AI与数据安全

AI and Data Security

宁　静，陈　飚，侯玉华（中讯邮电咨询设计院有限公司，北京 100048）
Ning Jing，Chen Si，Hou Yuhua（China Information Technology Designing & Consulting Institute Co.，Ltd.，Beijing 100048，China）

摘　要：
人工智能（AI）与数据安全正形成深度互构的相互促进的发展关系。一方面，AI技术通过智能威胁检测、自动化响应与行为分析，显著提升数据安全防护效率。机器学习模型可实时识别异常流量，预测潜在攻击路径，自然语言处理助力敏感信息分类与合规审查，大幅降低人工运维成本与响应延迟。另一方面，数据安全技术为AI发展提供可信基石，隐私计算（如联邦学习、差分隐私）保障模型训练中数据的“可用不可见”，加密技术与数据脱敏确保高价值训练集的合法流通。这种双向赋能不仅重塑了安全防御边界，更推动AI在更多敏感领域的合规落地，最终形成“以AI强化安全，以安全护航AI”的可持续发展生态。

关键词：
人工智能；大模型；数据安全；隐私保护；联邦学习；差分隐私；安全多方计算；可信执行环境
doi：10.12045/j.issn.1007-3043.2025.09.010
文章编号：1007-3043（2025）09-0052-06
中图分类号：TN915.08
文献标识码：A
开放科学（资源服务）标识码（OSID）：

Abstract：
Artificial intelligence （AI） and data security are forming a deeply intertwined and mutually reinforcing development relationship. On the one hand, AI technology significantly enhances the efficiency of data security protection through intelligent threat detection, automated response, and behavior analysis. Machine learning models can identify abnormal traffic in real time, predict potential attack paths, and natural language processing aids in sensitive information classification and compliance review, significantly reducing manual operation and maintenance costs and response delays. On the other hand, data security technology provides a trusted cornerstone for AI development: privacy computation （such as federated learning, differential privacy） ensures that data is “usable but invisible” during model training, and encryption technology and data desensitization ensure the legal circulation of high-value training sets. This bidirectional empowerment not only reshapes the boundaries of security defense but also promotes the compliant implementation of AI in more sensitive areas, ultimately forming a sustainable development ecosystem where “AI strengthens security, and security escorts AI”.

Keywords：
Artificial intelligence; LLM; Data security; Privacy protection; Federated learning; Differential privacy; Secure multi-party computation; TEE

引用格式：宁静，陈飚，侯玉华. AI与数据安全[J]. 邮电设计技术，2025（9）：52-57.

1 AI发展历程

人工智能（Artificial Intelligence，AI）的发展历程可以追到20世纪中叶，其演进经历了多次高潮与低谷。1956年夏天，在达特茅斯学院（DartmouthCollege）的一次会议上，一些专家把人工智能正式确立为计算机科学的一个全新研究领域[1]。

1956—1974年是人工智能发展的第一次高潮，人们发现计算机可以证明数学定理、学习使用语言，大量成功的初代人工智能程序和研究方向不断出现。然而，当时的人工智能研究还难以获得足够的支持，限制了其发展[2]。

收稿日期：2025-07-17

20世纪80年代以来，随着计算机性能的突飞猛进，计算机编程语言可以通过程序结构来实现逻辑功能。这促使一种模拟人类专家解决专业领域问题的计算机程序系统即专家系统得到发展，并引发关注。但是，短暂的热潮之后，专家系统暴露出如硬件存储空间的限制、系统维护成本的增加、不会自己学习等问题，人们对专家系统的研究陷入了困境。后来随着大数据技术、硬件技术和深度学习算法的发展，人工智能技术得到快速发展，2017年AlphaGo战胜世界围棋冠军柯洁后，显示了人工智能在局部领域已经超越人类水平；2017年，Google在其论文《Attention Is All Your Need》中提出自注意力机制，基于该机制提出了Transformer算法，基于该算法中的解码器部分结合预训练和微调技术，诞生了生成式预训练GPT系列模型，2023年，ChatGPT模型横空出世，该模型大幅改善了对话的逻辑性、交互性，2024年，DeepSeek大模型技术问世，人工智能技术迎来了前所未有的快速发展期[3-4]。

