中外法学 Peking University Law Journal
Vol. 37, No. 1(2025)pp. 65－84

人工智能时代联邦学习隐私保护的局限及克服

刘泽刚 *

摘 要 人工智能立法通常会对特定技术有所偏重。联邦学习属于主流的机器学习技术，最大的优势就在于其架构设计充分考虑了隐私需求。联邦学习在金融、数据公开等领域的应用已经比较广泛，并对自然人权益产生了重大影响。目前以隐私保护为目标的联邦学习不断暴露各种隐私揭示了个人数据隐私保护路径的法律缺陷：规范稀疏导致联邦学习缺乏明确隐私需求，“隐私设计”优势很难得到发挥；分布式架构导致联邦学习隐私保护责任难以落实；过度强调保密性和安全性，导致隐私保护的人格性被弱化和转化；技术权衡缺乏规范导致隐私保护缺乏透明性和确定性。这些问题揭示了人工智能隐私保护与个人数据保护在保护对象、保护流程、保护责任、保护框架等方面存在的巨大鸿沟。为了适应人工智能隐私保护的特殊要求，未来可在整合规范依据、调整规范重点、探索归责机制、构建沟通机制等方面对人工智能隐私保护规范进行升级和完善。

关 键 词 人工智能立法 联邦学习 隐私设计 差分隐私 隐私计算

人工智能训练、部署、使用过程存在许多隐私风险。但人工智能隐私保护目前没有一个整全性的方案。“人工智能”这个术语涵盖了许多路线迥异的技术方案。在科技界，包括杨立昆(Yann LeCun)和杰弗里·辛顿(Geoffrey Hinton)在内的科学家关于人工智能技术今后的理论发展方向和工程实践方案仍然存在较大的争议。在2010年之前，基于人工神经网络架构的深度学习曾经被广泛认定为一种没有前途的技术路径。但十多年来人工智能技术和产业的迅猛发展颠覆了之前的流行认知。欧盟的人工智能立法文本主要体现了当下的技术认知和实践的方向。比如欧盟《人工智能法》(Artificial Intelligence Act)专门强调了其与《通用数据保护条例》(General Data Protection Regulation，以下简称GDPR)的关系，并以机器学习这种比较

* 西南政法大学行政法学院副教授。

依赖数据驱动的技术路线为主要规范对象。欧盟人工智能立法的曲折过程及其最终版本对基础模型的专门强调也说明目前的人工智能立法并不完全是中立的,而是对特定技术有所偏重的。联邦学习(federated learning)是一种在隐私保护的前提下充分利用多个机构或主体的数据进行联合建模的机器学习基础技术。这种隐私保护机器学习框架既能稳定运行包括神经网络在内的各种主流算法又能兼容大模型技术,属于当前主流的机器学习技术。作为一种"隐私设计"(privacy by design),联邦学习最大的优势就在于其架构设计充分考虑了隐私需求,在所有任务中都能够保证"数据不出库"。〔1〕"数据不动模型动、数据可用不可见"的特征使得联邦学习在充分利用参与方数据协同训练模型的同时,还能很好地保护用户的隐私和数据安全。〔2〕这种依据个人数据(信息)保护法思路而设计的架构并不能一劳永逸地避免隐私风险。科技界已经指出了联邦学习的隐私设计仍有不足。2024 年 2 月,斯坦福大学人本人工智能研究院在题名为《在人工智能时代重新思考隐私》的白皮书中指出:个人数据保护权无法有效消除人工智能大量收集数据造成的隐私风险;现有和拟议中的隐私立法不足以解决人工智能的隐私问题。〔3〕目前联邦学习在金融、数据公开等领域的应用已经比较广泛,参与者多为机构和企业,不易为大众感知和认识。然而法学研究应该有超出普通和流行认知水平的高度,深入分析联邦学习隐私保护在法律层面的不足,将技术层面的固有缺陷呈现出来,也有助于揭示人工智能隐私保护的特殊性,并对人工智能隐私规范的发展方向提供启发。

本文对"隐私保护""个人信息保护"等术语的使用采取制度关联和现实发展的立场。目前法律领域广泛使用的"个人信息保护""隐私权""隐私保护"等概念,并没有哪个更加基础以至于可以成为其他概念的基础或者具有取代其他概念功能的必然性。相反,隐私权和个人信息保护在极其复杂的现实背景下以高度关联的方式共同发展。从现实发展的角度看,业界更多使用的是"隐私保护"(privacy preserving)这个具有高度包容性和弹性的概念。联邦学习提出者们在论文中用"隐私保护"这个术语来回应以欧盟 GDPR 为代表的个人数据保护的法律要求。但《中华人民共和国个人信息保护法》(以下简称《个人信息保护法》)等法律规范提出的隐私权保护等要求并不会在新兴的技术和产品上自动实现,而是需要包括法学界在内的相关主体更加有针对性的审视、反思与推动。本文从法律的规范立场审视现有联邦学习隐私保护的实际效果和不足,并由此分析其制度原因并提出相关法律对策。本文采用技术与规范结合的立场,用"隐私"一词来概括与确保数据免受意外或故意披露以保护人格尊严相关的各种权益。

一、联邦学习隐私保护的法律局限性

联邦学习本来就是个人数据保护法有效实施的成果。谷歌公司的技术团队在 2016 年首

〔1〕 陈凯、杨强:《隐私计算》,电子工业出版社 2022 年版,第 8 页。

〔2〕 杨强、黄安埠、刘洋、陈天健:《联邦学习实战》,电子工业出版社 2021 年版,第 6 页。

〔3〕 See Stanford Institute for Human-Centered AI, "Rethinking Privacy in the AI Era: Policy Provocations for a Data-Centric World," 2024, pp. 19-25, https://hai.stanford.edu/sites/default/files/2024-02/White-Paper-Rethinking-Privacy-AI-Era.pdf, last visited on 1 December 2024.

次提出联邦学习概念时声称其是一种充分考虑了隐私保护的、适合现代移动设备训练人工智能的分布式深度学习框架。移动设备上的人工智能训练数据通常是隐私敏感(privacy sensitive)或数量巨大的。这导致登录到数据中心进行训练的传统方式不再适合。作为一种替代方案,联邦学习将训练数据分布在移动设备上,通过聚集本地计算的更新来学习共享模型,从而解决安卓手机用户键盘输入法的本地更新优化问题。〔4〕这种基于设备的联合建模仅是联邦学习的一种形态。联邦学习的形态由跨设备联合训练逐步拓展至跨机构联合建模。广义的联邦学习强调各参与方的原始数据存储在本地,不进行交换或传输,而是使用即时聚合更新的方式来达到模型学习的目的。〔5〕联邦学习的愿景是在满足隐私保护需求的前提下,充分利用更多参与方的数据进行人工智能项目的开发和部署。法学界目前对 ChatGPT 这类预训练大型语言模型(Large Language Model,简称 LLM)的兴趣远高于联邦学习。实际上,联邦学习亦可有效促进 LLM 的稳步发展。有研究指出高质量语言数据存量可能将在 2026 年耗尽,低质量语言数据和图像数据的存量也将在未来 20 年中逐步耗尽。如果数据使用效率没有显著提高或找到新的数据源,机器学习的发展趋势可能放缓,LLM 的规模增长也会受到限制。〔6〕联邦学习为充分利用各种终端和机构的数据提供了合规架构,有助于突破数据瓶颈(data bottleneck),在隐私保护基础上进行 LLM 建构。〔7〕另外,由于端侧通信和算力限制,通用人工智能也可利用联邦学习这类分布式机器学习进行即时的端侧训练和模型更新。

在联邦学习提出的最初几年,各界对其隐私保护效果充满信心。由于欧盟隐私保护规范对谷歌这类跨国大公司来说利益攸关,因此 GDPR 等欧盟法规的制订和实施对联邦学习设计的影响是直接和明显的。有学者宣称:“联邦学习通过加密机制下的参数交换方式保护用户数据隐私,数据和模型本身不会进行传输,也不能反猜对方数据,因此在数据层面不存在泄露的可能,也不违反更严格的数据保护法案如 GDPR 等。”〔8〕但是,2019 年就有研究证明可以通过模型的输入输出以及中间梯度来反推参与模型训练的数据。〔9〕2020 年有研究展现梯度反

〔4〕 See H. Brendan McMahan, Eider Moore, Daniel Ramage and Blaise Agüera y Arcas, “Federated Learning of Deep Networks Using Model Averaging,” https://arxiv.org/pdf/1602.05629v1, last visited on 1 December 2024.

