WAIC
世界人工智能大会
WORLD ARTIFICIAL INTELLIGENCE CONFERENCE
2020 云端峰会 SUMMIT ONLINE

人工智能
数据安全治理报告

GOVERNANCE ON AI DATA SECURITY

SICSI
CYBER RESEARCH INSTITUTE
赛博研究院

观安

— 版权声明 —

COPYRIGHT STATEMENT

出品方：

上海赛博网络安全产业创新研究院

上海观安信息技术股份有限公司

编写组：

惠志斌　上海赛博网络安全产业创新研究院首席研究员

石英村　上海赛博网络安全产业创新研究院研究员

唐巧盈　上海赛博网络安全产业创新研究院高级研究员

李　宁　上海赛博网络安全产业创新研究院高级研究员

胡绍勇　上海观安信息技术股份有限公司首席技术执行官

夏玉明　上海观安信息技术股份有限公司研究院院长

魏国富　上海观安信息技术股份有限公司研究院副院长

陈一根　上海观安信息技术股份有限公司研究院研究员

谢　江　上海观安信息技术股份有限公司首席安全标准顾问

FOREWORD | 前言

当前，随着以“数字新基建、数据新要素、在线新经济”为特征的新一波数字经济浪潮全面来临，推动全球人工智能发展逐步从“探索期”向“成长期”过渡，在技术和产业上均进入重要的转型阶段。在此背景下，人工智能发展和数据安全问题日益深度交织融合，影响用户隐私、公民权益、商业秘密、知识产权、社会公平、国家安全等各个方面，数据安全问题已然成为人工智能全面新发展的重要制约瓶颈和亟需突破的关键挑战。

因此，报告从数据生命周期的视域，针对数据采集、数据处理、数据流通和数据使用阶段，重点聚焦并梳理了人工智能发展中较为独特或更突出的数据安全问题；并从人工智能发展战略、安全倡议和伦理规范、数据安全法律法规、相关行业标准、全球数据安全前沿技术和企业实践等维度，全面分析了当前全球人工智能数据安全治理的主要现状和最新动态。

基于人工智能发展的阶段性特点，以及人工智能数据安全挑战的特性，报告结合全球相关治理实践和我国实际情况，构建了综合性的人工智能数据安全治理框架，明确总体的治理思路和治理原则，并探索了顶层设计、标准体系、企业能力和安全供给四个维度的治理路径。最后，报告提出了通用场景下的人工智能数据安全风险评估平台，以及智能网联汽车、人脸识别和工业互联网三个人工智能主要应用场景的数据安全综合解决方案。

PART 1
人工智能发展与数据安全挑战

一 | 数字经济时代的人工智能发展趋势

1. 新一轮数字经济发展浪潮全面来临

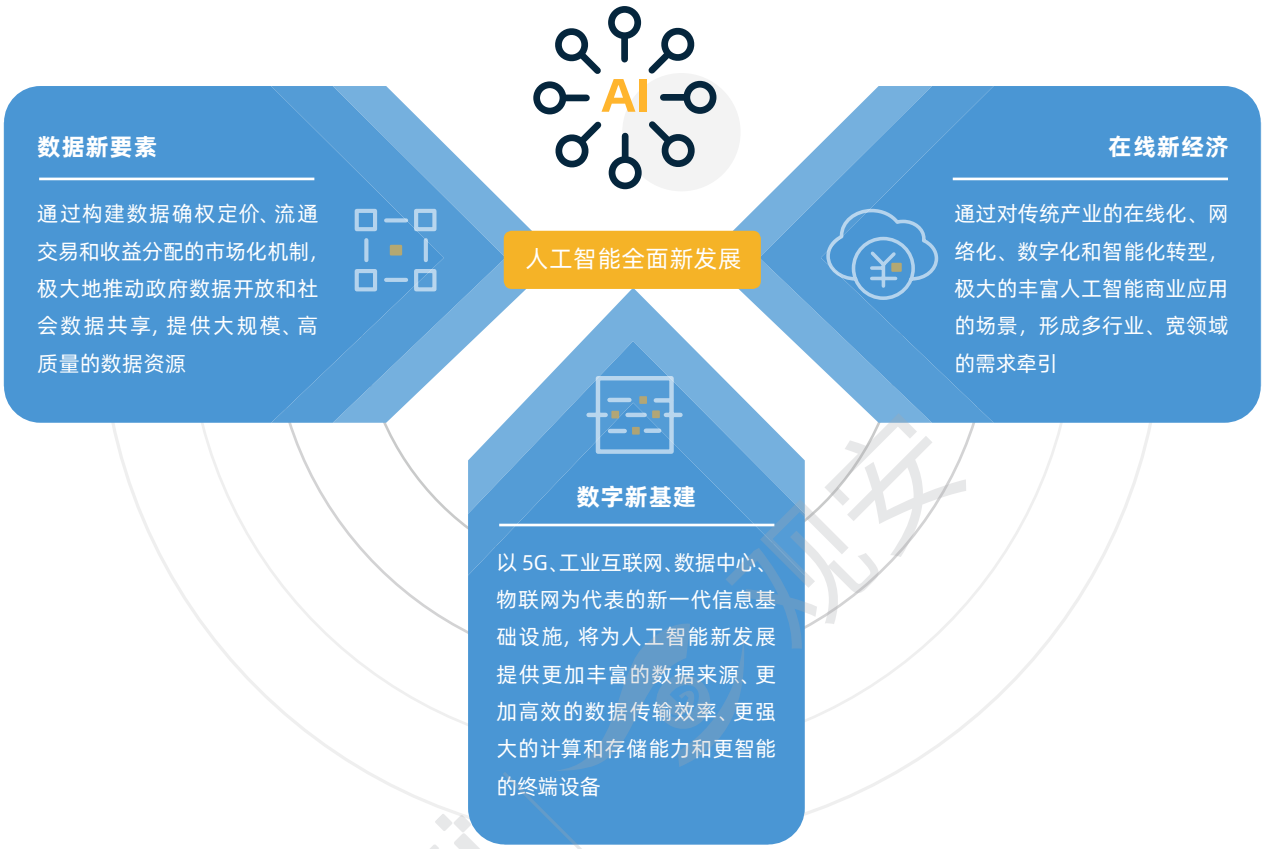
数字经济是指以数字化的知识和信息为关键生产要素，以现代信息网络为重要载体，以信息通信技术的广泛普及和有效使用为核心驱动，全面推动商业模式优化创新、生产消费效率提升和产业经济智能化升级的一系列经济活动[1]。近年来，随着大数据、云计算、物联网等为代表的数字技术带来了全球性的科技革命和产业变革，**以“数字新基建、数据新要素、在线新经济”为核心特征的新一轮数字经济发展浪潮全面来临，为基于算法、算力和数据驱动下的人工智能全面新发展注入了全新的强大动能。**

● **数字新基建成为人工智能新发展的坚实底座和基础支撑。**近年来，美、欧、日、英等全球主要经济体纷纷大力发展以 5G、物联网、工业互联网、云计算、数据中心、卫星互联网等为代表的新型数字基础设施建设，而中国自 2018 年 12 月首次提出“新基建”概念以来，至今已有 7 次中央级会议或文件明确表示加快新基建的建设速度，并在 2020 年 5 月将“加强新型基础设施建设”明确写入 2020 年《政府工作报告》。数字新基建的加快推进和不断完善将成为人工智能全面新发展的坚实数字底座。其中，物联网和工业互联网将极大拓宽人工智能的数据来源和应用场景，5G 和卫星互联网则能够大幅度提高人工智能数据传输、处理以及应用开发的效率，数据中心、云计算设施确保了人工智能发展所需要的巨大基础计算和存储需求，以人工智能芯片、智能终端、智能计算平台为代表的人工智能基础设施则为人工智能应用提供了高质量的硬件支撑。

● **数据新要素成为人工智能新发展的核心动能和强大驱动。**2020 年 4 月，中国发布《中共中央、国务院关于构建更加完善的要素市场化配置体制机制的意见》，将数据明确列为一种新型生产要素与土地、劳动力、资本和技术等传统要素并列，并强调要加快培育数据要素市场。随着全球各国不断加快数据市场的建设，将在未来形成包括数据要素确权定价、数据交易流通和收益分配等核心功能的数据要素市场改革驱动和政策赋能，能够极大地推动政府公共数据开放和社会企业数据共享，进一步打通数据壁垒，推动形成数据更大规模的有序、便捷、高效和安全流动交易的宏大数字空间，为人工智能全面新发展注入高质量的数据动能。

● **在线新经济为人工智能新发展提供丰富广阔的应用场景。**随着全球经济因疫情冲击而遭受重创甚至面临衰退，以在线新经济为代表的数字经济模式成为全球经济复苏和转型增长的核心驱动。在线新经济的本质是传统行业线上化、网络化、数字化和智能化转型，是指以大数据、云计算、区块链等新一代信息技术在办公、医疗、教育、金融、生产、物流等各个垂直领域的加速落地并形成新型的经济业态[2]。在线新经济的兴起将为人工智能发展提供丰富广阔的应用场景，不断推动人工智能的算法迭代优化，以及向更多行业和更多领域渗透落地，形成人工智能全面新发展庞大的、立体的需求牵引。

图 1 数字新基建、数据新要素和在线新经济推动人工智能全面新发展



2. 全球人工智能发展逐步从“探索期”向“成长期”转变

根据行业生命周期理论 (Industry Life Cycle) 和 Gartner 的技术成熟度曲线模型，报告认为当前全球人工智能发展正在逐步渡过“探索期”并进入“成长期”，且已进入了全面转型的关键节点，主要有以下四个关键特征。

● **人工智能专用技术迅速突破。**专用人工智能是一种面向特定领域的人工智能（即“弱人工智能”），由于其具备任务单一、需求明确、应用边界清晰、传统领域知识丰富和功能建模相对简单等特征，因此在重点领域形成了技术突破后，随即进入了快速的商业化应用阶段，成为人工智能迈向“成长期”的底层技术支撑。目前，人工智能主要的应用技术方向包括以深度学习为代表的机器学习算法，以计算机视觉、图像识别、语音识别为代表的智能感知技术，以及以无人驾驶、自动机器人等为代表的自主无人系统的三大领域。

● **人工智能产业生态蔚然成型。**从全球范围内看，围绕专用人工智能技术的人工智能产业已经初具规模。中国电子学会的《新一代人工智能产业白皮书（2019 年）》显示，2018 年全球新一代人工智能产业规模超过 555.7 亿美元，预计 2019 年产业规模将突破 718 亿美元[3]。而据德勤公司（Deloitte）预测，2025 年世界人工智能总体市场规模将超过 6 万亿美元，2017-2025 年复合增长率达 30%[4]。在人工智能产业

链上，形成了包括智能芯片、传感器、智能设备厂商的硬件层，数据分析处理、算法模型、软件开发和关键技术厂商的技术层，行业应用、解决方案、产品服务开发厂商的应用层等三大层级体系，整体产业生态发展开始从“探索期”的弥补市场空白向“成长期”的产业结构优化转型发展。

● **人工智能投融资日趋理性成熟。**自 2018 年全球人工智能投融资达到 784.8 亿美元高值以来，2019 年全球人工智能领域的投融资规模开始回落[5]。中国信息通信研究院的《全球人工智能产业数据报告》显示，2019 年一季度全球人工智能融资规模 126 亿美元，环比下降 7.3%[6]。创投研究机构 CB Insights 的《全球人工智能投资趋势年度报告》显示，AI 初创公司超过 70% 的投融资为早期投资或 A 轮融资，资金向头部初创企业集中的趋势明显加强[7]。伴随着“探索期”的风险投资甚至跟风投机泡沫的消除，核心技术、商业落地和可持续发展成为投资者最关切的决策因素，投融资整体趋向理性必然带来产业结构的优化，驱动人工智能从“探索期”向“成长期”发展。

● **人工智能应用场景向深层拓展。**目前，人工智能的应用场景包括金融、零售、医疗、教育、政务、制造、汽车、家居、智慧城市、数字内容、公共安全等多个垂直领域。相关行业场景的应用深度不一。IDC 发布《AI 驱动金融行业智能决策（2020）》显示，目前金融行业头部企业 AI 应用渗透率达 75% 以上，第二梯队的企业 AI 应用渗透率超过 50%，第三梯队的金融企业 AI 应用渗透率约 30%，成为当下人工智能渗透率最高的应用场景[8]。中国新一代人工智能发展战略研究院对 797 家中国人工智能骨干企业中的 581 家应用层企业进行了详细分析，提供企业技术集成与方案提供、智能机器人两个应用领域的人工智能企业数占比最高，分别为 15.43% 和 9.66%。紧随其后的是关键技术研发和应用平台、新媒体和数字内容、智能医疗、智能硬件、金融科技、智能商业和零售、智能制造领域[9]。相关研究预测，到 2022 年前后，医疗、公共安全、智能制造、无人驾驶和智慧城市等场景的整体人工智能渗透率都将超过 25%。可以说，人工智能发展将从“探索期”的向更多领域普及应用逐步过渡到“成长期”的向成熟领域更深层渗透。

3. 数据安全成为人工智能全面新发展的重要制约问题

数据安全是数字经济发展中最关键的安全挑战之一，随着人工智能在产业和技术两个方面都在加快渡过“探索期”，逐步进入“成长期”，**人工智能发展与数据安全将更加深度地交织在一起，数据安全问题已然成为人工智能突破关键转轨期所必须解决的重要制约瓶颈。**

一方面，**人工智能发展加剧了传统数据安全风险。**在以“数字新基建、数据新要素、在线新经济”为重要特征的数字经济发展大背景下，人工智能的新发展必然伴随着数据总量的井喷式爆发，各类智能化数据采集终端的加快增长，数据在多种渠道和方式下的流动更加复杂，数据利用场景更加多样，整体数字空间对于人类现实社会各个领域的融合渗透更趋于深层，这将使得传统数据安全风险持续地扩大泛化。

另一方面，**人工智能催生了各种新型的数据安全风险。**人工智能主要通过庞大的训练数据集来驱动算法更迭优化，从而实现输出结果的智能化，因其对于数据资源特有的处理方式，将会带来数据污染、数据投毒、算法歧视等一系列的新型数据安全问题。同时，人工智能在自动化网络攻击、数据黑产的应用，使得传统网络安全和数据安全威胁更加复杂，对国家和企业现有的数据安全治理能力形成巨大冲击。

二 | 人工智能全面新发展的数据安全挑战

人工智能全面新发展的数据安全挑战既有传统数据安全问题的普遍共性，更具有人工智能时代的独特烙印，影响领域覆盖用户隐私、公民权益、商业秘密、知识产权、社会公平、国家安全等各个方面。因此，报告基于数据生命周期的视域，从数据采集、数据处理、数据流动和数据使用阶段，重点聚焦分析人工智能发展中较为独特或更突出的数据安全问题。