在快速发展的人工智能领域，大模型的使用在推动技术进步和解决问题方面有着显著的潜力，但同时也伴随着一系列的网络安全风险和数据安全风险，一方面，大模型技术的背后是海量训练数据的支撑，大模型输出结果的准确性与参与训练模型的数据量息息相关，随着人工智能技术的不断发展，参与模型训练的数据量激增，这些数据可能成为黑客攻击的目标，从而对保护数据隐私和保证数据安全提出了新的挑战，另一方面，人工智能技术的发展为数据安全、数据治理提供更高效可靠的防护手段[5]。

数据是人工智能的输入，数据安全技术的发展与人工智能的发展存在相辅相成的关系。一方面，通过数据安全技术，可为人工智能的应用提供高质量的合规数据。另一方面，人工智能在整个数据生命周期的管理及数据隐私保护中发挥着重要作用[6]。

2 人工智能赋能数据安全

数据是指对客观事件进行记录并可以识别的符号，是对客观事物的性质、状态以及相互关系等进行记载的物理符号或是物理符号的组合。它是可识别的、抽象的符号[7]。

2.1 人工智能技术在数据生命周期管理中的应用

在当前万物融合的时代，数据的范畴得到了极大的外延，数据的数量急剧膨胀，伴随着大量数据的产生，如何对数据进行有效管理就变得越发重要，因此数据生命周期管理的概念被提出，可分为数据采集、数据存储、数据传输、数据处理、数据交换、数据销毁，数据的生命周期是指从数据的采集开始，直至数据使用寿命结束被销毁后数据生命周期结束[8]。

数据生命周期管理是对整个数据生命周期的数据流进行管理的过程。在整个数据生命周期中准确地识别和保护数据，是数据安全的基础要求。不同数据类型的不同生命周期阶段的安全防护需求也不相同。在数据采集阶段，人工智能技术可快速准确有效地进行数据识别、分类；在数据存储、使用和分享阶段，人工智能技术可快速有效进行攻击识别，从而保证数据安全；在数据归档阶段，人工智能技术可实现数据的自动识别和归档。下面对AI在数据生命周期中各阶段的应用进行介绍[9]。

在数据的采集阶段，不同的数据类型的采集方式可谓千差万别。根据数据安全的基本原则，在数据的采集阶段，应该对数据进行识别并做出合理的分类，并根据该类数据的保护要求控制其被访问和使用的范围。随着海量数据的涌现，使用传统人工标注分级分类方法已无法满足实际使用需求，而使用人工智能技术，例如机器学习、图像识别、自然语言处理等，可实现对文件及数据核心内容的提取，按照内容对数据进行梳理，实现对数据自动、精准地分级分类，同时这些数据可以被标注成为样本数据用于模型训练，提高模型的精准度。尤其对于隐藏在海量数据中的所有未被发现的敏感信息的评估以及对一些非结构化数据的处理，如果依赖人工处理，将无法大规模地开展数据识别工作，而采用人工智能技术可从非结构化数据中自动提取敏感信息，一旦敏感信息被有效识别，就可通过删除、编辑或更改其访问控制权限等操作进行保护，可有效消除数据泄露和勒索软件攻击带来的负面影响。

在数据的存储阶段，与采集阶段类似，数据的存储方式也有很大的差异，可能在本地存储设备上存储，也可能采用分布式存储、云存储。本阶段可利用AI技术自动实现必要的数据访问控制措施，以确保数据受到合适的保护，从而降低数据被泄露、被篡改和被破坏的风险。

在数据的使用阶段，数据安全防护的关键点为将安全控制机制应用于使用中的数据。一般而言，应能够监控对数据，特别是敏感数据的访问，并应用各类

安全控制技术,以确保数据的安全。

数据的分享阶段往往是传统的数据安全控制中最薄弱的环节。而对于数据这一信息载体而言,数据在分享后才能产生价值,但在一定程度上,数据安全和数据价值会产生冲突,这也是为什么在正确的时间采用正确的安全控制措施很重要的原因。

