Published in: *Computerrecht* 2021/116 (2021)

Data protection by design in AI? The case of federated learning

Stephanie Rossello,[1] Roberto Díaz Morales,[2] Luis Muñoz-González[3]

Summary: This article[4] investigates some of the data protection implications of an emerging privacy preserving machine learning technique, i.e. federated machine learning. First, it shortly describes how this technique works and focuses on some of the main security threats it faces. Second, it presents some of the ways in which this technique can facilitate compliance with certain principles of the General Data Protection Regulation as well as some of the challenges it may pose under the latter.

Keywords: artificial intelligence, federated learning, data protection by design, GDPR

1. Introduction

Similarly to what we have witnessed with the rise of other big data analytics tools,[5] the latest boom in Artificial Intelligence ("**AI**"), particularly machine learning, has been accompanied by warnings that the latter could imperil the core tenets of data protection. Machine learning creates predictive models by processing datasets in a process called "training". Since the development of accurate machine learning models is said to largely depend on the quantity of (qualitative) training data,[6] commentators have highlighted the tensions between machine learning and core data protection principles, such as data minimization and purpose limitation.[7] Recent advances in privacy preserving machine learning (i.e. a field of research intended to "allow data processing without revealing the data itself"[8]) indicate, however, that these tensions can, to a certain extent, be mitigated when designing the machine learning system. Below we focus on a nascent privacy preserving machine learning technique, federated learning. Although federated learning has not gone unnoticed by data protection authorities,[9] European legal scholars appear, so far, to have devoted little attention to it. This article, which is the result of the cooperation between machine learning experts and a legal researcher,

[1] Researcher at the Center for IT and IP Law at the KU Leuven.
[2] Head of Artificial Intelligence research at Tree Technology and Adjunct professor in the Department of Signal Theory and Communications at University Carlos III de Madrid.
[3] Research Associate in the Department of Computing at Imperial College London.
[4] The research leading to this publication has been funded by the European Union's Horizon 2020 research and innovation program under grant agreement No 824988 "Machine learning to augment shared knowledge in federated privacy preserving scenarios" (MUSKETEER).
[5] T. Z. Zarsky, 'Incompatible: The GDPR in the Age of Big Data', *Seton Hall Law Review* (47) 2017, pp. 995 - 1020.
[6] Y.A. de Montjoye and others. 'Solving Artificial Intelligence's Privacy Problem', *Field Actions Science Reports* (2017), p. 80.
[7] E.g.: CIPL, 'First Report: Artificial Intelligence and Data Protection in Tension', 2018, pp. 12 - 14 www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl_ai_first_report_-_artificial_intelligence_and_data_protection_in_te....pdf.
[8] G. A. Kaissis and others, 'Secure, Privacy-Preserving and Federated Machine learning in Medical Imaging', *Nature Machine Intelligence* (2) 2020, p. 307.
[9] E.g.: R. Binns and V. Gallo, 'Data Minimisation and Privacy-Preserving Techniques in AI Systems', 21 August 2019, www.ico.org.uk/about-the-ico/news-and-events/ai-blog-data-minimisation-and-privacy-preserving-techniques-in-ai-systems/.

contributes to filling this gap. It provides a non-exhaustive overview of the data protection implications of training machine learning models under the federated learning paradigm.[10] On the one hand, we evaluate how federated learning can help comply with certain data protection principles. On the other hand, we list some important but, so far, largely ignored questions that federated learning raises under the General Data Protection Regulation ("**GDPR**"). Preliminarily, we present the broader (technological) context[11] in which federated learning was developed, how it works and compare the security of federated learning models with that offered by models trained in a centralized manner.

2. Federated learning in a nutshell

2.1. How we got there?

As pointed out by de Montjoye and others,[12] the need to use personal data for analysis purposes on the one hand, and preserve privacy, on the other, has for a long time been met by means of data anonymization. Following what Ohm calls the "release and forget model",[13] before sending a dataset to the analyst, such dataset would first be modified by combining several (pseudonymization and de-identification) techniques to unlink "an individual's record from their identity in a particular dataset".[14] However, this approach failed to guarantee anonymity when it was applied to modern high-dimensional datasets (which, compared to older datasets, contain a significantly higher amount of information about a certain individual, collected from various sources such as mobile devices, IoT etc.).[15] This failure (together with an increase in the computational power of devices) contributed to the emergence of a novel approach, where the holders of the dataset grant the analyst (remote) access to the data under certain strict conditions.[16] Under this approach, the algorithm "visits"[17] the data and performs certain computations on that data at the 'edge' of the network.

