and the University of Toulouse have been active in the more specific field of ABAC.[92]

There are a number of legal provisions in the GDPR that address access control. Art. 25(2) on privacy by design states that the appropriate technical and organisational measures that must be implemented by the controller includes limiting the accessibility of data.[93] Further, under Art. 32(4), the GDPR states that

[89] Feng Chen et al., "Perfectly Secure and Efficient Two-Party Electronic-Health-Record Linkage," *IEEE internet computing* 22, no. 2 (2018)
[90] See e-SIDES D3.2 - Assessment of existing technologies
[91] Based on a Microsoft Academic query using the search term "access control"
[92] Based on a Microsoft Academic query using the search term "attribute based access control"
[93] Art. 25(2) of the GDPR states, "The controller shall implement appropriate technical and organisational measures for ensuring that, by default, only personal data which are necessary for each specific purpose of the

Grant Agreement number: 731873

"[t]he controller and processor shall take steps to ensure that any natural person acting under the authority of the controller or the processor who has access to personal data does not process them except on instructions from the controller, unless he or she is required to do so by Union or Member State law." In addition to organisational policy enforcement, this provision can be upheld through technological measures of access control. It should also be noted that notifications of data breaches to *data subjects* are not necessary if the controller has implemented appropriate technical and organisational protection measures, including access control, and those measures were applied to the data affected by the breach.[94]

In addition to operators of big data solutions, the GDPR also addresses data subjects. Art. 15(1) grants individuals the right to obtain a copy of their personal data as well as other related information, such as the purposes of the processing, the categories of the processing, and the recipients or categories of recipient to whom the personal data have been or will be disclosed.[95] According to Recital 63, "[w]here possible, the controller should be able to provide remote access to a secure system which would provide the data subject with direct access to his or her personal data", and all reasonable measures should be used to verify the identity of a data subject who requests access.[96] In addition to the right of access by data subjects, the related Art. 20 right to data portability – which states that individuals have the right to receive their personal data from a controller in a structured, commonly used and machine-readable format – may pose challenges in the design and implementation of access control technologies.

The fact that challenges of introducing access control technologies are intertwined with different legal regimes is also demonstrated by the different legal liability requirements that are applicable, for instance, within a commercial or citizen-government relationship context. Whereas tasks for securing citizens identities on commercial platforms are usually taken up by the commercial parties themselves, tasks for securing the digital identities of citizens are often outsourced to private companies by government authorities. Private companies in their position of providing privacy-preserving technology solutions in the form of access control are often fulfilling a crucial role as a big data technology stakeholder in protecting access to digital identity profiles. Depending on whether the given access control technology is used by public authorities or commercial parties, different liability regimes would apply. For instance, if a digital identity secured by access control belongs to a (non-law enforcement) public authority, beyond the relevant GDPR requirements that are applicable for all data processing parties (such as Art. 34 regarding data breach notifications), the given government has a positive obligation towards protecting their citizens under human rights law. This obligation compounds the liability of governments for securing the personal data of their citizens and should encourage them to tackle the challenges relating to the implementation and usage of access control solutions. Also, beyond the applicable GDPR requirements,

processing are processed. That obligation applies to the amount of personal data collected, the extent of their processing, the period of their storage and their accessibility. In particular, such measures shall ensure that by default personal data are not made accessible without the individual's intervention to an indefinite number of natural persons."

[94] Art. 34(3)(a) of the GDPR. Notification to the relevant data protection authority would still be necessary, however.

[95] Under Art. 15(3), additional copies may be obtained by the data subject, and the controller may charge a reasonable fee based on administrative costs.

[96] Recital 64 of the GDPR

commercial parties can be held liable for breaching any digital secrecy or confidentiality obligations outlined in business contracts.

Design challenges faced in the context of access control include:

- **Fine-grain and dynamic access control** – Big data applications typically require fine-grained and dynamic access control. Qualifiers of identity, groups and roles are often insufficient to express real-world access control policies.[97] Yan et al.[98] state that there is a need for a flexible, efficient and scalable access control system.
- **Manageability** – Access control becomes increasingly difficult if a higher number of devices is used. Traditional approaches, such as role-based access control and user-based access control, are becoming less and less manageable.[99] Practitioners have noted that this AC approach is often cumbersome to manage given the need to associate capabilities directly to users or their roles or groups.[100]
- **Different domains of trust** – Traditional access control architectures usually assume the data owner and the servers storing the data are in the same trusted domain.[101] A number of cloud data access control schemes have been proposed. However, existing solutions suffer from high computation complexity and cost and therefore few of them have been effectively deployed in practice.[102]

Another challenge that came up in an earlier phase of the project[103] is that it is usually one person that designs the access control system. Indeed, there often is a single point of failure but this is rather an organisational than a technical design challenge. In the literature, we could not find much evidence referring to this challenge. However, a single-point bottleneck was described with respect to the maintenance of the whole attribute set by only one authority in the context of earlier attribute-based encryption schemes.

Promising ways to overcome challenges faced in the context of access control include:

- **Attribute-based access control** – ABAC in an example for a set of approaches that can conceptually support fine-grain and dynamic access control policies based on attributes that are evaluated at run-time. ABAC is a flexible approach that can implement access control policies limited only by the computational language and the richness of the available attributes. This

---

[97] Vincent C. Hu, D. R. Kuhn and David F. Ferraiolo, "Attribute-Based Access Control," *Computer* 48, no. 2 (2015)
[98] Zheng Yan et al., "Flexible Data Access Control Based on Trust and Reputation in Cloud Computing," *IEEE Transactions on Cloud Computing* 5, no. 3 (2017)
[99] S. S. Manikandasaran and S. Sudha, "Data Access Control Techniques and Security Challenges in Cloud Computing: A Survey," *International Journal of Computer Sciences and Engineering* 6, Special Issue 2 (2018)
[100] Hu, Kuhn and Ferraiolo, "Attribute-Based Access Control"
[101] Shucheng Yu et al., "Achieving Secure, Scalable, and Fine-grained Data Access Control in Cloud Computing," in *Proceedings of the 2010 IEEE INFOCOM*, 1–9 (IEEE, 2010)
[102] Huaqing Lin, Zheng Yan and Raimo Kantola, "CDController: A Cloud Data Access Control System Based on Reputation," in *Proceedings of the 2017 IEEE International Conference on Computer and Information Technology*, 223–30 (Helsinki, Finland: IEEE, 2017)
[103] See e-SIDES D3.2 - Assessment of existing technologies

flexibility enables the greatest breadth of subjects to access the greatest breadth of objects without specifying individual relationships between each subject and each object.[104]

- **Hierarchical structures of users** – In order to simultaneously achieve fine-grain access control, scalability and to provide cloud users shared access privileges and flexibility on delegation of their access privileges, Ahuja & Mohanty[105] propose a scalable ABAC scheme for cloud storage. The scheme extends the ciphertext policy attribute-based encryption to achieve flexible delegation of access privileges and shared access privileges along with scalability and fine-grained access control. The scheme achieves scalability by employing hierarchical structure of users.
- **Multi-authority scheme** – Earlier attribute-based encryption schemes involve only one authority to maintain the whole attribute set, which can bring a single-point bottleneck on both security and performance. Li et al.[106] propose a threshold multi-authority ciphertext policy attribute-based encryption access control scheme for public cloud storage in which multiple authorities jointly manage a uniform attribute set. Taking advantage of (*t*, *n*) threshold secret sharing, the master key can be shared among multiple authorities.
- **Decentralised runtime monitoring** – In federated clouds, distributed components receive, exchange and process access requests and their corresponding access decisions with the possibility of being compromised. In order to promote accountability and reliability of a distributed access control system, Ferdous et al.[107] present a decentralised runtime monitoring architecture that is based on blockchain technology.
- **Reputation evaluation** – Reputation is a collective opinion on the trustworthiness of an entity. It is a measure of public trust put into an entity. Lin et al.[108], for instance, describe a cloud data access control system based on reputation. Trust management and reputation evaluation is integrated into a cryptographic system. The authors claim that their system allows to flexible control data access in an efficient way.

