SICSI
CYBER RESEARCH INSTITUTE
赛博研究院

AIGC

数据安全与算法治理报告

AIGC Data Security and
Algorithmic Governance Report

版权声明

COPYRIGHT STATEMENT

出品方

上海赛博网络安全产业创新研究院

本报告部分内容引用自上海赛博网络安全产业创新研究院和上海观安信息技术股份有限公司联合出品的《人工智能数据安全治理报告》，特此感谢。

前 言
PREFACE

当前，随着以“算力新基建、数据新要素、AI 大模型”为特征的算力经济时代来临，以 ChatGPT 为代表的 AIGC 应用掀起人工智能新一轮浪潮，全球人工智能发展逐步从“探索期”向“成长期”过渡，在技术和产业上进入重要转型阶段。在此背景下，人工智能发展和数据、算法安全问题深度交织融合，影响用户隐私、公民权益、商业秘密、国家安全等各个方面，AIGC 的创新应用和数据安全治理工作也进入全新的阶段。

本报告重点聚焦并梳理了人工智能应用中较为独特或更突出的安全问题，全面分析了当前全球人工智能数据安全和算法治理的主要现状和最新动态，基于全球相关治理实践和我国实际情况，构建了 AIGC 数据安全和算法治理框架。

CONTENTS

PART 1

人工智能发展与安全挑战

一、算力经济时代人工智能发展趋势

新一轮算力经济发展浪潮来临

近年来，随着大数据、云计算、人工智能等为代表的数字技术带来全球性的科技革命和产业变革，以**“算力新基建、数据新要素、AI 大模型”为核心特征的算力经济发展浪潮为人工智能全面发展注入了强大动能。**

算力新基建成为人工智能新发展的坚实底座和基础支撑。数字化时代的数据呈指数级爆发，随着人工智能等智能化应用的持续发展，用于人工智能训练和推理计算的智能算力需求也呈现出快速增长趋势。近年来，美、欧、日、英等全球主要经济体大力发展以云计算、边缘计算、智算数据中心等为代表的新型算力基础设施。2022 年，中国“全面启动东数西算”，拟在京津冀、长三角、粤港澳大湾区、成渝、内蒙古、贵州、甘肃、宁夏等 8 地启动建设国家算力枢纽节点。

2023 年，科技部发布《关于支持建设国家新一代人工智能公共算力开放创新平台的函》，批复 9 个平台建设国家新一代人工智能公共算力开放创新平台、16 个平台建设国家新一代人工智能公共算力开放创新平台，加快推进算力基础设施建设。算力作为数字经济时代的关键生产力，其基础设施建设保障了人工智能发展所需要的巨大基础计算和存储需求，为人工智能市场化应用提供高质量的硬件支撑。

数据新要素成为人工智能新发展的核心动能和强大驱动。2022 年 12 月，中国发布《中共中央 国务院关于构建数据基础制度更好发挥数据要素作用的意见》，从数据产权、流通交易、收益分配、安全治理等方面构建数据基础制度，提出 20 条政策举措。随着全球各国不断加快数据市场的建设，将进一步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，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，推动更大规模的数据有序、便捷、高效和安全流动交易，为人工智能全面发展注入高质量的数据动能。

AI 大模型成为人工智能新发展的创新引领和产业机遇。“大算力 + 大数据 + 强算法”构成 AI 大模型

的核心要素，其中包含“预训练”和“大模型”两层内涵，大模型在大规模数据集上完成数据预训练后，能够直接支撑各类应用。早期的大模型源于 NLP（自然语言处理）领域，随着人工智能多模态生成能力的演进，数据规模和参数规模的有机提升，目前多模态通用大模型逐渐成为发展主流。进入算力经济时代，基于 AI 大模型的各类应用有望带来生产力的颠覆性革新。

2. 以 AIGC 为代表的人工智能应用进入“成长期”

