中国联通

人工智能隐私保护白皮书

中国联通研究院

中国联通网络安全研究院

下一代互联网宽带业务应用国家工程研究中心

2023 年 11 月

版权声明

目录

前言

人工智能（Artificial Intelligence，简称AI）作为战略性新兴产业，作为新的增长引擎，日益成为科技创新、产业升级和生产力提升的重要驱动力量。生成式人工智能工具、人脸识别、智能工厂、智慧城市等人工智能技术现已广泛落地，这些令人难以置信的技术正在快速改变人们的生活，对经济社会发展和人类文明进步产生深远影响。与此同时，人工智能技术也带来难以预知的各种风险和复杂挑战，潜在的滥用对以前被认为是不可侵犯的个人敏感信息构成了前所未有的威胁，技术自身缺陷导致智能决策在多个领域存在不确定性和敏感信息泄露，系统被非法控制导致个人隐私被未授权的第三方获取和推理。因此，人工智能技术引发的隐私与安全问题已经成为时下的关注话题，也是当前人工智能领域所面临的最大挑战之一。

为了更好的推动新一代人工智能安全发展，让人工智能用的放心，各国政府和企业越来越重视人工智能隐私保护。人工智能隐私保护指的是在数据采集存储和数据使用共享，模型训练以及模型推理应用的全生命周期过程中有效的保护用户数据隐私不泄漏，不被未授权第三方获取或推理。因此，在人工智能处理大量个人数据和敏感信息的过程中，如何加强数据隐私管控；在人工智能训练过程中，如何保证数据质量，避免原始数据隐私泄露；在人工智能推理应用过程中，如何防御攻击引起的数据隐私推理，如何保护模型保密性与完整性日渐成为国际人工智能的重要议题。

本白皮书从人工智能隐私保护的内涵出发，从人工智能全生命周

期系统梳理人工智能通用隐私风险和生成式人工智能隐私风险。在此基础上，总结了国内外人工智能隐私保护法规政策标准化现状。然后分析了人工智能隐私保护技术和平台，包括管控技术、数据加密技术、攻击防御技术、隐私保护机器学习平台和人工智能安全检测平台等。最后以技术发展和隐私保护并重为原则，研究提出了多维度、负责任的人工智能隐私保护实施建议，让下一代人工智能用的放心。

本白皮书由中国联通研究院主笔，中国联通集团网络与信息安全部、中国软件评测中心（工业和信息化部软件与集成电路促进中心）、数据安全关键技术与产业应用评价工业和信息化部重点实验室、中国计算机行业协会数据安全专业委员会、三六零数字安全科技集团有限公司、中兴通讯股份有限公司联合编写。

编写组成员（排名不分先后）：

总策划：苗守野、李浩宇、叶晓煜

编委会：徐雷、陶冶、李慧芳、孙艺、陈泱、曹咪、傅瑜、唐刚、张德馨、白利芳、李尤、林青、杨晓琪、黄英男、李泽村、唐会芳、王雨薇、王继刚、陈靖

一 人工智能全生命周期隐私风险分析

在智能化变革的今天，技术的发展和变化都会对人们的生活带来空前的改变，互联网和大数据等相关技术的更迭加速了人工智能应用的步伐，使得人们的生产生活方式悄然的有了新的活力。技术的发展给社会带来机会的同时也同样不能忽略它的弊端和随之带来的一系列负面影响，尤其在今天这样无隐私的透明化的时代，人们在让渡出自己的部分权利来交换智能应用所带来的便利服务时，隐私泄露是人们必须要直面的问题。

最近几年，有关隐私受到侵犯的案件一再发生。例如，Facebook 未经用户允许将用户个人信息泄露给剑桥分析公司用于非正当目的，同时其利用网民的浏览行为来精准的投放广告，剑桥大学心理测量学中心从用户对哪些帖子和新闻进行阅读和点赞，来分析出每个人的性别、性取向、个性外向还是内向等，美食外卖企业“饿了么”、“大众点评”、“美团”会利用算法推送一些推荐食物和餐馆帮用户做出饮食决定，自动驾驶技术让人们可以轻松的出行，高德地图、百度地图等智能导航系统减少了人们寻找路线的时间和精力，ChatGPT 和其他生成式人工智能工具可以提高用户交互体验、提高员工的创作和办公效率，但这些信息都以数据的形式存储了下来，并被企业或其他主体收集和利用，一些智能手机应用甚至过度的收集并违规使用个人信息，使得个人隐私信息面临被泄露或被窃取的风险。

可以看到人工智能的普及与滥用使其面临越来越多的隐私与安

全威胁，社会各界也逐渐加大了对隐私风险的分析和隐私保护的关注度。**从隐私保护角度，数据隐私性、模型保密性、模型完整可用性是用户和服务提供商最为关心的问题。**因此，本章将先从数据、模型这两个不同的方面来揭示人工智能面临的通用隐私威胁。同时，由于生成式人工智能（Generative Artificial Intelligence，简称生成式AI）技术的快速发展和应用给人们带来了巨大的想象空间，但也增加了新的AI隐私风险，本章还将对生成式人工智能隐私风险进行揭示。

1.1 人工智能通用隐私风险

1.1.1 数据采集存储流通阶段隐私风险

数据不正当收集风险。人工智能算法尤其是在深度学习的开发测试过程中，需要大量训练数据作为机器学习资料、进行系统性能测试。在网上公开数据源和商务采购时，由于目前数据共享、交易和流通的市场化机制不健全，存在非法数据、买卖数据、暗网数据等不正当收集行为和一些未经授权的收集行为，这些数据缺乏用户知情同意，实际并没有获得数据的采集权限，很容易泄露用户隐私。

数据过度收集风险。在无人驾驶、智能家居、智慧城市等典型应用场景中，数据主要通过在公开环境中部署各类传感器或终端，并以环境信息为对象进行无差别、不定向的现场实时采集。现场采集由于难以提前预知采集的数据对象和数据类型，因此，在公开环境尤其是公共空间进行现场采集时，将不可避免地因采集范围的扩大化而带来

过度采集问题。比如，在智能网联汽车的无人驾驶场景中，自动驾驶汽车的传感器需要采集街景数据来支持智能驾驶系统的决策从而控制汽车行驶，但是这种无差别的街景数据采集必然会采集到行人的个人数据，其中包括行人的人脸数据等个人敏感信息，造成行人的隐私泄露风险，甚至还可能会采集到路边的重要基础设施、地理位置信息、军事营区等重要数据，给国家安全带来风险。

数据存储隐私泄露风险。一方面，在对数据进行保存时，如果没有对数据采取技术手段进行安全防护，容易被非法需求者通过网络攻击等黑客行为进行隐私数据窃取。另一方面，在数据存储过程中，由于对数据没有明确的隐私界定与标注，如果使用者无意中将涉及隐私的数据用于公开的人工智能训练分析中，个人隐私将在不经意间被泄露。再另一方面，在人工智能数据处理使用的过程中，涉及众多数据处理、保存步骤，对于种类多、数据量大的数据集，处理、保存操作难以规范与监管，潜藏被非法使用者利用、拷贝等安全隐患。

数据流通隐私泄露风险。由于大量人工智能企业会委托第三方公司或采用众包的方式实现海量数据的采集、标注、分析和算法优化，数据将会在供应链的各个主体之间形成复杂、实时的交互流通，可能会因为各主体数据安全能力的参差不齐，产生数据泄露或滥用的风险。此外，在全球数字经济发展不均衡的大背景下，大型科技巨头将人工智能的数据资源供给、数据分析能力、算法研发优化、产品设计应用等环节分散在不同的国家，数据跨境流动的场景也会对国家安全和个人信息保护造成不可控的隐私风险。

