联邦学习框架下的数据安全与利用合规路径*

孙绮雯
（清华大学　法学院，北京　100084）

摘　要：日趋严格的个人信息保护相关法律法规，在保护个人隐私的同时，增加了企业数据流通合规的难度和成本。在联邦学习框架中，数据不动模型动的隐私保护设计以技术促进法律的遵守，是打破数据孤岛壁垒、促进隐私保护前提下数据融合协作创新的可能解。将合法原则、数据最小化原则与目的限制原则嵌入到系统开发的技术中，联邦学习分布式协作框架以局部模型更新参数代替本地原始个人数据上传，实现数据本地训练存储，达到可用不可见的个人信息保护效果。由于潜在的网络安全攻击以及机器学习算法黑箱的固有缺陷，联邦学习仍然面临着质量原则、公正原则与透明原则的挑战。联邦学习不是规避合规义务的手段，而是减少个人信息合规风险的可行技术措施，使用时仍然存在需要履行的个人信息保护义务，数据权属与责任分配的确定需要综合考量各参与方角色和个人信息处理者类型。

关键词：联邦学习；个人信息保护；数据孤岛；网络安全攻击；协作共享
中图分类号：D922.174　　**文献标识码：**A　　**DOI：**10.19358/j.issn.2097-1788.2023.06.004
引用格式：孙绮雯. 联邦学习框架下的数据安全与利用合规路径［J］. 网络安全与数据治理，2023，42(6)：21-29.

Data security and utilization compliance path under the federated learning framework

Sun Qiwen
(School of Law, Tsinghua University, Beijing 100084, China)

Abstract: The increasingly stringent laws and regulations related to personal information protection have increased the difficulty and cost of compliance in data circulation of enterprises while protecting personal privacy. Under the framework of federated learning, the privacy protection design that does not transmit the original data but only transmits the model uses technology to promote legal compliance, which can be a possible solution for data fusion and collaborative innovation under the premise of breaking the barriers of data isolation and promoting privacy protection. The legal principles, data minimization principle and purpose limitation principle, are embedded into the technical process of the system development. The distributed collaborative framework of federated learning uploads the updated parameters of the local model instead of original personal data, realizing local training and storage of data, and achieving such a great personal information protection effect that data can be utilizable while at the same time invisible. Due to potential network security attacks and inherent defects of machine learning algorithms black box, federated learning still faces the challenges of the principles of quality, fairness, and transparency. Federated learning is not a way to evade compliance obligations, but a feasible technical measure to reduce compliance risks of personal information. There still exist personal information protection obligations to be fulfilled when using federated learning framework. The determination of data ownership and responsibility allocation requires comprehensively consideration of the roles of each participant and the types of personal information processors.

Key words: federated learning; personal information protection; isolated data island; network security attack; collaborate and share

* 基金项目：国家社会科学基金重大项目（18ZDA149）

0 引言

当前人工智能发展面临数据孤岛现象与数据融合需求的矛盾，联邦学习有助于破解数据协作创新与数据隐私保护的困境。作为基于设计隐私的分布式协作模型，联邦学习可以在保护个人信息的前提下，使得跨组织、跨设备、跨区域的不同特征维度数据合规共享、流通、融合。在联邦学习框架中还可以结合使用多种隐私计算技术，如多方安全计算、同态加密等，进一步加强对个人信息的保护，降低隐私泄露的安全风险。本文首先分析了联邦学习是基于设计隐私思想的分布式协作模型，然后对联邦学习框架在个人信息保护原则下的表现进行评价并提出建议，最后探讨了联邦学习如何促进数据合规并指出依然存在的合规风险。

1 基于设计隐私思想的分布式协作模型

1.1 数据孤岛现象与联邦学习框架

人工智能发展需要海量数据的收集与融合作为基础，但是有效数据往往难以获取或数据孤岛的形式存在：一是数据泄露等安全问题的频发以及隐私问题关注度的日益提高，让数据所有者共享数据的意愿越来越低；二是互联网巨头企业垄断大量数据[1]；三是世界范围内个人信息保护相关法律法规和监管日趋严格，数据的访问、处理、存储、流通有可能面临政策问责。

联邦学习是一种将训练数据分布在多个分别持有的设备上，并通过聚合本地计算的更新来学习共享模型的分布式机器学习方法[2]。各参与方协同训练模型而无需集中存储、共享原始数据，只在训练模型过程中将中间参数上传至中央服务器，以协调全局模型的更新，实现模型训练活动与直接访问处理、集中式数据存储的分离。中央服务器首先通过公开可用的中央训练数据进行预训练来初始化全局模型，并将该模型副本发送给参与训练的数据持有者。每个数据持有者用其个人数据本地训练从中央服务器接收到的模型，并将模型参数在脱敏后代替本地数据上传给中央服务器。中央服务器聚合协调不同的本地更新参数并全局更新模型。重复上述训练过程直到模型收敛或训练终止。

