Federated Learning: a Tangible Solution for GDPR Compliance?*

Matteo Johannes Bürgler

Table of contents

1. Introduction

The rapid evolution of artificial intelligence (AI) has outpaced the assumptions underlying European data protection frameworks. As the core data protection principles in art. 5(1) GDPR[1] stand in great tension with the technical realities of AI training, numerous issues arise. The AI Act[2] fails to resolve these, risking the stifling of European innovation. Merging both a legal and technological perspective, this article aims to provide pathways toward legally compliant and operationally sustainable AI development. First, it discusses AI development under the GDPR, examining inherent technical limitations and trade-offs of centralised AI training. Then it presents Federated Learning (FL) as a decentralised, privacy-preserving alternative. After explaining its distinct technical features, this article evaluates whether FL provides a tangible solution toward GDPR compliance. Additionally, it examines how this technology aligns with the AI Act, discussing how its technical characteristics complicate data governance, bias mitigation, and robustness while raising concerns regarding energy efficiency.

* Su determinazione della direzione, il contributo è stato sottoposto a referaggio anonimo in conformità all'art. 15 del regolamento della Rivista

[1] Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L 119/1 (GDPR).

[2] Regulation (EU) 2024/1689 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] OJ L 2024/1689 (AI Act).

2. Core GDPR Considerations for AI Development

Training trustworthy, human-centred AI systems[3] requires access to large-scale datasets, which frequently need to include personal data to achieve reliable, non-discriminatory outputs. Throughout AI training, personal data may be processed in various forms, including raw data collection; preprocessing; anonymisation, pseudonymisation, and synthetic data generation; training, validation, and testing; and runtime input.[4] Although AI systems are minimally addressed in the GDPR, art. 2(7) AI Act stipulates full applicability of the GDPR to AI development. Moreover, art. 47 AI Act obliges providers of high-risk AI systems to affirm GDPR-compliance in their conformity declarations.[5] However, legal issues arising from the overlap remain largely unaddressed. As the AI Act lacks clear data protection provisions, the existing GDPR – with its ambiguous provisions and broad discretion – must be applied to AI systems through statutory interpretation.[6]

Initially, the GDPR appears to oppose AI systems' characteristics. Particularly concerning are, for example, the data-intensity of AI systems along with their limited explainability, traceability, and predictability ("complexity"). Additionally, AI systems exhibit fragility in robustness and information security.[7] These issues create notable tensions and legal uncertainties along the AI value chain.[8]

Especially problematic is the correctness of outputs, as training data can be inaccurate and models may hallucinate – generating plausible yet incorrect results –, raising concerns under the data accuracy principle of art. 5(1)(d) GDPR. The GDPR's right to rectification (art. 16), right to be forgotten (art. 17), and right to withdraw consent (art. 7(3)) also pose obstacles. As it is difficult to determine how individual training data influenced parameter weighting, extracting or correcting specific information from trained AI systems ("machine unlearning"[9]) is complex and inherently limited. As training requires substantial energy and compute, retraining systems in response to individual requests is commercially unviable. Filters, therefore, are the best solution but must be designed data protection-friendly, as they also contain and process personal data[10]. Issues similarly arise regarding the informational obligations towards the data subject as they must be

---

[3] European Commission, *Ethics Guidelines for Trustworthy AI*, 2019.

[4] A. Anderl – A. Ciarnau, *Datenschutzrechtliche Herausforderungen bei Generativer KI*, in *ecolex*, 12, 2023, 1078 ff, 1079.

[5] Art. 16(g), annex V(5) AI Act.

[6] B.P. Paal, *Spannungsverhältnis von KI und Datenschutzrecht*, in M. Kaulartz – T. Braegelmann (eds.), *Rechtshandbuch Artificial Intelligence und Machine Learning*, Beck, 2020, paras 1 ff, para. 1.

[7] D. Bierbauer, *Datenschutzrechtliche Grundsätze bei der Entwicklung Künstlicher Intelligenz*, in D. Jahnel (ed.), *Jahrbuch Datenschutzrecht 2021*, Wien, 2022, 175 ff., 179.

[8] B.P. Paal, *Spannungsverhältnis*, cit., para. 2.

