CAICT 中国信通院

人工智能背景下数据安全八大发展趋势

中国信息通信研究院
云计算与数字化研究所

2026年7月

数据成为人工智能时代的动力基座，数据安全工作面临全新挑战

在人工智能时代，数据是驱动技术演进的核心要素。人工智能通过深度挖掘数据的内在价值，支撑着大模型的训练、推理与落地应用。随着数据成为人工智能发展的关键引擎，其在整个生命周期中的安全性，已成为保障人工智能持续健康发展与业务平稳运行的重要前提。然而，伴随数据价值的空前提升与人工智能技术的深度应用，数据安全也面临着诸多前所未有的新挑战。

数据是人工智能生态构建的底层驱动

随着人工智能模型趋于通用化、算法逐渐开源，数据将成为企业之间竞争的关键变量。掌握了最丰富、最真实、最有洞察力的数据，就拥有了未来的主动权。

人工智能领域的马太效应

数据价值与数据保护工作密切相关

数据被誉为人工智能时代的“石油”，不仅仅是训练模型的原始材料，更是驱动技术创新、未来决策的核心资源。然而与原油价值稳定不同，数据价值与数据安全息息相关，数据安全工作失误可能会导致数据价值的大幅下降。

数据泄露--模型竞争力根基崩塌，数据价值锐减

- 大模型高度依赖高质量、独特的数据训练模型参数，核心数据的泄露会导致竞争对手可以利用这些数据快速克隆或超越自身模型。

数据质量污染--模型性能劣化，“Garbage In, Garbage Out”

数据质量污染 → 模型泛化放大 → 影响模型参数 → 模型能力下降

人工智能时代数据安全工作面临的WWWH

What

哪些数据需要保护

旧（类型单一）：业务数据、日志数据、运维数据等；

新（类型复杂）：模型数据、临时数据、生成数据等。

Where

数据会在哪些场景暴露

旧（暴露面较少）：

传统场景中，数据存储在安全域内，暴露面较少。

新（暴露面激增）：

新增场景：

- 大模型训练：敏感数据寄存在模型“记忆”中；
- 大模型推理：不同数据关联推理暴露额外敏感信息。

Who

都有谁会参与数据流通

旧（没有数据流通）：

- 数据存储于企业内部，以拷贝、定向交换的方式共享数据，参与者固定。

新（频繁数据流通）：

- 数据在诸多需求方与提供方间流通，多方、无国界的合作共享面临数据滥用、数据合规等痛点。

How

数据会被怎么使用

旧（用途清晰）：

发起方：人 → 数据处理

决策逻辑清晰，流通链路可控，可溯源

新（用途模糊）：

发起方：智能体 → 数据处理

决策逻辑黑盒，链路不可控，难溯源

人工智能时代数据安全八大趋势

针对人工智能背景下的全新数据安全痛点，中国信通院汇聚业内智慧深入研究，从数据责任、数据使用 、数据流通、智能体安全四个视角出发，梳理人工智能时代数据安全的八大趋势，作为应对数据安全挑战的系统性策略，为行业高质量发展筑牢安全底座。

数据责任视角：明确数据的权责归属，敦促各方履行数据安全责任

趋势一：三权分置

落实数据三权分置，推动数据产权制度成为共识

趋势二：责任划分

数据安全责任从静态划分转向场景化动态适配

数据使用视角：以核心技术突破重塑安全边界，实现数据利用与保护并举

趋势三：机密计算

机密计算正在成为数据“使用中安全”的新基建

趋势四：模数共振

精准布控“模数共振”汇聚、训练、协同环节的数据安全

数据流通视角：保障安全可信进行数据流通，进一步释放数据价值

趋势五：数据空间

可信数据空间迈向人工智能原生驱动新阶段

趋势六：合规出海

以“来数加工”化解Token出海数据安全合规痛点问题

智能体安全视角：筑牢智能体内生安全防线，确定数据安全可信

趋势七：可信智能体环境

从身份、权限、行为三维度构建智能体可信数据使用环境

趋势八：智能体零信任架构

零信任理念正在缝合智能体协同工作引入的数据安全罅隙

趋势一：落实数据三权分置，推动数据产权制度成为共识

人工智能时代，数据类型向多模态扩展，参与主体日益多元，传统产权模式已制约数据价值释放。数据“三权分置”精准适配这一特点，针对模型、推理、训练、衍生等不同数据类型，根据采集者、处理者等不同角色，进行了差异化的权责分配。通过三权的灵活组合，该制度有效提升了产权体系的弹性，在满足多元实践需要的同时，为数据创新留足了空间。