在联邦学习框架中，整个训练过程只上传模型参数而不上传数据，允许数据处理而不泄露数据本身，数据可用不可见，从而在保障数据隐私和安全的情况下实现数据共享。联邦学习框架中博弈论的应用，如联邦学习激励机制利益分配，能够激励高质量数据的拥有者加入联邦学习联盟[3]。联邦学习模型在传递信息的过程中始终将原始数据保留在用户终端，使得各参与方能够在不披露底层数据和数据加密形态的前提下共建优化机器学习模型，有助于践行个人信息保护相关法律法规，降低个人信息合规难度和成本。联邦学习允许从跨数据所有者分布的数据中构建集合模型，通过加密机制下的参数交换，实现在各方数据本地化存储训练下的多方数据资源利用，从而解决数据来源匮乏、数据量不足、规模与质量不完备等问题，破解数据协作与隐私合规的难题[4]。

当数据中包括敏感个人信息时，传统分布式机器学习的中心调度将会给用户数据带来极大的隐私泄露风险。与之不同，联邦学习不需要将所有的训练数据从多个设备发送转移到一个中央存储库进行机器学习，数据始终保持本地化存储与训练，系统性风险和成本得以有效降低。与其他保护隐私的机器学习技术相比，如仅依赖于数据加密的机器学习技术，联邦学习的优势在于允许以更低的计算成本训练具有大量训练参与者的模型[5]。

1.2 设计隐私：数据不动模型动

联邦学习确保原始数据不出域，从源头上降低数据泄露风险，达到“数据不动模型动，数据可用不可见”的隐私保护效果。联邦学习体系下，各参与者身份、地位相同，实现了去中心或弱中心化。从理论角度看，分布式协作、本地化训练的联邦学习建模效果与将整个数据集集中在一起建模的效果相差甚小[6]。表 1 对联邦学习与传统分布式机器学习、其他隐私计算技术在技术特征、安全性等方面进行对比总结。

联邦学习通过设计隐私实施数据保护，将主动隐私保护措施整合到处理个人数据的系统设计阶段。隐私工程学者 Hoepman 在《设计隐私策略》中提出面向过程和面向数据的八种设计隐私策略。面向数据的策略侧重于对数据本身进行隐私友好的处理，本质上更具技术性，分别为最小化、分离、抽象、隐藏[7]。联邦学习尽可能地最小化和分离个人数据的处理，因此减少了各参与方为机器学习模型训练收集、处理、传输和存储的个人数据量，实现中央协调器与参与方之间的分离，还通过其分布式协作机器学习框架支持跨设备、跨组织分散数据的安全高效训练，实现参与方与参与方之间的分离。

表 1　联邦学习与传统分布式机器学习、其他隐私计算技术对比

	联邦学习	传统分布式机器学习	其他隐私计算技术	
			多方安全计算	可信执行环境
定义	将训练数据分布在移动设备上，并通过聚合本地计算的模型更新来学习共享模型	将所有的训练数据从多个设备发送转移到一个中央存储库进行机器学习	多方能够在不向彼此透露各自输入的情况下共同计算目标函数	通过软硬件方法在中央处理器中构建安全的隔离执行环境
技术特征	基于人工智能与隐私保护融合	机器学习	基于密码学设计特殊的加密算法和协议	基于可信硬件隔离可信机密空间
与联邦学习对比	—	• 系统性隐私泄露风险高 • 训练具有大量训练参与者的模型时计算成本高 • 模型精度并无显著差异 • 难以实现数据收集、训练等方面的最小必要	• 计算和通信开销大 • 性能相对较低	• 开发和部署难度大 • 需要信任硬件厂商
安全性	较高，但存在部分安全风险，可综合运用其他隐私计算技术提高安全性	低	高，其安全性有严格密码理论证明，但目前局限于半诚实模型	中，需要考虑安全设计，以及硬件漏洞、侧信道等安全问题

抽象策略要求尽可能地限制处理个人数据的细节，降低个人数据及其处理的详细程度。联邦学习中各参与方上传的数据为模型训练后更新的梯度等参数，与原始详细个人数据相比更粗粒度，能够有效降低信息风险级别、数据泄露风险以及合规难度。更新参数在没有充分保护的情况下容易受到隐私攻击，在联邦学习框架中应用安全聚合协议和差分隐私等技术对于保护用户信息免受组织和黑客攻击泄露至关重要。安全聚合协议通过加密，使得服务器只能在将传输的参数添加到数百或数千个其他用户的结果并对其进行平均后才能访问。在聚合且并入全局模型之前，差分隐私随机扰动局部模型的参数，在将各方模型参数发送到中央服务器前添加随机噪声，从而模糊结果。差分隐私技术的使用在为中央服务器提供用于算法训练的足够精确结果的同时，隐藏了实际具体传输的更新参数[8]。在联邦学习框架下，差分隐私等相关隐私计算技术的应用，增强了个人数据处理的模糊性与抽象性，降低了个人信息泄露的风险。

联邦学习以各参与方局部模型梯度更新的传输，代替各方原始个人数据的传输，通过衡量联邦学习要素之间通信中的隐私损失，根据所需的隐私级别向共享参数注入适当的噪声，从而掩盖梯度，提升数据隐私保护效果[9]。模型加密方法下，全局模型的参数在分发给各参与方进行本地训练前被加密。局部模型接收到加密参数后，返回可能进一步受到扰动的加密局部梯度。全局模型聚合并解密所有局部梯度以更新通用模型。