[9] H. Zu et al., *Machine Unlearning: A Survey*, in *ACM Comuting Surveys*, 56, 2023, 1 ff, 1.

[10] P.J. Pesch – R. Böhme, *Verarbeitung personenbezogener Daten und Datenrichtigkeit bei großen Sprachmodellen*, in *MMR*, 2023, 917 ff, 922.

fulfilled even when data have not been obtained directly from the data subject (art. 14 GDPR) like in "web scraping". Exemptions, such as when providing information entails disproportionate effort, require sensitive balancing.[11]

2.1 Legal Basis for AI Development

When raw training data contain personal data, processing must comply with art. 6 GDPR and, for special categories of personal data, with art. 9 GDPR. In AI development, lawfulness typically relies on data subjects' consent or controllers' legitimate interests. Particularly when "web scraping", legitimate interests are often invoked, as no contractual relationship exists and obtaining consent from all data subjects would be commercially unviable. However, the broad and ambiguous scope of legitimate interests under art. 6(1)(f) GDPR and the balancing of interests it entails make its application delicate. While CJEU case law[12] and opinions of the Article 29 Data Protection Working Party (WP 29)[13] provide guidance, legal uncertainty for controllers and unpredictability for data subjects persist.[14]

When data are collected close to data subjects, one might tend to precautiously obtain consent to address these concerns; however, consent is not a universal solution. arts. 4(11), 6(1)(a), 7, and 8 GDPR require consent to be given freely, in an informed manner, and for specific purposes. These criteria raise issues regarding the legal validity of consent. Moreover, when consent is invalid or withdrawn under art. 7(3) GDPR, it is disputed whether another legal basis can be subsidiarily invoked. Changing the legal basis may mislead data subjects, as obtaining consent implies they retain control over the processing through the right of withdrawal.[15] The European Data Protection Board (EDPB) has opposed switching to legitimate interests, arguing that controllers can only meet the obligation to disclose the legal basis at the time of data collection if the legal basis has been determined in advance.[16]

In principle, processing begins with data collection and ends when the purpose is achieved, triggering the obligation to erase the data (principle of storage limitation under art. 5(1)(e) GDPR). However, as the training of AI systems requires vast amounts of data, collecting data specifically for AI development purposes is often unfeasible. Thus, AI development frequently relies on secondary processing, raising concerns regarding the principle of purpose limitation, which prohibits processing incompatible

[11] L. Franke, *Datenschutzrechtskonformes Training von KI-Systemen mit öffentlich verfügbaren personenbezogenen Daten*, in *Rdi*, 2023, 565 ff., 568 ff.

[12] ECJ, C-708/18, *TK v. Asociația de Proprietari bloc M5A-ScaraA* (2019).

[13] 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*, WP 217, 2014, 1 ff.

[14] L. Franke, *Datenschutzrechtskonformes Training*, cit., 567 ff.

[15] D. Jahnel, *Artikel 6*, in D. Jahnel (ed.), *DSGVO*, Wien, 2020, paras. 1 ff, paras. 8 ff.

[16] European Data Protection Board, *Guidelines 05/2020 on consent under Regulation 2016/679*, Version 1.1, 2020, paras. 1 ff, para. 123.

with the purposes for which data were originally collected. To such secondary processing, the ambiguous conditions for re-use of personal data under art. 6(4) GDPR apply. Apart from consent, which has inherent limitations, practice will rely on demonstrating the compatibility of secondary processing with the purposes of initial collection. Relevant factors include the link between the original and secondary purpose, the context of data collection considering the relationship between the data subject and the controller, the nature of the personal data, the potential impact on data subjects, and the presence of appropriate safeguards. Recital 50 GDPR further emphasises the reasonable expectations of data subjects. Purpose compatibility is particularly likely when further processing is implied by the original purpose, constituting the next logical step. Demonstrating such purpose compatibility entails uncertainty, as requirements are ambiguous, further complicating GDPR-compliant training.[17]

