10-2024

Can federated learning solve AI's data privacy problem?: A legal analysis

Warren B. CHIK

Florian GAMPER

Forthcoming in Volume 52 Rutgers Law Record (2024)

Can Federated Learning Solve AI's Data Privacy Problem?: A Legal Analysis

Warren B. Chik[1] *& Florian Gamper*[2]

Abstract

Federated learning (FL) is a method of training AI systems on different datasets without sharing data. The promise of FL is to enable AI systems to be trained on data, including personal data, while preserving data privacy and confidentiality, and thus, inter alia, facilitate compliance with data protection legislation. FL has generated a considerable interest amongst the computer science community, yet there is a dearth of legal analysis of FL. This is a problem because the question of whether FL facilitates compliance with data protection legislation is a legal question. This article will fill this lacuna by providing a comprehensive legal analysis of FL through an examination of how the EU's General Data Protection Regulation (GDPR) applies to FL. This article postulates that, from a legal perspective, FL can be an effective method of facilitating compliance with data protection regulations. However, this article expresses doubt that, without support from policy makers and regulators, FL will be used sufficiently widely to make significantly more data available for the training of AI systems, than is currently the case.

.

Key words: Federated Learning, artificial intelligence, AI, machine learning, data sharing, AI data problem, privacy protection, GDPR

[1] Associate Professor of Law at Singapore Management University Yong Pung How School of Law, Tel: +65 6828 100, email: warrenchik@smu.edu.sg
[2] Research Associate at Singapore Management University Yong Pung How School of Law, Tel: +65 6828 100, email: fgamper@smu.edu.sg

This research/project is supported by the National Research Foundation Singapore and DSO National Laboratories under the AI Singapore Programme (AISG Award No: AISG2-RP-2020-018).

Table of Contents

I. INTRODUCTION

The speed of development in Artificial Intelligence (AI) in recent years has been breathtaking. Yet this comes with its own set of problems.[3] One of these problems is that the current iteration of AI requires vast quantities of data for the training of AI systems, and, currently, the demand for data is outstripping the supply of data.[4] This may hinder the further development and improvement of AI systems. This is sometimes referred to as AI's data problem.[5] However, making more data available to train AI raises various concerns, not least, in relation to data privacy, thus, AI has a data privacy problem.

[3] Thilo Hagendorff & Katharina, *15 challenges for AI: or what AI (currently) can't do*, 35 *AI & Soc*, 355, (2020).
[4] Tal Roded & Peter Slattery, *What Drives Progress in AI? Trends in Data*, FutureTech (March 19, 2024), https://futuretech.mit.edu/news/what-drives-progress-in-ai-trends-in-data. (last visited Aug. 17, 2024).
[5] Devika Rao, *All-powerful, ever-pervasive AI is running out of internet*, THE WEEK, June 5. 224;
S.E. Whang, et al. *Data collection and quality challenges in deep learning: a data-centric AI perspective*, 32, The VLDB Journal 79, (2023)

Federated learning (FL) may provide a solution to this problem.[6] The basic premise behind FL is simple: In the standard training of AI systems, data is collected and transferred onto a central server, where the AI system trains on the data. In contrast, in FL, data is not collected but remains in its original locations. Instead, each party receives the raw model, which is then trained on the dataset *in situ*. Upon completion of the training, the trained model is sent back, and is combined with other similarly trained models, into a single, integrated model. The result is that the integrated model has effectively been trained on all the datasets, but no data is transferred out from its original location.

Data is often held in 'silos'.[7] A data silo is anything that holds data (e.g. smartphones, laptops, hospitals, banks, etc.) but accessing data held in silos is challenging. Sometimes regulation, like data protection legislation or IP law, prevents data from being shared. Alternatively, there may be a reluctance to share data, for instance, due to concerns about data confidentiality or data integrity. The intended purpose of FL is to 'break open' these data silos, by enabling the training of AI systems while preserving data privacy and confidentiality. If FL can fulfill this promise, this could bring significant benefits. By way of example, "healthcare providers could train algorithms to develop new drugs based on patient data, while maintaining privacy and patient confidentiality, or researchers in different countries could train algorithms without transmitting data across jurisdictions."[8]

FL has generated significant interest amongst the computer science community, however, there is a dearth of writings on, and understanding of, FL among lawyers and legal academics.[9] This is a problem because the conceptualization of concepts like data and privacy may differ across disciplines[10] and to what extent FL can break open data silos created by regulation requires a legal analysis. This article aims to fill this lacuna by providing a comprehensive legal analysis of FL. This will be done by examining how the data protection principles – represented by the most stringent standards under the European Union's General Data Protection Regulation (GDPR)[11] – applies to FL. The argument will made that from a legal perspective FL can indeed be an effective method to ensure compliance with data protection regulation.

Although the legal analysis in this article focuses on the GDPR, the significance of the analysis extends beyond the EU. EU regulation has proven influential beyond the EU,[12] and many

---

[6] For an overview of FL, *see Federated Learning: Collaborative Machine Learning without Centralized Training Data*, 2017, Google Research Blog, posted by Brendan McMahan & Daniel Ramage, April 6, 2017, https://ai.googleblog.com/2017/04/federated-learning-collaborative.html (last accessed July 31, 2024)

[7] *See* Florian Gamper, *Federated Learning: What Lawyers Need to Know*, Law Gazette, (official publication of the Law Society of Singapore), June, 2024, https://lawgazette.com.sg/feature/federated-learning-what-lawyers-need-to-know/ (last accessed July 31, 2024) (makes a similar point as the above paragraph).

[8] *Id.*

[9] However, there is some legal analysis of FL. *See e.g.* S. Rossello et al., *Data Protection by design in AI? The case of federated learning*, 116 Computerr. (2021); Nguyen Truong et al. *Privacy Preservation in Federated Learning: An Insightful Survey from the GDPR perspective*, 110 Comput. Secur. J. 12402, 14-18, (2021).

[10] The same claim could be made in relation to many other concepts, like transparency, bias, fairness, to mention just a few.

[11] 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 O.J. (L119) [hereinafter GDPR].

[12] Anu Bradford, *The Brussels Effect*, 107(1) 1, Northwest. Univ. Law Rev.1, (2012) (agues that EU regulation impact jurisdictions outside the EU).

data protection regimes are modelled on the GDPR.[13] Moreover, FL raises an important conceptual question about the relationship between data protection and the development of AI; that is, whether the training of AI systems on personal data is in itself an infringement of data protection rights, or whether there is such an infringement only because of some feature of how the training is conducted, e.g. that data is collected to a central server or access to the data is given to a third party. In the standard training of AI systems, this question will seldom arise as data needs to be collected for the AI systems to be trained, and many data protection regimes regulate the collection of data.[14] However, because in FL no data is collected, this issue is brought into sharp focus, and in the age of AI, this is a question every data protection regime will need to answer.[15]

This article suggests that the training of AI systems itself does not infringe data protection rights, provided that the data is kept secure from abuse (i.e. the data being used for purpose other than training AI). The argument is that using personal data to train AI systems does not reveal information about an individual, such information is only revealed when the AI system is applied to a particular case. This article will show that the GDPR can be interpreted in this way, and if this interpretation is followed, the GDPR can provide for the protection of personal data, without hindering the development of AI systems.[16] For the legal analysis of FL this means that the question of whether the training of an AI system through FL is GDPR compliant will largely depend only on one factor, namely whether the data is kept secure, rather than the host of factors, which is typically required to assess GDPR compliance in standard training of AI systems. Thus, FL should make it easier for AI developers to train models on personal data.

Despite FL being a potential boon to AI's compliance with data protection regulation, this article will express doubt as to whether FL can make a significant contribution towards solving AI's data problem. Although, FL may be an effective way to deal with data protection, data protection is only one among other obstacles to data sharing. For instance, IP law may prevent data from being shared, and FL does not directly impact the application of IP law. Moreover, it is unlikely that FL will be used sufficiently widely to make significantly more data available, than is currently the case. There is also a lack of legal clarity in relation to FL, and without legal clarity it is unlikely that FL will be commonly adopted. Furthermore, currently, FL is not used widely across different organizations.[17] This means that a lot of data will remain inaccessible. This is an area where regulators and policy makers may be able to make a positive contribution. This paper will suggest that, if regulators and

[13] Graham Greenleaf, *Now157 Countries: Twelve Data Privacy Laws in 2021/22*, 176, Privacy L. & Bus. Int'l Rep.,1, 1 (2022).
[14] GDPR *supra* note 11, art.4.1(2), (states that collecting is a form of processing, inter alia GDPR arts. 5 and 6 regulate processing).
[15] Just to clarify, the question is also relevant for jurisdictions which currently do not have a data protection regime but are considering creating such a regime.
[16] *See*, Giovanni Sartor & Francesca Lagioia, Study, *The impact of the General Data Protection Regulation (GDPR) on artificial intelligence*, Study at the request of the Panel for the Future of Science and Technology (STOA) and managed by the Scientific Foresight Unit, within the Directorate-General for Parliamentary Research Services (EPRS) of the Secretariat of the European Parliament, (2020) [hereinafter EPRS Study], at 76
[17] Saikishore Kalloori & Abhishek Srivastava, *Towards cross-silo federated learning for corporate organizations*, 289, Knowledge-Based Systems, 2024, at 1.

policy makers decide to facilitate the use of FL, a possible tool is the creation of a FL regulatory regime, including an FL licensing regime, to facilitate data sharing across organizations.

This article will proceed as follows. First, an overview of FL will be provided. Second, this article will give a detailed analysis of how the GDPR applies to the training of AI systems. This analysis will take up considerable space, but it is crucial to understand how the GPPR applies to the training of AI systems as without such understanding, it is impossible to examine how the GDPR applies to FL. Third, this article will analyze to what extent FL can facilitate the sharing of non-personal data and examine the obstacles to FL being used more widely. Note that this article focuses on FL in relation to data protection regulation. Other issues, like IP law or antitrust law, will not be considered directly and are out of the scope for this article.

II. WHAT IS FEDERATED LEARNING?

The term FL was introduced in 2016,[18] and since then FL has been applied to multiple use cases. For instance, Google applied FL to Google keyboards (Gboard), and researchers used FL to analyze chest scans during the Covid-19 pandemic.[19] In a basic FL structure, a central server transmits the model to be trained (the 'raw model') to the various data silos (the 'local datasets')[20] and the raw model trains on data held in the local datasets (the 'raw data'). Once the training is completed, the trained model, or the relevant parameters of the model (together the 'model updates'), is transmitted back to the central server. The central server then amalgamates the different model updates into a final model. The result is that the final model is trained on the local datasets without transferring any of the raw data.[21] For the purpose of the analysis in this article, it will be assumed that that there is a single person or entity that controls the central servers, which will be referred to as the 'manager'. It is further assumed that the manager is also the person in charge of the overall FL project. The entities which control the local datasets and participate in the FL project will be referred to as the 'participants'. AI systems not trained using FL will be referred to as 'centrally trained AI systems'. There are different versions of FL. In 'decentralized FL' there is no central server coordinating the

[18] Priyanka Mary Mammen, *Federated Learning: Opportunities and Challenges*. In *Proceedings of ACM Conference (Conference '17)*, (2017), at 1. *But see*, Syreen Banabilah et al., *Federated learning review: Fundamentals, enabling technologies, and future applications*, 59, Information Processing and Management, Nove. 2022 at 2. (states that Google introduced FL in 2015).
[19] *Id*. 1.
[20] What this article refers to as local datasets is sometimes referred to as 'nodes'. *E.g.* Francesco Malandrino & Carla Fabiana Chiasserini, *Federated Learning at the Network Edge: When Not All Nodes are Created Equal* Jan 14 2021, arXiv 2101.05428, https://arxiv.org/abs/2101.01995 (last accessed July 26, 2024).
[21] FL is similar to learning. However, in distributed learning it is assumed that the local datasets are of similar size and contain data which is independently and identically distributed. By contrast FL does not make these assumptions. *See* Nguyen Truong et al. *supra* note 9, at 5.

actions of the local datasets.[22] Instead, the local datasets coordinate their action themselves, including the transmission of the raw model, as well as the amalgamation of the model updates into the final model. To keep the analysis of FL manageable, this article will only analyze FL where there is central server. The promise of FL is train AI models while preserving data privacy. Data privacy may also be preserved through other means, like anonymizing data before sharing or through differential privacy. However, data anonymization and differential privacy may entail some loss of information of the data, [23] and therefore, makes data potentially less suitable for the training AI systems. By contrast, FL is thought to improve accuracy.[24]

Currently, an important area of research in FL is improving security of FL.[25] One important security concern are, so called, "poisoning attacks".[26] In a poisoning attack a malicious participant may deliberately include wrong data in the data silos, or manipulate the model updates, in order to "decrease the overall performance of the system or to produce particular kinds of errors."[27] Another security concern are "privacy attacks" [28], where a malicious attackers tries to gain information about the raw data.[29] For instance, a third party may try to intercept the model updates to gain information about the local datasets or a malicious manager may attempt to gain information about the local datasets by analyzing the model updates. Technological solutions are available to address some of these issues, but thus far, these solutions can only ameliorate the problem rather than solve the problem.[30] This means that generally, FL provides more privacy protection than centralized training of AI systems, but FL's privacy protection is not absolute. In practice FL will often require an element of trust.[31] The participants need to trust that the manager does not attempt to gain information about the raw data through the model updates and the manager needs to trust that the participants not to engage in poisoning attacks.