表 1 人工智能数据安全风险挑战

数据 生命周期	数据 安全问题	风险 影响机制	安全 风险后果
数据采集阶段	用户权利保障	对个人数据采集时缺乏充分的用户知情和授权机制	侵犯用户隐私
	过度采集	现场无差别采集时，采集对象、数据类型范围扩大产生过度采集问题	侵犯用户隐私、危害国家安全和公共利益
数据处理阶段	数据污染	训练数据质量较差或缺乏标准化处理，使得数据与算法模型不相适配	导致算法反复优化、测试结果不稳定、成本激增甚至算法不可用等问题
	数据投毒攻击	恶意攻击者通过对训练数据的添加、篡改来定向干涉算法的决策和结果输出	直接导致算法决策的错误，并影响到算法关联的设备系统，引发实体物理层面的危害
	数据偏差和歧视	训练数据、样本数据或者算法设计本身存在偏差歧视，导致算法决策存在不准确性	涉及到公众的算法决策出现偏差，将会导致不公平的歧视现象
数据流通阶段	数据交互	人工智能在数据的采集、标注、分析和算法优化都会涉及到各个主体，数据产业链中的安全能力薄弱主体将会使得整个数据链路面临风险	可能会带来数据泄露、数据盗取和数据泄露的危害
	数据孤岛	目前合法、便捷、安全和低成本的数据交易流通市场尚未形成，企业之间、行业之间存在法律和技术上双重壁垒	制约人工智能发展，可能滋生数据黑产
	数据跨境	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，必然带来跨境数据流动	可能带来涉及个人敏感信息、重要数据的出境问题，从而威胁公民权益和国家安全
数据使用阶段	关联分析	人工智能对于分散数据项的关联分析和深度挖掘，能够将用户本无意公开的信息或特征暴露出来	将会严重侵害用户隐私和人身安全，甚至威胁国家安全
	还原攻击	人工智能在用于逆向还原攻击时，能够基本还原被攻击者的算法逻辑和训练数据特征集	恶意攻击者通常将其用于窃取企业商业机密
	对抗样本攻击	通过在网络空间或者物理世界的样本数据输入中添加细微、无法识别的干扰信息，导致人工智能模型在正常运转中输出一个错误的结果	使得人工智能的产生攻击者需要的定向输出结果，引发安全事故

1. 数据采集阶段的安全挑战

（1）人工智能数据采集时难以保障用户的数据权利

人工智能算法尤其是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和系统性能测试。目前，人工智能企业的数据采集主要包括现场无差别采集、直接在线采集、网上公开数据源和商务采购等方式。在现场无差别采集时，由于无法提前预知采集的用户，**难以获得用户的充分授权同意。**而在直接在线采集时，由于人工智能系统通常由训练好的模型部署而成，对用户数据需要进行连续性的处理分析，**因此很难保障用户的修改、撤回等权益。**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，因此存在一部分企业通过灰色渠道获得用户数据，**而这些数据缺乏用户知情同意。**

另外，随着计算机视觉、图像识别和语音识别技术的发展，以对**个人生物特征数据**进行采集分析的应用成为人工智能发展的重要应用方向。而在各国现行的法律规制下，人脸数据、基因数据、体型数据、语音特征等生物特征数据均属于个人敏感数据，**对此类数据的采集和处理存在较大的法律合规和隐私保护风险。**

案例 1：美国学术研究人员通常通过谷歌图片搜索、图片分享网站 Flickr 的授权（Creative Commons license）、公共 Instagram 帐户或者其他一些途径获取大量的图片，以供训练或测试人脸识别算法[10]。授权的协议显示这些图片数据仅用于学术研究。然而，随着微软、IBM、Facebook 和谷歌等公司与各大高校研究室的合作，大量用户仅授权用于学术研究的个人照片被用于商业领域。例如 2019 年 3 月，IBM 被爆出使用互联网上的照片作为人脸识别的“养料”，其中包含了图片分享网站 Flickr 上近 100 万张照片，但未获得用户许可，因此引发了国外媒体的高度关注和用户对隐私的广泛担忧[11]。

案例 2：2019 年 10 月，浙江理工大学特聘副教授郭兵因不愿意使用人脸识别，将杭州野生动物世界告上了法庭，成为国内消费者起诉商家的“人脸识别第一案”。杭州野生动物世界在未与郭兵进行任何协商亦未征得同意的情况下，通过短信的方式告知原告“园区年卡系统已升级为人脸识别入园，原指纹识别已取消，未注册人脸识别的用户将无法正常入园”。而郭兵认为，园区升级后的年卡系统进行人脸识别将收集他的面部特征等个人生物识别信息，该类信息属于个人敏感信息，一旦泄露、非法提供或者滥用，将极易危害包括原告在内的消费者人身和财产安全[12]。

（2）现场无差别采集可能会产生过度采集问题

现场采集是人工智能数据采集的重要方式，广泛应用于无人驾驶、智能家居、智慧城市等场景中。其主要通过在公开环境中部署各类传感器或采集终端，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来控制汽车行驶，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，甚至可能会采集到路边的重要基础设施分布、军事营区等重要数据而对国家安全带来风险。

案例 3：2018 年，亚马逊公司被曝出雇佣了数千名员工，聆听使用其智能音箱 Echo 的用户家中和办公室捕捉到的录音，以帮助改进 Alexa 语音助手，而这种录音的采集和处理无疑会涉及到用户的隐私和办公机密。

因此，智能音箱、智能电视、手机语音助手等智能硬件的“偷听”能力持续引发公众的普遍担忧。

2. 数据处理阶段的安全挑战

（1）数据污染可能会导致人工智能算法模型失效

数据污染的本质是数据质量的技术性治理问题，是指数据与人工智能算法不适配，从而导致算法模型训练成本激增甚至完全失效。数据污染产生的原因可能包括训练数据集规模过小、多样性或代表性不足、异构化严重、数据集标注质量过低、缺乏标准化的数据治理程序、数据投毒攻击等。在数据与模型算法适配度极低的情况下，在进行算法训练时将会明显带来反复优化、测试结果不稳定等问题，使得人工智能运行的成本大大提高，严重的数据污染甚至直接导致人工智能算法模型完全不可用。

（2）恶意数据投毒攻击导致人工智能决策错误

数据投毒是指恶意攻击者人为地在训练数据集中定向添加异常数据或是篡改数据，通过破坏原有训练数据的概率分布而导致模型产生分类或聚类错误，从而连续性引发人工智能的决策偏差或错误，最终产生恶意攻击者所期待的结果。在自动驾驶、智能工厂等对实时性要求极高的人工智能场景中，数据投毒对人工智能核心模块产生的定向干扰将会直接扩散到智能设备终端（如智能驾驶汽车的刹车装置、智能工厂的温度分析装置等），从而产生灾难性事故后果。

案例 4：伯克利人工智能研究人员 Nicholas Carlini 和 David Wagner 发明了一种针对语音识别 AI 的新型攻击方法，只需增加一些细微的噪音，就可以欺骗语音识别系统产生任何攻击者想要的输出，从而实现身份盗用、欺骗认证系统等非法行为[13]。

（3）数据偏差可能会导致人工智能决策带有歧视性

数据偏差是指人工智能算法决策中所使用的

的训练数据和样本数据，因地域数字化发展不平衡或社会价值的倾向偏见，而使得数据所承载的信息带有难以用技术手段消除的偏差，从而导致人工智能的决策结果带有歧视性。由于当下的专用人工智能主要是通过对训练样本数据的结构和概率进行特征统计，从而构建输入数据与输出结果的相关度，而并非通过抽象化的逻辑推演获取真正的因果关系，同时机器学习算法带有“黑箱”的不可解释性，因此这种因数据偏差导致的决策歧视难以使用技术性完全解决。

比如在政府基于大数据统计分析来进行决策时，其获取的网络数据可能会更多地体现经济发达地区或人群的特征，对于数字化程度较低的边缘地域以及老幼贫弱人群的特征无法有效覆盖，从而对政策制定的公平正义产生不利影响。同时，在金融征信、医疗教育和在线招聘领域，可能会因边远地区、弱势群体和少数族裔的数据量不足、数据质量不高等原因，导致自动化决策的准确率会基于人群特征形成明显的分化，从而产生实质性的歧视影响。

案例 5：由算法正义联盟发起的一项针对几款主流人脸识别分析服务（来自 IBM、微软、旷视科技等公司）的测试表明[14]，这些算法针对白肤色男性的识别率非常高，而针对黑肤色女性的识别率则要低很多，例如 IBM 和旷视科技的算法针对黑人女性的错误率都高达 35%。这种偏见部分是由于训练数据中黑肤色人种的代表性不足造成的。

3. 数据流通阶段的安全挑战

（1）人工智能多主体之间的数据交互存在泄露和滥用隐患

人工智能产业生态体系中各主体之间的数据交互而导致数据泄露或滥用主要包括三种类型。其一，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，因而数据将会在供应链的各个主体之间形成复杂、实时的交互流通链路，**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。**

其二，当下多数人工智能初创企业普遍使用开源学习框架，即通过谷歌、微软、亚马逊等互联网巨头公开的模块化基础算法进行应用开发，因此初创企业对于开源框架、第三方软件包、数据库和其他相关组件等均存在较大的依赖性，且由于缺乏严格的测试管理和安全认证，因此**将会面临不可预期的系统漏洞、数据泄露和供应链断供的安全风险。**

其三，通过边缘计算的方式进行人工智能系统开发及数据训练是目前企业较为流行的做法趋势，人工智能云服务平台和开发者、应用者的数据交互，**将会使部署在云侧和端侧的数据会面临比传统信息系统更加复杂的安全挑战。**

（2）数据孤岛和数据壁垒导致人工智能数据供给不足

由于人工智能发展处于“探索期”向“成长期”转型的阶段性特点，对于底层数据资源的竞争仍是人工智能企业最关键的市场竞争力体现。而由于成熟的数据要素市场尚未形成，数据合法、便捷、安全、低成本的交易流通机制仍是空白，远远无

法满足人工智能企业发展对于数据资源的需求。同时，在政府与企业之间、大企业与小企业之间、行业与行业之间，因数据确权问题、数据安全问题等存在着诸多法律和技术上的数据壁垒，形成了“数据孤岛”，**不仅极大制约着人工智能的发展，也成为滋生数据黑产的主要经济动因。**

（3）人工智能产生的跨境数据流动引发数据安全问题

在全球数字经济发展不均衡的大背景下，大型科技巨头在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，而小型初创企业也需要诸多第三方平台和数据分析公司的支撑。因此，无论是企业内部还是与第三方合作，**在人工智能技术研发和场景应用中均需要常态化、持续性、高速率、低延时的跨境数据流动。**

比如在智能网联汽车领域，智能汽车产生的路况、地图、车主信息等大量数据可能回传境外的汽车制造商，进行产品优化升级和售后服务支撑，将会带来个人敏感数据和重要数据出境后的安全不可控风险。这种人工智能发展引发的跨境数据流动，不仅因各国日益趋严的数据安全规制和本地化要求而面临极大的政策障碍，更将对主权国家的国家安全、数据主权等带来复杂的挑战。

4. 数据使用阶段的安全挑战

（1）智能化的深度挖掘将会威胁公民隐私和国家安全

深度挖掘是指人工智能技术能够对用户分散公开甚至匿名化的数据项进行关联分析，从而获得用户无意公开的信息特征和隐私。当前，随着大数据分析和用户画像技术的快速发展，个性化服务变得越来越普遍，各类平台和企业对于用户“数字轨迹”数据的采集成为其提供精准化产品服务的核心基础，这种对于用户习惯行为的长期跟踪和深度分析将使得公民隐私面临安全风险。同时，随着人工智能关联分析技术的发展，通过对公民分散的、单个无意义的数据项进行组合关联分析，能够形成对特定个体识别和特征画像的数据集（比如活动场所、行动习惯、政治态度、宗教信仰等）。这种技术不但本身会直接威胁到用户的人身安全和隐私，若被用于政党竞选和政治宣传，将对各国现行的政治制度产生极大的冲击和颠覆。

案例 6：2010 年 4 月，脸书公司（Facebook）为增强用户黏性启用了第一版 Graph API，此开放应用编程接口可允许第三方软件开发者在程序开发过程中通过与用户互动和授权而获得其个人数据，包括该软件用户的好友列表及好友个人信息。2013 年，英国剑桥大学教授亚历山大 · 科根（Aleksandr Koran）及其创办的数据挖掘公司剑桥分析（Cambridge Analytica）开发了一款专门针对选民的测试应用“这是你的数字化生活”（This is your digital life），对脸书和用户宣称是心理学研究，经用户授权后可收集用户年龄、住址、性别、教育背景等个人信息，还包括用户在网络中发表、阅读、点赞的内容以及各类好友数据。约 27 万脸书用户下载了这一应用，剑桥分析以这 27 万用户为基础衍生搜集了逾 8700 万用户的数据。在随后几年剑桥分析通过对这些数据进行分析向多国政要提供政治竞选服务。2018 年 3 月，《纽约时报》《卫报》和《观察者报》同时曝光脸书有超过 8700 万用户个人数据在未获得用户授权的情况下遭第三方剑桥分析公司 (Cambridge Analytica) 非法收集且用于大数据分析，从而精准对用户推送政治广告甚至假新闻，干预 2016 年美国总统大选和英国脱欧选举。

（2）对人工智能的逆向还原攻击将会侵犯商业秘密

逆向还原攻击是指攻击者通过人工智能应用的公开访问接口，利用一系列技术手段逆向还原出人工智能的算法模型和训练数据。由于算法模型在部署应用中通常需要将访问接口公共发布给用户使用，攻击者可以通过公共访问接口对算法模型进行黑盒访问，利用神经网络等人工智能算法对训练数据集的记忆[15]，从而分析系统的输入输出和其他外部信息，并推测系统模型的参数及训练数据中的隐私信息。甚至部分攻击者能够通过构造出与目标模型相似度非常高的模型，进行不断地优化逼近，从而实现对算法模型的窃取，进而还原出模型训练和运行过程。**逆向还原攻击对算法模型、参数特征的窃取将直接威胁企业的知识产权和网络资产安全，而其对训练数据隐私信息的窃取将对个人隐私构成安全威胁。**

案例 7：美国斯蒂文斯理工学院的研究人员证明任何以隐私保护为目的开发的联合深度学习方法（Collaborative Deep Learning）也是易于被攻破的。他们开发了一种攻击手段，利用学习过程的实时性，攻击者可以训练一个生成式对抗网络 (GAN)，生成目标训练集的原型样本，从而获取原数据集中的隐私数据[16]。

（3）对抗样本攻击将会导致人工智能决策错误

对抗样本攻击是指在样本数据输入中添加细微、无法识别的干扰信息，导致模型在正常运转中输出一个错误的结果。此类对抗样本攻击既可以是网络空间的虚拟信号错误，也可以是物理世界的实体识别错误。比如在智能网联汽车的无人驾驶中，通过对于实体停车或限速标志的精确更改，能够使得算法模型将其误识别为其他标识，从而引发交通事故。

PART 2
全球人工智能数据安全治理现状

人工智能数据安全是一个覆盖多主体、多维度的全球性安全挑战和治理议题，因此报告主要从宏观战略、法律法规、标准规范、前沿技术和企业实践等维度，梳理当前的全球人工智能数据安全治理现状。

