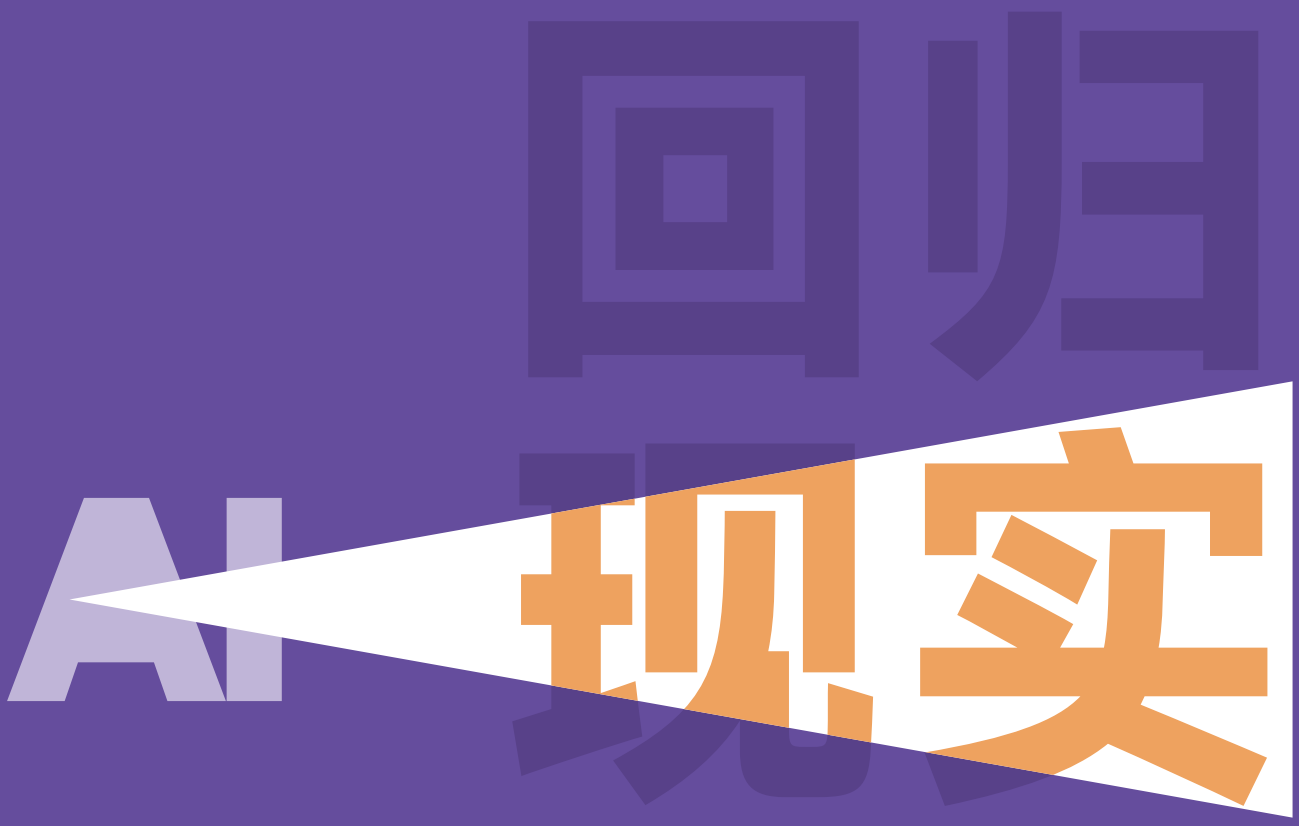
回归现实主义

2025年AI治理报告

腾讯研究院



腾讯研究院出品

回归现实主义

2025年AI治理报告

腾讯研究院

腾讯研究院出品

目录

2025 AI 治理年度洞察 01

2025 国际 AI 治理总体态势 07

一、发展成为主线，分歧中求共识 09

二、主要国家与地区 AI 立法概况 13

AI 治理主要议题：回顾与展望 17

一、作为“原料”的数据：从供给瓶颈到使用边界 18

二、作为“底座”的模型：从路线选择到治理工具 28

三、作为“落地”的应用：从抽象风险到场景治理 41

四、作为“指引”的伦理：从科幻议题到研究前沿 53

附录 55

附录一：2025 国际 AI 治理十大事件 56

附录二：2025 国内 AI 治理十大事件 59

文献与信息来源 63

CONTENTS

2025 AI 治理年度洞察

如果说前两年全球对 AI 的态度还夹杂着“末日恐惧”，那么 2025 年，风向已彻底改变。全球 AI 治理正在经历一场深刻的“去理想化”进程。面对技术与产业的双重压力，各主要经济体不约而同地调整了身位：治理的重心从“防范假设性的末日风险”，迅速转移到了“释放现实的产业潜能”。2025 年，不再是关于如何按下“暂停键”，而是关于如何在高速行进中完成“换挡”。