Privacy of genomic data is becoming an increasingly serious concern. However, according to Hunag et al.[109], there are no standard data storage solutions that enable compression, encryption and selective retrieval. Therefore, the authors present a privacy-preserving solution named SECRAM (Selective retrieval on Encrypted and Compressed Reference–oriented Alignment Map) for the secure storage of compressed aligned genomic data. The solution minimises information leakage, stores the sequence data in a lossless compressed format, enables selective retrieval of encrypted data and improves the efficiency of downstream analysis (e.g., variant calling). According to Hunag et al., SECRAM is a data storage format that is organised in position-based storage that enables random queries anywhere in the genome, highly

[104] Hu, Kuhn and Ferraiolo, "Attribute-Based Access Control"
[105] Rohit Ahuja and Sraban K. Mohanty, "A Scalable Attribute-Based Access Control Scheme with Flexible Delegation cum Sharing of Access Privileges for Cloud Storage," *IEEE Transactions on Cloud Computing* (2017)
[106] Wei Li et al., "TMACS: A Robust and Verifiable Threshold Multi-Authority Access Control System in Public Cloud Storage," *IEEE Transactions on Parallel and Distributed Systems* 27, no. 5 (2016)
[107] Md S. Ferdous et al., "Decentralised Runtime Monitoring for Access Control Systems in Cloud Federations," in *Proceedings of the IEEE 37th International Conference on Distributed Computing Systems*, 2632–3 (IEEE, 2017)
[108] Huaqing Lin, Zheng Yan and Raimo Kantola, "CDController" in *Proceedings of the 2017 IEEE International Conference on Computer and Information Technology*
[109] Zhicong Huang et al., "A privacy-preserving solution for compressed storage and selective retrieval of genomic data," *Genome research* 26, no. 12 (2016)

compressed through a combination of reference-based and general data compression techniques, and encrypted with standard secure cryptographic techniques and a fine-grained privacy control mechanism. By applying the solution, sequence data are securely protected in storage and can be efficiently retrieved for downstream analysis without any access to unauthorised information.

3.6. Policy enforcement

Policy enforcement focuses on the enforcement of rules for the use and handling of resources. Data expiration policies, for instance, are already enforced by some big data solutions.

Some classes of technologies including those focusing on access control and policy enforcement are threatened by a single point of failure and limited transparency. Usually, there is one person that designs the access control system and one that specifies the policies. Opacity is often intensified by trade secrets or practices that are not completely open. In such cases, non-technical measures such as specific processes complementing or replacing technologies become relevant, albeit sometimes cumbersome. For data protection authorities, it is particularly hard to exercise their enforcement power if foreign actors are involved. Policy enforcement technologies play a key role in the context of big data as chains of responsibility become longer and more geographically dispersed. Enforcement frameworks must be flexible and able to support different data processing requirements.[110] Automated policy enforcement mechanisms are important in the big data era as policies get lost easily when data is transferred between different systems. Technologies for policy enforcement are not yet mature.

In Europe, the University of Cambridge, the VU University of Amsterdam and the Imperial College London are particularly active in terms of research in the field of policy enforcement.[111] The European telecommunications companies Ericsson and the now-defunct Alcatel-Lucent have also been very active in research on policy enforcement.

In light of the extraterritorial effects of the GDPR, the prescriptions of the GDPR would even apply to geographically dispersed systems beyond the borders of the EU so long as these systems process the data of EU residents. With respect to automated policy enforcement mechanisms, and given those solutions that stick policies to personal data, different rights and principles could be triggered. If a policy, for instance, relates to the expiration of data then the right to erasure (Art. 17) of the GDPR would be applicable along with the principles enshrined in Art. 5, such as the principle related to data accuracy.[112] Given that policy enforcement mechanisms are used within dispersed chains, policies that are stuck to personal data are transferred from system to system. This requires that data subjects rely upon their right to obtain information regarding "the recipients or categories of recipient to whom the personal data have or will be disclosed".[113] Furthermore, within dispersed systems, the Article 29 Working Party Guidelines

[110] 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)
[111] Based on a Microsoft Academic query using the search term "policy enforcement"
[112] Art. 5(d) of the GDPR
[113] Art. 15(c) of the GDPR

Grant Agreement number: 731873

on consent under Regulation 2016/679[114] regarding joint controllership should also be considered. The Article 29 Working Party emphasised that if consent was used within processes where multiple data controllers are jointly involved, all of these controllers should be named for the data subject. Hence this would also apply for the exchange of policy enforcement mechanisms that rely upon distributed systems to exchange sticky policies.

The key design challenge in the context of policy enforcement is **need for a flexible enforcement framework**. Systems are increasingly geographically dispersed, and chains of responsibilities become longer. While much cloud security research focuses on enforcing standard access control policies typical of centralised systems, such policies often prove inadequate for the highly distributed, heterogeneous, data-diverse and dynamic computing environment of clouds.[115] Current approaches focus mostly on simple access control policies instead of complex organisation-wide policies that also include usage control.[116]

Another challenge that came up during a series of interviews conducted in an earlier phase of the project[117] is the need for automation. Automated policy enforcement mechanisms were considered particularly important in the big data era as policies get easily lost or neglected in the course of data being transferred between different systems. Researchers suggested to introduce sticky policies, where policies would attach to the data by using, for instance, digital signatures[118], but challenges related to such solutions can trigger different data protection prescriptions.

Promising ways to overcome challenges faced in the context of policy enforcement include:

- **Distributed policy enforcement** – Betge-Brezetz et al.[119] describe an approach of end-to-end privacy policy enforcement over the cloud infrastructure that is based on the sticky policy paradigm. Data protection is performed within the cloud nodes and is completely transparent for the applications. According to the authors, this approach offers a flexible and transparent way to

[114] Article 29 Working Party Guidelines on consent under Regulation 2016/679 (17/EN WP259 rev.01) https://www.datenschutz-praxis.de/wp-content/uploads/2018/06/20180416_Article29WPGuidelinesonConsent_publishpdf.pdf

[115] Kevin W. Hamlen, Lalana Kagal and Murat Kantarcioglu, "Policy Enforcement Framework for Cloud Data Management," *Bulletin of the Technical Committee on Data Engineering* 35, no. 4 (2012)

[116] Gabriela Gheorghe, Stephan Neuhaus and Bruno Crispo, "xESB: An Enterprise Service Bus for Access and Usage Control Policy Enforcement," in *Trust Management IV: 4th IFIP WG 11.11 International Conference, IFIPTM 2010, Morioka, Japan, June 16-18, 2010; Proceedings*, ed. Masakatsu Nishigaki, 63–78, IFIP Advances in Information and Communication Technology 321 (Berlin: Springer, 2010)

[117] See e-SIDES D3.2 - Assessment of existing technologies

[118] Kaniz Fatema, David W. Chadwick and Stijn Lievens, "A Multi-privacy Policy Enforcement System," in *Privacy and Identity Management for Life: 6th IFIP WG 9.2, 9.6/11.7, 11.4, 11.6/PrimeLife International Summer School, Helsingborg, Sweden, August 2-6, 2010, Revised Selected Papers*, ed. Simone Fischer-Hübner et al., 297–310, IFIP Advances in Information and Communication Technology 352 (Berlin, Heidelberg: IFIP International Federation for Information Processing, 2011)

[119] Stephane Betge-Brezetz et al., "End-to-end privacy policy enforcement in cloud infrastructure," in *Proceedings of the IEEE 2nd International Conference on Cloud Networking*, 25–32 (IEEE, 2013)

enforce various privacy constraints within the cloud infrastructure. Fatema et al.[120] propose a multi-policy authorisation infrastructure that provides privacy protection of personal data using multiple policy decision points, sticky policies and obligations.

- **Combination of mechanisms** – Policy management and enforcement is particularly difficult in the big data context. The performance of solutions is easily affected. Nevertheless, privacy policy needs to be integrated, flexible, context-aware and customisable. While most studies concentrate on a single mechanism, Al-Shomrani et al.[121] focus on applying more than one mechanism to secure user privacy. Gheorghe et al. [122] propose to implement flexible, instrumentable and highly configurable policy enforcement mechanisms. The combination of mechanisms allows not only to monitor and enforce preventive and reactive policies, both for access control and usage control policies, but also to do this both inside one domain and between domains.

3.7. Accountability and transparency

Accountability requires the evaluation of compliance with policies and the provision of evidence. A cornerstone of accountability in the context of big data is the provision of automated and scalable control, as well as auditing processes that can evaluate the level of compliance. Transparency calls for the explication of information collection and processing. In the age of big data, transparency is achieved by multichannel and layered approaches as well as standardised icons. Transparency is considered critical to allow data subjects informed choices.

Accountability and transparency technologies are highly 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 particular groups.[123] Therefore, a measure for fairness and a way to ensure that the learning is fair are needed. Moreover, explanations of certain decisions may be needed. For example, if a classifier is used to decide whether somebody receives a loan or not, it would be good to have an explanation of the decision. In the big data context, it is extremely difficult to explain what algorithms do with data or to get a preview of what the outcome of providing data may be.[124] Machine learning modules typically are “black boxes”. Similarly, if an Internet of Things (IoT) device performs in an undesired way, it is often almost impossible to find out why; the performance is not interpretable. There is a growing field of research that aims to explain why certain decisions are made. A key problem is that most IoT devices

[120] Kaniz Fatema, David W. Chadwick and Stijn Lievens, “A Multi-privacy Policy Enforcement System” in *Privacy and Identity Management for Life*

[121] Abdullah Al-Shomrani, Fathy Fathy and Kamal Jambi, “Policy enforcement for big data security,” in *Proceedings of the 2nd International Conference on Anti-Cyber Crimes*, 70–4 (IEEE, 2017)

[122] Gabriela Gheorghe, Stephan Neuhaus and Bruno Crispo, “xESB: An Enterprise Service Bus for Access and Usage Control Policy Enforcement” in *Trust Management IV*

[123] Toon Calders and Indrė Žliobaitė, “Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures,” in *Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases*, ed. Bart Custers et al., 43–57, Studies in applied philosophy, epistemology and rational ethics 3 (Berlin, Heidelberg: Springer, 2013)

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

do not have a screen, which makes it difficult for them to visually explain what they are doing. Accountability and transparency technologies are not yet mature.