一 | 各国在战略层面高度重视人工智能的数据安全问题

目前，世界主要国家均高度重视人工智能的数据安全和隐私保护问题，集中表现为：

1. 多国均在人工智能发展战略中明确提出要重视数据安全

● **美国**：2016 年 10 月，美国连续发布《为人工智能的未来做好准备》和《国家人工智能研究和发展战略规划》两份报告，提出实施“人工智能公开数据”计划，要确保联邦数据、模型和计算资源的高质量、完全可追溯和可访问性，支持人工智能的技术开发、模型训练和安全测试。2019 年 6 月，美国发布新版《国家人工智能研发与发展战略计划》，要求所有机构负责人审查各自负责的联邦数据和模型，注重保护数据安全、隐私和机密性。

● **欧盟**：2018 年 3 月，欧洲政治战略中心发布《人工智能时代：确立以人为本的欧洲战略》，战略认为欧洲人工智能发展面临数据短缺和数据偏见等问题，提出要扩大人工智能系统所需数据源，设计利于欧洲数据收集、使用和共享的监管方案，确保《通用数据保护条例》（GDPR）个人数据保护要求实施的建议。2018 年 4 月，欧盟委员会发布《欧盟人工智能》政策文件，要求公共部门应当遵守欧盟关于个人数据保护的法律政策，并表示欧盟委员会将修订公共部门信息

开放指令，出台私营部门数据分享指南来确保数据安全。2020 年 4 月，欧盟委员会发布人工智能白皮书——《面向卓越和信任的欧洲人工智能发展之道》，报告表示基于人工智能对社会产生重大影响以及建立信任的需要，欧洲人工智能必须以欧洲的价值观和包含人类尊严和隐私保护等在内的基本权利为基础，同时在“信任的生态系统”的人工智能监管框架中，强调要从技术伦理、网络和数据安全、消费者权益和公民基本权利的角度出发，对高风险的人工智能应用进行识别和加强监管。

● **中国**：2017 年 7 月，中国国务院发布《新一代人工智能发展规划》，高度关注人工智能数据安全风险。强调要“强化数据安全与隐私保护，为人工智能研发和广泛应用提供海量数据支撑”以及“促进人工智能行业和企业自律，切实加强管理，加大对数据滥用、侵犯个人隐私、违背道德伦理等行为的惩戒力度”。

● **英国**：2018 年 4 月，英国政府发布《产业战略：人工智能领域行动》，提出要“开发公平、安全的数据共享框架；与公私部门的主要数据持有者及数据科学社区合作，确定数据共享障碍；与业界合作探索安全、公平的数据传输框架与机制”。

● **日本**：2018 年 4 月，日本发布第五版《下一代人工智能和机器人核心技术开发计划》，要开展下一代人工智能框架与核心模块研究，研究兼顾数据安全与隐私保护的数据获取技术，探讨复杂问题和复杂场景下人工智能多模块融合效率与性能提升的方法。

● **印度**：2018 年 6 月，印度发布《人工智能国家战略》报告，认为数据偏差将导致的算法决策缺乏中立性，建议“识别内置偏差，评估其影响，并找到减少数据偏差的方法”。同时报告建议建立数据保护框架和部门监管框架，并促进采用国际标准。关于隐私保护，报告呼吁“采取适当的措施来缓解隐私泄露风险，并强调使用人工智能情况下采取更高标准的隐私保护的重要性”。

2. 多国均在人工智能安全倡议和伦理规范中强调数据安全

● **官方机构**：2017 年 1 月，美国公共政策委员会发布《算法透明和可责性》声明，强调要确保人工智能算法的可责性、可解释性和可审查性，并对数据来源和公民隐私提供充分的保护和事后救济。2019 年 1 月，欧洲委员会 108 号公约咨询委员会发布《人工智能和数据保护指南》，指南旨在确保人工智能应用程序不会损害数据保护的权利。该委员会强调，保护人权，包括保护个人数据的权利，应该成为开发或采用人工智能应用程序时的必要先决条件，特别是当人工智能在决策过程中使用时。2019 年 4 月，欧盟委员会发布了《可信赖人工智能伦理指南》，指出人工智能系统必须确保隐私和数据保护，这既包括用户提供的信息，也包括用户在和系统交互过程中生成的信息，同时确保收集的数据不会用于非法地或不公平地歧视用户的行为。2019 年 6 月，中国国家新一代人工智能治理专业委员会发布《新一代人工智能治理原则——发展负责任的人工智能》，将“尊重隐私”作为八项原则之一，要求人工智能发展应尊重和保护个人隐私，充分保障个人的知情权和选择权；在个人信息的收集、存储、处理、使用等各环节应设置边界，建立规范；完善个人数据授权撤销机制，反对任何窃取、篡改、泄露和其他非法收集利用个人信息的行为。

● **社会组织**：2016 年 3 月，IEEE 标准协会（IEEE-SA）发布了《合乎伦理的设计（EAD）：将人类福祉与人工智能和自主系统优先考虑的愿景》，在第五部分“个人数据与个体访问控制”中指出，数据不对称是个人信息保护的一个重大道德困境，在算法时代，人工智能系统对个人数据的使用不断增强，为解决不对称问题，需完善个人信息保护政策，要以尊重个人数据完整性的方式设计并应用自主和智能系统。2018 年 9 月，中国世界人工智能大会发布《人工智能安全发展上海倡议》，提出人工智能发展需要保障用户的数据安全，不得以牺牲用户隐私为代价，需要加强数据保护立法，丰富人工智能的技术路线，不断强化人工智能应用中的用户隐私保护。2018 年 10 月，第 40 届数据保护与隐私专员国际大会发布《人工智能伦理与数据保护宣言》，提出应通过默认应用隐私原则和设计即隐私原则来对人工智能系统进行负责任地设计和开发，确保在确定处理方式和数据处理时，使数据主体的隐私和个人数据得到尊重。

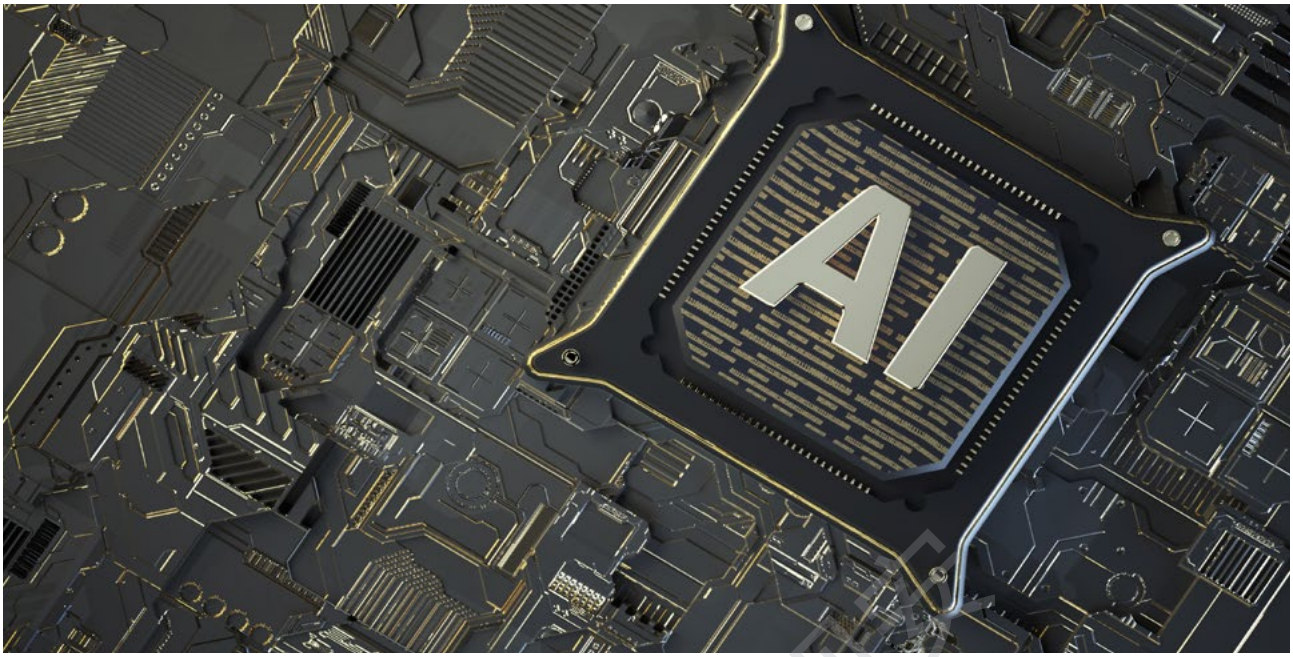
● **企业**：2016 年美国谷歌公司提出的人工智能“七原则”，在“隐私原则”中强调要给予用户通知和同意的机会，鼓励具有隐私保护的架构，并提供适当的透明度和对数据使用的控制。2017 年美国微软公司提出的人工智能“六原则”，在“隐私与保障”原则中强调设计人工智能时，必须要考虑智能隐私保护，必须要有先进的、值得信赖的保护措施，确保个人和群体的隐私信息安全。2017 年 1 月，阿西洛马人工智能 23 原则形成并发布，强调人工智能系统分析使用数据时，人类应当拥有对其自身产生的数据的访问、管理以及控制的权利；并且人工智能基于个人数据的应用不能削减人们真实的或者感知上的自由。

二 | 各国在人工智能数据安全法律层面上的立法模式差异较大

1. 美国：通过场景化立法规制人工智能数据安全

目前，虽然美国有多部联邦层级的数据安全草案在国会审议，但其短期内较难进入实质性立法程序，**因此美国当前依旧缺乏一部统一的、具备最高效力的国家数据安全保护法律。**在数据安全法律体系上，美国主要通过公民隐私保护、计算机通信安全、知识产权、金融商贸、医疗教育等不同法律部门法的相关条款和修正案对数据安全进行规制，同时依托州层面的地方立法和行业自律公约进行补充。**在人工智能数据安全方面，美国主要通过具体的场景化立法和州立法，对人脸识别、自动驾驶、隐私保护、精准推送、工业互联网等一系列人工智能行业应用进行数据安全监管。**

在人脸识别场景，美国近期立法动作频频，规制日趋严格。2020 年 2 月，美国加利福尼亚州众议院通过了《加州人脸识别技术法》，强调原则上不禁止私营主体与公共主体运用人脸识别技术，但要在保障公民隐私及自由与发挥人脸识别技术的公共服务优势方面寻求平衡。2020 年 3 月，美国华盛顿州议会通过了《人脸识别服务法》，法案旨在以造福社会的方



式促进人脸识别服务的使用，同时禁止相关应用威胁公民自由。由于华盛顿拥有微软和亚马逊两家全美最大的两家人脸识别软件开发公司总部，而加州则是美国谷歌、苹果等互联网巨头总部最重要的聚集地，**因此这两部地方性州立法将在事实上对美国人脸识别应用起到极大的规制作用。**此外，自 2019 年 5 月旧金山颁布全球首个禁止政府机构购买和使用人脸识别技术的法令以来，奥克兰、马萨诸塞州剑桥和波士顿等市议会也纷纷通过禁止政府使用人脸识别技术进行监控的相关法案。2020 年 6 月，美国参议院议员 Ed Markey 和 Jeff Merkley 共同向参议院提交了《2020 年人脸识别和生物特征识别技术禁令法案》，旨在禁止使用美国联邦资金采购联邦政府官员使用的人脸识别系统或“任何生物特征识别监控系统”。

在自动驾驶场景，自美国内达华州在 2011 年出台了全美第一部地方性自动驾驶安全法案以来，截止 2018 年底全美共有 36 个州通过州议会法案或是州长行政命令的方式对自动驾驶的认证、测试、部署和安全进行规制。2017 年 7 月，美国参众两院一致通过了《自动驾驶法案》，明确联邦政府和州政府在确保自动驾驶汽车安全方面的职责，并要求自动驾驶汽车生产商或者系统提供商需要向监管部门提交安全评估证明，以证明其自动驾驶汽车在数据、产品、功能等各个方面采取了足够的安全措施。同时要求自动驾驶汽车制造商必须制定隐私保护计划，明确对车主和乘客信息的收集、使用、分享和存储的相关做法，包括在收集方式、数据最小化、去识别化以及数据留存等方面的做法。

在隐私保护场景，截止 2019 年底，美国 50 个州都有不同完备程度的法律规制，通过规定在线互联网企业、电信增值业务企业的数据安全保护义务，来防止数据泄露、滥用和保护公民隐私。最具影响力的州隐私法案是加州 2020 年 1 月生效的《加利福尼亚消费者法》(CCPA)，因其能直接规制监管到像谷歌、脸书、易趣和推特等总部在加州的美国著名互联网平台企业，因此在美国数据安全和隐私保护方面的影响巨大。法案将可以基于合理分析推论的联系到个体的生物信息、能力智商、行为偏好和情感心理偏好等纳入个人信息范畴，并重点关注个人信息收集、买卖和共享三种活动，强调企业在进行个人信息处理分析和第三方共享利用的“opt-out”模式。

2. 欧盟：基于统一数据安全立法下的场景化监管

欧盟通过 2018 年 4 月生效的《通用数据保护条例》(GDPR) 框架构建了一整套统一完备的数据安全治理体系，对于全球各国数据规则制定都有着极大影响。**CDPR 通过对用户数据权利全面系统地明确梳理，对欧盟人工智能数据安全起到了基础性规制作用**，比如 GDPR 要求人工智能算法具有一定可解释性，同时第 22 条对包括用户画像在内的自动化决策行为的规定，明确当自动化决策产生的法律效力涉及数据主体或对其有重大影响时，数据主体有权随时反对企业使用其个人数据对其进行画像等自动化决策。如果自动化决策是合约必备且数据控制者征得数据主体的明示同意，数据控制者应当实施适当措施以保护数据主体的权利、自由和合法权益。**未来，欧盟将在 GDPR 框架下继续加快自动驾驶、人脸识别、精准推送和智能制造等人工智能重要应用场景的数据安全立法。**

2020 年 4 月，欧盟委员会发布《欧洲数据战略》，提出将就影响数据敏捷型经济体系中各主体关系议题探讨立法行动的必要性，解决包括企业间共生数据的共享 (物联网数据) 和建立数据池 (用于数据分析和机器学习) 的安全和信任问题。2020 年 6 月，欧盟委员会向欧洲议会和欧盟理事会提交《数据保护是增强公民赋权和欧盟实现数字化转型的基础——GDPR 实施两周年》报告，指出 GDPR 能够确保新技术的开发符合基本权利，尤其是在大型数字企业的在线广告和精准推送方面，GDPR 的有效实施是保护个人的重要要素。未来的挑战将在于阐明如何将行之有效的原则应用于需要持续监控的特定技术，包括人工智能、区块链、物联网和人脸识别等。2020 年 6 月，欧洲数据保护监管机构 (EDPS) 发布《EDPS 战略计划 (2020-2024) ——塑造更安全的数字未来》，在愿景中表示 EDPS 将积极关注可能对隐私和数据保护产生影响的数据处理实践和技术的发展，加强对特定新兴技术“发展前沿”的研究，包括生物识别技术、自动识别系统、量子计算、边缘计算和区块链等。同时，EDPS 要持续开发强大的技术工具和监督、审计、评估机制，针对数字生态系统中越来越“流行”的技术和工具，提供自动决策系统和人工智能处理个人数据的操作指南。最后，计划表示支持欧盟境内的公共场所暂停使用生物识别技术，并对此开展民主讨论。