全球人工智能发展正在逐步度过“探索期”进入“成长期”，其中，AIGC 作为人工智能行业在技术上的新突破、新发展，主要呈现两个关键特征。

● **AIGC 领域技术迅速突破。**人工智能相关技术发展经过三个阶段，“图灵测试”为代表性事件的早期萌芽阶段（20 世纪 50 年代至 90 年代中期），从实验性向实用性转变的沉淀积累阶段（20 世纪 90 年代中期至 21 世纪 10 年代中期），依托 NLP/ 多模态预训练大模型、深度生成模型、生成式对抗网络（GAN）、语言模型（Transformer）等 AI 技术的快速发展阶段（21 世纪 10 年代中期至今）。

目前，重点领域的技术突破催生了 AIGC 人工智能应用，在大模型、算力及 AI 技术的支撑下，AIGC 应用呈现出“专用人工智能”“通用人工智能”两个发展方向，前者具有任务单一、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，已进入快速的商业化应用阶段；后者基于理解、计划和问题解决能力，在不同的应用场景或任务领域均表现出色，已涌现了 ChatGPT、AutoGPT 等一系列新型 AIGC 应用，并进一步向“成长期”迈进，有望成为未来互联网的生产基础设施。

● **AIGC 产业生态初步成型。**人工智能产业链已形成包括智能芯片、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，整体产业结构进一步优化。

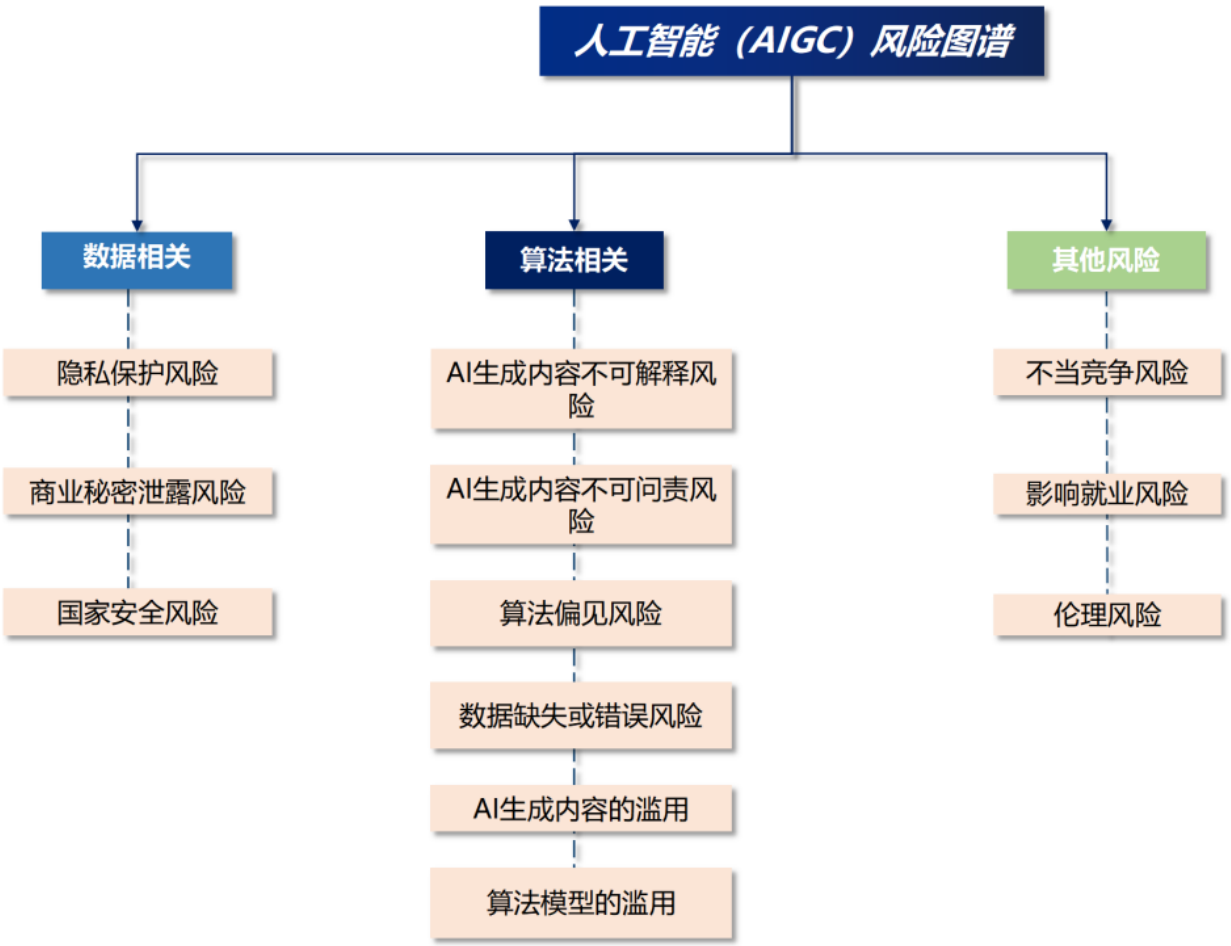
同时，AIGC 产业作为万众瞩目的蓝海，市场规模持续增长。根据对 610 个国内外应用的统计，AIGC 约包含 48 个分类。根据应用场景，可以分为文本对话、文案写作、图像生成、视频创作、音乐创作、特定实验（蛋白质序列）模拟类以及代码编写协助等特定人工智能类型。其中，全球范围内的图像生成、文案写作和代码编写三类 AIGC 产品年营收均已超过 1 亿美元，Stability、Jasper.ai 等 AIGC 独角兽估值大幅上升。