In Europe, the Max Planck Society, SINTEF, the University of Stavanger and the University of Oxford have been particularly active in accountability research.[125] With respect to transparency-related research, apart from the Max Planck Society and the University of Oxford, the Catholic University Leuven is among the most active institutions.[126]

With respect to accountability and transparency, one of the main design challenges relates to the fact that definitions for accountability and transparency among stakeholders within the big data value chain differ significantly. For instance, for business managers and computer scientists, accountability is often viewed as a requirement to evaluate the compliance with company policies and provisions of evidence for good conduct. Furthermore, transparency is often regarded by business managers as a goal that should be achieved and built through increasingly personalised profiles about customers in order to better target them. From this perspective, both computer scientists and business managers often have a business-oriented view regarding accountability and transparency. Accountability is often translated into taking out insurance to protect against potential liabilities, for instance in cases of data breaches, and transparency is often regarded as being desirable in order to better learn customer preferences.

Within the field of data protection law, and specifically the GDPR, accountability and transparency are main principles underlying the foundations of the regulation. As mentioned in section 2, the obligation of data controllers to demonstrate compliance with all the provisions of the GDPR, is currently perhaps the strongest accountability instrument introduced internationally in data protection law. Transparency is specifically mentioned under Art. 5 as key requirement, and throughout the text of the GDPR, transparency is woven into multiple prescriptions. These are all aimed at ensuring that data subjects can understand the technical language related to the processing and storage of their personal data. The right to information, for instance, can be regarded as a specific new incentive to ensure data controllers are and remain transparent about their processing of personal data. Accountability coupled with transparency under the GDPR were aimed at returning some power to the data subject versus big data companies as big data companies increasingly cultivate an asymmetrical power relationship towards data subjects. The GDPR empowers data subjects with new rights in order to place checks upon data controllers and data processors and to be better equipped with holding them accountable for what happens to data subjects' data. Therefore, the definition of accountability in the GDPR stretches beyond simple requirements for data controllers and processors to live up to data protection principles; it actually requires companies to demonstrate legal compliance with the new prescriptions. Some of the rights and prescriptions that are specifically relevant for improving transparency and accountability are, for instance, broadened conditions for consent (Art. 7), the right of access by the data subject (Art. 15), the right to erasure (Art. 17), the right not to be subject to profiling (Art. 22), security of processing (Art. 32), notification of a personal data breach (Art. 33) and of course privacy by design and privacy by default (Art. 25).

The fact is that according to the GDPR, computer scientists and business managers – if they are data controllers or processors – should all demonstrate compliance with the GDPR, including its principles of

[125] Based on a Microsoft Academic query using the search terms "accountability" and "computer science"
[126] Based on a Microsoft Academic query using the search terms "transparency" and "computer science"

accountability and transparency – can decrease the existing differences in definitions and conceptions and by this decrease these specific design challenges.

Design challenges faced in the context of accountability and transparency include:

- **Limitations of transparency** – Engineers have developed deep learning systems that work without necessarily knowing why they work or being able to show the logic behind a system's decision.[127] This can cause serious difficulties, such as when Google's Photo app unexpectedly tagged black people gorillas.[128] While the presumption was that the image recognition system hadn't been trained on enough black faces, the engineers of these systems could not precisely say why the problems were occurring, even though they had total access to the systems' designs and implementations.
- **Algorithmic accountability** – While algorithmic decision-making can offer benefits in terms of speed, efficiency, and even fairness, bias is routinely introduced into software systems in many ways, including the use of biased training data. On the one hand, transparency alone cannot create accountable systems. Ananny & Crawford[129] argue that making one part of an algorithmic system visible – such as the algorithm or the underlying data – is not the same as holding the assemblage accountable. Holding an assemblage accountable requires not just seeing inside any one component of an assemblage but understanding how it works as a system. On the other hand, there are several objections to full transparency.[130]
- **Integrity verification** – With the proliferation of cloud computing and the increasing needs in big data analytics, the verification of data integrity becomes increasingly important, especially on outsourced data.[131] Outsourced data and computation outcomes are not continuously trustworthy due to the lack of control and physical possession of the data owners.[132] A cornerstone of accountability in the context of big data applications thus is the provision of automated and scalable control and auditing processes that can evaluate the level of compliance with policies.

Identifying promising ways to overcome challenges faced in the context of accountability and transparency requires not only technological development but also clarity about what accountability actually is and how it is related to transparency and other concepts such as responsiveness, remediability, responsibility, verifiability, appropriateness and effectiveness. According to Jaatun et al.[133], an

[127] Mike Ananny and Kate Crawford, "Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability," *New Media & Society* 20, no. 3 (2017)
[128] Conor Dougherty, "Google Photos Mistakenly Labels Black People 'Gorillas'," http://bits.blogs.nytimes.com/2015/07/01/google-photos-mistakenly- (accessed January 21, 2019)
[129] Ananny and Crawford, "Seeing without knowing"
[130] Paul B. de Laat, "Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability?," *Philosophy & Technology* 31, no. 4 (2018)
[131] Chang Liu et al., "Public Auditing for Big Data Storage in Cloud Computing -- A Survey," in *Proccedings of the IEEE 16th International Conference on Computational Science and Engineering*, 1128–35 (IEEE, 2013)
[132] Mehdi Sookhak et al., "Towards dynamic remote data auditing in computational clouds," *The Scientific World Journal* (2014)
[133] Martin G. Jaatun et al., "Towards Strong Accountability for Cloud Service Providers," in *Proceedings of the IEEE 6th International Conference on Cloud Computing Technology and Science*, 1001–6 (IEEE, 2014)

accountable organisation must not only define, monitor and correct its data practices but also implement preventive, detective and corrective mechanisms. Accountability mechanisms may be technical tools. Machine learning models are particularly opaque, non-intuitive and difficult for people to understand. Algorithms often only confront data subjects with an automated decision without revealing the data and data processing method upon which that decision was made. Explainable AI will be essential to make sure that users understand, trust and effectively manage AI and in particular machine learning tools. Deep explanation, interpretable models and model induction are particularly promising approaches to increase explainability without sacrificing learning performance.[134]

3.8. Data provenance

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. Big data poses challenges for data provenance.[135] Problems are caused by the strong heterogeneity of the data. Additionally, the use of many analytics and storage solutions may result in prohibitively large amounts of provenance information. One of the experts we interviewed in an earlier phase of the project[136] 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 a unified model. The expert's statement makes the requirements clear that have to be met in the big data context. Each property that is related to an object, according to the expert, can come from a different source. The interviewee stressed that the company puts a lot of development effort into notions of data provenance and underlined the tight connection between data provenance, accountability and transparency. The measures taken by the company with focus on data provenance allow users to investigate 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. Provenance technologies are quite mature.

European institutions that are particularly active in the research field the University of Southampton, the Humboldt University of Berlin, the University of Oxford and the King's College London. Further institutions that have been active in the research field data provenance are the Hungarian Academy of Sciences and the German Aerospace Center.[137]

Art. 5(d) of the GDPR clearly sets out the obligation to strive for the highest data quality possible through the accuracy principle. In this sense, data provenance solutions, if implemented correctly, can ensure that the inputted data was high quality. Yet, the above-mentioned challenges of computer scientists can result in the degradation of data quality and also result in legal challenges that are related to maintaining data accuracy. For instance, the expiration of provenance data in highly heterogeneous systems could result in legal non-compliance if expired data remains used in different analytic and storage solutions.

---

[134] David Gunning, "Explainable Artificial Intelligence," (DARPA, 2017), https://www.darpa.mil/attachments/XAIProgramUpdate.pdf (accessed January 21, 2019)

[135] 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)

[136] See e-SIDES D3.2 - Assessment of existing technologies

[137] Based on a Microsoft Academic query using the search terms "provenance" and "computer science"

Furthermore, the heterogeneous system-to-system analytics can also impede the execution of such rights of data subjects such as the right to erasure (Art. 17) and the obligation to rectify erroneous data as soon as possible (Art. 5).

