CURRENT HOT TOPICS IN DATA PROTECTION

BDVA Position Paper

November 2022

BDV BIG DATA VALUE ASSOCIATION

Table of Contents

1 INTRODUCTION

This document is intended to summarize the current main challenges and developments in data protection, with a focus on the European Union (EU).

In 2019, the BDVA (Big Data Value Association) published a first position paper on data protection[1]. Since 2019, much has happened in the field of data protection: technical advances have created both new threats and new solutions, or made existing threats more severe or existing solutions more effective; the longer-term impact of the General Data Protection Regulation (GDPR) has started to become apparent, while already new relevant legislative acts are being introduced. Many of the findings and conclusions of the 2019 position paper are still valid, but the new developments necessitate a shift in focus, thus warranting this major update. What certainly remained true is that data protection is a crucially important topic (today more than ever!) with wide-ranging consequences for society, technology, policy, and business alike. In the following, we first review current challenges in data protection (Section 2), followed by promising data protection technologies (Section 3). The position paper ends with recommendations and conclusions in Section 4. The presented list of challenges and technologies is not meant to be exhausted; rather, it focuses on the ones that the authors deem most relevant and important.

2 CHALLENGES IN DATA PROTECTION

This section gives an overview about important current challenges in data protection. The identified challenges are grouped into three categories: technical challenges, legal challenges, and other important challenges (the latter category includes for example societal and market challenges).

2.1 Technical challenges

Privacy Enhancing Technologies engineering. Many Privacy Enhancing Technologies (PETs) are available to process data between multiple parties without having to exchange the data explicitly or in a readable format. We have also seen several startups delivering solutions in this domain. However, it is difficult to design such systems: selecting the right functionalities, combining them, incorporating them in products and services, and analyzing the potential information leakage is still a challenge for choosing and adapting the right PETs for a certain use case or application.

Scaling up. Technologies that are available to securely analyze data rely on advanced cryptographic mechanisms that can be very expensive in terms of computation. Advanced data analytics solutions, such as data-driven artificial intelligence (AI) techniques, require a large amount of calculations. Thus, developing large-scale AI solutions based on secure computation is a challenge. There is a strong need for more scalable PETs.

Interoperability. Actors that want to jointly analyze data based on Privacy Enhancing Technologies need to use the same technology to do so, also agreeing on parameters etc. Although work is being done to standardize protocols, solutions are needed that enable organizations to engage in secure computation with different partners that use different technologies from different vendors.

Anonymization and risk of re-identification. Anonymization is an irreversible process carried out through techniques such as the generalization of data or the addition of noise in a dataset. There is still a need for simple methods and tools to anonymize data so that they retain some utility.

[1] T. Timan, Z. Á. Mann (editors). Data protection in the era of artificial intelligence: Trends, existing solutions and recommendations for privacy-preserving technologies. BDVA, 2019, https://www.bdva.eu/node/1384

The re-identification of anonymized data remains a threat, especially as systems exchange data at a very large scale. We consider here the cases of mixed datasets, open data and location data, representative of the main challenges faced by current systems.

Mixed datasets. Even when a dataset is anonymous, if it is combined with one or more other datasets, this may result in an increase in the likelihood of re-identification. This is further exacerbated when at least one dataset contains an 'inextricable link'. GDPR applies to the whole matched dataset. This includes the datasets which contain an 'inextricable link' and are further matched, combined, or compared with anonymous datasets. Datasets that deal with highly sensitive data, such as healthcare datasets, are most at risk.

Open data. The Open Science movement has highlighted the importance of exchanging data to enhance technological progress, citizen science, and dissemination of research results. This is particularly important in the natural sciences domain and in medicine, as we have seen with the efforts to fight the Covid-19 pandemic. Sharing medical data is crucial to improve our understanding of diseases and to develop new drugs and treatments. However, because health records and clinical data are considered particularly sensitive data, this leads to increasing privacy challenges, such as the risk of re-identification and linking, especially when biometric and genetic data or brain scans are concerned, due to the difficulties in the anonymization process. Finding appropriate anonymization techniques for data that is considered 'inherently non-anonymous' is crucial.