1.1.2 模型训练与推理阶段隐私风险

（1）模型训练阶段数据污染风险

数据污染有失数据真实性。人工智能模型依赖海量数据，相比数据集大小，研发工程师更关注数据质量。知名学者吴恩达提出“80%的数据+20%的模型=更好的机器学习”，而数据污染和错误将降低模型精度，数据偏差和噪声将降低模型的泛化性和可靠性。数据是连接现实空间和虚拟空间的桥梁，如果数据质量出现问题，如数据内容失真、数据标注错误、数据多样性有限，则无法反映现实世界的真实情况，在此基础上建立的人工智能模型便会出现偏差，导致预测结果偏差或错误，甚至导致种族歧视或者性别歧视偏见，出现“垃圾进、垃圾出”的现象。如今的生成式 AI 模型也因静态数据的时效性，导致生成内容存在过时或者错误现象。

数据投毒攻击风险。数据投毒是指通过在训练数据集中故意添加污染数据（如错误样本或恶意样本），导致训练出来的模型在决策时发生偏差，从而影响模型的完整性或可用性。人工智能模型在训练过程中容易受到数据投毒攻击，攻击者可以通过实施标签翻转或添加后门等恶意行为来破坏训练数据的正确性。从而破坏模型决策的正确性。近年来，对人工智能模型的数据投毒问题已使得多个世界知名公司遭受重大负面影响，并造成了十分严重的后果。例如：美国亚马逊公司因其 Alexa 智能音箱学习了网络不良信息，发生了引导用户自杀的恶意行为。因此，训练数据的正确性问题已成为阻碍人工智能发展的重

大问题。

（2）模型推理应用阶段隐私风险

隐私被推理风险。人工智能模型推理产生的信息可能会间接暴露用户隐私。一方面，在对数据进行深度挖掘与分析时，所得到的结果数据可能将用户的个人隐私一并挖掘出来，并进一步进行数据应用，从而使数据中隐藏的个人隐私信息进行暴露。另一方面，在对去标识化的个人信息和行为模式进行融合和关联分析时，可能推理出与个人隐私相关的信息，比如政治倾向、财务状况等。

成员推理攻击风险。成员推断攻击是一种数据隐私攻击方法，该攻击通过判断输入数据是否是目标模型的训练数据来达到攻击效果。具体来说，攻击者不需要获取模型结构、模型参数、训练方法等，只需要向模型输入数据，从模型输出的置信度即可判断该输入是否为训练集中的数据。尤其对于过拟合模型，训练集数据与非训练集数据的置信度表现会有明显差异，如果目标攻击模型使用了个人敏感信息进行模型训练，成员推理攻击就会造成模型训练集中这部分敏感数据的泄漏。

模型逆向攻击风险。模型逆向攻击是一种通过还原训练数据造成数据隐私泄漏的攻击方法。攻击者可以在没有训练数据的情况下，通过模型输出的置信度不断调整输入数据，最终近似获得训练集中的数据。这一攻击如果使用在人脸识别系统、指纹识别系统等，则会造成用户生物识别信息的泄漏，例如随机构建一张图片，人脸识别模型给出用户名与置信度，结合置信度不断调整图片，最终就有可能将训练

集中的人脸恢复出来。

模型提取攻击风险。模型提取攻击是一种可以造成模型保密性被破坏与知识产权被侵犯的攻击方法。该攻击通过模型预测结果反推模型具体参数和结构，以达到训练出一个与目标模型相似度极高的模型的过程。企业训练一个机器学习模型往往要花费大量金钱，投入大量人力，通过模型提取攻击，攻击者可以在对模型不掌握任何信息的前提下，仅通过模型的输入与输出来训练一个替代模型，一定程度上侵犯了企业的知识产权，破坏了企业的商业模式。

对抗样本攻击风险。对抗样本攻击是一种在模型推理阶段破坏模型完整性的攻击方法，其通过对人工智能模型的输入数据加入微小噪声，以欺骗模型做出错误预测。人工智能模型并不总是稳定和可靠的，攻击者对输入数据加入难以察觉的细小扰动，可以使模型产生意想不到的错误。例如，对于一个猫和狗的图像分类器，攻击者可以在猫的照片上进行微调，使分类器错误地将该图分类为狗。对抗样本攻击的出现给人工智能模型的准确性和鲁棒性带来了挑战，这种攻击可能对身份识别系统这类关键应用产生严重影响，因此也对个人隐私产生极大威胁。

提示注入攻击风险。模型面临提示注入攻击，尤其对于语言模型，当模型无法区分系统指令与不受信任的用户输入指令时，用户攻击者就有机会绕过模型限制并违反模型的指导原则来劫持模型输出，注入攻击就有可能发生。这种攻击的思路是，通过注入指令来劫持模型输出，使模型忽略原始指令并执行注入的指令，从而偏离其原始行为，

造成信息泄漏或者生成违规内容等问题。提示泄露攻击是提示注入攻击的一种形式，该攻击用于泄露可能包含未经公开的机密或专有信息的提示的攻击。微软公布的 NewBing 对话机器人就被使用提示注入攻击发现了其聊天的初始提示，该提示通常对用户隐藏。

1.2 生成式人工智能隐私风险

随着生成式人工智能（Generative Artificial Intelligence，简称生成式 AI）技术的发展，AI 模型开始具备更通用和更强的基础能力，并从计算智能、感知智能进一步迈向认知智能。但同时，AI 模型能力的提升，也带来了新的隐私风险。生成式人工智能隐私风险可分为生成式人工智能内生隐私风险和生成式人工智能滥用导致的衍生风险。

生成式人工智能内生隐私风险主要是在使用生成式 AI 模型的过程导致的数据泄漏风险。一方面，当用户与以 ChatGPT 为代表的生成式 AI 模型进行问答交互时，有时会输入包含隐私数据的 prompt 指令，而这些指令都会被无差别地记录并存储。由于缺乏对相应数据的访问限制，这些指令中包含的用户隐私存在被泄漏的风险。另一方面，生成式 AI 模型通过对海量训练数据的学习来生成新的数据，且目前以 ChatGPT 为代表的生成式 AI 模型基本属于重组式创新，在进行前向推理时，模型存在将训练数据中包含的隐私数据变换、拼接后生成输出，暴露给无关用户的风险。

生成式人工智能滥用导致的衍生风险主要是指在缺乏约束和监

管的情况下，生成式 AI 技术可能被用于深度伪造虚假信息，从而进一步危害用户隐私安全。例如，ChatGPT 由于其强大的生成能力，可能被不法分子用于生成钓鱼短信和邮件，一些多模态大模型也可能被用于生成用户语音、图像和视频，进行诈骗攻击。这些行为不仅侵犯他人的肖像权、隐私权、名誉权，还可能被用来实行勒索诈骗等违法犯罪活动。

二 人工智能隐私保护法规政策和标准化现状

2.1 人工智能隐私保护国内外法律法规和政策

2.1.1 国内情况

随着《数据安全法》《个人信息保护法》与《网络安全法》三法的落地实施，我国数据安全领域法律框架基本搭建完毕，在人工智能安全领域，我国目前尚未对人工智能治理进行综合立法，但已有较多针对数据安全与信息保护的专门立法实践。

（1）法律层面

《个人信息保护法》确立了以“告知-同意”为核心的个人信息处理规则，详细规范了平台企业的大数据使用和用户画像行为，约束处理个人信息的行为。2021 年 1 月 1 日，《中华人民共和国民法典》正式施行，针对人工智能隐私相关问题，《民法典》规定人工智能技术的使用需要遵守相关法律法规，保护个人信息安全，并对其造成的损害承担相应的法律责任。

（2）部门规章层面

2017 年国务院印发《新一代人工智能发展规划》，其中明确指出要“确保人工智能安全、可靠、可控发展”“形成人工智能算法与平台安全性测试评估的方法、技术、规范和工具集”。2021 年 12 月，网信办发布《互联网信息服务算法推荐管理规定》，该管理规定主要对各类算法技术的适用场景和企业使用算法时需恪守的强制性义务及违反后的惩罚措施做了详细规定，明确指出算法推荐服务提供者应当建立数据安全和个人信息保护管理制度和技术措施。此外，最高人民法院于 2022 年 12 月发布了《关于规范和加强人工智能司法应用的意见》，要求人工智能建设要确保国家秘密、网络安全、数据安全和个人信息不受侵害。