Design challenges faced in the context of data provenance include:

- **Mass of provenance information** – The use of many different analytics and storage solutions quickly result in a prohibitively large amount of provenance information to be transferred between systems. When applied to big data, provenance problems become prohibitive. For instance, one of the most successful data provenance techniques consists in the so-called annotation-based approaches that propose modifying the input database queries in order to support data provenance tasks, while being able to access all the target data set. Obviously, the latter requirement becomes very hard when applied to big data repositories. [138] There have been proposals for provenance compression, but compression and decompression require computing effort and may shift the problem of storage overhead to performance overheads.[139] Fine-grain provenance tracking would require the computation of provenance that is several times larger than the original data.
- **Overhead** – There is an overhead linked to the collection of provenance data. The overhead varies based on the computation of provenance. Computational overhead is particularly high if provenance is computed every time data is transformed. This approach is also called eagerly provenance model. There is always an execution overhead when computing provenance on top of the computation cost related to the analysis.[140] Execution overhead and provenance storage size are workflow specific and increase with the input data size.[141]
- **Reproduction** – It is difficult to reproduce an execution from provenance data. Many existing provenance systems only record intermediate data generated during execution and their dependencies[142] or only when the provenance is required[143]. The latter approach is also called lazy provenance model. Execution environment information, which is also important for reproduction, is often neglected.
- **Lack of tools** – There is a need for flexible query tools and visualisation tools.[144] Query tools must support users who may be interested in tracking records generated by a particular person or detecting the confidentiality of tracked records. Visualisation tools are important for big data

[138] Alfredo Cuzzocrea, "Provenance Research Issues and Challenges in the Big Data Era," in *Proceedings of the IEEE 39th Annual International Computers, Software & Applications Conference*, 684–6 (IEEE, 2015)
[139] Thye W. Phua and Ryan K. L. Ko, "Data Provenance for Big Data Security and Accountability," in *Encyclopedia of Big Data Technologies*, ed. Sherif Sakr and Albert Zomaya (Cham: Springer, 2018)
[140] Jianwu Wang et al., "Big data provenance" in *Proceedings of the 2015 IEEE International Conference on Big Data*
[141] Daniel Crawl, Jianwu Wang and Ilkay Altintas, "Provenance for MapReduce-based data-intensive workflows," in *Proceedings of the 6th Workshop on Workflows in Support of Large-Scale Science*, 21–30 (New York, NY, USA: ACM, 2011)
[142] Jianwu Wang et al., "Big data provenance" in *Proceedings of the 2015 IEEE International Conference on Big Data*
[143] Alfredo Cuzzocrea, "Provenance Research Issues and Challenges in the Big Data Era" in *Proceedings of the IEEE 39th Annual International Computers, Software & Applications Conference*
[144] ibid.

Grant Agreement number: 731873

provenance techniques, as provenance is an interactive process that typically requires intelligent tools for visualising actual results and supporting next-step decisions.

Promising ways to overcome challenges faced in the context of data provenance include:

- **Secure provenance schemes** – Over the years, a number of secure provenance schemes have been proposed. A secure provenance scheme is used to provide important security properties and protect the collected provenance. Zafar et al.[145] provide a comprehensive overview and a comparative analysis of secure schemes. Most recently, dynamic Bayesian network and overlapped arithmetic coding, bloom filters and blockchains have received attention.
- **Fine-grain provenance** – Huq et al.[146] propose a tool which can infer fine-grained data provenance based on a given script. The tool visualises data dependencies among operations. Users can select individual values, which are generated by executing the script, and infer fine-grain provenance. Amsterdamer et al.[147] present a provenance framework that marries database-style and workflow-style provenance to capture internal state and fine-grained dependencies. The framework allows to choose the desired level of granularity in provenance querying and supports workflow analytic queries.
- **End-to-end tracking** – Ko & Will[148] present a kernel-space logger which potentially empowers all cloud stakeholders to trace their data. Logging from the kernel space empowers security analysts to collect provenance from the lowest possible atomic data actions and enables several higher-level tools to be built for effective end-to-end tracking of data provenance. The logger provides log tamper-evidence, prevention of fake or manual entries, an accurate timestamp synchronisation across several machines, efficient log space growth, and the accurate logging of root usage of the system.

3.9. Access, portability and user control

Access and portability facilitate the use and handling of data in different contexts. Having access to data means that people can view the data stored. Portability gives people 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 enables far-reaching user control, others are privacy preferences, sticky policies and personal data stores.

Access and portability foster competition. However, they can create problems as data may be brought from one domain where there are safeguards to another domain that is riskier. Moreover, access and

[145] Faheem Zafar et al., "Trustworthy data: A survey, taxonomy and future trends of secure provenance schemes," *Journal of Network and Computer Applications* 94 (2017)
[146] Mohammad R. Huq, Peter M. G. Apers and Andreas Wombacher, "ProvenanceCurious: A tool to infer data provenance from scripts," in *Proceedings of the 16th International Conference on Extending Database Technology*, 765–8 (ACM, 2013)
[147] Yael Amsterdamer et al., "Putting Lipstick on Pig: Enabling database-style workflow provenance," *Proceedings of the VLDB Endowment* 5, no. 4 (2011)
[148] Ryan K.L. Ko and Mark A. Will, "Progger: An Efficient, Tamper-Evident Kernel-Space Logger for Cloud Data Provenance Tracking," in *Proceedings of the 7th International Conference on Cloud Computing*, 881–9 (Anchorage, AK, USA: IEEE, 2014)

portability need 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 between the parties. As service providers do not only face a higher risk of losing users but also see the opportunity to gain users, many will eventually accept the paradigm of access and portability. With respect to user control, the informational asymmetry between users and data collectors is considered a central challenge. The fact that people get control over their data does not resolve the asymmetry. The user 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 user has greater control does not change this. Consequently, providing the user with more information to make the consent more informed does not really address the underlying challenge. People should be empowered, but to some extent, this also means that responsibility is pushed onto them. Therefore, big data solutions should be designed 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 to understand what the outcome of certain decisions will be, especially where complicated algorithms are used. Technologies for user control are currently neither mature nor user-friendly enough to be used by millions of people.

Due to the broad use of the terms, it was not possible to reliably identify key institutions doing research on access, portability and user control using Microsoft Academic.

From the perspective of the GDPR, data subjects gained additional rights with respect to access, portability and user control. Specifically, as mentioned earlier regarding how to learn what happens to data subjects' data, the right of access by the data subject (Art. 15) is a new right provided for data subjects. Furthermore, the right to data portability is another new right that enables the data subject to "have the personal data transmitted directly from one controller to another, where technically feasible". To this right there are also limitations attached; for instance, a data subject cannot exercise this right if processing is "necessary for the performance of a task carried out in the public interest" or, if the data controller carries out its duty as an "official authority". With respect to user control, from the legal perspective of the GDPR, new conditions for consent were introduced in Art. 7 that were aimed at addressing the problem that one consent given to one company in practice became interpreted and used as a consent given to all other companies that were linked to the first one. In an era where the dispersed network of big data systems is only growing, Art. 26 of GDPR also introduced the concept of joint controllership. This article specifies that big data companies that jointly determine the purpose and means of processing should all be considered controllers. In the Article 29 Working Party Guidelines on consent under Regulation 2016/679,[149] all of these controllers should obtain explicit consent from the data subjects. This measure can be understood as being aimed at increasing data subjects' user control.

However, without addressing the challenges that are presented to computer scientists, possibilities for compliance are limited and these rights cannot achieve their full potential. Hence any tool that aims to address these challenges also fosters the enforceability of these rights.

[149] Article 29 Working Party Guidelines on consent under Regulation 2016/679 (17/EN WP259 rev.01) https://www.datenschutz-praxis.de/wp-content/uploads/2018/06/20180416_Article29WPGuidelinesonConsent_publishpdf.pdf

Design challenges faced in the context of access, portability and user control include:

- **Legal unclearness** – Legal unclearness is particularly relevant with respect to access and portability. For instance, even though the GDPR entered into force, the object of data portability is still considered unclear and likely to have a too restrictive interpretation.[150] It is unclear in this context if the object of data portability is restricted to the withdrawal of data explicitly given to a controller or if it also includes data observed by the controller as well as the direct data transfer to another controller. Legal unclearness is not directly a technical challenges but it can make the design of legally compliant software solutions difficult.
- **Lack of tools for personal data management** – Privacy and legal concerns, as well as the lack of technical solutions for personal data management prevent personal data from being shared and reconciled under the control of the individual. [151] The lack of access and control fuels growing concerns as it prevents individuals from understanding and managing the risks associated with the collection and use of their data.
- **Large amount of personal data** – Especially in online social networks, people are sharing a lot of sensitive personal information. While such networks do permit users to control what they share with whom, access control policies are notoriously difficult to configure correctly.[152] Moreover, many people do share personal information not only through one social network but trough several ones as well as numerous other online services. This raises the question of whether the users' privacy settings match their sharing intentions.

Another challenge that came up in an earlier phase of the project[153] is the erosion of user control in the big data context because data that has been released cannot easily be controlled anymore. In the literature, we could not find much evidence referring to these challenges.

Promising ways to overcome challenges faced in the context of access, portability and user control include:

- **Re-balancing power** – Chaudry et al.[154] present a networked platform that collates and mediates access to personal data. The authors see the platform as a first step to re-balancing the power between data subjects and the corporations that collect and use their data. De Montjoye et al. [155]

[150] Paul de Hert et al., "The right to data portability in the GDPR: Towards user-centric interoperability of digital services," *Computer Law & Security Review* 34, no. 2 (2018)
[151] Yves-Alexandre de Montjoye et al., "openPDS: Protecting the Privacy of Metadata through SafeAnswers," *PloS one* 9, no. 7 (2014)
[152] Michelle Madejski, Maritza Johnson and Steven M. Bellovin, "The Failure of Online Social Network Privacy Settings," (Columbia University, 2011), https://academiccommons.columbia.edu/doi/10.7916/D8QV3VB4/download (accessed January 21, 2019)
[153] See e-SIDES D3.2 - Assessment of existing technologies
[154] Amir Chaudhry et al., "Personal Data: Thinking Inside the Box," *Aarhus Series on Human Centered Computing* 1, no. 1 (2015)
[155] Montjoye et al., "openPDS"

describe a personal metadata management framework with high user control. Heitmann et al.[156] present an architecture for privacy-enabled user profile portability, based on technologies from the emerging Web of Data.

