Acta Univ. Sapientiae, Legal Studies, **11**, 2 (2022) 35–64

DOI: 10.47745/AUSLEG.2022.11.2.03

Mitigating the Privacy Risks of AI through Privacy-Enhancing Technologies[1]

Barnabás SZÉKELY
LL.M, Certified Data Protection Officer
PrivacyPro Ltd. (Cluj-Napoca, Romania)
e-mail: szekelybarnabas@gmail.com

Abstract. The development and operation of an AI solution generally requires large amounts of data. This may involve processing of personal data, which implies privacy risks for the data subjects and the obligation to comply with data protection rules for data controllers. Privacy-enhancing technologies (PETs) can help enhance data collection and mitigate privacy risks posed by the development of AI solutions. In this context, this thesis proposes to present a set of emerging technologies that address privacy risks characteristic to machine learning models and enable privacy-preserving machine learning. The essay will highlight three state-of-the-art PET solutions: homomorphic encryption, secure multi-party computation, and differential privacy.

Keywords: artificial intelligence, data protection, privacy, European Union, differential privacy

1. Introduction

Artificial intelligence (AI) is becoming a key element for digital transformation, shaping the future of humanity in almost every industry and evolving at an accelerating pace. According to *The One Hundred Year Study on Artificial Intelligence* 2021 Study Panel Report led by Stanford University, the field of AI has made remarkable progress in almost all its standard sub-areas between 2016 and 2021. This includes vision, speech recognition, natural language processing, image and video generation, multi-agent systems, planning, decision-making, and integration of vision and motor control for robotics.[2] The speed of development can be linked to the recent advances in computing power and the increasing availability

1 The following study constitutes the first publication, in abridged form, of the author's LL.M dissertation submitted in the course of the Master's Programme in Artificial Intelligence for Public Services of the Madrid Polytechnic University (2022).

2 Stanford University – Littman–Ajunwa–Berger–Boutilier–Currie–Doshi-Velez–Hadfield–Horowitz–Isbell–Kitano–Levy–Lyons–Mitchell–Shah–Sloman–Vallor–Walsh 2021.

of vast swathes of data, also boosted by the evolution of AI investments. In 2019, investment in AI in the European Union (EU) grew by 64%, then by 37% in 2020, and the overall level of AI investments was estimated to reach €10.7 billion. If this trend is maintained, the EU will exceed its annual AI investment target of €22 billion by 2030. In the United States, the growth was 55% in 2019 and 50% in 2020, reaching €21.2 billion. On the contrary, in the United Kingdom, investment in AI grew at a higher rate in 2020 (46%) than in 2019 (40%).[3]

Through new products and services, AI is increasingly present in our daily lives. Besides the innovation, opportunities, and potential value to society, AI systems also pose a potential risk to the fundamental rights, health, and safety of citizens. Discrimination, privacy and data protection harms (for example, loss of confidentiality), lack of transparency, explainability and accountability became intensely discussed and debated issues of AI systems. As the High-Level Expert Group on Artificial Intelligence (AI HLEG) set up by the European Commission in June 2018 emphasizes in its Ethics Guideline on Trustworthy AI, privacy is a fundamental right particularly affected by AI systems.[4] The privacy implications of AI depend to a large extent on the specific use cases, the sensitivity of the training data, the social groups the system is deployed on, the overlapping legal requirements,[5] and social, cultural, and political aspects. In the second chapter of this thesis, titled *Artificial Intelligence vs. Privacy*, the main privacy and data protection risks raised by AI systems will be explored.

The development and use of AI systems often involve the processing[6] of personal data.[7] The General Data Protection Regulation (EU) 2016/679 (General Data Protection Regulation, GDPR) is built in a technologically neutral manner and does not refer specifically to AI. In order to be able to face any technological evolution, GDPR regulates the processing of personal data regardless of the technology used,

3 European Commission Joint Research Centre – Evas–Sipinen–Ulbrich et al. 2022.

4 European Commission Directorate-General for Communications Networks, Content and Technology 2019.

5 The European Data Protection Board and the European Data Protection Supervisor (EDPS) have raised their concern that 'the (AI Act) Proposal is missing a clear relation to the data protection law as well as other EU and Member States law applicable to each "area" of high-risk AI system' listed in the Annex III of the Regulation. Joint Opinion 5/2021 on the proposal for a Regulation of the European Parliament and of the Council laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) 2021.

6 Article 4(2) GDPR: any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organization, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction.

7 Article 4(1) GDPR: any information relating to an identified or identifiable natural person (data subject). An identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural, or social identity of that natural person.

as the European Data Protection Board (EDPB) highlighted in a response to Sophie in't Veld's (Member of the European Parliament) letter on unfair algorithms. Also, the EDPB commented that any processing of personal data through an algorithm falls within the scope of the GDPR.[8] In conclusion, whenever the processing of personal data performed by an AI system falls within the territorial scope[9] of the GDPR, all provisions of the Regulation will apply to such processing. In the second chapter, the main challenges posed by GDPR requirements will be presented.

The GDPR, similarly to the Artificial Intelligence Act (AI Act),[10] applies a preventive risk-based approach. The basis of this approach is the 'data protection by design and by default principle' (DPbDD).[11] Data protection by design[12] requires the implementation of appropriate organizational and technological measures prior to and during the whole lifecycle of data processing activities. This ensures that privacy and data protection risks are identified and addressed in the early stages of the AI system's lifecycle, also that the data protection principles[13] and necessary safeguards are embedded in the AI system's entire lifecycle. Implementing these principles at a systemic level and ensuring that, by default, only personal data which are necessary for each specific purpose of the processing are processed requires new technical approaches. The technologies that are designed to support the implementation of data protection principles are covered by the term Privacy-Enhancing Technologies (PETs). According to Borking and Raab, PETs 'are a coherent system of ICT measures that protects privacy by eliminating or reducing personal data or by preventing unnecessary and/or undesired processing of personal data, all without losing the functionality of the data system'.[14] The EDPB underlines that 'PETs that have reached the state-of-the-art maturity can be employed as a measure in accordance with the DPbDD requirements if appropriate in a risk-based approach' but in themselves do not necessarily satisfy the obligations under GDPR Art. 25 on data protection by design.[15] In the third chapter, an overview of emerging PETs that address the most common data security risks and challenges posed by big data and AI developments will be provided.

8 EDPB 2020b.

9 Article 3 of Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data and repealing Directive 95/46/EC (General Data Protection Regulation) 2016.

10 Proposal for a Regulation of the European Parliament and of the Council laying down harmonized rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts 2021.

11 Article 25 GDPR.

12 The concept was developed in the 1990s, brought forward by Ann Cavoukian, former Information and Privacy Commissioner of Ontario. At that time, the term Privacy by Design (PbD) was used.

13 Article 5 GDPR.

14 Borking–Raab 2001.

15 EDPB 2020a.

PETs are a promising set of techniques that can support privacy-preserving machine learning (PPML), facilitating the use of a powerful form of data analysis.[16] By 2025, 60% of large organizations will use one or more privacy-enhancing techniques in analytics, business intelligence, or cloud computing – as Gartner predicts.[17] In the third chapter of this thesis, three promising areas of PETs will be analysed in depth: homomorphic encryption, secure multi-party computation, and differential privacy. Also, real-world use cases will be discussed to demonstrate how these PETs contribute to privacy-preserving machine learning.

2. Artificial Intelligence vs. Privacy

As the penetration of AI is increasing, a growing number of sectors are transformed. Besides the benefits of AI, specific privacy and data protection risks arise in the case of AI systems that process large datasets of personal data or combine non-personal data that can lead to the re-identification of individuals.

2.1. Privacy and Data Protection Risks

Depending on the particular context, varying likelihood and severity of the risks, personal data processing could lead to physical, material, or non-material damage.[18] In recital 75, GDPR addresses among the potential consequences a broad range of harms and emphasizes that the processing of personal data may give rise to 'discrimination, identity theft or fraud, financial loss, damage to the reputation, loss of confidentiality of personal data protected by professional secrecy, unauthorized reversal of pseudonymization, or any other significant economic or social disadvantage'. The following subchapter will provide an overview of privacy risks in relation to compliance with data protection principles.