3. 数据安全和算法问题成为人工智能发展的掣肘

随着人工智能在产业和技术两个方面加快度过“探索期”，逐步进入“成长期”，人工智能（AIGC）发展与安全更加深度地交织在一起，数据和算法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。**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数量的加快增长，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和算法风险。**人工智能通过训练数据集构造和优化的算法模型，因其对于数据资源特有的处理方式，将带来更多的隐私保护忧患、智能代替人工造成的就业担忧、算法对于市场竞争带来的不公平以及科技伦理等一系列问题。同时，人工智能在自动化网络攻击、数据黑产的应用，使得网络安全和数据安全威胁更加复杂，对国家和企业现有的安全治理能力形成巨大冲击。



二、人工智能发展的安全挑战

人工智能（AIGC）发展过程中面临的安全挑战既有传统数据安全问题，也具有人工智能特性带来的新型风险，其影响覆盖用户隐私、公民权益、商业秘密、国家安全等方方面面。

1. 数据采集阶段：难以保障数据所有者权利

AIGC 等新一代人工智能应用在模型训练阶段所需的数据呈现指数级增长趋势，因此通过数据采集获取海量数据是人工智能发展的重要前提。从数据采集方式来看，目前主要包括现场无差别采集、直接在线采集和商务采购等方式。**现场无差别采集时，**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，当采集的对象为用户时，也难以获得用户的充分授权同意。**直接在线采集时，**由于是通过技术手段对于网络公开数据进行扫描爬取，而人工智能系统通常由训练好的模型部署而成，需要对数据进行连续性的处理分析，因此很难保障数据所有者的修改、撤回等权益，并且可能涉及知识产权问题。**数据交易时，**由于目前数据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数据，数据收集存在违法问题。最后，随着 AI 生成技术的发展，有效网络数据的增长将跟不上训练模型所需数据量的增速，与此同时数据获取的成本也不断上涨，因此合成类数据（Synthetic Data）将是未来人工智能企业主要的数据源之一。

案例一

2023 年 1 月 23 日，在加州北区法院，美国三名漫画艺术家针对包括 Stability AI 在内的三家 AIGC 商业应用公司发起集体诉讼，指控 Stability AI 研发的 Stable Diffusion 模型以及三名被告各自推出的、基于 Stable Diffusion 模型开发的付费 AI 图像生成工具构成版权侵权。