2.2. How does federated learning work?

Federated learning is an instance of the aforementioned 'visiting algorithm' paradigm.[18] In machine learning literature, federated learning is defined as a "distributed machine learning approach which enables training on a large corpus of decentralized data residing on devices like mobile phones"[19] or other type of devices, such as, for example, a company's data centers. Unlike the traditional centralized machine learning paradigm, federated learning does not require transferring all the training data from

[10] For a similar endeavor by data science scholars: N. Truong and others, 'Privacy Preservation in Federated learning: Insights from the GDPR Perspective', 2020, pp.1-21, arxiv.org/abs/2011.05411.
[11] Note that this article provides the technical information necessary to support the legal analysis presented in sections 3 and 4 below. The language and concepts have, however, been simplified in order to reach a broad (legal) audience.
[12] de Montjoye and others (n 6) p. 81.
[13] P. Ohm, 'Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization', *UCLA Law Review* 2010, p. 1716.
[14] de Montjoye and others (n 6), p. 81.
[15] Idem, p. 82; Ohm (n 13), pp. 1716–1722.
[16] de Montjoye and others (n 6), p. 83.
[17] Idem.
[18] K. Bonawitz and others, 'Towards Federated learning at Scale: System Design', 2019, pt. 1, arXiv:1902.01046.
[19] Idem.

several devices to one single central repository (typically, the cloud), where the machine learning model is trained (see figure 1 below).[20] Rather, under the federated learning paradigm, the machine learning model is trained locally, where the training data originally reside, under the coordination of a central server (also called "central node" or "aggregator").

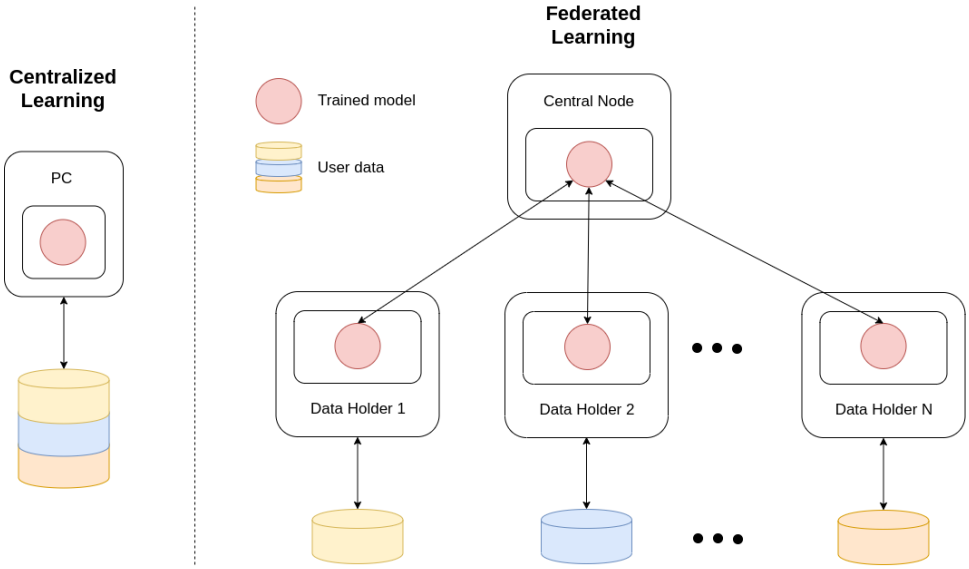

Figure 1: Centralized Learning vs Federated learning

As shown in figure 2 below, a basic federated learning process typically follows four steps. First, every data holder participating in the training receives a copy of the central model. Second, each data holder updates the model received from the central server with its data.[21] Third, the data holders send the updates to the model back to the central server. Fourth, the central server aggregates the different local updates and globally updates the model. Subsequently, the training process goes back to step one and repeats the process until the training is completed.

[20] In this figure the data holders and the central server are honest, therefore no leakage of information from the data to the central server is allowed.