〔5〕 参见王力、张秉晟、陈超超:《隐私保护机器学习》,电子工业出版社 2021 年版,第 166 页。

〔6〕 See Pablo Villalobos, Jaime Sevilla, Lennart Heim, Tamay Besiroglu, Marius Hobbhahn and Anson Ho, “Will We Run Out of Data? An Analysis of the Limits of Scaling Datasets in Machine Learning,” https://arxiv.org/pdf/2211.04325v1, last visited on 1 December 2024.

〔7〕 2023 年 4 月,联邦学习开源平台 FATE 发布联邦大模型 FATE-LLM 功能模块。FATE-LLM 在参与方数据不出域的前提下,根据各方数据量进行算力投入,通过 FATE 内置的预训练模型进行横向联邦,利用各自数据进行联邦大模型微调,从而提升大模型的效果和稳健性。See Release v1.11.0, https://github.com/FederatedAI/FATE/releases/tag/v1.11.0, last visited on 1 December 2024.

〔8〕 杨强、刘洋、陈天健、童咏昕:“联邦学习”,《中国计算机学会通讯》2018 年第 11 期,第 53 页。

〔9〕 See Ligeng Zhu, Zhijian Liu and Song Han, “Deep Leakage from Gradients,” https://proceedings.neurips.cc/paper/2019/file/60a6c4002cc7b29142def8871531281a-Paper.pdf, pp. 14774-14784, last visited on 1 December 2024.

转(inverting gradients)攻击可以重构参与方训练数据。〔10〕很明显,联邦学习推出时的乐观判断无法成立。实际上,攻击者可以根据联邦学习系统的结构特征对其进行数据中毒、模型攻击、推理攻击、服务器漏洞等形式的攻击,其中一些还会导致严重的隐私风险。常见的针对联邦学习的隐私攻击包括:模型重建攻击、恶意服务器攻击、基于 GAN 的推理攻击、推断成员攻击等。〔11〕

目前,关于联邦学习隐私风险的研究仍处于方兴未艾的阶段。正如本文即将指出的那样,技术领域常常混用"隐私""数据"与"安全"概念。本文基于规范与技术结合的立场,从法学角度对与法律紧密相关的四个方面的联邦学习隐私风险进行梳理,对与法律关联不大的"隐私"技术问题并没有深入探讨。这种梳理很可能是不完整的,但也基本覆盖了最为迫切的隐私规范问题。总体而言,目前联邦学习隐私保护在法律规范层面的局限性主要都是以个人数据保护为核心的隐私保护路径造成的,主要表现在以下四个方面。

(一)隐私保护规范稀疏导致隐私需求匮乏

隐私需求(privacy requirement)是与隐私相关的系统要求。隐私需求的直接来源主要有法律(law)、法规(regulation)、标准(standard)、最佳实践(best practice)以及利益相关方的期待等。〔12〕隐私设计(privacy by design)的一般流程是首先确定隐私需求;接下来进行隐私风险评估,选择恰当的隐私控制方法;最后进行程序的开发和集成。隐私需求是隐私设计的出发点和成败的关键。隐私需求不清晰或不完备会导致隐私设计的功效大打折扣。目前联邦学习的隐私需求来源单一,主要由开发者对系统隐私性能进行主动探索和改进。这充分体现了"隐私设计"的主动(proactive)和预防(preventive)特征:设计者应主动预估系统潜在的弱点和可能发生的隐私威胁,然后选择恰当的技术和管理措施对相关风险进行预防。除了来自开发者和研究者的隐私需求外,目前联邦学习领域的其他隐私需求非常匮乏。具体来说,表现在如下几个方面。

首先,法律对联邦学习这类机器学习架构的隐私保护并无专门规定或特殊要求。目前针对人工智能和大数据领域的法律规范方式是软硬法结合、传统立法与各种标准、指南、最佳实践等规范共同发挥作用。这些不同性质的规范之间互相参照、紧密联系,共同构成人工智能广义的法律规范环境。由于全球人工智能立法仍处于探索阶段,现有法律中对人工智能隐私保护最具操作性的规范是个人数据保护法。以 GDPR 为例,联邦学习"数据可用不可见"的特征使其能轻松地符合其六大原则中的准确性、存储限制、完整性和保密性原则,对目的限制原则和数据最小化原则的遵从度也非常高,对合法公平透明原则的遵从程度则与其他机器学习系

〔10〕 See Jonas Geiping, Hartmut Bauermeister, Hannah Dröge and Michael Moeller, "Inverting Gradients-how Easy is it to Break Privacy in Federated Learning?" https://papers.nips.cc/paper/2020/file/c4ede56bbd98819ae6112b20ac6bf145-Paper.pdf, pp. 16937-16947, last visited on 1 December 2024.

〔11〕 See Rémi Gosselin, Loïc Vieu, Faiza Loukil and Alexandre Benoit, "Privacy and Security in Federated Learning: A Survey," *Applied Sciences*, Vol. 12, No. 19, 2022, p. 1.

〔12〕 参见(美)威廉·斯托林斯:《信息隐私工程与设计》,王伟平译,机械工业出版社 2021 年版,第 42 页。

统基本相同。[13] 欧盟人工智能立法没有对联邦学习这类架构设置专门规范。依据该法场景规范的路径，联邦学习适用的各种场景大多符合低风险的情形，即便涉及高风险人工智能应用，联邦学习也比其他人工智能架构更容易符合该法的各种隐私合规要求。联邦学习在现有规范框架下像一个高度自觉的隐私保护“优等生”，来自法律法规的强制性隐私需求是匮乏的。

其次，由于联邦学习正处于高速发展阶段，技术框架和细节仍有很大完善空间，相关共识有待逐步形成。即便有研究者或机构提出一些评价标准，也都是阶段性和探索性的。总体来看，目前关于联邦学习隐私保护的最佳实践和标准是稀少的。

最后，利益相关方的隐私期待也是匮乏的。在现有个人数据保护法框架下，数据主体对联邦学习的隐私期待是模糊的。从规范层面看，宪法、民法和其他部门法中的人格权和隐私权规范过于抽象，基本不具备特定场景的可操作性。从实际效果来看，隐私保护实际被数据保护替代。隐私权对应的人格权主体被降格为毫无能动性的受保护的数据主体。从事实层面看，由于联邦学习的架构和技术非常复杂，大部分普通数据主体根本无从了解其相关权利可能受损的场景和原理。联邦学习的训练和推理流程对原始数据的保护总体来说是优于其他架构的，因此数据主体的维权忧虑也较弱。权利主体隐私期待需求的匮乏导致在商业部署中，联邦学习的隐私保护目标很容易在与数据合规利用、模型效率或者网络安全的权衡中被牺牲。

（二）隐私保护法律责任模糊

联邦学习隐私保护法律责任的确定与追究的困难主要来自其松散的“联邦”关系和特殊的“联合”学习过程。具体表现在如下几个方面。

首先，联邦学习在主体关系方面的“松散性”导致归责困难。在松散的联邦结构中，各数据持有方之间是一种比较平等的关系。一旦出现追责问题，很难从规范层面简单地确定具体责任主体。以横向联邦学习为例，在客户端—服务器网络结构下，如谷歌 Gboard 之类的跨设备端联邦学习应用出现隐私泄露问题，比较容易确定部署应用的大公司的主要责任。但如果是跨机构的联邦学习，服务器则既可能是主导联邦学习系统建构的主体设置的，也可能是各个客户端主体共同信任的第三方提供的。且服务器控制者未必应对隐私保护负更多责任。对等网络结构联邦学习各参与方之间的关系更接近松散的“邦联”：各参与方无须借助第三方即可直接通信，安全性虽得到了进一步提高，但出现问题后更难确定责任归属。

其次，联邦学习联合学习建模的特征使其参与方的法律性质难以确定。各参与方不一定就是欧盟个人数据保护法上的“数据控制者”或者我国《个人信息保护法》上的“个人信息处理者”。GDPR 界定的数据控制者是指能够单独或与他人共同决定个人数据处理目的和方式的组织或个人，其在个人信息处理活动中发挥核心决策作用，并对该决策负责。很明显，个人数

〔13〕 See Nguyen Truong, Kai Sun, Siyao Wang, Florian Guitton and YiKe Guo, “Privacy Preservation in Federated Learning: An Insightful Survey from the GDPR Perspective,” *Computers & Security*, Vol. 110, 2021, pp. 8-16.