数据产权

指对特定数据享有的财产性权利

数据持有权

自行持有或委托他人代为持有合法获取的数据的权利。

数据使用权

以加工、聚合、分析等方式，将数据用于优化生产经营、形成衍生数据等的权利。

数据经营权

以转让、许可、出资或依法设立担保等有偿或无偿的方式对外提供数据的权利。。

人工智能时代数据产权发生结构性变化

变化	传统数据权责	人工智能时代数据权责
权利形态重构	沿用物权逻辑 持有、使用、收益四权合一	推崇“非竞争性、多源共生” 持有、使用、经营灵活组合
价值产生来源改变	价值来源：独占、存储	价值来源：“流动”和“加工” 不转移所有权的前提下被多方主体 高效利用，释放乘数效应
确权逻辑演进	侧重于界定“数据归谁所有”	侧重界定“谁对数据有什么权利” 差异化配置权利，平衡各方利益。

数据产权分置应根据数据类型不同针对性配置规则

在人工智能时代，数据具有非竞争性、多源共生及价值动态衍生等独特属性，只有根据各参与方对不同数据类型的实际投入与贡献进行差异化确权，才能精准平衡各方权益、降低流通成本，最大化激发数据要素的创新潜能与乘数效应。

	模型数据	推理数据	衍生数据	生成数据
数据所有者	数据持有权 数据使用权	数据持有权 数据使用权 数据经营权	数据持有权 数据使用权 数据经营权	数据持有权 数据使用权 数据经营权
模型拥有者	数据持有权 数据使用权 数据经营权	——	数据持有权 数据经营权 ➢需提前协商	数据持有权 数据使用权 ➢需提前协商
智能体服务商	数据使用权	——	——	数据使用权 ➢需提前协商

趋势二：数据安全责任从静态划分转向场景化动态适配

传统数据安全责任建立在“谁掌握数据谁负责”的清晰前提下。人工智能时代，模型与数据的由谁负责，随场景（自训练、API调用、平台工具链、智能体应用等）动态变化，人工智能时代数据安全的责任边界愈发模糊、治理难度大幅上升。

人工智能时代下的数据安全责任划分动态化、场景化

维度	传统数据安全责任	人工智能时代数据安全责任
责任主体	用户+服务商	用户+服务商+大模型拥有者
典型场景	自研业务系统处理数据、云存储归档数据、购买软件采集数据等	使用智能体应用处理数据、调用大模型API对数据推理、用人工智能工具链训练模型、自训练模型
保护对象	个人数据、业务数据、系统数据	新增：模型参数、训练数据、生成内容、提示词等
典型风险	DDos、数据泄露、漏洞利用、入侵	模型投毒、后门攻击、对抗样本、模型偏见、输出违规内容
责任边界	**较清晰，谁掌控数据，谁对数据负责**	**较模糊，谁提供模型、谁管模型参数、谁管用户输入、谁管模型输出，要根据场景具体分析责任**

识别关键安全事项 分析责任的动态转移

人工智能时代，“模型、数据、算力”组合向用户提供服务，数据安全责任责任跟随：**模型归谁控制 和 数据位置**，动态适配。

以“模型输出内容安全”为例，分析不同场景下“输出内容安全” 的责任归属情况：

典型场景	主责方	变化原因
智能体应用	智能体服务商	整个应用由智能体服务商运营，输出内容安全是智能体服务商责任
API调用大模型服务	大模型拥有者+用户	• 大模型拥有者提供基础输出过滤 • 用户设计业务逻辑，负最终责任
用平台工具链训练模型	云服务商+用户	• 云服务商提供基础模型安全工具 • 用户训练模型，负最终责任
自训练模型	用户	模型为用户自己搭建并训练，输出内容完全由用户控制