[21] In machine learning, learning algorithms are designed to find the maximum or minimum of an objective mathematical function. This occurs through iteratively updating the model using several optimizing algorithms.

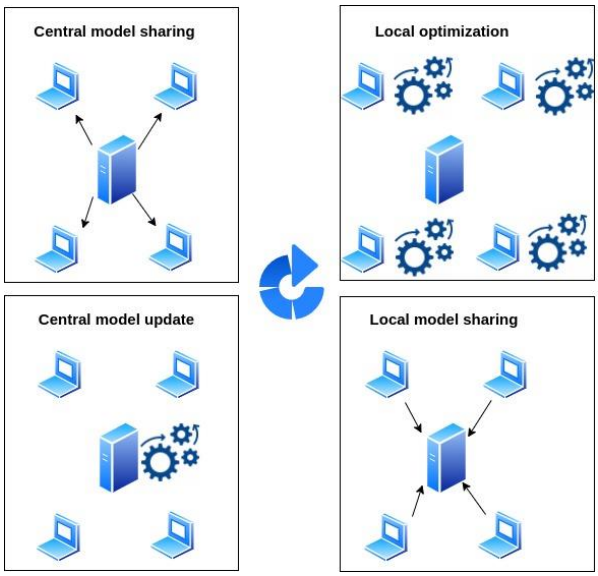

Figure 2: Federated learning process

Under certain circumstances, the updates that each data holder shares with the central server could leak information about the underlying (personal) training data to the central server or a third party.[22] Therefore, a major line of research in federated learning is devoted to combining the latter with other privacy preserving technologies based on data encryption.[23] This combination limits the capacity to extract the (personal) training data from the updates sent to the central server but is typically associated with a higher computational cost.[24]

The iconic problem that federated learning intends to solve is to obtain "a *single, global* statistical model from data stored on tens to potentially millions of remote devices".[25] Indeed, compared to other privacy preserving machine learning techniques such as those relying only on data encryption, federated learning offers the advantage of allowing to train a model with a larger number of training participants, at lower computational costs. Tellingly, federated learning is applied by Google to support next-word prediction in its mobile keyboard application, Gboard.[26] Moreover, the technique is gaining

[22] See research cited in L. Lyu, H. Yu and Q. Yang, 'Threats to Federated learning: A Survey', 2020, pt 1.2, arXiv:2003.02133.
[23] T. Li and others, 'Federated learning: Challenges, Methods, and Future Directions', 2019, p. 11, arXiv:1908.07873; Q. Yang and others, 'Federated Machine learning: Concept and Applications', *ACM Transactions on Intelligent Systems and Technology* (10) 2019, pp. 3–4, dl.acm.org/doi/10.1145/3298981.
[24] Y. Liu and others, 'Secure Federated Transfer Learning', 2020, pp. 1-9, arXiv:1812.03337.
[25] Li and others (n 23), p. 3.
[26] Bonawitz and others (n 18). Federated learning also seems to occupy a core role in Google's recently announced plan to replace third party cookies with more privacy preserving web advertising, see: C. Bindra, 'Building a privacy-first future for web advertising', 25 January 2021 ,www.blog.google/products/ads-commerce/2021-01-privacy-sandbox/.

traction in healthcare applications and medical research,[27] as it is believed it can help overcome legal and ethical barriers to the sharing of health data.[28]

2.3. Security implications of training models with federated learning

An important research line in federated learning is focused on improving the security of federated learning systems against certain type of attacks. For the purpose of this contribution we focus on two types of attacks, poisoning and certain privacy attacks.

2.3.1. Poisoning attacks

Research shows that poisoning attacks are one of the main threats when training federated learning algorithms. [29] Through data poisoning attacks, an attacker aims at "subverting the entire learning process in a nondiscriminatory or targeted way, i.e., aiming to decrease the overall performance of the system or to produce particular kinds of errors." [30] Depending on the attacker's capabilities and knowledge, the amount of training data available and the machine learning algorithm used, different poisoning attacks are possible. We distinguish data poisoning from model poisoning attacks.