Location data. Anomaly Six, a US government contractor, claims to be able to monitor the movements of billions of identified people by analyzing the exact GPS measurements gathered through covert partnerships with smartphone apps producers[2,3]. This also concerns cars and other mobile devices with built-in SIM cards. Currently, there is no option in smartphones to block the access to the location data to all apps installed at once. Most data subjects do not read the terms and conditions of smartphone apps and therefore do not realize that they are giving consent to be tracked. Due to the technical challenges of anonymizing location data and the risk of linking between different sources, major privacy issues arise. Even with anonymized location data, recent research has shown that more than 90% of people could be uniquely identified in a dataset of 60M people using four points of auxiliary information. The privacy of individuals is very unlikely to be preserved even in country-scale location datasets.[4]

Pseudonymization. Pseudonymization is a light reversible process that consists in keeping the mapping between the identifiers appearing in the pseudonymized dataset and the real persons. This mapping is under the control of a trusted entity.

In practice, rather than anonymizing data, companies prefer to apply pseudonymization and some generalization. In many cases, such pseudonymous data are considered as anonymous data. A classic example is where hashing is frequently mistaken as an anonymisation technique. It is to be noted that pseudonymous data remain personal. To make the de-pseudonymization and reidentification process more difficult, companies and cloud storage operators are increasingly using double pseudonymization.

Explainability and interpretability. To fulfil the transparency principle and the right to explanation, as well as the informed consent requirement, stipulated by the GDPR, the new AI Act, and international legal instruments, technical efforts should be made to enhance the development and use of white boxes in crucial fields, such as healthcare. The challenge here is to be transparent and explainable, without risking

[2] J. Frith, M. Saker. It Is All About Location: Smartphones and Tracking the Spread of COVID-19. Social Media + Society, vol. 6, nr. 3, 2020, https://doi.org/10.1177/2056305120948257

[3] https://theintercept.com/2022/04/22/anomaly-six-phone-tracking-zignal-surveillance-cia-nsa/?s=09

[4] A. Farzanehfar, F. Houssiau, Y.-A. de Montjoye. The risk of re-identification remains high even in country-scale location datasets. Patterns, vol. 2, nr. 3, 100204, 2021, https://doi.org/10.1016/j.patter.2021.100204

unwanted information leakage. For instance, when explainability is implemented as ‘explanation by example’, the example provided by the system must not reveal sensitive data. Furthermore, privacy-preserving techniques in AI, such as federated learning, should consider addressing explainability to avoid issues related to data biases and quality degradation. Explainability should be also enhanced in a human-centric way; for example, in healthcare, AI model predictions and explanations have to be understandable by medical professionals.

Blockchain and privacy. Preserving privacy in blockchains and other distributed ledger technologies is a challenge. Lately, “novel privacy-preserving solutions for blockchain based on crypto-privacy techniques are emerging to deal with those challenges and empowering users with mechanisms to become anonymous and take control of their personal data during their digital transactions of any kind in the ledger, following a Self-Sovereign Identity (SSI) model”[5]. In addition, the immutability of blockchains makes it difficult to comply with the provisions of data protection regulation, such as the right to be forgotten, stipulated by the GDPR.

Quantum computing. Quantum computers that perform operations on qubits will be much faster than classical computers. The strength of today’s cryptographic algorithms which are usually based on hard mathematical problems will be impacted by two fundamental quantum computing algorithms: Grover’s Algorithm and Shor’s Algorithm. Quantum computing may render many cryptographic algorithms insecure. This threat needs to be considered before the existence of commercial quantum computers, due to the long lifecycle of some products, and the threat that existing sensitive data protected by current cryptographic algorithms can be stored to crack the cryptographic protection in the future.

Artificial Intelligence and cybersecurity. The recent technological progress in Artificial Intelligence increases threats in at least two ways. First, AI technology can provide novel instruments for intentional attacks. And second, AI systems in operation, especially mission-critical AI systems, are of particularly high interest to attackers as they are of high economic or strategic value. In particular, data security, model security and system security are critical security aspects of AI. For example, training data for machine learning can be attacked by poisoning the data set to cause a harmful result for the AI system. Also, there are attacks from input path to the AI system, such as adversarial samples attack, membership inference attack etc. The training data set for AI is also a crucial asset for its business owners, and its leakage could impact the competency of the business owners.