在各个成员国方面，各国政府也不断推出针对人工智能具体应用场景的数据安全法律和政策。在自动驾驶领域，2017 年 5 月，德国联邦议会和联邦参议院共同通过了《道路交通法》修正案，允许“按规定使用”自动驾驶功能，同时明确了驾驶员使用该功能的权利义务以及驾驶数据的采集、存储、使用及删除规则。2017 年 8 月，英国政府于发布了《网联汽车和自动驾驶汽车的网络安全关键原则》，对数据和个人信息的安全存储、传输、处理和删除提出了明确要求。2019 年 2 月，欧盟成员国达成共识，共同签定自动驾驶指导文件，确定了包括行驶数据记录、网络安全及安全评估测试等在内的八项原则。在人脸识别领域，2018 年 7 月，比利时政府出台相关规定，禁止非授权的公共部门和私营部门使用人脸识别或其他基于生物特征的视频分析摄像机，主要针对非警方的私营企业和公共部门使用人脸识别摄像机带来的侵害公民隐私问题。

3. 中国：加快数据安全统一立法和人工智能场景化立法

2020 年以来，我国开始加快国家层面数据安全统一立法的速度。5 月 28 日，第十三届全国人民代表大会第三次会议通过《中华人民共和国民法典》，确立了数据和虚拟财产依法受到保护、公民个人信息和隐私权保护的基本原则。比如在《法典》第四编“人格权”的第六章“隐私权和个人信息保护”中，对自然人的隐私权，侵犯隐私的行为方式，自然人的个人信息定义，收集、处理自然人个人信息的原则、方式和限制，自然人的个人信息权利，信息收集、控制者的责任、义务和豁免等进行了明确规定。同时，针对人工智能人脸识别技术的应用，《法典》第 1019 条明令禁止“利用信息技术伪造侵害他人肖像权”，第 1023 条首次将声音作为人格权的肖像权保护客体。

6 月 1 日，十三届全国人大常委会第五十八次委员长会议审议通过了调整后的《全国人大常委会 2020 年度立法工作计划》，将《个人信息保护法》和《数据安全法》明确作为初次审议的法律案。6 月 28 日，第十三届全国人大常委会第二十次会议对《中华人民共和国数据安全法（草案）》进行了审议，并在 7 月 3 日正式对外发布征求意见。**《数据安全法（草案）》作为我国数据安全领域的顶层立法，将数据安全明确纳入到国家整体安全观中，对国家数据安全制度和主体数据安全保护义务进行了全面规定，将为我国人工智能数据安全治理奠定坚实的上位法基础。**

同时，**我国还高度关注人工智能重点应用场景的数据安全立法。**在 2017 年 7 月国务院发布的《新一代人工智能发展规划》中，明确提出要“制定促进人工智能发展的法律法规和伦理规范”，包括“重点围绕无人机、服务机器人等应用基础较好的细分领域，加快研究制定相关安全管理法规，开展与人工智能应用相关的民事与刑事责任确认等法律问题研究，建立追溯和问责制度”。

在国家人工智能发展战略的指引下，**我国相关部门在金融科技、智慧城市、无人机、自动驾驶等应用领域纷纷出台了相应的规范性文件，强调要加强人工智能的相关数据安全研究和管控。**如在金融科技领域，2018 年 4 月，中国人民银行、中国银监会、中国证监会和国家外汇管理局共同发布了《关于规范金融机构资产管理业务的指导意见》，在第二十三条对运用人工智能技术开展投资业务进行了相应规定，要求金融机构应当向金融监督管理部门报备人工智能模型的主要参数以及资产配置的主要逻辑，并向投资者充分提

示人工智能算法的固有缺陷和使用风险。2019 年 8 月，中国人民银行发布《金融科技（FinTech）发展规划（2019-2021 年）》，提出要“加强金融领域人工智能应用潜在风险研判和防范，完善人工智能金融应用的政策评估、风险防控、应急处置等配套措施，健全人工智能金融应用安全监测预警机制，研究制定人工智能金融应用监管规则，强化智能化金融工具安全认证，确保把人工智能金融应用规制在安全可控范围内”。

整体来看，我国目前尚未形成体系完善的人工智能数据安全法律法规。虽然《数据安全法》和《个人信息保护法》出台在即，但其落实尚需要一系列配套法规、部门规章和规范性文件提供支撑。同时由于上位法尚未出台，人工智能场景化立法的步伐也相对滞后，数据安全并未在相关人工智能应用行业的规范性文件中得到足够的重视和明确的规制要求。**未来，我国还需要在《网络安全法》、《数据安全法》（待出台）和《个人信息保护法》（待出台）的体系框架下，加快生物特征识别、工业互联网、智能网联汽车、数字内容精准推送等人工智能重点应用领域的场景化立法，构建完备的人工智能数据安全法律体系。**

三 | 全球纷纷加快人工智能数据安全的标准制定

1. 国际标准组织：加快推动国际人工智能数据安全标准制定

● **ISO/IEC JTCI：**2017 年 10 月，ISO/IEC JTCI 在俄罗斯召开会议，成立人工智能分委员会（即JTC1 SC42），负责人工智能标准化工作。目前，SC42 已经成立了包括基础标准（WG1）、大数据（WG2）、可信赖（WG3）、用例与应用（WG4）、系统计算方法和计算特征（WG1）等五个工作小组。此外，SC42 还包括人工智能管理系统标准咨询组（AG1）和智能系统工程咨询组（AG3），不断加快人工智能数据安全方面的国际标准制定工作。

ISO/IEC 主导的系列标准有：1）ISO/IEC TR 24027《信息技术 人工智能 人工智能系统中的偏差和人工智能辅助决策》，系统梳理了人工智能系统和自动化决策中可能引起算法偏见的要点，并提出解决方案。2）ISO/IEC TR《信息技术 人工智能 人工智能可信度概述》，提出人工智能可信赖的内涵，梳理分析人工智能系统的典型工程问题和威胁可信赖的风险，并提出解决方案。3）ISO/IEC TR《信息技术 人工智能 风险管理》，系统梳理了人工智能多维度的风险，包括数据安全、隐私保护、数据质量等，并提出人工智能风险管理的流程方法。

● **IEEE**：IEEE 标准协会设立了相关工作小组开展一系列人工智能数据安全标准制定工作，如 IEEE P3652.1 联邦学习基础框架与应用工作组负责联邦学习的相关安全标准化工作。

IEEE 主导的系列标准：1）IEEE P7002《数据隐私处理》：为如何对收集个人信息的系统和软件的伦理问题进行管理提出指导性意见，提供了规范系统 / 软件的工程生命周期中隐私问题的最佳实践，支撑企业对自身隐私实践进行隐私影响的合规性评估。2）IEEE P7003《算法偏差注意事项》：提供了在创建算法时消除负偏差问题的步骤，包括基准测试程序和选择验证数据集的规范，为智能系统的开发者提供如何避免代码、模型中的负偏差的参考。3）IEEE P7006《个

人数据人工智能代理标准》：梳理了在机器自动化决策时，创建和授权防伪个人化人工智能所需要的技术要点，包括由个人控制的输入、学习、伦理、规则和价值。允许个人为其数据创建“个人条款和条件”，为人工智能代理者如何为用户提供一种管理和控制其数字身份的方式提供参考。

● **ITU-T：**ITU-T 在就人工智能的生物特征识别、智能汽车、内容治理、智慧医疗等应用中的安全问题开展标准制定工作。包括 ITU-T SCG 多媒体研究组、ITU-T SG17 安全研究组、ITU-T SG1 安全标准工作组下设的 Q9“远程生物特征识别问题组”和 Q10“身份管理架构和机制问题组”等。其中，Q9 重点关注生物特征数据的隐私保护、可靠性和安全性等问题。

2. 美国：强调通过标准制定确保其全球人工智能的领导地位

2019 年 2 月，美国总统特朗普发布 13859 号行政令，指示联邦机构应确保美国保持在人工智能中的领导地位，提出“确保技术标准能够反映联邦在创新、公众信任和公众对使用 AI 技术的系统的信任方面的优先事项，并制定国际标准以促进和保护这些优先事项。”2019 年 8 月，美国国家标准与技术研究院（NIST）发布了《美国如何领导人工智能：联邦参与制定技术标准及相关工具的计划》，旨在落实总统行政令的要求。《计划》认为美国在人工智能领域的全球领导地位取决于联邦政府在人工智能标准制定中发挥积极的推动作用，并确定了人工智能标准的九个重点领域，包括概念和术语、数据和知识、人机互动、指标、网络、性能测试和报告方法、安全、风险管理和可信赖。因此，美国积极参与 ISO/IEC 等国际标准组织的人工智能、大数据等工作组的相关标准研究、制定和推广工作，主导和召集了一系列涉及人工智能数据安全标准的工作。

2020 年 1 月，NIST 发布《隐私框架 1.0 版：通过企业风险管理来提升隐私的工具》，旨在为相关组织对隐私风险进行评估管理，提升自身个人数据保护能力提供支撑。隐私框架由 NIST 召集的众多利益相关者共同开发，包括：1）核心层，帮助组织确定隐私保护的预期目标和各行动的优先级；2）概况层，帮助组织识别并管理隐私风险，满足组织的隐私保护目标和业务风险防范的需求；3）实施层，帮助组织通过对资源和流程的协调管理，实现隐私保护。基于 NIST 在全美标准届的权威地位，《隐私框架 1.0》将成为美国各大互联网企业开展隐私保护工作的主要参考，为美国人工智能数据安全治理提供基础性的标准支撑。

3. 欧盟：欧盟和成员国共同参与人工智能数据安全标准制定

目前，欧盟的人工智能数据安全标准包括欧盟层面的指南参考和各成员国具体制定的国家标准。2017 年 12 月，欧盟网络与信息安全局（ENISA）发布了《移动应用中的隐私和数据保护——应用开发生态系统与 GDPR 技术实施研究》，基于移动应用程序中的数据安全和隐私目标提出了在移动 APP 中实施“设计即隐私”理念的建议，其中涉及众多用户画像和自动化决策场景。2018 年 12 月，欧盟网络与信息安全局（ENISA）发布了《自动代理中的安全和隐私——为网络安全政策制定形成框架》，对人工智能和自动代理（Autonomous Agents）中存在的数据安全和隐私问题进行了梳理，并提出了对策建议。2019 年 6 月，欧盟基本权利局（FRA）发布了《数据质量和人工智能 - 减轻偏见和错误，以保护基本权利》，为评估数据质量提供了指导。

同时，2020 年 2 月，欧盟理事会发布人工智能白皮书《面向卓越和信任的欧洲人工智能发展之道》，强调欧洲人工智能治理结构必须是和成员国国家的主管部门开展合作，从而避免责任的碎片化，提升成员国能力，包括识别新兴趋势、标准化和认证活动等。以德国为例，2018 年 7 月，德国联邦政府通过了《联邦政府人工智能战略要点》文件，在十项目标的第八点中明确提出要“在国际标准化委员会中强有力地代表欧洲共同利益；致力于公开和国际标准的制定。”2019 年 8 月，德国标准协会（DIN）表示目前正加紧研制人工智能标准化路线图。标准化路线图将总结出人工智能领域内现有的规范和标准，特别是将会提出人工智能未来必要发展的建议，旨在提高德国在国际层面的影响力和竞争力，促使欧洲价值观成为全球基准。

4. 中国：高度重视人工智能行业场景化的数据安全标准建设

目前，我国工信部、全国信息安全标准化技术委员会（SAC/TC260）、中国通信标准化协会（CCSA）等国家部委和标准化组织，高度重视人工智能数据安全的相关标准制定工作。2018 年 1 月，国家标准化管理委员会正式成立国家人工智能标准化总体组，承担人工智能表转化工作的统筹协调和规划布局。

2020 年 3 月，国家工信部发布《网络数据安全标准体系建设指南（意见征求稿）》，明确将人工智能列为数据安全标准体系建设的重点内容。2020 年 3 月，全国信息安全标准化技术委员会发布《全国信息安全标准化技术委员会 2020 年度工作要点》，明确提出要“积极应对新技术新应用带来的国家网络安全挑战，研制 5G 安全、人工智能安全、物联网安全、区块链安全等领域新技术标准。”

同时在人工智能具体场景上，2019 年 5 月，国家工信部组织全国汽标委编制了《2019 年智能网联汽车标准化工作要点》，在重点内容的第三项中明确提出要“有序推进汽车信息安全标准制定。完成汽车信息安全通用技术等基础通用及行业急需标准的制定，研究提出汽车软件升级、信息安全风险评估等应用类标准的立项，系统开展汽车整车及零部件信息安全测试评价体系研究。”

目前，我国涉及人工智能数据安全的在研或已发布标准大致有三类，一种是基于数据安全、隐私保护视域下的数据安全标准，能够对人工智能数据安全治理提供基础性技术参考；一种是基于人工智能技术开发和工程应用的视域下的安全标准，针对终端、环境、平台和算法等不同工程环节提出数据安全要求；一种是针对生物特征识别、自动驾驶、工业互联网、物联网、智能家居等人工智能各种应用场景的行业性标准规范。具体见下表 2

表 2：中国人工智能数据安全相关标准（包括立项、在研和已发布）

标准类型		标准名称	归口单位	相关内容
通用性标准	数据安全和隐私保护	《信息安全技术个人信息安全规范》	SAC/TC260	对个人信息的识别、采集、保存、共享、使用等全生命周期的提出了整体性合规要求，为人工智能的数据安全和隐私保护提供重要参考
		《信息安全技术大数据服务安全能力要求》	SAC/TC260	对企业数据安全能力建设提出了整体性要求，为人工智能企业提高自身数据安全能力提供借鉴
		《信息安全技术数据安全能力成熟度模型》	SAC/TC260	从安全能力维度、能力成熟度维度和数据安全过程维度构建了 DSMM 架构。并具体分析了数据采集安全、数据传输安全、数据存储安全、数据处理安全、数据交换安全、数据销毁安全，以及通用安全的相关内容
	人工智能技术开发和工程应用	《信息安全技术 移动智能终端个人信息保护技术要求》	SAC/TC260	为移动智能终端的个人信息采集、处理、存储等一系列活动的安全保护和隐私合规提供参考
		《信息安全技术人工智能算法安全指南》	SAC/TC260	为人工智能算法模型的设计、优化、迭代和运行的安全提供参考
		《信息安全技术人工智能应用安全指南》	SAC/TC260	为人工智能应用的产品和服务安全提供参考
		《人工智能终端产品个人信息保护要求和评估方法》	CCSA	对人工智能终端产品的数据采集原则、隐私框架设计、安全评估要求提供参考
		《人工智能终端设备安全环境技术要求》	CCSA	为人工智能终端设备的部署、环境、提升等环节提供参考
		《人工智能服务平台数据安全要求》	CCSA	对提供人工智能算法开发优化、数据标注分析等平台机构的数据安全管理与评估提出要求
场景标行业应用准	生物识别领域	《信息安全技术生物特征识别信息的保护要求》	SAC/TC260	系统梳理了生物特征识别系统的威胁，对生物特征信息与身份主体之间的绑定关系、系统正常运行、应用模型开发以及隐私保护提出安全要求
		《信息安全技术 网络人脸识别认证系统安全技术要求》	SAC/TC260	规定了安全防范领域的利用人脸识别进行视频监控系统的基本构成、功能要求、性能要求和测试方法
		《信息安全技术虹膜识别系统技术要求》	SAC/TC260	基于信息安全等级保护的要求，对使用虹膜识别系统进行身份鉴别的系统设计、实现、测试和管理提出整体安全要求
		《信息安全技术指纹识别系统技术要求》	SAC/TC260	对指纹识别系统的安全威胁、安全目的进行分析，并提出指纹识别系统的安全技术要求
		《信息安全技术基于可信环境的生物特征识别身份鉴别协议》	SAC/TC260	规定了可信环境的标准，并提供可信环境下的生物特征识别身份鉴别协议的模板，包括协议框架、协议流程、协议要求和协议接口