2023 年 7 月 13 日，国家网信办联合国家发展改革委、教育部、科技部、工业和信息化部、公安部、广电总局发布了《生成式人工智能服务管理暂行办法》。该办法的出台既是促进生成式人工智能健康发展的重要要求，也是防范生成式人工智能服务风险的现实需要。办法明确要求，参与生成式人工智能服务安全评估和监督检查的相关机构和人员对在履行职责中知悉的个人隐私和个人信息应当依法予以保密，不得泄露或者非法向他人提供，且要尊重他人合法权益，不得危害他人身心健康，不得侵害他人隐私权和个人信息权益。

另外，工信部发布了《“十四五”信息化和工业化深度融合发展规划》，对智能产品在工业、交通、医疗、教育等重点行业的应用推广进行了系统性的部署。2021 年 9 月，科技部发布的《新一代人工

智能伦理规范》围绕管理、研发、供应、使用和组织五个环节提出了18项具体规范，将抽象的伦理原则以具体规范的形式融入了人工智能全生命周期，推动形成具有广泛共识的人工智能治理框架和标准规范。市场监督管理总局组织起草的《互联网平台分类分级指南（征求意见稿）》和《互联网平台落实主体责任指南（征求意见稿）》也于2021年年末开始向社会征求意见，有望通过合理划分平台等级，推动超大型平台承担更多责任与义务，形成更为细致合理的平台责任规范。其他部委如人民银行、人力资源社会保障部、卫健委等也在具体领域积极出台政策文件，共同促进人工智能治理在我国落地生根。

（3）地方层面

2022年，上海、深圳等地发布促进人工智能产业发展相关条例，湖北、四川等地发布了省人工智能相关发展规划。2023年5月，北京市人民政府印发《北京市加快建设具有全球影响力的人工智能创新策源地实施方案（2023-2025年）》，提出充分发挥北京市在人工智能领域的创新资源优势，持续提升全球影响力，进一步推动人工智能发展，到2025年，人工智能核心产业规模达到3000亿元，持续保持10%以上增长，辐射产业规模超过1万亿元。此外，《深圳经济特区人工智能产业促进条例》于2022年9月6日正式公布，《条例》创新性地规定市政府应当设立人工智能伦理委员会，并明确指出从事人工智能研究和应用的组织或者个人，应当遵守人工智能伦理安全规范，不得从事危害国家安全和社会公共利益、侵犯个人隐私和个人信息权益、实施价格歧视或者消费欺诈等七类行为。

从国内的发展情况来看，党和国家高度重视人工智能时代下隐私保护面临的新挑战，近年来人工智能隐私保护相关的法律法规、政策正在紧锣密鼓的制定，步伐夯实稳健，我国人工智能隐私保护顶层设计正在不断构建。

2.1.2 国外情况

近年来人工智能在深度学习、人机协同等关键领域呈现出蓬勃的发展态势，但其日益凸显的隐私与安全问题也为人类权益、社会公平和国际格局带来了风险和挑战。为此，世界各地的监管机构持续致力于人工智能系统的规范使用，近年来起草、讨论、通过、发布和生效了大量法律法规、原则性文件与行政命令。

（1）欧盟和英国

欧盟和英国关于监管人工智能的建议在范围和方法上有明显差异，并考虑了部署人工智能系统的组织可以采取哪些步骤，以确保这些建议从最初阶段能符合 GDPR（通用数据保护条例）。该条例被称为史上最严格的条例，于 2018 年生效，规定了数据控制者和处理者的责任和义务，设立了数据保护官，增大了处罚力度。人工智能的设计和运行需要获取的所有数据，都受到该条例的有效约束。

欧盟选择了一个广泛的立法框架，而英国选择依靠部门监管机构，并将推行一项侧重于促进技术进步和保持人工智能“超级大国”地位的议程。

英国人工智能战略于 2021 年 9 月 22 日公布(《人工智能战略》)。

该计划是一项十年计划，英国打算通过该计划“促进人工智能的商业应用，吸引国际投资，培养下一代科技人才”，同时将自己定位为人工智能治理的全球领导者。

根据《人工智能战略》，人工智能代表了“最高水平的经济、安全、健康和福祉优先”。英国政府认为“在人工智能领域保持竞争力对我们的国家雄心至关重要……”，同时，英国人工智能办公室（Office for Artificial Intelligence）打算发布一份关于人工智能促进创新立场的白皮书（人工智能白皮书）。作为人工智能白皮书的前奏，英国政府于 2022 年 7 月 20 日发布了一份题为《建立有利于创新的方法来监管人工智能》的政策文件（《人工智能政策文件》）。这份人工智能政策文件阐述了英国政府的愿景，即建立一个“相称、轻触和前瞻性”的监管框架，使英国能够跟上全球竞争对手的步伐。同时英国中央数字办公室等机构在 2021 年 11 月发布算法透明度标准，涵盖了数据要求、透明度模板和行动指南等内容，为政府部门和公共机构利用算法进行决策提供支持。

与英国相反，欧盟选择了广泛的立法方式，正在敲定发布世界首部人工智能法案。作为人工智能监管在全球范围内的“第一推动者”，2021 年欧盟继续采用全面立法的方式规范人工智能在各行业的应用，并于 4 月 21 日发布《人工智能法》提案，这是世界范围内首部对人工智能进行综合性立法的法案。该法案将人工智能应用划分为不可接受的风险（Unacceptable risk）、高风险（High risk）、有限风险（Limited risk）和极低风险（Minimal risk）四类。其中，涉及操

纵人类行为的“潜意识技术”，利用儿童和残疾人脆弱性或可能影响社会信用评分的人工智能应用被认为具有不可接受的风险，法案禁止该类应用上市；涉及公共基础设施、社会福利、医疗服务、教育培训等领域的人工智能应用被认为具有高风险，法案提出应在其上市前进行充分的风险评估，确保算法偏见最小化、活动记录可追溯，并引入合适的人力监管措施以最大限度地减少风险；聊天机器人等对人类生存安全与基本权利具有有限风险的人工智能应用，法案仅明确了其向用户公开透明的义务；电子游戏和垃圾邮件识别软件等对人类安全与权利不产生影响或者影响甚微的极低风险的人工智能应用，法案并未进行干预。

2019 年 4 月，欧盟委员会发布了正式版的人工智能道德准则《可信赖人工智能的伦理准则》，提出了实现可信赖人工智能（Trustworthy AI）全生命周期的框架。该准则提出，可信赖 AI 需满足 3 个基本条件：合法的（lawful），即系统应该遵守所有适用的法律法规；合伦理的（ethical），即系统应该与伦理准则和价值观相一致；稳健的（robust），即从技术和社会发展角度来看，可信赖 AI 必须是鲁棒的。2022 年 6 月生效的《数据治理法案》是《欧洲数据战略》框架下的第一份立法草案。《数据治理法案》强调规则创新，鼓励数据共享、提高数据利用效率，进而让数据资源的流转利用服务更高的公共政策目标：一是建立公共部门持有数据的再利用机制，二是建立框架以促进数据中介机构的发展，三是对于数据利他行为做出规范化的引导。2023 年 6 月 14 日，《人工智能法案》授权草案在欧

洲议会高票通过，《人工智能法案》授权草案的一个突出特点是注重基于风险来制定监管制度，以平衡人工智能的创新发展与安全规范。草案严格禁止对人类安全造成不可接受风险的人工智能系统，包括部署潜意识或有目的操纵技术、利用人们弱点或用于社会评分的系统，并扩大了人工智能高风险领域的分类，将对人们健康、安全、基本权利或环境的危害考虑在内。

（2）美国

2016 年 10 月，美国政府发布《国家人工智能研究和发展战略规划》和《国家人工智能研究和发展战略规划》（2016）两项报告，提出实施“人工智能公开数据”计划。2019 年 2 月，美国总统特朗普签署《人工智能倡议》发展规划，进一步指示加强联邦政府、机构的数据、算法和计算机处理资源对人工智能研发人员和企业的开放。2019 年 6 月，美国政府发布《国家人工智能研发与发展战略计划》（2019），新版本中要求所有机构负责人审查各自联邦数据和模型，注重保护数据安全、隐私和机密性。2020 年 1 月，美国政府发布《人工智能应用监管指南》，要求联邦机构继续促进技术和创新进步的同时保护美国的技术、经济和国家安全、隐私、公民自由和其他美国价值观。