2.2 Remedies for Centralised AI Development

A potential remedy lies in the exclusion of any personal reference according to recital 26 GDPR, rendering the GDPR inapplicable. This is the case when training sets consist solely of anonymous data, which have never related to an identified or identifiable person, or of anonymised data, where personal reference has been removed to the extent that individuals are not identifiable given all means reasonably likely to be used.[18] As the CJEU clarified, anonymisation does not require the absolute or irreversible impossibility of re-identification; rather, the relevant actors must not have a reasonable likelihood of accessing re-identification information.[19] Notably, because anonymisation itself is a processing operation, it remains subject to the GDPR, though usually constituting compatible secondary processing.[20] Similarly, synthetic data – artificially generated data statistically resembling real datasets – offer a workaround because they do not, on their face, relate to real persons and thus render the GDPR inapplicable.[21] In contrast, if re-identification remains reasonably possible, data are at best pseudonymised.[22] Pseudonymisation replaces direct identifiers with other values to prevent immediate attribution to a person. However, quasi-identifiers or other attributes can still enable re-identification when combined with additional information. The key distinction from anonymisation lies in whether relevant actors have reasonable access to such re-identification

[17] D. Jahnel, *DSGVO*, cit., paras 100 ff; Article 29 Data Protection Working Party, *Opinion 03/2013 on purpose limitation*, WP 203, 2013, 1 ff, 23 ff; see art. 5(1)(b) GDPR.

[18] D. Jahnel – A Pallwein-Prettner, *Datenschutzrecht*, Wien, 2021, 57.

[19] ECJ, C-413/23 P, *EDPS v. SRB* (2025).

[20] Article 29 Data Protection Working Party, *Opinion 05/2014 on Anonymisation Techniques*, WP 216, 2014, 1 ff., 7 ff.

[21] B.P. Paal, *Spannungsverhältnis*, cit., para. 28; M. Kaulartz, *Personenbezug von KI-Modellen*, in M. Kaulartz –T. Braegelmann (eds.), *Rechtshandbuch Artificial Intelligence und Machine Learning*, Munich, 2020, paras. 1 ff, paras. 22 ff.

[22] D. Jahnel – A. Pallwein-Prettner, *Datenschutzrecht*, cit., 57; art. 4(5) GDPR.

information; if so, the GDPR remains applicable.[23] Nevertheless, pseudonymisation lowers risks, which supports controllers in the balancing of interests and the purpose-compatibility test. It also advances the principles of data minimisation and integrity/confidentiality while constituting an appropriate technical and organisational measure, thereby furthering "data protection by design" and "by default".[24]

Academics doubt data can be both effectively and durably anonymised, particularly when their informational value must be preserved.[25] The GDPR's absence of technical detail on anonymisation further complicates the issue as applicable criteria must be inferred by way of *argumentum e contrario* from the Regulation and relevant CJEU case law[26]. Accordingly, anonymisation is only sufficient when data subjects cannot be singled out at all or only with disproportionate effort. Recital 26 GDPR requires an assessment of all means reasonably likely to be used, considering cost, time, and the state of the art. The illegality of a re-identification method does not, by itself, exempt it from consideration. Future technological developments throughout the entire processing period must also be factored in. As a result, anonymity is often only temporary.[27] The WP 29 has identified three key re-identification risks: singling out, where individual or all data records are isolated, making data subjects identifiable; linkability, where datasets are combined to determine whether they relate to the same person; and inference, where an attribute's value is deduced with significant probability from other data points.[28]

Even if successful, anonymisation typically reduces data quality, leading to inaccuracies and representativeness gaps that can impair AI system performance – a notable concern under the AI Act. Moreover, most anonymisation methods are neither scalable nor automatable, with reliable results requiring significant manual effort. After anonymisation, data become static, meaning that once personal reference is removed, later enrichment with new subject attributes is impossible.[29] Also, there is no one-off anonymisation given cumulative re-identification risks: Each future extension of the dataset with other (anonymised) data, the availability of new publicly accessible data, or the creation of additional datasets obtainable with reasonable effort may enable fresh combinations that allow re-identification. Controllers must therefore continuously monitor and ensure that such

[23] ECJ, C-413/23 P, *EDPS v. SRB* (2025).

[24] Article 29 Data Protection Working Party, WP 216, cit., 20 ff.; BP Paal, *Spannungsverhältnis*, cit., para. 27a.