2.2. AI Meets the Data Protection Principles

The fundamental building blocks of the GDPR are the seven key data protection principles:[19]

a. Lawfulness, fairness, and transparency;
b. Purpose limitation;
c. Data minimization;
d. Accuracy;

16 The Royal Society 2019.
17 Gartner 2021.
18 Recital 75 GDPR.
19 Article 5 GDPR.

e. Storage limitation;
f. Integrity and confidentiality (security);
g. Accountability.

Regardless of the purposes and means of personal data processing, compliance with these principles is an essential requirement. Failure to comply with the principles at the heart of the GDPR can result in heavy fines. Infringements on the principles are subject to the highest tier of administrative fines, meaning financial penalties of up to €20 million or 4% of the total worldwide annual turnover, whichever is higher.[20] The principles that have particular relevance to AI systems are discussed in detail below.

2.2.1. Fairness

According to the EDPB, 'fairness is an overarching principle which requires that personal data shall not be processed in a way that is detrimental, discriminatory, unexpected or misleading to the data subject'.[21] Fair data processing presumes that data have not been collected or processed through unfair means, without the data subject's knowledge, or by misleading or deception of data subjects. Also, fairness implies that data is processed in ways that data subjects would reasonably expect and the continuous assessment of how the processing affects the interests of individuals.[22]

Fair processing requires that AI systems do not produce discriminatory effects. The AI HLEG's guidelines quoted earlier draw the attention that 'data sets used by AI systems (both for training and operation) may suffer from the inclusion of inadvertent historic bias, incompleteness and bad governance models.'[23] A freshly published study on auditing the quality of datasets used in algorithmic decision-making systems is pointing out that 'the possibility of obtaining biased AI outcomes is strongly related to the characteristics of the data and the quality of the data management process, including data gathering, cleaning, annotation and processing'.[24] If the datasets used for training are unbalanced or biased, the system may produce outputs that have discriminatory effects on individuals without objective justification. In order to mitigate these risks, high-quality training and testing data are necessary that are representative of the population the AI system is deployed on.

Frequent incorrect outputs of the AI systems would also breach the fairness principle. The performance of the trained model should be measured by statistical

20 Article 83(5) GDPR.
21 EDPB 2020a.
22 Information Commissioner's Office – The Alan Turing Institute 2020.
23 European Commission Directorate-General for Communications Networks (AI HLEG) 2019.
24 Panel for the Future of Science and Technology – European Parliamentary Research Service 2022.

accuracy measures such as precision, recall, accuracy, and F1 score. In some cases, where high-quality test data is unavailable or the output is subjective, measuring statistical accuracy would not be appropriate.

PETs such as secure multi-party computation (SMPC) and federated learning (FL) can facilitate compliance with the fairness principle. The technologies can be used to prevent and restrict data usage for purposes that negatively impact an individual. These technologies will be discussed in detail in the third chapter, titled *Privacy-Enhancing Technologies*.

2.2.2. Transparency

This requirement was explicitly included in the data protection legislation for the first time by the authors of the GDPR. Information about how personal data is collected, used, consulted, or otherwise processed should be transparent, easily accessible, and easy to understand. GDPR highlights that individuals should be made aware of implications, risks, safeguards, and rights in relation to the processing of personal data.[25] Recital 60 adds that data controllers[26] 'should provide the data subject with any further information necessary to ensure fair and transparent processing taking into account the specific circumstances and context in which the personal data are processed'.

In the case of AI-assisted decision-making, it presumes to provide meaningful information about the logic involved, the significance, and the envisaged consequences of the AI decision. Where the decision is based solely on automated processing, information should be provided also about the right to object and the right to obtain human intervention.[27] This brings to life the obligation to explain the technical logic or reasoning behind a particular output of the AI system and the related human decision for AI-assisted decision making. The AI HLEG stated that 'technical explainability requires that the decisions made by an AI system can be understood and traced by human beings'. Also, they project that trade-offs might have to be made between enhancing a system's explainability and increasing its accuracy.

Transparency, human agency and oversight, and accountability play an important role as three key principles for trustworthy AI. AI HLEG underlines that if 'an AI system has a significant impact on people's lives, it should be possible to demand a suitable explanation of the AI system's decision-making process'. Also, 'the explanation should be timely and adapted to the expertise of the stakeholder concerned'.[28]

25 Recital 39 GDPR.

26 Article 4(7) GDPR: the natural or legal person, public authority, agency or other body that, alone or jointly with others, determines the purposes and means of the processing of personal data.

27 Article 13(2) b) and f) GDPR.

28 AI HLEG 2019.

Explainability relies on the level of interpretability, which depends on the model or set of models used by the AI system. For example, the usage of support vector machine (SVM) models may result in low levels of interpretability. 'Black box' techniques such as artificial neural networks (ANNs) can produce very low levels of interpretability. This may also apply to random forest models in some cases.[29]

2.2.3. Purpose Limitation

The purpose limitation principle is considered the cornerstone of data protection and is strongly linked to other data protection requirements. Purpose limitation requires data to be collected for specified, explicit, and legitimate purposes and not further processed in a manner that is incompatible with those purposes. The principle implies that purposes for processing personal data should be determined from the outset of the data processing lifecycle, at the time of the collection of the personal data.[30] Processing data only for the purposes defined beforehand could be challenging for AI systems because the purpose may change as the model learns and develops.

Data supposed to be used as training data is frequently collected originally for other purposes. If the latter data processing purposes are incompatible with the original purpose, then the purpose limitation principle could be a barrier to the development of an AI system. When assessing if the purpose of further processing is compatible with the purpose for which the personal data was initially collected, the data controller should take into account the following:

> [A]ny link between those purposes and the purposes of the intended further processing; the context in which the personal data have been collected, in particular the reasonable expectations of data subjects based on their relationship with the controller as to their further use; the nature of the personal data; the consequences of the intended further processing for data subjects; and the existence of appropriate safeguards in both the original and intended further processing operations.[31]

Recital 50 states that irrespective of the compatibility of the purposes, further processing should be allowed if the data subject has given consent or the processing is based on Union or Member State law. The latter could apply when the processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority. Also, further processing of data is considered to be compatible with the original purpose if it takes place in connection with scientific or historical research or for statistical and archival purposes in the public interest.

29 Information Commissioner's Office – The Alan Turing Institute 2020.

30 Working Party was set up under Article 29 of Directive 95/46/EC (WP29), Opinion 03/2013 on purpose limitation.

31 Recital 50 GDPR.

Scientific research purposes should be interpreted in a broad manner, including, for example, technological development and demonstration, fundamental research, applied research, and privately funded research.[32] In some cases, the development of artificial intelligence may be considered to constitute scientific research, but this presumes that the model learns something new – identifying trends or correlations – from the processed personal data.[33]

The data strategy of the European Commission encourages data exchange between the public sector and businesses, and reuse for data-driven innovation.[34] However, the complexity and uncertain application of the purpose limitation principle could hinder the reuse of personal data for AI-related technologies.

Similarly to the fairness principle, secure multi-party computation and federated learning are the two PETs which can play a role in satisfying the requirements of the purpose limitation principle.

2.2.4. Data Minimization

This principle requires identifying and processing the minimum amount of relevant and adequate personal data necessary to fulfil purposes. Minimization is referenced as an organizational measure for data protection by design and by default.[35] As AI systems generally involve the processing of large amounts of data, particularly during the training phase, at first sight, it may seem incompatible with the minimization principle. In some cases, it is also impossible to determine from the beginning which features of the training data may be relevant.

Data minimization in practice means preventing excessive data collection and using only the data necessary for the purposes of the processing. Instead of limiting data processing to a specific volume of data, it limits 'the amount of detail included in training or in the use of a model'.[36] The level of accuracy of the AI systems' output is the main factor in reducing the available data to the subset included in the final AI model.