2019 年，美国颁布了《2019 年国防授权法案》，依据此授权法，美国成立国家人工智能安全委员会（National Security Commission on Artificial Intelligence），研究人工智能和机器学习方面的进展，以及它们在国家安全和军事方面的潜在应用。此外，依据《2019

年国防授权法案》，美国国防部创建了联合人工智能中心（JAIC），作为开发和执行总体人工智能战略的责任机构。2021 年 1 月，美国正式颁布《2020 年国家人工智能倡议法案》，旨在确保美国在全球人工智能技术领域保持领先地位。该法案强调要进一步强化和协调国防、情报界和民用联邦机构之间的人工智能研发活动；同时，设立国家人工智能倡议办公室，承担“监督和实施美国国家人工智能战略”等职责。2020 年 5 月，《生成人工智能网络安全法案》出台，该法案要求美国商务部和联邦贸易委员会明确人工智能在美国应用的优势和障碍，调查其他国家的人工智能战略，并与美国进行比较；评估找出对应的供应链风险及解决方案，制定国家人工智能战略的建议。2022 年 6 月 3 日，美国参议院和众议院发布了《美国数据隐私和保护法》（the American Data Privacy and Protection Act, ADPPA）的草案，该立法草案是第一个获得两党两院支持的美国联邦全面隐私保护提案。这项具有分水岭意义的隐私保护法案，将为数据隐私保护引入一个美国联邦标准。

2023 年 5 月下旬，拜登政府采取了几项额外措施，进一步明确其人工智能治理方法。同时，美国白宫科技政策办公室（OSTP）发布了修订后的《国家人工智能研发战略计划》，以“协调和集中联邦研发投资”。OSTP 还发布了一份信息征询书，征求关于“减轻人工智能风险，保护个人权利和安全，利用人工智能改善生活”的意见。

2023 年 10 月下旬，拜登政府签署了一项旨在改善人工智能安全的行政命令，这是美国迄今为止最全面的人工智能规则和指南，具有

里程碑意义。行政令规定了“人工智能安全新标准”、“保护美国公民隐私”等内容，明确要求开发高级人工智能系统的公司应与美国政府分享安全测试结果、加强对隐私保护技术的支持和研究等。

（3）日韩

2022 年 4 月 22 日，日本政府在第 11 届综合创新战略推进会上正式发布《人工智能战略 2022》，作为指导其未来人工智能技术发展的宏观战略。该战略提出推进数据合作和标准化，防止数据偏差、人工智能技术滥用的风险；确保数据真实性和数据所有人的知情权；构建数据存储的基础设施，确保供应链的安全性。2023 年 4 月 21 日，日本政府决定设立新的“战略会议”及“AI 战略小组”，负责讨论与人工智能相关的国家战略。对于正在迅速普及的聊天机器人 ChatGPT 等整个人工智能领域，上述“战略会议”将发挥指挥塔作用，指明政策的基本方向。针对人工智能方面课题，该会议将从促进应用、研究开发和强化规则两方面进行讨论。除精通人工智能技术的学者和研究人员外，法律相关领域的专家和政府相关人士也将加入上述“战略会议”。

2020 年 1 月，韩国科技部公布 2020 年度工作计划，正式启动《人工智能国家战略》，意图推动韩国从“IT 强国”发展为“AI 强国”，计划在 2030 年将韩国在人工智能领域的竞争力提升至世界前列。其中提到发展人工智能，必须对数据进行收集以及利用。数据作为人工智能发展的核心，《人工智能国家战略》表示数据安全具有多种隐患，需要对相关法律法规进行修改。2020 年 12 月 24 日，韩国科学和信

息通信技术部和国家事务协调办公室共同发布了人工智能立法路线图，包括了 11 个领域的 30 项立法任务，旨在奠定人工智能时代的法律基础。韩国认识到，在大力推动人工智能发展的同时，需要尽快推动现有立法的变革。为此，韩国科学信息通信技术部成立了立法研究小组，组织法学、人文、社会科学和哲学等多领域的人员，共同草拟这个立法线路图。算法决策对政治、社会、经济和文化具有重大影响，因此有必要确保算法的透明性和公平性，以建立应用人工智能技术的信任基础。2020 年 12 月 22 日，韩国科学与信息通信技术部发布了《国家人工智能伦理标准》，通过制定人工智能伦理规范，打造安全的人工智能使用环境，为韩国未来人工智能发展和负责任使用提出了方向和指引。

随着人工智能技术的快速发展和应用，个人隐私保护面临着不断演变的挑战。从世界范围观察，近年来各国尤其是发达国家正不断加强人工智能隐私保护法规政策的制定和执行，建立有效的监管机制，以确保个人隐私权得到充分的保护，为人工智能隐私保护相关技术在全球范围内的应用及全球化合作交流提供保障和支持，逐渐在人工智能发展中平衡创新与隐私保护的需求。

2.2 人工智能隐私保护国内外相关标准化情况

2.2.1 国内相关标准研究情况

我国人工智能领域标准建设由国家牵头统一布局，2020 年 7 月，

国家标准委、中央网信办、国家发展改革委、科技部、工信部联合印发《国家新一代人工智能标准体系建设指南》，旨在加强人工智能领域标准化顶层设计、推动人工智能产业技术研发和标准制定、促进产业健康可持续发展。在标准体系结构中，“安全/伦理”单独成块，贯穿于其他部分，为人工智能建立合规体系。指南明确指出，安全与隐私保护标准包括基础安全、数据、算法和模型安全、技术和系统安全、安全管理和服务、安全测试评估、产品和应用安全六个部分。

我国人工智能安全国家标准，主要归口于全国信息安全标准化技术委员会（简称“信安标委”或 TC260）。截止 2023 年 7 月，已有处于征求意见阶段的《信息安全技术 人工智能计算平台安全框架》（20230249-T-469）以及处于批准阶段的《信息安全技术 机器学习算法安全评估规范》（20211000-T-469）两项人工智能安全国家标准，另有基因识别、声纹识别、步态识别、人脸识别数据安全要求四项国家标准，规定了生物特征识别数据处理的基本安全要求、全生命周期中的安全要求以及应用场景中的安全要求。此外，我国企业、高等院校等也积极参与国际安全标准制定工作，在人工智能安全领域，清华大学在 IEEE 标准协会牵头立项了《生成式预训练 AI 模型的安全性和可信性技术要求》（P7018）国际标准。在行业人工智能安全领域，已有面向特定行业的人工智能算法安全、算力安全处于征求意见阶段，例如《信息通信领域人工智能算法安全评估指南》、《电信领域人工智能算法安全要求》、《互联网深度合成信息服务标识通用安全要求》、《算力网络计算节点安全能力要求》等。

2.2.2 国外相关标准研究情况

2017 年 10 月 ISO/IEC JTC1 在俄罗斯召开会议，决定新成立人工智能的分委员会 SC42，负责人工智能标准化工作。SC42 目前已成立 5 个工作组，包括基础标准（WG1）、大数据（WG2）、可信赖（WG3）、用例与应用（WG4）、人工智能系统计算方法和计算特征工作组（WG5）。其中，SC42 WG3 人工智能可信标准组已经开展人工智能风险管理、人工智能的可信度概览、算法偏见、伦理等标准研制。

IEEE 标准协会主要聚焦于涉及人工智能伦理道德规范的标准研究，已经发布了多项人工智能伦理标准和研究报告。IEEE P7002《数据隐私处理》指出如何对收集个人信息的系统和软件的伦理问题进行管理，将规范系统/软件工程生命周期过程中管理隐私问题的实践，也可用于对隐私实践进行合规性评估（隐私影响评估）。