Ransomware attacks. In recent years, ransomware attacks have become prevalent, causing organizations around the world huge problems[6]. In contrast to other types of malware that were dominant in the past, the target of ransomware attacks is not the machine, but the data hosted on the machine. One threat resulting from such attacks is that criminals expose the (personal) data when the company does not pay the ransom. An interesting property of ransomware attacks relates to the fundamental security properties confidentiality – integrity – availability (the CIA triad). In the past, data protection was mostly concerned with confidentiality (e.g., precluding eavesdropping) and integrity (e.g., precluding tampering). Ransomware attacks, on the other hand, often threaten the availability of data by making the data unavailable until ransom is paid.

User-centric data control. The increased value of the data generated by the citizens and the emergent opportunities associated with the data economy require that individuals have control on how these data are used, the final purpose, and how long the data can be used. In this regard, user consent is an important aspect to be considered. Ensuring data sovereignty becomes a crucial challenge for facilitating data sharing.

[5] J. B. Bernabe, J. L. Canovas, J. L. Hernandez-Ramos, R. T. Moreno, A. Skarmeta. Privacy-preserving solutions for blockchain: Review and challenges. IEEE Access, vol. 7, pp. 164908-164940, 2019, https://doi.org/10.1109/ACCESS.2019.2950872

[6] https://www.techtarget.com/searchsecurity/feature/Ransomware-trends-statistics-and-facts

2.2 Legal challenges

Interpretation of technology. Advanced technological solutions are available, but assessing a technological solution from a legal point of view is often challenging. For instance, when personal data is collectively analyzed using a Fully Homomorphic Encryption scheme, who is the controller? From a legal point of view, encrypted personal data are still personal data, and as such, they cannot be considered as anonymized data. But how to ensure that every processing performed over these encrypted data is compliant to the legitimate interest or user's consent? Answering such questions is difficult, but would be absolutely necessary to enable sound reasoning about the legal consequences of advanced technologies.

Data Act proposal. The new proposal of the Data Act grants users a new right to accessibility and portability of IoT (Internet of Things) data, both personal and non-personal, and introduces the possibility for public bodies to have access to that data in situations of public emergency. Because this entails the movement of a large amount of data, some of which may relate to individuals, privacy issues may arise. The Data Act states that manufacturers of IoT devices do not have exclusive rights over the data obtained from or generated by the use of a product or related service. It also mentions that the Commission has undertaken to review the 96/9/EC Database Directive so that the sui generis right should not apply to databases containing data obtained from or generated by the use of a product or a related service. The development of this new piece of legislation and its impact on data subjects should be monitored because of the potential data protection implications.

Digital Markets Act (DMA). The DMA introduces new requirements and obligations for gatekeepers, such as social networks and search engines, that is, companies with a market value of more than 75 billion euros or an annual turnover of 7.5 billion euros. The new regulation allows the processing of personal data for targeted advertising only with a user's explicit consent and it also prohibits gatekeepers "from combining personal data sourced from these core platform services with personal data from any other services offered by the gatekeeper or with personal data from third-party services, and from signing in end users to other services of the gatekeeper in order to combine personal data, unless the end user has been presented with the specific choice and provided consent in the sense of Regulation (EU) 2016/679".

However, security concerns have been raised regarding the forced interoperability and portability requirements[7]. This may impact data protection.

Digital Services Act (DSA). The new DSA aims at regulating online intermediary services by introducing many new obligations[8] to enhance transparency and stop illegal online content and malicious advertisement. According to the Explanatory Memorandum, "the measures concerning advertising on online platforms complement but do not amend existing rules on consent and the right to object to processing of personal data. They impose transparency obligations towards users of online platforms, and this information will also enable them to make use of their rights as data subjects. They also enable scrutiny by authorities and vetted researchers on how advertisements are displayed and how they are targeted". Article 31.2 provides that "Upon a reasoned request from the Digital Services Coordinator of establishment or the Commission, very large online platforms shall, within a reasonable period, as specified in the request, provide access to data to vetted researchers who meet the requirements in paragraphs 4 of this Article, for the sole purpose of conducting research that contributes to the identification and understanding of systemic risks as set out in Article 26(1)". Overall, the DSA has the potential to have significant impact on the data protection landscape.