Data minimization requires extensive testing. The predictive model built by the Norwegian Tax Administration that helps identify risk errors of tax returns is considered a good example of best practice. From the tested five hundred features, only the thirty that proved the most relevant were used.[37]

Data minimization is also about minimizing the risks of processing. Data controllers developing or using AI systems should assess the impacts on data

32 Recital 159 GDPR.
33 Norwegian Data Protection Authority (Datatilsynet) 2018.
34 European Commission 2020.
35 Article 25 GDPR.
36 Norwegian Data Protection Authority (Datatilsynet) 2018.
37 Ibid.

protection by performing a data protection impact assessment (DPIA)[38] and 'consider how to achieve the objective in a way that is least invasive for the data subjects'.[39] Risk reduction can be reached by reducing the degree of identification by perturbation, adding 'noise', or anonymization. Also, PETs such as synthetic data generation, federated learning, and differential privacy can be effective solutions for data minimization. The last one will be further explored in the next chapter.

2.2.5. Accuracy

The data accuracy principle requires that processed personal data is accurate and, where necessary, kept up to date. This was already present in Convention 108,[40] the first legally binding international instrument in the data protection field, and has been maintained after the GDPR replaced the Data Protection Directive 95/46/EC[41] in 2016. The principle means that the number of inaccurate data elements in training data should be limited. Also, hidden biases should be prevented and representativeness ensured in order to have accurate outputs. In a big data context, keeping personal data up to date could be a mission impossible to achieve. Accuracy and fairness principles together raise the standard for AI systems that make inferences about people. Such a system can be deployed only if it is sufficiently statistically accurate to fulfil its purposes.

Data accuracy has particular relevance for AI. As the French Data Protection Authority highlights in its report on ethical matters raised by AI, 'the matter of the quality of the data processed by algorithms and AI is the most straightforward. It is not difficult to understand that incorrect data or data that is quite simply out of date will lead to errors or malfunctions of varying gravity depending on the sector in question'.[42] Inaccurate data could have a significant impact on individuals in the deployment phase also, resulting in an erroneous output or unjustified decision.

2.2.6. Storage Limitation

The principle of storage limitation prohibits keeping personal data longer than needed for the purposes of the processing. In order to ensure that the personal data

38 According to the WP29 Guidelines on DPIA, endorsed by the EDPB, innovative use or applying new technological or organizational solutions or matching or combining datasets in a way that would exceed the reasonable expectations of the data subjects can trigger the need to carry out a DPIA. WP29, *Guidelines on Data Protection Impact Assessment* (DPIA) and determining whether processing is 'likely to result in a high risk' for the purposes of Regulation 2016/679 (WP 248 rev.01) 2017.

39 Norwegian Data Protection Authority (Datatilsynet) 2018.

40 Council of Europe, Convention for the protection of individuals with regard to the processing of personal data, opened for signature on 28 January 1981.

41 Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data 1995.

42 French Data Protection Authority (CNIL) 2017.

are not kept longer than necessary, time limits should be established for erasure or for a periodic review.[43] Defining appropriate data retention periods assumes that we identified all the purposes of data processing in advance. This is undoubtedly challenging in the case of AI-based processing, as the purpose of processing may change during the development and due to the high level of data replication.

GDPR leaves room for exceptions: if the personal data is processed solely for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes, the data may be stored for longer periods.[44]

2.2.7. Security

The principle focusing on confidentiality and integrity states that personal data must be processed in a manner that ensures their appropriate security, 'including protection against unauthorised or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organisational measures'. Security requirements apply explicitly to data controllers and processors[45] also. Two sections of the GDPR, articles 32–36 are dedicated to security requirements. These include the newly introduced requirement to notify personal data breaches to the supervisory authority and in certain cases to the data subjects too.

In order to maintain security, organizations should evaluate the risks inherent in the processing and implement measures to mitigate those risks.[46] The measures put in place should ensure an appropriate level of security. This implies the implementation of risk-based technical and organizational measures both at the time of the determination of the means for processing and at the time of the processing itself. In order to implement measures which ensure a level of security appropriate to the risk, organizations should take into account the state of the art and the costs of implementation in relation to the risks and the nature of the personal data to be protected.[47] This requires that organizations developing, deploying, or using AI assess and mitigate the security risks personal data processing may raise.

In addition to the obligation to ensure confidentiality and integrity of data processing, Article 32 provides the ongoing availability and resilience of processing systems and services and the monitoring and testing of processing activities. AI HLEG stresses technical robustness, which is a critical component of achieving trustworthy AI. The expert group emphasizes that 'technical robustness requires that AI systems be developed with a preventative approach to risks and in a manner

43 Recital 39 GDPR.
44 Article 5(1) e) GDPR.
45 Article 4(8) GDPR: a natural or legal person, public authority, agency or other body that processes personal data on behalf of the controller.
46 Recital 83 GDPR.
47 Article 32(1) GDPR.

such that they reliably behave as intended while minimizing unintentional and unexpected harm and preventing unacceptable harm'.[48]

Ensuring security of personal data implies more than preventing re-identification of data subjects, unauthorized disclosure of training data or model outputs, or inferences about individuals represented in the training data. As AI HLEG highlights, 'AI systems, like all software systems, should be protected against vulnerabilities that can allow them to be exploited by adversaries' and should also 'have safeguards that enable a fallback plan in case of problems'.[49]

This presumes security capabilities against third-party malicious activities such as hacking, alteration of the training data, model poisoning or evasion, or model inversion attack. The next subchapter, titled *Attacks on AI Models*, provides an overview of the security issues raised by AI systems.

The EDPB and EDPS draw the attention in their joint report that the wording of Article 83 of AI Act[50] does not include a reference to changes in external risks. They recommend that 'a reference to changes of the threats-scenario, arising from external risks, e.g., cyber-attacks, adversarial attacks and substantiated complaints from consumers therefore should be included in Article 83 of the Proposal'.[51]

PETs can partially support the compliance of AI systems with the data security requirements. Employing differential privacy, homomorphic encryption, multi-party computation, federated learning or using synthetic data for training purposes can contribute to security goals and a privacy-preserving AI system. More on some of these technologies will be explained in the third chapter.

2.2.8. GDPR Fines

The importance of data protection principles is also reflected by the enforcement actions taken by the national supervisory authorities. The Hungarian Data Protection Authority (NAIH) imposed a fine of €665,000 in February 2022 for the unlawful use of artificial intelligence. A Hungarian bank applied an AI-powered emotion analysis on voice recordings of calls conducted between its customers and a call centre. The bank failed to comply with the transparency and purpose limitation principles. Further, the Authority considered the solution's inefficiency in predicting the customers' emotions accurately, as well as the risk of processing

48 AI HLEG 2019.

49 Ibid.

50 This Regulation shall apply to the high-risk AI systems, other than the ones referred to in Paragraph 1, that have been placed on the market or put into service before [date of application of this Regulation referred to in Article 85(2)], only if, from that date, those systems are subject to significant changes in their design or intended purpose.

51 EPDP 2022; Joint Opinion 5/2021 on the proposal for a Regulation of the European Parliament and of the Council laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) 2021.

data by AI. The decision of the Authority highlighted that the information, the data protection impact assessment (DPIA), and balancing test documentation provided by the bank were not in compliance with the GDPR.[52]

Also, Clearview, the controversial facial recognition service provider, has been heavily fined for infringing GDPR by scraping images of individuals from public web sources and generating biometric data. After the data protection authorities in the United Kingdom and Italy, the company has been hit with another sanction from the Hellenic Data Protection Authority.[53] The value of fines received totalled €49 million.