2019 年 5 月 1 日，美国国家标准与技术研究院（NIST）发布人工智能标准化计划纲要，将人工智能数据安全与隐私保护相关标准化纳入人工智能可信标准领域。2022 年 2 月，全球移动通信系统协会（GSMA）发布了人工智能安全指南第一版，概述了人工智能应用潜在的风险并提供了相对应的防护措施。2023 年 1 月 26 日，美国国家标准与技术研究院（NIST）为加强对人工智能（AI）相关个人、组织和社会风险的管理，通过与私营和公共部门合作，制定了《人工智能风险管理框架》，该框架将可信度考量纳入设计、开发、使用和评估 AI 产品、服务和系统中，并基于其他机构的 AI 风险管理工作，确保

制定过程的公开、透明。除此之外，与该框架相关的其他资源也包含在《人工智能风险管理框架》中。

在人工智能隐私保护标准化方面，国内外还存在一定的差异。例如，国际标准化组织（ISO）发布的标准具有全球通用性，在全球范围内均适用。而国内印发的相关标准主要适用于我国境内的企业和组织，与国际标准存在一定的差异。此外，由于在文化和社会背景等方面存在差异，国内外对于人工智能隐私保护的要求和做法也有所不同，例如欧洲的 GDPR 在个人数据保护方面比较严格，而我国则更注重技术标准和应用指南的制定。

为了解决这些差异，国际标准化组织（ISO）和国内标准化工作组间已经开始展开交流与合作，以推动国际标准与国内标准的对齐。通过积极开展国际合作与对话，可以加强国内外标准的互认和对齐，形成更加全面和统一的人工智能隐私保护体系标准。

三 人工智能隐私保护技术和平台

实践中，针对人工智能隐私保护和数据安全问题，科技公司/金融科技在尝试“以子之矛攻己之盾”，即运用技术手段解决技术带来的挑战。目前，常用的人工智能隐私保护技术包括管控技术、数据加密技术、攻击防御技术以及一些新兴技术。在此基础上，业界推出了隐私保护机器学习平台和人工智能安全检测平台。

3.1 人工智能隐私保护技术

3.1.1 人工智能隐私保护管控技术

权限管理是根据预设的规则或者策略限制用户访问被授权的资源，可以保护系统安全和数据完整性。访问控制是一种确保数据处理系统的资源只能由经授权实体以授权方式进行访问的手段。对人工智能系统实施访问和使用权限控制机制，只有授权人员可以访问和使用特定的数据，可以确保人工智能数据与模型的隐私与安全。

分类分级保护可以理清楚保护需求及重点，并针对不同等级，采取相应的保护措施。这种精细化的、分级化的管控手段，有助于降低系统隐私泄露带来的负面影响。智能化程度越高的人工智能应用，数据隐私风险越高。因此，可以根据人工智能应用场景和功能，对人工智能应用进行分类分级，然后定制差异化的人工智能隐私保护机制。例如，针对初级的基于人工智能技术的数据分析，可按权限申请数据调取和共享，保证数据可信共享。针对智能化程度更高的生成式人工智能应用，可采用可溯源的解决方案，应对图片、视频等生成内容进行标识，发现违法内容及时采取处置措施等。

3.1.2 人工智能隐私保护数据加密技术

差分隐私（Differential Privacy）是一种数据匿名化技术，其最早是针对统计数据库的隐私泄露问题提出的一种隐私定义。该定义要求数据集的计算处理结果对于具体某记录的变化是不敏感的，即攻

击者无法通过观察计算结果来获取准确的个体信息。差分隐私保护技术通过添加噪声使敏感数据失真但同时保持某些数据或数据属性不变，来保证处理后的数据仍然可以保持某些统计方面的性质，以便进行数据挖掘等操作。Laplace 机制和指数机制是两种基础的差分隐私保护实现机制，分别适用于对数值型结果的保护和对非数值型结果的保护。近年来，基于机器学习的数据发布和数据挖掘技术成为热点研究方向。为了保护机器学习应用中的用户数据隐私，研究者将差分隐私技术和机器学习算法结合，提出了基于差分隐私的机器学习隐私保护方案，主要包括基于输入扰动的隐私保护方案、基于中间参数扰动的隐私保护方案、基于目标扰动的隐私保护方案和基于输出扰动的隐私保护方案。机器学习中部署差分隐私技术时仅需要通过随机化和利用随机噪声扰动数据，因此并不会带来过多额外的计算开销。

同态加密（Homomorphic Encryption）是一种加密形式，允许用户直接对密文进行特定的代数运算，得到的数据仍是加密的结果，且与对明文进行同样的操作再将结果加密一样。同态加密技术最早用于对统计数据进行加密，由算法的同态性保证了用户可以对敏感数据进行操作但又不泄露数据信息。同态加密可以进一步分为部分同态加密、稍微同态加密和全同态加密。其中，部分同态加密技术仅支持对密文进行部分形式的计算，以 BGN 算法为代表的稍微同态加密支持有限次数的计算，全同态加密则可以对密文进行无限次数的任意同态操作。在机器学习领域，为了实现用户数据机密性，需要结合加密技术对数据进行保护。但传统的密码学方法计算复杂性非常大，而全同态加密

由于允许在加密数据上执行任意操作且无需解密，在计算成本上优势明显。基于同态加密的机器学习隐私保护方案分为无需多项式近似的同态加密隐私保护方案和基于多项式近似的同态加密隐私保护方案。

多方安全计算（Secure Multi-Party Computation，MPC）是图灵奖得主姚期智教授于 1982 年提出的一个密码概念，能够在加密值上进行计算。通过使用 MPC，多个数据库可以联手做计算，却又不透漏各自的数据。MPC 主要基于混淆电路、密密分享与不经意传输技术，汇聚多方数据，以实现高质量的学习，同时能保证各方数据隐私。目前，MPC 对金融科技、生物识别、医疗、保险等 AI 应用中非常有用。例如，在金融科技领域，可以应用安全多方计算，实现跨实体欺诈分析，实现数据的安全处理和共享。对于已有的生物识别系统，可以应用安全多方计算，使得生物特征在密文状态下进行计算，并将最终结果恢复成明文，从而包含原始生物特征的隐私与安全。

3.1.3 人工智能隐私保护攻击防御技术

针对人工智能模型训练和推理阶段面临的数据与模型隐私安全风险，研究者根据不同的攻击类型提出了相应的防御措施。针对数据投毒攻击的防御，一般考虑通过鲁棒性机器学习和数据清洗来改变正常训练数据的分布。针对成员推理攻击的防御，研究人员发现通过在模型中添加正则项或者使用 model stacking 可以显著减少成员推理攻击。针对模型逆向攻击的防御，常见的方式是利用差分隐私技术来实现对数据的隐私保护，也有研究者提出利用联邦学习建立虚拟共有

模型进行多方共同训练，从而降低本地训练数据泄露的风险。针对模型提取攻击的防御，一种最直接的方式是对模型参数或输出结构进行近似处理，也有研究者利用模型水印技术来保护模型数据的知识产权，降低模型被盗用的风险。针对对抗样本攻击的防御已经有较多方法：直接对抗训练是将对抗样本及正确标签重新输入到模型中进行重训练，梯度掩模通过隐藏梯度使基于梯度的对抗样本攻击失效，对抗样本检测即直接检测是否存在对抗样本。

近期，随着大语言模型的快速兴起和应用，研究者提出了提示攻击防御方法和生成内容检测过滤防御方法，预防大模型的提示攻击威胁和生成内容隐私泄露。对于提示注入攻击防御，一种简单直接的提示注入攻击防御策略就是将防御策略添加到指令中，增加指令的鲁棒性来强制执行期望的行为。常用的技术有调整提示位置、用特殊符号标识等。同时，研究者提出构建提示检测器对提示进行检测、分类或过滤，以防止敏感和有害的提示输入。目前，OpenAI 的 ChatGPT、微软的 NewBing 等，都采用了这种防御策略。对于生成内容过滤防御，其目标是识别并避免输出隐私内容。生成内容检测方法主要包括构建规则集合的方法和构建审核模型的方法。通过这些方法先对输出内容进行检测和识别，再并根据检测识别结果进行隐私内容屏蔽和过滤，可以避免生成敏感信息和风险内容。