Data poisoning attacks imply that attackers manipulate the training data provided by one or more data holders. For example, attackers could contaminate the datasets provided by some federated learning participants. These attacks are not unique to federated learning. They can also be performed in a centralized machine learning setting by an external actor, i.e. someone other than the data holders or the central server. However, in a federated learning scenario, these attacks could also be carried out by internal actors, i.e. one of the training participants. Indeed, when training with possibly "millions of participants, it is impossible to ensure that none of them are malicious".[31] Fortunately, assuming that there is a significant number of training participants, the effect of this type of attacks can be limited. Indeed, often the effect of the poisoned dataset provided by the malicious training participant is diluted by the (non-poisoned) datasets contributed by the honest training participants.

Model poisoning attacks imply that attackers directly manipulate the information sent to the central server. In other words, some participants send malicious model updates to the central server. Model poisoning attacks are unique to federated learning. Like data poisoning attacks, they can be performed by both external and internal actors. However, compared to data poisoning, the consequences of model poisoning attacks can be more dramatic. Indeed, these attacks can allow malicious participants to acquire full control of the training process and neutralize the contributions provided by the honest

[27] N. Rieke and others, 'The Future of Digital Health with Federated learning', *npj Digital Medicine* 2020, pp. 1-7, doi.org/10.1038/s41746-020-00323-1.
[28] Willem G van Panhuis and others, 'A Systematic Review of Barriers to Data Sharing in Public Health', *BMC Public Health* (14) 2014, doi.org/10.1186/1471-2458-14-1144, as cited in Rieke and others (n 27).
[29] E.g.: A.D. Joseph and others, 'Machine learning Methods for Computer Security', Dagstuhl Perspectives Workshop 12371, *Dagstuhl Manifestos* (3) 2013, pp. 1–30; L. Muñoz-González and others, 'Towards Poisoning of Deep Learning Algorithms with Back-gradient Optimization', *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security* 2017.
[30] L. Muñoz-González and E. Lupu, 'The Security of Machine learning Systems', in L. F. Sikos (ed.), *AI in Cybersecurity*, Springer International Publishing: 2019, p. 54.
[31] E. Bagdasaryan and others, 'How to backdoor federated learning', 2019, pt. 1, arXiv:1807.00459.

participants. Research[32] shows that, in certain cases,[33] these attacks can successfully compromise the performance of the final federated learning model.

In light of the above, there is a very active line of research on the protection of federated learning models from data and model poisoning attacks. Recent works have shown that it is possible to mitigate the effect of both data and model poisoning attacks, in cases where the attacker aims to compromise the overall system's performance.[34] However, protection against more targeted attacks[35] is more challenging and may, in some cases, not be possible.[36]

2.3.2. Membership and property inference attacks

While federated learning models can, in certain cases, [37] be more vulnerable to poisoning attacks than centralized machine learning models, they are arguably less vulnerable to certain privacy attacks. In adversarial machine learning,[38] privacy attacks are defined as attacks aimed at "obtaining private information about the machine learning system, the data used for training, or the users of the system".[39] For the purposes of our discussion, we focus on membership and property inference attacks.

Membership inference attacks occur at test time (when the model is deployed) and are aimed at inferring whether a specific data point (that is, a record of an individual with some characteristics, e.g. a picture of Adam Smith) has been used for training the machine learning model. If the properties of some of these data points are unique to certain individuals, the attacker could be capable to indirectly identify these individuals with a high degree of confidence. In order to perform these attacks, the adversary needs to build a surrogate machine learning model (similar to the target machine learning model) and systematically query the target machine learning model in order to infer, with a certain degree of confidence, whether a specific data point (or the data from a specific individual) has been used to train the target model. For example, the attacker could query the machine learning model with different pictures of Adam Smith, observe the target model's output and compare it with the outputs provided by the surrogate model. It can be argued that federated learning models are less vulnerable to these attacks than centralized machine learning models. The reason for this is that federated