以过程为导向的策略侧重于处理个人数据的过程，分别为通知、控制、实施、证明[7]。这四种设计隐私策略也为建立联邦学习技术使用的合规框架提供新思路。

2　联邦学习框架中的个人信息保护原则

2.1　合法、最小必要与目的限制

联邦学习助力数据合规，促进个人信息处理的合法原则实现，特别是在处理敏感个人信息或者个人信息跨境传输等场景时，联邦学习能减轻相关严格法律法规带来的合规挑战，减少需要合规的内容，降低个人信息合规成本，使个人信息处理更加符合相关法律法规的要求。例如在医疗保健领域，病症、病理报告、检测结果等病人隐私数据常常分散在多家医院、诊所等跨区域、不同类型的医疗机构。严格法律法规的出台，增加了医院在隐私法律、行政法规、道德等方面

面临的约束，使得数据共享聚合存在诸多风险。联邦学习可以在数据不出本地的前提下实现跨机构协作，多方合作建立的预测模型能够更准确地预测癌症、基因疾病等疑难病。例如，在2020年，中科院泛在计算系统研究中心、中国科学院大学、深圳鹏城实验室和微软亚洲研究院联合提出了FedHealth架构。该框架通过联邦学习收集不同组织拥有的数据，显著提高了可穿戴设备行为识别的准确率和精度，并通过迁移学习为医疗保健提供个性化服务：一个医院的部分疾病诊疗信息可以通过联邦迁移学习传输到另一家医院，以帮助其他疾病的诊断[10]。在促进个人信息处理合规的同时，联邦学习实现了个人信息保护、数据安全与技术创新的平衡，以合法、安全、高效的方式打破数据孤岛壁垒，为大数据时代个人信息处理的合规提供新思路。

数据最小化原则与目的限制原则关系紧密，要求在收集个人信息时，应当以处理目的的实现为基准，来确定收集个人信息的最小范围。目的限制原则要求处理者只能对个人信息实施符合初始目的的相应处理活动，不得从事与处理目的无关的个人信息处理[11]。《个人信息保护法》第6条规定，处理个人信息应当具有明确合理的目的，并应当与处理目的直接相关，采取对个人权益影响最小的方式；收集个人信息，应当限于实现处理目的的最小范围，不得过度收集。在传统的机器学习中，数据最小化难以实现，一方面是因为难以预先判断训练机器学习模型需要的数据量，另一方面数据越多模型训练效果越好。大多数机器学习算法严重依赖数据质量和数量，因此研究人员倾向于收集尽可能多的相关数据。联邦学习系统不需要收集和处理原始训练数据，中央服务器只需要从参与者处收集本地机器学习模型参数来更新全局模型。以模型参数的收集代替原始个人数据这一特性，让联邦学习系统减少数据收集量的同时增强收集必要性。联邦学习避免将原始训练数据传输到中央服务器，也减少了对此类数据不必要的复制和集中，从而降低了这些数据被重新用于最初收集目的以外的风险，间接地促进目的限制原则的遵守。

2.2 质量原则与自助化决策原则

在个人信息处理活动中，质量原则对个人信息处理者提出的要求是，个人信息处理者应当积极采取各种技术措施和组织措施来检查所处理的信息的质量，确保信息的准确和完整，从而最大限度地减少错误的风险，避免因个人信息不准确、不完整对个人权益造成不利影响[11]。在联邦学习框架下，参与者之间数据的异质性程度未知，信息准确程度难以判断。数据存储在本地，中央服务器无法访问原始个人数据，联邦学习框架很难确保数据标签的正确性。此外，当训练有数百万参与者时，不可能确保他们没有恶意。联邦学习可能受到如数据投毒与模型投毒等数据安全网络攻击，攻击者或者以非歧视或有针对性的方式颠覆整个学习过程，降低系统的整体性能或产生特定类型的错误[12]，或者向中央服务器发送恶意的模型更新以获得对训练过程的直接完全控制，并抵消诚实参与者提供的贡献。潜在的网络安全攻击加大了满足质量原则的难度。因此一方面可以事前进行风险评估，基于半诚实或恶意的参与方和中央服务器的假设，订立合作协议以约定各参与方在各环节的权利、责任和合规义务[6]，并形成全流程监督机制。另一方面通过技术手段，综合运用其他隐私计算技术，如运用多方安全计算、同态加密等保护梯度等参数交换，降低网络攻击导致数据泄露的风险。

《个人信息保护法》第58条要求，提供重要互联网平台服务、用户数量巨大、业务类型复杂的个人信息处理者应当遵循公开、公平、公正的原则；第24条要求，个人信息处理者利用个人信息进行自动化决策，应当保证决策的透明度和结果公平、公正。联邦学习中各参与方的平等地位可以避免海量数据持有者操纵模型，从而增强模型训练的公平公正[13]；但机器学习固有的算法黑盒偏见，仍然会使应用联邦学习的自动决策系统存在导致歧视和不公正的可能性。联邦学习系统应适当地对各个数据集进行加权并采用边缘计算，以降低有偏见或不安全数据带来的风险[14]。