案例二

2023 年 2 月 15 日，《华尔街日报》消息，Open AI 公司未经授权大量使用路透社、纽约时报、卫报、BBC 等国外主流媒体的文章，用以训练 ChatGPT 模型，构成侵权。

2. 数据处理阶段：可能导致决策错误或算法歧视

数据污染可能导致人工智能算法模型失效。数据污染产生的原因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。

数据投毒可能导致人工智能决策错误。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，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，将通过破坏原有训练数据的概率分布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。数据投毒对人工智能核心模块产生的定向干扰还可能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），产生灾难性事故后果。

数据偏差可能导致人工智能决策歧视。人工智能算法决策中所使用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存在难以消除的偏差时，将造成最终决策结果的歧视性。比如在对话生成场景中，ChatGPT 等应用可能因为训练数据的不足，生成带有政治偏见的信息，在金融征信、医疗教育和在线招聘领域，可能因为边远地区、弱势群体和少数族裔的数据量不足、数据质量不高等因素，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。

案例三

2022 年 5 月，美国平等就业机会委员会（EEOC）起诉 iTutorGroup 旗下公司（iTutorGroup, Inc.、Tutor Group Limited 等），指控 iTutorGroup 的在线招聘软件的算法会自动拒绝年龄较大的应聘者，55 岁以上的女性和 60 岁以上的男性将被取消考虑资格。这违反了美国《年龄歧视法》（ADEA），即保护 40 岁及以上的人免受歧视。

3. 数据流通阶段：流通交互的合法性及安全性

海量数据在存储及交互中泄露和滥用隐患。部分人工智能企业采取委托第三方或众包的方式进行海量数据的采集、标注、分析和算法优化，数据在供应链的各个主体之间形成复杂、实时的交互流通链路，考虑到各主体数据安全能力的参差不齐，可能产生数据泄露或滥用的风险。此外，人工智能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，但由于缺乏严格的测试管理和安全认证，因此将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。

案例四

2020 年 2 月 27 日，面部识别应用服务公司 Clearview AI 向美国福克斯新闻网证实，公司所有的客户列表、账户数量以及客户进行的相关搜索数据遭遇了未经授权的入侵。Clearview AI 的面部识别应用客户包括了美国移民局、司法部、银行，FBI，ICE，梅西百货，沃尔玛、NBA、阿拉伯联合酋长国的主权财富基金等 2228 多家机构和公司。

数据孤岛和流通壁垒导致人工智能数据供给不足。底层数据资源的竞争是人工智能企业最关键的市场竞争力体现。以 GPT-3 为例，该模型有 1750 亿个参数，预训练数据量高达 45TB。然而，数据需求和供给之间的不对称、成熟的数据要素市场尚未形成，这些因素都严重影响了行业的发展。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、数据安全等问题也存在着诸多法律和技术上的数据流通壁垒，间接形成了数据黑产滋生的经济动因。

AIGC 应用导致的数据出境流动合法性问题。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领域的数据资源供给、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，无论是出于企业自身需要还是第三方合作，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。在各国日益趋严的数据出境安全评估要求下，数据出境流动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：可能影响国家、企业及个人安全

算法推荐、模型滥用将威胁公民隐私和国家安全，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，这种为开展算法推荐而对用户习惯行为长期跟踪和深度分析的行为，将使公民隐私面临安全风险，甚至可被用于政党竞选和政治宣传等目的，对各国现行的政治制度产生极大的冲击。此外，基于大模型和算法的自动代码编写、内容自动生成使得网络攻击、舆论引导等恶意行为的成本下降。由于 AIGC 通过大型语料库的无监督训练而完成语言生成，模型恶意使用可能导致攻击者通过隐蔽的算法也可能完成意识形态输出与渗透，部分国家可能凭借其科技先发优势，主导数字领域话语权，