- **Keep an overview** – The amount of user-generated media uploaded to the web is expanding rapidly and it is beyond the capabilities of any person to sift through it all to see which media impacts his or her privacy. Smith et al.[157] present a concept by which users can stay informed about which parts of the big data deluge is relevant to them. The authors propose a watchdog service that can be operated in different ways.
- **Decentralised personal data management** – Often, individuals still have little or no control over the data that is stored about them and how it is used. Zyskind et al.[158] present a model that uses blockchains to protect personal data. They implement a protocol that turns a blockchain into an automated access-control manager that does not require trust in a third party.

Besmer & Richter Lipford[159] present a tool that allows users that are tagged in photos to send a request to the owner to hide the linked photo from certain people. This approach follows the ideas that forewarned is forearmed and that creating awareness of critical content is the first step towards the solution of the problem. The work relies on direct technical tags. Besmer & Richter Lipford list several design considerations that are relevant in the domain of their tool.

[156] Benjamin Heitmann et al., "An architecture for privacy-enabled user profile portability on the web of data," in *Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems - HetRec '10*, ed. Peter Brusilovsky et al., 16–23 (New York, New York, USA: ACM Press, 2010)

[157] Matthew Smith et al., "Big data privacy issues in public social media," in *Proceedings of the 6th IEEE International Conference on Digital Ecosystems and Technologies* (IEEE, 2012)

[158] Guy Zyskind, Oz Nathan and Alex Pentland, "Decentralizing Privacy: Using Blockchain to Protect Personal Data," in *Proceedings of the 2015 IEEE Security and Privacy Workshops*, 180–4 (San Jose, CA, USA: IEEE, 2015)

[159] Andrew Besmer and Heather Richter Lipford, "Moving beyond untagging: photo privacy in a tagged world," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ed. Elizabeth Mynatt et al., 1563–72 (ACM, 2010)

4. Conclusion – Requirements for the design and use of privacy-preserving big data solutions

This section describes requirements for the design and use of big data solutions, which can be considered privacy preserving because they include privacy-preserving technologies or are used in an environment in which non-technical measures were taken to preserve privacy. In most cases, both points are likely to apply. By design, big data solutions can only lay the foundation for privacy-preserving data sharing and usage. It is equally important that the solutions are used in an appropriate manner and environment. The purpose of this section is to lay the foundation for the assessment of big data solutions under development and the formulation of recommendations in WP5. However, the requirements are quite general, and may have to be adjusted to specific contexts and operationalised to be properly assessable.

The requirements are based on previous e-SIDES deliverables (especially, the gap analysis documented in D4.1 and the assessment of classes of privacy-preserving technologies presented in D3.2), a review of related previous work (presented in section 2**Error! Reference source not found.**) and an analysis of design challenges in the context of privacy-preserving technologies (presented in section 3). For each requirement, we describe why it is important and what aspects be taken into account when taking measures to meet it. The requirements primarily target developers of big data solutions, but they are also relevant for developers of privacy-preserving technologies as well as to operators of big data solutions.

4.1. Embed security and privacy features

The importance of privacy-preserving technologies grows continuously with information systems becoming increasingly networked within organisations and across organisational boundaries, and datasets becoming larger and more heterogeneous. Security and privacy features based on these technologies need to be embedded in big data solutions rather that provided as extras or optional add-ons. Moreover, the features need to be activated and, if possible, configured so that they provide a high level of privacy protection by default. Not only because a tight integration increases the probability that they work effectively and efficiently, but also because operators of big data solutions are rather unlikely to purchase and use privacy add-ons. The pressure exerted by their clients or users, data protection authorities, or the legal system is still too low. Irrespective of external pressure, some organisational cultures do not seem to be ready for a truly privacy-preserving business conduct.

As reported in D3.2, privacy-preserving technologies are integrated in today's big data solutions only to a limited extent. There are strong solutions in research but there is a big gap when it comes to deployment. One would expect that the handling of personal data and privacy protection are very important for operators of big data solutions that are highly networked and process personal data as well as their clients or users. However, it seems that clients and users are blinded by the benefits they get in return for their personal data. Consequently, it is not surprising that there is a rather low demand from the customer side for big data solutions that help protecting privacy. The privacy paradox has long been a heavily discussed topic in research. Therefore, but also because a fundamental right is concerned, people shouldn't have to pay extra for privacy preservation.

D4.1 discusses several reasons for the implementation gap. Concerning the embedding of security and privacy features into big data solutions, the cost-benefit ratio, the value of privacy for individuals and the

Grant Agreement number: 731873

skills needed are most relevant. Moreover, cultural values have at least some relevance. Adding privacy-preserving technologies to big data solutions leads to additional costs for both solution developers and operators of solutions. To make sense form an economic point of view, these costs must be offset by the expected benefits. It is important that the long-term effects of higher trust in the data economy are considered as part of the equation. Moreover, it is important to note that the GDPR has only recently changed the relevance of fines significantly. The lack of pressure from the end users (i.e., the clients or users of the big data solution operators) and the difficulty to find staff that can properly integrate privacy-preserving technologies into big data solutions are important barriers to be kept in mind. Embedding security and privacy features into big data solutions may help operators of this solutions to overcome a culture that currently keeps them from putting more emphasis on data protection.

The protection goals introduced in section 2.5 may be useful to keep the embedding of security and privacy features focused. The goals provide a scheme for addressing the legal, technical, economic, and societal dimensions of privacy. As some of the goals are in conflict, it is impossible to achieve all goals at the same time. In sections 2.6, we showed that system design can benefit a lot from previous work. Privacy patterns, for instance, proved to be extremely useful to translate concerns about the flow of personal data into technical artefacts that address them. Comprehensive lists of such patterns are available. Moreover, there are data-oriented as well as process oriented design strategies that do not only help achieving the protection goals but are also related to the privacy patterns. Last but not least, insight into common design mistakes allows to doing it right at the first time.

Particularly since adding privacy-preserving technologies to big data solutions leads to additional costs, it needs to be stressed that privacy-preserving technologies need to be combined to be effective. There is neither one most important technology nor a most important class of technologies. Moreover, 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). The optimal combination of technologies will definitely vary from one setting to another. As section 3 clearly shows, the technologies differ greatly in their maturity. This needs to be kept in mind. Relying on a less mature technology generally means taking a higher risk, but it may pay off due to resulting competitive advantages.

If organisations dealing with personal data are reasonably transparent about their practices and know to use big data solutions that are inherently privacy preserving, trust in the data economy will most likely increase in the long run. Eventually, privacy preservation will thus lead to a win-win situation.

4.2. Take preventive measures

Because of the serious negative implications that privacy breaches may have on the affected individuals, privacy breaches need to be prevented before they happen. Further reasons include significant fines that are more and more often the consequence of data breaches, considerable negative effects on the reputation of organisations involved in data breaches and the general reduction of trust in the data economy. Recent data breaches clearly show the limitations of reactive approaches that typically prescribe measures to be taken when privacy is violated or when certain rules are broken. Technologies are considered proactive in the sense that they prevent incidents or rule violations in the first place. Privacy by design, which is closely related to the concept of privacy-preserving technologies as it requires

Grant Agreement number: 731873

privacy safeguards to be integrated in technological solutions, plays a key role with respect to the prevention of breaches before they happen.

In D3.2, we stressed that organisations collecting, using and distributing data are generally responsible for data management and anonymisation. The GDPR, for instance, quite clearly defines the responsibilities of data controllers and data processors dealing with personal data of people living in the EU. Regardless of the legal requirements, there seems to be wide agreement that the strongest party should have the biggest responsibilities to prevent breaches and data misuse. Still, it seems that end-users need to protect themselves as nobody else will do it for them. Giving the control of their data back to the user, is often considered a viable approach. However, it is important that organisations do not consider data protection as somebody else's task or problem as this point of view passes the responsibility from one hand to another. The responsibility should not be pushed to the end users who may not fully understand what they are doing.

Privacy by design received quite some attention in D4.2. Already in 2011, seven foundational principles of privacy by design were developed and all of them are highly relevant in the big data context. Although the principles still create the bedrock for privacy by design innovation, critics argue that their flexible interpretation is both a blessing and a curse for practitioners. So far, privacy by design has received considerable attention in policy circles but the actual design, implementation and integration remains an open question. Therefore, we discussed it in the context of the implementation gap with respect to privacy-preserving technologies. Data breaches have continued to occur despite laws and regulations that require organisations to implement security measures. Researchers have found evidence for a statistically significant negative impact of data breaches on a company's market value. However, as the impact seems to be neither very large nor sustainable, the resulting pressure to act for organisations dealing with personal data is rather small. It may be that the potential for competitive advantages through taking privacy and transparency seriously, and making this public is a more significant for organisations to prevent breaches before they happen than the fear of negative effects of data breaches.