Web scraping and illegal collection of personal data. Lately, authorities have faced the legal issue of web scraping, that is, the large-scale acquisition of personal data from websites that are publicly available, such

[7] https://www.neweurope.eu/article/the-digital-markets-act-is-a-security-nightmare/
[8] https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountable-online-environment_en

as LinkedIn. Different countries have ruled in a different way depending on their legal system. The District Court for the Northern District of California in the US issued in 2019 a preliminary injunction that ordered LinkedIn to stop blocking hiQ Labs from scraping information from LinkedIn member profiles. The company was using publicly available LinkedIn data to create reports for its customers, in order to help them identify which of their employees were most likely to resign or to be targeted by recruiters[9]. In 2019, the Court ruled in favor of hiQ Labs, therefore Microsoft appealed to the Supreme Court, which dismissed the case but ordered the Court to reconsider the case. In April 2022, the Court upheld its previous decision in favor of hiQ Labs. This ruling poses a risk for EU citizens who use LinkedIn, as US companies may collect their data without their knowledge and use them in a way that would be considered illegal under the GDPR. In previous years, other similar cases were brought to court in the US by citizens that were not consenting to web scraping of their data: the case of IBM scraping faces on Flickr[10] and ClearviewAI scraping photos on Facebook[11]. In the latter case, ClearviewAI was recently significantly fined by EU authorities[12,13], which ordered the company to destroy the data.

NIS 2 Directive (in progress)[14] is going to stipulate cybersecurity risk management and reporting obligations for essential and important entities and "rules and obligations on cybersecurity information sharing". It also requires that "the exchange of information shall preserve the confidentiality of that information and protect the security and commercial interests of essential or important entities". The challenge will be in the future detailed procedure and protection measures, which should be carefully designed, to avoid causing significant damage to the industry. For example, any natural or legal person may report, possibly anonymously, a vulnerability to the designated member state CSIRT (Cyber Security Incident Response Team). This may happen before the manufacturer starts to give remedies or develop patches to fix the vulnerability, so there needs to be a very high security standard on how the member state CSIRT obtains, coordinates and processes the vulnerability information from reporters.

eIDAS regulation under revision[15]**.** The objective of the revised version of the eIDAS regulation is to provide at least 80% of European citizens with an identity wallet by 2030 and to extend the usage of the wallets to any online and offline services, whether administrative or commercial. Although at its early stage, eIDAS 2.0 is questioning the robustness of the wallet against possible data leakages or corruptions, Member States' sovereignty and users' freedom to present themselves as they wish over the Internet. Indeed, with valuable personal data inside, e.g., civil status and diploma, wallets may become the target of numerous attacks, so there is a strong need to build a certification methodology to evaluate wallet implementations and to guarantee the adequate security measures to meet the challenges. Beyond the security dimension, as a matter of sovereignty, there is a challenge for Member States to carefully select the stakeholders that are involved in the design, integration, operation, maintenance, governance, audit etc. of the wallets. Finally, it is important, for most Internet usages – when high level identification/authentication is not necessary with real identities – to let users freely decide the identity that best represents them according to the context, and as such to keep the alternative of pseudonymity.

2.3 Other important challenges

Trust. Privacy Enhancing Technologies are notoriously hard to explain to laypersons, since they often involve advanced mathematical operations and intricate technical concepts. Therefore, it can be hard to present solutions in such a way that the public has trust in their operation. For example, in the case of medical data, even though it is possible these data can only be used for predetermined analyses and can

[9] https://www.forbes.com/sites/zacharysmith/2022/04/18/scraping-data-from-linkedin-profiles-is-legal-appeals-court-rules/

[10] https://www.nbcnews.com/tech/internet/facial-recognition-s-dirty-little-secret-millions-online-photos-scraped-n981921

[11] https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html

[12] https://www.bbc.com/news/technology-61550776

[13] https://edpb.europa.eu/news/national-news/2022/facial-recognition-italian-sa-fines-clearview-ai-eur-20-million_en

[14] https://digital-strategy.ec.europa.eu/en/library/proposal-directive-measures-high-common-level-cybersecurity-across-union

[15] https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0281&from=EN

be guaranteed to be untraceable towards the patient, how can the patient really be convinced that their data will *not* be shared with, say, their employer or health insurance company?

Profiling. Although the GDPR specifically limits the allowed possibilities of profiling, the incentives for extensive profiling are growing. The Cambridge Analytica scandal clearly showed some of those incentives, as well as the danger of profiling, especially on social media, up to the point that even democracy is in danger. This shows how important the GDPR's provisions on profiling are, and how challenging it is to enforce them.