[25] B.P. Paal, *Spannungsverhältnis*, cit., paras. 26 ff.; M. Valkanova, *Trainieren von Modellen als Zweckänderung*, in M. Kaulartz –T. Braegelmann (eds), *Rechtshandbuch Artificial Intelligence und Machine Learning*, Munich, 2020, paras. 1 ff., paras. 23 ff.; C. Winter – V. Battis – O. Halvani, *Herausforderung für die Anonymisierung von Daten*, in *ZD*, 2019, 489 ff., 489.

[26] ECJ, C-582/14, *Breyer v. Bundesrepublik Deutschland* (2016); ECJ, C-434/16, *Nowak v. Data Protection Commissioner* (2017).

[27] P. Hacker, *A Legal Framework for AI Training Data - From First Principles to the Artificial Intelligence Act*, in *Law, Innovation and Technology*, 13, 2021, 257 ff., 265 ff.; M. Valkanova, *Trainieren als Zweckänderung*, cit., paras. 17 ff., 31.

[28] Article 29 Data Protection Working Party, WP 216, cit., 11 ff.

[29] M. Kaulartz, *Personenbezug*, cit., paras. 20 ff.

combinations remain infeasible.[30]

Synthetic data similarly raise significant caveats. Although such data are artificially generated and not directly linked to natural persons, their utility frequently depends on preserving statistical correlations with the underlying real dataset. This creates a residual risk of re-identification, particularly where synthetic records maintain unique outlier structures or enable linkage when combined with auxiliary information.[31] Scholars therefore emphasise that synthetic data cannot be assumed to be anonymous *per se* but require the same recital 26 GDPR assessment of identifiability as anonymised data.[32]

Consequently, anonymised or synthetic data only exempt AI training from the GDPR when true anonymity is achieved – a technically demanding, often temporary, and ultimately impractical threshold at scale. Even then, such data may degrade model performance, undermining the AI Act's objectives. Decentralised methods, however, offer a promising alternative, enabling training on large volumes of high-quality personal data while upholding data protection requirements.

3. Federated Learning as a GDPR Friendly Alternative

Introduced by McMahan et al.,[33] FL is a distributed machine learning paradigm that trains a single global model through decentralised compute at the edge of the network, e.g., end-user devices or companies' data centres. In contrast to centralised AI training, FL keeps data locally stored rather than transferring it to a single entity.

FL involves four steps: (1) The coordinating server ("central node" or "aggregator") broadcasts the current centralised model to client devices. (2) Each client locally trains an updated version of the model with their data. (3) Clients return the updated model to the central node while their data remain siloed. (4) The central node aggregates the updates into a new global model, which is then redistributed for the next round.[34]

FL is no longer merely a concept but powers real-world applications. A prominent example is healthcare, where FL allows hospitals to collaboratively train diagnostic models, especially in medical imaging, by leveraging large, diverse datasets without exchanging sensitive medical records.[35]

30 M. Valkanova, *Trainieren als Zweckänderung*, cit., para. 27. J Schefzig, *Big Data = Personal Data? Der Personenbezug von Daten bei Big Data-Analysen*, in *DSRITB*, 2014, 103 ff, 110 ff.

31 T. Stadler – B. Oprisanu – C. Troncoso, *Synthetic Data -- Anonymisation Groundhog Day*, arXiv:2011.07018, 2022, 1 ff.

32 M. Kaulartz, *Personenbezug*, cit., paras 21 ff.

33 H. Brendan McMahan et al., *Communication-Efficient Learning of Deep Networks from Decentralized Data*, arXiv:1602.05629, 2016, 1 ff.

34 S. Rossello – R.D. Morales – L. Muñoz-González, *Data protection by design in AI? The case of federated learning*, in *Computerrecht*, 116, 2021, 1 ff., 3.

35 D. Ciupek – M. Malawski – T. Pieciak, *Federated Learning: A new frontier in the exploration of multi-institutional medical imaging data*, arXiv:2503.20107, 2025, 1 ff.

3.1 GDPR Implications

Training of local models is still subject to the GDPR when personal data are processed. Therefore, data holders who are not the data subjects themselves may only engage in FL with personal data if processing is lawful and adheres to the data protection principles. While the ambiguous provisions on legitimate interests and compatible secondary processing still apply, this section will outline FL's key benefits.