据控制者本身就是一个与责任紧密关联的概念。欧盟数据保护机构也认同:“控制者是一个功能概念,旨在根据事实影响分配责任。”〔14〕控制者必须确定应为哪些预期目的处理哪些数据。联邦学习的各参与方往往不能达到控制者的标准。这是因为联邦学习关于个人数据处理的主要目的是训练模型,而不是获取更多关于个体的信息。更重要的是在联邦学习架构下个人数据是可用不可见的,且数据可用性不是以流动性作为前提的,相反是在不流动的情况下发挥作用。不可见本身就符合个人数据保护框架下的隐私期待。如果不能将联邦学习参与方定性为个人数据控制者或个人信息处理者,从个人数据法角度对其进行责任分配和追究也会成为无根之木。

(三)隐私保护的人格权益被转换和弱化

法律隐私概念包含两个重点,即人格性(自主性、身份和尊严)和遮蔽性(免于侵入、限制观察)。联邦学习提出时考虑了个人数据保护法的要求,但落实在设计和工程实践上,则主要参考现有网络安全和信息安全规范的要求。联邦学习隐私保护其实经过了两次转换:首先在立法层面被转换为个人数据保护,然后又在更具体的技术标准层面被转换为信息安全防护。由此,联邦学习隐私保护过分强调遮蔽性(保密性),而忽视了人格性的维度。

在设计和工程实现上,隐私保护是信息系统安全防护框架下的一个分支领域。例如国际标准组织(ISO)在数据通信和网络安全领域制定的标准就首先是安全标准,并在安全标准基础上逐步开始推出隐私标准,其中最重要的是 ISO/IEC 29100 系列隐私框架标准。〔15〕这些隐私标准是基于安全标准和安全管理的框架建构起来的。大部分隐私标准的要求最终都通过贯彻安全标准得以实现。在人工智能安全标准化的研究组 SC42 推出的众多标准中,没有一项是专门的隐私保护标准。〔16〕这也从侧面说明目前人工智能标准制订领域中安全的级别优先且远高于隐私。

美国国家标准与技术委员会(National Institute of Standards and Technology,以下简称 NIST)〔17〕编号为 SP 800-53 的《信息系统和组织的安全和隐私控制》(Security and Privacy Controls for Information Systems and Organizations)提供了详细全面的隐私控制事项,并具

〔14〕 Article 29 Data Protection Working Party, “Opinion 1/2010 On the Concepts of ‘Controller’ and ‘Processor’,” https://ec. europa. eu/justice/article-29/documentation/opinion-recommendation/files/2010/wp169_en. pdf, p. 9, last visited on 1 December 2024.

〔15〕 ISO 在数据和网络安全领域的标准大都是与国际电工委员会(IEC)共同制定,因此标准代号通常包含两个标准组织的缩写,如 ISO/IEC 27001,ISO/IEC 29100。ISO 中信息安全工作组 SC27 制订的关于人工智能隐私保护的专门标准 ISO/IEC WD 27091(Cybersecurity and Privacy—Artificial Intelligence—Privacy Protection)仍处于工作草案(Working Draft)的阶段,尚未正式推出。

〔16〕 See ISO/IEC JTC 1/SC 42, ISO-International Organization for Standardization, https://www.iso.org/committee/6794475.html, last visited on 1 December 2024.

〔17〕 NIST 直属美国商务部,前身为国家标准局(NBS),从事物理、生物和工程方面的基础和应用研究,以及测量技术和测试方法方面的研究,提供标准、标准参考数据及有关服务。NIST 的安全和隐私标准在全球具有引领作用。

有明确的组织和结构。其中控件分为 20 个族(families),绝大部分都是隐私与安全共同的控制项目。而且在 SP 800-53 第四版中还增加了个体参与(individual participation)和隐私授权(privacy authorization)两个专门的隐私控制项目。〔18〕但文件的制定者认为"没有信息安全的基本基础,组织就不可能拥有有效的隐私"。〔19〕于是,在 2020 年第五版对控制项目进行了大量调整,包括用其他项目取代了个体参与和隐私授权两个专门的隐私控制项目,以及对第四版的一些控制措施元素进行分解细化,〔20〕从整体上提高了控制项目的集成性,加强了隐私和安全控制的一体化。〔21〕但隐私保护的需求进一步被安全防护吸收了。其他的相关标准包括 IEEE 标准委员会(SASB)于 2021 年推出的联邦学习标准(IEEE P3652.1)以及中国信息通信研究院牵头制定的《隐私计算联邦学习产品安全要求和测试方法》(YD/T 4691-2024)和《隐私计算联邦学习产品性能要求和测试方法》(YD/T 4692-2024)等,其关注重心在于产品的研发、评估、测试和验收,而非权益保护和法律监管。还有一些所谓国际标准本来就是国内企业(如微众银行)主导制订和推动通过的。由于这些标准不具有强制性且与本文的主题无密切关系,不在此多论。

更重要的是,单纯强调安全防护无法杜绝高危害的隐私风险。目前联邦学习安全和隐私防护大部分的研究都是基于诚实但好奇的模型假设。有学者基于这种假设系统研究了横向联邦学习的半诚实安全性。〔22〕半诚实的攻击者会在遵守联邦学习的密码安全协议的基础上,试图从协议执行过程中产生的中间结果推断或者提取出其他参与方的隐私数据。〔23〕由于数据法律法规等因素的约束,加之恶意行为会导致模型质量下降损害攻击方自身利益,联邦学习模型训练的参与方通常符合半诚实但好奇的假设,不会尝试极端的恶意攻击。可信执行环境等安全计算技术的引入,也可以在一定程度上限制此类攻击者的影响,使其很难从服务器返回的参数中推断出其他参与方的隐私信息。但从根本上说,联邦学习很难杜绝"拜占庭将军问

〔18〕 参见斯托林斯,见前注〔12〕,第 58 页。

〔19〕 See National Institute of Standards and Technology, "Withdrawn NIST Technical Series Publication (SP 800-53 Rev. 4)," p. J-1, last visited on 1 December 2024.

〔20〕 参见斯托林斯,见前注〔12〕,第 58 页。

〔21〕 SP 800-53 于 2020 年更新为 SP 800-53 Rev. 5 (09/23/2020),文件详情参见 https://doi.org/10.6028/NIST.SP.800-53r5, last visited on 1 December 2024.

〔22〕 See Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal and Karn Seth, "Practical Secure Aggregation for Privacy-Preserving Machine Learning," in *CCS '17: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, New York: ACM, 2017, pp. 1175-1191.

〔23〕 被动攻击是指通过监听网络通信来获取信息的攻击行为。这种攻击通常不会对目标系统造成明显损害,但可能会收集到大量敏感信息,造成隐私泄露风险。常见的被动攻击形态包括网络监听攻击、协议分析攻击、ARP 欺骗攻击、DNS 欺骗攻击、中间人攻击等。主动攻击破坏信息的真实性、完整性及系统服务的可用性,常见形态包括拒绝服务、篡改和伪造。

题”(Byzantine Generals Problem)的困扰。〔24〕一旦出现合作关系崩溃导致半诚实模型假设失效,尤其是中心服务器成为恶意攻击者时,完全可以通过模型推测出参与方相关数据。更严重的情况是各方相互默许通过技术手段获取其他参与方的数据。这种看似极端的假设实际上也具有利益上的现实性:联邦学习各参与方若非竞争关系,完全可能利用联邦学习的架构共谋突破法律限制进行数据交换,在获取可用人工智能模型的同时非法盗取数据。由于联邦学习架构复杂,对此类恶意串通行为进行外部监管的难度非常大。跨机构联邦学习中即便涉及个人数据,自然人也很难进行参与和干预,数据主体在浑然不觉中就已遭受隐私侵害。