除了第24条，《个人信息保护法》第7条提出了个人信息处理的公开透明原则，即处理个人信息应公开个人信息处理规则，明示处理的目的、方式和范围。联邦学习框架中存在机器学习共有的算法黑箱问题，机器学习开发人员通常依靠对训练数据集的分析来解释模型行为。但是防止中央服务器访问原始训练数据集的隐私设计以及无法检查用户端本地训练模型的安全聚合机制，使得联邦学习系统的开发人员无法访问完整的训练数据集。因此模型的可解释性降低，联邦学习系统面临着透明度、解释能力和偏差的挑战。因此应当引入算法影响评估、告知义务、外部算法问责，以积极的方式促进人工智能的可解释性，提高算法透明度以构建公共信任的基础[15]。表2对联邦学习框架在上述个人信息保护原则下的表现进行评价和建议。

表 2　个人信息保护原则下对联邦学习框架的评价和建议

	合规评价	应对建议
合法原则	• 减少需要合规的内容 • 降低个人信息合规成本 • 降低传输信息风险级别	联邦学习只能促进合规，并非规避合规义务的手段，需要注意依然存在应当履行的个人信息保护义务
最小必要原则	• 以模型参数的收集代替原始个人数据 • 减少数据收集量 • 增强必要性	
目的限制原则	• 无需将原始训练数据传输到中央服务器 • 减少数据不必要的复制和集中 • 降低数据被重新用于与最初收集目的不相符的目的的风险	
质量原则	• 参与者之间数据的异质性程度未知 • 信息准确程度难以判断 • 很难确保数据标签的正确性 • 恶意参与者与潜在的网络安全攻击	• 基于半诚实或恶意的参与方和中央服务器的假设，订立合作协议以约定各参与方在各环节的权利、责任和合规义务 • 综合运用其他隐私计算技术，如运用多方安全计算、同态加密等保护梯度等参数交换，降低网络攻击导致数据泄露的风险
公平公正原则	• 各参与方的平等地位可以避免海量数据持有者操纵模型 • 提升了具有小数据集参与方的地位和联合训练的动力 • 最大限度地减少有偏见的数据集的影响 • 机器学习固有的算法黑盒偏见，仍然会使应用联邦学习的自动决策系统存在导致歧视和不公正的可能性	联邦学习系统应适当地对各个数据集进行加权并采用边缘计算，以降低有偏见或不安全数据的风险
透明原则	• 存在机器学习共有的算法黑箱问题，训练过程、决策结果的透明度有限、可解释性低 • 联邦学习系统的开发人员无法访问完整的训练数据集	应当引入算法影响评估、告知义务、外部算法问责，以积极的方式促进人工智能的可解释性，提高算法透明度

3　联邦学习框架中的个人信息保护法律适用

3.1　使用联邦学习应当履行个人信息保护义务

联邦学习框架中数据流通的合规基础与合规风险如图 1 所示。

个人数据与个人信息的定义都强调了直接或间接的可识别性，匿名化处理后的信息不在个人信息相关法律法规的适用范围内，决定了处理数据的实体是否需要遵守条例施加的各种义务。但是一方面“合理可能”“可识别性”“关联性”“不可复原性”等概念模糊，关于个人数据、个人信息、匿名化的定义争议较多，在实践中存在很大的法律不确定性空间。GDPR 采用从数据中去除足够的元素，使得数据主体不再能够被识别。该方法的核心在于使其不能再被控制者或第三方通过使用“所有可能合理使用的手段”（考虑现有技术和实施成本等因素）来识别自然人。另一方面，个人数据的分类存在动态性，初始阶段是匿名数据，但因网络攻击导致数据直接泄露，或是可以相结合识别个人信息的其他数据的泄露等原因，导致该匿名数据转变为个人数据。通过数据链接重新识别的风险是不可预测的。

使用联邦学习训练机器学习模型并不能使联邦学习框架中所有的处理操作免受相关个人信息保护法律法规的规制。首先，考虑各参与方提供的原始训练数据符合个人信息条件的情况，对这些数据执行的某些操作，如数据预处理时数据规范化或数据对齐等自动化处理操作，将受到个人信息保护相关法律法规的规制。

其次，考虑各参与方和中央服务器间交换模型更新参数的操作是否在个人信息保护相关法律法规的规制范围内。在联邦学习中，更新参数的整个过程可能受到投毒攻击、模型提取攻击、模型反转攻击、成员和属性推理攻击等，因此可能导致个人信息的泄露。基于对联邦学习数据安全以及相关争议概念的考量，使用联邦学习应当履行个人信息保护义务。下面从梯度泄露、不可靠的参与方与中央服务器角度分别分析。