第一步

典型场景识别

识别数据流向、模型控制权

第二步

风险映射

识别场景下特有的安全风险，风险映射关键安全事项

第三步

控制权矩阵分析

对安全事项的算力、数据、模型、应用分析实际控制权

第四步

场景化责任矩阵

为关键安全事项指定主要责任方

趋势三：机密计算正在成为数据“使用中安全”的新基建

过去数据安全聚焦“存储”与“传输”，但在数据被CPU加载、运算、输出的“使用”阶段，以明文形态暴露在内存环境，成为安全链路最薄弱环节。机密计算通过硬件级可信执行环境，在计算过程中将数据与代码完全隔离于受保护区域，即使操作系统、平台管理员也无法窥探，实现数据“可用不可见、结果可验证不可伪造"。

基于机密计算实现数据“可用不可见”，依托硬件可信根保障模型推理“部署不裸奔”

✗	✓
数据共享不敢用 多方数据协作中，原始数据交付后，无法约束使用范围与频次。	**TEE飞地隔离计算** 数据在独立TEE内运算，原始数据零暴露。
云上计算不敢信 云平台管理员、宿主OS均有权限读取内存明文。	**硬件可信根+远程证明** TEE启动前远程证明飞地完整性，云平台无法伪造或篡改可信环境。
合规审计难闭环 "使用中"阶段黑盒无法举证合规，监管与审计缺乏技术抓手。	**全流程可验证链路** 附带远程证明报告，实现全链路可审计。
模型资产易泄露 推理部署后模型权重可被内存抓取逆向工程。	**模型在TEE中运行** 模型权重在TEE内加密加载与推理，外部无法读取模型参数。

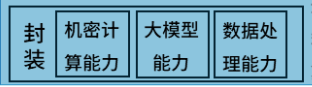
机密计算根据应用形态划分五大应用等级，保障人工智能不同应用场景数据安全

L5 机密智能体

将智能体部署在机密硬件上，提供覆盖数据、凭证、运行时计算及大模型推理的全流程机密计算保护。

具体形态：独立TEE架构的密态“龙虾盒子”，具备机密计算协处理器与全栈机密计算架构。

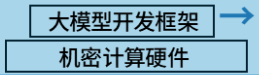
L4 机密模型平台



通过API提供开箱即用的机密计算能力。

具体形态：可信机密计算平台、MaaS机密计算专区、多方安全计算（MPC）与机密计算融合平台。

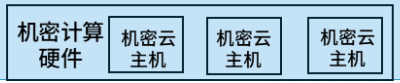
L3 机密模型开发框架



模型与数据“可用不可见”

具体形态：机密推理引擎、机密大模型服务框架、人工智能隐私过滤中间件等。

L2 机密云主机



具体形态：机密计算虚拟机（如Azure Confidential VMs）或机密容器（Kata Containers + GPU TEE）等。

L1 机密计算硬件

具体形态：支持机密计算的芯片，如B200机密计算GPU、Intel TDX、华为昇腾NPU机密计算环境等。

趋势四：精准布控“模数共振”汇聚、训练、协同环节的数据安全

2026年4月，工业和信息化部、国家数据局联合印发《关于联合实施2026年"模数共振"行动的通知》，面向制造业领域20个重点行业，推动人工智能模型与数据资源协同互促、同频共振，提出"以模引数、用数赋模、模数共振"核心机制。

除新型工业化外，农业、能源、交通、医疗等传统行业也在积极借鉴"模数共振"经验推进本行业的模数协同。然而，模数协同在释放数据价值的同时，也带来了数据汇聚融合、模型训练流转、多方协同共享等环节的新型数据安全风险，传统行业在借鉴过程中应重点关注并提前应对。。