Another aspect of FL is that AI systems trained through FL tend to be less transparent than centrally trained AI systems.[32] The distinguishing feature of FL is that the raw data, as well as the model updates, are inaccessible. This makes it more difficult to understand how the AI system works or why the system made certain decisions. A consequence of the AI system being less transparent is that it is usually more difficult to ensure fairness or lack of bias of the AI system because the data on

[22] Sometimes decentralized FL is referred to as 'gossip learning. *See e.g.*, Hegedűs István, Danner Gábor, Jelasity, Márk, *Decentralized learning works: An empirical comparison of gossip learning and federated learning*, 148, J PARALLEL DISTR COM, 109 (2021).
[23] *See* Nguyen Truong et al. *supra* note 9, 3-5
[24] Aimin Yang et al., *Review on application progress of federated learning model and security hazard protection*, 9 Digital Communications and Networks, 146, 149, (2023) (state that: "The use of federated learning ensures that the trained model results in better accuracy than traditional machine learning").
[25] S. Rossello, et al., *supra* note 9, 5.
[26] *Id*. 5, *see also id*. 5-6 (provides more details of poisoning attacks).
[27] *Id*. 4-5.
[28] *Id*. 6, *see also id*. 6-7 (provides more details of privacy attacks). Note that this article uses the term 'privacy attack' as generic reference for all cyber-attacks, which try to gain information about data.
[29] *Id*. 6-7.
[30] *Id*. 6.
[31] *See* Shuaicheng Ma et. al, Conference Proceedings, *Transparent Contribution Evaluation for Secure Federated Learning on Blockchain*, 2021 IEEE 37th International Conference on Data Engineering Workshops (ICDEW), 88 (Discussed issue of trust in FL, and proposes a solution based on blockchain to overcome trust issue).
[32] Nguyen Truong et al. *supra* note 9, at 15.

which the system was trained cannot be inspected to check for biases. Lack of transparency is an issue for all AI systems, not just for systems trained through FL.[33] However, because the raw data cannot be inspected, the transparency issues is more severe for FL. Technological solutions can ameliorate this problem to some extent.[34] Furthermore, a good governance structure can help to address to some extent the issue of transparency and fairness in FL. For instance, participants should keep comprehensive and accurate records on how the data silos were compiled. Nevertheless, it remains the case that AI systems trained using FL will be less transparent than centrally trained AI systems.

III. HOW DOES THE GDPR APPLY TO AI?

The previous part argued that, although FL does not ensure absolute data privacy, generally, FL is superior in preserving data privacy than centrally trained AI systems. However, the question whether FL preserves privacy for the purpose of the GDPR requires a legal analysis, but to examine how the GDPR applies to FL it is first necessary to analyze how the GDPR applies to the use of personal data in AI systems generally.

A. *GDPR overview*

When analyzing the GDPR it is important to note that EU legislation is interpreted purposively.[35] If a legislative text is not clear or ambiguous "it must be interpreted in light of the objectives it pursues and the courts should always seek to give effect to the legislative purpose/objective behind the law."[36] To determine the purpose/objective behind the law, one of the factors that a court may consider are the recitals of the legislation.[37] The recitals are not legally binding, they are part of the preamble of the legislation, and "set out concise reasons for the chief provisions of the enacting terms, without reproducing or paraphrasing them."[38]

---

[33] *Id.*
[34] *Id.*
[35] *E.g.*, Lord Sales, Justice of the Supreme Court of the United Kingdom, argues that: "Another important aspect derives from the open-textured form of EU legislation. This has the effect of casting responsibility onto the Court to identify the purposes underlying specific legislative regimes adopted by the EU institutions in order to give determinate meaning to the words enacted." Lord Sales, Justice of the Supreme Court of the United Kingdom, Annual Lecture of the UK Association for European Law, *Retained EU Law: Purposive Interpretation when the Constitutional Architecture Changes*, (Nov. 20, 2023), https://www.supremecourt.uk/docs/speech-231120.pdf (last accessed Jul. 17 2024); Practical Law (UK), *Recital (EU)*, Glossary, (2024). https://uk.practicallaw.thomsonreuters.com/Glossary/UKPracticalLaw/I13f404e5785211e79bef99c0ee06c731?transitionType=Default&contextData=(sc.Default)&firstPage=true (last accessed July 18, 2024).
[36] Llio HUMPHREYS et. al., *Mapping Recitals to Normative Provisions in EU Legislation to Assist Legal Interpretation*, in A. Rotolo, Legal Knowledge and Information Systems: JURIX 2015: the Twenty-Eighth Annual Conference, Nove. 24, 2015), 41, 41.
[37] Nial Fennelly, Legal Interpretation at the European Court of Justice, 20 Fordham Int'l L.J. 656 (1996), 666.
[38] *See*, Practical Law (UK), *supra* note 35.

Further, the courts may consider interpretive guidance issued by authoritative bodies. The GDPR established the European Data Protection Board (EDPB),[39] which has the task to issue non-legally binding opinions and guidance on the application and interpretation of the GDPR.[40] Similarly, the General Data Protection Directive (DPD),[41] the predecessor of the GDPR, established the Article 29 Working Party (WP29),[42] which like the EDPB was tasked with issuing opinions and guidance on the application and interpretation of the DPD. Although the WP29 was replaced by the EDPB in 2018,[43] the EDPB has endorsed a number of guidelines and other documents issued by WP29.[44] Thus, some of the opinions and guidance issued by the WP29 continue to be relevant.

The GDPR distinguishes between data subject, data controller and data processer. The data subject is the natural person to whom personal data relates.[45] The data controller is the person or entity which "determines the purposes and means of the processing of personal data".[46] The data processor is the person or entity "which processes personal data on behalf of the controller".[47] If the requirements for the territorial scope of the GDPR, as set out in article 3 are met, the processing of personal data must comply with the data protection principles, set out in article 5 of the GDPR, namely: lawfulness, fairness and transparency; purpose limitation; data minimization; accuracy, storage limitation, integrity and confidentiality. In order for the processing of personal data to be lawful, the processing must meet at least one of the grounds set out in article 6.1(a). If personal data is sensitive personal data,[48] the processing must further comply with the requirements set out in article 9. In addition, the GDPR sets out various rights of the data subject, including, rights to information of how data is process as well as rights of access to personal data,[49] rights in relation to profiling and automatic decision making,[50] a right to rectification and erasure,[51] and a right to demand a restriction of processing.[52]

---

[39] GDPR *supra* note 11, art. 68; *See also*, EDPB, *Legal Framework*, https://www.edpb.europa.eu/about-edpb/about-edpb/legal-framework_en (last visited Jul. 18, 2024).
[40] GDPR *supra* note 11, art. 70.1.
[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 O.J. (L 281) [hereinafter DPD].
[42] *Id.* art.29.
[43] EDPB, *Legacy: Art. 29 Working Party*, https://www.edpb.europa.eu/about-edpb/who-we-are/legacy-art-29-working-party_en (last accessed July 18. 2024).
[44] EDPB, *Endorsement 1/2018*, 2018, https://www.edpb.europa.eu/sites/default/files/files/news/endorsement_of_wp29_documents_en_0.pdf (last accessed July. 18. 2024).
[45] GDPR *supra* note 11, art. 4(1).
[46] *Id*. art. 4(7).
[47] *Id.* art. 4(8).
[48] Recital 10 of the GDPR states that sensitive personal data is personal data which falls within special categories. Article 9.1 of the GDPR defines special categories of personal data as "[p]ersonal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation."
[49] GDPR *supra* note 11, arts. 12-15.
[50] *Id.* arts. 21-22.
[51] *Id.* arts. 16-17.
[52] *Id.* art. 18.

B. *The difference between the training and application phase of AI systems*

Although, the GDPR does not contain the term 'artificial intelligence', there is an apparent tension between he GDPR's data protection principles and using personal data to train AI systems, in particular the principles of purpose limitation and data minimization.[53] As will be further analyzed below, [54] the purpose limitation principles means that if personal data is collected, it must be collected for a specific purpose. The data minimization principle means that no more data shall be collected than what is necessary for the particular purpose.[55] These principles are problematic for the training of AI systems because:

> [T]he requirements of purpose limitation and data minimisation are not always feasibly carried out in ML-based systems. The majority of ML algorithms heavily rely on data quality and quantity, thus researchers tend to collect as much related data as possible. Therefore, determining 1) the purposes of data collection as well as 2) what data is adequate, limited, and relevant only to the claimed purposes before executing such ML algorithms are problematic challenges.[56]

The severity of the tension between the GDPR and AI is a matter of ongoing debate, with some commentators arguing that the using personal data for the training of AI systems is incompatible with the GDPR;[57] while others, argue to that the GDPR does not pose substantial obstacles to the training of AI systems.[58] If the GDPR is indeed incompatible with the development of AI systems, then this could have severe consequences for the AI industry in Europe,[59] which may mean that the GDPR would need to be amended or replaced.[60] For the purpose of this article, if the GDPR is incompatible with the training of AI systems, it would mean that FL is also incompatible with the GDPR because FL is a way of training AI systems.

This article will argue that the GDPR can be interpreted to be compatible with the training of AI systems on personal data and that it is indeed likely that the GDPR will be interpreted in this way in order "to reconcile both desiderata: protecting data subjects and enabling useful applications of AI."[61] The key to resolving the tension between the GDPR and AI development is to make a distinction between the use of personal data when an AI system is trained (the 'training phase') and

---

[53] *E.g.* EPRS Study *supra* note 16, *Executive Summary*, at II.
[54] *See infra* section on *Purpose limitation.*
[55] *See infra* section on. *Data minimization*.
[56] Nguyen Truong et al. *supra* note 9, at 5 (note, 'ML' means machine learning).
[57] *E.g.*, Tal Z. Zarsky, *Incompatible: The GDPR in the Age of Big Data*, 47 Seton Hall L. Rev, 995, (2017).
[58] *E.g.*, Viktor Mayer-Schönberger & Yann Padova, *REGIME CHANGE? ENABLING BIG DATA THROUGH EUROPE'S NEW DATA PROTECTION REGULATION*, 17 SCI. & TECH. L. REV., 315 (2016).
[59] EPRS Study supra note 16, at 76.
[60] *Id.*
[61] *Id.*

using personal data as input into an AI system when the AI system is used or applied (the 'application phase'), as for several reasons the data protection issues in the two phases are very different.

First, the sheer volume of data on which AI systems are typically trained means that a single data point is usually of little significance. This means that, whether or not a single individual's personal data is included in the training dataset of an AI system, usually makes little difference to the way the AI system will operate. This conclusion was also reached by a study on "The Impact of the General Data Protection Regulation (GDPR) on AI", [62] managed by the Scientific Foresight Unit, within the Directorate-General for Parliamentary Research Services (EPRS) of the Secretariat of the European Parliament (the 'EPRS Study'), which states that:

> In general, the inclusion of a person's data in a training set is not going to affect to a large extent that particular person, since the record concerning a single individual is unlikely to a make a difference in a model that is based in a vast set of such records.[63]

Second, the overall objective of training AI systems is that the system can be applied to data outside the dataset on which it was trained,[64] as "[m]achine learning systems discover correlations between data and build corresponding models, which link possible inputs to presumably correct responses (predictions)."[65] Thus. for inferences to be drawn about an individual, an AI system must be applied to a particular case, i.e. an input must be given to the AI systems, and the AI system then links the input to an output, which the systems has 'learned' by being trained on vast amounts of data. For example, assume an AI system is trained on MRI scans of a certain people to predict the likelihood of a person developing cancer. This AI system by itself does not contain information about a particular individual's likelihood to develop cancer. In order to gain information about an individual's likelihood to develop cancer, the trained AI systems must be applied to the MRI scan of and individual. An analogy is to regard AI systems as a tools, and the training of an AI system as forging a tool. A tool by itself does not do anything, a tool only has an impact once it is used. Similarly, the training of an AI system, does not reveal information about an individual, unless there is a successful privacy attack. Instead, the training of an AI system, creates a tool, which may allow information about an individual to be revealed, when the AI system is used.

Nevertheless, the inclusion of personal data in a training set, exposes the risk that the personal data may be abused i.e. the personal data is used to draw inferences about an individual.[66] To continue with the previous example, if a person's MRI scan is used in the training set of an AI system, this exposes said person to the risk that her MRI scan is used for a purposes other than to train the AI

[62] EPRS Study *supra* note 16.
[63] *Id.* at 46.
[64] *E.g.* Yun Xu &Royston Goodacre1, *On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning*, 2(3) J Anal Test., 249, (2018) (gives an overview of how training and validation datasets are used in machine learning).
[65] EPRS Study *supra* note 16, at 1.
[66] *See Id.* at 44 (the EPRS Study makes the same point)

system. In particular, the person is exposed to the risk that the trained AI system is applied to her scan to make a prediction about her likelihood of developing cancer, which could, inter alia, impact the person's insurability or the insurance premium. Thus, if personal data is used in the training phase of an AI system the relevant data protection issue is whether the personal data is kept secure from abuse, but the training of AI system *per se* does not raise data protection issues.