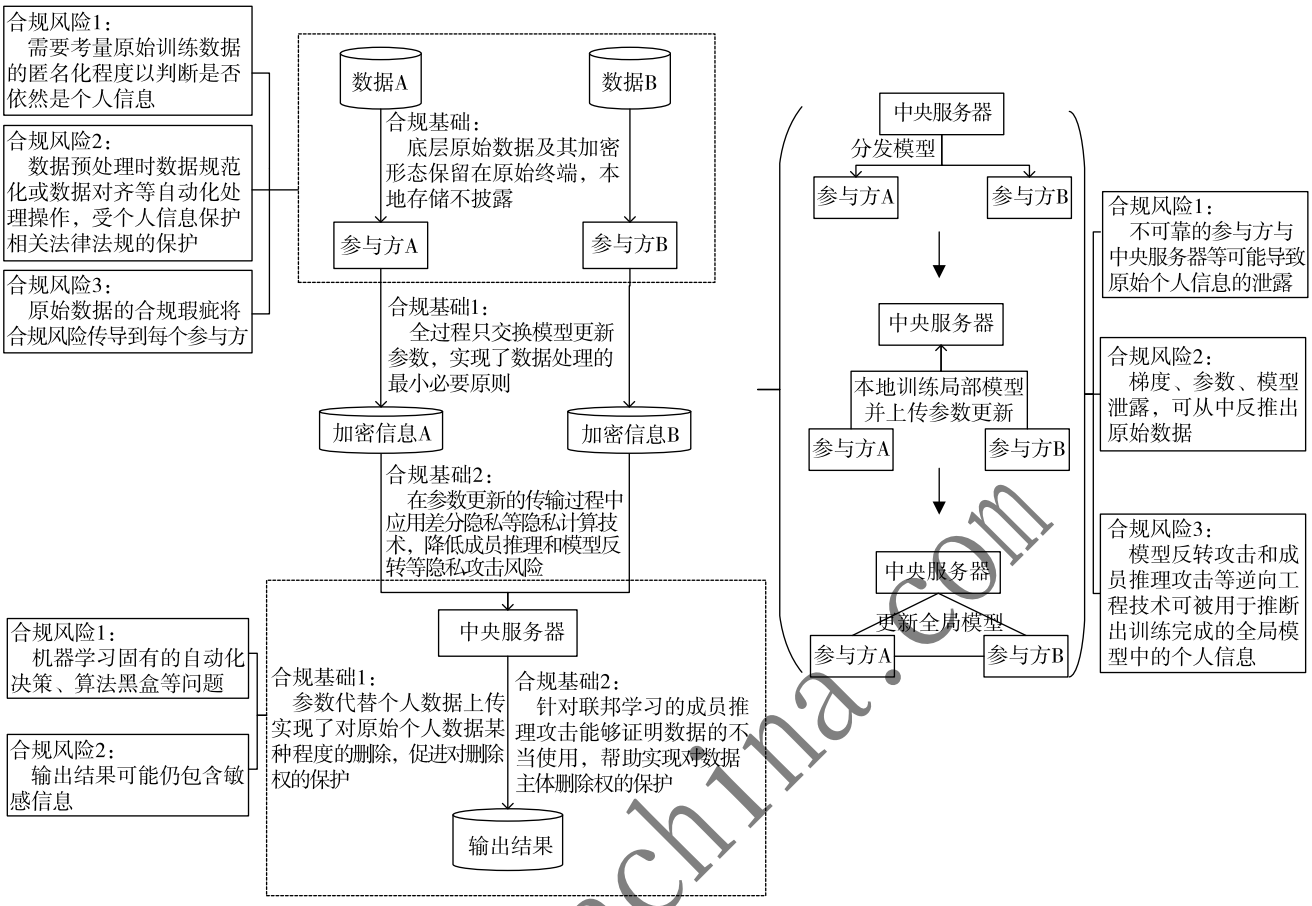


图 1　联邦学习框架下数据流通的合规基础与合规风险

传输训练参数避免了个人数据以其原始形式被直接发送到协调服务器，但是由于训练数据样本的某些特征固有地编码到本地机器学习模型中，此类模型参数仍然包含敏感信息，因此依然存在大量个人信息泄露的风险，与其他相关信息相链接、结合可能导致身份可识别。有研究证明，可以从公开共享的梯度中获取私有训练数据[16]，如在计算机视觉联邦学习应用程序中可以根据参数梯度的相关信息重建用户输入的高分辨率图像[17]。

不可靠的参与方会加剧个人信息泄露的风险。半诚实的参与方诚实地运行协议，但是会根据合法获取的中间参数额外推断出其他参与方的标签或数据。恶意参与方可能以拜占庭方式运行，发送训练模型更新以有针对性地塑造、恶意取代全局模型[18]，通过上传精心设计的有害信息诱导其他参与方暴露更多自身数据，或者不遵守隐私协议进而影响全局的隐私性[19]。此外，恶意参与方还可能执行协作学习环境中的推理攻击，观察更新后的全局模型中的更改以从共享参数中推断出来自其他方的敏感信息，不仅能够推断出成员资格，即其他参与者训练数据中确切数据点的存在，还能推断出训练数据子集及其属性[20]。

如果中央服务器是恶意的，它会通过推理攻击识别更新参数的来源，并通过参与者反复反馈的参数进一步非法推断目标参与者的数据集信息，从而导致参与者的隐私泄露。在某些协作深度学习协议中，一个诚实但好奇的中央服务器可以通过检查发送到服务器的本地训练的模型参数与原始数据样本之间的比例关系，利用简化设置从共享梯度更新中部分恢复参与者的数据点[21]。恶意中央服务器可以通过分析从相关本地节点获得的周期性参数更新，还可能故意要求受害者训练具有对抗性影响的修改模型[22]。