表 2：中国人工智能数据安全相关标准（包括立项、在研和已发布）

标准类型		标准名称	归口单位	相关内容
场景标行业应用准	智能网联汽车	《信息安全技术 汽车电子系统网络安全指南》	SAC/TC260	为汽车电子系统的数据存储、流动和处理等安全管理提供实践指导
		《信息安全技术 车载网络设备信息安全技术要求》	SAC/TC260	旨在解决智能网联汽车行业中关于车载网络设备信息安全技术要求标准问题。
		《车联网信息服务 数据安全技术要求》	CCSA	针对智能网联汽车行业的信息服务，提供数据安全管理参考
		《车联网信息服务 用户个人信息保护要求》	CCSA	针对智能网联汽车对用户个人信息采集、处理和加工，提出安全保护要求
		《基于移动互联网的汽车用户数据应用与保护技术要求》	CCSA	针对移动互联网环境下的汽车行业对用户数据采集、处理和应用，提供安全管理和技术部署的参考
		《基于移动互联网的汽车用户数据应用与保护评估方法》	CCSA	为监管方和第三方对移动互联网环境下汽车行业的用户数据应用和保护规范，提供评估框架参考
	工业互联网 / 物联网	《工业互联网数据安全保护要求》	CCSA	针对工业互联网整体的数据安全提出要求
		《工业互联网安全能力成熟度评估规范》	CCSA	对工业互联网企业的安全能力评估提供框架参考，高度强调数据安全
		《信息安全技术 物联网数据传输安全技术要求》	SAC/TC260	对物联网环境下的数据传输、交互和共享，提出安全技术要求

四 | 全球人工智能数据安全前沿技术方向与实践

数据安全和隐私保护技术的突破研发和落地应用，能够极大地提高政府和企业人工智能应用中的数据安全能力。目前国际上致力于此类技术研究的主体主要有两类，一是以谷歌、微软、腾讯、阿里巴巴为代表的全球互联网巨头，投入建设了大量人工智能实验室和研究所，如 Microsoft Research、Google Brain、Intel AI、Visa Research 等。二是以伯克利大学、斯坦福大学、麻省理工学院、清华大学、上海交大等为代表的全球知名理工类高校。

下表 3 为目前全球数据安全前沿技术方向，报告重点聚焦于目前企业实践中应用较多的前沿技术，包括全同态加密、多方安全计算、差分隐私、联邦学习和区块链。同时，报告还关注了当前人工智能数据安全检测和对抗领域的专用技术方向，包括数据偏见检测、防训练数据集污染和防对抗样本攻击技术。

表 3：全球人工智能数据安全前沿技术方向与实践

技术路线	核心原理	技术特点	应用场景	存在问题
全同态加密	通过加密算法设计，确保对加密数据计算后的加密结果与明文计算结果一致，并向合作双方输出加密结果	可在不解密的情况下对密文进行计算和分析	数据敏感性较高，缺乏可信第三方的简单数据分析场景	全同态加密的消耗计算成本较高；技术研发难度高；验证速度慢，大规模落地可行性较差
多方安全计算	在无可信第三方的情况下，通过计算协议和约定函数的设计，确保双方分别输入原始数据得到共同的正确计算结果	输入隐私性、计算正确性、去中心化性	多方联合数据计算、数据安全查询、数据可信交换等	难以确保参与方的诚实性，当恶意参与方超过一定比例时无法得出正确结果；系统协调和验证效率较低
差分隐私	将数据和代码混淆成无法理解的形式，但保留原有功能和可计算性	本质是密码技术，对计算环境要求较低	不可信环境下的数据传输、计算	影响数据准确性，验证成本高
联邦学习	本地进行 AI 模型训练，然后仅将模型更新的部分加密上传到数据交换区域，并与其他各方数据的进行整合	数据隔离、模型质量无损、地位对等、共同获益	AI 联合训练	对于数据与算法模型交互传输的带宽、延时要求更高
区块链	区块链是建立在互联网之上一个点对点的公共账本，由区块链网络的参与者按照共识算法规则共同添加、核验、认定账本数据	去中心化性、自信任性、防篡改性	数据的声明发布、授权使用等	区块链自身的安全缺陷，比如针对区块链的安全攻击，缺乏有效技术监管
零知识证明	通过算法逻辑设计，确保证明者能够在不向验证者提供任何有用信息的情况下，使验证者相信特定验证目标的准确性	验证正确性、数据保密性	多方联合数据计算、数据安全查询、数据可信交换等	业务场景下算法设计难度较大，验证速度慢，可扩展性差
可信执行环境	包括可信硬件（TEE）和可信计算协议（TCP），通过划分内存区域等硬件手段构造计算沙盒，提供整体的可信保密计算环境	数据隔离、可信执行环境 TEE、操作监控和历史回放	监管沙箱、计算沙箱、多方数据联合计算	过于依赖硬件厂商（Intel 的 SGX 和 AMD 的 PSP）

1. 隐私计算

中国科学院信息工程研究所副总工程师李凤华团队在 2016 年率先提出了“隐私计算(Privacy Computing)”概念，并将其定义为“**面向隐私信息全生命周期保护的计算理论和方法，是隐私信息的所有权、管理权和使用权分离时隐私度量、隐私泄漏代价、隐私保护与隐私分析复杂性的可计算模型与公理化系统。**”[17]

在当下的安全实践中，隐私计算通常是指**在数据全程保密或无接触的情况下，确保合作双方能够对数据进行计算、比对、运行等并读取和利用结果，并保证任何一方均无法得到除应得的计算结果之外的其他任何信息。**目前，隐私计算包括全同态加密(Full Homomorphic Encryption, FHE)、多方安全计算(Secure Multi-Party Computation, sMPC)、差分隐私(Differential Privacy)、联邦学习(Federated Learning)、零知识证明(Zero-knowledge Proof)等多种技术方向。

● **同态加密：同态加密是指对加密数据进行处理得到一个输出，将此输出进行解密，其结果与用同一方法处理未加密原始数据得到的结果一致。**在同态映射下，先运算后加密和先加密后运算，得到的结果相同。同态加密算法从功能上可分为部分同态算法和全同态算法。1) 部分同态是指支持加法同态或者乘法同态或者两者都支持但是操作次数受限；2) 全同态算法则可简单理解为能不受限制地同时支持加法和乘法操作，从而完成各种加密后的运算(如加减乘除、多项式求值、指数、对数、三角函数等)。

> **实践 1：**英特尔于 2018 年 12 月推出的开源版 HE-Transformer，即利用同态加密技术，使得机器学习算法能够处理加密隐私数据。但目前该技术在应用过程中面临的挑战是会大大延长计算时间，导致训练过程低效[18]。

● **多方安全计算**：多方安全计算主要针对无可信第三方情况，安全地进行多方协同计算问题。**即在一个分布式网络中，多个参与实体各自持有秘密输入，各方希望共同完成对某函数的计算，而要求每个参与实体除计算结果外均不能得到其他用户的任何输入信息。**从计算场景上，可以将安全多方计算分为特定场景和通用场景。特定场景是指针对特定的计算逻辑，比如比较大小，确定双方交集等。具体场景可以采用多种不同的密码学技术设计协议。通用场景是指安全多方协议的设计要具有完备性，可以理论上支持任何计算场景。目前采用的方法主要是加密电路，不经意传输以及同态加密。通用的两方计算已经具备了商用的条件。多方计算在某些特定场景下也已经没有太多的性能瓶颈，而通用计算协议在可扩展

性层面依然不成熟，这也是学术界一直在探索的方向。

当前，多方安全计算的主要适用场景包括：1）数据安全查询。使用安全多方计算技术，能保证数据查询方仅得到查询结果，但对数据库其他记录信息不可知。同时，拥有数据库的一方，不知道用户具体的查询请求。2）联合数据分析。改进已有的数据分析算法，通过多方数据源协同分析计算，使得敏感数据不被泄露。

● **差分隐私：差分隐私是一种被广泛认可的隐私保护技术，通过对数据添加干扰噪声的方式保护数据中的隐私信息。**当对用户数据进行训练时，差分隐私技术能够提供强大的数学保证，保证模型不会学习或记住任何特定用户的细节。在许多场景下机器学习涉及基于敏感数据进行学习和训练，例如个人照片、电子邮件等。理想情况下，经过训练的机器学习模型的参数代表的应该是一般模式，而不是关于特定训练示例的事实。为了确保训练数据中的隐私得到有效的保护，可以使用差分隐私技术。2016 年，研究者提出基于差分隐私的深度学习算法，利用随机梯度下降过程中对梯度增加扰动来保护训练敏感数据 [19]。但在某些情况下，由于添加了噪声，差分隐私技术可能会导致精度受到影响 [20]。

实践 2：Apple 通过差分隐私技术，保护用户共享给 Apple 的信息 [21]。具体而言， Apple 分析用户数据之前，利用差分隐私技术为数据添加随机信息，使得 Apple 无法将这些数据与用户设备进行关联。只有当单个用户的数据与大量其他用户的数据相结合，平均掉随机添加的信息时，相关的模式才会显现。而这些模式，能够帮助 Apple 深入了解人们如何使用他们的设备，同时避免收集与个人相关的信息。

● **联邦学习：联邦学习是指本地进行 AI 模型训练，然后仅将模型更新的部分加密上传到数据交换区域，并与其他各方数据的进行整合。**其技术特点包括：1）数据隔离：数据不会泄露到外部，满足用户隐私保护和数据安全的需求；2）模型质量无损：不会出现负迁移，保证联邦模型比割裂的独立模型效果好；3）地位对等：合作过程中，合作双方是对等的，不存在一方主导另外一方；4）共同获益：无论数据源方，还是数据应用方，都能获取相应的价值。

当前，联邦学习技术的主要类型包括：1）横向联邦学习：数据集共享相同特征空间但样本不同。例如，两个区域银行可能具有与其各自区域不同的用户组，并且它们的用户的交集非常小。但是，它们的业务非常相似，因此要素空间相同；2）纵向联邦学习：两个数据集共享相同的样本 ID 空间但特征空间不同；3）迁移联邦学习：两个数据集不仅在样本上而且在特征空间上都不同的情况。考虑两个机构，一个是位于中国的银行，另一个是位于美国的电子商务公司。由于地理位置的限制，两个机构的用户群体之间的交叉点很小。另一方面，由于业务不同，双方的特征空间只有一小部分重叠。

目前，联邦学习主要应用于 AI 联合训练。通过利用联邦学习的特征，为多方构建机器学习模型而无需导出企业数据，不仅可以充分保护数据隐私和数据安全，还可以获得到更好的训练模型，从而实现互惠互利。

实践 3：谷歌推出 TensorFlow Federated learning 在用户设备上进行模型训练。针对基于用户与移动设备的交互进行训练的机器学习模型，2017 年，谷歌发布了应用于移动设备的联邦学习算法，可以将模型训练引入移动设备中，同时确保所有用于模型训练的用户数据保存在设备上。

2. 区块链

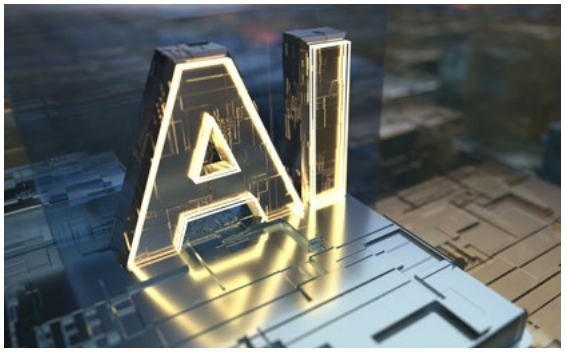
● **技术框架**：区块链是建立在互联网之上一个点对点的公共账本，由区块链网络的参与者按照共识算法规则共同添加、核验、认定账本数据。网络中每个参与者都拥有一个内容相同的独立账本，且账本数据是公开透明的。目前区块链应用主要有三种模式：1）公有链是运行在互联网的完全分布式区块链；2）联盟链则是由多个关联机构共同发起和运营，带有准入机制；3）私有链是公有链的私有化部署，往往由单个机构主持运行。

● **技术特点**：1）去中心化：在区块记录生成过程中，区块链参与方的权利和义务平等。去中心化同时也是多中心化，即在部分节点失效，甚至恶意化的情况下，仍能保证区块链的正常运行；2）自信任：区块链所有节点之间无需信任也可以进行交互。因为区块链账本的存储是多副本的，合约执行和记录添加是基于公开的机器代码和共识机制，所以节点的任何行为都是可预期的；3）防篡改：已生成的区块链记录由全体成员共同保存。而且任何节点的本地账本都自动与共识版本对齐。在一定的规则和时间范围内，区块记录的更改行为都是不可实现的。

● **适用场景**：区块链技术因其独特的密码学机制和共识机制可实现虚拟资产的确权、授权、停权。在数据流动领域，可用于数据确权、数据完整性校验、和数据的追踪溯源等。

3. 人工智能数据偏见检测

训练数据的不足和偏见会导致 AI 系统产生偏见。当前，已有许多企业和学术机构开始研究如何检测和解决训练数据中的偏见问题，并已取得了一定成果。例如，麻省理工学院的研究人员开发了一种算法来减轻训练数据中隐藏的、以及潜在未知的偏见[22]。这种算法将原始学习任务与变分自编码器相融合，以学习训练数据集中的潜在结构，然后自适应地使用所学习到的潜在分布，在训练过程中重新加权特定数据点的重要性。通过无监督的方式学习潜在的数据分布可以帮助发现训练数据中隐藏的偏见，例如训练数据集中代表性不足的数据种类，再通过增加算法采样这些数据的概率来避免偏见被引入 AI 系统中。研究人员通过该技术有效解决了人脸识别系统中的种族和性别偏见问题。



实践 4：为有效检测和消除机器学习模型和数据集中的偏见，IBM 开发了一个开源工具包 AI Fairness 360。这个可扩展的开源工具包可以帮助用户在整个 AI 应用生命周期中检查、报告和减轻机器学习模型中的歧视和偏见，包括 70 多个用于测试偏见的数据集和模型度量指标，以及 10 个用于减轻数据集和模型中偏见的算法[23][24]。在 10 个算法中，有 4 个算法是用来消除数据集偏见的，包括 Optimized Preprocessing、Reweighing、Disparate Impact Remover 和 Learning Fair Representations。其中 Reweighing 算法通过改变不同训练样本的权重来消除训练数据集中的偏见；Optimized Preprocessing 算法通过改变训练数据的特征和标签来消除数据集中的偏见。