That there is an important distinction between the data protection issues in the training phase and application phase is increasingly being recognized. The WP29 states that in relation to Big Data two phases need to be distinguished. "In the first one, the organisations processing the data want to detect trends and correlations in the information. In the second one, the organisations are interested in individuals."[67] Similarly, the taskforce set up by the EDPB to investigate ChatGPT,[68] argues that different phases need to be distinguished when analyzing how the GDPR applies to AI, namely: "i) collection of training data (including the use of web scraping data or reuse of datasets) ii) pre-processing of the data (including filtering), iii) training, iv) prompts and ChatGPT output as well as v) training ChatGPT with prompts."[69] The importance of distinguishing between the use of personal data in the training phase and application phase of AI systems has also been recognized by non-European regulators. For instance, the Singapore Personal Data Protection Commission (PDPC) issued the "Advisory Guidelines on use of personal data in AI Recommendations and Decision Systems"[70] in which the PDPC argues that distinction should be made between personal data being used in the training phase and application phase of AI systems.[71]

C. *Compatibility of AI development with the GDPR principles*

This section will analyze in detail some of the provision of the GDPR which are thought to cause difficulties for the development of AI system. This analysis will show that if one takes into account that the data protection issues of using personal data in the training phase of AI systems are different than the data protection issues of using personal in the application phase of AI systems, it is possible to reconcile the GDPR with the development of AI systems.

---

[67] ARTICLE 29 DATA PROTECTION WORKING PARTY, *Opinion 03/2013 on purpose limitation*, 00569/13/EN WP 203, at 46 [hereinafter WP29 Opinion on Purpose Limitation]
[68] *See* European Data Protection Board,, *Agenda 89th EDPB meeting*, at 1 (Jan. 16, 2024) https://www.edpb.europa.eu/system/files/2024-01/20240116plenagenda_public.pdf (last accessed July 26, 2024)
[69] ChatGPT taskforce, *Report of the work undertaken by the ChatGPT Taskforce*, at 6 (May 23, 2004) https://www.edpb.europa.eu/system/files/2024-05/edpb_20240523_report_chatgpt_taskforce_en.pdf (last accessed July 18, 2024) [hereinafter ChatGPT Report]
[70] Personal Data Protection Commission, *Advisory Guidelines on use of personal data in AI Recommendations and Decision Systems*, (May 1, 2024) https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/advisory-guidelines/advisory-guidelines-on-the-use-of-personal-data-in-ai-recommendation-and-decision-systems.pdf (last accessed July 18, 2024)
[71] *Id.* at Part III and Part IV

1. Purpose limitation

One of the GDPR provisions that is thought to be difficult to reconcile with the development of AI is the purpose limitation principle, in article 5.1(b) of the GDPR, which states that data shall be:

> [C]ollected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes shall, in accordance with Article 89(1), not be considered to be incompatible with the initial purposes.

The issue is that to train AI systems requires vast amount of data and often it will not be clear at the time of collection of the data what the specific purpose for the collection is.[72] AI systems may also discover new connections in the data that may not be anticipated or foreseen at the time when the data was collected,[73] which may be regarded as contrary to the purpose limitation principle.

A possible way to reconcile the purpose limitation principle with the development of AI systems is to obtain the data subject's consent to a broad purpose, at the point when the personal data is initially collected, to ensure that any further purpose is compatible with the initial purpose.[74] However, for consent to be valid under the GDPR, consent needs to be specific,[75] and it is unlikely that a broad or generic consent, such as for personal data to be used for the training of AI systems, will be valid.[76] Moreover, the purpose limitation principle requires that personal data shall be "collected for specified, explicit and legitimate purposes".[77] Thus, even if consent were given for personal data to be used for an unspecified purpose this would not satisfy the principle of purpose limitation. Further, the sheer number of data subjects who would need to give consent, makes it difficult to rely on consent in practice.[78] There will be some cases where consent will be an appropriate way to reconcile the training of an AI system with the purpose limitation principle, but generally, consent to a wide purpose will not be sufficient to meet the purpose limitation principle.

A better way to reconcile the purpose limitation principle with the development of AI systems is to rely on the qualification of the purpose limitation principle in article 5.1(b), which allows further

---

[72] Nguyen Truong et al. *supra* note 9, at 5.
[73] *See* Zarsky *supra* note 56,1006 (Zarsky makes a similar point about in relation to big data analytics)
[74] *See* Viktor Mayer-Schönberger & Yann Padova *supra* note 57, at 325-326 (The authors point out that the giving of broad consent could be a solution to the problems posed by the purpose limitation principle but also argue that it will be difficult for companies to attain such broad consent).
[75] GDPR *supra* note 11, art. 4(11).
[76] WP29 Opinion on Purpose Limitation *supra* note 66, at 17-18, (WP29 states the importance for the purpose to be specific, inter alia so that individual can make an informed choice).
[77] GDPR *supra* note 11, art. 5.1(b).
[78] *See* Viktor Mayer-Schönberger & Yann Padova *supra* note 57, at 325-326 (The authors point out that the giving of broad consent could be a solution to the problems posed by the purpose limitation principle but also argue that it will be difficult for companies to attain such broad consent).

processing provided the purpose of the further processing is not incompatible with the initial purpose for which the data was collected. Note that article 5.1(b) phrases this qualification as a double negative, requiring that further processing 'not to be incompatible' with the original purpose. In contrast, recital 50 of the GDPR states that the further processing exemption requires that the processing is "compatible" with the initial purpose. Nevertheless, the WP29 opined that the use of the double negative indicates "that the legislators intended to give some flexibility with regard to further use."[79]

Article 6.4 of the GDPR provides guidance on how to determine whether the processing is compatible with the initial purpose. It states that the controller shall, inter alia, consider the following:

> (a) any link between the purposes for which the personal data have been collected and the purposes of the intended further processing;
> (b) the context in which the personal data have been collected, in particular regarding the relationship between data subjects and the controller;
> (c) the nature of the personal data, in particular whether special categories of personal data are processed, pursuant to Article 9, or whether personal data related to criminal convictions and offences are processed, pursuant to Article 10;
> (d) the possible consequences of the intended further processing for data subjects;
> (e) the existence of appropriate safeguards, which may include encryption or pseudonymisation.[80]

When applying article 6.4 it is important to take into account the distinction between the training phase and the application phase of an AI system. According to article 6.4(d) "the possible consequences of the intended further processing for data subjects" is a relevant factor for whether further portions is not incompatible with the initial purpose. As stated above, in the training phase, provided the personal data is kept secure from abuse, the impact on data subjects' interests is minimal or not existing. Thus, according to article 6.4(d), provided appropriate safeguards are in place,[81] using

[79] WP29 Opinion on Purpose Limitation *supra* note 66, at 21.
[80] *See also*, recital 50 of the GDPR which states that to assess whether purpose limitation principle is complied with, the data controller should take into account "any 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."; *See also*, WP29 Opinion on Purpose Limitation *supra* note 66, at 3, which suggests the following criteria to be relevant for the purpose limitation principle:

> - the relationship between the purposes for which the personal data have been collected and the purposes of further processing;
> - the context in which the personal data have been collected and the reasonable expectations of the data subjects as to their further use;
> - the nature of the personal data and the impact of the further processing on the data subjects;
> - the safeguards adopted by the controller to ensure fair processing and to prevent any undue impact on the data subjects.

[81] Whether there are "appropriate safeguards" is also specifically mentioned as a factor to be taken into account pursuant 6.4(e) of the GDPR.

personal data to train AI system is not incompatible with the initial purpose.

To what extent the criteria stated in articles 6.4(a) to 6.4(c) are met will depend on the specifics of each case. The EPRS Study argues that because in the training phase of an AI system the impact on a data subject is minimal, article 6.4(d) should take priority over the other factors listed in article 6.4.[82] It is submitted that this is the correct interpretation. Article 6.4 does not set out an exhaustive list of factors to be considered,[83] and neither does article 6.4 state how the different factors are to be balanced against each other. Article 6.4 also does not state that for the further purpose not to be incompatible with the original purpose all the factors in 6.4 need to be met, rather article 6.4 states factors that shall be taken into account. Further, the overarching objective of the GDPR is "[t]he protection of natural persons in relation to the processing of personal data",[84] but if the processing of personal data has no impact on a natural person, then no protection is required.[85] A purposive interpretation of article 6.4, supports the conclusion reached by the EPRS Study, that in case of using personal data in the training phase of an AI system, the fact that the impact on the data subject is minimal ought to be given priority over other the other factors.

This interpretation of the purpose limitation principle is further supported by article 5.1(b), which states that data processing for "statistical purposes shall, in accordance with Article 89(1), not be considered to be incompatible with the initial purposes". The reason that the use of personal data for statistical purposes is not incompatible with the initial purposes is that the result of processing of personal data for statistical purpose is "aggregate data", [86] which means that it is not personal data.[87] A similar argument can be made in relation to using personal data in the training phase of AI systems. When training on data, an AI system 'learns' correlations between data points, for the purpose of applying these correlations to new cases. Thus, when training on personal data the AI systems 'learns' aggregate data and not personal data.[88] Therefore, if processing for statistical purpose is not incompatible with the initial purpose, it should also be the case that processing of personal data to train AI systems, should be regarded as not incompatible with the original purpose, provided adequate safeguards for the personal data are in place.

Zarsky, however, points out a potential problem with this argument, which is that pursuant to recital 162, "[t]he statistical purpose implies that the result of processing for statistical purposes is not personal data, but aggregate data, and that this result or the personal data are not used in support of measures or decisions regarding any particular natural person." Zarsky argues that this means that many AI or Big Data systems cannot be regarded as processing for statistical purposes, because their

[82] EPRS Study *supra* note 16, at 47 states the "original purpose should not be a primary concern, nor should be the data subject's expectations".
[83] Article 6.4. of the GDPR states the list of factors which the data controller need to 'take into account, inter alia"
[84] GDPR *supra* note 11, recital 1.
[85] WP29 states in WP29 Opinion on Purpose Limitation *supra* note 66, at 25, that: "Data protection law … has after all been designed to protect individuals against the impact of improper or excessive use of their personal data." 5
[86] GDPR *supra* note 11, recital 162.
[87] *Id.*
[88] *E.g*. Srinivas Martha, Martha Nuthana Priya, *Role of Statistics in Artificial Intelligence*, 7 IJEAST, 96 (2022) (discussing the use of statistics in AI).

results are often "in support of measures or decisions regarding . . . natural person[s]".[89] For instance, a system that statistically analyses the shopping patterns of individuals to offer them different prices for certain items, would not fall within the statistical purpose exemption because the result "is used in support of measures or decisions regarding . . . natural person[s]", namely the decision what price to offer to an individual.[90]

As argued previously, the principle behind the statistical purpose exemption is the same principle, which states that using personal data for training AI systems is not incompatible with the initial purpose. Thus, if the statistical purpose cannot be "in support of measures or decisions regarding any particular natural person", then the same should apply for the training of AI systems. Therefore, if Zarsky's argument is correct, then recital 162 could seriously limit the extent to which the training of AI systems on personal data may fall within the 'not incompatible' exemption. Yet, it is unlikely that recital 162 can be given the interpretation suggested by Zarsky.

First, it is important to note that wording, on which Zarsky's argument is based on, is in a recital. Recitals are not legally binding and to be used as an aid to interpretation.[91] If the phrase "in support of measures or decisions regarding any particular natural person" is given a broad interpretation, then almost no statical analysis would fall within the 'statistical purpose exemption'. Little statistical analysis is done purely for the sake of gaining knowledge, most statistical analysis is used "in support of measures or decisions regarding . . . natural person[s]." For instance, even a statical analysis to forecast the weather is usually not done purely to gain knowledge, as weather forecasts may impact decisions, such as a person's decision whether to go hiking or not.

A typical example where the statistical purpose exemption applies is a shop using customer data to determine whether the shop gained or lost customers over a certain period of time.[92] Yet, this analysis may also be used "in support of measures or decisions regarding any particular natural person." For instance, if the shop gained more customers, the business may decide to hire a builder to enlarge its business, or the shop may use these statistics in the appraisal of managing staff. If the 'statistical purpose exemption' excludes any statistics used "in support of measures or decisions regarding any particular natural person", then barely any statistics would fall within the 'statistical purpose exemption'. If one applies a purposeful interpretation to the phrase "in support of measures or decisions regarding any particular natural person" a more plausible interpretation is that the statistical purpose exemption does not apply when the statistics is used "in support of measures or decisions regarding" the persons, whose data was used to create the statistical analysis.

Second, Zarsky's interpretation of recital 162 is not supported by the WP29, which states "that data used for statistical purposes or other research purposes should not be available to support measures or decisions' *that are taken with regard to the individual data subjects concerned* (unless

[89] *See* Zarsky *supra* note 56, at 1008. (Zarsky analysis Big Data analytics but his arguments equally apply to the training of AI systems).
[90] *See* Viktor Mayer-Schönberger & Yann Padova *supra* note 57, 323 (gives a similar example).
[91] *See supra*, section *GDPR overview*.
[92] Viktor Mayer-Schönberger & Yann Padova *supra* note 57, 323 (paper gives a similar example).

specifically authorized by the individuals concerned)."[93] Note that the WP29 uses the expression "individual data subjects concerned" rather than "any particular natural person", the former expression suggesting a narrower interpretation than the latter. The WP29 seems to suggest that the relevant individuals for the statistical purpose exemption are the data subjects, whose data was used to create the statistics. The WP29 further states that "[t]o comply with this requirement, controllers need to guarantee the security of the data".[94] To require security of data only makes sense, if the expression "in support of measures or decisions" refers to the data subjects, whose data was used to compile the statistics. Security of data means that the data cannot be used for a purpose, different from the intended purpose, in particular it means that personal data cannot be used to draw inferences about data subjects, to whom the personal data pertains. This indicates that the WP29 regards the no "in support measures or decisions" requirement to mean that "no support measures or decisions" shall be taken about an individual, whose personal data was used to compile the statistics.