在数据存储、使用和分享阶段,数据可能会面临恶意软件攻击。传统的恶意软件检测主要通过规则匹配以及hash值计算、运行截图比较等方式对恶意软件进行检测,但这些传统的恶意软件检测手段都需要通过人工分析发现恶意软件,然后人为地将其总结为规则或计算hash值,在此基础上恶意软件检测系统才能对这些已知的恶意软件进行检测。对新出现的恶意软件,传统检测方式无能为力。同时,在实际的业务场景中,过多复杂的规则匹配会给系统带来无法承担的压力,而通用性较高的规则库匹配则会带来大量的误报,从而放过可能存在的攻击点。再者,随着人工智能技术的发展,可由AI自动大量生成恶意软件,如果仍依赖传统检测方式将无法保证恶意软件识别的及时性及准确率,因为表层对抗是无穷无尽的,检测需要由表及里,尽可能挖掘出更本质的检测方法,这就是人工智能在恶意软件攻击识别中所能起到的作用。AI对海量数据间特征的分析要比人类强大许多,利用AI技术可对恶意软件做各种层面的分析,能够学习到恶意软件的深层特征,能够对海量数据间的特征进行关联分析。对AI模型投喂恶意攻击训练数据即可学习新的攻击模式,从而使其具有对未知恶意攻击软件的识别能力,可有效助力对恶意攻击的检测识别,从而对数据安全防护起到关键作用。

随着互联网业务场景的不断发展,可获取到的数据越来越多,数据维度越来越丰富,而在获取到大量不同维度的数据后,需要对数据的关联关系进行有效分析。由于数据使用场景及数据间的关联关系严重影响数据可被使用范围的定义,依靠人工是无法衡量和评估如此复杂的关联关系及不同场景对数据的需求,与专家经验相比,人工智能技术具有强大的逻辑推理能力,能更快速、更有效地挖掘和识别不同维度数据之间的关联关系,从而提升大量不同维度数据在实际使用中的价值,也可针对不同的使用场景对数据采取不同的访问控制措施。

在数据的归档阶段,需要关注不同法律法规以及公司对数据留存的约束性要求,并应根据相关政策和用户需求,决定数据的归档方式、保护机制和隔离要求,基于人工智能技术可实现对政策法规及用户需求的自动识别,同时结合人工智能技术在数据识别分类中的应用,可实现数据的自动归档,提高数据归档的准确性和效率。

在数据的销毁阶段,数据的归档时限已过,或者根据监管或业务需要删除数据。此时,数据安全的关注重点是采取合理的数据销毁机制,以确保数据无法被恢复。即使数据并非以传统的文件类型存储,也需要考虑相应的销毁机制。基于人工智能技术可自动识别监管要求,然后对所存储的数据采取相对应的销毁措施进行删除,确保在销毁期采用合理合规的销毁机制对应销毁数据予以销毁。

在数据生命周期管理中应用人工智能技术,可有效提升数据识别、数据分类、数据销毁等方面的智能化水平,从而提升数据安全防护能力[10]。

2.2 人工智能技术在数据隐私保护中的应用

人工智能被广泛认为是第4次工业革命的核心驱动力,正在重塑经济、社会和技术的各个方面,如今数据安全与隐私保护却可能会成为制约其发展的“绊脚石”。数据对于大模型的重要性不言而喻,它是构建和优化这些模型的基石。大模型如GPT-4、ChatGPT或DeepSeek等,依赖于大量、多样化且高质量的数据来训练其复杂的算法。用于进行训练的数据的数量和质量直接影响模型的性能,包括其理解、推理和生成文本的能力。因此,采取必要的保护数据安全和数据隐私技术及措施,不仅是确保大模型达到最佳性能的关键,也是利用数据为业务增长提供动力的关键。

敏感数据是指一旦泄露可能对个人、企业或社会造成严重危害的数据,按照来源不同,其可分为个人敏感数据(包括姓名、身份证号、银行账户、医疗记录等直接关联个人身份与权益的信息,这些数据在用户注册、交易、就医等场景中自然产生)和企业机构敏感数据(包括经营数据、网络架构、IP地址列表等,这些数据多在企业运营、技术系统运行中生成)[11]。