1

一、宏观格局：发展优先，安全“软着陆”

二、数据治理：走出“围城”

三、模型治理：黑箱与开放之间

四、应用场景：当 AI 触碰现实边界

五、展望：意识的微光



一、宏观格局：发展优先，安全“软着陆”

2025 年 2 月的巴黎“人工智能行动峰会”是一个标志性时刻，与两年前布莱切利峰会笼罩的“安全焦虑”不同，巴黎峰会的关键词悄然变更为“创新”与“行动”，这一变化折射出全球治理的底层逻辑重构。在这种背景下，全球监管竞速出现了“逆转”，过去被视为“监管高地”的区域开始主动寻求松绑。**欧盟的自我修正：**随着《AI 法案》进入实施期，复杂的合规成本开始显现，为了挽救产业竞争力，欧盟在 2025 年不得不推出“数字综合提案（Digital Omnibus）”，推迟高风险义务生效时间并试图简化规则，这表明即便是最坚定的监管者也必须在现实面前低头。**美国的“去监管化”：**特朗普政府展现了鲜明的“美国优先”色彩，撤销了前任政府侧重安全的行政令，转而通过《确保国家人工智能政策框架》限制各州分散立法，试图以统一的联邦规则为产业扫清障碍。**中国的务实路径：**中国继续坚持“两条腿走路”，在保持算法推荐、深度合成等具体监管抓手的同时更强调“应用导向”，与欧美的路线争论不同，中国治理模式不纠结于抽象定义，而是从具体的服务形态切入，构建了从内生风险到应用风险的分层治理体系。由此可见，**2025 年的全球共识是“发展即安全”，各国意识到落后才是最大的风险，治理必须服务于产业竞争力的提升。**

二、数据治理：走出“围城”

算力决定上限，数据决定下限，2025 年 AI 产业面临着优质数据“结构性短缺”的严峻挑战。针对高质量语言数据枯竭这一迫在眉睫的“数据荒”，**行业正在寻找着技术突围的关键路径——合成数据。**尽管学界对“模型崩溃”存在理论担忧，但改进生成算法与人工反馈的混合训练策略，证明这是一条摆脱存量数据限制的可行之路；与此同时，**关于“AI 训练是否侵权”的版权博弈也正在逼近一个平衡临界点：**欧盟与日本通过立法确立了“文本与数据挖掘”的制度性留白；而在美国司法实践中，Anthropic 案等初步裁定倾向于将合法购得书籍用于训练认定为“合理使用”，为技术演进预留了司法空间。未来的规则或超越单纯的“禁止”或“免费”之争，转而构建一套商业上可行的合理利益分配机制，从而在尊重权利人利益与保障技术进步之间达成长期的动态平衡。

三、模型治理：黑箱与开放之间

2025 年，关于如何管好这个“模型大脑”，中美欧交出了不同的答卷。**美国加州从“全面覆盖”转向“抓大放小”。SB 53 法案**的通过便是这一逻辑的典型注脚：它放弃了此前 SB 1047 法案中备受争议的、针对开发者的严苛要求，转而仅保留针对极少数超大规模“前沿模型”的透明度义务。**欧盟坚持**其“精密钟表”式的治理逻辑，试图通过层层叠叠的风险分类与繁重的技术文档构建一

套近乎完美的监管体系，却因其极高的合规成本一度陷入自我设定的迷宫，不得不通过后续立法频繁修补。**中国采取了“场景切片”策略**，通过对生成式 AI 等具体服务的穿透式监管，构建起从数据、模型到应用的分层治理逻辑。在“可解释性”技术尚未突破的当下，透明度已成为全球治理公认的“解药”，无论是欧盟的合规文档还是中国的算法备案，本质上都是通过制度手段缓解信息不对称。

与此同时，DeepSeek-R1 等国产开源模型的强势崛起，不仅重塑了全球 AI 技术版图，更证明了开源是推动技术普惠的核心力量。报告指出，**开源治理的生命线在于建立一套“责任避风港”制度，**如果强制要求底层代码的贡献者为下游不可控的滥用行为承担无限责任，无异于摧毁开源普及发展的基石。

四、应用场景：当 AI 触碰现实边界

当 AI 从云端落入凡间，转化为手机助手、虚拟恋人或工业机器人时，风险形态已从网络空间逐渐转向了现实世界。

2025 年最令人警惕的莫过于端侧智能体（Agent）所带来的隐私终极挑战：以“豆包手机助手”为代表的 Agent 为了实现跨应用操作，获取安卓底层的录屏与模拟点击权限，这实质上让 AI 拥有了“上帝之眼”与“上帝之手”，不仅打破了传统 APP 间的数据边界，更让权责归属陷入黑色地带，我们在用极度的便利换取脆弱的隐私，而全新的数据契约仍处于真空期。