The WP29 also provides specific guidance on the application of the GDRP to big data analytics,[95] and due to the similarity with the training of AI systems, [96] the guidance is also relevant for AI systems. The WP29 introduces a distinction between two scenarios. "In the first one, the organizations processing the data want to detect trends and correlations in the information. In the second one, the organizations are interested in individuals."[97] This is similar to the distinction between the training phase and application phase of an AI system. The WP29 continues as follows:

> In the first scenario, the concept of *functional separation* is likely to play a key role, and the extent to which this may be achieved could be an important factor in deciding whether further use of the data for (marketing or other) research can be considered compatible. In these cases, data controllers need to guarantee the confidentiality and security of the data, and take all necessary technical and organisational measures to ensure functional separation.[98]

According to the WP29 the concept of functional separation refers "full or partial anonymisation, pseudonymisation, or aggregation of the data, privacy enhancing technologies, as well

[93] WP29 Opinion on Purpose Limitation *supra* note 66, at 30 (italics added.) Note, however, that the WP29 opinion refers to the provision in the DPD not the GDPR. *See* discussion *supra* section *GDPR overview.* Also note that EDPB states that: "The Article 29 Working Party provided guidance for the understanding of the principle of purpose limitation under Directive 95/46/EC. Although the Opinion is not adopted by the EDBP, it may still be relevant as the wording of the principle is the same under the GDPR." European Data Protection Board Guidelines, *Guidelines 4/2019 on Article 25 Data Protection by Design and by Default*, at 19, (Oct. 20, 2020, version 2.0) [hereinafter EDPB Guidelines 4/2019].
[94] WP29 Opinion on Purpose Limitation *supra* note 66, at 30.
[95] WP29 Opinion on Purpose Limitation *supra* note 66, Annex II.
[96] The WP29 states that "Big data' refers to the exponential growth in availability and automated use of information: it refers to gigantic digital datasets held by corporations, governments and other large organisations, which are then extensively analysed using computer algorithms. Big data relies on the increasing ability of technology to support the collection and storage of large amounts of data, but also to analyse, understand and take advantage of the full value of data (in particular using analytics applications). The expectation from big data is that it may ultimately lead to better and more informed decisions." (WP29 Opinion on Purpose Limitation *supra* note 66, at 45). This is very similar to the current iteration of AI.
[97] WP29 Opinion on Purpose Limitation *supra* note 66, 46.
[98] *Id.* (internal footnotes omitted).

as other measures to ensure that the data cannot be used to take decisions or other actions with respect to individuals".[99] The idea of functional separation is important, as it suggests that a dataset, which comprises personal data, contains different 'types' of information. One type of information is information about identifiable individuals. For instance, a dataset containing MRI scans, which have personal identifiers attached to them (e.g. patient's name, data of birth, ZIP code) contains information about individuals' health. Yet, the dataset also contains another 'type' of information, namely information about possible correlations between the MRI scans. The GDPR protects the former type of information but not the latter. Usually if a person controls a dataset, the person has access to both types of information. If the types of information can be functionally separated – for example through "anonymization, pseudonymization or the aggregation of the data" – then the GDPR only applies to information which relates to an identifiable individual. AI systems are essentially tools "to capture the relevant knowledge originally embedded in the training set, namely the correlations between cases and responses."[100] Therefore, provided the data on which AI systems are trained is kept secure (i.e. there is effective functional separation), when AI systems train on personal data they train on the type of information, which consists of the correlations between data points, and not on the type of information, which is information about individuals. Thus, the training of AI systems on personal data should not be regarded as incompatible with the original purpose.

2. Data minimization

Another GDPR provision which is thought to be incompatible with the deployment of AI is the data minimization principle contained in Article 5.1(c), which requires that data collection must be "adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed". The data minimization principle further requires that personal data can only be retained for as long as necessary for a specific purpose.[101] Yet, AI developers may want to retain data for longer in order to use it to train other AI systems. Furthermore, it seems difficult to apply the standard that data should be limited to what is necessary to the training of AI systems because when AI systems are trained on data there is no clear cut-off point for how much data is necessary to train AI systems. Often it will be the case that more data is better for the training of AI systems but one cannot clearly say that a certain amount of data is necessary or sufficient. Nevertheless, if one makes the distinction between personal data used in the training phase and application phase of AI systems, the data minimization principle can be interpreted as compatible with the training of AI systems.

Different from the purpose limitation principle, the data minimization principle is not subject to an explicit qualification, such that data for statistical or research purpose may be maintained for a

[99] *Id*. at 27.
[100] EPRS Study *supra* note 16, at 11.
[101] Note that there is an overlap with the purpose limitation principle.

longer period of time. A possible reason for the lack of explicit qualification is that the data minimization principle already contains a qualification by using the phrase "adequate, relevant and limited to what is necessary". This phrase makes it unnecessary to have a qualification because each type of processing has a different standard for how much data is adequate, relevant and necessary. Thus, the absence of an explicit qualification for the data minimization principle should not be regarded as implying a legislative intention to have more stringent standards for data minimization than for the legitimate purpose principle.

In practice, it is likely that the data minimization principle needs to be applied using a proportionality test.[102] According to the data minimization principle, data has to be "adequate, relevant and limited to what is necessary", but this "does not exclude the inclusion of additional personal data in a processing, as long as the addition of such data provides a benefit".[103] Yet, there has to be a limit to how much data can be collected, even if collecting of additional data provides a benefit. If this were not the case, any amount of data could be collected as long as the additional data has some benefit, which would make the data minimization principle meaningless. Recital 4 of the GDPR states that "[t]he right to the protection of personal data is not an absolute right; it must be considered in relation to its function in society and be balanced against other fundamental rights, in accordance with the principle of proportionality." If the right to protection of personal data as a whole is subject to a proportionality principle, which involves balancing competing rights, then *a fortiori* the data minimization principle is subject to principle of proportionality. A straightforward way to apply the principle of proportionality to the data minimization principle is to balance the benefits of collecting the additional data with the impact such collection has (or may have) on the data subject. A similar conclusion was also reached by the EDPB, stating that:

> Controllers should consider both the volume of personal data, as well as the types, categories and level of detail of personal data required for the processing purposes. Their design choices should take into account the increased risks to the principles of integrity and confidentiality, data minimisation and storage limitation when collecting large amounts of detailed personal data, and compare it to the reduction in risks when collecting smaller amounts and/or less detailed information about data subjects.[104]

Furthermore, even Zarsky, who argues that the GDPR and AI development are incompatible, agrees that article 89.1[105] and recital 156 of the GDPR imply that for "statistical purpose, the GDPR

[102] EPRS Study *supra* note 16, at 47.
[103] *Id.*
[104] EDPB Guidelines 4/2019 *supra* note 92, at 12.
[105] Article 89.1 of the GDPR states: "Processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes, shall be subject to appropriate safeguards, in accordance with this Regulation, for the rights and freedoms of the data subject. Those safeguards shall ensure that technical and organisational measures are in place in particular in order to ensure respect for the principle of data minimisation. Those measures may include pseudonymisation provided that those purposes can be fulfilled in that manner. Where those purposes can be fulfilled by further processing

concedes that data minimization could be achieved by pseudonymization; applying technological and statistical safeguards which will not allow for the identification of the data subjects."[106] If for statistical purposes data minimization can be achieved through pseudonymization, then a similar principle should be applied to the training of AI systems. As stated above, the reasons why statistical analysis is largely outside the scope of the GDPR[107] is because statistical analysis deals with aggregate not personal data. Therefore, if statistical purposes "may be subject to looser minimisation requirements"[108] then the same considerations should also apply to the training of AI systems. This means that if it can be demonstrated that including more data in the training set for an AI system has benefits (e.g. making the AI system more accurate), and these benefits outweigh the impact on the data subject, then the inclusion of this additional data does not infringe the data minimization principle. As was argued previously, provided the data is kept secure, using personal data in the training phase of an AI system has a minimal impact on the data subject. Therefore, training of AI systems on personal data, provided the data is secure, will usually confirm to the data minimization principle.

3. Lawfulness of processing data

In order to process data lawfully, the processing must meet at least one of the six criteria stated in article 6.1 of the GDPR. Pursuant to article 6.1(a), processing of personal data is lawful if it is done with the data subject's consent. However, for consent to be valid under the GDPR it has to be "freely given, specific, informed and unambiguous",[109] and as previously argued, attaining valid consent for the use of personal data for the training of AI systems poses serious challenges.[110] The grounds listed in articles 6.1(b) – 6.1(e) are grounds where the processing of personal data is necessary for specific purposes, such as fulfilling contractual obligations (article 6.1(b)), compliance with legal obligations (article 6.1(c)) or to protect the vital interests of data subject (article 6.1(d)). These grounds can be relied on for the development of AI systems in specific circumstances, but due to their specificity they are unlikely to be widely used in relation to training of AI systems.

This leaves article 6.1(f) (legitimate interest) as the most likely ground for the lawful processing of personal data for the purposes of training AI systems.[111] Article 6.1(f) states that the processing of personal data is lawful provided that:

---

which does not permit or no longer permits the identification of data subjects, those purposes shall be fulfilled in that manner."

[106] Zarsky *supra* note 56, at 1011. The EPRS Study similarly states that "the processing of personal data for merely statistical purposes may be subject to looser minimisation requirements" (EPRS Study *supra* note 16, at 47)

[107] *E.g.* EPRS Study *supra* note 16 at 48 (makes a similar point).

[108] EPRS Study *supra* note 16, 47.

[109] GDPR supra note 11, art. 4.1.

[110] *See supra*, discussion in section on *Purpose limitation*.

[111] For instance, OpenAI uses legitimate interest as the basis processing personal data collected. *See*, EDPB ChatGPT Report, *supra* note 68, at 6.

> processing is necessary for the purposes of the legitimate interests pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of personal data, in particular where the data subject is a child.

The WP29 makes it clear that the legitimate interest ground for the purpose of the DPD[112] is not a ground that can "only be used sparingly . . . as 'a last resort' ".[113] On the other hand, the legitimate interest ground must not be interpreted so widely as to create a loophole in the GDPR, allowing almost all forms of processing of personal data.[114] The WP29 opined that the legitimate interest ground requires a balancing test, where the legitimate interest of the data controller or third party is to be balanced against the impact on the data subject.[115] A similar conclusion was also reached by the EDPB, which states that for article 6.1.(f) to apply, the data controller needs to show that: i) there was an interest; ii) the interest was legitimate; iii) the processing was necessary and iii) the fundamental rights of the data subject and the legitimate interest of the controller must be balanced.[116]

The WP29 suggests a broad interpretation of what constitutes an interest,[117] the essential requirement being that it must "represent a real and present interest (i.e. not be speculative)."[118] The training of AI systems clearly qualifies as a real and present interest. Similarly, the WP29 suggests a wide interpretation for what constitutes a legitimate interest, stating that "legitimate interest could include a broad range of interests, whether trivial or very compelling, straightforward or more controversial"[119] and it includes "economic interests of a company".[120] Furthermore, the WP29 states that "an interest can be considered as legitimate as long as the controller can pursue this interest in a way that is in accordance with data protection and other laws. In other words, a legitimate interest

---

[112] Article 7.1(f) DPD provides:

> processing is necessary for the purposes of the legitimate interests pursued by the controller or by the third party or parties to whom the data are disclosed, except where such interests are overridden by the interests for fundamental rights and freedoms of the data subject which require protection under Article 1.

Article 6.1(f) of the GDPR provides:

> processing is necessary for the purposes of the legitimate interests pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of personal data, in particular where the data subject is a child.

Thus article 7.1(f) of the DPA and article 6.1(f) of the GDPR are substantially similar, therefore, the WP29 opinion can be regarded as persuasive. The EPDB also specifically referred to the WP29 opinion. *See* e.g. European Data Protection Board Guidelines, *Guidelines 3/2019 on processing of personal data through video devices*, version 2.0, at 10 (Jan. 29, 2029) [hereinafter EPDB Guidelines on Video Devices].

[113] ARTICLE 29 DATA PROTECTION WORKING PARTY *Opinion 06/2014 on the notion of legitimate interests of the data controller under Article 7 of Directive 95/46/EC*, 844/14/EN WP 217 (Apr. 9, 2014) [hereinafter WP29 Legitimate Interest Opinion].

[114] Irene Kamara & Paul De Hertat, Working Paper, UNDERSTANDING THE BALANCING ACT BEHIND THE LEGITIMATE INTEREST OF THE CONTROLLER GROUND: A PRAGMATIC APPROACH. Brussels Privacy Hub, Vol 4 Num. 12, at 5 (Aug. 2018).

[115] WP29 Legitimate Interest Opinion *supra* note 112, at 23.