Related work on privacy by design, which might help to make progress with respect to the open question regarding its actual design, implementation and integration, was discussed in section 2.7. The features we introduce can make a first step in this direction. They go beyond the foundational principles by detailing what more exactly has to be done with the data and how concrete methods should be designed. Apart from the features, key criteria to be considered in a privacy by design methodology were outlined. The recommendations and risk mitigation strategies describe good practices that are specifically tailored to the particular needs of big data contexts. Implementing the recommendations and/or strategies may be useful to keep the implementation process focused and help to avoid common mistakes. Having the protection goals, which were introduced in section 2.5, and their specific in mind cannot hurt when putting privacy by design to work.

Technologies for user control play an important role to prevent breaches – or, to be more precise, misuse from the end user's point of view – before they happen, particularly if it is not a question of legal or illegal. Many online services, particularly ones that offer their services free of charge, provider their users with a plethora of settings to define precisely what may be shared with how. The design challenges discussed in section 3.9 show that making the right settings is not easy for individuals. Technologies for user control are currently neither mature nor user-friendly enough to be used by millions of people.

Just as making sure that security and privacy features are embedded in big data solutions, preventing breaches before they happen has the potential to significantly contribute to achieving a level of trust in the data economy that is considerably higher than the one that is observed today.

4.3. Connect people, processes and technology

A combination of technical and non-technical measures is essential. To ensure privacy, knowledgeable people as well as proper processes are important complements to privacy-preserving technologies. For instance, a particular need for increased awareness, improved usability and education targeting was identified. Currently, most big data solutions, no matter if they have particular security and privacy features, can only be used correctly by experts. In general, technologies have limitations and alone are not sufficient. At some level, non-technical measures will always be necessary to make sure a given technology functions as expected. Moreover, data protection officials are needed that are aware and able to assess the impact on privacy or risks regarding personal data that is collected and used by organisations. Therefore, there is a need to consider people, process and technology when designing or using privacy-preserving big data solutions.

The assessment of existing privacy-preserving technologies provided in D3.2 supports the view that technical and non-technical measures need to be combined. However, it remains unclear how this combination is best realised, particularly as the technology landscape continuously evolves. D4.1 pays particular attention to the limit of technological approaches. Technical and non-technical measures are considered as equally important, complementary and partly overlapping with respect to their potential. Consequently, insight into the role that non-technical measures can play is essential not only for developing reasonable design requirements for future big data solutions, but also for better understanding the design of existing big data solutions. Across application contexts, privacy-preserving technologies as well as big data solutions have weaknesses that still need to be addressed. For the time being, its people and processes that have to mitigate the consequences of these weaknesses. Cultural values in an organisation or geographic region determine to some extent how the combination of people, processes and technologies will look like.

The importance of considering people, process and technology can perfectly be seen when looking at well-established privacy principles. The principles provided by the OECD (see section 2.1), ISO/IEC (see section 2.2) or the US Federal Trade Commission (see section 2.4) and taken up also by legal regulations such as the GDPR (see section 2.3), cannot be put into practice properly without dealing with people, process and technology. The major sets of privacy principles differ only in the details. Each of the privacy preserving technologies introduced in D3.1 is useful with respect to at least one of the principles and hardly any of the principles can be properly implemented without people and processes.

The analysis of design challenges provided in section 3 also clearly shows that technological approaches have limits. Without doubt, privacy-preserving technologies are essential since information systems are becoming increasingly networked and datasets are becoming larger and more heterogeneous, but people and processes are at least as relevant to make privacy-preserving big data solutions a reality.

The combination of technical and non-technical measures is essential to detect, investigate and prevent data breaches and misuse. Awareness of the fact that people and processes are there to make sure that a given technology functions as expected with help increasing trust in the data economy.

Grant Agreement number: 731873

4.4. Comply with laws and corporate policies

Big data solutions need to comply with laws and corporate policies. At the same time, they must be flexible enough to be adapted to changing demands, not only with respect to the business side but also with respect to the regulatory side. The heavy discussions that recently accompanied the entry into force of the GDPR show that achieving legal compliance is not always trivial, even if the legal text has been available for a long time.

Regulations are not always in line with business models or seem to hamper scientific progress. As discussed in detail in D4.1, regulations with respect to sensitive or inferred data, for example, do not find agreement without exception, which might lead to circumvention attempts. For instance, healthcare researchers argue that strict privacy protection and the robust implementation of privacy-preserving technologies could hamper epidemiology research as well as the value of big data for the advancement of healthcare research. Currently, regional differences in data protection and antitrust or competition law regimes between the EU and the US also influences the extent to which liabilities of big data stakeholders are made transparent and the extent to which privacy-preserving technologies are embraced and implemented.

In section 2.3, the privacy principles described in Art. 5 of the GDPR are presented. These principles should lie at the heart of every approach to processing personal data of people living in the EU. The principles do not give hard and fast rules but rather embody the spirit of the GDPR and, as such, there are very limited exceptions. Compliance with the spirit of these key principles is therefore a fundamental building block for good data protection practice. Takin the principles seriously is the key to being compliance with the detailed provisions of the GPDR. Failure to comply with the principles may result in substantial fines. Art. 83(5)(a) of the GDPR states that infringements of the basic principles are subject to the highest tier of administrative fines.

The e-SIDES classes of privacy-preserving technologies are well suited to put the principles into practice. Of particular relevance for all principles but the one focusing on integrity and confidentiality are technologies for user control, access and portability, data provenance, transparency and accountability, and policy enforcement. Policy enforcement technologies can play a key role with respect to the achievement and proof of compliance with laws and corporate policies. With respect to integrity and confidentiality, technologies for encryption, anonymisation, sanitisation, deletion, access control are policy enforcement are most relevant. The technologies are relevant to put the principles into practice but there are limits, for instance, with respect to the assessment of lawfulness or legitimacy. Here, non-technical measures definitely have to complement technical ones.

For big data solutions, it is essential to be not only legally compliant but also in line with corporate policies. Organisation must be able to demonstrate that they have full control over their data data-related activities. This is a prerequisite for trust in the data economy.

Grant Agreement number: 731873

Bibliography

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).

Ahuja, Rohit, and Sraban K. Mohanty. “A Scalable Attribute-Based Access Control Scheme with Flexible Delegation cum Sharing of Access Privileges for Cloud Storage.” *IEEE Transactions on Cloud Computing* (2017): 1.

Aïmeur, Esma. “Personalisation and Privacy Issues in the Age of Exposure.” In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*, 375–6. ACM, 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.

Al-Shomrani, Abdullah, Fathy Fathy, and Kamal Jambi. “Policy enforcement for big data security.” In *Proceedings of the 2nd International Conference on Anti-Cyber Crimes*, 70–4. IEEE, 2017.

Amsterdamer, Yael, Susan B. Davidson, Daniel Deutch, Tova Milo, Julia Stoyanovich, and Val Tannen. “Putting Lipstick on Pig: Enabling database-style workflow provenance.” *Proceedings of the VLDB Endowment* 5, no. 4 (2011): 346–357.

Ananny, Mike, and Kate Crawford. “Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability.” *New Media & Society* 20, no. 3 (2017): 973–989.

Archer, David W., Dan Bogdanov, Yehuda Lindell, Liina Kamm, Kurt Nielsen, Jakob I. Pagter, Nigel P. Smart, and Rebecca N. Wright. “From Keys to Databases—Real-World Applications of Secure Multi-Party Computation.” *The Computer Journal* 9 (2018): 192.

Archer, David W., Dan Bogdanov, Benny Pinkas, and Pille Pullonen. “Maturity and Performance of Programmable Secure Computation.” *IEEE Security & Privacy* 14, no. 5 (2016): 48–56.

Barocas, Solon, and Andrew D. Selbst. “Big Data’s Disparate Impact.” *California Law Review* 104 (2016): 671–732.

Besmer, Andrew, and Heather Richter Lipford. “Moving beyond untagging: photo privacy in a tagged world.” In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Edited by Elizabeth Mynatt et al., 1563–72. ACM, 2010.

Betge-Brezetz, Stephane, Guy-Bertrand Kamga, Marie-Pascale Dupont, and Aoues Guesmi. “End-to-end privacy policy enforcement in cloud infrastructure.” In *Proceedings of the IEEE 2nd International Conference on Cloud Networking*, 25–32. IEEE, 2013.

Calders, Toon, and Indrė Žliobaitė. “Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures.” In *Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases*. Edited by Bart Custers et al., 43–57. Studies in applied philosophy, epistemology and rational ethics 3. Berlin, Heidelberg: Springer, 2013.