Automated Decision-Making (ADM) systems. As highlighted by a number of NGOs[16] and activists, ADM is becoming more and more popular, especially in the public sector. This poses new challenges and risks to freedoms and rights of individuals. Recent cases, such as SyRi in the Netherlands, showed the dangers of an indiscriminate use of personal data for ADM. Similarly to profiling, also automated decision-making is strongly limited by the GDPR, but the recent trends show that these limitations are not enforced effectively.

Risk governance. The International Risk Governance Council (IRGC) pointed out that technological advances may lead to emerging risks when there are no "appropriate prior scientific investigations or post-release surveillance of the resulting public health, economic, ecological and societal impacts."[17] Such advances are ahead in human-systems integration concepts like Cobots, Hybrid Intelligence, and in the context of autonomous vehicles. Data protection will have to address the linking of personal data with machine data and the related conflict regarding data sovereignty, i.e., who can use what data and how.

Data minimization. The "principle of data minimization" means that a data controller should limit the collection of personal information to what is directly relevant and necessary to accomplish a specified purpose. AI and other digital systems are sustainable in terms of data protection if they rest on deliberate, proactive assessment of minimal data to collect. This requires extensive domain expertise to combine with deep data and algorithmic knowledge. Hence the need for hybrid approaches and their investigation, e.g., as to the validity and performance of the human-machine interaction.

Data disclosure. Personal and other sensitive data must be protected at any time. The disclosure risk of this kind of data implies avoiding unauthorized data disclosure and also minimizing the amount of data to be revealed under the user consent. Mechanisms for preventing unauthorized disclosure, minimal data disclosure such as selective disclosure, zero knowledge property or Attribute Based Access Control (ABAC) must be taken into account for addressing these risks. The challenge is to provide mechanisms assuring minimal data disclosure while keeping the utility of disclosed data.

Public awareness of data protection issues. Since the number of different accounts and associated passwords for an average internet user is still increasing[18], and password-less login scenarios are still rare, there is a growing risk of data leakage. Cyber criminals are exploiting the common strategy of many internet users to reuse passwords across different accounts. Public awareness of data protection is still too low, allowing cyber criminals to maintain their modus operandi. This is especially important when looking at user-centric data control (as introduced above) and the eIDAS regulation.

3 PROMISING TECHNOLOGIES

This section reviews technologies that have been proposed and seem promising for maintaining or improving data protection in the future.

[16] https://algorithmwatch.org/en/

[17] International Risk Governance Council. The Emergence of Risks: Contributing Factors. International Risk Governance Council, 2010, ISBN 978-2-9700672-7-6, https://irgc.org/wp-content/uploads/2018/09/irgc_ER_final_07jan_web.pdf

[18] https://www.lastpass.com/resources/ebook/psychology-of-passwords-2020

3.1 Synthetic data

Synthetic data generation is the process to produce datasets that do not contain any data of real persons, but that still have the statistical features that are characteristic of real-life data. Typically, machine learning is used to learn the features of the original, real dataset, and then to produce data that, when it is well done, is indistinguishable from real data. Often, Generative Adversarial Networks are applied for this purpose: one AI system (the generator) tries to generate data while another AI system (the verifier) is trained to distinguish between real data and fake data. The generator learns how to adapt its data generation until the verifier system starts to fall for the generator's tricks. Synthetic data generation has made significant progress in recent years; intriguing examples can be found on www.thisxdoesnotexist.com. However, there are still many challenges; for instance to create synthetic data for tabular data, texts (despite impressive results reported by GPT-3), or to create multiple datasets that are consistent with each other.

3.2 Cryptographic solutions

Secure multiparty computation. Different technologies exist that enable parties to collectively analyze data, without sharing the data in a readable form. At any time during the exchange of data between parties, it is impossible for anyone to extract information from that data. Relevant technologies include Secret Sharing and Homomorphic Encryption. Carefully designed cryptographic solutions enable specific computations based on encrypted data. The available technologies have different limitations and constraints, for example only one type of computation is possible (e.g., adding two numbers), or specific conditions have to be met (e.g., at least three parties are taking part in the computation). Some of the technologies can be mathematically proven to guarantee some security properties, potentially offering strong privacy guarantees.