[116] EDPB ChatGPT Report *supra* note 68, at 6.

[117] WP29 Legitimate Interest Opinion *supra* note 112, at 24.

[118] *Id.*

[119] *Id.*

[120] *Id.*

must be 'acceptable under the law'."[121]

The WP29's interpretation of article 6.1(f) may seem deficient because had the legislature intended to establish that data processing in accordance with the law is a ground for processing data lawfully, the legislature would have used language to that effect. Instead, the legislature used the phrase "legitimate interest". This apparent deficiency finds its solution in a footnote, where the WP29 states that it uses a very wide definition of term law, including "customs, codes of conduct, codes of ethics, contractual arrangements, and the general context and facts of the case."[122] Thus, an interest which is 'acceptable under the law' is not simply any interest, which is legal and can be pursued in a legal manner, but it must also be an interest in conformity with, customs, relevant code of ethics and so on. The 'necessity requirement' has been interpreted to mean that "that processing of personal data is the least restrictive measure to the rights of the data subjects."[123] Taken together, it is clear that the notion of legitimate interest is to be interpreted widely. Thus, many instances of training of AI systems will meet the definition of legitimate interest.

The last step to determine whether article 6.1(f) applies is to balance the legitimate interest of the data controller or third party with the interests of the data subject. Note that, different from the interest of the data controller or third party, there is no requirement that the data subject's interest needs to be legitimate.[124] The WP29 suggests that one of the relevant factors in the balancing test is how compelling the legitimate interest of the data controller or third party is. [125] For instance, if personal data is used for research purposes, likely to benefit society as whole, this will constitute a compelling interest. Further, the WP29 suggests that another relevant factor for the applicability of article 6.1(f) is the extent of the impact on the data subject.[126] The EDPB also states that in, what this article calls the training phase,[127] "adequate safeguards play a special role in reducing undue impact on data subjects and can therefore change the balancing test in favor of the controller".[128]

This has implications for the application of article 6.1(f) to the use of personal data for the training of AI systems. According to the criteria stipulated by the WP29, the training of AI systems is likely to constitute a legitimate interest, either because it is for the economic benefit of a company or for the benefit of research. Further, if the data is kept secure from abuse, the impact on the data subject of having her personal data included in the training set for an AI system is minimal. This makes it largely unnecessary to decide whether the training of an AI system constitutes a compelling interest or not as regardless of whether the training of AI system is regarded as compelling, the

[121] *Id.* at 25 (internal references omitted).
[122] *Id.* at 25, footnote 48.
[123] Irene Kamara & Paul De Hertat *supra* note 113, at 14.
[124] *Id.* at 12.
[125] WP29 Legitimate Interest Opinion *supra* note 112, at 26.
[126] *Id.*
[127] The EDPB refers to "Collection of training data, pre-processing of the data and training". EDPB ChatGPT Report *supra* note 68, at 6-7.
[128] *Id.*, WP29 Legitimate Interest Opinion *supra* note 112, at 31, 42-43 (also emphasizes the significance of adequate safeguards).

impact on the data subject is minimal, provided the data is kept secure. Therefore, the result of the balancing test should favor allowing AI system to be trained on personal data.

4. Information obligations

Articles 13 and 14 places a duty on the data controller to provide the data subject with certain information when the personal data is collected or obtained. Article 13 applies if the data is collected or obtained directly from the data subject. Article 14 applies if the data is not collected or obtained from the data subject. In both cases, the data controller must, inter alia, provide "the identity and the contact details of the controller" (articles 13.1(a) and 14.1(a)) as well as "the purposes of the processing for which the personal data are intended as well as the legal basis for the processing" (articles 13.1(c) and 14.1(c)). If the processing is based on the legitimate interests ground pursuant to article 6.1(f), the data subject must be informed about the nature of legitimate interests (articles 13.1(d) and 14.2(b)).

Article 14(b) provides that the data controller is not required to provide such information to the data subject if it "proves impossible or would involve a disproportionate effort". Surprisingly, article 13 does not contain such a provision. The EPRS Study opined that "[i]t is hard to understand why this is the case. In fact, the reasons that justify an exception to the information obligation when the data were not obtained from the data subject, should also justify the same exception when the data were collected from him or her."[129] A possible reason why article 13 does not contain an exception to the information obligations is that the legislator may have presumed that if the data is collected from the data subject, then the data controller is in direct contact with the data subject, and therefore, it will not be the case that it is "impossible or would involve a disproportionate effort" to provide the data subject with the required information. If this is the case, then the scope of article 13 should be construed narrowly, meaning that article 13 applies if the data controller and data subject are in direct contact with each other, and therefore, the data controller can provide the data subject with the required information without undue effort. In all other cases, article 14 applies and the data controller can rely on the exception to the information obligation pursuant article 14.4(b). This should make it easier for AI developers to rely on the information obligation exemption. For instance, if data is collected through, what is sometimes referred to as, 'web scraping', which is "the automated collection and extraction of certain information from different publicly available sources on the Internet (such as websites)", [130] it should be possible to rely on the information obligation exemption. As the data is not collected directly from the data subjects but from websites. Nevertheless, it is important to point out the "impossible or would involve a disproportionate effort" exemption to the information obligations

[129] EPRS Study *supra* note 16, at 54.
[130] EDPB ChatGPT Taskforce, supra *note* 68, at 6.

must not be treated as a loophole i.e. AI developers must take seriously the obligations to provide information to the data subjects.

5. Right to object and right to erase

Article 21.1 of the GDPR allows the data subject to object at any time to the processing of personal data "on grounds relating to his or her particular situation" if the data processing is based on grounds 6.1(e) ('processing in the public interest') or 6.1(f) ('legitimate interest'). According to article 21.6 the right to object also applies to the processing for statistical purposes. In relation to processing on grounds of legitimate interest, article 21.1 provides that data processing may continue if the data controller can demonstrate a "compelling legitimate grounds for the processing which override the interests." Note that for the balancing test pursuant article 21, the data controller must show a *compelling legitimate* interest, instead of a *legitimate* interest which is required for the balancing test pursuant article 6.1.(f), implying a higher threshold for the former than the latter. Whether a data controller has a compelling legitimate interest will depend on the facts of each case. However, as suggested above, a single data point usually has negligible impact on the functioning of an AI system. This may make it difficult to argue that there is a compelling legitimate interest for the processing of particular individual's personal data to continue.[131] On the other hand, removing a data point, or even multiple data points, will usually not have a significant impact on the training of an AI system. Thus, there may not be a particularly significant impact from the right to object on the development of AI systems, provided it is technically feasible and not too onerous for AI developers to stop the processing of a particular data points, if requested to do so.

Article 17 of the GDPR (right to be forgotten) gives the data subject the right to demand that the data controller erases her personal data if one of the grounds listed in article 17.1 applies, e.g. the data is no longer necessary for the purposes for which it was collected.[132] In relation to using personal data for training AI systems, an important question is whether the obligation to delete personal data extends to data derived from such personal data.[133] If the answer to this question is in the affirmative, this could create significant difficulties for the development of AI because it may entail that a data subject not only has the right to demand the deletion of her personal data on which the AI system was trained but also the AI system itself or those parts which are derived from her personal data. However, it is submitted that the relevant test is whether an individual can be identified from the data.[134] If from the trained model an individual can be identified then it is personal data, if no individual can be

[131] EPRS Study *supra* note 16, at 59 (makes a similar point in relation to right to be forgotten in article 17 of the GDPR).
[132] GDPR *supra* note 11, art. 17.1(a).
[133] EPRS Study *supra* note 16, at 57.
[134] *E.g.* C-434/16, *Peter Nowak* v. *Data Protection Commissioner*, 34. (where it was held that comments made by an examiner on an exam script are personal data). *Contrast*, joint cases c-141 and 372/12. See Joined Cases C–141 & 372/12, *YS, M and S* v. *Minister voor Immigratie, Integratie en Asiel,* 2014 E.C.R. I- 2081 (where it was held that the legal analysis of the application for a resident's permit was not personal data).

identified then it is not personal data. This means that pursuant article 17 an AI developer has an obligation to delete raw data but there is only an obligation to delete the trained model if the model itself qualifies as personal data.

IV. HOW DOES THE GDPR APPLY TO FEDERATED LEARNING?

The analysis above suggests that the GDPR can be interpreted in a way that is compatible with the development of AI systems. Furthermore, it was suggested that the distinction between the training phase and application phase of AI systems is being taken into account by the relevant regulatory bodies, like the EDPB or WP29. In this part, we will utilize the previous analysis to examine how the GDPR applies to FL. We will first analyze how the GDPR applies to the training phase of AI systems, and then how it applies to the application phase of AI systems trained through FL. Note that the analysis in this part is based on the assumption that that all the relevant parties are subject to the GDPR (unless the contrary is explicitly stated) and that the data in the data silos is personal data for the purpose of the GDPR.

A. *GDPR and the training phase of FL*

The question of how the GDPR applies to FL has been considered by other scholars.[135] For instance, Rossello et al. concluded that "[a]s is usually the case with privacy preserving technologies, when considered in isolation, federated learning is no silver bullet."[136] Similarly, Truong et al. posited that "FL-based systems are not naturally compliant with the GDPR",[137] both set of commentators seemingly suggesting that FL represents an evolutionary, rather than revolutionary, development. The conclusion that FL is "no silver bullet" or "naturally compliant with the GDPR" is correct. Yet, we will argue that FL has a more significant impact on the way the GDPR applies than these commentators seem to suggest. The argument will be made that, in case of FL, the question whether the training of an AI system on personal data is GDPR compliant is largely based on whether the personal data is kept secure.

Upon the assumption that the data in the data silos is personal data, the first question to be answered is whether FL is a form of processing data. Processing of personal data is defined in article 4(2) GDPR as "any operation . . . such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction". If personal

[135] *E.g.* S. Rossello, et. al *supra* note 9; Nguyen Truong et al. *supra* note 9.
[136] S. Rossello et. al *supra* note 9, at 12.
[137] Nguyen Truong et al., *supra* note 9, at 1.

data is collected for the purpose of using it for FL, then this will constitute processing of personal data, because the collection of data is a form of processing data. If the data was collected for a purpose other than using it for FL, the question to address is whether using the data for FL constitutes processing of data. Rossello et al. argue that "the processing operations performed on . . . data in the context of the federated learning process (e.g. data normalization, data alignment etc.) are likely to qualify as a "processing" under article 4(2) GDPR."[138] Considering the wide definition of the term processing in the GDPR, this conclusion seems correct.

If FL is a form of processing personal data, then the data protection principles in article 5 of the GDPR must be complied with, and as FL is form of training AI systems the data minimization principle (article 5.1(c)) is particularly relevant. Further, if the personal data was collected for a purpose other than for the training of an AI system, the purpose limitation principle (article 5.1(b)) will also be relevant. As was argued in part III above, to determine compliance with the data minimization and the purpose limitation principle, one needs to determine what impact the processing has on the data subject. In the training phase of an AI system the impact on the data subject is negligible, expect for the risk that the personal data may be abused. Therefore, in the training phase of an AI to determine whether the data minimization and purpose limitation principle has been complied with depends on whether the raw data is kept secure i.e. whether it is possible to use the data for a purpose other than FL. Whether a particular FL project will meet the required standards for keeping data secure is a question of fact, which needs to be determined in each case. However, to meet the required standard it is not necessary to show that there is no risk of abuse of the data; but rather, a balancing of risk and benefit is required.[139] In FL, personal data is usually more secure than in centrally trained AI systems; therefore, it will be easier for FL to be compliant with the data minimization and purpose limitation principles than centrally trained AI systems.

A similar analysis also applies to the question of whether using personal data for FL meets the criteria for the processing data based on the legitimate interest ground, pursuant article 6.1.(f) of the GDPR. As argued in part III above, the key question for determining whether the training of an AI system on personal data meets the legitimate interest requirements is how the usage of the personal data impacts the data subject. As in the training phase of AI Systems the impact on the individual is negligible, provided the data is secure, the relevant question is to what extent the personal data is kept secure. For the legitimate interest ground to apply it is not necessary to show that there is no impact on the data subject, which means that it is not necessary to establish that it is impossible for a malicious actor to access the data subject's personal data. Instead, all that needs to be established is that the impact on the data subject is outweighed by the legitimate interest of the data controller or third party. Therefore, similar to the data minimization and purpose limitation principles, the legitimate interest ground for lawful processing establishes a balancing test, between the legitimate

[138] S. Rossello et al., *supra* note 9, at 9 and 12.
[139] *See supra* discussion in section on *Purpose limitation* and *Data minimization*.

interests of the data controller or third party on the one hand, and the interest of the data subject on the other. Therefore, the conclusion is the same as for the data minimization and purpose limitation principles: Training AI systems using FL, generally, means that the data is more secure, therefore, it is easier for FL to satisfy the requirements for processing on the basis of the legitimate interest ground, than centralized training of AI systems.