3.1.4 人工智能隐私保护新兴技术

近年来，研究者提出了一些新兴的隐私保护技术，如联邦学习技

术、区块链技术、可信执行环境、机器遗忘技术、模型数字水印技术等，这些技术由于其创新性和实用性，吸引了学术界和产业界的极大关注。

联邦学习（Federated Learning）由 Google 在 2016 年提出，是一种多个参与方在不交互数据的情况下，通过安全机制交互模型参数信息或者梯度信息，从而达到协同训练效果的分布式机器学习方法。与传统的集中式存储与训练模型相比，联邦学习具有“去中心化”的特点，可以实现数据隐私保护与数据共享分析的平衡，实现“数据可用不可见”。常用的联邦学习技术按照数据集合维度可以分为三大类，包括横向联邦学习、纵向联邦学习和联邦迁移学习，其中代表算法和架构有 FedAvg 算法，微众银行的 FATE 架构。目前联邦学习技术被 Facebook、亚马逊、苹果等科技公司广泛使用，国内金融科技企业和高校也在发力数据隐私安全技术，第四范式公司也将迁移学习算法应用到公司核心产品“先知”平台，并在医疗领域实现落地应用。在通信领域，可以利用联邦学习和各网络设备的数据联合训练模型，优化网络站点规划，另外还可以催生以通信运营商为中心的跨领域生态合作。

联邦学习与多方安全计算有些类似，都可以保障多参与协作时的数据隐私，但联邦学习作为新兴的隐私保护范式，面向机器学习模型，通过原始数据在本地训练模型，只交互模型的中间计算结果，实现“数据可用、不可见”、“数据不动、模型动”，而多方安全计算面向数据，通过构建一系列基础运算操作，将多方原始数据转换为密文后实

现流动和协同计算。

区块链（Blockchain）是随着数字加密货币而逐渐兴起的一种去中心化的分布式存储架构与计算范式。区块链按照时间顺序将数据区块以顺序连接的方式形成一种链式的、分布式数据结构，每个区块头保存前一个区块的哈希地址以保证各个区块相连。区块链通过共识协议在分布式节点上生成和同步数据，保证数据存储的一致性，利用密码学手段保证了数据不可篡改和不可伪造，借助可编程脚本实现合约条款的自动执行和数据操作。通过综合运用数据加密、时间戳、分布式共识、P2P 通信和经济激励等手段，区块链突破了传统中心式架构的缺陷，具有去中心化、去信任化和防篡改的安全特性。在数据隐私保护方面，区块链能够有效避免中心化平台产生的隐私泄露风险；在数据存储方面，能够实现不可篡改、不可删除伪造的数据存储安全。与此同时，区块链技术可以为人工智能提供大量的数据，还可以将人工智能模型在基于区块链的平台上运行、存储和共享，创造一种更加安全、透明和可信的去中心化人工智能。区块链为数据隐私与安全、网络安全、人工智能安全等问题提供了解决方案。

可信执行环境（Trusted Execution Environment，TEE）是一种新兴的系统安全与隐私保护技术，该技术从底层硬件和操作系统开始出发，提供一个隔离的运行环境，保护代码和数据不被攻击者攻击。其工作原理分为以下几个步骤：隔离、加密、完整性检查和远程认证。TEE 在处理器内部创建一个独立的执行环境，与其他应用程序和操作系统隔离，通过硬件加密技术来保护数据和代码的安全，通过完整性

检查来确保代码和数据在执行过程中没有被篡改。TEE 还支持远程认证，允许用户通过安全通道验证 TEE 的真实性和完整性。将 TEE 与人工智能系统融合，有助于更好地保障模型训练和推理过程中的保密性。训练阶段，TEE 中的数据处理都处于加密状态；推理阶段，TEE 则可保护用户输入和模型结果的隐私。同时，其硬件隔离和安全验证机制可以更有效地防止未经授权的访问和攻击，增强模型运行时的安全性。

机器遗忘（Machine Unlearning）技术可以从训练数据集和已训练的模型中完全且快速地移除样本及其影响。出于隐私、法规和法律的需要，有些特定样本的信息需要从模型中移除，在移除这些样本的同时还需要从已经训练过的模型中删除这些样本的影响。这是因为成员推理攻击和模型逆向攻击可以揭示关于训练数据集特定内容的信息。更重要的是，一些立法要求强制删除私人信息。目前两种主流的遗忘方法是数据重组和模型操作。数据重组聚焦于修改训练数据、重新训练模型，考虑到模型重新训练的开销，这类方法往往会加入一些剪枝策略来减小重训练的成本。而模型操作则通过直接调整模型的参数，消除遗忘数据对于模型的影响。模型操作的遗忘速度更快但遗忘效果有效，数据重组的遗忘效果较好但遗忘速度较慢。

模型数字水印（Digital Watermarking）技术可以保护数据和模型隐私、以及模型知识产权。机器学习和人工智能模型在各个行业、各个领域中发挥着越来越重要的作用，特别是预训练大模型，具有极高的商业价值，模型的保密性和知识产权保护也越来越受到关注，模型溯源和模型版权保护的需求日益渐起。研究人员发现，可以在模型

训练阶段向模型植入一系列秘密信息达到给模型添加“水印”的目的。当从可疑模型中提取相同或者相似的水印时，就可以验证模型的所有权。目前，水印技术主要分为两类：一种是将水印直接嵌入模型权重参数中，另一种是在模型预测信息中嵌入水印。

3.2 人工智能隐私保护平台

为了实现人工智能系统和数据隐私保护，目前业界的人工智能隐私保护平台可主要分为隐私保护机器学习平台和人工智能安全检测平台。

隐私保护机器学习平台主要是基于联邦学习、安全多方计算、区块链、可验证计算等技术打造的数据安全共享基础设施。其中，通过安全多方计算、联邦学习可以打通数据孤岛，将计算环节移动到数据端，实现数据可用不可见，解决多家机构数据合作过程中可能存在的数据安全风险和隐私泄露问题；通过联邦区块链可以保证过程的不可篡改行和可溯源性，实现数据可用不可见和计算可信可链接；通过联邦大模型可以突破数据和算力的壁垒，实现多方数据的融合和增值，同时保护数据隐私和安全。目前，隐私计算机器学习平台已经在政府、金融、医疗等领域进行了应用落地，主要包含金融风控、晶总营销、政府服务、保险定价、医疗健康等实现。除此之外，隐私计算机器学习平台还在智慧能源、智慧城市、工业互联网等探索性应用中发挥着重要作用。

人工智能安全检测与评估平台还处于起步阶段，目前相关平台主

要是基于深度学习、智能博弈对抗等技术打造的人工智能安全与隐私保护工具集。相关平台主要包含模型漏洞识别与挖掘、模型能力测评、模型攻击防御测评、模型生成内容检测与评估等功能，未来可持续增加模型可解释性、算法公平性、隐私保护等特性服务。

四 人工智能隐私保护建议

4.1 建立健全的人工智能隐私保护合规监管机制

如今，人工智能技术在很多个行业都有所应用，不可避免的会由于各种原因造成个人隐私泄露，在这种情况下必须出台强有力的举措来应对这样的风险，才能在利用人工智能应用换取个人数据享有经济利益和社会价值的同时，真正有效的保护个人隐私。

建立健全法律法规与监督机制。安全与发展是一体之两翼，驱动之双轮，必须牢筑人工智能安全法律基础，明确人工智能安全责任与义务，才能从根本上保证人工智能产业持续健康发展。而隐私保护又是安全的重中之重，建议持续完善顶层设计，出台隐私和数据保护相关政策法规和标准化设计，从上至下建立人工智能安全法律体系。做到“有法可依”。建立监管和执法机制，监督人工智能产品数据处理和隐私保护情况，做到“有法必依”。监管机构需定期进行检查和评估，做到“执法必严”，对违法违规行为进行严厉处罚，做到“违法必究”。确保人工智能公司在收集和使用用户数据时遵守相关规定，保护用户的合法权益。同时，监管机构也需要密切关注人工智能技术