(四)缺乏规范约束的技术权衡导致隐私保护缺乏确定性

联邦学习在技术层面具有显著的复杂性,主要体现在以下几个方面。首先,联邦学习的类型和算法具有复杂性。联邦学习通常可分为横向、纵向和迁移学习三类。横向联邦学习和纵向联邦学习是根据客户端数据的不同属性来进行分类的。联邦迁移学习则是联邦学习与迁移学习的结合。客户端之间的数据特征和分类标签差异较大,在进行训练时需要进行对齐工作。随着联邦学习框架的发展,越来越多的传统机器算法可在联邦学习上实现。这导致联邦学习在框架和算法类型上具有很高的复杂性。其次,联邦学习是一种分布式学习。分布式机器学习比集中式学习在架构上更复杂。而联邦学习则比一般的分布式机器学习面临更严峻的挑战:数据的多源异构(Multi-Source Heterogeneous Data);设备不稳定;通信成本高昂。传统机器学习的训练数据是独立同分布(independent and identically distributed,以下简称 IID)的。〔25〕在联邦学习中,数据分布在多个设备或服务器上,每个设备或服务器上的数据可能来自不同的用户群体或环境,导致数据分布不一致。这与 IID 假设相违背。因为 IID 假设要求所有数据样本都来自同一个分布。而在联邦学习中,不同设备上的数据可能存在相关性。例如,用户的手机和电脑上的数据可能有关联,但这些数据是在不同的环境中收集的,这两个设备上的数据可能具有不同的特征分布。根据不同地区或者不同的用户行为,这些数据之间必然产生相关性。而这种相关性违反了独立性的要求。因此 IID 数据不满足同分布的要求。在联邦学习中,为了保护用户隐私,通常需要采用差分隐私等技术。这些技术在非 IID 数据(Non-IID)上的应用更加复杂。高度的复杂性使得开发者必须在多种目标和技术参数中做出权衡(trade off)和选择。但目前这些权衡却没有实质性的法律约束。这导致有利于隐私保护

〔24〕 拜占庭将军问题是一个虚构模型,用于讨论分布式系统中在少数节点恶意伪造信息的情况下达成共识的问题。拜占庭容错(Byzantine Fault Tolerant)讨论的是容忍拜占庭错误的共识算法。拜占庭是东罗马帝国首都,由于地域宽广,守卫边境的多个将军需要通过信使来传递消息,达成某些一致决定。但将军中可能存在叛徒,试图发送虚假消息干扰共识达成。

〔25〕 独立同分布是指变量序列或者其他随机变量有相同的概率分布,并且互相独立。对机器学习而言,数据独立同分布意味着输入空间中的所有样本服从一个隐含未知的分布,训练数据所有样本都是独立地从这个分布上采样而得。数据驱动的人工智能的有效性假设训练样本集与预测样本独立同分布,而且二者都与真实总体样本服从同一分布。

的技术指标容易在权衡中被牺牲，从而造成联邦学习隐私保护在规范层面的不确定性增加。这种由权衡造成的不确定性在差分隐私上体现得尤为突出。

差分隐私的原理由辛西娅·德沃克（Cynthia Dwork）于2006年给出严格的数学证明。〔26〕从本质上来说，差分隐私是一种扰动技术（perturbation techniques），其思路是在原始数据上添加噪声，使从扰动数据上计算出来的统计信息与从原始信息上计算出来的信息难以区分。扰动技术简单高效，但易受概率性攻击，这导致一种两难境地：噪声添加过多，会严重影响学习的准确度和效率；添加太少，又达不到隐私保护的效果。典型的差分隐私又被称为ε-差分隐私。ε被称为差分隐私的隐私预算（privacy budget）。当隐私预算ε足够小时，隐私保护程度较高，但数据可用性较低，机器学习效果较差。提高隐私预算ε，情况则相反。隐私预算就是一种在效率和隐私保护之间的权衡参数。差分隐私的不同部署方式会造成差异巨大的效用、性能、隐私损益状况，必须根据各种需求和目标进行权衡。而联邦学习中需要权衡的环节远不止差分隐私。例如横向联邦学习隐私保护标准可以通过秘密共享（secret sharing）、秘钥协定（key agreement）、认证加密（authenticated encryption）或同态加密等方式实现，但需要投入较大的计算和通信开销，所以在实践中往往通过差分隐私实现弱化的横向联邦学习的半诚实隐私安全性要求。〔27〕这些缺乏规范的权衡增加了联邦学习隐私保护的不确定性。

二、人工智能隐私保护的特殊性

人工智能对包括隐私权在内的各种权利产生的影响有极大的不确定性。以GDPR为代表的个人数据保护法是针对计算机、互联网和大数据的技术和产业特征建构的。个人数据保护法较为有效地实现了大数据条件下的隐私保护。这容易导致一种错误认识：个人数据保护法完全能够胜任人工智能的隐私保护。然而，从联邦学习隐私保护存在的问题来看，个人数据保护法并不能充分适应人工智能隐私保护的挑战。综合来看，联邦学习隐私保护的法律缺陷揭示了人工智能隐私保护和个人数据保护之间的鸿沟，凸显了人工智能隐私保护的特征。

（一）保护对象鸿沟

从人工智能的本质和发展趋势来看，个人数据保护法无法捕捉最重要的人工智能隐私保护因素。人工智能的目标是接近人类整体的智能水平，而非个体的智能表现，因此关注的是训练数据集作为整体的统计学特征。机器学习的本质是从有限数据中自动学习规律，并利用规律对未知数据进行预测。从本质和目标上看，人工智能本来就更依赖公共数据而非个人数据，并无获取个人数据的现实利益需求。随着人工智能应用的推广，未来人工智能发展可能更多利用通过数据增强技术产生的数据，以及人类使用人工智能系统产生的数据。传统意义上的

〔26〕 See Cynthia Dwork, "Differential Privacy," https://mathpicture.fas.harvard.edu/files/mathpicture/files/2021-10-05_dwork_seminar.pdf, pp. 1-12, last visited on 1 December 2024.

〔27〕 参见陈凯等，见前注〔1〕，第118—122页。

个人数据对人工智能的价值是有限的。

联邦学习联合多源数据进行模型训练，其实质是通过打破数据孤岛，更加全面地学习人类行为的整体特征。虽然联邦学习模型训练时传输的权重和梯度等信息依然反映数据特征，但包括个人数据在内的各类训练数据确实没有离开数据控制者的本地数据库。在更广泛的意义上，当前人工智能主流技术路线的机器学习中参与训练的数据只是“炼金术”的原料，最终产品是模型而非新的个人数据。模型与个人数据间并无严格映射。即便攻击者通过技术手段获取了机器学习训练数据，被泄露的也是参与方层面的所谓“隐私数据”，即参与方有权存储和处理但无权或不愿分享的数据。这些数据在用于机器学习训练前都已进行预处理，以便机器从其中学习到有价值的特征。经过预处理后的数据集与传统意义上的个人数据存在很大差异。攻击者要想还原出能够识别出特定个人的数据，往往还需要其他数据和技术条件的支持。因此，从个人数据保护法合规的角度规范机器学习的隐私保护实际上缺乏适恰的“规范接口”。

代表人工智能发展趋势的各类大模型实质也是机器学习，其与个人数据的规范联系稀松模糊。例如，当前主流 LLM 都采用规模巨大的公开数据集进行训练。LLM 开发者虽然通常会公开项目的数据集构成和数据收集的标准，但对数据来源、数据集大小、数据标注数等技术指标却往往讳莫如深。〔28〕 尽管 Open AI 在训练时是否非法使用了个人数据遭到了质疑，但这些数据中绝大部分都是通过公开渠道获取的，与一般意义上的个人数据概念相去甚远。〔29〕作为与 Open AI 闭源项目并行的强大 LLM 开源项目，Meta 的 LLaMA 的训练也只使用公开的数据。〔30〕 当然，LLM 在具体领域的应用推理阶段涉及数据保护和隐私保护问题。如果有用户在使用过程中将涉及隐私的数据上传给模型，很可能造成隐私泄漏。但这并非开发者和企业的责任，而是用户自身疏漏造成的泄漏。从联邦学习对数据使用的本质特性到 ChatGPT 这类 LLM 对个人数据依赖性低的情况来看，以个人数据保护法为代表的数据法规并不完全

〔28〕 有研究者根据 OpenAI 论文公开的数据情况，推测了 GPT-3 预训练数据集大小一共有 753.4GB，训练数据均来自公开渠道，包括 11.4GB 维基百科数据；21GB 古腾堡图书（Gutenberg Book）语料库；101GB 互联网最大的电子书站点 Bibliotik Journey 的数据；50GB 流行社交媒体平台 Reddit 数据；570GB 来自公开爬取整理的 Common Crawl 数据集。See Alan D. Thompson，“What’s in My AI? A Comprehensive Analysis of Datasets Used to Train GPT-1，GPT-2，GPT-3，GPT-NeoX-20B，Megatron-11B，MT-NLG，and Gopher，” https://LifeArchitect.ai/whats-in-my-ai，last visited on 1 December 2024.