风险点一：数据汇聚融合安全风险

■ 安全风险

模数共振推动跨主体、跨行业数据大规模汇聚融合，打破了原有数据安全边界：

- 敏感信息交叉关联产生“1+1>2”的泄露风险；
- 数据出域流转过程中暴露面扩大。

解法一：筑牢数据汇聚融合安全防线

■ 布控应对

制度保障：数据分类分级

✓ 明确敏感数据流通边界，从出口进行数据管控。

技术保障：可信数据流通、机密计算

✓ 实现数据“训练不出域”“可用不可见”，破解"不愿给、不敢出"壁垒。

风险点二：模型训练数据泄露风险

■ 安全风险

专用模型训练和特色智能体开发需要大量高质量行业数据，训练数据本身包含敏感信息：

- 模型“记忆”训练数据导致数据泄露；
- 智能体自主调用数据存在权限失控风险；

解法二：强化模型训练与智能体数据安全

■ 布控应对

✓ 安全审查：确保入模数据合规脱敏；

✓ 安全防护：加强模型逆向攻击防护，防止训练数据泄露；

✓ 权限管控：实施智能体最小权限原则。

风险点三：多方协同流转确权风险

■ 安全风险

“模数共振空间”涉及多主体协同与利益分配：

- 数据权属不清导致流通纠纷；
- Token化确权体系面临伪造篡改风险；

解法三：构建可信协同流转与确权机制

■ 布控应对

✓ 建立多方安全计算与身份可信验证体系，实现“流转可控、使用可审、责任可究”。

✓ 建设可信数据流通基础设施，明确数据权属与责任边界。

趋势五：可信数据空间迈向人工智能原生驱动新阶段

随着大模型和智能体与可信数据空间结合探索走深向实，其底层技术架构、数据治理流程等与人工智能深度融合，智能体在受控的可信数据空间环境中调用高价值数据，可信数据空间不再仅被当作“数据流通的管道”，而是升级成为“支持智能协作的环境”。

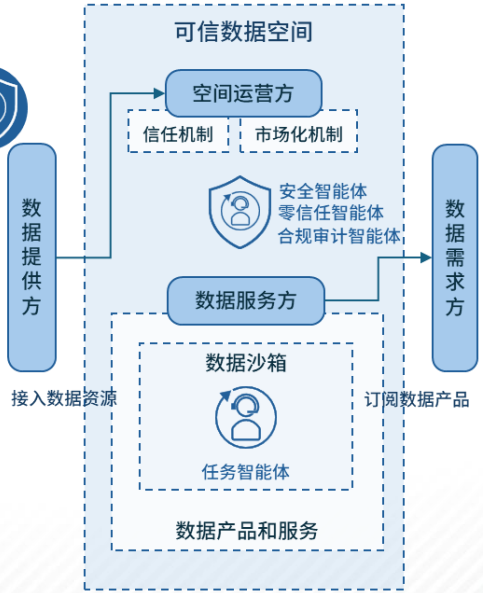
可信数据空间从“规则驱动”迈向“原生智能”

智能体成为可信数据空间的**任务执行者**，根据任务目标自主检索、调用数据，跨数据集组合、推理。

智能体成为可信数据空间的**安全执行者**。诸如零信任智能体、安全智能体、合规审计智能体在空间中运行。

	传统可信数据空间	人工智能可信数据空间
数据怎么用	被动：签了合约后人工查询	主动：任务智能体自主检索、组合、推理
谁管安全	合约是静态的，违规靠事后追责	安全智能体在空间内实时感知、动态处置、全程留痕
谁执行任务	无任务，只有数据和规则	智能体是空间内的"数字居民"，能感知、决策、执行
空间是什么	**数据流通的合约管道**	**智能协作的运行环境**

智能体以智能治理、模型推理等手段反哺可信数据空间



治理能力从“人治”到“自治、智治”

以前

合约约定谁在什么条件下使用什么数据，出现问题查询日志、找责任人、走追责流程。

现在

安全类智能体驻扎可信数据空间持续运营：

- 安全智能体：实时感知异常、联动处置
- 零信任智能体：持续验证与授权
- 合规审计智能体：识别主体责任

推理能力从“人工规划”到“自主执行”