[32] P. Blanchard and others, 'Machine learning with Adversaries: Byzantine Tolerant Gradient Descent', *NeurIPS* 2017, pp. 119–129.
[33] This is the case when standard federated learning aggregation schemes (e.g. model averaging) are used.
[34] Blanchard and others (n 32), pp. 119–129; E.M.E., Mhamdi, R. Guerraoui and S. Rouault, 'The Hidden Vulnerability of Distributed Learning in Byzantium', *ICML* 2018, pp. 3518–3527; L. Muñoz-González, K.T. Co, and E.C. Lupu, 'Byzantine-Robust Federated Machine learning through Adaptive Model Averaging', 2019, arXiv preprintarXiv:1909.05125.
[35] A.N. Bhagoji and others, 'Analyzing Federated learning through an Adversarial Lens', *ICML* 2019, pp. 634–643.
[36] Reasons for this are that the effect of the attacks is very subtle, only affects a small set of inputs, and the manipulated model updates are very similar to the non-manipulated ones.
[37] This is the case, as already mentioned, when using standard aggregation schemes.
[38] Adversarial machine learning is the "research area that lies at the intersection of machine learning and cybersecurity, which aims to understand the vulnerabilities of existing machine learning algorithms and the development of new, more secure algorithms", Muñoz-González and Lupu (n 30), p. 48.
[39] Idem, p. 50.

learning models typically allow to use more training data. This implies that, in order to build the surrogate machine learning model, the adversary will typically need to gather a larger dataset to attack a federated learning model, compared to the dataset required to attack a centralized machine learning model. Moreover, in federated learning, even if attackers can infer whether a specific data point has been used to build the target model, they cannot, in most cases, identify the data source.[40]

Research on privacy attacks against federated learning models has devoted significant attention to property inference attacks.[41] Property inference attacks are aimed at inferring properties of the training data (i.e. broader categories, for example, gender or race), rather than a specific data point. As these attacks are directed against the updates exchanged between the training participants and the central server, they can only occur in federated learning settings, not in centralized ones. They can typically be performed by malicious training participants.[42] Although research[43] has shown that these attacks can be successful, the circumstances in which they can be effective are very limited and, in most cases, unrealistic. More specifically, a property inference attack allowing the attacker to identify an individual would require the following conditions to be met:

- the number of training participants should be very small, typically two (the adversary and an honest participant);
- the properties of the datasets provided by the participants should be very different;
- the training procedure (at each iteration) is performed using very specific settings.

The abovementioned conditions are, however, unrealistic in most cases since:

- there is always some overlap between the datasets provided by all the participants. In other words, the datasets provided by the different data holders are expected to have some properties in common, otherwise, if the datasets were completely different, the resulting federated learning model could perform poorly;
- when the number of honest participants increases (i.e. it is more than one), the attacker could only infer properties of the aggregated dataset. In other words, the attacker would, in most cases, not be able to identify the specific properties of the data used by a specific participant;
- the settings typically used to train federated learning algorithms do not correspond to the settings the adversary needs to perform most of these attacks.

3. Data protection by design in action

Federated learning is a technique aimed at implementing data protection by design, as enshrined in article 25.1 GDPR. Generally speaking, this principle requires controllers to implement technical and

[40] This would only be possible in the particular case where just two participants perform the federated learning task and the adversary is one of those participants, i.e. an insider.
[41] L. Melis and others, 'Exploiting Unintended Feature Leakage in Collaborative Learning', *IEEE Symposium on Security and Privacy*, 2019, pp. 691-706.
[42] More specifically, malicious training participants can perform such attacks by leveraging the knowledge of their own dataset and the parameters of the federated learning model that are sent to all the participants at each iteration during the training phase.
[43] Melis and others (n 41).

organizational measures to translate data protection principles when designing a system. To borrow the language of privacy engineering scholars like Hoepman, it can be said that federated learning implements the "privacy by design strategies" to "minimize" and "separate" the processing of personal data as much as possible.[44] The first strategy is intended to reduce the amount of personal data used, and also implies that one should minimize the quantity of personal data shared with third parties.[45] The second strategy is intended to make "[…] it harder to combine or correlate data […]". [46]

It follows that federated learning can facilitate compliance with the principle of data minimization (article 5.1(c) GDPR).[47] This principle requires the data controller to limit the personal (training) data used to build the machine learning model to what is adequate, relevant and necessary to achieve the purposes for which those data are processed. Indeed, in a federated learning setting, the data holder does not transfer the 'raw' personal data used to train the machine learning model to another entity. Therefore, training a model under the federated learning paradigm allows to avoid the duplication of personal data. It can also be argued that, by avoiding the centralization of personal data, federated learning contributes to reducing the likelihood that these data are combined with one another (and/or other data) and, subsequently, processed for purposes incompatible with the original purpose of collection. Consequently, federated learning could facilitate compliance with the principle of purpose limitation (article 5.1(b) GDPR). Another advantage of federated learning is that, as mentioned above, the resulting models are arguably less vulnerable to membership inference attacks than models trained in a centralized fashion. This reduces the ambiguity, highlighted by some data protection scholars,[48] regarding the possible qualification of machine learning models amenable to membership inference attacks as personal data.