2.3. Attacks on AI Models

Given the complexity of big data processing, AI systems can pose specific threats in addition to the security issues associated with any IT system. Besides human errors or omissions, data breaches can also be caused by external attacks, which 'may target the data, the model or the underlying infrastructure, both software and hardware'.[54]

Big data systems are increasingly becoming targets of more elaborate and specialized attacks.[55] Attacks on AI systems are increasing constantly, but the AI industry is alarmingly unprepared for these.[56] The attacks resulting in data breaches can lead to financial losses in addition to privacy and data protection harms. The average data breach cost has climbed 12.7% in the last two years, reaching $4.35 million in 2022.[57]

This subchapter will concentrate on the threats that can target machine learning models and present potential security challenges for personal data.

2.3.1. Poisoning

Poisoning attacks target classification algorithms and are likely to occur in several stages of the lifecycle from data collection to monitoring. In such an attack, an adversary is able to manipulate data (insert malicious data into the training or validation data) or model (replacing model file with an altered one) in order to modify the algorithm's behaviour in a chosen direction at a later point in time.[58] This may cause intentional misclassification or discrimination affecting the model's accuracy, compromising the integrity and trustworthiness of the AI system.

Federated learning and homomorphic encryption are PETs discussed in relation to the prevention of poisoning.

52 Hungarian Data Protection Authority 2022.
53 Ibid.
54 AI HLEG 2019.
55 European Union Agency for Network and Information Security (ENISA) 2016.
56 Advisa – The Road to Secure and Trusted AI Report – The Decade of AI Security Challenges 2021.
57 IBM 2022.
58 ENISA 2016.

2.3.2. Model Inversion

We usually think in relation to machine learning models that the training dataset cannot be recovered from the trained model. As several studies have shown, even without access to the dataset used for training, the output of the machine learning algorithms can be extremely revealing.

The exception is provided by model inversion attacks, which aim at classification algorithms. This attack can take place when the attacker has access to certain personal data belonging to specific individuals included in the training data, and by observing the inputs and outputs of the machine learning model it can infer further personal information about those same individuals. An excess of training data and the possibility to repeatedly query the model can contribute to the success of these attacks, and in some cases the re-identification of the data subjects.

One of the earliest successful model inversion attacks were deployed on recommender systems, 'a demonstration that collaborative filtering systems, where item recommendations are generated for a user based on behavioural patterns of other users, can end up revealing the consumption patterns of individual users'.[59] Also, researchers demonstrated the possibility of reverse engineering on a medical model designed to predict the correct doses for an anticoagulant using patient data including genetic biomarkers. They proved that an attacker having access to some demographic information of the patients included in the training data could infer their genetic biomarkers from the model.[60] Model inversion proved effective for attacking Facial Recognition Technology (FRT) systems. Researchers could reconstruct the facial images associated with the individuals included in the training data and match these (by humans) with 95% accuracy.[61]

Model inversion attacks can produce unauthorized disclosure of some personal data processed for training, causing loss of confidentiality and affecting the trustworthiness of the system.

In order to deploy an AI system in a way which prevents model inversion attacks, we should use secure multi-party computation or synthetic data if possible.

2.3.3. Membership Inference

This attack targets regression and classification algorithms in the deployment phase. Membership inference attacks allow malicious third parties to determine whether a given individual was present in the training data or not.[62] The attack itself does not cause disclosure of personal data, but using the model in combination

59 Veale–Binns–Edwards 2018.
60 Fredrikson–Jha–Ristenpart 2015.
61 Ibid.
62 Shokri–Stronati–Song–Shmatikov 2017.

with other data about a particular individual could directly lead to data breaches.[63] For example, if patients' clinical records are used to train a model associated with a disease, attackers knowing that a certain patient's data was used to train the model could reveal that the patient has this disease.

Membership inference attacks exploit confidence scores specific to prediction models. The score is disproportionately high in a prediction about an individual that was in the training dat, because the model has seen the data before. This allows us to determine if the individual was in the training data.

Exposure to membership inference depends on the amount of information 'remembered' by the used machine learning algorithm from the training datasets and on the degree of overfitting the model.[64] The level of risks associated with the attack depends on the sensitivity that the information membership may reveal. For example, if the model is trained with data on a vulnerable population like people suffering from mental disorder, addictions, or HIV, a membership inference attack could have a high-risk impact.

Making inferences about individuals represented in the training data through black-box or white-box inference can lead to breach of the principle of confidentiality. Employing differential privacy during model training can provide defence against membership inference attacks. More about trade-offs and limitations in the next chapter.

EDPB underlines the importance of implementing appropriate safeguards to identify, measure, and mitigate the risks that are specific to some machine learning such as data poisoning, model inversion, and white-box inference. Also, it considers essential to put in place monitoring processes to monitor (logging and collecting information on accuracy and fairness) the AI systems once in use.[65] The conclusion of ENISA could be the right one for this chapter as well: 'AI permeates every aspect of our daily lives, and therefore it is of paramount importance to ensure the cybersecurity of AI to ensure that AI and the set of associated technologies will be trustworthy, reliable and robust.'[66]

3. Privacy-Enhancing Technologies

In the context of AI, the security dimension of data protection bears a leading role in managing threats in a multi-party ecosystem and implementing specific controls to ensure that the AI system is secure. This implies that the necessary technical and organizational safeguards are put in place in the design stage of new

63 Ibid.
64 Shokri–Stronati–Song–Shmatikov 2017.
65 EDPB 2022.
66 ENISA 2016.

AI applications.[67] This is the scope of the notion of data protection by design and default, introduced as a legal requirement by Article 25 of the GDPR.

3.1. AI with Privacy

In 2015, ENISA expressed the need for a conceptual shift from 'big data versus privacy' to 'big data with privacy', 'adopting the privacy and data protection principles as an essential value of big data'[68] – mainly due to the 'scale of big data processing which brings existing privacy risks into a whole new (and unpredictable) level'.[69] Therefore, following a data protection by design approach may ask for an innovative solution since AI systems can have multiple levels of data processing and different techniques. Several techniques and technologies were proposed, developed, and improved in the last four decades that aim to support the deployment and configuration of appropriate technical and organizational measures in order to satisfy specific data protection principles. The group of these emerging technologies and techniques is commonly known as Privacy-Enhancing Technologies (PETs). Although the implementation of PETs may be necessary to comply with the data protection legal requirements, these alone cannot ensure compliance. To achieve compliance, PETs should always be used in conjunction with data protection policy and governance systems and frameworks.

In relation to GDPR compliance, there are several techniques to enhance data protection for AI systems. The most common ones are homomorphic encryption, secure multi-party computation, differential privacy, and synthetic data. Applying these promising PETs alone or combined facilitates the privacy-preserving applications of AI. These will be highlighted in the following chapter, having a special focus on technologies supporting the compliance with the data security principle.

3.2. Homomorphic Encryption

In order to prevent unauthorized parties from accessing the processed personal data or to safely provide access, it is necessary to mask the data in three different states: at-rest, in-use, and in-transit. During the typical encryption techniques which secure data at-rest and in-transit, the original data once encrypted becomes obscured or unintelligible. The challenge was to develop an encryption technique which protects data in-use, while keeping it intelligible for processing.

In 1978, Ronald L. Rivest, Len Adleman, and Michael L. Dertouzos laid down the theoretical foundations of homomorphic encryption,[70] which allows computation

67 Ibid.
68 ENISA 2015.
69 Ibid.
70 Rivest–Adleman–Dertouzos 1978.

to be performed directly on encrypted data without requiring to decrypt it first. This method was further developed by Craig Gentry, who was the first to describe a construction for a fully homomorphic (typically asymmetric) encryption scheme.[71]

3.2.1. Benefits

Partially homomorphic encryption (PHE) allows only additions or only multiplications, somewhat homomorphic encryption can support a limited number of both additions and multiplications, and fully homomorphic encryption (FHE) enables multiple operations to be performed over encrypted data.[72] The end-result of the operation(s) remains also encrypted, and it can be decrypted by the owner of the key.[73] The result is equivalent to the results obtained working with the original unencrypted data directly.[74] In essence, homomorphic encryption ensures the secure outsourcing of specific operations on confidential data and safely provides access to them, being an important component of the defence against poisoning attacks.