案例五

2023 年 5 月，甘肃平凉市公安局崆峒分局网安大队在日常网络巡查中发现，某百度账号出现一篇标题为“今晨甘肃一火车撞上修路工人致 9 人死亡”的文章，并判断为虚假信息。犯罪嫌疑人洪某通过 ChatGPT 将搜集到的新闻要素修改编辑后，使用“海豹科技”软件上传至其购买的百家号上非法获利，并且造成恶劣的社会影响。

模型攻击、不当使用将威胁商业秘密，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，攻击者可以利用神经网络等人工智能算法对训练数据集的记忆，通过公共访问接口对算法模型进行黑盒访问，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，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。此外，办公环境中对于 AIGC 的不当使用，也可能造成企业商业秘密泄露——输入应用的数据被人工智能企业用于模型再训练。

案例六

2023 年 4 月，三星曝出三起员工因使用 ChatGPT 而泄露敏感信息的事件。由于三星允许其半导体部门的工程师使用 ChatGPT 进行源代码修复，在修复过程中，员工向 ChatGPT 输入了包括新程序的源代码本体、与硬件相关的内部会议记录等在内的机密数据，最终导致相关数据被 ChatGPT 获取，用于进一步的模型训练。

PART 2

全球人工智能数据和算法安全治理

在人工智能领域数据安全、算法歧视、科技伦理等风险隐患涌现的背景下，人工智能数据和算法安全成为一个覆盖多主体、多维度的全球性安全挑战和治理议题。

一、各国在 AIGC 领域的治理现状

世界主要国家均高度重视人工智能的安全问题，并将其提升至国家战略层面。聚焦 AIGC 领域，各国近一年加快采取监管举措，规范 AIGC 应用过程中的数据处理活动，并且将数据和算法治理放在首要位置。

1. 美国：鼓励 AIGC 发展，以场景化立法规制安全

美国积极主动地通过立法来谋求其全球科技主导地位，对于 AIGC 持“全面鼓励、拥抱发展”的态度。虽然目前暂无专门规范 AIGC 的法律文件，仅 OpenAI CEO Sam Altman 在 2023 年 5 月举行的美国参议院听证会上，敦促立法者对制造先进 AI 的组织实施许可要求和其他法律要求，但人工智能的整体治理仍在推行中，因此 AIGC 的治理工作可以参考现行法律法规。

在人工智能及算法治理方面，美国长期以来关注人工智能及算法对于公民权利、自由及数据和隐私安全的影响，并且通过发布一系列政策文件逐步完善治理。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。2022 年 2 月，美国国会议员提出《2022 年算法责任法案》，要求对软件、算法和其他自动化系统实施新的透明度和监管举措。2022 年 10 月，白宫发布《人工智能应用监管指南备忘录（草案）》，提出加强人工智能数据来源的合规透明度，并且进一步围绕治理提出“灵活监管”方式，比如以风险评估和管理为主要举措。2023 年 1 月，NIST 发布《人工智能风险管理框架 1.0》，保证在风险控制的前提下，促进可信赖、负责任 AI 系

统的开发与应用。

在人工智能数据安全方面，美国聚焦人脸识别、自动驾驶等人工智能场景加快场景化的数据安全立法和州立法。在人脸识别场景，《加州人脸识别技术法》《人脸识别服务法》这两部地方性州立法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，截至 2020 年，美国约 18 个城市颁布了相关法律，2021 年，美国再度通过五项禁止警察和政府使用人脸识别技术的市级禁令。在自动驾驶场景，根据美国高速公路安全协会（GHSA）数据，截至 2021 年，已有 38 个州及华盛顿特区实施自动驾驶相关立法或行政命令。除了确保自动驾驶汽车安全方面的职责，还要求自动驾驶汽车生产商或者系统提供商向监管部门提交安全评估证明，以证明其在数据、产品、功能等各个方面采取了足够的安全措施并且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。除了场景化的安全规制，美国出台一系列隐私保护相关法律法规，对于人工智能企业同样起到监管作用。目前，美国缺乏统筹性的数据安全法案，主要以州立法为主。仅 2021 年，美国 38 个州就出台了 160 多项与消费者隐私相关的法案。2022 年 6 月 3 日，美国众议院和参议院发布《美国数据隐私和保护法案》讨论稿，作为首个获得两党两院支持的全面的联邦隐私立法草案，有望推动美国分散的隐私立法走向统一。