4. Key questions

While the advantages of using federated learning have already been noticed by some data protection authorities,[49] little consideration has been devoted to three crucial questions, which we believe call for the attention of organizations intending to deploy federated learning.

4.1. Which processing is covered by the GDPR?

A first question relates to the extent to which training a machine learning model with federated learning qualifies as a "processing of personal data wholly or partly by automated means […]" (article

[44] J.-H. Hoepman, *Privacy Design Strategies (The Little Blue Book)*, 2020, pp. 5–9, www.cs.ru.nl/~jhh/publications/pds-booklet.pdf.
[45] Idem, pp. 5–7.
[46] Idem, p. 8.
[47] See also: Binns and Gallo (n 9); Datatilsynet, 'Artificial Intelligence and Privacy', 2018, p. 26, www.datatilsynet.no/globalassets/global/english/ai-and-privacy.pdf.
[48] For authors expressing the view that machine learning models susceptible to membership inference attacks could qualify as personal data: M. Veale, R. Binns and L. Edwards, 'Algorithms That Remember: Model Inversion Attacks and Data Protection Law', *Phil. Trans. R. Soc. A 376* 2018, pp. 1-15. For authors arguing against this view: MR Leiser and F. Dechesne, 'Governing Machine-Learning Models: Challenging the Personal Data Presumption', *International Data Privacy Law* (10) 2020, pp. 187- 200.
[49] See e.g. (n 47).

2.1 GDPR). We distinguish between processing operations on the 'raw' training data and operations on the model updates.

There is little doubt that, if the 'raw' training data provided by the data holder(s) qualify as personal data, the operations performed on such data in the context of the federated learning process will fall under the material scope of the GDPR. Let us, for example, consider the case of a group of hospitals training a machine learning model for diagnostic healthcare using federated learning. If the 'raw' training data provided by the hospitals consist of MRI scans displaying the patient's unique identifier number, these data are likely to qualify as a "special category of personal data" (albeit pseudonymous) under articles 4(1) and 9.1 GDPR. Moreover, the processing operations performed on these 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. Training a machine learning model using the federated learning approach does, hence, not necessarily exempt all the processing operations occurring in the federated learning framework from the application of the GDPR. In the abovementioned example, this implies that hospitals will be allowed to train a federated learning model with personal health data, only if the arguably ambiguous[50] conditions concerning legal re-use of personal data for AI training are met. These require, among others, that the processing rests on one of the legal grounds listed in articles 6.1 *juncto* 9.2 GDPR and that the purpose of the training is compatible with the purpose for which the data were originally collected, pursuant to articles 5.1(b) and 6.4 GDPR.

It is more complicated to ascertain whether the operations performed on the model updates derived from the 'raw' training data fall under the material scope of the GDPR. As mentioned above, these updates may, in certain cases, reveal (information about) the underlying (personal) training data. This means that they may qualify as data relating to an "identifiable natural person" under article 4(1) GDPR. Whether data relates to an identifiable individual must be determined on the basis 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", taking into account objective factors such as costs, amount of time required for identification, the technology available at the time of the processing and future technological developments (Recital 26 GDPR). Considering, on the one hand, the potentially broad interpretation of the notion of identifiability,[51] and, on the other, the possibility of updates leaking the underlying training data as well as their (theoretical) vulnerability to property inference attacks, it cannot be excluded that, in certain specific settings,[52] these updates may qualify as personal data.[53] Should that be the case, the controller(s) responsible for the processing operations on these data will also have to ensure that the processing of model updates complies with the GDPR. This means, among others, ensuring that the updates are processed securely by the entity running the central server, as required under article 32 GDPR. Moreover, if this entity qualifies as a processor under

[50] P. Hacker, 'A Legal Framework for AI Training Data', *SSRN* 2020, p. 27.
[51] M. Finck and F. Pallas, 'They Who Must Not Be Identified—Distinguishing Personal from Non-Personal Data under the GDPR', *International Data Privacy Law* (10) 2020, pp. 11–20.
[52] This could be the case especially if federated learning is not combined with other privacy preserving technologies.
[53] For a similar argument raised in relation to certain machine learning models trained centrally: Veale, Binns and Edwards (n 48).

the GDPR, appropriate (contractual or statutory) confidentiality obligations will have to be imposed, as envisaged by article 28.3 (b) GDPR.