机器学习依赖对数据样本的规律挖掘以提高准确性，可能使得训练完成的全局模型参数及结构“记住”训练样本的细节。模型携带训练数据集的敏感信息，这些敏感信息可以通过对模型的黑盒访问来提取，因而存在较高敏感个人信息泄露风险[23]。攻击者可以通过反复查询模型的预测接口，来推测某条记录是否存在于训练集、推测模型的具体参数，而根据模型发布的参数能够进一步推测训练集成员或训练集具体样本[19]。

3.2　数据保护中的角色、责任和权利

在《民法典》和《个人信息保护法》中，有个人信

息处理者、共同处理者和受托处理者等多种个人信息处理者类型。在联邦学习框架中，提供原始训练数据的参与者众多，辨别和分配每个参与者遵守个人信息保护相关法律法规的责任仍然有较大难度。如果无法明确，可能会导致缺乏透明、公平的处理，违反相关个人信息保护原则。因此探讨联邦学习框架下每种类型参与者在数据保护中的角色和义务并进行责任分配具有重要意义。

实施联邦学习系统的服务提供商是共同处理者还是受托处理者，需要根据具体情况分类讨论。如果仅提供第三方托管服务，完全无自决性和主动性，无法决定处理目的和处理方式，此时服务提供商为受托处理者，需要根据各参与方需求协助参数上传及下发，并保障其安全性。联邦学习客户端应用程序应为客户端提供多个选项，允许客户端完全控制本地训练以及与中央服务器交换机器学习模型更新，并且只在用户端准备好时进行参数传输，而非中央服务器直接访问和获取原始训练数据以及各参与方本地训练的机器学习模型。原始数据并无合规瑕疵的情况下，若中央服务器为恶意的，其将通过推理攻击识别更新参数的来源，推断目标参与者的个人信息，甚至故意要求其进行对抗性模型训练。此时恶意中央服务器违反协助义务以及保障个人信息安全的义务，实质上自主决定了新的处理目的和处理方式，应当转化为个人信息处理者，承担个人信息处理者的所有义务和责任，并对其过错承担侵权责任；若与其他恶意参与方共同实施网络攻击，应当与委托人承担共同侵权产生的连带责任。若原始数据有瑕疵，由于联邦学习以更新参数代替原始个人数据上传的特性，服务提供商难以尽到对原始个人数据的合理审查义务。中央服务器应当实施适当的技术措施和安全措施，以证明训练客户端节点、通信和模型聚合等数据处理活动已按照规定进行。

更多情况下，提供中央服务器的联邦学习系统服务提供商扮演着个人信息的共同处理者，决定收集哪些信息（局部模型的更新参数）、如何使用个人信息（用于更新全局模型并下发）以及与谁共享个人信息（各参与方）等，而并非仅仅根据各参与方指示的目的、方式处理个人信息。欧盟法院在 FashionID 等案件裁决中对联合控制概念的广泛解释，可能导致每一个使处理个人数据成为可能的行为者都有资格成为共同控制人。共同处理者对个人信息处理活动是否发生以及如何发生具有决定性影响，即多个主体共同作出同一决定或多个主体作出合作决策。联邦学习系统的服务提供商指示各参与方使用其本地训练数据训练机器学习模型并共享参数更新，处理收到的各参与方的本地模型更新参数，在聚合和更新全局模型后发送给所有参与方并要求更新。服务提供商与各参与方各自决策互补，对于个人信息的处理目的和方式的确定具有不可或缺的实际影响，应当共同承担连带责任。

在联邦学习框架中，个人参与方享有知情权、查询权、更正权、删除权、限制处理权、可携带权、自动化决策解释权等相关个人信息保护权利。由于联邦学习算法模型内部结构复杂、运行过程自主性较强且人工无法干预等因素，联邦学习也存在算法黑箱问题，在数据输入、模型训练、结果输出等方面[24]存在人类无法完全理解人工智能的决策过程、也无法预测人工智能的决策或输出等问题。为保障数据主体的知情权，应当以用户为导向，达到一定程度的可理解性，实现社会的普遍接受。一方面使用户知情并理解对其数据的使用方式，数据控制器应当简明、清晰地对联邦学习训练中个人数据将如何处理、处理什么内容、处理本地机器学习模型以构建全局模型的预期训练目标、数据保留期等内容进行一般性解释。为增强可审计和可追溯能力，应当在算法生命周期中撰写算法关键决策环节日志文档。另一方面，通过交互界面使用户能够调整算法相关参数，以理解其对自己的可能影响。

针对训练过程以及决策的透明度、结果的公平公正性难以保障，影响数据主体的合法权利的问题，可以采用模拟安全攻击、查阅记录文档等方式，查看是否已经采取了充分、必要、有效的个人信息保护手段和措施，例如同态加密、传输通道加密等措施，以保障关键场景中的个人信息安全，参与方也可以在评估风险的基础上通过切片化、标签化及脱密处理等方式控制输入模型的梯度和参数信息，增强对数据流通涉及信息的控制力[6]。