A further aspect that requires analysis is how the GDPR applies to the model updates. Recall that in FL algorithms are trained *in situ* on the individual datasets, and the trained algorithms or the relevant parameter (i.e. the model updates) are transmitted to a central repository. The first question to answer is whether the model updates are personal data. Personal data is defined in article 4.1 of the GDPR as "any information relating to an identified or identifiable natural person". Recital 26 states that "[t]o determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly." Whether it is reasonably likely that a natural person can be identified from the model updates is a question of fact, which needs to be determined in each case. If a natural person can be identified from the model updates, then the model updates will be personal data.[140] Transmitting the model updates will also likely qualify as processing personal data pursuant to article 4(2) of the GDPR. Nevertheless, as with the processing of the raw data, the relevant issue for the processing of the model updates is whether they are kept secure; and the risk of the model updates revealing information about an identifiable individual needs to be balanced against the interest of the model updates being proceed.

Article 15.1 of the GDPR grants data subjects "the right to obtain from the controller confirmation as to whether or not personal data concerning him or her are being processed, and, where that is the case, access to the personal data." This could be a problem for FL because neither the raw data nor the model updates are easily accessible. However, article 15.1 does not pose an insurmountable obstacle to FL. First, it is important to note that article 15.1 does not create a right for the data subject to directly access his personal data but rather a "right to obtain from the controller . . . access to the personal data". Therefore, to analyze how article 15.1 applies it first needs to be determined who the data controller is.

According to article 4(7), the data controller is the entity, "which, alone or jointly with others, determines the purposes and means of the processing of personal data". In the case of FL, because of the multiplicity of parties, it can be a complex question to determine who the data controller is,[141] and the answer to the question who the data controller is will depend on the facts of each case. However, the most likely candidates to be the data controller in an FL project are the manager and the

[140] *See* EPRS Study *supra* note 16, at 38 (discussing the status of 'inferred personal data' under the GDPR). But *see supra* note 133 (In *YS, M and S* it was held that legal analysis of the application for a resident's permit was not personal data. But contrast this with the later case of *Nowak* where it was held that comments made by an examiner on an exam script are personal data. *Nowak* did not explicitly overrule *YS, M and S*, thus there is some uncertainty on the status of inferred personal data.

[141] S. Rossello et. al, *supra* note 9, at 10-11.

participant, who controls a particular data silo, as these two entities are likely to qualify as an entity who "determines the purposes and means of the processing of personal data". If the manager and/or the relevant participant are indeed the data controllers, then with regards to the raw data it should not be too difficult to give the data subject access to his personal data because the participant, who controls a data silo, has access to that data silo. For instance, assume that an FL project consists of several hospitals, which use MRI scans to train an AI system to detect cancer. The MRI scans controlled by each hospital will form a data silo and each hospital controlling this data silo will be the data controller for the relevant MRI scans. Thus, if a data subject requests to access his personal data, the hospital should be able to follow up on this request.

The issue is more complicated in relation to the model updates because the participant and the manager may find it difficult to access the model updates. However, in principle, there is no reason why it should not be possible to develop an FL protocol that allows the manager (or another entity) access to the model updates. The problem is that if the manager has access to the model updates, this will decrease the privacy of the raw data because it increases the risk of a privacy attack.
It has long been noted that there is a trade-off between privacy and transparency,[142] and making the model updates accessible is an example of this trade-off. However, thus far, there has been a lack of regulatory guidance on how to balance this trade-off. However, recital 62 provides that the right to access "should not adversely affect the rights or freedoms of others." Thus, an argument could be made that the right to access should not extend to the model updates because granting such a right may reduce the security of other individuals' personal data and thus, may "adversely affect the rights or freedoms of others." Another argument is to claim that the data subject's "right to obtain from the controller . . . access to the personal data" only applies if the controller has the ability to give such access. In FL, the manager or the participant may not be able to gives access the model updates, and the law cannot require a person to do something, that the person cannot do.[143] These arguments have not been tested in court and neither has regulatory guidance on these points been issued. Therefore, how the right to access applies to the model updates is currently a moot point and regulatory guidance on this point is needed.

Article 15.3 gives the data subject the right to obtain a copy of her personal data. Yet, this right is limited by article 15.4, which states that the right granted in article 15.3 "shall not adversely affect the rights and freedoms of others." Therefore, if a data subject requests a copy of the model updates, and the model updates allow for the identification of individuals, other than the data subject, who is making the request, then it should be possible to deny the request on the basis that it would "adversely affect the rights and freedoms of others."

---

[142] *E.g.* Lars Vilhuber, *Reproducibility and transparency versus privacy and confidentiality: Reflections from a data* editor, 235, J. Econom, 2285, (2023), (provides an overview of the trade-off between transparency and confidentiality but also suggests a more nuanced approach to the trade-off).
[143] This is sometimes expressed as the legal maxim '*lex non cogit ad impossibilia*' (the law does not compel the impossible).

An aspect where FL may be thought to make a difference but does not, is in relation to transfer provision set out in chapter V of the GDPR. Inter alia, chapter V states that personal data can only be transferred to a third country (i.e. non-EU country) if the EU Commission has designated such country as providing an "adequate level of protection" (article 45.1). The term transfer is not defined in the GDPR, but the EDPB states that transfer includes "making available". [144] Therefore, in an FL project, where a data silo is located in the EU, and a raw model is transmitted to this data silo from outside the EU, it is likely that this will be regarded as making the data available, and therefore constitute a data transfer. Therefore, in relation to the transfer provisions, whether FL or centralized training of AI systems is used, will not make a substantial difference.

The conclusion is that if an AI system is trained on personal data using FL, the GDPR will apply. Nevertheless, FL significantly changes the way that the GPDR applies to the training of AI systems because the main factor that determines compliance with the GDPR is whether the raw data (and, if applicable, the model updates) are secure. The significant difference between FL and centralized training of AI systems is that FL constitutes a form of what the WP29 calls "functional separation".[145] Functional separation is the separation of the function of using personal data to learn something about individuals, to whom the personal data relates, from the function of using data to learn correlations between data points, for the purpose of applying these correlations to cases outside the training set. In centrally trained AI systems, there is no effective functional separation because AI developers can access the personal data, and therefore, can use the data to gain information about individuals or use the data to learn about the correlations between data points. Therefore, it is important to have regulation, like the GDPR, in place to prevent developers from using the data to gain information about individuals. By contrast, FL separates the function of learning correlations, from the function of using personal data to gain information about individuals. In FL, the AI developers do not have access to the raw data, and the trained model only has information about correlations between data points. Information about individuals can only be gained if the model is applied to a particular case. The separation of these two functions in FL is not absolute because privacy attacks remain a possibility. Therefore, the security of data remains an issue relevant for the GDPR. However, to the extent that FL achieves functional separation, the training of AI system does not raise data protection issues. This means that in the training phase, it will be easier for AI developers to ensure compliance with the GDPR if they use FL rather than centralized training of AI systems, because, if FL is used, ensuring adequate data security will largely ensure compliance with the relevant GDPR provisions for the training of AI systems.[146]

[144] European Data Protection Board Guidelines, *Guidelines 05/2021 on the Interplay between the application of Article 3 and the provisions on international transfers as per Chapter V of the GDPR*, (Feb. 14, 2023, version 2.0), at 8.
[145] WP29 Opinion on Purpose Limitation *supra* note 66, at 46.
[146] However, note that data security will not ensure compliance with all GDPR provisions. For instance, FL does not make a difference to how the transfer provisions apply.

B. *GDPR and application phase of FL*

The section above analyses how the GDPR applies the training phase of AI systems, however, FL also has also implications for how the GDPR applies to the application phase of AI systems. As explained previously, in FL the raw data and model updates are not accessible, which makes AI system trained through FL less transparent than centrally trained AI systems, and this has implication for the application of the GDPR.

Article 5.1(a) of the GDPR requires personal data to be processed "fairly and in a transparent manner".[147] Fairness in this context can be analyzed as either formal or substantive fairness.[148] As a formal requirement, article 5.1(a) requires that data subjects should be informed about the "existence of the processing operation and its purposes."[149] This requirement does not pose a special challenge for FL as compared to centrally trained AI systems because data controllers using FL can fulfil this requirement the same way as data controllers using standard training of AI systems.

With regards to substantive fairness recital 71 is relevant, which states that:

> [I]naccuracies in personal data are corrected and the risk of errors is minimised, secure personal data in a manner that takes account of the potential risks involved for the interests and rights of the data subject and that prevents, inter alia, discriminatory effects on natural persons on the basis of racial or ethnic origin, political opinion, religion or beliefs, trade union membership, genetic or health status or sexual orientation, or that result in measures having such an effect.

It is more challenging for FL to comply with the substantive fairness requirements than for centrally trained AI systems because the raw data and model updates are not accessible. This makes it more difficult to detect potential biases and errors in the data, or whether a decision was based on a person's racial or the ethnic origin, political opinions, which might therefore be discriminatory. Furthermore, the WP29 states, that, within the context of automated decision making, "[c]ontrollers should carry out frequent assessments on the datasets they process to check for any bias, and develop ways to address any prejudicial elements, including any over-reliance on correlations."[150] Within the context of FL it is challenging to comply with this requirement because the raw data cannot be inspected.

---

[147] GDPR *supra* note 11, art. 5.1(a).
[148] EPRS Study *supra* note 16, at 44 (EPRS Study makes similar distinction).
[149] GDPR *supra* note 11, recital 50.
[150] ARTICLE 29 DATA PROTECTION WORKING PARTY Guidelines, *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679*, 17/EN 7/EN WP 251, (Oct. 13, 2017), at 16, [hereinafter WP29 Guidelines on automated decision-making].

Another issue is that the GDPR contains provision about automatic decision making. Articles 13.2.(f) and 14.2(g) of the GDPR give the data subject the right to be informed about "the existence of automated decision-making, including profiling" and to give "meaningful information about the logic involved". There has been a lot of debate about what exactly "meaningful information about the logic involved" entails.[151] However, Chesterman observes, and the authors agree, that "[t]he argument degenerated into a semantic dispute over the difference between a 'right to explanation' and a 'right to … meaningful information', but it appears to have been settled by the European Data Protection Board",[152] as the WP29 states that "[t]he controller should find simple ways to tell the data subject about the rationale behind, or the criteria relied on in reaching the decision without necessarily always attempting a complex explanation."[153] For the purpose of this article the significant point is that compliance with articles 13.2.(f) and 14.2(g) is more difficult for FL than for centrally trained AI systems. As AI systems, trained using FL, are usually less transparent, it is more challenging to understand the logic involved in the decision-making process, and hence more difficult to give meaningful information about the logic involved.

Transparency of AI systems is also an important issue beyond compliance with the GDPR. For instance, the issues of transparency and fairness are also important for the EU's Artificial Intelligence Act ('AI Act').[154] Article 13(1) of the AI Act imposes a requirement on AI systems, which are classified as "high-risk AI systems"[155], that such systems "shall be designed and developed in such a way as to ensure that their operation is sufficiently transparent to enable deployers to interpret a system's output and use it appropriately". In addition, the AI Act requires that high risk AI system need to have an appropriate risk management system,[156] data governance[157] and appropriate management practices.[158]

The fairness and transparency provisions of the GDPR and the AI Act undoubtedly pose a challenge for FL, but it is not impossible for AI systems trained through FL to comply with the relevant provisions. Neither the GDPR nor the AI Act require that the data on which an AI system is trained must be accessible. Rather the GDPR requires that in the case of automated decision-making,

---

[151] *E.g.* EPRS Study *supra* note 16, at 57 (makes the point that article 15 is very unclear). *See also*, L. Edwards, L. and M. Veale, *Slave to the algorithm? Why a 'right to an explanation' is probably not the remedy you are looking for*, DLTR, 16, (2019), (arguing that there is a right to an explanation in the GDPR, although the right is vague and potential counterproductive); *Contrast*, S. Wachter, B. Mittelstadt, and L. Floridi, *Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation*, 7, Int. Data Priv. Law, 76, (2016) (arguing the that there is no right to explanation in the GDPR). *See also*, WP29 which states in the WP29 Guidelines on automated decision-making *supra* note 49, at 26: "The controller should find simple ways to tell the data subject about the rationale behind, or the criteria relied on in reaching the decision without necessarily always attempting a complex explanation."

[152] Simon Chesterman, *We, the Robots?: Regulating Artificial Intelligence and the Limits of the Law*, Cambridge University Press, 2021, at 159 (internal footnote omitted).

[153] WP29 which states in the WP29 Guidelines on automated decision-making *supra* note 149, at 26

[154] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), 2024 O.J (L 1689), [hereinafter the AI Act].

[155] *Id*. art. 6(1).

[156] *Id.* art. 9.

[157] *Id.* art 10(2).

[158] *Id.*

meaningful information can be given about the logic involved,[159] and the AI Act requires that AI systems are sufficiently transparent such that an AI system can be interpreted appropriately.[160] It is potentially possible to meet all these requirements by means other than accessing the raw data,[161] and research is currently ongoing on how to achieve AI fairness even if the AI system was trained using FL.[162]

V. IS THE IMPACT ON THE DATA SUBJECT ALL THAT MATTERS?

The analysis above is based on the argument that there is an important difference between data protection issues in the training phase and application phase of AI systems because using personal data in the training phase has only a negligible impact on the data subject. A possible counterargument is that using personal data to training AI systems should be considered an infringement of a person's data protection rights, regardless of whether such training has an impact on this person or not.

Thus, the relevant question is whether using personal data *per se* should be considered an infringement of data protection, or whether using personal data only infringes a person's data protection rights if the usage of the data has an impact on this person. This is an important question, which does not only require legal analysis but also involves issue of public policy, and each jurisdiction needs to come to its own conclusion how to answer this question. Nevertheless, as will be set out below, for reasons of principle and policy, we caution against holding that the use of personal data for the training of AI systems is itself an infringement of data protection.