As personal data are never disclosed, the resulting data lineage – including where the data originated, where they reside, and how they are used – is simpler to trace. This supports GDPR-compliance by promoting data subjects' right to obtain information under art. 15 GDPR. By decentralising training, FL also strengthens compliance with the security requirements under art. 32(2) GDPR: It eliminates vulnerabilities inherent in conventional AI development, such as man-in-the-middle attacks during data transfers, unauthorised access to centralised storage, and accidental disclosure due to human error in data handling.[36]

Beyond data security, FL exemplifies "data protection by design" in action.[37] By avoiding duplication, FL supports the principle of data minimisation under art. 5(1)(c) GDPR, as only strictly necessary information – namely, the updated model parameters – is shared and subject to further processing. By preventing centralisation, FL also reduces risks of unrelated data being combined and processed in a manner incompatible with the original purposes of collection, thereby arguably promoting purpose limitation under art. 5(1)(b) GDPR.[38]

FL allows training on raw, high-value personal data without prior anonymisation or pseudonymisation while remaining privacy-friendly. While local training is still subject to the GDPR, FL increases the chances of satisfying the balancing of interests under art. 6(1)(f) GDPR and passing the purpose compatibility test under art. 6(4) GDPR. Despite reduced risks from decentralised training and avoiding data duplication, shared local updates or the global model may still constitute personal data if re-identification is possible.[39] Thus, FL research focuses on integrating "privacy-enhancing technologies" (PETs) during aggregation to prevent such an extraction of personal data.[40] Common approaches include:

Secure aggregation: Cryptographic protocols such as Bonawitz et al.'s secure aggregation scheme let the server learn only the sum of all client updates, hiding every individual contribution even if the server itself is curious or compromised.[41]

Homomorphic encryption (HEC): In homomorphic schemes, clients en-

[36] H. Woisetschläger et al., *Federated Learning Priorities Under the European Union Artificial Intelligence Act*, arXiv:2402.05968, 2024, 1 ff, 3.

[37] S. Rossello – R.D. Morales – L. Muñoz-González, *Data protection by design in AI*, cit., 7.

[38] *Ibid.,* 8.

[39] M Kaulartz, *Personenbezug*, cit., paras. 26 ff.

[40] Ibid., 4; Y. Liu et al., *Secure Federated Transfer Learning*, arXiv:1812.03337, 2020, 1 ff.

[41] K. Bonawitz et al., *Practical Secure Aggregation for Federated Learning on User-Held Data*, arXiv:1611.04482, 2016, 1 ff.

crypt under a public key, and the server aggregates directly on ciphertexts, seeing neither intermediate values nor plaintexts, but only the final aggregate. Multi-key schemes (e.g., xMK-CKKS) avoid the need for a shared secret key and remove a major trust assumption by eliminating reliance on server-side confidentiality, albeit at the cost of higher compute and communication.[42]

Secure multiparty computation (SMPC): Each update is split into secret shares distributed among multiple non-colluding servers; only the joint aggregate is ever reconstructed. It is especially suitable for financial applications.[43]

Differential privacy (DP): By adding Gaussian (or Laplace) noise before the transmission, DP provides guarantees that the inclusion or exclusion of any single data subject changes the aggregate's distribution only negligibly, thereby only having a bounded impact on the output.[44]

However, it must be noted that FL further complicates "machine unlearning" and therefore compliance with data subjects' rights to erasure and rectification. Typical approaches such as "teacher–student frameworks" or "amnesic unlearning" typically require access to training data or even the entire training trajectory with client-level snapshots; resources that, by design, are never available in FL. Existing FL-specific unlearning research[45] therefore focuses on the removal of entire clients to support GDPR-compliance without accessing training data.[46]

3.2 Bridging Data Silos

Building large-scale datasets for centralised AI development is complex and often entails crawling the entire internet.[47] However, internet-scale data collection raises serious concerns about lawful processing and compliance with informational obligations. In AI development, data quality is as crucial as access,[48] yet much high-quality, representative personal data remain siloed and inaccessible. Even where valuable data are identified and processing is GDPR-compliant, the costs of transferring and storing such vast datasets must be considered.