2. 欧盟：重塑 AIGC 治理规范，基于风险识别强监管

欧盟在数字领域和人工智能领域一直走在前列，而且在战略层面予以高度重视。2020 年 2 月，欧盟委员会发布《欧洲数据战略》和《欧洲人工智能白皮书》，描绘了构建一个健康共通的数据空间的蓝图。同年 12 月，欧盟委员会和外交与安全政策联盟高级代表发布了新的《欧盟数字十年的网络安全战略》，为未来欧洲数字领域的发展定下十年计划。在 AIGC 监管方面，欧盟持续一贯的“强监管”态度，加快推出人工智能领域的“GDPR”。

在人工智能及算法治理方面，欧盟现行的人工智能立法聚焦于泛化的人工智能领域，意图重塑人工智能的整体治理性规范，对于 AIGC 的监管与治理要求则处于草案阶段。2023 年 5 月，欧盟通过《人工智能法案》的谈判授权草案，该法案提出“基于风险识别”（Risk-based Approach）为不同类型的人工智能系统施加不同的要求和义务。不同风险层级的人工智能所承担的责任和义务将有所区分。其中，ChatGPT 等生成式大模型产品和服务被归于高风险层级，因此不仅需要遵循《人工智能法案》的相关条款，还需要遵循额外的透明度的要求，譬如 AI 标识、训练数据来源合法性等。

在人工智能数据安全方面，欧盟通过 2018 年 4 月生效的《通用数据保护条例》（GDPR）框架构建了一套统一完备的数据安全治理体系。GDPR 对用户数据权利进行全面系统的梳理，进而对欧盟人工智能数据安全起到基础性规制作用。近年来，围绕在线平台的非法内容打击、个人数据保护以及推动数据共享机制完善，欧盟进一步推动《数字服务法案》《数字市场法案》《数据治理法案》的颁布与实施，为人工智能企业发展过程中的数据安全监管奠定法律基础。

3. 中国：实施 AIGC 适度监管，多层级、多角度规范

2023 年以来，我国 AIGC 市场进一步释放需求、监管体系逐步完善。从国内人工智能的发展历程来看，初步发展阶段（2013–2016 年）主要是在产业、经济的发展规划的文件中提及人工智能，同时以技术赋能产业的方式鼓励人工智能发展；飞速发展阶段（2016–2017 年），国务院印发“十三五”国家战略性新兴产业发展规划，增加人工智能相关内容，首个中国人工智能国家战略《新一代人工智能发展规划》发布，为后续国标和行标的制定和印发奠定了基础。针对性发展阶段（2017 年至今），人工智能产业进入以技术找场景、以技术找产业的阶段，人工智能安全治理受到高度重视。2017 年国务院发布的《新一代人工智能发展规划》明确指出“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”。2023 年 5 月，习近平总书记在主持召开二十届中央国家安全委员会会议时再次强调，要提升网络数据人工智能安全治理水平，以新安全格局保障新发展格局。

在人工智能及算法治理方面，我国虽然未出台一部针对人工智能的通用性的监管法律，但采用了多部法律法规衔接的监管方式。2023 年 7 月，我国出台首部针对 AIGC 的部门规章《生成式人工智能服务管理暂行办法》，明确对生成式人工智能服务采取“包容审慎”和“分类分级监管”的治理策略，虽并未对如何进行分类分级提供具体的规定，但根据国务院办公厅于 2023 年 5 月 31 日发布的《国务院 2023 年度立法工作计划》，其计划在 2023 年提请全国人大常委会审议人工智能法草案，该法案或针对人工智能、AIGC 具备更高的法律效力和更为广泛的适用范围，为我国 AIGC 在安全合规中发展提供指引。