当前欧盟的《通用数据保护条例》(GDPR)对企业处理用户数据的方式提出了严格的要求,企业若违反规定将面临巨额罚款。例如脸书和谷歌在欧洲因违规使用用户敏感数据而被处以巨额罚款。而在国内,《中华人民共和国国家安全法》《中华人民共和国网络安全法》《中华人民共和国数据安全法》《中华人民共和国个人信息保护法》等法律法规以及其他规范对数

据的产生、分类、使用提出了明确要求，若发现有违规使用行为，将面临严重处罚。

在敏感数据的产生阶段，使用AI技术对数据类别、数据敏感性、数据可使用范围等关键约束项进行识别和标注，能够有效控制数据应用范围，提高数据流通效率和价值。同时，AI具备强大的数据分析和处理能力，能够实时监测数据的访问和使用情况。通过对大量历史数据的学习，AI可以建立起正常数据行为的模型。一旦数据的访问模式或使用出现偏离正常模型的异常行为，AI系统能够迅速发出警报。例如，在系统网络中，AI可以实时监测用户对数据的访问请求，若发现某个用户在短时间内频繁访问大量敏感数据，且访问行为与日常访问模式不符，AI系统就能及时检测到这一异常行为，并通知安全管理人员进行进一步调查。这种实时监测和异常检测能力能够帮助系统管理员及时发现潜在的数据安全威胁。

当检测到隐私泄露事件发生时，应在第一时间采取措施进行防范，对于隐私泄露事件，及时响应将有效减少恶意攻击所带来的安全影响，利用AI系统可以实现自动化的安全响应，大大缩短响应时间，降低安全事件造成的损失。一旦检测到异常行为，AI系统能够自动采取一系列预先设定好的措施，如隔离受影响的数据、阻断可疑的网络连接、启动数据备份和恢复流程等。例如，在遭受网络攻击时，AI系统能够迅速识别攻击类型和来源，并自动调整防火墙策略，阻止攻击的进一步蔓延。同时，AI系统还可以根据攻击的特点和历史数据，预测下一步可能的攻击行为，提前做好防范准备。采用这种自动化的安全响应机制能够在面对复杂多变的数据安全威胁时，迅速做出反应，保障数据的安全。

3 数据安全赋能AI

AI在数据层、模型层及应用层面临的安全和隐私威胁呈现多样性、隐蔽性和动态演化的特点。AI在数据层所面临的安全攻击主要围绕数据的机密性、完整性和可用性展开。

AI数据层所面临的针对数据机密性的攻击主要有数据窃取攻击、特征推理攻击、成员推理攻击和模型窃取攻击；AI所面临的针对数据完整性的攻击主要有数据投毒攻击和对抗样本攻击；AI模型所面临的可用性攻击可能发生在AI系统的推理或运行阶段，在推理阶段，AI模型可能会接收并处理大量的异常输入甚至是恶意输入，导致模型无法提供正常的服务。

大模型的一个重要影响因素就是数据，只有汇聚大量优质的数据，才能保障模型的训练和应用效果，然而在实际中，数据在使用过程中会面临针对机密性、完整性和可用性的攻击，同时各类不同数据可能存在于多个企业、多个部门、甚至是多个地域，受制于数据安全隐患、合规监管要求和公司商业利益保护，无法被直接汇聚供模型训练使用。基于数据可能会面临的安全攻击以及数据孤岛问题，采用有效的数据安全和隐私保护措施，可在保护数据安全的同时连接数据孤岛，实现多方协同释放数据要素价值，促进人工智能技术的发展。

数据安全保护技术大致可分为基于密码学的数据安全防护技术、基于硬件可信计算的数据安全防护技术以及融合衍生技术联邦学习，在实际部署中可多项技术融合使用以达到最佳效果[12]。