发展情况，及时调整和完善监管政策和措施，以适应不断变化的技术与风险。

人工智能公司需合规化建设。人工智能企业应树立发展和隐私保护并重意识，积极履行合规建设主体责任，依据相关法律法规和标准建设人工智能系统。企业内部应建立人工智能隐私保护管理体系，规范产品的数据采集、存储和使用行为，接受监管机构的监督和检查。研发人工智能产品时，应使用技术手段保证人工智能产品在隐私保护方面合法合规，如实施访问控制、权限管理和日志审计等，以监控和控制数据的访问；使用安全加密技术保护产品中的数据传输和存储，以防止未经授权的访问和数据泄露；产品在使用云服务或其他外部存储提供商时，企业要合理评估其安全性和隐私保护措施。同时，企业还应提供针对产品透明的算法和模型信息，向用户说明该产品是如何通过学习和训练生成对应的结果以及做出对应的决策的。

4.2 加强可操作的人工智能隐私保护标准建设

健康有序的行业发展离不开标准的先行引领，标准化、规范化发展才能行稳致远。目前我国已有《国家新一代人工智能标准体系建设指南》对人工智能安全与隐私保护标准做出了规划，但具体标准数量仍旧不足。未来，在持续完善标准体系顶层设计的同时，人工智能及隐私安全相关的科研院所、企事业单位、普通高等院校、职业院校等各类主体应瞄准人工智能安全基础性标准、关键急用型标准，集中力量抓紧研制，在各类人工智能新技术标准中增加隐私保护相关内容，

切实将发展与安全一体化。在全力丰富相关标准的同时，应注意满足各标准之间的一致性与协调性，以及所研标准与国际标准之间的衔接；还需注重标准的可操作性，准确描述安全要求，避免模糊的定性描述，保证落地执行时的结果一致性。应尽快构建人工智能安全评估体系，从算法模型、系统平台、数据保护等方向构建测试评估指标体系，指导人工智能产品安全发展。科研院所、企事业单位等应加大贯标力度，推广应用已研制标准，持续发挥标准的基础性、规范性作用。

4.3 构建全周期的人工智能隐私保护体系

人工智能全生命周期大体可包括数据采集存储、模型训练推理应用等阶段，如何把隐私保护纳入人工智能研发各个阶段具有重要的意义。

数据采集阶段的隐私保护。对人工智能所需数据的采集操作进行认证与授权工作，确保人工智能数据采集者拥有全局唯一标识符，明确角色分级分类与权限控制规则，确保采集者可以访问并且只能访问自己角色级别所能访问的数据。同时建立人工智能数据安全采集制度，对恶意采集行为进行判定与持续追踪，防止恶意行为引发数据与隐私泄露。

数据存储阶段的隐私保护。建立数据接口规范，对数据在处理、保存环节中涉及到的数据对接等操作进行管理，避免数据在处理、保存过程中形成隐私泄露风险。在存储隐私相关的数据时，明确数据的使用范围，进行分级分类存储，并采用数据加密技术进行安全存储。

防止使用者无意间对隐私数据进行操作引起隐私泄露问题。对于数据量庞大的数据进行存储时，若采用分布式存储技术，应建立健全的网络安全机制，防止数据通过网络传输被非法使用者访问、窃取等。

模型训练中实现数据使用隐私保护。一是对数据进行清洗等预处理，进行基本的防数据投毒、侵权数据、有害数据的分析判断，保证人工智能使用合法的、高质量的数据进行训练。二是保证个人数据隐私，为了预防从模型中推断出隐私信息，数据隐私必须进行去隐私化处理，屏蔽掉无关信息。三是根据使用场景选择数据使用框架，如多方协同场景下，可以选择多方安全技术、联邦学习等分布式框架，在数据不出域的情况下实现模型训练，并实现数据源隐私保护。

模型训练中内置隐私保护机制。目前，OpenAI 的 ChatGPT、微软的 BingChat 和谷歌的 BARD 都有基本的隐私与安全保护机制。但是，总体上看，其隐私与安全保护机制还比较弱，容易被轻易饶过。因此，模型训练过程中要加强输入控制，防范网络层面和内容层面的有害输入，对语言模型要拦截各类提示注入攻击。同时要加强模型输出控制，对生成模型，需在生成内容输出前，对生成内容进行合法合规检测和过滤，防止输出存在数据隐私安全问题的内容。另外要加强模型特征、模型训练和模型结果的透明性、可解释性，深度神经网络由大量的节点之间相互联接构成，用户目前还不清楚不同数据、特征对模型参数的影响，且网络的输出则依网络参数值、激励函数、连接方式的不同而不同。需要通过模型解释性研究，从根本上确保人工智能隐私保护。

模型推理应用中的安全测评。建议从两个维度对人工智能平台和

服务进行安全测评。一是网络安全维度，通过渗透测试、模糊测试等安全性测试手段，检测模型、算法插件等有无安全漏洞。这类漏洞通常会导致平台失控或产生有害内容。一旦发现，应及时通知厂商修复。二是决策安全维度，因为人工智能技术存在较大的不确定性和不可控性，所以，需要通过精心设计和定制化的输入等，如对抗样本、恶意指令，检测平台是否会产生有偏见的决策，或是生成有害的、有偏见的、侵权的、与事实不符的内容，并进一步检测平台和服务是否在训练数据集、模型、安全模块、二次开发调用接口或者算法插件上出现问题，从而给出平台和服务的整改建设方案。

模型推理应用中的安全监测与预警。建议推进针对人工智能的态势感知能力建设，打造大范围人工智能应用风险和数据安全风险识别的威胁发现能力。通过梳理人工智能系统、应用、接口的资产清单，针对性地建立定向流量实时监测、统一威胁事件与关联日志收集，建立一体化态势感知监测体系，有效识别和预防人工智能环境面临的安全威胁。推进针对人工智能的威胁情报能力建设，打造精准化威胁情报共享体系，通过结合已有威胁情报中心建设情况，融合多源合作伙伴情报源，构建情报驱动的威胁分析应用，增强人工智能应用及重要场景的防护能力。

4.4 发展多维度的人工智能隐私保护技术

加强技术创新和应用，技术是保护数据隐私和安全的重要手段，它可以提高数据的可靠性、可信度和可控性。目前，已经有一些技术

被用于保护数据隐私和安全，如密码学技术、差分隐私技术、多方安全计算技术等。这些技术可以实现数据的加密、扰动、隔离、去标识化等功能，使得数据在传输、存储或使用过程中不被泄露或滥用。然而，这些技术也存在一些局限性和挑战，如性能损失、兼容性问题、成本增加等。因此，本白皮书建议加强多层次、多维度的技术创新和应用，例如，从算法本身出发，需要加强特征、模型训练和模型结果的可解释性研究，创新机器模型遗忘和数字水印技术，保护数据隐私、模型保密性与完整性。从算法计算范式和计算环境出发，需要优化联邦学习、区块链、可信执行环境等新兴技术，推出基于人工智能和新兴技术融合的隐私保护技术，以提高人工智能隐私和安全的保障水平。

4.5 培养复合型的人工智能隐私保护人才

人工智能正在引发新一轮智能化浪潮，尤其是以 ChatGPT 为代表的生成式 AI 模型的快速发展与应用，其引发的“AI 即服务”趋势拓展了更大的业务空间，同时其引发的“AI 隐私与安全保护”风险，也昭示着 AI 隐私保护保障迫在眉睫。AI 隐私保护涉及人工智能、隐私计算、区块链、密码学、数据安全、网络安全、政治学等多领域的交叉融合。对于人才培养，一方面，企业应培育一支人工智能复合型人才队伍，加强对人工智能技术人员的隐私保护技术、对抗防御技术、安全合规等培训，拓展人才的知识广度。另一方面，企业应加强与高校、科研机构在人工智能安全领域的对接合作，企业作为“AI 服务”的提供者，高校、科研机构作为人工智能、机器学习隐私保护基础理