此外，我国聚焦人工智能重点技术领域，发布《互联网信息服务深度合成管理规定》《人工智能 深度合成图像系统技术规范》《生成式人工智能服务内容标识要求（征求意见稿）》等规范性文件。聚焦算法安全领域，发布《互联网信息服务算法推荐管理规定》《信息安全技术 机器学习算法安全评估规范（征求意见稿）》，通过算法备案和安全评估“双监管”完善算法安全治理。相关规范之间形成交叉，构建现有的 AIGC 监管体系。

在人工智能数据安全方面，2020 年以来，我国加快国家层面数据安全统一立法的进程。目前已经形成以《网络安全法》《数据安全法》《个人信息保护法》为基础的综合性、全局性法律。其中包含“针对小型个人信息处理者、处理敏感个人信息以及人脸识别、人工智能等新技术、新应用，制定专门的个人信息保护规则、标准”等。

二、AIGC 治理前沿技术发展趋势

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 AIGC 治理能力。目前重点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，以及人工智能数据安全检测和对抗领域的专用技术，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等。

1.AIGC 数据安全——隐私计算

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。隐私计算的技术方向包括：

同态加密：对加密数据进行处理从而得到一个输出，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。

多方安全计算：针对无可信第三方情况，安全地进行多方协同计算。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。前者是指针对特定的计算逻辑，比如比较大小，确定双方交集等。后者则可以采用多种不同的密码学技术设计协议。当前，多方安全计算的主要适用场景包括：1）数据安全查询。2）联合数据分析。

差分隐私：通过对数据添加干扰噪声的方式保护数据中的隐私信息。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，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。

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。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须导出企业数据，不仅可以充分保护数据隐私和数据安全，还可以获得更好的训练模型，从而实现互惠互利。

2.AIGC 数字信任——区块链

区块链作为建立在互联网之上的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。区块链本质是一个去中心化的共享数据库，其具备“不可伪造”“全程留痕”“可以追溯”“公开透明”“集体维护”等特征。

目前区块链已形成三种模式：1）公有链（运行在互联网的完全分布式区块链）；2）联盟链（由多个关联机构共同发起和运营，带有准入机制）；3）私有链（公有链的私有化部署，由单个机构运营）。

随着区块链被应用于广泛应用于供应链金融、商品溯源、版权存证、司法存证等领域，区块链凭借其技术特性为数据流动过程中的安全提供保障，包括用于数据确权、数据完整性校验和数据的追踪溯源等。

3.AIGC 模型安全——对抗训练

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。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。

针对现场数据的对抗样本攻击，目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。

4.AIGC 反歧视——偏见检测

训练数据的不足和偏见将导致 AI 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。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种族和性别偏见问题。

5.AIGC 内容安全——鉴别及标识

随着人工智能生成能力的提升，人类对于 AIGC 生成的内容无法直接感知，AIGC 的滥用将导致虚假新闻、AI 生成图片及音视频泛滥，网络空间的信息可信度进一步下降，对于打造清朗的网络空间带来新的挑战。针对这一现象，已有一批企业研究并应用 AIGC 内容的识别、标记技术。譬如基于深度伪造视频图像检测鉴定引擎，对 AI 生成的内容进行鉴别；通过自动添加显式水印标识，实现人类可感知，通过空域水印或变换域水印的方式添加隐式水印标识，实现人类无感但技术手段可抓取标识。

PART 3

AIGC 数据和算法安全治理框架

人工智能的发展是长期的、动态的过程，AIGC 作为人工智能现阶段的重要产物，其安全治理框架也是目前人工智能发展所重点考虑的议题。我国需抢占先机，充分发挥“安全”对“发展”的赋能作用，总结算力经济背景下具有中国特色的治理范式，提高我国在人工智能安全领域的国际话语权和影响力。

一、治理原则

● **健康有序，鼓励创新应用：**应当以鼓励数据安全、便捷、低成本地共享流动，促进人工智能健康发展为前提和根本，在法律法规和政策层面关注主要风险和企业合规成本，避免因过于严苛的治理举措抑制技术创新和产业发展。