3.1 基于密码学的数据安全技术

作为数据安全领域中的一个核心技术，密码学在确保安全的方面扮演了重要角色，并为数据隐私、完整和认证提供理论基础和技术支撑。基于密码学的数据安全保护技术有多方安全计算、同态加密及差分隐私[13]。

3.1.1 多方安全计算

多方安全计算满足了以安全的方式处理隐私保护场景中几个拥有隐私数据的参与方执行协作计算的问题。可保证2个或者多个拥有隐私数据输入的参与方在不知道其他方隐私数据的情况下，使用通过多方安全计算后得到的隐私计算结果。目前主要有4种计算框架来实现多方安全计算，包括不经意传输、混淆电路、秘密共享和零知识证明，目前多方安全计算已用于各种传统的机器学习模型，图1所示为多方安全计算的实现原理[14]。

3.1.2 同态加密

同态加密是一种满足同态运算性质的加密技术，在对密文进行特定形式的代数运算后，得到的仍然是加密的结果，将其解密所得到的结果与对明文进行同样运算后得到的结果一样。同态加密主要关注在数据处理时如何保护数据安全不被泄露。按照支持同态运算的类型和深度，可以将同态加密分为部分同态加密、类同态加密和全同态加密。图2所示为同态加密过程中明文数据和密文数据的等效性。

3.1.3 差分隐私

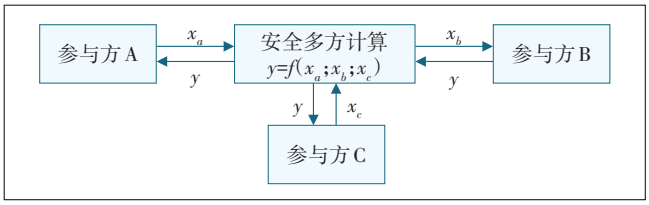

图1 多方安全计算原理

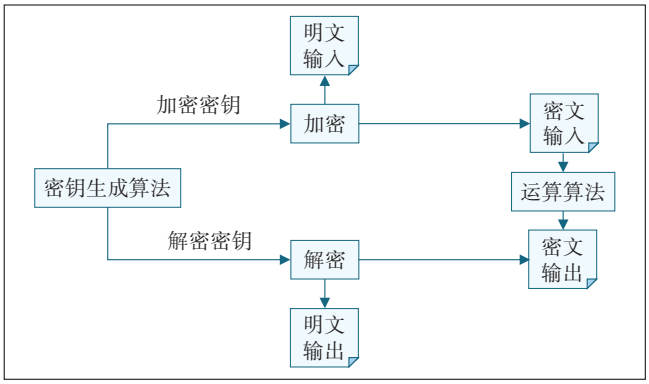

图2 同态加密系统示意

差分隐私技术是通过在数据中添加随机噪声(如拉普拉斯噪声等),使攻击者无法判断某一样本数据是否在数据集中,例如在模型训练数据中的隐私信息(如用户身份证号)可能被模型记忆并在推理时泄露。该漏洞一旦出现将严重违反隐私法规,可通过在模型训练前进行训练数据脱敏(去除或加密敏感字段)或通过差分隐私技术在模型训练数据中加入噪声数据,降低数据关联性从而达到对数据隐私保护的目的。

3.2 基于可信计算的数据安全技术

基于硬件技术的数据安全防护技术——可信执行环境(Trust Execution Environment,TEE)是打造一个用于执行计算的独立区域,其核心思想是隔离,就是将TEE与操作系统分割开并行运行。在CPU为TEE提供的特定区域内,数据和代码的执行完全与外部环境隔离,从而保证了机密性和完整性,TEE内的应用程序可以正常访问CPU和内存的全部功能,但不会受到操作系统中其他应用程序的影响[14]。

TEE与普通操作系统(Rich Execution Environment,REE)一同运行,TEE确保重要数据在不可信环境下不被篡改或窃取,REE则用于执行操作系统中其他应用程序,双方通过TEE提供的特定接口进行数据交互,不能随意互相访问,共同组成终端所有的硬件。