4.2. Who controls what?

A second question is related to the allocation of data protection responsibilities in relation to processing operations occurring in the framework of complex, distributed systems such as federated learning. The GDPR relies on the binary controller-processor distinction. The controller determines "alone or jointly with others, the purposes and the means of the processing" (article 4 (7) GDPR) and the processor "processes personal data on behalf of the controller" (article 4 (8) GDPR). As remarked by Van Alsenoy, it is often uncertain how the criteria for identifying these actors should be applied in practice.[54]

This uncertainty has arguably been amplified by the broad interpretation of the notion of joint control[55] in the recent rulings of the Court of Justice of the European Union ("**CJEU**") in, among others, Fashion-ID.[56] Indeed, as argued by Advocate General Bobek in *Fashion ID*, taken at its extreme, this ruling may result in every actor that makes the processing of personal data possible qualifying as a joint controller.[57]

Recent guidance by the European Data Protection Board ("**EDPB**") on the concepts of controller and processor under the GDPR has not cleared away the ambiguity and potential for broad interpretation. According to the EDPB, "the overarching criterion for joint controllership to exist is the joint participation of two or more entities in the determination of the purposes and means of a processing operation".[58] A 'joint determination' means (among others) a common decision, which implies that the actors decide together and have a common intention.[59] A 'jointly determined purpose' is a purpose that is either (i) identical, (ii) common, (iii) closely linked or (iv) complementary to the purpose pursued by another entity.[60] This could be the case "when there is a mutual benefit arising from the same processing operation […]".[61] 'Jointly determining the means' can follow from a situation in which a given entity makes use of a technology developed by another entity for its own purposes.[62] The EDPB confirms that access to personal data is not a prerequisite for (joint) control.[63] Critically, the EDPB adds, control "may extend to the entirety of the processing at issue" or "be limited to a particular stage in

[54] B. Van Alsenoy, *Data Protection Law in the EU: Roles, Responsibilities and Liability,* KU Leuven Centre for IT and IP Law, 1st edn, Intersentia, 2019, p. 9.
[55] C. Millard and others, 'At This Rate, Everyone Will Be a [Joint] Controller of Personal Data!', *International Data Privacy Law* (9) 2019, pp. 217-219.
[56] CJEU 29 July 2019, C-40/17, ECLI:EU:C:2019:629 ('*Fashion ID*'), *Computerr.* 2019/219 .
[57] AG M. Bobek, *Opinion in Fashion ID,* 2018, ECLI:EU:C:2018:1039, para. 74.
[58] EDPB, 'Guidelines 07/2020 on the Concepts of Controller and Processor in the GDPR', 2 September 2020, p. 17.
[59] Idem, p. 18.
[60] Idem, p. 19.
[61] Idem.
[62] Idem, p. 20.
[63] Idem, p. 16.

the processing".[64] However, like the CJEU in Fashion-ID,[65] the Board fails to specify how to identify the relevant (part of) the processing operation in relation to which control should be assessed. These are evidently open-ended and ambiguous legal criteria that are challenging to apply in practice. As pointed out by other scholars,[66] it is, for instance, highly unclear how granularly the notion of purpose should be defined when assessing control.