关于删除权，联邦学习从框架设计上一定程度上可以缓解删除权带来的问题，因为联邦学习框架收集的是本地模型参数更新而非原始个人数据，原始个人数据始终在本地而非中央服务器存储处理。参数代替个人数据上传这一联邦学习框架的特性实现了对原始个人数据某种程度的删除。在某些情境下，针对联邦学习的成员推理攻击能够证明数据的不当使用，帮助实现对删除权的保护[20]。给定联邦学习模型和一个确切的数据点，采用针对聚合统计的成员推理攻击[25]，或者针对机器学习模型的黑盒成员推理[26]，可以推断出该数据点是否用于训练模型，从而可以检验数据控制者履行删除义务的效果。但是服务提供商履行删除义务后，若通过全局模型参数以及其他未行使删除权的用户的局部参数，反向还原出被删除参数，此时联邦学习框架下的删除权可能失去实际价值。

4 结论

在联邦学习框架中，自有数据不出本地，数据不动模型动，各参与方与中央服务器间只传输模型更新参数，进行分布式联合建模训练。该框架不仅解决数据孤岛问题，促进数据融合技术创新，提升机器学习模型训练效果，还以技术促进数据安全合规，降低个人信息合规成本和难度，实现对个人信息更大范围、更深层次的保护。面对训练过程中潜在的如投毒攻击、模型提取攻击、模型反转攻击、推理攻击等网络安全攻击，联邦学习框架可以融入其他隐私保护技术，如安全聚合协议、同态加密、安全多方计算等。但在应用过程中，这些隐私保护技术也可能带来新的个人信息保护义务。联邦学习框架中机器学习固有的算法黑箱，其难以解释性、难以预测性等问题带来对质量原则、自动化决策原则的挑战，并对数据处理者提出相应的数据保护责任要求。如何在技术上增强联邦学习框架安全系数以降低网络攻击风险，如何在立法上禁止对输出结果、过程参数的反向识别，如何建立可解释的算法黑箱制度以提高联邦学习框架透明度、可靠性、安全性与合规性，需要进一步研究。

参考文献

[1] 杨强. AI 与数据隐私保护：联邦学习的破解之道 [J]. 信息安全研究，2019，5 (11)：961 - 965.

[2] MCMAHAN H B，MOORE E，RAMAGE D，et al. Communication-efficient learning of deep networks from decentralized data [C]// International Conference on Artificial Intelligence and Statistics，2017.

[3] KANG L N，Chen Zichen，Liu Zelei，et al. A multi-player game for studying federated learning incentive schemes [C] //Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI'20)，2021：5279 - 5281.

[4] 王健宗. 数据隐私保护新曙光——联邦学习的机遇，挑战与未来 [J]. 联数，2019：1 (8).

[5] ROSSELLO S，MORALES R D，MUÑOZ-GONZÁLEZ L. Data protection by design in AI? the case of federated learning [J]. Computerrecht：Tijdschrift Voor Informatica En Recht，2021：1 - 11.

[6] 闫树，袁博，吕艾临，等. 隐私计算：推进数据“可用不可见”的关键技术 [M]. 北京：电子工业出版社，2022.

[7] HOEPMAN J H. Privacy design strategies (the little blue book) [M/OL]. [2023 - 05 - 10]. http：//www. cs. ru. nl/ ~ jhh/publications/pds-booklet. pdf.

[8] LAKE J. Federated learning：is it really better for your privacy and security? [EB/OL]. (2019 - 10 - 25) [2023 - 05 - 10]. https：//www. comparitech. com/blog/information-security/federated-learning/.

[9] CHALAMALA S R，KUMMARI N K，SINGH A K，et al. Federated learning to comply with data protection regulations [J]. CSI Transactions on ICT，2022，10：47 - 60.

[10] Chen Yiqiang，Wang Jindong，Yu Chaohui，et al. FedHealth：a federated transfer learning framework for wearable healthcare [J]. IEEE Intelligent Systems，2020，35 (4)：83 - 93.

[11] 程啸. 论我国个人信息保护法的基本原则 [J]. 国家检察官学院学报，2021，29 (5)：3 - 20.

[12] MUÑOZ-GONZÁLEZ L，LUPU E C. The security of machine learning systems [J]. AI in Cybersecurity. Intelligent Systems Reference Library. Springer，Cham，2019，151.

[13] LI L，FAN Y X，TSE M，et al. A review of applications in federated learning [J]. Computers & Industrial Engineering，2020，149. DOI：10. 1016/j. cie. 2020. 106854.

[14] NEWTON W. Can federated learning unlock AI in clinical trials without breaching privacy? [EB/OL]. (2022 - 10 - 14) [2023 - 05 - 10]. https：//www. clinicaltrialsarena. com/features/federated-learning/.

[15] 刘云. 论可解释的人工智能之制度构建 [J]. 江汉论坛，2020 (12)：113 - 119.

[16] Zhu Ligeng，Liu Zhijian，Han Song. Deep leakage from gradients [C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems，2019：14774 - 14784.

[17] GEIPING J，BAUERMEISTER H，DRÖGE H，et al. Inverting gradients-how easy is it to break privacy in federated learning? [C]// Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS' 20)，2020：16937 - 16947.

[18] TRUONG N，SUN K，WANG S Y，et al. Privacy preservation in federated learning：an insightful survey from the GDPR perspective [J]. Computers & Security，2021，110. DOI：10. 1016/j. cose. 2021. 102402.