TEE通过3层架构实现安全。3层架构包括可信应用软件层(Trusted Application,TA),应用软件是需要被保护的应用软件;安全操作系统层,对上层应用软件TA提供支持;硬件层安全组件,提供硬件支撑。图3所示为ARM的TRustZone可信环境的实现架构。

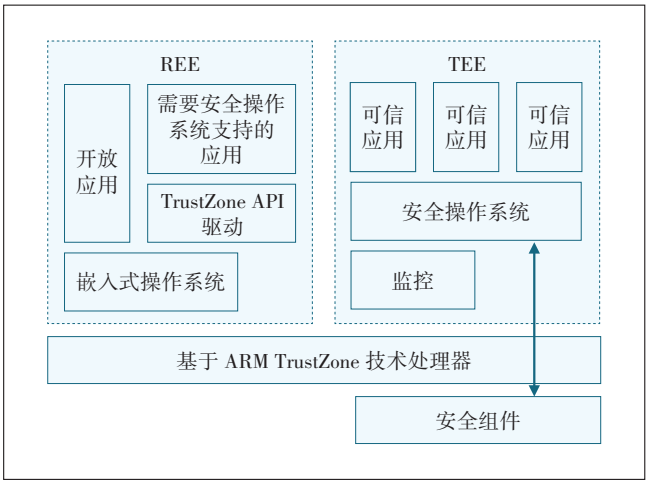

图3 ARM TrustZone TEE可信计算架构

3.3 联邦学习

联邦学习是一种多个参与方在各自数据不出本地的前提下共同完成某项AI模型训练任务的技术。通过联邦学习,不同的数据拥有方可以在不交换彼此数据的情况下,建立一个虚拟的共有模型,这个虚拟模型的效果等同于各方把数据聚合在一起建立的模型。联邦学习可通过交互各参与方本地计算的中间因子来实现模型的聚合和优化。联邦学习通过交互中间因子来替代原始数据,但直接交互明文的中间因子也有泄露和反推原始数据的可能性。根据各参与方数据的不同特点,联邦学习具体分为横向联邦学习、纵向联邦学习和联邦迁移学习。图4所示为纵向联邦学习实现架构[15]。

联邦学习有着去中心化以及保证原始数据不出域的优势,在理论上可有效解决数据孤岛和数据安全问题,但也还面临一些需要攻破的问题,比如,如何降低在迭代训练过程中因复杂的通信和计算消耗而带来的性能损失,如何建立激励机制吸引更多参与方加入以及如何防止由于梯度信息的泄露而导致原始数据泄露。

为加强建模过程中对数据隐私的安全保护,实际方案大多是在联邦学习的基础上结合多方安全计算、同态加密、差分隐私等密码学技术,对交互的中间因子进行加密保护,其中利用同态加密技术对中间因子进行加密交互,是目前应用最多的方案。另外,还有

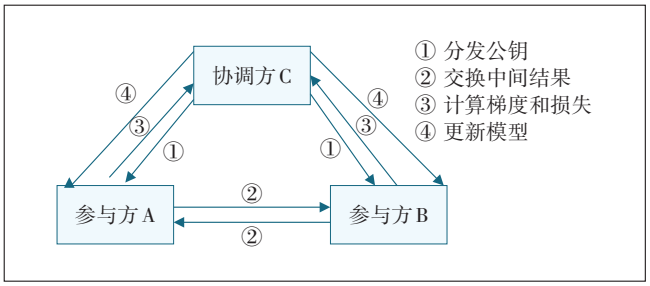

图4　纵向联邦学习架构

方案结合了可信执行环境，实现基于可信硬件的中间因子安全交互。

4 结束语

在AI飞速发展的同时，我们也要认识到其在发展过程中所面临的问题。

a）AI的发展可能引发一般性伦理问题和社会问题，AI未来会在某些领域取代人类工岗位，从而导致失业问题，但我们应该认识到AI的发展也会创造新的岗位，整体提升社会生产效率。

b）AI的发展可能会引发隐私泄露问题，大量数据参与模型训练，有效提升模型解决问题的能力，大量数据不止是对数据的规模的要求，还对数据的多样性和实时性有要求，在大数据技术和设备的支持下，用户的某些个人信息可能会被实时采集并存储，利用大数据挖掘技术，数据使用者可在不完全获取数据的情况下提取出有用的个人信息，使得个人隐私处于随时被窥探的状态。