〔29〕 美国的人工智能和大数据产业监管比较宽松。即便如此，针对人工智能开发部署者的诉讼却时有发生。2023 年 6 月 OpenAI 和微软遭遇一场要求赔偿 30 亿美元的集体诉讼。16 名匿名原告声称 OpenAI 秘密地从互联网上收集了 3000 亿个单词以及数百万人消费者的数据，并挪用这些数据来开发不稳定未经测试的人工智能技术。See https://storage.courtlistener.com/recap/gov.uscourts.cand.414754/gov.uscourts.cand.414754.1.0.pdf，last visited on 1 December 2024.

〔30〕 预训练数据总计 1.4T 的 token，其中 Common Crawl 的数据占比 67%，C4 数据占比 15%，Github、Wikipedia、Books 三项数据均都各自占比 4.5%，ArXiv 占比 2.5%，StackExchange 占比 2%。See Hugo Touvron et al.，“LLaMA：Open and Efficient Foundation Language Models，” https://arxiv.org/abs/2302.13971，last visited on 1 December 2024.

适合人工智能的产业逻辑，通过个人数据法规范人工智能隐私保护的空间也在逐步收紧。

（二）保护重点鸿沟

在联邦学习从早期跨设备形态向跨机构形态扩展的过程中，“隐私保护”的意义发生明显转化：系统安全成为压倒性的保护重点，自然人的人格利益遭到忽视。从法理上讲，具有人格的自然人才谈得上隐私，但在联邦学习领域中，隐私保护的权益主体却被逐步偷换成了参与训练的各个机构主体。谷歌最初提出联邦学习方案时针对的是手机终端的输入法用户。这些用户都是自然人，隐私保护的权益主体也是自然人。但跨机构联邦学习首先保护的是参与机构的各种安全防护权益，这些权益往往也被泛称为隐私保护。由此造成了隐私保护重点的重大偏移，“隐私保护”被转换为保密和安全防护。

人工智能领域的安全防护和隐私保护存在明显差异。第一，保护目标不同。隐私保护的最终目标是人格权益，是为了保护自然人免受不必要的干扰而创建屏障和管理边界，从而促进人格自主发展和尊严保障。安全防护的目标是保障系统的机密性、完整性和可用性。第二，防范的行动者不同。安全防护主要防范未经授权的外部行动者，而隐私保护则对系统内外的行动者都要加以防范。第三，防御的攻击类型不同。人工智能的隐私保护与机密性关联更为紧密，机密性针对的攻击模式包括重构攻击、模型反演攻击、成员推理攻击。人工智能的安全防护则更多与完整性和可用性相关，针对的是投毒攻击、对抗攻击、询问攻击等攻击模式。第四，保护涉及的防御方法不同。隐私保护的防御方法包括安全多方计算、同态加密、差分隐私等。安全防护的防御方法包括防御精馏、对抗练习、正则化等。

在专门规范匮乏的情况下，人工智能隐私保护依然主要依赖信息安全和网络安全规范。人工智能专门隐私规范和标准也会对现有安全规范和标准形成路径依赖。尽管隐私与网络信息安全紧密相关，但毕竟存在本质差异。隐私保护经过个人数据保护和安全防护两个层次的“压缩”处理，逐步远离其人格权益保护重点，过于偏重“遮蔽性”或保密性目标。隐私保护的人格权益目标被弱化和偏移。

（三）保护责任鸿沟

以 GDPR 为代表的个人数据保护法创造了以“数据控制者”为核心的一系列主体概念，而这些主体概念的区分标准就是个人数据处理责任的分担。个人数据保护法采用了一套直观的数据流程描述框架，包括数据收集、存储、处理、传输、携带、删除等流程。但这种相对简单的责任划分方式已经不适应人工智能隐私保护的复杂场景。虽然当前机器学习是由数据驱动的，但数据并非人工智能隐私保护的主要矛盾。以联邦学习为例，由于数据并不发生法律意义上的流动或迁徙，参与方的责任也无法用数据处理流程进行划分和确定。在这种情况下继续沿用个人数据保护法的责任划分规范，很可能导致个人数据保护法成为人工智能项目逃避隐私保护义务的合规屏障。因此，从数据出发很难从技术和规范上厘清人工智能项目的隐私保护责任。与此相应，数据保密或安全责任也不能完全覆盖人工智能的隐私保护责任。这造成了个人数据保护责任和人工智能隐私保护责任的鸿沟。数据控制者、处理者等大数据场景下的责任主体概念已不太适合人工智能场景的责任划分。

联邦学习的形态丰富、技术复杂,而且已在金融、医疗等领域发挥重要作用,但大多数时候以一种底层架构的方式在底层默默发挥作用,不像 ChatGPT 和 Sora 这类大模型容易引发关注和焦虑。社会公众对联邦学习的运用以及其对各种重要权益的影响也关注不足。然而,从实际影响的广度和深度来看,联邦学习对普通人的意义并不比基础模型小,更需要防范风险和规范责任。但联邦学习松散的网络结构和联合训练模式导致从技术和规范上对参与方进行责任划分异常困难。为了在联邦学习这类模型中实现隐私保护,有必要结合技术特征进行相对具体的责任规范。

(四)保护框架鸿沟

机器学习从本质上说就是计算机将数据转变为知识和智能,转变方式就是"计算"。从现实角度来看,人工智能是基于计算机实现的计算以及模式判别。而模式判别其实也可归结为广义的计算问题。当然,没有数据驱动,机器学习就无法获得发展,但过分强调数据的重要性,不仅很难真正理解人工智能隐私保护的本质,甚至难以准确描述人工智能隐私保护的过程。但是,从数据出发而不从计算的角度思考,根本无法理解当前人工智能隐私保护的"权衡性"特征:只要不在乎计算成本和可用性,可以非常完美地保护隐私。如果不考虑计算因素,现有密码学技术中的同态加密(homophobic encryption)技术不仅准确度很高,还能够消除数据效用和数据隐私之间的权衡。〔31〕但同态加密的计算成本很高。有研究表明,在典型应用场景和参数设置下,最先进的全同态加密方案实现比多方安全计算的计算效率低几千倍。〔32〕如果不是基于降低计算成本和提高经济效率的考虑,完全可以使用密码学技术全面保护隐私。正是计算成本和通信成本的权衡导致隐私风险的增加。但这种权衡却是现实和必须的,否则一些人工智能项目只能停留于数学原理层面,很难在工程上得到实现。

既然计算才是数据转变为信息、知识和智能的关键,人工智能隐私保护也只有从计算视角才能得到和全面的把握。从计算角度看,人工智能的训练和推理(应用)具有非常不同的数据特性,对隐私保护的意义也存在很大不同。以联邦学习为例,其训练过程更加复杂,隐私风险更加突出。在客户端架构中,提供聚合功能的往往是可信计算平台。横向联邦学习的参与方都有完整的模型,因此可以在本地进行推理。在推理过程中,可以采取恰当技术进行隐私保护。在纵向联邦学习中,由于参与各方没有完整的模型,推理过程依然要通过可信第三方的计算与协调,隐私保护难度更大。

因此,只有以计算为中心,从系统整体出发,才能完整描述和全面规范人工智能隐私保护问题。这也是"隐私计算"(privacy preserving computation)技术兴起的根本原因。〔33〕隐私

〔31〕 参见彭南博、王虎等:《联邦学习技术及实战》,电子工业出版社 2021 年版,第 33 页。

〔32〕 See Ilaria Chillotti, Nicolas Gama, Mariya Georgieva and Malika Izabachène, "Faster Packed Homomorphic Operations and Efficient Circuit Bootstrapping for TFHE," in Takashi Takagi and Thierry Peyrin (eds.), *Advances in Cryptology – ASIACRYPT 2017*, Cham: Springer, 2017, pp. 351-368.