实践 5：在上海市人工智能产业安全咨询委员会的指导下，2020 年 7 月，上海观安信息技术股份有限公司联合多家高校、厂商和测评机构，开发了人工智能数据安全风险评估平台，针对特定人工智能应用场景中的数据安全风险进行第三方评估评级。平台通过设定安全基线，开发用于敏感数据探测、数据质量检测、数据差异检测、漏洞检测以及脆弱性检测的工具，对基于检测工具汇集的数据实现数据安全风险信息实时收集、自动推送、智能分析、量化评估与诊断分级，针对人工智能应用场景中的数据安全实现多层级、多维度风险评估，为企业对人工智能系统开展自评以及第三方测评机构针对人工智能项目开展风险评估和产品认证提供技术、工具和平台。

4. 人工智能数据安全对抗

● **防训练数据集污染：**针对通过污染训练数据集以达到影响算法决策的攻击类型，目前存在三种技术可以防御此类攻击，包括训练数据过滤、回归分析和集成分析方法[25]。其中训练数据过滤是通过检测和净化的方法实现对训练数据集的控制，防止训练数据集被注入恶意数据；回归分析是基于统计学方法，检测数据集中的噪声和异常值；集成分析是通过采用多个独立模型构建综合 AI 系统，来减少综合 AI 系统受数据污染的影响程度。

● **防对抗样本攻击：**应对针对现场数据的对抗样本攻击当前可采用的防御方法包括：网络蒸馏、对抗训练、对抗样本检测、输入重构、深度神经网络模型验证等。其中对抗训练技术可通过在模型训练阶段，使用已知的攻击方法生成的对抗样本，对模型进行重训练，改进模型的抗攻击能力；对抗样本检测技术是在模型运行阶段，通过特殊的检测模型对现场数据进行判断，检测现场数据是否包含对抗样本；输入重构技术是指在模型运行阶段，对样本进行重构转化，以抵消对抗样本的影响。

PART 3

我国人工智能数据安全治理框架

治理 (Governance) 是公共政策领域的概念。1995 年联合国全球治理委员会 (CGG) 将治理定义为 “各种公共或私人的机构或个人管理其共同事务的方法总和,使相互冲突的或不同利益得以调和,并采取联合行动的持续过程,既包括权威性的秩序和正式制度, 也包括各种非正式的规则。” 根据现代治理理论, 治理一般具有以下特征: 1) 治理不仅是一整套规则或活动, 而是一个持续性的互动过程; 2) 治理的基础不是权威控制, 而是利益协调; 3) 治理涉及到公共部门、私人部门等多个主体。目前, 治理概念被广泛应用到公共政策制定的各个领域, 全球相关公共部门、私营企业、社会研究机构和学者均提出过 “人工智能治理” 、“数据安全治理” 等类似概念。

中国软件评测中心发布《电信和互联网行业数据安全治理白皮书 (2020 年) 》认为, 数据治理是多元治理主体以数据生产要素为对象, 以释放数据价值为目标, 以守住数据安全为底线, 以建立健全数据全生命周期秩序规则为核心, 以推动数据有序管理和流转为主要活动, 以强化数据管理技术手段为支撑的一系列活动, 具有综合性、复杂性和长期性等特征。而数据安全治理是数据治理的一个重要组成部分, 贯穿数据治理各个过程及数据全生命周期, 聚焦数据的 “安全” 属性, 即 “让数据使用更安全” 。

报告基于人工智能发展转型的阶段性特点, 以及人工智能数据安全挑战的特性, 充分吸收现有的人工智能治理和数据安全治理理念, 在 2019 年《人工智能数据安全风险与治理》报告的基础上, 结合全球最新治理实践和我国实际情况, 构建了包括治理思路、治理原则和治理路径在内的综合性人工智能数据安全治理框架, 如下图 2。

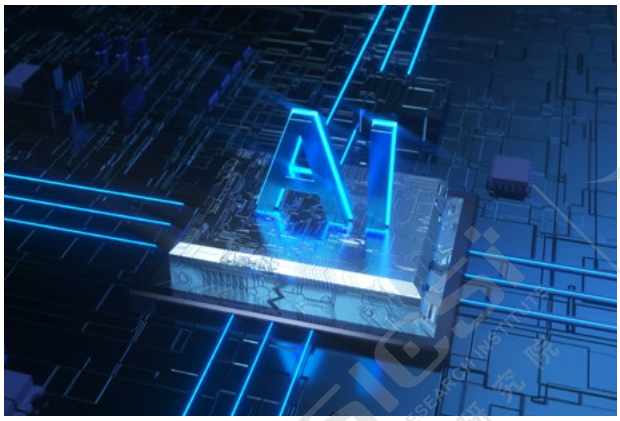
图 2：人工智能数据安全治理框架

一 | 治理思路

人工智能数据安全挑战会随着人工智能技术的发展突破、应用行业的不断深入扩大等因素加快演变，其对于人类现实社会的外溢和威胁将是一个复杂的长期过程。因此，我国必须要在**人工智能的动态发展中实现对数据安全风险整体的可知可控，确保人工智能数据在采集、标注、处理、存储、流动、共享和场景应用的全生命周期安全，不断提高人工智能企业的数据安全能力，增强人工智能数据安全供给链的连续性和可用性。**

同时，我国要充分发挥“安全”对“发展”的赋能作用，通过完备的安全治理降低数据流动中的技术和法律壁垒，实现数据安全、便捷、低成本的互通和利用，**赋能人工智能技术和产业的全面新发展，总结具有中国特色的人工智能数据安全治理范式**，提高我国在人工智能数据安全领域的国际话语权和影响力，引领全球人工智能和数据安全规则制定。

二 | 治理原则



● **鼓励 AI 发展：**加强数据安全管理是为了鼓励数据流动、促进人工智能发展而非限制发展，因此我国人工智能数据安全治理应当以鼓励数据安全、便捷、低成本的共享流动，促进人工智能健康发展为前提和根本，同时相关法律法规和规制政策应当关注政府的执法监管成本和企业的合规成本，避免因过于严苛的安全管理抑制技术创新和产业发展。

● **保护主体权益：**我国人工智能数据安全治理应当充分保障各主体的数字权益，包括公民的隐私权和知情权，企业的知识产权和商业秘密等，确保人工智能发展应用中的公平正义，增进社会整体福祉。

● **重视技术赋能**：人工智能数据安全治理应当高度重视前沿数据安全和隐私技术的赋能作用，因此我国应当鼓励推动人工智能技术本身在隐私保护、数据检测、网络安全和攻防对抗中的创新应用，构造内嵌安全式的人工智能发展。

● **聚焦行业场景：**人工智能数据安全挑战带有人工智能发展阶段性和应用场景化的独特性，因此我国应当充分结合人工智能技术发展趋势、产业成熟度和应用渗透度，针对行业场景化应用的数据安全特点进行分步推进治理措施，避免“一刀切”。

● **推动多元参与：**人工智能数据安全治理涉及多个环节的不同主体，因此我国要充分调动各类主体的积极性。比如在人工智能数据安全的标准制定和人工智能安全供给方面，都需要吸纳人工智能产业生态的各类企业、网络安全和数据安全厂商、科研院所和智库机构等各个不同主体的共同参与。

● **加强国际合作：**随着数字经济全球化和大型互联网跨国公司的发展，人工智能数据安全已经成为了一个全球性的挑战，需要各国在尊重和开放基础上充分加强在形成国际共识、制定国际标准、对接法律框架、监管跨国公司等领域的国际合作。因此我国应坚持对外开放和国际合作，探索引领全球人工智能和数据安全规则制定。

三 | 治理路径

1. 加快完善人工智能数据安全治理的顶层设计

顶层设计是指我国决策层从国家全局和长期的角度，围绕人工智能数据安全治理议题形成整体的治理目标、治理策略和治理体系。具体而言，应当包括以下三个层次：

● **宏观战略**：在全球各国普遍高度重视人工智能发展的情况下，人工智能数据安全作为人工智能安全的核心组成构建，我国决策层应当将人工智能数据安全纳入到国家整体安全观当中，作为人工智能战略和发展规划的重要板块。

● **法律法规**：人工智能数据安全问题是人工智能安全发展与数据安全两个议题的深度交汇，我国应当加快推进《数据安全法》和《个人信息保护法》立法，同时重点关注人工智能在重要行业领域的前沿应用，通过“场景化立法”的策略解决具备特殊性的人工智能数据安全挑战，当下可重点聚焦于人脸识别、平台监管、自动化决策等。

● **监管机制**：由于人工智能数据安全问题将会威胁到用户隐私、公民权益、商业秘密、知识产权、社会公平、国家安全等各个领域，我国决策层应当构建中央国家安全领导机构的统一指导下，由各部门、地方的主管部门负责主管行业、地域的人工智能数据安全的监管机制，建立包括指导、检查、整改、约谈、罚款等立体式的执法监管工具箱。

2. 重点聚焦人工智能数据安全的标准体系建设

在我国数据安全法律体系尚处于加快建设和完善的情况下，人工智能数据安全标准因其政策位阶更偏落地，在制定程序上更加便捷、探索试错成本更低，对于我国政府人工智能数据安全监管和企业数据安全能力提高更具现实意义和指导作用。因此，**人工智能数据安全标准应当作为我国在人工智能当前的发展阶段下，进行数据安全治理的主要政策工具。**具体而言，标准体系应当包括以下三个层次：

● **标准计划**：国家标准化部门应当在人工智能安全标准制定规划和数据安全标准制定规划，重点关注人工智能安全与数据安全的交汇处，形成一系列清晰明确、可执行落地的标准制定计划。

● **通用标准**：即针对人工智能技术形成一整套的数据安全通用标准，包括人工智能芯片、人工智能终端、算法模型框架、人工智能平台的产业链维度，和人工智能数据采集、人工智能数据标注、人工智能数据共享流通、人工智能数据利用的数据生命周期维度。

● **行业标准**：即根据人工智能行业渗透和应用深度，在人工智能技术普及较广、产业发展较为成熟的领域，形成行业级别的人工智能数据安全标准，包括生物特征识别、工业互联网、智能网联汽车、数字内容、智慧城市、金融信用等。

3. 不断提高人工智能企业自身的数据安全能力

企业是人工智能数据安全的主体，人工智能产业生态的各类企业厂商应当在自身传统数据安全治理框架下，针对人工智能数据安全问题的特殊性，形成企业级的人工智能数据安全治理框架，提高数据安全能力。具体而言，应当包括以下五个层次：

● **组织建设**：设立负责数据安全工作的组织机构和专职人员，进行明确的职责分配。数据安全治理组织机构应负责制定组织内部的数据安全制度规范，负责数据安全风险管理、数据安全策略管理、数据安全合规管理、数据安全标准管理等，并对组织内部的数据安全管理活动进行指导和监督。

● **制度规范**：数据安全制度规范包括数据质量管理规范、数据安全管理规范和安全管理流程。其中数据质量管理规范应当包括对训练数据样本的完整性、结构性、代表性进行整体的安全检测；数据安全管理规范应包括数据安全采集、数据安全传输、数据安全存储、数据安全共享、数据安全销毁等规范；数据安全管理流程是指通过规范化的流程确保制度规范的落实。

● **技术能力**：根据企业自身情况，通过内部研发、外部采购、托管服务等方式，部署必要的数据安全产品和服务，通过技术手段辅助数据安全制度规范的实施。

● **人员能力**：通过内部培训、外部招聘等多种方式提升组织内部相关人员的数据安全意识和能力，构建一支覆盖企业管理、人工智能发展、数据安全法律合规、数据安全运维等多个专业能力的安全队伍。

● **企业文化**：人工智能开发企业要将数据质量、数据安全和隐私保护融入企业组织文化建设，要以此为理念开展人工智能产品的设计、开发和测试，人工智能系统应用企业要以此为理念管理好应用过程中产生的数据。

4. 打造全面立体的人工智能数据安全能力供给

人工智能数据安全涉及到技术、法律、产业发展等各个领域，因此政府及其公共部门、产业界和社会机构应当共同形成一个覆盖技术、产品、服务、法律、咨询的整体安全能力供给。具体而言，应当包括以下五个层次：

● **技术研发**：我国政府应当加大对数据安全和隐私保护技术的政策和资金支持，鼓励企业、安全厂商和科研院所加大对保护数据安全、提高数据质量技术的投入，形成一批高技术研发能力的数据安全技术实验室和研究所，重点研究方向可包括多方安全计算、联邦学习、区块链、同态加密、零知识证明、基于隐私的机器学习、基于小数据的人工智能算法、数据偏见检测等。

● **产品服务**：网络安全厂商和数据安全厂商应当针对人工智能多个应用场景，形成包括测试工具、安全防护、人员培训、场景落地、安全运维、安全托管等一整套的安全解决方案，形成高质量、全方位的市场化安全能力菜单。

● **测评认证**：相关测评机构、认证机构、企业应当共同推动人工智能安全测评认证能力的提升，建立人工智能测评服务平台，为企业提供数据质量、数据安全、算法模型的专业测试和评级服务；并围绕人工智能产品服务，形成统一专业的数据安全和隐私保护认证。

● **合规咨询**：打造人工智能数据安全法律合规的供给集合，引导律师事务所、咨询公司为人工智能企业的法律合规提供包括风险评估、协议设计、流程管理、跨境流动、纠纷解决等一整套的法律合规服务。同时我国相关公共部门、高校和智库研究机构应当通力合作，围绕全球人工智能发展、人工智能数据安全形势和人工智能数据安全治理实践，形成一批专业性的研究成果和公开报告，充分向社会和国际宣传我国人工智能数据安全治理的模式经验和最佳实践，为企业提高数据安全能力提供充足的公共知识储备和行动参考。

● **安全生态**：加强人工智能数据安全供给的上下游协同和合作，包括人工智能数据安全供给与人工智能应用企业之间，人工智能数据安全供给的各个主体之间，企业与政府公共部门、社会主体之间等，共同打造一个“安全赋能发展，发展驱动安全”的良性生态闭环。