c）未来随着AI道德自主性的提升，AI将具备自主做出道德选择和行动策略的能力，行为将不受人类控制，这势必会给现有的社会秩序和伦理规范带来强烈的冲击。

在AI发展的道路上会面临各种问题，数据安全作为其中的关键问题，关乎每个人的切身利益和整个社会的稳定发展。我们必须清醒地认识到AI时代数据安全面临的新挑战，积极探索利用AI技术保障数据安全的新路径。通过应用AI技术，采取有效的应对策略，在AI飞速发展的浪潮中筑牢数据安全的坚固防线，让AI技术更好地服务于人类社会的发展。AI技术的发展与数据安全技术的发展也将呈现更加紧密的关系，两者相互促进、共同发展。一方面，AI技术将不断提升数据安全防护的智能化水平，提供更加高效、精准的安全保障；另一方面，也需要不断创新数据安全防护技术，确保应用于AI技术发展的数据被安全合规使用。随着AI技术和数据安全保护技术的发展进步，我们能够在享受AI技术带来便利的同时，有效地保障数据安全，让数字世界更加安全、可靠、美好。

参考文献：

[1] 亨利基辛格，埃里克施密特，丹尼尔胡滕洛赫尔. 人工智能时代与人类未来[M]. 北京：中信出版集团，2023：205-210.
[2] 苏秉华，吴红辉，刘梦菲，等. 人工智能（AI）应用从入门到精通[M]. 北京：化学工业出版社，2020：55-73.
[3] 刘兆峰. 多模态大模型：算法、应用与微调[M]. 北京：机械工业出版社，2024：158-173.
[4] 邱才明，凌泽南，冯湛搏，等. ChatGpt漫谈[M]. 北京：清华大学出版社，2024：56-69.
[5] 余来文，林晓伟，刘梦菲，等. 智能时代：人工智能、超级计算与网络安全[M]. 北京：化学工业出版社，2018：55-73.
[6] Gartner. 生成式人工智能安全[EB/OL]. [2025-05-10]. https://www. gartner. com/cn/newsroom/press-releases/2024-cte-gen-ai-attact.
[7] Gartner. 数据安全平台市场指南（中国篇）[EB/OL]. [2025-05-10]. https://www. doc88. com/p-49819740154887. html.
[8] 国家工业信息安全发展研究中心. 数据安全白皮书[EB/OL]. [2025-05-10]. https://www. cics-cert. org. cn/etiri-edit/kindeditor/attached/upload/2021/05/27/1b44236fed303bd2864f30cba909421b. pdf.
[9] 刘隽良，王月兵，覃锦端，等. 数据安全实践指南[M]. 北京：机械工业出版社，2022：143-159.
[10] 张莉. 数据治理与数据安全[M]. 北京：人民邮电出版社，2024：78-89.
[11] 李凤华，牛犇，李晖. 隐私计算理论与技术[M]. 北京：人民邮电出版社，2021：35-43.
[12] 张曙光，涂锟，陆阳，等. 隐私计算与密码学应用实践[M]. 北京：电子工业出版社，2023：165-173.
[13] 陈左宁，卢锡城，方滨兴. 人工智能安全[M]. 北京：电子工业出版社，2024：110-123.
[14] 范渊，刘博. 数据安全与隐私计算[M]. 北京：电子工业出版社，2023：113-119.
[15] 耿佳辉，牟永利，李青，等. 联邦学习原理与算法[M]. 北京：机械工业出版社，2023：21-24.

作者简介：
宁静，毕业于天津大学，工程师，硕士，主要从事终端安全、密码通信安全等安全领域的技术研究工作；陈飓，毕业于北京理工大学，高级工程师，硕士，主要从事移动终端信息安全、云手机安全业务相关的研究工作；侯玉华，毕业于沈阳工业大学，高级工程师，硕士，主要研究方向为移动信息安全、终端操作系统。