FL overcomes these issues, minimising attack surfaces while simplifying GDPR-compliance. By reducing exposure to potential regulatory penal-

---

42 J. Ma et al., *Privacy-preserving Federated Learning based on Multi-key Homomorphic Encryption*, arXiv:2104.06824, 2021, 1 ff.

43 V. Mugunthan et al., *SMPAI: Secure Multi-Party Computation for Federated Learning*, at 33rd Conference on Neural Information Processing Systems, Vancouver, 2019, 1 ff.

44 A. Banse – J. Kreischer – X.O. Jürgens, *Federated Learning with Differential Privacy*, arXiv:2402.02230, 2024, 1 ff.

45 A. Halimi et al., *Federated Unlearning: How to Efficiently Erase a Client in FL?* arXiv:2207.05521, 2022, 1 ff.

46 H. Woisetschläger et al., *Federated Learning Priorities*, cit., 3.

47 *Ibid,* 7.

48 S.E. Whang et al., *Data collection and quality challenges in deep learning: a data-centric AI perspective*, in *The VLDB Journal*, 32, 2023, 791 ff.

ties, FL may encourage data providers to share otherwise siloed personal data. By facilitating broader access to high-quality, up-to-date datasets, FL bridges the gap between data availability and regulatory constraints, enabling more representative and robust AI training. This, in turn, helps prevent discriminatory outcomes and negative effects on fundamental rights.[49]

4. Accommodating Federated Learning within the AI Act

The AI Act recognises the potential of FL, endorsing FL as a promising approach to safeguarding privacy and data protection throughout the entire AI lifecycle. The AI Act emphasises the importance of data minimisation and "data protection by design" and "by default", noting that compliance measures may extend beyond anonymisation and encryption to « the use of technology that permits algorithms to be brought to the data and allows training of AI systems without the transmission […] of […] data themselves ».[50]

Notably, most applications that benefit from FL, such as healthcare or critical infrastructure, are classified as high-risk under art. 6(2), Annex III AI Act, entailing obligations such as data and data governance (art. 10) and accuracy, robustness, and cybersecurity (art. 15).[51] Yet the AI Act, drafted primarily with centralised training in mind, does not fully reflect FL's distinctive characteristics. As the following analysis demonstrates, FL encounters significant challenges in meeting certain core obligations while also raising concerns over energy consumption. Accordingly, fully accommodating FL requires further research on energy efficiency as well as a flexible interpretation of the AI Act's essential requirements; absent this, a dedicated framework for FL may be necessary.

4.1 Data Governance

Art. 10 AI Act imposes strict data and data governance requirements throughout the entire model lifecycle. Practices shall include an « examination in view of possible biases that are likely to affect the health and safety of persons, have a negative impact on fundamental rights or lead to discrimination »[52] and « appropriate measures to detect, prevent and mitigate possible biases »[53]. In AI systems, bias refers to systematic deviations producing unfair outcomes, often disproportionately disadvantaging certain groups due to historical inequities in training data and system design choices. In FL, such bias may also stem from non-IID (non-independent

[49] H. Woisetschläger et al., *Federated Learning Priorities*, cit., 7 ff.

[50] Recital 69 AI Act.

[51] H. Woisetschläger et al., *Federated Learning Priorities*, cit., 2.

[52] Art. 10(2)(f) AI Act.

[53] Art. 10(2)(g) AI Act.

and identically distributed) data, meaning that data distributions vary significantly between participants, leading to models favouring dominant groups.[54] However, studies have broadened the understanding of bias in FL beyond non-IID data. Bias may also result from clients not representing the true population distribution, biased local data, and systemic societal bias. Technical sources of bias may include disparities in compute, data quantity, and device capability (e.g., varying sensor qualities), as well as communication constraints, convergence difficulties, and misaligned incentives among competing participants.[55]

Although FL enables access to more high-value, up-to-date datasets and thereby supports the development of more representative models, it must comply with art. 10 AI Act. As training data are inaccessible by design in FL, data quality and bias must be assessed through indirect methods. However, art. 10 AI Act does not explicitly consider these unique characteristics, raising concerns about how FL can effectively meet its requirements. In particular, it remains unclear whether techniques combating non-IID data[56] provide enough robustness for compliance or whether additional safeguards are necessary.[57]