以前

数据产品的类型与使用方法以人参与为主：

- **数据集**：需求方拿走用
- **分析模型**：需求方调用
- **API接口**：按需取字段
- **报告服务**：固定产出物

现在

- 任务智能体按任务需求**自主决定**如何开展推理
- 任务智能体可以**跨多个数据产品组合推理**

趋势六：以“来数加工”化解Token出海数据安全合规痛点问题

人工智能时代，Token出海核心在于将国内的电力、算力通过大模型转化为高附加值的数字服务出口，是一种典型的服务出海模式，其安全合规不仅考虑传统数据跨境传输，更叠加了模型服务调用、智能体协同等维度，通过设立“数据保税区”实现“来数加工”全链路闭环。

从出海结构化文件升级为动态语义交互 引入数据安全合规新风险

	数据出海	模型出海	Token出海
本质	数据出境	模型权重文件出境	模型路由能力+多模型服务出海
数据流向	单向出境	单向出境	**入境再出境：**调用境外用户输入→路由决策→推理（境内）→返回境外
发生频率	批次传输	一次性部署	**高频调用：**每次调用即跨境事件，持续高频
数据落脚点	单一境外服务器	单一境外部署节点	**数据多落脚点：**路由决策使得数据在多个模型节点间跳转

风险一：高频跨境调用让传统按次审批的监管机制失效

Token出海是持续运行的服务，每天数万次Token调用都构成跨境事件，难以实现逐笔审批。

风险二：路由决策产生多落脚点，数据处理路径难举证

Token出海的路由引擎可在多个模型间跳转，用户难以向监管提供完整的数据处理活动记录。

风险三：境外监管调取审计记录，可能触发数据二次出境

日志存储于境内，交付给境外监管机构，又构成一次数据出境事件，触发一个新的合规问题。

风险四：多司法管辖区数据权利要求冲突

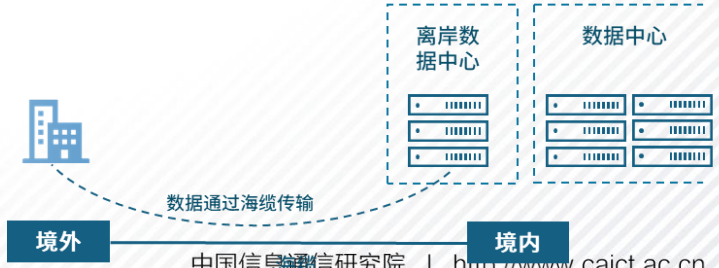
Token云服务面向全球市场，不同辖区对“删除、保留等”定义和执行力度的要求存在冲突。

“来数加工”政策助力Token出海从概念走向落地

2026年4月，全国首个城市级“Token出海”全链路在广东汕头闭环验证

- 坐拥60千兆瓦海上风电资源
- 国家批复“来数加工”试点政策
- 海缆登陆站承载全国半数出口带宽
- 百年侨乡，生态天成

通过设立“数据保税区”，境外数据经合规通道进入境内专用专区的算力集群处理，处理结果经海缆直连返回境外，与国内数据隔离。全链路闭环验证的模式如下图所示，“海外调用→离岸数据中心处理→Token消耗→生成数据→数据合规传输”即为典型的离岸数据加工模式。



趋势七：从身份、权限、行为三维度构建智能体可信数据使用环境

一个执行单元的视角：解决一个智能体跑起来后，数据安全如何管控

人工智能智能体具备自主规划、主动调用工具的“主动性”特征，其在身份形态、授权模式、行为可预测性三方面与传统应用存在差异，传统安全防护手段难以应对。通过身份可信、权限可信、行为可信三层体系，构建智能体运行可信环境，化解引入的新型数据安全风险。

智能体的身份、授权、行为三方面特点引入新的数据安全风险

与传统应用相比，智能体应用在三方面有显著不同：一是身份形态，智能体可代表人类执行任务；二是授权模式，智能体是动态、任务级授权，权限可膨胀；三是行为可预测性，智能体具有自主行动能力。这些特点为引入了新的数据安全风险。