Cavoukian, Ann, and Jeff Jonas. “Privacy by Design in the Age of Big Data.”. https://jeffjonas.typepad.com/Privacy-by-Design-in-the-Era-of-Big-Data.pdf (accessed December 14, 2017).

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.

Chaudhry, Amir, Jon Crowcroft, Heidi Howard, Anil Madhavapeddy, Richard Mortier, Hamed Haddadi, and Derek McAuley. “Personal Data: Thinking Inside the Box.” *Aarhus Series on Human Centered Computing* 1, no. 1 (2015): 4.

Chen, Deyan, and Hong Zhao. “Data Security and Privacy Protection Issues in Cloud Computing.” In *Proceedings of the 2012 International Conference on Computer Science and Electronics Engineering*, 647–51. IEEE, 2012.

Chen, Feng, Xiaoqian Jiang, Shuang Wang, Lisa M. Schilling, Daniella Meeker, Toan Ong, Michael E. Matheny, Jason N. Doctor, Lucila Ohno-Machado, and Jaideep Vaidya. “Perfectly Secure and Efficient Two-Party Electronic-Health-Record Linkage.” *IEEE internet computing* 22, no. 2 (2018): 32–41.

Chen, Lily, Stephen Jordan, Yi-Kai Liu, Dustin Moody, Rene Peralta, Ray Perlner, and Daniel Smith-Tone. “Report on Post-Quantum Cryptography.”. Internal Report 8105. https://nvlpubs.nist.gov/nistpubs/ir/2016/NIST.IR.8105.pdf (accessed January 20, 2019).

Crawl, Daniel, Jianwu Wang, and Ilkay Altintas. “Provenance for MapReduce-based data-intensive workflows.” In *Proceedings of the 6th Workshop on Workflows in Support of Large-Scale Science*, 21–30. New York, NY, USA: ACM, 2011.

Custers, Bart H. M. “Data Mining and Group Profiling on the Internet.” In *Ethics and the internet*. Edited by A.H Vedder, 87–104. Antwerpen, Groningen: Intersentia, 2001.

——— . “Effects of Unreliable Group Profiling by Means of Data Mining.” In *Discovery science: 6th international conference, DS 2003 Sapporo, Japan, October 17-19, 2003 : proceedings*. Edited by Gunter Grieser, Y. Tanaka and Akihiro Yamamoto, 291–6. Lecture Notes in Computer Science 2843. Lecture notes in artificial intelligence. Berlin, New York: Springer, 2003.

Cuzzocrea, Alfredo. “Provenance Research Issues and Challenges in the Big Data Era.” In *Proceedings of the IEEE 39th Annual International Computers, Software & Applications Conference*, 684–6. IEEE, 2015.

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).

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

Domingo-Ferrer, Josep, and Krishnamurty Muralidhar. “New directions in anonymization: Permutation paradigm, verifiability by subjects and intruders, transparency to users.” *Information Sciences* 337-338 (2016): 11–24.

Dougherty, Conor. “Google Photos Mistakenly Labels Black People ‘Gorillas’.”. http://bits.blogs.nytimes.com/2015/07/01/google-photos-mistakenly- (accessed January 21, 2019).

Drozd, Olha. “Privacy Pattern Catalogue: A Tool for Integrating Privacy Principles of ISO/IEC 29100 into the Software Development Process.” In *Privacy and Identity Management. Time for a Revolution? 10th IFIP WG 9.2, 9.5, 9.6/11.7, 11.4, 11.6/SIG 9.2.2 International Summer School, Edinburgh, UK, August 16-21, 2015, Revised Selected Papers*. Edited by David Aspinall et al., 129–40.

IFIP Advances in Information and Communication Technology 476. Cham: Springer International Publishing, 2016.

Duan, Yitao, and John Canny. “From Commodity to Value: A Privacy-Preserving e-Business Architecture.” In *Proceeding of the 2006 IEEE International Conference on e-Business Engineering*. Edited by Wei-Tek Tsai, 488–95. Los Alamitos, CA, USA: IEEE, 2006.

Elliot, Mark, Elaine Mackey, Kieron O’Hara, and Caroline Tudor. “The Anonymisation Decision-Making Framework.”. http://ukanon.net/wp-content/uploads/2015/05/The-Anonymisation-Decision-making-Framework.pdf (accessed January 20, 2019).

Fatema, Kaniz, David W. Chadwick, and Stijn Lievens. “A Multi-privacy Policy Enforcement System.” In *Privacy and Identity Management for Life: 6th IFIP WG 9.2, 9.6/11.7, 11.4, 11.6/PrimeLife International Summer School, Helsingborg, Sweden, August 2-6, 2010, Revised Selected Papers*. Edited by Simone Fischer-Hübner et al., 297–310. IFIP Advances in Information and Communication Technology 352. Berlin, Heidelberg: IFIP International Federation for Information Processing, 2011.

Ferdous, Md S., Andrea Margheri, Federica Paci, Mu Yang, and Vladimiro Sassone. “Decentralised Runtime Monitoring for Access Control Systems in Cloud Federations.” In *Proceedings of the IEEE 37th International Conference on Distributed Computing Systems*, 2632–3. IEEE, 2017.

Geng, Quan, and Pramod Viswanath. “Optimal Noise Adding Mechanisms for Approximate Differential Privacy.” *IEEE Transactions on Information Theory* 62, no. 2 (2016): 952–969.

Gheorghe, Gabriela, Stephan Neuhaus, and Bruno Crispo. “xESB: An Enterprise Service Bus for Access and Usage Control Policy Enforcement.” In *Trust Management IV: 4th IFIP WG 11.11 International Conference, IFIPTM 2010, Morioka, Japan, June 16-18, 2010; Proceedings*. Edited by Masakatsu Nishigaki, 63–78. IFIP Advances in Information and Communication Technology 321. Berlin: Springer, 2010.

Ghosh, Arpita, Tim Roughgarden, and Mukund Sundararajan. “Universally Utility-maximizing Privacy Mechanisms.” *SIAM Journal on Computing* 41, no. 6 (2012): 1673–1693.

Graux, Hans, Jef Ausloos, and Peggy Valcke. “The Right to Be Forgotten in the Internet Era.”. ICRI Research Paper 11.

Gunning, David. “Explainable Artificial Intelligence.”. https://www.darpa.mil/attachments/XAIProgramUpdate.pdf (accessed January 21, 2019).

Hamlen, Kevin W., Lalana Kagal, and Murat Kantarcioglu. “Policy Enforcement Framework for Cloud Data Management.” *Bulletin of the Technical Committee on Data Engineering* 35, no. 4 (2012): 39–45.

Hansen, Marit. “Top 10 Mistakes in System Design from a Privacy Perspective and Privacy Protection Goals.” In *Privacy and identity management for life: 7th IFIP WG 9.2, 9.6/11.7, 11.4, 11.6 international summer school, Trento, Italy, September 5-9, 2011; revised selected papers.* vol. 375. Edited by Jan Camenisch, 14–31. IFIP Advances in Information and Communication Technology 375. [Berlin], [Heidelberg]: Springer, 2012.

Hansen, Marit, Meiko Jensen, and Martin Rost. “Protection Goals for Privacy Engineering.” In *Proceedings of the 2015 IEEE Security and Privacy Workshops*, 159–66. San Jose, CA, USA: IEEE, 2015.

Heitmann, Benjamin, James G. Kim, Alexandre Passant, Conor Hayes, and Hong-Gee Kim. “An architecture for privacy-enabled user profile portability on the web of data.” In *Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems - HetRec '10*. Edited by Peter Brusilovsky et al., 16–23. New York, New York, USA: ACM Press, 2010.

Hert, Paul de, Vagelis Papakonstantinou, Gianclaudio Malgieri, Laurent Beslay, and Ignacio Sanchez. "The right to data portability in the GDPR: Towards user-centric interoperability of digital services." *Computer Law & Security Review* 34, no. 2 (2018): 193–203.

Hoepman, Jaap-Henk. "Privacy Design Strategies." In *ICT systems security and privacy protection: 29th IFIP TC 11 International Conference, SEC 2014, Marrakech, Morocco, June 2-4, 2014. Proceedings.* vol. 428. Edited by Nora Cuppens-Boulahia et al., 446–59. IFIP Advances in Information and Communication Technology 428. Heidelberg: Springer, 2014.

Hu, Vincent C., D. R. Kuhn, and David F. Ferraiolo. "Attribute-Based Access Control." *Computer* 48, no. 2 (2015): 85–88.

Huang, Zhicong, Erman Ayday, Huang Lin, Raeka S. Aiyar, Adam Molyneaux, Zhenyu Xu, Jacques Fellay, Lars M. Steinmetz, and Jean-Pierre Hubaux. "A privacy-preserving solution for compressed storage and selective retrieval of genomic data." *Genome research* 26, no. 12 (2016): 1687–1696.

Huq, Mohammad R., Peter M. G. Apers, and Andreas Wombacher. "ProvenanceCurious: A tool to infer data provenance from scripts." In *Proceedings of the 16th International Conference on Extending Database Technology*, 765–8. ACM, 2013.

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.