Post-quantum cryptography. Migration from existing cryptographic algorithms to Post Quantum Cryptography (PQC) algorithms in the standard development process will be an effective way to address the quantum computing challenges (cf. Section 2.1). Quantum computers capable of cryptographic applications are expected around 2033[19], and some security-critical products have a long lifecycle that can go beyond 2033. Thus, it is recommended to identify high-priority scenarios, and then start to design and implement PQC capabilities (both hardware and software) in order to support a smooth migration to PQC algorithms in the future, starting with the identified high-priority scenarios as soon as possible.

3.3 Federated learning

Federation is based on the general concept of decentralization, aiming to prevent central concentration of power. In the context of AI, federated learning (FL) and/or federated data spaces are considered suited to address the issue of data protection: distributed data spaces under control of different actors instead of central accumulated data pools under single actor control. Federated machine learning that builds upon the use of data from distributed, independent sources constrains the risk of undesired exploitation of assembled information. In recent years, there has been significant scientific progress in federated learning, enabling learning from data held by different actors without the need to share those data with the other participants. The security and privacy properties of federated learning have been assessed, potential attacks identified, and counter-measures against those attacks developed. Issues that remain to be resolved include accessibility, interoperability, communication, and quality control.

Moreover, to realize FL it is crucial to provide a solution which mitigates the black-box nature of AI models integrated in the distributed learning setting. The data quality, bias, standardization and related statistics or key features for a successful model, remain unseen from the central server and clients. This makes it

[19] M. J. D. Vermeer, E. D. Peet. Securing Communications in the Quantum Computing Age: Managing the Risks to Encryption. RAND Corporation, 2020, https://www.rand.org/content/dam/rand/pubs/research_reports/RR3100/RR3102/RAND_RR3102.pdf

unclear whether the model updates of certain data populations from local sites are relevant and optimal for each participant. This is currently missing in FL because data from other parties are private. Furthermore, local sites cannot retrieve any information related to data quality, such as explanation and interpretability from the local and global models.

Finally, applications of federated learning for solving challenging real-world problems (e.g. clinical needs in healthcare) are lacking in use cases that showcase the unique advantage of the approach in different domains. Once these issues are adequately addressed, the wide adoption of FL may lead to higher trust and increased real-world adoption of AI applications.

3.4 Identity management

In the current digital world where the number of smart devices, applications and online services is steadily rising, personal and sensitive data are often shared with third parties. Citizens' concern about trust and data privacy issues are increasing as well. Unlike centralized or federated identity management models, the Self-Sovereign Identity (SSI) approach allows the data owner to take control of their own identity and data when sharing with a third party. SSI can be implemented through the identity wallet which is expected to be regulated by the revised version of the eIDAS European regulation[20], known as eIDAS 2.0. SSI can rely on a decentralized infrastructure, for instance based on blockchain technology. SSI gives data owners the possibility to decide what data is delivered, who can access this data, the purpose for using the data, and how long the data can be used. SSI has also the capability to certify to third parties the authenticity of the citizen's identities like a social security ID number, a Regalian identity, a professional identity etc. This requires that some identity issuers are recognized as trustful in the identity management system for issuing related identities, e.g., National Health Service issuing the social security numbers. The certified data are provided to the third party through signed verifiable credentials, cryptographically secure, privacy-respecting, and machine-verifiable[21], controlled by the user. SSI helps to comply with the GDPR and to protect data particularly when sensitive data are shared.

3.5 Data protection engineering

Usage control. Usage control[22] was designed as a more sophisticated access control, providing users with the technical basis to monitor the usage of their data. Usage control, in addition to authorizations based on users' attributes and rights, introduces obligations and conditions. Obligations are actions to be performed to be granted access, such as accepting cookies to enter a web page. Conditions are related to the system and are not under direct control of the users, like the time of the day. Usage control is useful in dynamic environments where access must be continuously monitored according to complex criteria. As an example, the diversity of IoT data and devices requires fine-grained policies. Post-authorizations and post-conditions can be used to monitor the user after the access has been granted, which is not possible with classic access control. Besides, it helps to comply with the GDPR by enforcing continuous control of the data controller's obligations.