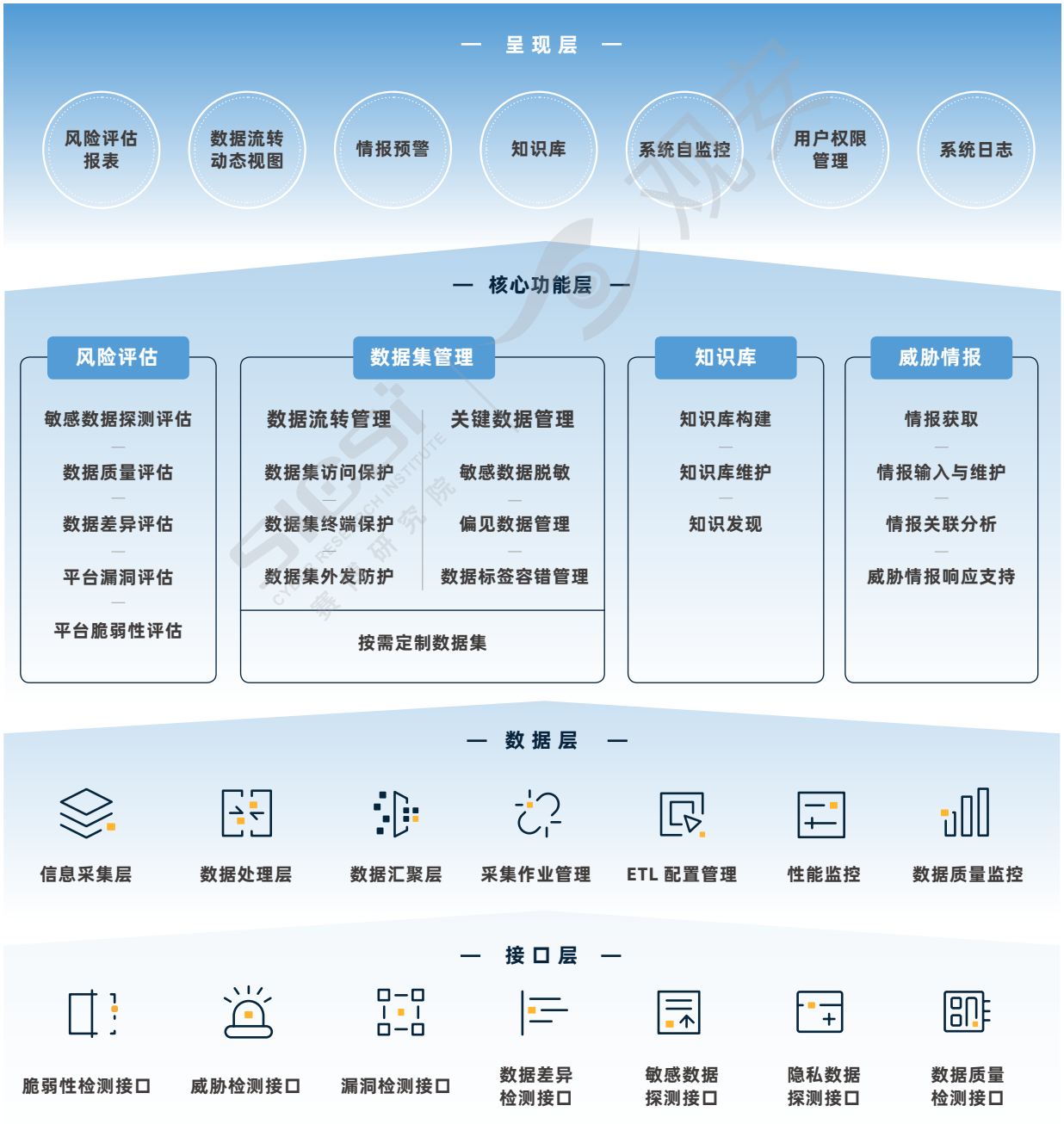
PART 4

人工智能场景的数据安全技术解决方案

一 | 通用场景的人工智能数据安全风险评估平台

人工智能数据安全风险评估平台能够针对特定人工智能应用场景中的数据安全风险进行总体评估和评级，以及数据集管理和知识库建设。平台主要包括风险评估、数据集管理、知识库管理、威胁情报等功能。（见下图 3）

图 3：人工智能数据安全风险评估平台总体架构图



人工智能数据安全风险评估平台通过设定安全基线，开发用于敏感数据探测、数据质量检测、数据差异检测、漏洞检测以及脆弱性检测的工具，基于检测工具汇集的数据实现数据安全风险信息实时收集、自动推送、智能分析、量化评估与诊断分级，针对人工智能应用场景中的数据安全实现多层级、多维度风险评估，为企业对人工智能系统开展自评以及第三方测评机构针对人工智能项目开展风险评估和产品认证提供技术、工具和平台。

同时，人工智能数据安全风险评估平台能够利用数据积累形成的数据资源池，根据不断变化的应用场景需求，构造出用于各类人工智能模型的训练、测试及验证标准数据集、人工智能数据安全知识库及威胁情报库。

1. 功能呈现层

功能呈现层包含了人工智能数据安全风险评估平台的主要用途，包括风险评估报表、数据流转动态视图、情报预警和知识库。

● **风险评估报表：**风险评估报表是通过对人工智能数据暴露面和风险面的全面检测，利用系统格式自动生成相应的人工智能数据安全风险评估报表。

● **数据流转动态视图：**数据流转动态视图是为了保证业务终端查看数据平台各种数据的安全，以及维护终端在系统维护过程中访问数据平台数据的安全。数据流转动态视图能够对接入到数据平台的终端进行准入控制，对导出平台的数据进行跟踪管理，以及对提供在线数据集服务的终端进行管控，对终端查看的内容进行防泄漏处理。

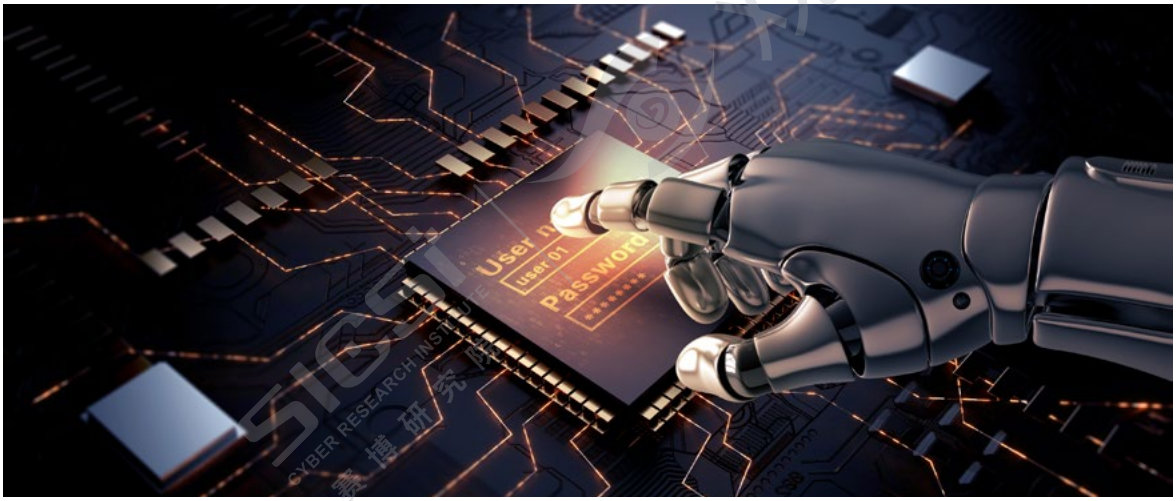
● **情报预警：**威胁情报是一种基于证据的知识，描述了现存的、或者是即将出现针对资产的威胁或危险，包括情境、机制、检测指标、影响和可行的建议，并可以用于通知企业针对相关威胁或危险做出决策。

● **知识库：**人工智能数据安全知识库汇聚了全球市场与技术研发的动态信息，实现人工智能数据安全技术、应用实践和各国监管、合规政策的知识库，为厂商和研究机构服务，促进人工智能和人工智能安全等产业的发展。将量化分析各国人工智能数据安全领域的伦理规范、法律法规、政策、标准、最佳实践等，建设专业的政策数据库；以高质量的知识库建设，综合外部实时动态信息，实现情报检索、专利分析、态势简报等功能。

2. 核心功能层

核心功能层包括风险评估模块、数据集管理模块、知识库管理模块和威胁情报管理模块，定义了人工智能数据安全风险评估平台的分析能力。

- **风险评估模块**：基于各检测工具的检测结果和数据进行综合的数据安全风险评估，其功能包括：敏感数据探测评估功能、数据质量检测评估功能、数据差异检测评估功能、AI 平台漏洞检测管理功能、AI 平台脆弱性检测管理功能。
- **数据集管理模块**：根据业务场景以及相应的算法模型需要，设置条件从数据资源池抽取人工智能深度学习所需的标准数据集，为了保证业务终端查看数据平台各种数据的安全，以及维护终端在系统维护过程中访问数据平台数据的安全，应对接入到数据平台的终端进行准入控制，且对终端查看的内容进行防泄漏处理，还需对人工智能平台数据集进行数据流转、脱敏管理。
- **知识库管理模块**：将量化分析各国人工智能数据安全领域的伦理规范 法律法规 政策标准、最佳实践等，建设专业的政策数据库，实现包括量化和评估风险等级、风险类别以及风险处置等功能，并根据应用场景需求动态更新与迭代。
- **安全威胁情报管理**：用于支撑外部威胁的分析和定位，功能包括威胁情报获取、威胁情报输入与维护、外部资产发现和监控、情报关联分析。



3. 数据层

数据层实现各类检测工具的数据的采集、处理、汇聚、存储、检索能力。该层以接口形式向人工智能数据安全风险评估平台的分析评估提供输入数据。数据采集和补全的流程是一个完整高效的事件处理过程，具体流程包括：

- **Input**：用来处理数据输入的接口层，并形成初始的数据流。
- **Decode**：对数据进行格式匹配或者解码。
- **Filter**：对数据进行过滤或；转换；补全或者关联。
- **Encode**：对输出数据进行整形和转换，decode 和 encode 对某种格式都是成对存在的。
- **Output**：数据输出接口。

4. 接口层

接口层实现对各检测工具的任务指令下发和检测结果、数据的上传。具体包括：

● Flume 提供对数据进行简单处理，并写到各种数据接受方（可定制）的能力； Flume 提供了从 console（控制台）、RPC（Thrift-RPC）、text（文件）、tail（UNIX tail）、syslog（syslog 日志系统，支持 TCP 和 UDP 等 2 种模式），exec（命令执行）等数据源上收集数据的能力。

● 支持 Syslog 协议，接收通过 Syslog 发出的日志，并具备按级别、Facility 过滤功能。

● 支持通过 FTP、SFTP 采集日志文件，支持文本文件和 csv 日志格式文件。

● 支持 SNMP 协议，可接收通过 SNMP Trap 发出的设备日志，统一采集平台可通过 SNMP 查询设备性能，如 CPU、内存等。

● 支持 ODBC/JDBC 协议，可通过 JDBC/JDBC 采集设备日志信息。

● 支持 Oracle、DB2、Informix、Sybase、Sqlserver、Mysql 主流数据库和 SQL 语句取数据。

二 | 人工智能场景：智能网联汽车的数据安全解决方案

智能网联汽车中的数据来源于用户、ECU、传感器、IVI 及操作系统、第三方应用及车联网服务平台等，种类包括用户身份信息、汽车运行状态、用户驾驶习惯、地理位置信息、用户关注内容等敏感信息，在车辆保险、用户行为分析等方面具备很大价值，将是未来车联网安全重点。

1. 数据安全风险分析

● **传输和存储环节存在数据被窃风险。**目前，车联网相关数据主要存储在智能网联汽车和车联网服务平台上，存储和传输方案主要由整车厂商、车联网服务商设计实现，由于数据的采集传输、存储等环节没有统一的安全要求，可能因访问控制不严、数据存储不当等原因导致数据被窃。如汽车端数据可能被 OBD 外接设备非法读取、IVI 系统数据可能被第三方应用越界读取、网络传输数据可能被攻击者嗅探或遭受中间人攻击、车联网服务平台端数据可能被非法和越权访问。数据被窃通常与业务设计、技术实现有关，将是车联网安全防护的重要内容。

● **数据过度采集和越界使用成为隐私保护主要问题。**车联网信息服务所采集的如车主身份信息（如姓名、身份证、电话）、车辆静态信息（如车牌号、车辆识别码）、车辆动态信息（如位置信息、行驶轨迹），以及用户的驾驶习惯等，都属于用户个人隐私。个人信息的采集一般需遵循“知情同意”、

“最小必要”、“目的限定”三大原则，但由于车联网属于新兴行业，管理还在完善中，对于哪些数据可被采集、数据如何利用、是否可以分享给第三方等关键问题，各国目前还缺乏细化管理要求，因此目前数据采集和使用还存在侵犯用户隐私的风险。

● **数据跨境流动问题成为威胁国家安全潜在隐患。**车联网数据包含道路地理等信息涉及国家安全，应加强管理，目前车联网数据汇总于车联网服务平台，存在云平台数据跨境流动管理问题，主要体现在两个方面：一是存在境外车联网服务商跨界服务隐患。境外进口汽车的网络服务及后台服务可能由境外通信企业和整车企业提供，通信数据及车联网数据传往境外，可能泄露国家地理位置信息，危害国家安全。二是存在境内外云平台数据共享隐患。由于目前整车厂大多为合资企业，车联网服务以境内云平台为主，但其外资公司通常负责全球车联网运营，境内平台与境外平台是否互联，是否存在数据传输共享，是国家数据管理需要关注的重点内容。

2. 数据安全解决方案

● **加强车载端访问控制、实施分域管理，降低数据安全风险。**建立安全分级访问机制，智能网联汽车通常配备有两个 APN 接入网络。APN1 负责车辆控制域 (CleanZone) 通信，主要传输汽车控制指令及智能汽车相关敏感数据，通信对端通常是整车厂商私有云平台，安全级别较高。APN2 负责信息服务域(Dirty Zone)通信，主要访问公共互联网信息娱乐资源，通信对端可能是整车厂公共云平台或者第三方应用服务器，IVI 系统中的车载应用，如新闻、娱乐、广播等通常通过 APN2 进行通信。车辆控制域和信息服务域采用隔离的方式来加强安全管理。一是网络隔离，APN1 和 APN2 之间网络完全隔离，形成两个不同安全等级的安全域，避免越权访问。二是车内系统隔离，车内网的控制单元和非控制单元进行安全隔离，对控制单元实现更强访问控制策略。三是数据隔离，不同安全级别数据的存储设备相互隔离，并防止系统同时访问多个网络，避免数据交叉传播。四是加强网络访问控制，车辆控制域仅可访问可信白名单中的 IP 地址，避免受到攻击者干扰，部分车型对于信息服务域的访问地址也进行了限定，加强网络管控。

● **基于 PKI 和数据通信加密，构建可信“车 - 云”数据通信。**目前企业普遍重视通信加密，部分厂商在软加密基础上建设 PKI 系统，搭建更便捷的“车 - 云”通信，采取的防护措施具体包括：一是基于证书的车载端身份认证，传统的‘车 - 云”通信通过车机编码绑定的方式进行认证，易被伪造绕过。目前较完备的方式是基于 PKI 证书身份认证，智能网联汽车首次启动进行通信连接时，云平台签发可信证书写入车载安全芯片，用于“车 - 云”通信，确保仅有认证后的车辆可与私有云通信，同时基于 PKI 技术使得云平台具备证书撤销更新的功能。二是基于证书的传输加密，智能网联汽车在获取可信证书后，后续通信通过

证书进行密钥协商并加密通信数据，加密协议通常采用 HTTPS 应用层加密或者 SSL、TLS 传输层加密，增加攻击者窃听破解的难度，保障数据的通信安全。

● **网络侧进行异常流量监测，提升车联网网络安全防护能力。**通过采用异常流量监测对车联网业务进行流程监测，提供安全监测预警及应急处置服务，具体分为监测预警、网络控制两个方面：监测预警功能包括：定制监控服务，对安全事件进行探测，提供流量监控优化、异常流量告警、历史数据留存等；网络控制包括：定义受保护的 IP 地址 / 范围、阻止点对点通信、借助防火墙和入侵检测系统中断异常 IP 通信。

● **企业制定内部数据分级管理要求，加强敏感信息管理。**车联网整车厂商对用户数据进行分级保护，对于涉及驾驶员信息、驾驶习惯、车辆信息、位置信息等敏感数据采取较高级别的管理要求，仅被整车厂商签名认可的应用才可读取相关数据，其他非签名认证应用仅可读取非敏感数据。针对敏感数据实行单独的存储要求，通过加密提升数据安全级别。