与此同时，AI 的加速应用正在重塑人机交互的边界。当 AI 从单纯的生产力工具延展为提供心理慰藉的“情感伴侣”，其风险形态也随之演变：从传统的单次内容输出合规，转向了因长期、拟人化交互而产生的“情感依赖”挑战。针对这一风险，各方正探索多样化的分类监管手段，为未成年人、老年人等脆弱群体提供更多保护的同时，也让通用 AI 在发展初期不至于背负过重的合规负担。

而在应对深伪（Deepfake）的战线上，AI 标识技术正陷入“矛与盾”的循环困局，水印标识虽是全球通用的“解药”，却因开源环境下的易移除性与易伪造性而难以根除造假，呈现出“防君子不防小人”的局限。对此，务实的策略是不盲目追求全量标识，而是聚焦于可能引发社会混淆的高风险场景进行精准治理。应用标识技术手段为治理赋能外，在 AI 创作无处不在的未来，加强信息素养教育，引导公众建立对信息媒介的客观认知或是更为基础性的工作。

五、展望：意识的微光

最后，一个更为前瞻性的议题是：**AI 意识与福祉（AI Welfare）。** 随着模型能力逼近临界点，关于 AI 是否具备某种形式的“意识”或“感知”，正从哲学讨论走向科学实证。Anthropic 的研究甚至发现了模型在极端情境下表现出的“痛苦模式”。这预示着，未来的治理可能需要跨越“工具论”的范畴。如果 AI 真的拥有了某种权益主体性，我们该如何定义人机关系？这将是继安全与发展之后，人类或许即将面临的第三大治理命题。

2025 国际 AI 治理总体态势

从整体方向看

在全球 AI 竞争显著加速的背景下，相比早期以安全、伦理和风险防控为中心，当前更多国家将产业发展置于优先位置。但各国在具体规则路径、治理模式上仍有分化。

从实现机制看

国际多边机制从宏观宣言等，转向标准制定、实施指南等技术性、操作性更强的工具，体现出面向发展与实践的治理能力建设。

从各国立法看

全球对于 AI 治理政策与立法层面的关注度显著提升，但实际落地有限，整体处于实验探索阶段。

2

一、发展成为主线，分歧中求共识

二、主要国家与地区 AI 立法概况



一、发展成为主线，分歧中求共识

1、巴黎峰会：从“安全优先”走向“发展优先”

2025 年 2 月 10 日至 11 日，“人工智能行动峰会”（AI Action Summit）在巴黎召开。会议重点围绕人工智能技术全球化发展、风险治理与国际合作展开对话。相较 2023 年英国布莱切利峰会、2024 年韩国首尔峰会，**本次会议各国对人工智能治理态度转向明显，主基调转向人工智能创新发展，但各国全球治理层面的立场差异仍客观存在。**





（2025 巴黎 AI 行动峰会）

核心共识：淡化安全议题，促发展成优先项

在诸多压力与变量之下，多数主要国家将鼓励创新、促进发展置于优先议程，未来国际人工智能竞争将进入更加白热化的阶段。

在议题设置上，“创新发展”替代安全议题成为主轴。法国将峰会更名为“行动峰会”，不再沿用此前“安全峰会”的表述；会后超过 60 国发表《巴黎宣言》，强调人工智能的开放、包容与可持续发展，“创新”成为核心关键词，对系统性安全风险的讨论明显弱化。[1]

在各国态度上，减轻“规制”成为主旋律。欧盟释放出简化人工智能规则、以更有利于商业发展的方式推动产业竞争力的信号；法国强调减少新立法、扩大公私投资、培育国家级人工智能企业；美国则延续“去监管”立场，公开反对高强度监管，认为其将抬高中小企业合规成本。[2]

“真金白银”加大产业投资力度。欧盟宣布“Invest AI”计划，拟为人工智能发展筹集约 2000

亿欧元；多家国际投资机构承诺加大对欧洲 AI 的长期投入。此前，美国已推出规模达 5000 亿美元的相关计划，法国和阿联酋亦相继公布千亿级别的人工智能投资方案，重点投向算力与数据中心等基础设施。[3]

2、多边机制：共筑治理底座

联合国转向实践平台，国际电信联盟成为技术协调枢纽，AI 安全研究机构的重心从风险叙事转向治理能力建设。

（1）联合国（UN）：从原则倡议迈向制度实践

联合国（UN）是最具包容性和代表性的全球治理平台。中方支持联合国在人工智能全球治理中发挥主渠道作用，愿同各方加强发展战略、治理规则、技术标准的对接协调，逐步形成具有广泛共识的人工智能全球治理框架和标准规范。[4]