Information flow control. Commonly used with usage control, information or data flow control[23] is a mechanism designed to monitor the dissemination of the data to other users. It prevents copying the data elsewhere and processing them, or sharing the data with unauthorized users, thus avoiding data breaches. Monitoring the information flow can be an intrusive process as it requires monitoring the system and network calls of the monitored users. It is therefore necessary to rely on *confidential computing* to ensure

[20] Proposal for a Regulation of the European Parliament and of the Council amending Regulation (EU) No 910/2014 as regards establishing a framework for a European Digital Identity, COM(2021) 281 final 3.6.2021, https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0281&qid=1628524321657

[21] https://www.w3.org/TR/vc-data-model/

[22] J. Park, R. Sandhu. The UCON$_{ABC}$ usage control model. ACM Transactions on Information and System Security, vol. 7, nr. 1, pp. 128-174, 2004, https://doi.org/10.1145/984334.984339

[23] A. C. Myers, B. Liskov. A Decentralized Model for Information Flow Control. ACM SIGOPS Operating Systems Review, vol. 31, nr. 5, pp. 129-142, 1997

the network and system calls are processed privately, e.g. using *Trusted Execution Environment* (see next paragraph).

3.6 Confidential computing

Various types of encryption techniques can be used to protect data at rest (i.e., data stored in files or databases) and data in transit (i.e., data being transmitted over the network). However, protecting *data in processing* (i.e., data that is being processed by a CPU, GPU, or other processing element) is challenging. With traditional encryption schemes, the data must first be decrypted before it can be processed, making it vulnerable to security breaches. There are promising novel encryption schemes that allow some kind of processing of the encrypted data without first having to decrypt it (e.g., homomorphic encryption), but such encryption schemes are often not yet practical, especially because of their computational overhead. Confidential computing offers a different solution for protecting data in processing. Confidential computing uses hardware-based *trusted execution environments*: processors that guarantee certain security features for the memory or for parts of the memory[24]. For example, a process may create a *secure enclave* in RAM, and the processor guarantees that no other process can access the data in the secure enclave, not even processes of the operating system. Thus, using confidential computing, the *trusted computing base* can be significantly reduced. Leading chip manufacturers increasingly offer confidential computing solutions.

3.7 Artificial intelligence

As described in Section 2.1, the recent uptake of artificial intelligence (AI) techniques leads to data protection challenges. On the other hand, various AI techniques can also be used to enhance data protection. An important example is the already mentioned use of generative adversarial networks to generate synthetic data (see Section 3.1). Another example is the use of automated reasoning techniques for automatically identifying threats to data protection in a running system, or even automatically devising mitigation measures to the found threats[25].

3.8 Privacy measures

The utilization of massive data in the context of digitization affects both the origin of data and the control of data. Thus, both data privacy and data sovereignty need to be increasingly considered. Data protection must address both while drawing on the quality of data at the same time. Regulative provisions set the frame respectively, as does the European Data Innovation Board for the ongoing monitoring and assessment of data protection measures. In this context, also with view on a European certification for AI that is being considered, there is an increasing need for quantitative methods to evaluate data protection. For this purpose, a set of KPIs (key performance indicators) aiming at the assessment of the data protection level of an enterprise or an application should be developed and made available to the public. In recent years, many different privacy KPIs have been proposed and evaluated, paving the way to a better understanding of the usefulness and expressiveness of the different KPIs. In this context, the use of privacy metrics tools to measure the degree of data privacy in a system, helps to understand how the users' data privacy is protected.

[24] https://confidentialcomputing.io/

[25] Z. Á. Mann, F. Kunz, J. Laufer, J. Bellendorf, A. Metzger, K. Pohl. RADAR: Data protection in cloud-based computer systems at run time. IEEE Access, vol. 9, pp. 70816-70842, 2021, https://doi.org/10.1109/ACCESS.2021.3078059

4 CONCLUSIONS AND RECOMMENDATIONS

In this position paper, we have reviewed current challenges and promising solutions for data protection. Looking forward, we provide a list of recommendations for different stakeholder groups in the following table.