The complexity of a typical federated learning ecosystem amplifies the challenges. First, as mentioned under paragraph 4.1, there is the arduous task of discerning which specific processing operations, out of the bundle of operations occurring in the federated learning framework, relate to personal data. Second, there is the complexity due to the multitude of actors (potentially, millions) providing 'raw' training data. In such an extremely complex environment, discerning and allocating each actor's responsibility for compliance with the GDPR, particularly vis-à-vis the data subjects, may be difficult and, if not done clearly, result in a lack of transparent and fair processing contrary to article 5.1(a) GDPR.[67]

4.3. Who controls the (raw training data provided by the) controllers?

A third question that partially follows from the second one relates to the specific problem that federated learning intends to solve: training a global model with potentially tens to millions of training participants, where the 'raw' training data provided by each training participant can "by design" not be inspected by other actors than the holder of the data.[68] As mentioned above, this may render federated learning vulnerable to poisoning attacks by training participants, which can, in certain cases, significantly impair the performance of the final federated learning model.

This obviously implies that, if the final federated learning model is used to infer new personal data, those data risk being inaccurate. This could, depending on the purpose for which the data are processed,[69] result in an infringement of the accuracy obligation laid down in article 5.1(d) GDPR.[70] This vulnerability, coupled, in some cases, with the difficulty to detect (the source of) the poisoned contribution, renders the need for *ex ante* accountability measures (particularly, those concerning the 'quality' of the training data)[71] all the more important in the context of federated learning. On the basis of these measures, each training participant should be able to show continuous compliance with the

[64] Idem, p. 15.
[65] R. Mahieu and J. van Hoboken, 'Fashion-ID: Introducing a Phase-Oriented Approach to Data Protection?', *European Law Blog*, 30 September 2019.
[66] C. Ducuing and J. Schroers, 'The recent case law of the CJEU on (joint) controllership: have we lost the purpose of 'purpose'?' (2020) 6 Computerrecht: Tijdschrift voor Informatica, Telecommunicatie en Recht 429.
[67] Article 29 Working Party, 'Opinion 1/2010 on the Concepts of "Controller" and "Processor"', 16 February 2010, p. 24
[68] Bagdasaryan and others (n 31).
[69] CJEU 20 December 2017, Case C-434/16 ('*Nowak*'), ECLI:EU:C:2017:994, para. 53, *NJ* 2018/314.
[70] D. Hallinan and F. Zuiderveen Borgesius, 'Opinions Can Be Incorrect (in Our Opinion)! On Data Protection Law's Accuracy Principle', *International Data Privacy Law* 2020, pp. 1-10; Article 29 Working Party, 'Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679', 3 October 2017, pp. 17–18.
[71] On the topic of (among others) quality risks in AI training data: Hacker (n 50) pp. 1-35.

GDPR [72] It will, for example, be particularly important that each training participant carefully documents each training dataset used.[73] Moreover, clear protocols should be established specifying which requirements the training data should meet, in light of (among others) the purpose and target population to which the federated learning model will be applied.

Whether the relation between training participants is qualified as a controller-processor or a (joint) controllership relationship, each training participant would also be recommended to conduct a careful due diligence investigation[74] of the other training participants' compliance with the GDPR, before venturing into a federated learning scheme. Again, considering the very large number of training participants that may be involved in a federated learning scheme, carrying out this *ex ante* screening process may not come without practical difficulties.

5. Conclusion

As is usually the case with privacy preserving technologies, when considered in isolation, federated learning is no silver bullet. Although it can, under certain circumstances, help facilitate compliance with some data protection principles, it does not, as such, exempt organizations from the GDPR's application, especially if the raw training data qualifies as personal data. Moreover, in light of the difficulty to audit, at a system level, the data used for training the federated learning model, the use of federated learning requires an increased attention to *ex ante* (technical and organizational) accountability measures. These measures should demonstrate that each entity participating in the federated learning scheme is continuously willing and able to comply with the GDPR. Further interdisciplinary research should be devoted to investigating which measures are suitable and recommended for adoption into large machine learning environments, such as the ones in which federated learning is typically intended to be used.

[72] As required under article 5.2. GDPR for controllers and article 28.1 GDPR for processors

[73] This obligation can, in relation to controllers, arguably be derived from article 30 GDPR. On the need for documentation of machine learning training datasets in general: T. Gebru and others, 'Datasheets for Datasets', 2020, pp. 1-24, arXiv:1803.09010.

[74] M. Hintze, 'Data Controllers, Data Processors, and the Growing Use of Connected Products in the Enterprise: Managing Risks, Understanding Benefits, and Complying with the GDPR', *Journal of Internet Law* 2018, pp. 20–21.