Research has proposed several paths forward. For instance, “agnostic federated learning” can improve fairness and robustness by minimising worst-case loss across heterogeneous client data. This means that a model is trained to perform reliably across all possible user distributions instead of optimising for average performance across all clients. Crucially, this does not correct biases in the training data itself but tweaks the model to prevent any single group from experiencing disproportionately poor performance.[58] Similarly, “PrivFairFL” promotes fairness by decentralised computation of group-level performance metrics through SMPC while DP adds noise. These aggregated metrics enable bias detection and fairness auditing, allowing model adjustments without exposing individual identities.[59] Ultimately, such methods offer a pathway to meet the art. 10 AI Act obligations while preserving the decentralised nature of FL. However, they may entail higher energy costs compared to centralised AI training.[60]

4.2 Robustness and Quality Management

Arts. 15, 17 AI Act introduce obligations to ensure accuracy, robustness,

---

[54] H. Woisetschläger et al., *Federated Learning Priorities*, cit., 3.

[55] M. Benmalek – A. Seddiki, *Bias in Federated Learning: Factors, Effects, Mitigations, and Open Issues*, in *Ingénierie des Systèmes d'Information*, 29, 2024, 2137 ff.

[56] Y. Zhao et al., *Federated Learning with Non-IID Data*, arXiv:1806.00582, 2018, 1 ff.

[57] H. Woisetschläger et al., *Federated Learning Priorities*, cit., 3, 8.

[58] M. Mohri – G. Sivek – A.T. Suresh, *Agnostic Federated Learning*, arXiv:1902.00146, 2019, 1 ff.

[59] S. Pentyala et al., *PrivFairFL: Privacy-Preserving Group Fairness in Federated Learning*, arXiv:2205.11584, 2022, 1 ff.

[60] H. Woisetschläger et al., *Federated Learning Priorities*, cit., 3, 8.

and cybersecurity as well as quality management. These requirements necessitate close monitoring during training to ensure consistent, reliable performance. In FL, however, fulfilling these monitoring requirements may require frequent validation and communication, threatening increased energy consumption and further widening the performance gap to centralised learning.[61]

A common approach to ensure robustness is to track model performance with a validation dataset – that is, data not used during training and seen by the model during validation for the first time. When performance in validation data decreases while performance in training data still increases, training is stopped early. This technique prevents "overfitting", which occurs when a model adapts too closely to the training data, thereby reducing its ability to generalise to new data.[62] Notably, "overfitting" also increases the risk of reproducing training data, potentially enabling re-identification of data subjects and rendering the model itself personal data.[63]

In FL, server-side validation, in which model updates are validated at the central node before proceeding with the next training round, is common. The resulting idle times typically account for approximately 31% of total power usage.[64] This creates a trade-off; either models are validated centrally before the next training round, causing clients to idle, or training continues in parallel, increasing the risk of "overfitting" and wasting the energy.[65] To address this, several strategies have been proposed: (1) validating only after completed training, which saves energy but risks creating a non-functional AI model, wasting all energy consumed, and conflicts with art. 17 AI Act; (2) validating after every *i*th aggregation round, balancing robustness and energy efficiency by adjusting validation frequency and therefore idle time; or (3) validating asynchronously by starting the next aggregation round while validation is still ongoing, minimising idle time but risking "overfitting".[66] Given these trade-offs, research focuses on improving validation efficiency per data sample through techniques like dataset distillation.[67]

When server-side validation is not possible, as in HEC-architectures, validation must be performed client-side. This approach, however, depends on the trustworthiness of clients and the secure distribution of validation datasets. Notably, FL providers remain responsible under the AI Act for ensuring that the final model meets robustness and quality requirements.[68]

---

61 *Ibid.,* 3 ff.

62 L. Prechelt, *Early stopping-but when?* in G.B. Orr – K.R. Müller (eds.), *Neural Networks: Tricks of the Trade*, 2nd edn, Berlin, 2012, 53 ff.

63 M. Kaulartz, *Personenbezug*, cit., paras 10 ff.

64 H. Woisetschläger et al., *Federated Learning Priorities*, cit., 7.

65 *Ibid.,* 4.