Regarding reasons of principle, holding that training AI systems on personal data in itself infringes data protection, seems contrary to basic principles of the GDPR. Recital 1 of the GDPR states that:

> The protection of natural persons in relation to the processing of personal data is a fundamental right. Article 8(1) of the Charter of Fundamental Rights of the European Union (the 'Charter') and Article 16(1) of the Treaty on the Functioning of the European Union (TFEU) provide that everyone has the right to the protection of personal data concerning him or her.

[159] GDPR *supra* note 11, arts. 13.2.(f), art. 14.2(g).
[160] AI Act *supra* note 153, art. 13(1).
[161] *E.g.* Nathalie Baracaldo *et. al.*, *Towards an Accountable and Reproducible Federated Learning: A FactSheets Approach*, IBM Research, (9 Feb. 25, 2022), (discusses some method how transparency and trustworthiness can be incorporated into FL).
[162] *E.g.* Yahya H. Ezzeldin, *FairFed: Enabling Group Fairness in Federated Learning*, The Thirty-Seventh AAAI Conference on Artificial Intelligence, (2023), AAAI-2, (provides an example of how fairness in FL might be achieved).

Article 8(1) of the Charter[163] and article 16(1) of the TFEU[164] create a right to data *protection*, not a right for the individual to control her data. Control of data may be an important tool to achieve data protection, but data protection and data control are not the same thing. Further, if using personal data *per se* constitutes a violation of the right to data protection, even if there is no impact on the individual, then this would turn the right to data protection into a quasi-property right.[165] A property right, inter alia, entails that a violation of such right is actionable in court, even if the violation of the property right caused no harm to the right holder.[166] In contrast, if a person brings a claim in tort, the plaintiff usually needs to show, not only that the defendant violated some right, but also that the violation of the right caused the plaintiff harm.[167] Yet, there is no indication that the framers of the Charter or the TFEU intended to give individuals property rights in data, and neither are such rights explicitly granted in the GDPR.[168]

Moreover, even if using personal data for training AI systems *per se* is considered as a violation of data protection rights, article 8.1 of the Charter must be read in conjunction with article 8.2, which provides that "data must be processed fairly for specified purposes and on the basis of the consent of the person concerned or some other legitimate basis laid down by law." Thus, the right to data protection does not give individuals absolute control over their data because article 8.2 provides that data may be processed if there is a legitimate basis to do so. Further, article 52 of the Charter provides that any limitation of a right laid down in the Charter is subject to a proportionality test.[169] Therefore, even if using personal data for training AI systems *per se* infringes the right to data protection; this right is, nevertheless, limited by article 8.2, and the proportionality test established by article 52. This means that if personal data is used to train AI systems, one needs to balance the right to data protection with other considerations, such as the benefits that can be derived from using such data to train AI systems. If the training of AI systems has only a negligible effect on the individual,

---

[163] Art. 8(1), Charter of Fundamental Rights of the European Union, Dec. 18, 2000, 2000 O.J. (C 364) 1. [hereinafter the Charter] states: "Everyone has the right to the protection of personal data concerning him or her."

[164] Art. 16(1), Consolidated version of the Treaty on the Functioning of the European Union, 2012 O.J. (C 326) [hereinafter the TFEU] states: "Everyone has the right to the protection of personal data concerning them."

[165] *See* Guido Calabresi & Douglas Melamet, *Property Rules, Liability Rules and Inalienability: One View of the Cathedral*. 85, Harv. L. Rev. 1089, (1972) (provides an overview of the difference between property rules and liability rules, which roughly corresponds to property rights and right to protected by tort law.

[166] Guido Calabresi & Douglas Melamet state that: "An entitlement is protected by a property rule to the extent that someone who wishes to remove the entitlement from its holder must buy it from him in a voluntary transaction in which the value of the entitlement is agreed upon by the seller." *Id*. at 1092. E.g. if person A steals property x from person B, B has a restitutionary claim against A for the return of x.

[167] Guido Calabresi & Douglas Melamet state that: "Whenever someone may destroy the initial entitlement if he is willing to pay an objectively determined value for it, an entitlement is protected by a liability rule." E.g. If A negligently damages property x which belongs to person B, B has a claim that A pays for the damaged caused. *Id.* 1092.

[168] *But see* Andreas Boerding et. al., *Data Ownership—A Property Rights Approach from a European Data Ownership—A Property Rights Approach from a European Perspective*, 11, 323 J. Civ.L. Stud. (2018). (Andreas Boerding et. al. make the argument "that European property law provides a sufficient framework for establishing a theoretical concept of data ownership." [at 324]. Nevertheless, "a sufficient framework for establishing a theoretical concept of data ownership" does not mean that there are property rights. Furthermore, in relation to the GDPR, Andreas Boerding et. al. argue that "the GDPR does not provide specific regulations for any form of data ownership" [at 331].

[169] Charter *supra* note 162, art.52:

> Any limitation on the exercise of the rights and freedoms recognised by this Charter must be provided for by law and respect the essence of those rights and freedoms. Subject to the principle of proportionality, limitations may be made only if they are necessary and genuinely meet objectives of general interest recognised by the Union or the need to protect the rights and freedoms of others.

the balancing test should favor allowing the usage of personal data for the training of AI systems. The conclusion is that the intention behind creating a right of data protection in the Charter and the TFEU is to protected data from abuse, rather than giving individuals control rights over their data. 'Protection' implies protection from something, which means that there must be some impact on the data subject. This does not mean that the impact must be tangible. Impact may also consist of something non-tangible, like inferences being drawn about an individual. But if there is no impact on the data subject, then there is nothing from which the data, or the data subject, needs to be protected from.[170] Therefore, use of personal data, which has no impact on the individual (e.g. using the data for statical purposes or to train AI systems), should be considered outside the ambit of data protection.

The policy reasons for rejecting the idea that using data to train AI systems *per se* infringes data protection is that there seems to be no practical benefit to such a move because most (possibly all) of the risk associated with AI would still exist even if AI systems were not trained on personal data. Risks associated with AI may be divided into existential risks and societal risks. Existential risk deals with the risk of AI eradicating humanity, or similar risks. Societal risks are risks such as discriminatory AI systems (amplifying/perpetuating existing prejudices), that AI may lead to widespread unemployment, or similar risks. It is quite clear that regardless of whether AI is trained on personal data or not, has very little impact, if any at all, on the existential risk of AI.

Regarding societal risk, according to the GDPR, the AI Act, as well as many standards for ethical AI around the world, transparency is an important aspect of ethical AI.[171] An obstacle to achieving transparency is that AI systems often operate as a 'black box',[172] i.e. it is difficult to understand why an AI system made a certain decision. Yet, the black box problem is not something that can be addressed through data protection because an AI system will operate as a black box, regardless of whether an AI system is trained on personal or non-personal data. There would be a black box problem even if AI system would not require any training on data.[173] The black box it is simply the problem that it is difficult to understand how AI systems operate. Therefore, it is not problem, which involve issues of data protection.

Another ethical standard is that AI systems should be fair and not biased. In the current iteration of AI, whether an AI system is biased or not often depends on the type of data the system is trained on.[174] However, this does not mean it is an issue of data protection. AI systems not trained on personal data, or not trained on data at all, may also be biased. Moreover, if an AI system exhibits

---

[170] The WP29 seems to be of a similar opinion, as it states that: "Data protection law … has after all been designed to protect individuals against the impact of improper or excessive use of their personal data." WP29 Opinion on Purpose Limitation *supra* note 66, at 25.
[171] *See* Simon Chesterman supra note 151, ch. 7 (provides an overview of different ethical codes for AI).
[172] *See* Nguyen Truong *supra* note 9, at 15 (discusses the black box problem).
[173] It is important to note the AI is not the same as machine learning. "Artificial Intelligence is the field of developing computers and robots that are capable of behaving in ways that both mimic and go beyond human capabilities." In contrast, "[m]achine learning is a pathway to artificial intelligence. This subcategory of AI uses algorithms to automatically learn insights and recognize patterns from data, applying that learning to make increasingly better decisions." *Artificial Intelligence (AI) vs. Machine Learning*, Columbia Engineering, 2024, https://ai.engineering.columbia.edu/ai-vs-machine-learning/ (last accessed July 29, 2024).
[174] *E.g.* Nguyen Truong *supra* note 9, at 15 (discusses the black box problem).

bias because it was trained only on certain type of data, making more data available may help to reduce the bias of such a system.[175] Therefore, to achieve AI systems, which are fairer and less biased, rather than restricting the use of personal data for the purpose of training AI systems, it should be encouraged that AI systems are trained on more data as well as ensuring that such data is representative.

Another issue may the concern that the widespread adoption of AI may lead to a highly uneven distribution of the profits generated from AI. One problem may that the control of large amounts of data, including personal data, by some organizations, gives such organizations substantial market power, which may be used to prevent competition, [176] and may lead to extensive profits. One may want to question such market power, but this is an issue for competition or antitrust law and not for data protection.[177] Fairness, transparency and an equitable distribution of profits are important issues, but they are not data protection issues, and, as stated above, during the training phase of AI systems no information about individuals is revealed. Therefore, holding that training AI systems on personal data *per se* constitutes an infringement of data protection does not seem to have any discernable practical benefit.

This does not mean that data protection issues play no part in relation to AI. On the contrary, data protection issues are very important, but they are important in the application phase of AI systems. In the application phase, an AI system may draw inferences about individuals, and this is the issue that data protection needs to address. Therefore, data protection regulators should focus their resources and energy on the application phase of AI systems, not the training phase. In the training phase of AI systems, the only relevant issue for data protection is whether personal data is secure. It is important to point out that there may be cases where there is no clear distinction between training phase and application phase of an AI system. For instance, a person may use an AI system by feeding prompts into the system, and based on these prompts the AI system delivers results to the user, but the AI system also trains on these prompts.[178] In this case, there may be no clear separation between training phase and application phase. Nevertheless, the principle that the data protection issues are different in the training phase and application phase remains. If the AI systems trains on prompts, provided the data is secure, this does not infringe data protection rights, even if the prompts are personal data, but if the AI system delivers results, which consist of personal data, then this relevant for data protection.

---

[175] A problem with the data on which AI systems are trained that they may over-or underrepresent certain people within society. *See*, O*pen source data science: How to reduce bias in AI*, World Economic Forum, (Oct. 14, 2022), (discuss making AI systems less biased by using open source, which means making more data available).

[176] *COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS* A European strategy for data, COM (2020) 66 final, (Feb. 19, 2020), at 3, [hereinafter EU strategy paper for data].

[177] In the alternative, the issues may be dealt with through legislation, specifically designed for this purpose. For, instance, the EU enacted specific legislative acts to deal with this issue. *See*, Digital Services Act. See, Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive, recital 4, 2022, 2000/31/EC, O.J. L (277) [hereinafter Digital Services Act].

[178] For instance, ChatGPT operates this way.

VI. FEDERATED LEARNING AND THE SHARING OF NON-PERSONAL DATA

Thus far, this article has analyzed FL in relation to the sharing of personal data. However, what is sometimes less appreciated, is that there are also important issues in relation to the sharing of non-personal data, in particular the question of why non-personal data is not shared more widely. According to one estimate data-sharing could increase GDP by 1% to 2.5%.[179] There is also increasing awareness of the importance of sharing non-personal data among governments and government agencies. For instance, the EU Commission formulated a strategy paper for the development of a data driven economy, [180] stating that "non-personal industrial data and public data in Europe . . . will constitute a potential source of growth and innovation that should be tapped."[181]

The EU has also taken legislative steps to facilitate data sharing. To mention just a few, the EU has enacted the 'Digital Market Act' (DMA),[182] which aims make the digital economy fairer and more competitive.[183] Inter alia, the DMA does not allow large technology to combine data, which they collected from different services.[184] 'The Data Governance Act',[185] includes provisions for the reuse of data held by the public sector[186] and requirements for data intermediaries.[187] The 'Free Flow of Non-Personal Data Regulation' aims to facilitate sharing of data between EU member states.[188] The 'Open Data Directive' aims to make the data held by public bodies more accessible.[189] There is also sector specific legislation like the revised 'Payment Services Directive' (PSD2) which applies to payment services and aims to foster competition by making it easier for consumers to switch between service providers.[190] The 'Data Act'[191], aims to "to remove barriers to a well-functioning internal market for data"[192] by granting users access to the data they generate and the right to share this data with third parties.[193]

---

[179] Enhancing Access to and Sharing of Data - Reconciling Risks and Benefits for Data Re-use across Societies, (2019), OECD, at 11, available at https://www.oecd-ilibrary.org/sites/90ebc73den/index.html?itemId=/content/component/90ebc73d-en (last accessed July 22,, 2024).
[180] EU strategy paper for data *supra* note 175.
[181] *Id*. at 1.
[182] Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act), 2022, O.J. L (265).
[183] *Id.* article 1.1.
[184] *Id.* articles 5.1(b) and 5.1(c) .
[185] Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act), 2022, O.J. L (152).
[186] *Id.* Ch. II.
[187] *Id.* Ch. III.
[188] Regulation (EU) 2018/1807 of the European Parliament and of the Council of 14 November 2018 on a framework for the free flow of non-personal data in the European Union, 2018, O.J. (L 303).
[189] Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information (recast), 2019, O.J. (L 172).
[190] Directive (EU) 2015/2366 of the European Parliament and of the Council of 25 November 2015 on payment services in the internal market, amending Directives 2002/65/EC, 2009/110/EC and 2013/36/EU and Regulation (EU) No 1093/2010, and repealing Directive 2007/64/EC, 2015, O.J. (L 337).
[191] Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act), O.J. (L 2023/2854).
[192] *Id.* recital 4.
[193] *Id.* Ch. II.