3.2.2. Use Cases

Homomorphic encryption is a useful data protection by design measure in use cases when processing – for the time being, simple arithmetical operations such as addition and/or multiplication – is performed by a third party. For example, a cloud service provider as a processor can perform operations on behalf of the data controller without accessing the content of the personal data. A real-world use case is the collection of data from connected devices with the purpose of obtaining aggregated values. The aggregator service receives individual encrypted data and adds them up, resulting in the final data encrypted accumulated values.[75] Enabling privacy-preserving data aggregation, homomorphic encryption is especially suitable for smart meter systems.[76]

In the last few years, electronic homomorphic encryption has been extended to new use cases such as smart contracts, electronic voting,[77] genome privacy,[78] fraud detection,[79] or password breach monitoring,[80] and it has the potential to be used

71 Gentry 2009.
72 The Royal Society 2019.
73 Industry, Government and Academic Consortium to Advance Secure Computation 2017.
74 Spanish Data Protection Agency 2020.
75 Ibid.
76 Wang–Heb–Zhangc 2022.
77 AEPD 2020.
78 Industry, Government and Academic Consortium to Advance Secure Computation 2017.
79 Maass 2020.
80 Lauter–Kannepalli–Cruz Moreno 2021.

in a wide range of applications. It can also support privacy-preserving machine learning. For example, it can underpin privacy-preserving predictions.[81]

3.2.3. Limitations

Due to the diversity of operations performed on the encrypted data, FHE is currently inefficient besides the higher level of protection and utility. Application of FHE is held back currently since it is highly computationally intensive, suffers from bandwidth and latency issues, and running time can exponentially increase depending on security parameters.[82] PHE and SHE, on the other hand, provide good performance and protection, but with very limited utility.

As encrypted data is typically much larger, more storage and processing resources are needed to encrypt, store, and decrypt the data. Therefore, homomorphic encryption is extremely computationally expensive and impractically slow.[83] As cloud computing evolves, performance will increase, and this PET will become more accessible for commercial applications.[84]

AEPD highlights the risk arising from 'using the same key on the data that are going to be processed may entail a vulnerability in the encryption system'. Risk increases as the volume of the processed data and the period of access grows. AEPD stresses that 'the use of an additional encryption layer in communications is essential together with the need to minimise the information encrypted under the same key, which must be limited to the groups of data that operate with one another'.[85] Standardization of homomorphic encryption techniques is also a major ongoing challenge.

3.2.4. Potential

The level of maturity differs for variations of homomorphic encryption. There are products based on PHE offered on the market, SHE is piloted, but FHE is just at research level.[86] However, homomorphic encryption can already enable other PETs such as secure multi-party computation, private data aggregation, or federated machine learning.[87] The use of homomorphic encryption opens new doors for the secure processing of personal data 'such as servers based on the IoT, Cloud Computing and automated learning or Machine Learning'.[88]

81 The Royal Society 2019.
82 Ibid.
83 Ibid.
84 AEPD 2020.
85 Ibid.
86 The Royal Society 2019.
87 Ibid.
88 AEPD 2020.

Some companies started to use homomorphic encryption for end-to-end encrypted systems' data processing. Meta is experimenting with this technique to detect child sexual abuse material on end-to-end encrypted messaging platforms. However, it 'is not yet technically feasible to implement in messaging at scale' because it would take more than seven months to run on each message.[89]

3.3. Secure Multi-party Computation

Secure multi-party computation (SMPC) is a more mature PET enabling computation on encrypted data without losing data utility. Introduced in 1986 and the first prototypes developed in 2004,[90] SMPC is a subfield of cryptography concerned with enabling private distributed computations.

3.3.1. Benefits

SMPC allows 'computation or analysis on combined data without the different parties revealing their own private input'.[91] Lack of trust between two or more parties, also data protection restrictions (ex. appropriate legal basis, ensuring confidentiality) or technical constraints of data sharing between parties who intend to carry out analyses on their combined data, are addressed by this cryptographic technique providing data masking.[92] It is important that the original, distributed data existing across several parties does not need to be gathered to a central repository. Practically, SMPC enables operations to be performed on the input data of two or more parties, without revealing the input data of one party to the other parties, and ensures parties to jointly form the obtained results.[93] Unlike homomorphic encryption which protects data in storage and also during computing, SMPC supports only the latter.

SMPC can be used together with federated machine learning to leverage the benefits of stronger confidentiality with greater scale of computation.[94]

3.3.2. Use Cases

Commercial products for secure multi-party computation first appeared in 2010.[95] Some real-world applications preceded them. These permit auctions where participants could place bids without revealing them. A good example

89 Business for Social Responsibility 2022.
90 Malkhi–Nisan–Pinkas–Sell 2004.
91 The Royal Society 2019.
92 AEPD 2022.
93 Ibid.
94 Mugunthan–Polychroniadou–Byrd–Hybinette Balch 2019.
95 The Royal Society 2019.

is provided by Denmark, where SMPC was used to redistribute the country's EU-fixed production quota among sugar beet producers without the need for a central auctioneer or revealing commercially sensitive information.[96] This system allows bidders to identify the winner of the auction without revealing information related to the actual bid.

Pooling personal data from different governmental departments to gain insights for policy makers involves high privacy risks. SMPC can enable us to put information in a wider context without revealing citizens' personal data. For example, Estonia used SMPC to analyse encrypted income tax records and higher education records to determine if students who work during their studies are more likely to fail to graduate in time than fellow students who are focused exclusively on their studies.[97]

3.3.3. Limitations

Although SMPC is constantly evolving, like all PETs, it faces a number of challenges. The widespread use of the SMPC is limited by the relatively high costs of computation and bandwidth. However, where high-bandwidth settings are available (devices connected within a data centre), SMPC significantly outperforms FHE.[98] If all the participants outsource their computation to the same cloud provider, bandwidth costs can be reduced, 'but it requires a strong trust model'.[99] Also, data structures need to be standardized in order to perform data analysis with SMPC.

Since the output is revealed in the case of SMPC, 'the output must be controlled to limit what an adversary can infer about the private data from the output'.[100] This leakage can be addressed in the best way by combining SMPC with differential privacy.[101]

Beyond the SMPC execution itself, the protection of the cryptographic keys is a challenge that has to be tackled by involved parties.[102] Thus, parties implementing SMPC should have a high level of security capabilities.

3.3.4. Potential

SMPC is promising for operations that require large amounts of data as the training of machine learning models. SMPC can be used to allow private multi-party machine learning. Different parties send encrypted data to each other to train a

96 Bogetoft–Lund–Damgård–Geisler 2009.
97 Bogdanov–Kamm–Kubo–Rebane–Sokk–Talviste 2016.
98 Evans–Kolesnikov–Rosulek 2018.
99 Ibid.
100 Ibid.
101 Pettai–Laud 2015.
102 The Royal Society 2019.

machine learning model, eliminating the need for a trusted central authority that would perform the computation by gathering all the data and decrypting it.[103]

SMPC has a great potential for machine learning systems trained on health data. SMPC could facilitate the use and sharing of health data for R&D purposes, as it tackles the problem that data is distributed across several organizations, and gathering all the data to a central repository is rarely permitted by the data protection requirements.

By keeping the input data of each party private and providing the correct output to each of them, SMPC can support the compliance with the purpose limitation and prevent model inversion attacks. However, it is necessary to implement additional data protection measures to guarantee GDPR compliance.