〔33〕 "Privacy Preserving Computation"直译为"隐私保护计算",但通常都将其简称为"隐私计算"。

计算并不是单一技术，而是在保护隐私的前提下进行数据计算和分析的技术的总称。多方安全计算、联邦学习、可信执行环境是现阶段主要的隐私计算技术方案。三类方案各具优势，在部分应用实践中可实现能力互补。从人工智能的角度看，联邦学习框架可以整合多方安全计算和可信执行环境的优势。隐私计算的概念充分体现了当前隐私保护在技术层面发展的趋势：从计算的角度更现实地描述和更全面地保护隐私。

三、人工智能隐私保护规范的完善方向

人工智能隐私保护与个人数据保护是互补而非互斥的关系。这也是欧盟《人工智能法》将个人数据权作为其数据规范基础的理据。〔34〕《人工智能法》说明部分第(3)项指出："如果本条例包含关于在处理个人数据方面保护个人的具体规则，涉及限制为执法目的使用人工智能系统进行远程生物特征识别、为执法目的使用人工智能系统对自然人进行风险评估以及为执法目的使用人工智能系统进行生物特征分类，则就这些具体规则而言，本条例宜以《欧盟运作条约》第16条为依据。"〔35〕《人工智能法》的出台仅仅是人工智能法律规范的开始，而远非结束。这一方面是因为立法除了规范和支持欧盟人工智能发展外，还有暗含的抵御美国人工智能行业在欧洲取得垄断地位的强大动机。另一方面人工智能的不确定性使得任何法律规制的计划都不可能是完备的。从实际效果来看，在人工智能领域领先的美国实际上采取了非常宽松的监管政策和法律框架。欧盟正在进行的人工智能立法尚未充分重视人工智能隐私保护与个人数据保护之间的一些重大差异。我国目前已启动的人工智能立法对欧美经验可以借鉴，却无需亦步亦趋。综合来看，若我国隐私保护规范做出适度调整，我国可以在人工智能隐私保护领域做出引领性的贡献。

(一)整合规范依据

从制度环境来看，当前全球科技立法重点已经从数据法和竞争法转向人工智能立法。人工智能成为隐私权保护的基本制度环境。以个人数据保护法为代表的的数据立法产生于互联网发展早期，以欧盟1995年《数据保护指令》为代表，还包括一系列与数据安全、数据使用有关的规范。随着互联网产业和技术的迅猛发展，数据法也逐步升级但始终处于一种跟随状态。

〔34〕 See Texts Adopted-Artificial Intelligence Act, https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.html and https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf, last visited on 1 December 2024. 此后文中引述和论及法案文本不再一一注明出处。

〔35〕《欧盟运行条约》(Treaty on the Functioning of the European Union，简称TFEU)第16条规定："1.每个人都有权保护自己的个人数据。2.欧洲议会和理事会按照普通立法程序行事，应制定关于在欧盟机构、团体、办事处和机关以及成员国开展属于欧盟法律范围的活动时处理个人数据方面保护个人的规则，以及关于此类数据自由流动的规则。对这些规则的遵守应受独立当局的控制。根据本条通过的规则不得损害《欧洲联盟条约》第39条规定的具体规则。"See Treaty on the Functioning of the European Union, https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/? uri=CELEX:12012E/TXT, last visited on 1 December 2024.

为了扭转这一局面，欧盟推出了内容更加全面更适应大数据时代的 GDPR。然而，事实证明数据保护法已经不能有效调整相关对象。为此，以欧盟《数字市场法》《数字服务法》为代表的竞争法出台，弥补了数据法的不足。但随着人工智能成为引领产业发展的主导技术，数据立法加竞争立法的格局已不能有效规范相关产业和活动。以欧盟为代表的人工智能专门立法应运而生。与此相应，单纯以数据法，尤其是以个人数据保护法作为主要规范基础对隐私权进行保护，已不符合现实。随着联邦学习和区块链等技术的高速发展，当前数据保护方法显得更加过时和陈腐。尽管我国《数据安全法》《个人信息保护法》的出台意义十分重大，但若想有效保护人工智能领域的隐私，现有《个人信息保护法》必须进行调整。

短期内，应充分发挥“隐私保护”(privacy preserving)的规范整合作用。由于数据保护无法取代隐私的需求。“隐私保护”成了整合各种隐私规范和技术的有效概念。这在奉行实用主义的美国首先得到实践。2023 年 3 月 31 日，美国白宫科技政策办公室(OSTP)发布《促进数据共享与分析中的隐私保护国家战略》(National Strategy to Advance Privacy-Preserving Data Sharing and Analytics，以下简称《战略》)，提出了数据共享与分析的隐私保护战略(Privacy-Preserving Data Sharing and Analytics，以下简称 PPDSA)。该战略确立了四个指导支柱，代表了其隐私和数据方法的基础：精心制作保护公民权利的 PPDSA 技术；在促进平等的同时促进创新；建立具有问责机制的技术；以及最大限度地减少弱势群体的风险。更重要的是，PPDSA 用“隐私保护”这个概念整合了各类社会、法律和技术手段。〔36〕 在隐私立法迟滞的情况下，隐私保护作为整合性的概念，不仅具有技术上的可操作性，也具有规范上的必要性。

中期内，应顺应人工智能产业现实，在法律上认真区分数据、信息、隐私概念。大数据时代的数据隐私保护规范将数据作为隐私保护的核心要素。但联邦学习却提示我们，紧盯数据很可能会导致人工智能系统中的隐私保护目标落空。这种情况在 LLM 领域也很明显。在人工智能场景中，数据、信息、隐私可通过“计算”进行转换，彼此动态关联。在大数据时代边界模糊的数据、信息和隐私的区分在人工智能时代变得更加必要。我国现有立法也是“数据”“信息”“隐私”三个术语并用。我国人工智能立法完全可以顺应技术事实，创新性地对人工智能场景下的数据、信息和隐私进行必要区分。

长期来看，应该发挥隐私权的作用，提升隐私保护效果并回归人格权的初衷。个人数据保护法兴起后，隐私权的重要性明显降低，明显有“退居二线”的现象。在大数据产业和人工智能产业领域，隐私权被拆解为若干个人数据保护方面的权利。在科技和产业高速发展的情况下，隐私权的规范内涵和实现路径显得晦暗不清。但从根本上说，隐私权是一种关乎自然人尊严

〔36〕 See National Science and Technology Council, “National Strategy to Advance Privacy-Preserving Data Sharing and Analytics(March 2023)”, https://www.whitehouse.gov/wp-content/uploads/2023/03/National-Strategy-to-Advance-Privacy-Preserving-Data-Sharing-and-Analytics.pdf, last visited on 1 December 2024.

和发展的人格权。人工智能时代更需弘扬人的尊严。随着产业技术的逐步稳定，加之规范层面的不断探索，未来应发挥隐私权概念的潜力，通过更具整合性的法律规范加强人工智能隐私保护。

(二)探索归责机制

人工智能隐私保护的归责问题非常复杂。尤其是联邦学习这类参与方众多且架构松散的人工智能项目，无论在技术上还是在法律上归责都是困难的。从技术上讲，隐私保护的归责审计关键在于解释。人工智能的解释是为了应对理论缺陷、应用缺陷与监管要求。〔37〕解释的目标也不仅是协调人与机器之间的关系，更重要的是加强人工智能参与者之间的沟通和信任。内在可解释模型的归责审计相对简单，但联邦学习等复杂模型都不是内在的可解释模型，这类复杂模型通常都是通过事后解析方法获得可解释性的。常见的复杂模型事后解析方法有部分依赖图、累积局部效应图、LIME 事后解析方法、SHAP 事后解析方法等。

除了可解释性这个方向外，人工智能隐私保护的归责审计问题还可以通过区块链技术加以实现。区块链并非单一技术，而是 P2P 动态组网、基于密码学的共享账本、共识机制、智能合约等技术的组合。区块链的最大意义在于搭建可信网络，在没有中心化背书的情况下，实现点对点的价值转移。由于区块链上的任何动作和数据变化都会被忠实记录下来，这无疑为各种监管的审计工作提供了良好的支撑。区块链与联邦学习集成不仅能进一步优化联邦学习的架构和流程，还可以加强联邦学习的激励和监管审计机制。〔38〕当然，区块链与人工智能的集成发展目前还处于探索阶段，且部署的成本相对高昂，但区块链人工智能在隐私保护归责机制方面的优越性是毋庸置疑的。