● **加强数据传输、利用环节管理，避免数据外泄。**敏感数据传输通过 APN1 在车辆控制域中加密传输，避免外泄。加强数据使用限制，部分车企将车联网数据仅作为内部数据使用，用于车辆故障诊断，拒绝与任何第三方企业共享用户数据，尽可能确保户私密数据安全可控。

三 | 人工智能场景：人脸识别的数据安全解决方案

1. 数据安全风险分析

● **网络安全防护能力不足，存在数据泄露隐患。**首先，在个人影像数据采集环节，智能监控设备目前存在严重安全问题，极易导致数据泄露。其次，在数据的网络传输环节，个人影像数据即使配合指纹进行多重生物特征识别，也必须转换为二进制代码进行网络传输,面临失窃风险。同时，在数据分析环节，人脸识别的人工智能系统模块仍旧运行在传统信息系统上，任何一个环节有漏洞，都可能被黑客攻破。最后，在数据存储环节，个人影像数据库的防护能力不足，数据泄露风险严峻。

● **企业内部数据安全制度不健全。**首先，各类社交平台、电子商务、自助服务、拍摄软件等商业领域广泛使用人像采集功能，智能摄像头随时随地采集不特定人群的个人影像数据，企业不断大规模收集、积累的个人影像数据，出于盈利目的滥用个人数据的现象比比皆是。其次，在识别核心算法上拥有自主知识产权的企业极少，市场上人脸识别产品质量良莠不齐、安全防护技术不统一，系统安全漏洞大量存在，个人影像数据随时面临泄露危险。最后，人脸识别在商业领域展开应用，将使得个人影像数据成为身份认证的关键生物信息。伴随着互联网企业、金融机构等私

营企业不断获取海量的个人影像数据，将对个人隐私、信息安全保护造成极大考验。

总体而言，人脸识别场景的数据安全威胁根源在于：一是缺乏能充分适应各种干扰环境的人脸预处理算法。特别在移动互联网下，个人影像数据的拍摄环境更加复杂多变，对于人脸识别算法精度的考验更大。二是受深度学习框架中的软件漏洞、生成恶意样本、训练数据遭受污染等影响，用于比对的模板数据库产生遗漏或者误差。三是人工智能系统在辅助决策过程中存在计算错误，产生错误信息。

2. 数据安全解决方案

● **加强准入认证与控制**：对视频平台及数据存储区域构建双因子认证体系，实现合法用户通过合法进程的准入控制，有效防止非法人员在获悉监控系统账号及密码的情况下访问监控系统，非法调阅、篡改、删除视频数据。

● **确保数据链路安全**：视频或人脸识别数据从前端接入后进行加密到目的节点处进行解密，整个过程通过安全的加密算法进行加密传输，保证在链路传输过程中不被非法窃听、劫持。

● **重视数据共享安全**：需要提供给外部单位共享使用的视频和人脸识别数据，通过技术手段制作外发数据，设定操作权限（编辑、拷贝、打印）和使用期限次数，超期自动删除，防止外单位人员因保管不当造成泄密，确保数据安全。

● **实现可追踪溯源**：对于监控视频或人脸识别图像，运用透明水印技术，显示用户自定义内容，显示实时时间；或采用数据标签技术，做到任何一个视频，任何一张图片，可全程追溯，并利于事后追责。

总之，对于视频监控或人脸识别系统来说，必须在前端就保护好监控或采集终端免受攻击，不被利用、劫持或控制；其次，监控或采集终端即使被攻陷，成为攻击的跳板，也要保证能有效识别和有效阻断，对系统发动的网络攻击能够被阻止；最后，必须保证在使用过程中，视频监控或人脸识别数据的访问严格受限，操作动作严格受限，对于拍摄、访问等行为能够有效追溯，对外部门的合法调阅或共享使用，也能做到追溯防护。

四 | 人工智能场景：工业互联网的数据安全解决方案

1. 数据安全风险分析

● **边缘层数据安全**：边缘层安全是指工业互联网平台与工业企业接入过程中数据采集、协议转换、边缘计算的安全。由于智能传感器、边缘网关等边缘终端设备计算资源有限，安全防护能力薄弱，工业互联网平台在数据采集、转换、传输的过程中，数据被侦听、拦截、篡改、丢失的安全风险更高，攻击者可利用边缘终端设备漏洞对平台实施入侵或发起大规模网络攻击。

● **IaaS 层数据安全**：工业 IaaS 安全是指工业互联网平台云基础设施的安全，包括主机设备物理安全、虚拟化系统安全、虚拟化网络安全、虚拟化管理安全、工业数据存储安全等。工业 IaaS 是虚拟化、资源池化的信息基础设施，面临着虚拟机逃逸、跨虚拟机侧信道攻击、镜像篡改等新型攻击方式的威胁。另外，多数平台企业使用第三方云基础设施服务商提供的 IaaS 服务，存在数据安全责任边界不清晰等安全问题。

● **PaaS 层数据安全**：工业 PaaS 是基于工业知识显性化、模型化、标准化的赋能使能开发环境，其安全包括通用 PaaS 平台、工业应用开发工具、工业微服务组件、工业大数据分析平台的安全。通用 PaaS 平台感染病毒、木马，可造成平台瘫痪、服务中断、数据丢失等严重后果。工业应用开发工具、微服务组件存在漏洞，将影响工业 APP 的正常开发和使用。工业大数据分析平台汇聚海量工业企业的工艺参数、产能数据等高价值数据，被黑客入侵可能导致敏感信息泄露，威胁平台数据安全。

● **SaaS 数据安全**：工业 SaaS 安全是指工业互联网平台应用层的应用服务安全。其中，工业 APP 涉及专业工业知识、特定工业场景，集成封装多个低耦合的工业微服务组件，功能复杂、安全设计规范缺乏，可能存在安全漏洞和缺陷，面临的工业 APP 漏洞、API 通信安全、用户管控、开发者恶意代码植入等应用安全问题更为突出。

● **平台层数据安全**：平台数据安全涉及接入平台、平台运行、平台退出三个阶段中的数据安全。其中，在接入平台阶段，包括上述边缘层接入以及工业 APP 接入到平台过程中数据面临的侦听、拦截、篡改、丢失、窃取等安全风险；在平台运行阶段，主要面临数据存储安全风险；在平台退出阶段，涉及用户迁移平台或完全退出平台时数据泄露与备份的安全风险。

2. 数据安全解决方案

针对工业互联网数据收集、存储、处理、转移、删除等各环节的数据安全问题，通过实现数据分级分类管理、统一认证、隐私保护等能力等方式，构建工业互联网整体的数据安全解决方案，共由五个子系统构成：数据安全采集传输子系统、数据安全交换子系统、数据泄露防护子系统、数据安全审计子系统、数据脱敏子系统。

● **数据泄露防护子系统**：数据泄露防护子系统实现识别工业互联网机密数据的内容，正确地认识客户的业务流程，梳理出合理的业务流程保证敏感数据正确的流动为目标。针对数据泄露的各个风险面和暴露面，提供统一解决方案，促进核心业务持续安全运行。

● **数据安全审计子系统**：随着工业互联网发展进程不断深入,企业的业务系统变得日益复杂，由内部员工违规操作导致的安全问题变得日益突出起来。防火墙、防病毒、入侵检测系统等常规的安全产品可以解决一部分安全问题，但对于内部人员的违规操作却无能为力。工业互联网在发展的过程中，因为战略定位和人力等诸多原因，越来越多的会将非核心业务外包给设备商或者其他专业代维公司。如何有效地监视设备厂商和代维人员的操作行为，并进行严格的审计是企业面临的一个关键问题。数据安全审计子系统是针对企业内网的运维操作和业务访问行为进行细粒度控制和审计的合规性管理系统。它通过对运维人员和业务用户的身份进行认证，对各类运维操作和业务访问行为进行分析、记录、汇报，以帮助用户事前认证授权、事中实时监控、事后精确溯源，加强内外部网络行为监管、促进核心资产（数据库、服务器、网络设备等）的正常运行。

● **数据脱敏子系统**：数据脱敏子系统是数据安全防护全生命周期中的重要环节。工业互联网中的数据脱敏，对工业互联网中某些生产、运营、销售等敏感信息通过脱敏规则进行数据的变形，实现敏感隐私数据的可靠保护。这样，就可以在开发、测试和其它非生产环境以及外包环境中安全地使用脱敏后的真实数据集。数据库脱敏子系统可实现自动识别敏感数据和管理敏感数据，提供

灵活的策略和脱敏方案配置，高效可并行的脱敏能力，同时保证数据的有效性和可行性，使脱敏后的数据能够安全的应用于测试、开发、分析和工业互联网的第三方使用环境中。

● **数据安全交换子系统**：工业互联网的典型网络现状，是将内部工业生产网络与外部公共信息通信网络分开，形成所谓的“内网”和“外网”，有些部门由于业务种类众多，数据敏感度不一的原因，在内网中还人为隔离几个不同的网络，用于处理不同的业务或者存放不同敏感程度的数据。因此，数据安全交换子系统是数据安全支撑体系中的重要一环。要真正能够实现信息共享，一定要实现跨网络、跨安全域的数据交换。但是由于黑客攻击的大肆蔓延，数据泄密不时发生，用户对于数据交换过程中的安全非常重视，网络越复杂，交换数据越敏感，其对安全的重视也就越高。安全数据交换子系统提供数据库（支持国内外所有主流数据库）、文件（支持国内外各种主流文件系统）、流数据安全交换，并提供文件格式检查、病毒查杀、数据加密、交换审计等安全服务。

● **数据安全分析与预警平台建设**：工业互联网数据安全分析与预警平台是对工业互联网数据安全威胁、安全状态整体呈现并对数据安全防护子系统中所有系统数据以及业务数据进行安全关联分析，对正在发生的数据安全事件进行告警，对未来可能出现的数据安全事件提供预警的平台。工业互联网数据安全分析与预警平台采用机器学习等技术，针对全量数据进行安全分析，既可利用已知知识，同时可对行为识别，利用机器学习等技术发现未知异常情况，通过OODA（观察 - 定位 - 决策 - 行动）、情报、目标等线索启动数据安全分析。

PART 5
参考文献

1、2016 年 G20 杭州峰会《G20 数字经济发展与合作倡议》. 01

2、赛博研究院 . 非接触新经济安全治理报告 [EB/OL]. (2020-05-20) https://mp.weixin.qq.com/s/o_lYQKm8N7l87pnW4_W3lw. 01

3、中国电子学会 . 新一代人工智能产业白皮书 (2019 年) [EB/OL]. (2019-05-20) .https://mp.weixin.qq.com/s/UMhAEbFlhaKJ1vySI93wjw. 02

4、德勤 . 全球人工智能发展白皮书 [EB/OL]. (2019-09-22) .https://www2.deloitte.com/content/dam/Deloitte/cn/Documents/technology-media-telecommunications/deloitte-cn-tmt-ai-report-zh-190919.pdf. 02

5、乌镇智库 . 全球人工智能发展报告 (2018) [EB/OL]. (2019-04-25) .http://www.199it.com/archives/869189.html. 03

6、中国信息通信研究院 . 全球人工智能产业数据报告 (2019Q1)[EB/OL]. (2019-05-23) http://m.caict.ac.cn/sytj/201905/P020190523542892859794.pdf. 03

7、CB Insights. 全球人工智能投资趋势年度报告 [EB/OL]. (2020-03-12) .http://www.199it.com/archives/1019471.html. 03

8、IDC.AI 驱动金融行业智能决策 (2020) [EB/OL]. (2020-04-21) .https://www.idc.com/getdoc.jsp?containerId=prCHC46233820. 03

9、中国新一代人工智能发展战略研究院 . 中国新一代人工智能科技产业发展报告 (2020)[EB/OL]. (2020-06-24) .http://www.xinhuanet.com/info/2020-06/30/c_139176781.htm. 03

10、If your image is online, it might be training facial-recognition AI.https://www.cnn.com/2019/04/19/tech/ai-facial-recognition/index.html.(最后访问时间: 2020-06-30). 05

11、IBM didn't inform people when it used their Flickr photos for facial recognition training. https://www.theverge.com/2019/3/12/18262646/ibm-didnt-inform-people-when-it-used-their-flickr-photos-for-facial-recognition-training.(最后访问时间: 2020-06-30). 05

12、新华网 . 中国人脸识别第一案 杭州一动物园被起诉 . 发布时间 2019-11-04.http://www.xinhuanet.com/legal/2019-11/04/c_1125188289.htm.(最后访问时间: 2020-06-30). 05

13、Nicholas Carlini& David Wagner.Audio Adversarial Examples: Targeted Attacks on Speech-to-Text. University of California Berkeley.https://arxiv.org/pdf/1801.01944.pdf.(最后访问时间: 2020-06-30). 06

14、Photo algorithms id white men fine—black women, not so much.https://www.wired.com/story/photo-algorithms-id-white-men-fineblack-women-not-so-much/.(最后访问时间: 2020-06-30). 07

15、Privacy-Preserving Machine Learning 2018: A Year in Review.https://medium.com/dropoutlabs/privacy-preserving-machine-learning-2018-a-year-in-review-b63.45a95ae0f.(最后访问时间: 2020-06-30). 09

16、Briland Hitaj,Giuseppe Ateniese, Fernando Perez-Cruz.Deep Models Under the GAN: Information Leakage from Collaborative Deep Learning.https://arxiv.org/abs/1702.07464.(最后访问时间: 2020-06-30). 09

PART 5 参考文献

17.李凤华,李晖,贾焰,俞能海,翁健.隐私计算研究范畴及其发展趋势[J].通信学报,2016,37(04):01-11. 22

18、孟书海.基于同态加密的机器学习研究综述[J].电脑知识与技术.2019. 22

19、Deep learning with differential privacy[C]//2016 ACM Sigsac Conference on Computer and Communications Security. 23

20、Evaluation of Privacy-Preserving Technologies for Machine Learning https://medium.com/outlier-ventures-io/evaluation-of-privacy-preserving-technologies-for-machine-learning-8d2e3c87828c. 23

21、This is how we protect your privacy. https://www.apple.com/lae/privacy/approach-to-privacy/.(最后访问时间: 2020-06-30). 23

22、Uncovering and Mitigating Algorithmic Bias through Learned Latent Structure http://www.aies-conference.com/wp-content/papers/main/AIES-19_paper_220.pdf. 24

23、AI Fairness 360 (AIF360 v0.2.0) https://github.com/ibm/aif360.(最后访问时间: 2020-06-30). 25

24、AI Fairness 360 Open Source Toolkit http://aif360.mybluemix.net/.(最后访问时间: 2020-06-30). 25

25、华为.AI安全白皮书[EB/OL].(2019-09-19).https://www-file.huawei.com/-/media/corporate/pdf/cyber-security/ai-security-white-paper-cn.pdf. 25

WAIC
世界人工智能大会
WORLD ARTIFICIAL INTELLIGENCE CONFERENCE
2020
云端峰会
SUMMIT ONLINE

SICSI
CYBER RESEARCH INSTITUTE
赛博研究院

观安

web-**www.sicsi.org.cn**
phone-**021-61432693**
e-mail-**public@sicsi.org.cn**