The relevant question for the purpose of this paper is to what extent FL can facilitate the sharing of non-personal data. Before answering this question, it is important to understand some of the reasons for why non-personal data is not shared more widely. IP law is certainly one reason preventing data sharing. However, the IP law issues in relation to FL are substantially similar to the IP issues in relation to centralized training of AI systems, and thus, IP law is outside the scope of this article. With regards to other reasons why non-personal data is not shared more widely, the EU Commission has identified several obstacles to data sharing.[194]

An important obstacle, according to the EU Commission, is the market power of companies, which control large amounts of data. The EU commission argues that "[c]urrently, a small number of Big Tech firms hold a large part of the world's data. This could reduce the incentives for data-driven businesses to emerge, grow and innovate".[195] Another obstacle, identified by the EU Commission is "lack of trust between economic operators that the data will be used in line with contractual agreements."[196] Sharing of data entails risk for the organization making the data available, because the organization receiving the data may use the data in a way which harms the organization that shared the data. For instance, a group of banks may find it beneficial to share customer data to train an AI system to detect fraud. Yet, if a bank shares customer data with a competitor bank, there is a risk that the competitor uses the data to 'poach' clients. Contractual provisions may be able to ameliorate this problem to some extent, but it is unlikely to solve the problem. As it is very difficult to detect for which purpose data is used, once data is handed over to a different organization. Other obstacles to effective data sharing, identified by the EU Commission, is legal uncertainty about what data can be shared legally,[197] lack of data sharing infrastructure[198] as well as the lack of interoperability and common standards for sharing data. [199]

FL can be a useful tool to overcome some of these obstacles but not others. If market power by data controllers is a significant reason for why data is not shared more widely, then it is unlikely that FL will be able to promote more sharing of data. FL does not impact the market power of on organizations and neither does FL affect whether it is more profitable for a data controller to restrict the supply of data, than to share the data more widely. FL also does not directly impact the problem of lack of data sharing infrastructure, interoperability or common standards for sharing data.

However, FL has the potential to deal with the risks associated with data sharing. As FL does not require the transfer of data itself, there is less risk that the dataset will be used for nefarious purposes. As stated above,[200] FL cannot guarantee absolute confidentiality of the raw data, as privacy attacks are a possibility, but FL can lower the risk that data is used in a way that harms the organization that shared the data. Furthermore, FL may lower legal uncertainty about what data can

[194] EU strategy paper for data *supra* note 175.
[195] *Id*., at 3.
[196] *Id*., at 7.
[197] *Id.*
[198] *Id.* 9.
[199] *Id.* 8.
[200] *See supra* part II, *What is Federated Learning?*

be shared legally. As argued above,[201] if personal data is made available through FL in relation to the GDPR the main issue is whether the personal data is secure. Therefore, if an organization uses FL to make data available, the organization needs ensure that the data is kept secure and does not need to deal with the question of whether data is personal or non-personal data.

Nevertheless, it is unlikely that FL will make substantially more data available for the training of AI systems. As noted above, FL was first introduced in 2016 and since then FL has been applied to different uses cases. [202] However, it is notable that FL is mainly used within organizations and among researchers, rather than being used across organizations, let alone between competitor organizations.[203] A detailed analysis of the reasons for the lack of usage of FL across organizations is outside the scope of this article. But possible reasons may be a lack of awareness of FL among industry players, or that it is more challenging and costly to use FL than centralized training of AI systems. Another possible reason, is that FL requires an element of trust.[204] The manager needs to trust the participants will not engage in poisoning attacks and the participants need to trust the manager will not to engage in privacy attacks.[205] Models in computer science often assume the existence of trusted third party,[206] but in the real world a trusted third party may not exist, and this is a potential problem for FL.[207] For instance, competitor companies may find it beneficial to train an AI system collectively using FL but they may be reluctant to appoint a competitor company as manager due to concern that the competitor company may engage in privacy attacks. A potential solution to this problem is to appoint an independent entity as manager. For instance, a company that provides FL services on a commercial bases, and a number of such companies exist.[208] Nevertheless, besides a number of commercial providers offering FL services, it remains the case that FL is not widely used across different organizations, and if FL is not widely used across organizations, it is unlikely that a significant amount of additional data will be made available through FL.

VII. RECOMMENDATIONS

Before providing recommendations, it is useful to give some historical perspective. The predecessor of the GDPR, the DPD, came into effect in 1995.[209] The DPD did not contain any

[201] *See supra* part IV, *How does the GDPR apply to FL?*
[202] *See supra*, p.4.
[203] Saikishore Kalloori & Abhishek Srivastava, *supra* note 17, at 1.
[204] Priyanka Mary Mammen, *supra* note 18, at 1.
[205] *Id.*
[206] Alexander Iliev & Sean Smith, *Towards Tiny Trusted Third Parties*, Dartmouth Computer Science Technical Report TR2005-547, (2005).
[207] *E.g.* Shuaicheng Ma, Yang Cao, Li Xiong, Conference Proceedings, *Transparent Contribution Evaluation for Secure Federated Learning on Blockchain*, 2021 IEEE 37th International Conference on Data Engineering Workshops (ICDEW), 88 (Discussed issue of trust in FL, and proposes a solution based on blockchain to overcome trust issue).
[208] *E.g.*: Apheris (https://www.apheris.com/platform); Owkin (https://owkin.com/collaborate); Intellegens (https://intellegens.com/products-services/ichnite/) (last accessed July 22, 2024).
[209] DPD *supra* note 40.

references to the Internet or related concepts e.g. social networks. This is to be expected because in the mid-1990s the Internet was not as important as it came to be.[210] By contrast, the GDPR contains references to the Internet,[211] reflecting the awareness that data collection in the digital age raises different issues than in the pre-digital age.[212] Yet, the GDPR does not contain references to AI. This too is to be expected because at the time the GDPR was drafted, AI and big data were not as important as they are now. At the time of the drafting of the GDPR, the main usage of personal data was to draw inferences about individuals, to whom the data pertains.[213] To find correlations between data points was less commercially important. Thus, it is to be expected that the GDPR does not contain specific references to the concept of functional separation, although the GDPR 'indirectly' recognizes the importance of functional separation because the GDPR prescribes a different treatment for anonymized data.[214]

Nevertheless, as was argued above, the GDPR is flexible enough to be interpreted in a way to accommodate the concept of functional separation and the difference between personal data used in the training phase and application phase of AI systems.[215] The problem is that the analysis of the GDPR provided in this article is a *possible* interpretation of the GDPR, and reasonable legal opinion may differ on how the GDPR applies to AI and FL. It was noted on several occasions in this article,[216] that the EDPB and the WP29 seem to be willing to embrace functional separation and to interpret the GDPR to accommodate the distinction between personal data used in the training phase and application phase of AI systems. Nevertheless, more clarity is needed.[217] Thus far, the EDPB has not provided a clear statement which explicitly sets out the importance of functional separation and the difference between personal data used in the training phase and application phase of AI systems.[218] More clarity is particularly needed in relation to how the GDPR applies to FL. As noted in this article, there is some uncertainty how the GDPR applies to FL, for instance, there is no clear guidance in relation to the right of access and the model updates, and how to balance the trade-off between privacy and transparency in FL systems.[219] Without clarity of how the law applies to FL, it is unlikely that FL will be used widely by the industry.

Therefore, we recommend that regulators and policy makers should evaluate how FL fits into their wider data strategy, and whether, FL can be a tool to achieve their objectives. AI faces a data

[210] EPRS Study *supra* note 16, at 35.
[211] *E.g.* GDPR *supra* note 11, recitals 24, 30, 32, and 65.
[212] EPRS Study *supra* note 16, at 35.
[213] For instance, companies were interested in the information that a particular person is looking for a Caribbean holiday, mainly to show that person advertisements about Caribbean holidays or otherwise sell something to that person in relation to holidays.
[214] GDPR *supra* note 11, recital 26.
[215] In FL the difference between using data in the training phase and application phase is significant because, unless there is a successful privacy attack, the trained AI system cannot be applied to the data on which the AI system was trained. Therefore, legal analysis of FL is important because it brings functional separation into sharp focus.
[216] *See supra* part III *How does the GDPR apply to AI?*
[217] EPRS Study *supra* note 16, at III-IV, (EPRS Study also argues that in some areas of how the GDPR applies to AI more clarity is desirable).
[218] The investigation of ChatGPT could be used as an opportunity to clarify the issues. See ChatGPT Report, *supra* note 68.
[219] *Supra*, section *GDPR and the training phase of FL*.

problem and, particularly in relation to non-personal data, policy makers are searching for ways to facilitate more data sharing, and FL could be one of these ways, and there is a lot that regulators and policy makers could do to facilitate the use of FL. FL has some significant shortcomings, not least, that, currently, FL is not widely used across organizations. As stated previously, a possible reason for this is that FL requires an element of trust. Models in computer science often assume the existence of a trusted third party, but such a party may not exist in the real world, but appropriate regulation may be able to create 'trusted third parties', for instance, by creating a regulatory regime for FL. A regulatory regime should set out minimum standards, which commercial providers of FL projects need to adhere to, and a licensing regime for such providers to ensure compliance with these standards. This should help to improve trust in such providers.

Another problem is that AI systems trained through FL are less transparent than centrally trained systems. An FL regulatory regime could help in this regard, for instance, by stating clearly how to balance the trade-off between privacy and transparency.[220] Regulation could also play a part in making the privacy/ transparency trade-off less pertinent. For example, the regulatory regime could provide that only certain regulatory agencies may inspect the raw data and the model updates, if there is reasonable suspicion that an AI system is biased, discriminatory or otherwise unfair. The regulatory regime should also establish a protocol setting out how regulatory agencies are to conduct inspections of the raw data, to keep the raw data and model updates secure. Furthermore, a regulatory regime may be able to set common standards for data sharing through FL. This is important because the lack of common standards is one of the factors identified by the EU commission as a reason that inhibits data sharing. Lastly, the creation FL regulatory regime, or even just a statement from regulators or policy makers how they view FL, would raise awareness of FL in the industry and give the industry more confidence when using FL.

A regulatory regime is, of course, no panacea. It may be the case that future research will show that the lack of transparency is too sever an issues and not easily overcome. In this case, regulators and policy makers may want to consider restricting the use of FL to the training of AI systems, where transparency is less important. The exact details, as well as the feasibility, of a FL regulatory regime would need to be analyzed further and creating such a regime is not cost-free. Thus, just to reiterate, the recommendation is *not* that regulators and policy makers should endorse FL and create an FL regulatory regime. The recommendation is that regulators and policy makers should evaluate how FL fits into their overall data strategy. If they decide that the use of FL is something they want to encourage, then creating an FL regulatory regime could be an effective tool to do so. However, the most significant tool available to regulators and policy makers to facilitate the use of FL is to create clarity about the legal effects of FL. FL has some clear benefits in relation to data privacy but without legal clarity, it is unlikely that FL will be used more widely.

[220] *Supra*, section *GDPR and the training phase of FL*.

VIII. CONCLUSION

This article has shown that using data to train AI systems does not infringe data protection, provided the data is kept safe. This article has further shown that FL is a way to create functional separation between using data to draw inferences about individuals and using data to train AI systems. In this regard, FL has some significant advantages over centralized training of AI systems. Yet, without legal clarity of how data protection regulation applies to FL, and ways to overcome the 'trust' problem, it is unlikely that FL will be used sufficiently widely to make a significant difference to availability of data to train AI. Regulators and policy makers may be able to facilitate the wider use of FL by providing legal clarity and, possibly, by creating an FL regulatory regime. Without such measures it is doubtful whether FL will be able to make a substantial contribution toward solving AI's data problem.

However, even if FL does not succeed in solving AI's data problem, there are important lessons that can be gleaned from FL. The most important being that FL highlights the importance of functional separation and bringing into sharp focus that the use of data to train AI systems by itself does not infringe data protection, provided that the data is kept secure. In the future, researchers may find a way overcome the problems that currently prevent FL from being used more widely, or researchers may find a different and more effective method of implementing functional separation, different from anonymizing data or FL. In this case, the lessons learned from analyzing FL in relation to functional separation and that the training AI systems on personal data *per se* does not infringe data protection, should prove useful for analyzing and evaluating future scenarios.

There is a further point that the analysis of FL highlights, namely the importance of interdisciplinary work in relation to AI. The conceptualization of a certain concepts like privacy may differ across disciplines and the same is true for other concepts like bias, transparency, explainability, etc. To properly assess AI and AI's impact these differences must be taken into account, and computer scientist, lawyers, policy makers and many more need to work together in a meaningful way. The analysis of FL shows that 'interdisciplinarity' is not just buzzword – in relation to AI it is a necessity.