3.4. Differential Privacy

While homomorphic encryption and SMPC deal with privacy during computation, differential privacy addresses privacy in disclosure.[104] This PET was introduced in 2006 by C. Dwork[105] and her team,[106] and it is based on the Law of Large Numbers.[107]

Differential privacy is a 'strong, mathematical definition of privacy in the context of statistical and machine learning analysis', and 'it is used to enable the collection, analysis, and sharing of a broad range of statistical estimates based on personal data'.[108] The differential privacy mathematically guarantees that the result of a differentially private analysis provides the same inference about any individual's personal data, regardless of whether that particular individual's personal data was included in the input to the analysis.[109] Differential privacy preserves the usefulness of data by allowing statistical analysis and identification of trends on larger datasets, but in a way that protects individuals' privacy by 'establishing data protection guarantees by design through the practical implementation of information abstraction strategies'.[110]

Depending on the stage when the data analysis is applied, differential privacy can be local (distributed) or global (centralized). In the case of local differential privacy, random noise is added at the data collection stage 'so that users get a "plausible deniability" type of guarantee with respect to data being collected about them'.[111] This may result in reducing accuracy by adding more noise than

103 Ibid.
104 Ibid.
105 Dwork 2006.
106 Dwork–McSherry–Nissim–Smith 2006.
107 Law of Large Numbers 2020.
108 Wood–Altman–Bembenek–Bun–Gaboardi–Honaker–Nissim–O'Brien–Steinke–Vadhan 2018.
109 Ibid.
110 AEPD 2021.
111 The Royal Society 2019.

the global approach, as adding noise at an early stage of the data lifecycle does not permit optimizing the amount of noise to a specific analysis.[112] Global differential privacy assumes that noise is added to the output, taking away the possibility to determine if a particular data record was included in the dataset used to produce the output.[113]

The adjustability of the amount of noise added to the original dataset is an important feature of differential privacy. By increasing the amount of noise, privacy risks decrease, but data utility may decline too. The challenge is to calculate the value of noise in a way 'that preserves the result within the utility range'.[114]

3.4.1. Benefits

Differentially private mechanisms can provide a way to query datasets containing private data while mitigating 'the risk of revealing whether a specific individual or organisation is present in a dataset or output'.[115]

One of the major benefits of differential privacy is the strong protection provided against membership inference attacks if the training process is differentially private.[116] Also, differential privacy is the best practice against re-identification attacks performed by combining different datasets.[117]

Another benefit of differential privacy is the possibility to quantify the privacy loss and compare it among different techniques. This enables the control and analysis of cumulative privacy losses when running multiple differentially private analyses on a particular dataset. Also, measuring privacy loss acquired by groups is possible.[118] Immunity to post-processing is also an important property of differential privacy. This allows to arbitrarily transform a differentially private output using some data-independent function, but without impacting its privacy guarantees.[119]

Dwork and her team have shown that differential privacy can improve generalization in machine learning algorithms.[120] In particular, 'if a differentially private learning algorithm has good training accuracy, it is guaranteed to have good test accuracy'.[121]

112 Ibid.
113 Ibid.
114 AEPD 2021.
115 Ibid.
116 Shokri–Stronati–Song–Shmatikov 2017.
117 Chin–Anne Klinefelter 2012.
118 Nguyen 2019.
119 Zhu–Van Hentenryck–Fioretto 2020.
120 Dwork–Feldman–Hardt–Pitassi–Reingold–Roth 2015.
121 Papernot–Guha Thakurta 2021.

3.4.2. Use Cases

Differential privacy is PET which has a wide range of applications from linear regressions and cumulative distribution functions to machine learning.[122]

After having implemented differential privacy for other services, the United States Census Bureau took the decision to replace data swapping, the previously applied disclosure avoidance mechanism, with differential privacy for the 2020 census. This was motivated by its goal 'to publish a specific, higher number of tables of statistics with more granular information than previously'[123] and at the same time to protect against emerging technology threats such as re-identification attacks.[124]

High-profile tech companies such as Apple, Google, Microsoft, and Uber also implemented differential privacy in practical applications. Apple used local differential privacy for collecting statistics from hundreds of millions of users in order to identify popular emojis, popular health data types, and media playback preferences in Safari.[125]

Google implemented differential privacy with a similar goal: such, to collect statistics from end-users in a privacy-preserving way.[126] Similarly, Uber in collaboration with the University of California implemented this method to perform analytics on user data and determine the average trip distance for users.[127]

3.4.3. Limitations

Adding noise to a dataset can cause accuracy and robustness issues. Especially for smaller datasets, this can harm utility. Practically, the trade-off of utility and privacy improves proportionally with the size of the dataset, where less noise is needed, as 'the more individuals included in a dataset, the harder it might be to identify that a specific individual was included'.[128]

In the case of local models, the utility is usually affected because the distributed data requires more noise to achieve differential privacy. In order to obtain highly accurate aggregate statistics, large datasets are essential. But working on large datasets does not automatically lead to great utility. The algorithms transforming a dataset to differentially private need to be designed to the specific use case to ensure that the output meets utility expectations.[129]

122 AEPD 2021.
123 The Royal Society 2019.
124 United States Census Bureau 2022.
125 Differential Privacy Team of Apple 2017.
126 Erlingsson–Pihur Korolova 2014.
127 Johnson–Near–Song 2018.
128 The Royal Society 2019.
129 Ács–Castelluccia 2014.

The privacy preservation effect of differential privacy is heavily dependent on the 'privacy budget', the 'quantitative measure of by how much the risk to an individual's privacy may increase by, due to that individual's data inclusion in the inputs to the algorithm'.[130] Setting the 'privacy budget' is key to ensuring privacy guarantees, and it requires expertise. It is crucial to take into consideration 'the statistical inferences that might happen after the release of results and how, for example, outsiders might be able to link data with side information'.[131]

For example, differential privacy could lose its privacy guarantees where differentially private data collection from the same individuals is continuous over time. It is not possible 'to collect differential privacy protected data from a community of respondents an indefinite number of times with a meaningful privacy guarantee'.[132] This is why the previously detailed user data collection performed by Apple and Google presents shortcomings.

3.4.4. Potential

Differential privacy provides great performance for datasets where the number of individuals is large but the weight of each individual to the output is limited. This privacy-enhancing technology contributes substantially to enable privacy-preserving machine learning.

Overfitting is a typical mistake in machine learning. This can be mitigated by achieving differential privacy, which 'goes hand in hand with preventing overfitting to particular examples'.[133]

Differential privacy supports a wide range of techniques used in statistics and machine learning such as classification, clustering, and also statistical disclosure limitation techniques such as synthetic data generation. The generated synthetic data retains statistical properties of the original data but at the same time protects against model inversion attacks.[134] We can state without doubt, differential privacy will have a central role in deploying privacy-preserving machine learning.

4. Conclusions

The adoption of AI and machine learning has skyrocketed since the pandemic. As AI systems are becoming increasingly widespread in the public and private sectors, the majority of organizations realize that privacy challenges can be hardly

130 The Royal Society 2019.
131 Ibid.
132 Domingo-Ferrer–Sánchez–Blanco-Justicia 2020.
133 The Royal Society 2019.
134 Ibid.

overcome. But despite the fact that privacy is considered the fourth most relevant AI risk after cybersecurity, regulatory compliance, and explainability,[135] 52% of business decision makers say that their company is not safeguarding data privacy through the entire lifecycle,[136] thus failing to meet an important condition for trustworthy AI. This creates exposure for the data involved in training, testing, or deploying the system. Attacks targeting machine learning models can exploit these vulnerabilities in novel ways, as has been shown in the second chapter, increasing privacy risks for individuals and the odds of a hefty GDPR fine.

Due to the volume of the data, the complexity of the processing, and the unforeseen consequences for data subjects, applying the data protection principles in a machine learning context is far from straightforward. Practically, data protection compliance would require translating these principles into concrete requirements and system design specifications and then finding and implementing appropriate technical and organizational measures throughout all of the stages of the data processing lifecycle. As GDPR asks for state-of-the-art and risk-based safeguards implemented from the design phase, innovative solutions and approaches are needed, which are better suited to the personal data processing performed by machine learning models in order to unlock the full potential of these data-driven AI technologies.