"ISO/IEC 29100: Information technology - Security techniques - Privacy framework.".

Jaatun, Martin G., Siani Pearson, Frederic Gittler, and Ronald Leenes. "Towards Strong Accountability for Cloud Service Providers." In *Proceedings of the IEEE 6th International Conference on Cloud Computing Technology and Science*, 1001–6. IEEE, 2014.

Jost, Christine, Ha Lam, Alexander Maximov, and Ben J. M. Smeets. "Encryption Performance Improvements of the Paillier Cryptosystem.". IACR Cryptology ePrint Archive. https://www.iacr.org/cryptodb/data/paper.php?pubkey=26497 (accessed January 20, 2019).

Karnouskos, Stamatis, and Florian Kerschbaum. "Privacy and Integrity Considerations in Hyperconnected Autonomous Vehicles." *Proceedings of the IEEE* 106, no. 1 (2018): 160–170.

Ko, Ryan K.L., and Mark A. Will. "Progger: An Efficient, Tamper-Evident Kernel-Space Logger for Cloud Data Provenance Tracking." In *Proceedings of the 7th International Conference on Cloud Computing*, 881–9. Anchorage, AK, USA: IEEE, 2014.

Laat, Paul B. de. "Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability?" *Philosophy & Technology* 31, no. 4 (2018): 525–541.

Li, Wei, Kaiping Xue, Yingjie Xue, and Jianan Hong. "TMACS: A Robust and Verifiable Threshold Multi-Authority Access Control System in Public Cloud Storage." *IEEE Transactions on Parallel and Distributed Systems* 27, no. 5 (2016): 1484–1496.

Lin, Huaqing, Zheng Yan, and Raimo Kantola. "CDController: A Cloud Data Access Control System Based on Reputation." In *Proceedings of the 2017 IEEE International Conference on Computer and Information Technology*, 223–30. Helsinki, Finland: IEEE, 2017.

Liu, Chang, Rajiv Ranjan, Xuyun Zhang, Chi Yang, Dimitrios Georgakopoulos, and Jinjun Chen. "Public Auditing for Big Data Storage in Cloud Computing -- A Survey." In *Proccedings of the IEEE 16th International Conference on Computational Science and Engineering*, 1128–35. IEEE, 2013.

Madejski, Michelle, Maritza Johnson, and Steven M. Bellovin. “The Failure of Online Social Network Privacy Settings.”. https://academiccommons.columbia.edu/doi/10.7916/D8QV3VB4/download (accessed January 21, 2019).

Manikandasaran, S. S., and S. Sudha. “Data Access Control Techniques and Security Challenges in Cloud Computing: A Survey.” *International Journal of Computer Sciences and Engineering* 6, Special Issue 2 (2018): 87–95.

Martin, Kirsten E. “Ethical Issues in the Big Data Industry.” *MIS Quarterly Executive* 14, no. 2 (2015): 67–85.

Montjoye, Yves-Alexandre de, Erez Shmueli, Samuel S. Wang, Alex S. Pentland, and Tobias Preis. “openPDS: Protecting the Privacy of Metadata through SafeAnswers.” *PloS one* 9, no. 7 (2014): 1-9.

Morris, Liam. “Analysis of Partially and Fully Homomorphic Encryption.”. http://gauss.ececs.uc.edu/Courses/c6056/pdf/homo-outline.pdf (accessed January 20, 2019).

Naehrig, Michael, Kristin Lauter, and Vinod Vaikuntanathan. “Can homomorphic encryption be practical?” In *Proceedings of the 3rd ACM Workshop on Cloud Computing Security*, 113–24. ACM, 2011.

Ohm, Paul. “Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization.” *UCLA Law Review* 57 (2010): 1701–1777.

Phua, Thye W., and Ryan K. L. Ko. “Data Provenance for Big Data Security and Accountability.” In *Encyclopedia of Big Data Technologies*. Edited by Sherif Sakr and Albert Zomaya. Cham: Springer, 2018.

Reardon, Joel, David Basin, and Srdjan Capkun. “On Secure Data Deletion.” *IEEE Security & Privacy* 12, no. 3 (2014): 37–44.

“Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC.”.

Sankar, Lalitha, S. R. Rajagopalan, and H. V. Poor. “Utility-Privacy Tradeoffs in Databases: An Information-Theoretic Approach.” *IEEE Transactions on Information Forensics and Security* 8, no. 6 (2013): 838–852.

Sharma, Sagar, Keke Chen, and Amit Sheth. “Toward Practical Privacy-Preserving Analytics for IoT and Cloud-Based Healthcare Systems.” *IEEE internet computing* 22, no. 2 (2018): 42–51.

Shor, Peter W. “Polynomial-Time Algorithms for Prime Factorization and Discrete Logarithms on a Quantum Computer.” *SIAM Review* 41, no. 2 (1999): 303–332.

Smith, Matthew, Christian Szongott, Benjamin Henne, and Gabriele von Voigt. “Big data privacy issues in public social media.” In *Proceedings of the 6th IEEE International Conference on Digital Ecosystems and Technologies.* IEEE, 2012.

Sookhak, Mehdi, Adnan Akhunzada, Abdullah Gani, Muhammad Khurram Khan, and Nor B. Anuar. “Towards dynamic remote data auditing in computational clouds.” *The Scientific World Journal* (2014).

Soria-Comas, Jordi, and Josep Domingo-Ferrer. “Big Data Privacy: Challenges to Privacy Principles and Models.” *Data Science and Engineering* 1, no. 1 (2016): 21–28.

Spiekermann, S., and L. F. Cranor. “Engineering Privacy.” *IEEE Transactions on Software Engineering* 35, no. 1 (2009): 67–82.

Sweeney, Latanya. “Discrimination in online ad delivery.” *Communications of the acm* 56, no. 5 (2013): 44.

"The OECD Privacy Framework.". https://www.oecd.org/sti/ieconomy/oecd_privacy_framework.pdf (accessed October 23, 2018).

Vallor, Shannon. *Technology and the virtues: A philosophical guide to a future worth wanting*. First issued as an Oxford University Press paperback. New York, NY, United States of America: Oxford University Press, 2016.

van den Hoven, Jeroen. "Value Sensitive Design and Responsible Innovation." In *Responsible innovation: Managing the responsible emergence of science and innovation in society*. Edited by Richard Owen, J. R. Bessant and Maggy Heintz, 75–83. Chichester: Wiley, 2013.

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).

Wang, Wei, Yin Hu, Lianmu Chen, Xinming Huang, and Berk Sunar. "Exploring the Feasibility of Fully Homomorphic Encryption." *IEEE Transactions on Computers* 64, no. 3 (2015): 698–706.

Wright, David. "The state of the art in privacy impact assessment." *Computer Law & Security Review* 28, no. 1 (2012): 54–61.

Xue, Liang, Ni, Jianbing, Li, Yannan, and Jian Shen. "Provable data transfer from provable data possession and deletion in cloud storage." *Computer Standards & Interfaces* 54, no. 1 (2017): 46–54.

Yan, Zheng, Xueyun Li, Mingjun Wang, and Athanasios V. Vasilakos. "Flexible Data Access Control Based on Trust and Reputation in Cloud Computing." *IEEE Transactions on Cloud Computing* 5, no. 3 (2017): 485–498.

Yang, Changsong, Xiaofeng Chen, and Yang Xiang. "Blockchain-based publicly verifiable data deletion scheme for cloud storage." *Journal of Network and Computer Applications* 103 (2018): 185–193.

Yu, Shucheng, Cong Wang, Kui Ren, and Wenjing Lou. "Achieving Secure, Scalable, and Fine-grained Data Access Control in Cloud Computing." In *Proceedings of the 2010 IEEE INFOCOM*, 1–9. IEEE, 2010.

Zafar, Faheem, Abid Khan, Saba Suhail, Idrees Ahmed, Khizar Hameed, Hayat M. Khan, Farhana Jabeen, and Adeel Anjum. "Trustworthy data: A survey, taxonomy and future trends of secure provenance schemes." *Journal of Network and Computer Applications* 94 (2017): 50–68.

Zhao, Chuan, Shengnan Zhao, Minghao Zhao, Zhenxiang Chen, Chong-Zhi Gao, Hongwei Li, and Yu-an Tan. "Secure Multi-Party Computation: Theory, practice and applications." *Information Sciences* 476 (2019): 357–372.

Zhou, Changli, Chunguang Ma, and Songtao Yang. "An Improved Fine-Grained Encryption Method for Unstructured Big Data." In *Intelligent computation in big data era: International Conference of Young Computer Scientists, Engineers and Educators, ICYCSEE 2015, Harbin, China, January 10-12, 2015. Proceedings.* vol. 503. Edited by Hongzhi Wang, 361–9. Communications in computer and information science 503. Heidelberg: Springer, 2015.

Zyskind, Guy, Oz Nathan, and Alex Pentland. "Decentralizing Privacy: Using Blockchain to Protect Personal Data." In *Proceedings of the 2015 IEEE Security and Privacy Workshops*, 180–4. San Jose, CA, USA: IEEE, 2015.