66 *Ibid.,* 7.

67 S. Lei and D. Tao, *A comprehensive survey to dataset distillation*, arXiv:2301.05603, 2023, 1 ff.

68 H. Woisetschläger et al., *Federated Learning Priorities*, cit., 7.

4.3 Energy–Privacy Trade-off

FL's data privacy comes at a cost: Privacy-preserving methods such as SMPC require substantial compute and communication, reducing energy efficiency. While DP avoids heavy cryptographic costs, small-scale FL requires stronger noise, degrading model performance and slowing training. As a result, total energy consumption also increases as more training iterations are needed.[69]

Beyond technical efficiency, the environmental impact of FL depends on deployment context. While state-of-the-art data centres are increasingly energy-efficient and often use renewable energy, FL operates at the network edge, where devices rely on the local energy mix, which is typically more carbon-intensive. Although specialised edge hardware can match data centre GPUs in throughput ("samples per watt"), this advantage fades when considering "time-to-accuracy". FL also raises questions of energy accountability, as devices are typically not owned by the model provider.[70] Nonetheless, FL avoids transmission and duplicated storage costs associated with centralised AI training, partially offsetting its environmental footprint.

Despite its greater energy demands, FL faces no binding consumption limits under the AI Act. Instead, the AI Act introduces voluntary pathways such as art. 40 harmonised standards and art. 95 codes of conduct, which FL deployments may struggle to satisfy due to compute and communication overhead. However, training a general-purpose AI (GPAI) model via FL could still trigger regulatory scrutiny. The significant compute demands of FL (which are indicated by estimated training energy consumption) may lead to its classification as GPAI with systemic risk.[71] Finally, since art. 112 AI Act mandates periodic reviews on energy-efficient development and empowers the Commission to propose binding amendments, FL's energy demands escape hard caps today but may face genuine constraints in the next AI Act revision.

5. Conclusion

FL is "data protection by design" in action.[72] It supports art. 5(1) GDPR principles, mitigates ambiguous provisions on lawful processing such as legitimate interests and compatible secondary processing, and enables simpler data lineage. Integrating PETs into FL reduces re-identification risks and prevents local or global models from being personal data. Additionally, FL's privacy-friendly nature reduces exposure to regulatory penalties, thereby promoting cooperation and unlocking siloed datasets. This enables access to high-value, up-to-date data that promise to improve model robustness and reliability. However, as training data, by design, are

[69] *Ibid.*, 6 ff.

[70] *Ibid.,* 3.

[71] Art 51(1)(b), annex XIII (c) AI Act.

[72] S. Rossello – R.D. Morales – L. Muñoz-González, *Data protection by design in AI*, cit., 7.

never available in FL, this comes with shortfalls in data governance as indirect methods to assess data quality and bias are necessary. While FL already introduces a privacy and energy efficiency trade-off, complying with arts. 15 and 17 AI Act monitoring requirements threatens to further increase energy consumption. Ultimately, FL provides a compelling path to reconcile technical realities of AI development with data protection. However, fully realising FL's potential requires further technical research, particularly on energy efficiency including the energy–privacy trade-off. Given its distinctive features, FL demands a flexible interpretation of the AI Act's essential requirements; absent this, the adoption of a dedicated framework may prove necessary.[73]

[73] H. Woisetschläger et al., *Federated Learning Priorities*, cit., 8 ff.

Abstract

This article analyses the tensions between the GDPR and the AI Act in the context of AI development, focusing on the technical realities of centralised training. It evaluates Federated Learning (FL) as a decentralised, privacy-preserving alternative that advances compliance with data protection principles, unlocks siloed data, and exemplifies "data protection by design". Yet at the same time, FL complicates compliance with the AI Act's obligations for high-risk AI systems, in particular data governance, bias mitigation, and robustness, as these were drafted with centralised training in mind. Further analysis details the energy–privacy trade-off inherent in FL. It concludes that while FL provides a credible pathway to trustworthy, human-centric AI development, its distinctive features demand further technical research and either a flexible interpretation of the AI Act's essential requirements or a dedicated regulatory framework.

Keywords

Federated Learning – GDPR compliance – AI Act – energy-privacy trade-off – privacy-enhancing technologies