● **技术向善，保障主体权益：**应当充分保障多方主体的权益，包括公民的隐私权和知情权，企业的商业秘密和公平竞争权，国家的数据安全等。确保人工智能发展应用中的公平正义、机会均等，增进社会整体福祉。

● **敏捷治理，强化技术赋能：**积极拥抱前沿科技，灵活运用智能治理工具赋能数据在合规中高效流动，加强智能化技术手段在算法风险全流程治理过程中的创新应用。

● **多元包容，推动国际合作：**加强企业间、行业间交流合作，持续探索企业自治和行业自律的最佳实践。在尊重和开放基础上，加强国际合作，加快形成国际共识、制定国际标准、对接法律框架。

二、治理路径

1. 完善治理顶层设计

● **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，决策层应当将人工智能数据安全和算法治理纳入国家整体安全观当中，组成人工智能战略和发展规划的重要板块。

● **法律法规：**加快推进人工智能整体立法，完善人工智能数据安全和算法治理的专门立法。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”的策略解决人脸识别、人工智能生成等特定应用的安全挑战。

● **监管机制：**人工智能风险将威胁用户隐私、公民权益、商业秘密、国家安全等，应当构建由各部门、地方的主管部门负责主管行业、地域的数据安全和算法治理的监管机制，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. 建设系列标准体系

● **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中，重点关注人工智能安全与数据安全、算法治理的交汇处，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：**针对人工智能技术形成一整套通用标准，包括人工智能芯片、人工智能终端、算法模型框架等维度，以及人工智能数据采集、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等维度。

● **行业标准：**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和算法治理标准。

3. 提高人工智能企业自身治理能力

● **组织建设：**设立负责人工智能安全治理工作的组织机构和专职人员，进行明确的职责分配。制定组织内部的数据安全制度规范、数据安全风险管理、算法治理策略等，对组织内部的大规模数据训练、数据处理活动、算法设计等进行指导和监督。

● **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，部署必要的安全产品和服务，通过技术手段辅助数据安全制度规范、数据安全风险管理、算法治理策略的实施。

● **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、数据合规处理、数据安全运维、算法测试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公平透明、数据安全融入企业组织文化建设，要以此为理念开展模型设计、人工智能产品的研发和测试，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. 打造全面的安全治理能力供给

人工智能安全治理涉及技术、法律、产业发展等各个领域，因此政府、产业界和社会机构应当共同形成一个覆盖法律、技术、产品、服务、咨询的整体安全能力供给。具体而言，应当包括以下四个层次：

● **技术研发：**我国政府应当加大对大模型、数据安全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投入，形成一批高技术研发能力的技术实验室和研究所，研究方向包括多方安全计算、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：**人工智能企业、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，构建包括测试工具、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、企业应当协同提升人工智能安全测评认证能力，建立人工智能测评服务平台，为企业提供有关数据质量、数据安全、算法模型的专业测试和评级服务；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、算法认证。

● **合规咨询：**打造人工智能领域“合规供给箱”，为人工智能企业提供包括风险评估、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，围绕全球人工智能发展、人工智能安全风险形势和人工智能安全治理实践，形成一批专业的研究成果和公开报告，为企业提供模式经验和最佳实践参考。

参考文献

1. 赛博研究院 , 观安信息 . 人工智能数据安全治理报告 [EB/OL]. (2020-07-10)
2. 赛博研究院 . 非接触新经济安全治理报告 [EB/OL]. (2020-05-20)
3.Privacy-Preserving Machine Learning 2018: A Year in Review.
4. 华为 .AI 安全白皮书 [EB/OL]. (2019-09-19) .
5.Uncovering and Mitigating Algorithmic Bias through Learned Latent Structure

SICSI
CYBER RESEARCH INSTITUTE
赛博研究院

AIGC数据安全与算法治理报告

AIGC Data Security and
Algorithmic Governance Report