维度	传统应用vs智能体应用	引入新的数据安全风险
身份形态	固定服务账号 vs 代表人类	身份边界模糊，低权限智能体向高权限智能体转发恶意请求，导致**越权执行数据操作**
授权模式	静态、长期 vs 动态、临时	权限过度泛化、任务结束未及时收回，当**智能体被利用时导致数据被窃取**
行为可预测性	可穷举 vs 无法预测	攻击者**诱导智能体执行非预期操作**，将敏感数据伪装成日志发送至外部，难以被识别

构建三层智能体可信环境 从源头降低数据安全风险

根据梳理的三大数据安全风险根源——身份边界模糊、权限过度泛化、行为不可验证，智能体可信环境需要从对应三个方面系统性构建，身份可信解决谁在访问数据问题，权限可信解决能访问什么数据问题，行为可信解决数据去了哪里问题。

- 谁在访问数据？
  - **身份可信－从源头消除身份不可见**
  - 为智能体分配唯一身份，绑定自然人
- 能访问什么数据？
  - **权限可信－从源头杜绝权限过宽**
  - 仅授予当前任务所需最小权限
- 数据去哪里？
  - **行为可信－从源头阻断非预期数据外发行为**
  - 所有数据调用在沙箱环境中进行

趋势八：零信任理念正在缝合智能体协同工作引入的数据安全罅隙

多智能体协同视角：解决多个智能体间如何建立信任架构

智能体可信环境主要站在一个执行单元视角，观察由智能体本身特性引入的数据安全风险。回归实际业务场景，多智能体协同工作，数据在智能体间流动，信任在智能体间默认传递，人与应用间可用制度约束，而智能体间缺乏信任机制，基于零信任理念为智能体数据安全构建“新边界”。

多智能体协同引入数据流转失控&共享数据污染风险

1\数据传递过度暴露，缺少数据流转校验

上游智能体将全部上下文传递给下游智能体，下游智能体被动接收大量无用原始数据。

2\数据路由不可见，缺少可见性

编排引擎自动分发任务给不同智能体，管理者缺乏数据经过哪些智能体的全链路视野。

3\数据来源无法追溯

多个智能体串联加工后，最终输出敏感信息难以追溯产生的源头。

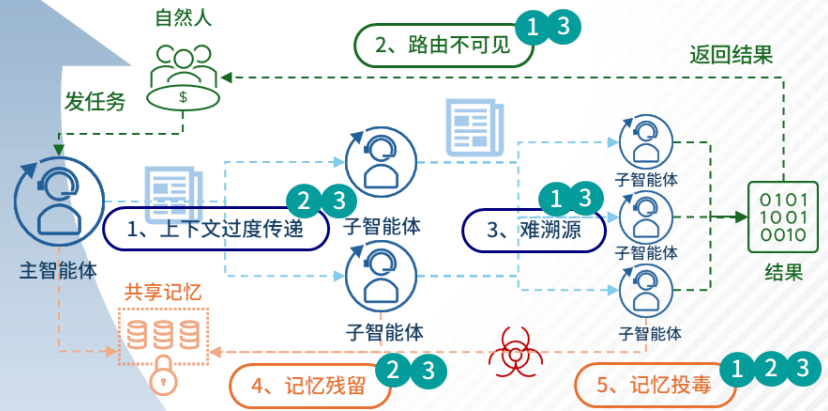
4\数据残留，记忆缺乏过期删除机制

前次会话接触的敏感数据写入共享记忆，被后续无关任务调度，数据过期不删。

5\数据投毒传播，缺少隔离机制

攻击者在一次会话中注入的恶意内容写入共享记忆，潜伏后扩散至多个下游智能体。

零信任理念在智能体协同场景下得以扩展



原则一：永不信任，始终验证	从“会话级”扩展至“每次工具调用级”
原则二：假设失陷	从“系统级”扩展至“智能体行为级”
原则三：从最小权限到最小代理	从“用户级”扩展至“操作级”