[19] 刘艺璇，陈红，刘宇涵，等. 联邦学习中的隐私保护技术 [J]. 软件学报，2022，33 (3)：1057 - 1092.

[20] MELIS L，SONG C，DE CRISTOFARO E，et al. Inference attacks against collaborative learning[EB/OL]. (2018 - 05 - × ×) [2023 - 05 - 10]. https：//www. researchgate. net/publication/325074745_ inference_ attacks_ against_ collaborative_ learning.

[21] PHONG L T，AONO Y，HAYASHI T，et al. Privacy-preserving deep learning：revisited and enhanced [C]//International Conference on Applications and Techniques in Information Security，2017.

[22] WANG Z，SONG M，ZHANG Z，et al. Beyond inferring class representatives：user-level privacy leakage from federated learning [C]//IEEE Conference on Computer Communications. IEEE，2019：2512 - 2520.

[23] SONG C Z, RISTENPART T, SHMATIKOV V. Machine learning models that remember too much [C]//Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, 2017.
[24] 叶晓亮，王丽颖，李晓婷. 人工智能算法“黑盒”安全风险及治理探讨 [J]. 互联网天地，2022（1）：32－35.
[25] PYRGELIS A, TRONCOSO C. DE CRISTOFARO E. Knock knock, who's there? membership inference on aggregate location data [EB/OL]. (2017－11－29) [2023－05－10]. https://arxiv.org/abs/1708.06145.
[26] SHOKRI R, STRONATI M, SONG C, et al. Membership inference attacks against machine learning models [C]//IEEE Symposium on Security and Privacy, 2017：3－18.

（收稿日期：2023－04－05）

作者简介：

孙绮雯（2000－），女，硕士研究生，主要研究方向：计算法学。

（上接第8页）

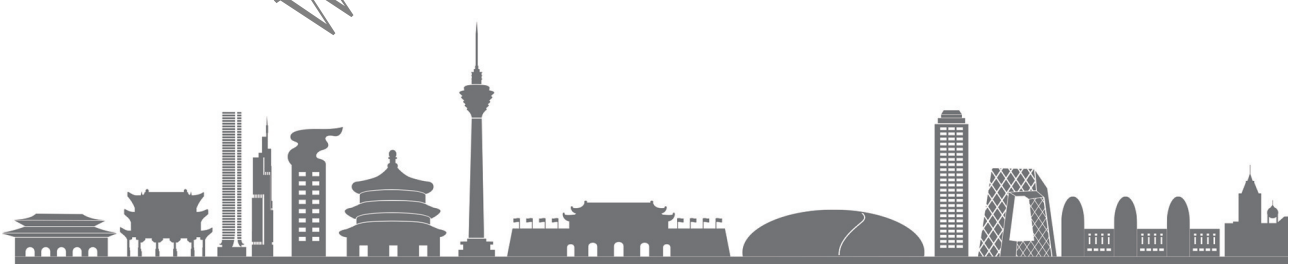
[21] 孙瑜晨. 反垄断中价格协同行为的认定及其规制逻辑 [J]. 北京理工大学学报（社会科学版），2017，19（5）：128－136.
[22] 许光耀.“经济学证据”与协同行为的考察因素 [J]. 竞争政策研究，2015（1）：91－96.
[23] 周秉瀛. 价格跟随行为的反垄断规制 [J]. 中国物价，2016（8）：25－27，43.
[24] 刺森. 算法共谋中经营者责任的认定：基于意思联络的解读与分析 [J]. 现代财经（天津财经大学学报），2022，42（3）：101－113.
[25] 周围. 算法共谋的反垄断法规制 [J]. 法学，2020（1）：40－59.
[26] 柴始青. 算法合谋反垄断规制路径探索 [J]. 社会科学战线，2022（1）：43－48，101.
[27] 唐要家，尹钰峰. 算法合谋的反垄断规制及工具创新研究 [J]. 产经评论，2020（3）：5－16.
[28] 李振利，李毅. 论算法共谋的反垄断法规制路径 [J]. 学术交流，2018（7）：73－82.
[29] 李丹. 算法共谋：边界的确定及其反垄断法规制 [J]. 广东财经大学学报，2020，35（2）：103－112.
[30] 刘学. 人工智能时代算法合谋反垄断的适法困境与逻辑依归 [J]. 西北民族大学学报（哲学社会科学版），2022（1）：87－94.
[31] 刘佳. 人工智能算法共谋的反垄断法规制 [J]. 河南大学学报（社会科学版），2020，60（4）：80－87.

（收稿日期：2023－03－29）

作者简介：

王聚兴（1998－），男，硕士研究生，主要研究方向：民法、经济法。

李晗（1999－），女，硕士研究生，主要研究方向：民法、经济法。



版权声明

凡《网络安全与数据治理》录用的文章，如作者没有关于汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权等版权的特殊声明，即视作该文章署名作者同意将该文章的汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权授予本刊，本刊有权授权本刊合作数据库、合作媒体等合作伙伴使用。同时，本刊支付的稿酬已包含上述使用的费用，特此声明。

《网络安全与数据治理》编辑部