从规范上来看，目前欧盟和美国采用的都是基于风险的路径。但风险本身是难以测算的。人工智能归责机制的重点是发展富于弹性的责任主体、归责原则和责任框架。在这方面可以借鉴欧盟的经验。

1. 责任主体

目前经济合作与发展组织(OECD)、欧盟和美国采取了渐趋一致的适应人工智能的责任主体概念体系，尤其是欧盟《人工智能法》中的责任主体描述更加细致和完善，值得借鉴。欧盟《人工智能法》采取了完全不同于个人数据保护法的主体和责任概念框架。不再以数据为中心，而是结合人工智能在产品、服务和市场方面的特征界定了一整套主体概念并对其责任进行了详细规定。《人工智能法》的主体概念采取的是“参与者”(actor)这个表述，最核心的参与者

〔37〕 参见杨强、范力欣、朱军、陈一昕、张拳石、朱松纯等：《可解释人工智能导论》，电子工业出版社 2022 年版，第 7—9 页。

〔38〕 参见薄列峰、(美)黄恒、顾松庠、陈彦卿等：《联邦学习：算法详解与系统实现》，机械工业出版社 2022 年版，第 288—290 页。

是“提供者”(provider)和“部署者”(deployer)。〔39〕“提供者”是指开发人工智能系统或通用目标人工智能模型(general-purpose AI models),或已开发人工智能系统或通用目标人工智能模型,并将其投放市场或以自己的名义或商标提供服务的自然人或法人、公共机关、机构或其他团体,无论有偿还是无偿。;“部署者”是指在其授权下使用人工智能系统的任何自然人或法人、公共机关、机构或其他团体。除了提供者和部署者之外,人工智能法案还有更加宽泛的“运营商”(operator)概念,包括提供者、部署者、授权代表、进口商和经销商。很明显,与 GDPR 的数据中心的视角相比,《人工智能法》用更加广阔的市场化视角细致描述了人工智能领域的各种主体的责任。

2.归责原则

《人工智能法》对参与方的义务和责任并非泛泛而论,而是结合人工智能系统的开发、部署、使用等流程进行了非常具体的描述,并提出了三类责任原则:风险对应责任原则、公平分担责任原则和目标相称责任原则。

风险对应责任原则,即不同风险的人工智能系统对应不同责任。《人工智能法》根据“基于风险的方法”(risk-based approach)对人工智能系统进行分类并提出不同的要求和义务。存在“不可接受”(unacceptable)风险的人工智能系统在欧盟被禁止。“高风险”(high-risk)人工智能系统必须满足一系列要求和义务才能进入欧盟市场。“有限风险”(limited risk)的人工智能系统主要接受非常轻的透明度义务约束。“低风险或最小风险”(low or minimal risk)的人工智能系统无需遵守任何额外的法律义务便可以在欧盟开发和使用。在《人工智能法》制订过程中,高风险人工智能系统的范围逐步扩大。《人工智能法》第三章第 6 条描述了界定“高风险”人工智能系统的标准。《人工智能法》第 16 至 27 条规定了高风险人工智能系统的参与者必须履行相关义务,这些义务也构成责任区分和追究的重要根据。为了保证高风险人工智能系统的责任不会落空,《人工智能法》说明部分(79)项指出:由被定义为提供者的特定自然人或法人承担将高风险人工智能系统投放市场或投入服务的责任是适当的,无论该自然人或法人是否是设计或开发该系统的人。

公平分担责任原则,即按人工智能价值链(the value chain for AI)公平分担责任。《人工智能法》说明部分(97)到(117)项针对可能造成重大安全和基本权利影响的通用目标人工智能模型的各类问题进行了非常谨慎、务实和细致的立场阐述。为了确保沿着人工智能价值链公平分担责任,通用目标人工智能模型应符合《人工智能法》规定的相称且更具体的要求和义务

〔39〕“人工智能参与者”(AI actors)这个表述目前已成为经济合作与发展组织(OECD)、欧盟以及美国 Nist 共同采用的概念。本文将“actor”译为“参与者”而非“行动者”是因为 OECD 将人工智能参与者定义为“在人工智能系统生命周期中发挥积极作用的人,包括部署或操作人工智能的组织和个人”。这一概念强调的是发挥积极作用这一社会属性,而非特定的行动模式。将“provider”译为“提供者”而非“供应商”,是因为有很多人工智能系统都是开源和免费提供的。这也是采取“部署者”(deployer)译名的主要原因。我国《生成式人工智能服务管理暂行办法》中也采用了“提供者”这一概念,并将“生成式人工智能服务提供者”界定为“利用生成式人工智能技术提供生成式人工智能服务(包括通过提供可编程接口等方式提供生成式人工智能服务)的组织、个人”。参见《生成式人工智能服务管理暂行办法》第 22 条第(2)项。

(requirements and obligations),包括通过适当的设计、测试和分析来评估和减轻可能的风险和危险;实施数据治理措施;遵守技术设计要求;符合环境标准等。

目标相称责任原则,即为实现《人工智能法》目标必须对一些人工智能系统提出特定要求和责任。尽管《人工智能法》对通用目标人工智能模型规定了非常具体的要求和义务,但这并不意味着凡是使用通用目标人工智能模型的人工智能系统就一定是高风险的,只是为了实现《人工智能法》的目标才对其进行更为具体的要求。这其实显示了欧盟关于人工智能系统责任分配的原则除了按价值链分配外,还隐含着为实现《人工智能法》目标而承担相应责任的原则,即相称原则。

我国 2023 年发布的《生成式人工智能服务管理暂行办法》中也采用了"生成式人工智能服务提供者"的概念。但该办法对其他相关主体的描述过于简单,只有"生成式人工智能服务使用者"一种,没有"部署者""授权代表""进口商""经销商"等具体主体,也缺乏"参与者""经销商"等统称。全国网络安全标准化技术委员会 2024 年 2 月发布的《生成式人工智能服务安全基本要求》中也只有"服务提供者"一种主体描述。我国人工智能立法理应采用更为健全的责任主体概念并应探索恰当的归责机制。随着归责理论和规范的完善,人工智能隐私保护责任的难题也会逐步得到化解。

3.责任框架

试图采用一套简单流程框架覆盖人工智能的隐私保护的责任和义务,无异于痴人说梦。然而,不结合隐私保护流程的描述,就无法对具有高度复杂架构的人工智能系统进行合理的责任划分。理智的做法是采取一种双层责任框架,即在对一般人工智能系统的责任采取原则性规定的同时对特殊类型人工智能技术方案进行专门具体的规范。这种立法方式看似权宜之策,却是符合人工智能复杂性和发展现状的做法。欧盟的人工智能立法已经采取了这种双层责任框架。《人工智能法》对人工智能系统(AI systems)和人工智能模型(AI models)进行了区分。虽然模型是系统的重要组成部分,但模型本身并不构成系统,需添加各种组件才能构成系统。《人工智能法》结合人工智能开发、训练、部署、应用的一般流程对不同风险的人工智能系统责任做了原则性规定。与此同时,《人工智能法》说明部分(97)到(117)项结合开发、部署、使用的流程特征,对通用目标人工智能模型参与者提出了非常具体的要求和义务。《人工智能法》还在第 3 条(65)项中界定了通用目标人工智能模型的"系统性风险"(systemic risks),即"通用目标人工智能模型的高影响能力所特有的风险,因其影响范围广泛而对内部市场产生重大影响,并对公众健康、安全、公共安全、基本权利或整个社会产生实际或可合理预见的负面影响,可在整个价值链中大规模传播"。《人工智能法》第 53、54 条规定了通用目标人工智能模型提供者的义务,第 55 条规定了具有系统风险的通用目标人工智能模型提供者的义务。

这种双层框架增加了人工智能立法的弹性和适应力。但欧盟《人工智能法》的双层框架并不彻底。实际上,之所以对通用目标人工智能模型专门进行规范,正是因为其具备重大的影响力容易导致更大的风险。未来完全可能出现其他具备重大影响力的人工智能技术方案,而且既有的技术也可能因为部署应用规模扩大从而具有重大影响的能力。完全可以将影响力作为标准,通过较为严格的程序确定需要专门规范的人工智能技术方案,并规定更有针对性的义务