Stakeholder group	Relevant issues	Recommendations
Legal authorities	Are current regulations respected? How to ensure regulations are aligned with the needs of all stakeholders? How to ensure end-users are well-informed of risks and solutions?	Monitor how regulations are understood and applied. Co-design regulations with all stakeholders to ensure both industrial innovation and data protection. Raise awareness of end-users on risks and solutions (e.g., web scraping by non-European companies).
Industry	How to innovate while protecting data? How to ensure high level of data protection?	Assess suggested solutions thoroughly. Enforce users' trust with relevant data protection solutions by relying on multiple security/privacy layers for higher protection.
Public administration	How can public administrations comply with current regulations on data protection (e.g., GDPR) and guarantee the rights of citizens when processing personal data? How to make the right decisions in terms of technological orientations and uses to build a digital society that is trustworthy, secure and respectful of individuals and their freedom (e.g.: Self-sovereign identities)?	Work on improvement of aspects for assessing, managing and minimizing risks to rights and freedoms. Identify the risks associated with the processes and services managing the data of citizens. Listen to independent and multidisciplinary sources to clarify all issues before making decisive orientations for the digital society.
Research institutions and universities	How to accompany technological progress on distributed ledgers, data sanitization, AI, anonymization and pseudonymization, secure sharing etc.? How to provide relevant analysis on risks to data privacy and state-of-the-art solutions for data protection?	Research on relevant issues, contribute to the definition of regulation challenges as regards new technologies, e.g., blockchains. Experiment with current technologies and give feedback to all stakeholders.

Stakeholder group	Relevant issues	Recommendations
Individuals	How to stay informed on risks and solutions for data protection? How to have tools answering needs?	Follow guidelines provided by legal authorities. Reinforce public awareness campaigns that address data protection.

Overall, our review of current hot topics in data protection gives rise to the following main conclusion. The rapid development of information technology leads to ever improving means to both protect data and attack data. Therefore, data protection is an activity of sustained vigilance, in which all stakeholder groups have an important role to play.

As data spaces are the cornerstone of the European Data Strategy for creating a single European data market, the challenges and technologies highlighted through this position paper also give valuable insights into the data spaces ecosystem. As indicated, all the stakeholders involved in data spaces – citizens, public and private organizations, industry, and policy makers – need to take into consideration the uptake of the indicated technologies for overcoming the described challenges, to ensure the continued protection of data, which is a prerequisite for the uptake of data spaces. The appearance of new technologies and unforeseen challenges implies that companies (both SMEs and large enterprises) and regulators need to be flexible enough to adapt to new scenarios. The focus on relevant aspects such as data sovereignty, and the creation of a secure and privacy-preserving ecosystem where data are protected, will be the basis for building a trusted environment for sharing data.

ABOUT THIS DOCUMENT

Editor: Zoltán Ádám Mann (University of Amsterdam, Netherlands)

Contributors:

- Freek Bomhof (TNO, Netherlands)
- Sophie Chabridon (Télécom SudParis, France)
- Nathanaël Denis (Télécom SudParis, France)
- Chiara Gallese Nobile (Eindhoven University of Technology, Netherlands)
- Norbert Jastroch (MET Communications, Germany)
- Maryline Laurent (Télécom SudParis, France)
- Zoltán Ádám Mann (University of Amsterdam, Netherlands)
- Juan Carlos Pérez Baún (Atos, Spain)
- Haibin Song (Huawei, Germany)

Reviewers:

- Aindrias Cullen (Trilateral Research, Ireland)
- Francesca Manni (Philips, Netherlands)
- Robert Seidl (Nokia, Germany)
- Rob Smeets (Philips, Netherlands)

This paper has been produced under the **BDVA Task Force 6 Subgroup 4 – Data Protection Technologies**.

Note: This document should be referenced as follows:

Bomhof, F., Chabridon, S., Denis, N., Nobile, C. G., Jastroch, N., Laurent, M., Mann, Z., Á., Pérez Baún, J.C, Song, H. Current hot topics in data protection. BDVA. 2022

ABOUT BDVA

The Big Data Value Association is an industry-driven international not–for-profit organisation with more than 230 members all over Europe and a well-balanced composition of large, small, and medium-sized industries as well as research and user organizations.

BDVA focuses on enabling the digital transformation of the economy and society through Data and Artificial Intelligence by advancing in areas such as big data and AI technologies and services, data platforms and data spaces, Industrial AI, data-driven value creation, standardisation, and skills.

BDVA has been the private side of the H2020 partnership Big Data Value PPP, it is a private member of the EuroHPC JU, it is also one of the founding members of the AI, Data and Robotics Partnership and a partner in the Data Spaces Business Alliance. BDVA is an open and inclusive community and is always eager to accept new members who share its ambitious objectives.

Contact for further information: info@core.bdva.eu

BDV BIG DATA VALUE ASSOCIATION

www.bdva.eu