Privacy-enhancing technologies specifically tailored to mitigate privacy risks characteristic to machine learning models were presented in the third chapter. Homomorphic encryption, secure multi-party computation, and differential privacy not only provide shield against attacks targeting machine learning models but can bring a significant contribution to comply with the fairness, purpose limitation, and data minimization principles besides the data security principle. PETs enable multiple applications and bring new possibilities for data analysis.[137] These technologies evolving at an accelerating pace since 2000 could open the horizon for privacy-preserving machine learning. However, PETs do not transform personal data processing compliant with data protection regulations in one fell swoop, and they should be used together with process controls, high-standard policy, and data governance systems.

At this stage, significant barriers are still present, PETs being limited by high computational demand, data-interoperability, data utility, and accuracy issues. Security risks resulting from reverse engineering are also highly debated. Maturity level and early-stage use of PETs also make the road to market-wide penetration meandering. This could be enhanced by promoting good practices, initiating standardization, and developing certification mechanisms for mature PETs by the responsible bodies. These all point to the need of further research and development

135 McKinsey 2021.
136 IBM 2022.
137 The Royal Society 2019.

to explore the potential benefits and impact of PETs on processing personal data by AI systems.

Hopefully, the pace of innovation in this field will be maintained, and in the near future we will be able to enjoy the benefits of AI systems deployed by private and public organizations together with top-notch privacy safeguards.

References

ÁCS, G.–CASTELLUCCIA, C. 2014. *A Case Study: Privacy Preserving Release of Spatio-Temporal Density in Paris*. http://www.crysys.hu/~acs/publications/AcsC14kdd.pdf.

AEPD (SPANISH DATA PROTECTION AGENCY (AGENCIA ESPAÑOLA DE PROTECCIÓN DE DATOS). 2020. *Encryption and Privacy III: Homomorphic Encryption*. https://www.aepd.es/en/prensa-y-comunicacion/blog/encryption-privacy-iii-homomorphic-encryption.

2021. *Anonymisation and Pseudonymisation (II): Differential privacy*. https://www.aepd.es/en/prensa-y-comunicacion/blog/anonymisation-and-pseudonymisation-ii-differential-privacy.

2022. *Privacy by Design: Secure Multi-Part Computation: Additive Sharing of Secrets*. https://www.aepd.es/en/prensa-y-comunicacion/blog/privacy-by-design-secure-multi-part-computation-additive-sharing-secrets.

BOGDANOV, D.–KAMM, L.–KUBO, B.–REBANE, R.–SOKK, V.–TALVISTE, R. 2016. *Students and Taxes: A Privacy-Preserving Study Using Secure Computation*. https://www.researchgate.net/publication/302065845_Students_and_Taxes_a_Privacy-Preserving_Study_Using_Secure_Computation.

BOGETOFT, P.–LUND, D. H.–DAMGÅRD, I.–GEISLER, M. 2009. *Secure Multiparty Computation Goes Live*. https://www.researchgate.net/publication/220796917_Secure_Multiparty_Computation_Goes_Live.

BORKING, J.–RAAB, C. 2001. *Laws, PETs and Other Technologies for Privacy Protection, Journal of Information, Law and Technology*. https://www.researchgate.net/publication/220667925_Laws_PETs_and_Other_Technologies_for_Privacy_Protection.

BUSINESS FOR SOCIAL RESPONSIBILITY 2022. *Human Rights Impact Assessment: Meta's Expansion of End-to-End Encryption – Executive Summary*. https://about.fb.com/wp-content/uploads/2022/04/BSR-E2EE-HRIA-Executive-Summary.pdf.

CHIN A.–KLINEFELTER, A. 2012. *Differential Privacy as a Response to the Reidentification Threat: The Facebook Advertiser Case Study*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2062447.

COUNCIL OF EUROPE 1981. *Convention for the Protection of Individuals with Regard to the Processing of Personal Data, Opened for Signature on 28 January 1981*. https://rm.coe.int/convention-108-convention-for-the-protection-of-individuals-with-regar/16808b36f1.

DIFFERENTIAL PRIVACY TEAM OF APPLE 2017. *Learning with Privacy at Scale*. https://docs-assets.developer.apple.com/ml-research/papers/learning-with-privacy-at-scale.pdf.

DOMINGO-FERRER, J.–SÁNCHEZ, D.–BLANCO-JUSTICIA, A. 2020. *The Limits of Differential Privacy (and Its Misuse in Data Release and Machine Learning)*. https://arxiv.org/pdf/2011.02352.pdf.

DWORK, C. 2006. *Differential Privacy*. https://audentia-gestion.fr/MICROSOFT/dwork.pdf.

DWORK, C.–FELDMAN, V.–HARDT, M.–PITASSI, T.–REINGOLD, O.–ROTH, A. 2015. *Generalization in Adaptive Data Analysis and Holdout Reuse*. https://arxiv.org/pdf/1506.02629.pdf.

DWORK, C.–MCSHERRY, F.–NISSIM, K.–SMITH, A. 2006. *Calibrating Noise to Sensitivity in Private Data Analysis*. https://link.springer.com/content/pdf/10.1007/11681878_14.pdf.

EDPB 2020a. *Guidelines 4/2019 on Article 25 Data Protection by Design and by Default, Version 2.0, 2020*. https://edpb.europa.eu/sites/default/files/files/file1/edpb_guidelines_201904_dataprotection_by_design_and_by_default_v2.0_en.pdf.

2020b. *Response Letter to Sophie in't Veld*. https://edpb.europa.eu/sites/edpb/files/files/file1/edpb_letter_out2020_0004_intveldalgorithms_en.pdf.

2022. *Guidelines 05/2022 on the Use of Facial Recognition Technology in the Area of Law Enforcement*. https://edpb.europa.eu/system/files/2022-05/edpb-guidelines_202205_frtlawenforcement_en_1.pdf.

ERLINGSSON, Ú.–PIHUR, V.–KOROLOVA, A. 2014. *RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response*. https://arxiv.org/pdf/1407.6981.pdf.

EUROPEAN COMMISSION. 2020. *A European Strategy for Data*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0066

EUROPEAN COMMISSION DIRECTORATE-GENERAL FOR COMMUNICATIONS NETWORKS (AI HLEG). 2019. *Content and Technology, Ethics Guidelines for Trustworthy AI*. https://data.europa.eu/doi/10.2759/346720.

EUROPEAN COMMISSION JOINT RESEARCH CENTRE–EVAS, T.–SIPINEN, M.–ULBRICH, M. et al. 2022. *AI Watch: Estimating AI Investments in the European Union*. https://data.europa.eu/doi/10.2760/702029.

EUROPEAN UNION AGENCY FOR NETWORK AND INFORMATION SECURITY (ENISA).

2015. *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/@@download/fullReport.

2016. *Big Data Threat Landscape and Good Practice Guide*. https://www.enisa.europa.eu/publications/big-data-security/at_download/fullReport.

2020. *AI Cybersecurity Challenges – Threat Landscape for Artificial Intelligence*. https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges/at_download/fullReport.

EVANS, D.–KOLESNIKOV, V.–ROSULEK, M. 2018. *A Pragmatic Introduction to Secure Multi-Party Computation*. https://www.cs.virginia.edu/~evans/pragmaticmpc/pragmaticmpc.pdf.

FREDRIKSON, M.–JHA, S.–RISTENPART, T. 2015. *Model Inversion Attacks That Exploit Confidence Information and Basic Countermeasures*. https://rist.tech.cornell.edu/papers/mi-ccs.pdf.

FRENCH DATA PROTECTION AUTHORITY (CNIL) 2017. *How Can Humans Keep the Upper Hand – The Ethical Matters Raised by Algorithms and Artificial Intelligence*. https://www.cnil.fr/sites/default/files/atoms/files/cnil_rapport_ai_gb_web.pdf.

GARTNER 2021. *Top Strategic Technology Trends for 2022: Privacy Enhancing Computation*. https://www.gartner.com/doc/reprints?id=1-27VY7GL1&ct=211103&st=sb.