和责任,从而形成更完整的双层责任框架。

(三)调整规范重点

人工智能的隐私保护需求很难在安全框架下得以全面实现。人工智能立法需要区分安全和隐私保护需求,让隐私保护重回人格权益的本质。而做到这一点必须坚持系统观点和工程观念,以“计算”为核心对隐私保护要素进行全面规范。

人工智能隐私保护必须重视系统观念。当前特别需要强调人工智能隐私保护不能只盯住数据,而应综合考虑整个人工智能系统开发和应用周期的各种要素,以隐私保护需求的实现为导向,对最重要的环节进行恰当的规范。充分重视法律对各种技术标准、指南、最佳实践的引导作用,引导制定各种可行的人工智能隐私保护标准。

人工智能隐私保护还需重视工程观念。作为一种上世纪九十年代提出的理念,隐私设计原则提出时面对的技术和产业环境与当前存在很大差异。人工智能产品和服务开发的流程也与大数据产业存在明显不同。虽然隐私设计也强调全生命周期的规范,但其重点毕竟在设计,对更复杂的人工智能系统开发的穿透性有限,很难保证隐私需求的充分实现。以深度学习为代表的人工智能技术架构实际上存在较多“黑箱”环节,设计需求必须经过工程调适才能得到逐步实现。在此情况下,隐私工程(privacy engineering)可以发挥更大的作用。隐私工程的重点在于实施一系列技术来降低隐私风险,并确保组织能够就资源如何分配和如何有效地实施信息系统控制做出有目的的决策。从原理上来看,隐私设计应该是先于隐私工程的步骤。但正如有学者指出的那样,隐私工程实际上通常包含了整个系统开发生命周期与隐私相关的活动。〔40〕

在系统观点和工程观念的指导下,人工智能立法应该以“计算”而非“数据”为核心描述技术和产品流程,并以此为基础对人工智能涉及隐私的训练和推理过程进行有针对性的重点规范。在训练阶段除了注重数据处理的合法基础外,更需要对人工智能提供者的隐私保护责任加以引导,使其更关注数据合规之外的人格性的隐私保护风险。在推理应用阶段,需要进一步开拓自然人参与人工智能项目的渠道,让自然人有更多的工具和手段维护自身人格性的隐私权益。

(四)建构沟通规范

人工智能是一种典型的“社会技术”(socio-technical)。人工智能系统受到社会动态和人类行为的影响。人工智能的风险和收益受到技术因素与社会因素的相互作用,这些社会因素与系统的使用方式、系统与其他人工智能系统的交互、操作系统的人以及系统部署的社会背景有关。〔41〕社会技术特性为有效规范人工智能提供了“沟通”(communication)路径。在人工智能的有效规范中,沟通比设计更加重要。如果没有沟通,隐私保护的需求很容易在复杂的技术权衡中被牺牲掉。实际上,透明性和可解释性更应该被理解为一种沟通要求。这种沟通涉

〔40〕 参见斯托林斯,见前注〔12〕,第 32 页。

〔41〕 See Artificial Intelligence Risk Management Framework (AI RMF 1.0), https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf, p. 1, last visited on 1 December 2024.

及人工智能的开发者、应用方、监管方、研究者、普通用户以及权益可能受影响的自然人。对类似联邦学习和 LLM 这类还处于发展期的技术架构，研究者的参与显得尤为重要。从软件开发的角度来看，沟通也是最重要的技能之一。在技术迭代速度更快的人工智能领域，只有通过良好的沟通，开发者才能从客户和最终用户那里获得准确和详细的隐私保护需求信息，从而降低项目失败和客户不满的风险。各种隐私增强技术也可以从沟通的渠道进行理解。对人工智能进行强介入规范是不符合技术和产业规律的。隐私增强技术领域的规范和标准也不是为了高强度介入式规制相关技术。但为了对实质影响隐私保护效果的技术权衡过程进行约束，制订有利于沟通的规范则是必要的。通过这类能显著提升参与性和沟通效果的法律规范和技术标准，能够更好地保证隐私需求在工程层次得到实现。

我国人工智能立法也应该“风险路径”和“沟通路径”并重，全方位地促进隐私保护的规范完善。立法可对系统应具备合理透明度这样的法律机制进行强化和细化。只有人工智能系统的参与度和透明度得到提升，隐私保护的有效性才能得到保证。

四、结　论

前欧洲数据保护监督员（European Data Protection Supervisor）乔瓦尼·布塔雷利（Giovanni Buttarelli）〔42〕指出：隐私悖论（privacy paradox）并不是人们拥有的隐藏与暴露需求之间的矛盾，而是我们还没有找到恰当的方法去应对由快速数字化带来的可能性与脆弱性问题。〔43〕与大数据相比，人工智能技术的猛烈变革对人类社会的冲击更加深入。但当前对人工智能风险的反应却呈现出“远忧”过度，但“近虑”不足的状态。人工智能可能取代人类或控制人类等各种危机论层出不穷。与此同时，对正在发生影响的人工智能隐私风险的关注却严重不足。人们没有根据地默认大数据时代的技术和规范能够顺利解决人工智能的隐私问题。有研究表明，人类在没有任何概率分布信息的不确定性状态下进行决策时，大脑激活的主要脑区是框前额叶和杏仁核。前者激发的是计划和对本能的抑制，而后者激发的是惧怕的情绪。〔44〕当前大众对人工智能的认知偏差很大程度上来自对极端不确定性的情绪反映。对普通人而言，这或许无可厚非。但学者不应随波逐流，更不应放大和滥用消极情绪。乌卡时代

〔42〕 欧洲数据保护监督员（简称 EDPS）是欧盟设立的独立数据保护机构。EDPS 由监督员及办公室构成。根据 GDPR 第 68、75 条等规定，EDPS 与欧盟各成员国数据保护机构代表共同组成欧洲数据保护委员会（European Data Protection Board）。Giovanni Buttarelli 于 2014 年 12 月担任 EDPS，一直在此岗位工作至 2019 年 8 月去世。

〔43〕 Alan Charles Raul (ed.), *The Privacy, Data Protection And Cybersecurity Law Review*, 6th ed., London: Law Business Research Ltd, 2019, p. 3.

〔44〕 See Ming Hsu, Meghana Bhatt, Ralph Adolphs, Daniel Tranel and Colin F. Camerer, “Neural Systems Responding to Degrees of Uncertainty in Human Decision-Making,” *Science*, Vol. 310, No. 5754, 2005, pp. 1680-1683.

(VUCA)〔45〕的法学研究应该摆脱情绪阻碍,克服认知障碍,在了解行业和技术动态的前提下,探寻规范层面的定力和基石。

Abstract: Artificial intelligence legislation often tends to focus on specific technologies. Federated learning, a mainstream machine learning technique, is distinguished by its architecture, which is designed with privacy needs in mind. Federated learning has been widely applied in fields such as finance and data sharing, significantly impacting individual rights. However, its privacy-centric design has also exposed various privacy risks, highlighting deficiencies in the legal framework for personal data protection: sparse regulations leave federated learning without clear privacy requirements, limiting the effectiveness of its "privacy by design" advantage; its distributed architecture makes assigning privacy protection responsibilities difficult; an excessive emphasis on confidentiality and security weakens and transforms the concept of privacy as a personal right; and a lack of regulatory guidance for technical trade-offs undermines transparency and certainty in privacy protection. These issues reveal significant gaps between artificial intelligence privacy protection and personal data protection in terms of their objects, processes, responsibilities, and frameworks. To meet the unique demands of privacy protection in artificial intelligence, future efforts could focus on integrating regulatory foundations, adjusting regulatory priorities, exploring liability mechanisms, and establishing communication frameworks to enhance and refine privacy protection standards.

Key Words: Artificial Intelligence Legislation; Federal Learning; Privacy by Design; Differential Privacy; Privacy Preserving Computation

(责任编辑:彭 錞)

〔45〕 是指人们生活在一个不稳定性、不确定性、复杂性、模糊性的时代、境况或者世界中。VUCA 是 volatility(易变性),uncertainty(不确定性),complexity(复杂性),ambiguity(模糊性)的缩写。