论的引领者，推动人工智能与经济社会安全、可信、可控融合，拓展人才的深度。

4.6 构建负责任的人工智能隐私保护生态

人工智能隐私保护与安全需要产学研用各方推动负责任的人工智能隐私保护与安全生态。在基础能力与理论技术上，要加强人工智能可解释性、公平性、鲁棒性等基础理论研究，夯实人工智能安全技术底座，建设安全的行业大模型；人工智能企业可联合行业各方共同研制人工智能算法与平台安全性测试、安全防御、安全监测与预警方法、技术、规范和工具集，以实现自动化、全面化的安全评估；鼓励人工智能安全能力领先企业开放安全能力，以平台化方式服务中小微企业发展。在应用场景上，针对不同行业特点，研发相适应的人工智能安全产品；推动人工智能安全技术与产品在各行业各领域的深度应用，拥有成功应用案例的企业可以积极申报各类应用示范遴选，推广成功经验，提升企业影响力。在产品落地部署与运行上，企业要重视人工智能产品的隐私与安全保护，一方面要正向保证人工智能模型本身是保密的、安全的，避免各类安全威胁，保护用户隐私；另一方面，企业也应关注人工智能模型不被恶意攻击者利用，在训练数据集、模型参数与结构、人工智能产品的安全性上加大投入，提升保护力度。

参考文献

[1]陈宇飞, 沈超, 王骞, 等. 人工智能系统安全与隐私风险[J]. 计算机研究与发展,

2019, 56(10):16.

[2]杨子祺,吴正阳,任奎. 一种基于模型水印的机器学习模型版权保护方法[P].浙江省：CN116244669A,2023-06-09.

[3]Xu, H., Zhu, T., Zhang, L., Zhou, W., & Yu, P. Machine Unlearning: A Survey. ACM Computing Surveys. 2023.

[4]中国信息通信研究院, 清华大学, 蚂蚁集团. 可信 AI 技术和应用进展. 2023.

[5]大数据协同安全技术国家工程研究中心. 大语言模型提示注入攻击安全风险分析报告. 2023.

[6]袁勇,王飞跃. 区块链技术发展现状与展望[J]. 自动化学报, 2016, 42(4):14.

[7]王群,李馥娟,王振力,等. 区块链原理及关键技术[J]. 计算机科学与探索, 2020, 14(10):23.

[8]李宗维,孔德潮,牛媛争等. 基于人工智能和区块链融合的隐私保护技术研究综述[J]. 信息安全研究,2023,9(06):557-565.

[9]张夏明,张艳. 人工智能应用中数据隐私保护策略研究[J]. 电子科学技术, 2020, 000(004):76-84.

[10]The GSM Association. Artificial Intelligence Security Guidelines Version 1.0. 2022.

[11]俞巍，李志强，李青青，龚奇源. AIGC 大模型为什么需要可信执行环境（TEE）？. https://mp.weixin.qq.com/s/BnoFbMdqe5cBi8Reh0oeVw. 2023.

[12]李杨,温雯,谢光强. 差分隐私保护研究综述[J]. 计算机应用研究, 2012, 29(9):6.

[13]熊平,朱天清,王晓峰. 差分隐私保护及其应用[J]. 计算机学报, 2014, 37(1):22.

[14]胡奥婷,胡爱群,胡韵,等. 机器学习中差分隐私的数据共享及发布:技术,应用和挑战[J]. 信息安全学报, 2022, 7(4):16.

[15]钱萍,吴蒙. 同态加密隐私保护数据挖掘方法综述[J]. 计算机应用研究, 2011, 28(5):5.

[16]崔建京,龙军,闵尔学,等. 同态加密在加密机器学习中的应用研究综述[J]. 计算机科学, 2018, 45(4):7.

[17]谭作文，张连福. 机器学习隐私保护研究综述[J]. 软件学报， 2020, 31(7):30.

[18]李浪，余孝忠，杨娅琼，等. 同态加密研究进展综述[J]. 计算机应用研究，2015， 32(11):6.

[19]科技日报. 我国人工智能五大开放创新平台集体亮相. http://www.cac.gov.cn/2019-05/10/c_1124475056.htm. 2019.

[20]国家知识产权局. 二〇二二年中国知识产权保护状况. 2023.

[21]中国信息通信研究院. 联邦学习常见应用研究报告. 2022.

[22]中国移动研究院. 联邦学习技术发展与应用白皮书. 2021.

[23]李鉴， 邵云峰， 卢燚， 吴骏. 联邦学习及其在电信行业的应用. 信息通信技术与政策. 2020.

[24]中国信息通信研究院. 可信人工智能产业生态发展报告. 2022.

[25]中国发展网. 我国首部人工智能产业专项立法《深圳经济特区人工智能产业促进条例》正式公布. https://baijiahao.baidu.com/s?id=1743295101400323317&wfr=spider&for=pc. 2022.

[26]第十三届全国人民代表大会三次会议. 中华人民共和国民法典. 2020

[27]李秋娟. ChatGPT 风靡背后，美国联邦人工智能治理现状：法律、政策和策略. http://www.python88.com/topic/156519. 2023.

[28]36 氪. 史上最严数据保护法 GDPR 生效：保护用户隐私，但将拖慢创新 . https://baijiahao.baidu.com/s?id=1601327537454164540&wfr=spider&for=pc. 2018.

[29]孙洁、郄晓航. 《网络安全法》、《数据安全法》和《个人信息保护法》：三法联动开启企业新一轮数据合规浪潮. http://www.jiayuan-law.com/CN/news_content.aspx?Lan=CN&MenuID=00000000000000000006&KeyID=00000000000000002065&Type=00000000000000000081. 2021.

[30]申拓律师事务所. 人工智能技术在法律上的应用和监管_相关_保护_处理. https://it.sohu.com/a/674282479_121441272. 2023.

[31]北京市人民政府. 北京市人民政府关于印发《北京市加快建设具有全球影响力的人工智能创新策源地实施方案（2023-2025 年）》的通知. 2023.

[32]上海市经济和信息技术委员会.《上海市促进人工智能产业发展条例》全文公布.2022.

[33]王鹏.保护与开放：加速推进和规范人工智能立法.https://baijiahao.baidu.com/s?id=1768838850725404353&wfr=spider&for=pc.2023.

[34]Paul Voigt,Daniel Tolks.欧洲数据经济的”首项法案”|《数据治理法案》.https://mp.weixin.qq.com/s?__biz=Mzl5MDQzOTkyMQ==&mid=2247486285&idx=1&sn=9835aa299a3c90b2f3182b65f67dbed2&chksm=ec1e9cd3db6915c56161c91c210fbd29685ac61658443e8f4fde1f9a854614c8e3884602d3ee&scene=27.2022.

中国联通研究院是根植于联通集团（中国联通直属二级机构），服务于国家战略、行业发展、企业生产的战略决策参谋者、技术发展引领者、产业发展助推者，是原创技术策源地主力军和数字技术融合创新排头兵。联通研究院致力于提高核心竞争力和增强核心功能，紧密围绕联网通信、算网数智两大类主业，按照 4+2+X 研发布局，开展面向 C3 网络、大数据赋能运营、端网边业协同创新、网络与信息安全等方向的前沿技术研发，承担高质量决策报告研究和专精特新核心技术攻关，致力于成为服务国家发展的高端智库、代表行业产业的发言人、助推数字化转型的参谋部，多方位参与网络强国、数字中国建设，大力发展战略性新兴产业，加快形成新质生产力。联通研究院现有员工 700 余人，85%以上为硕士、博士研究生，以“三度三有”企业文化为根基，发展成为一支高素质、高活力、专业化、具有行业影响力的人才队伍。

战略决策的参谋者
技术发展的引领者
产业发展的助推者

态度、速度、气度

有情怀、有格局、有担当

中国联合网络通信有限公司研究院
地址：北京市亦庄经济技术开发区北环东路 1 号
电话：010-87926100
邮编：100176