GENTRY, C. 2009. *Fully Homomorphic Encryption Using Ideal Lattices*. https://www.cs.cmu.edu/~odonnell/hits09/gentry-homomorphic-encryption.pdf.

HELLENIC DATA PROTECTION AUTHORITY. 2022. *Press Release about Fining Clearview AI Inc., 20 July 2022*. https://edpb.europa.eu/news/national-news/2022/hellenic-dpa-fines-clearview-ai-20-million-euros_en.

HUNGARIAN DATA PROTECTION AUTHORITY.

2022. *Press Release about a Fine Imposed in Connection with the Use of Artificial Intelligence, 20 May 2022*. https://edpb.europa.eu/news/national-news/2022/data-protection-issues-arising-connection-use-artificial-intelligence_en.

IBM. 2022. *Cost of a Data Breach – Report*. https://www.ibm.com/downloads/cas/XZNDGZKA.

IBM–MORNING CONSULT. 2022. *IBM Global AI Adoption Index 2022*. https://www.ibm.com/downloads/cas/GVAGA3JP.

INDUSTRY, GOVERNMENT AND ACADEMIC CONSORTIUM TO ADVANCE SECURE COMPUTATION. 2017. *Homomorphic Encryption Standardization, Basics of Homomorphic Encryption*. 2017. https://homomorphicencryption.org/introduction.

INFORMATION COMMISSIONER'S OFFICE – THE ALAN TURING INSTITUTE. 2020. *Explaining Decisions Made with AI*. https://www.pdpjournals.com/docs/888063.pdf.

JOHNSON, N.–NEAR, J. P.–SONG, D. 2018. *Towards Practical Differential Privacy for SQL Queries*. https://arxiv.org/pdf/1706.09479.pdf.

LAUTER, K.–KANNEPALLI, S.–MORENO, R. C. 2021. *Password Monitor: Safeguarding Passwords in Microsoft Edge*. https://www.microsoft.com/en-us/research/blog/password-monitor-safeguarding-passwords-in-microsoft-edge.

MAASS, E. 2020. *Fully Homomorphic Encryption: Unlocking the Value of Sensitive Data While Preserving Privacy*. https://securityintelligence.com/posts/fully-homomorphic-encryption-next-step-data-privacy.

MALKHI, D.–NISAN, N.–PINKAS, B.–SELL, Y. 2004. *Fairplay – A Secure Two-Party Computation System*. https://www.usenix.org/legacy/event/sec04/tech/malkhi/malkhi.pdf.

MCKINSEY 2021. *Global Survey: The State of AI in 2021*. https://www.mckinsey.com/business-functions/quantumblack/our-insights/global-survey-the-state-of-ai-in-2021.

MUGUNTHAN, V.–POLYCHRONIADOU, A.–BYRD, D–HYBINETTE BALCH, T. 2019. *SMPAI: Secure Multi-party Computation for Federated Learning*. https://www.jpmorgan.com/content/dam/jpm/cib/complex/content/technology/ai-research-publications/pdf-9.pdf.

NGUYEN, A. 2019. *Understanding Differential Privacy – From Intuitions behind a Theory to a Private AI Application*. https://towardsdatascience.com/understanding-differential-privacy-85ce191e198a.

NORWEGIAN DATA PROTECTION AUTHORITY (DATATILSYNET). 2018. *Artificial Intelligence and Privacy*. https://www.datatilsynet.no/globalassets/global/english/ai-and-privacy.pdf.

PANEL FOR THE FUTURE OF SCIENCE AND TECHNOLOGY AND EUROPEAN PARLIAMENTARY RESEARCH SERVICE. 2022. *Auditing the Quality of Datasets Used in Algorithmic Decision-Making Systems*. https://www.europarl.europa.eu/RegData/etudes/STUD/2022/729541/EPRS_STU(2022)729541_EN.pdf.

PAPERNOT. N.–GUHA THAKURTA, A. 2021. *How to Deploy Machine Learning with Differential Privacy*. https://www.nist.gov/blogs/cybersecurity-insights/how-deploy-machine-learning-differential-privacy.

PETTAI, M.–LAUD, P. 2015. *Combining Differential Privacy and Secure Multiparty Computation*. https://eprint.iacr.org/2015/598.pdf.

RIVEST, R. L.–ADLEMAN, L.–DERTOUZOS, M. L. 1978. *On Data Banks and Privacy Homomorphisms*. https://people.csail.mit.edu/vinodv/6892-Fall2013/RAD78.pdf.

SHOKRI, R.–STRONATI, M.–SONG, C.–SHMATIKOV, V. 2017. *Membership Inference Attacks against Machine Learning Models*. https://arxiv.org/pdf/1610.05820.pdf.

STANFORD UNIVERSITY–LITTMAN, M. L.–AJUNWA, I.–BERGER, G.–BOUTILIER, C.–CURRIE, M.–DOSHI-VELEZ, F.–HADFIELD, G.–HOROWITZ, M. C.–ISBELL,

C.–KITANO, H.–LEVY, K.–LYONS, T.–MITCHELL, M.–SHAH, J.–SLOMAN, S.–VALLOR, S.–WALSH, T. 2021. *Gathering Strength, Gathering Storms: The One Hundred Year Study on Artificial Intelligence (AI100). Study Panel Report.* http://ai100.stanford.edu/2021-report.

THE ROYAL SOCIETY. 2019. *Protecting Privacy in Practice: The Current Use, Development and Limits of Privacy Enhancing Technologies in Data Analysis.* https://royalsociety.org/-/media/policy/projects/privacy-enhancing-technologies/privacy-enhancing-technologies-report.pdf.

UNITED STATES CENSUS BUREAU. 2020. *Census Data Products: Next Steps for Data Releases*. https://www.census.gov/newsroom/blogs/random-samplings/2022/04/2020-census-data-products-next-steps.html

VEALE, M.–BINNS, R.–EDWARDS, L. 2018. *Algorithms That Remember: Model Inversion Attacks and Data Protection Law*. https://arxiv.org/pdf/1807.04644.pdf.

WANG, B.–HEB, S–ZHANGC, S. 2022. *Privacy Protection Data Aggregation Scheme with Batch Verification and Fault Tolerance in Smart Grid Communication*. http://166.62.7.99/assets/default/article/2022/04/19/article_1650361841.pdf.

WOOD, A.–ALTMAN, M.–BEMBENEK, A.–BUN, M.–GABOARDI, M.–HONAKER, J.–NISSIM, K.–O'BRIEN, D. R.–STEINKE, T.–VADHAN, S. 2018. *Differential Privacy: A Primer for a Non-technical Audience*. https://scholarship.law.vanderbilt.edu/cgi/viewcontent.cgi?article=1058&context=jetlaw.

ZHU, K.–VAN HENTENRYCK, P.–FIORETTO, F. 2020. *Bias and Variance of Post-processing in Differential Privacy*. https://arxiv.org/pdf/2010.04327.pdf.

*** *Advisa – The Road to Secure and Trusted AI Report – The Decade of AI Security Challenges.* 2021. https://adversa.ai/download/1220.

*** *Directive 95/46/EC (WP29), Opinion 03/2013 on Purpose Limitation.* 2013. https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf.

*** *Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data.* 1995. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A31995L0046.

*** *Joint Opinion 5/2021 on the Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act).* 2021. https://edps.europa.eu/system/files/2021-06/2021-06-18-edpb-edps_joint_opinion_ai_regulation_en.pdf.

*** *Law of Large Numbers.* 2020. https://encyclopediaofmath.org/index.php?title=Law_of_large_numbers.

*** *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act).*

and Amending Certain Union Legislative Acts. 2021. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206.

*** *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation).* 2016. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679.

*** *WP29, Guidelines on Data Protection Impact Assessment (DPIA) and Determining Whether Processing Is "Likely to Result in a High Risk" for the Purposes of Regulation 2016/679 (WP 248 rev.01).* 2017. https://ec.europa.eu/newsroom/article29/items/611236/en.