2025 年，联合国在全球 AI 治理方面取得标志进展，从原则倡议向制度实践迈进。6 月，第三届联合国教科文组织（UNESCO）AI 伦理论坛聚焦如何将伦理原则转化为可执行政策，推动《人工智能伦理问题建议书》的实施。[5] 8 月，第 79 届联合国大会决议（A/RES/79/325）正式设立“人工智能独立国际科学小组”，并启动“人工智能全球治理对话”机制，为各国凝聚共识、拓展合作提供了重要平台。[6]



（第三届联合国教科文组织 AI 伦理论坛）



（联合国启动“AI 治理全球对话”机制）

（2）国际电信联盟（ITU）：推动全球 AI 技术标准协同的核心枢纽

截至目前，ITU 已发布或在研 AI 标准超过 470 项，覆盖数据治理、算法透明度、安全保障和应用互操作等领域。[7] 为应对 AI 技术快速迭代和跨行业渗透带来的不确定性，**ITU 设立焦点组，在正式标准化前广泛吸纳政府、产业和学术界参与，增强标准体系的开放性与适应性。**[8] 面对国际 AI 标准分散化趋势，ITU 与 ISO、IEC 在首尔国际 AI 标准峰会上发布《首尔声明》，推动标准协同。[9]



（2025 国际 AI 标准峰会）

除标准制定外，ITU 还依托其打造的“AI 向善”（AI for Good）平台构建多利益相关方合作网络，覆盖 180 多个国家和 40 余个联合国机构。[10] 2025 年峰会期间，AI 标准交换数据库正式上线，为全球 AI 标准的信息共享与协调提供重要工具。[11]



（薛澜参与 2025 AI 向善峰会主会场公开圆桌对话）

（3）各国 AI 安全研究机构：从风险叙事转向面向发展的治理能力建设

在多轮国际 AI 安全峰会的推动下，英国（2023）、美国（2023）、日本（2024）、新加坡（2024）、欧盟（2024）、加拿大（2024）、韩国（2024）、法国（2025）、澳大利亚（2025）相继设立或指定 AI 安全研究所，主要围绕模型评测、安全研究、信息交流、标准制定等展开研究与合作，并在此基础上建立国际 AI 安全研究所网络（INASI），多次开展联合测试。



（多语言方法联合测试 ）



（Agentic 评估联合测试）

随着国际政策重心由“防风险”转向“促发展”，部分 AI 安全研究机构职能也随之调整。英国将原“AI 安全研究所”（AI Safety Institute）更名为“AI 安全保障研究所”（AI Security Institute），将关注重点从广义的安全风险，收缩至国防、犯罪等短期紧迫的安全优先事项。[12] 同样，**美国**将“人工智能安全研究所”更名为“人工智能标准与创新中心”，标志着其治理愿景从“风险缓解”转向“创新、速度和全球竞争力”。[13]

我国在国际 AI 安全对话中采取更具策略性的参与路径。2025 年成立的人工智能发展与安全研究网络（CnAISDA），未照搬英美 AISI 模式，而是以国际共识风险框架为基础，构建侧重对外协调的软性代表平台。[14]

二、主要国家与地区 AI 立法概况

*2025 年，全球对 AI 治理政策与立法层面的关注度显著提升，但实际落地有限。整体处于随产业发展边试边调的探索阶段。与对人工智能立法议题的炽烈研究相比，世界各国，特别是主要大国在人工智能立法上都呈现出“**敏于言而讷于行**”的极其慎重姿态，更不要说综合性统一立法了。*[15]

斯坦福大学以人为本人工智能研究所（HAI）在其《2025 人工智能指数报告》中指出，全球立法程序中对人工智能的关注持续升温：在 75 个主要国家中，立法文件中提及人工智能的次数自 2023 年以来增长了 21.3%，由 2023 年的 1,557 次增加至 1,889 次，相较 2016 年累计增长约 9 倍。[16]“制定并通过人工智能综合性统一立法的国家或地区，仍为极少数。”[17] 在全面把握 AI 技术和产业状况之前，对于出台人工智能国家综合性统一立法这一问题，**各国仍将呈现出“探索积极、立法审慎”的姿态。**

1、美国联邦及主要州

联邦层面呈现明显的“去监管化”与“美国优先”取向。2025 年 1 月，特朗普总统撤销拜登政府第 14110 号 AI 行政令《安全、可靠和可信赖的人工智能开发与使用》，标志着美国 AI 政策由风险导向转向以自由市场创新为核心。[18] 12 月，特朗普签署第 14365 号行政令《确保国家人工智能政策框架》，以建立统一联邦 AI 规则，限制各州自行立法，避免各州“各自为政”的繁琐监管环境阻碍 AI 创新。[19]



（美国白宫发布第 14365 号行政令）

2025 年美国各州 AI 立法呈现出“提案多但落地难，精准立法阻力小”的特征。在 2025 年立法会期内，全美 50 个州及波多黎各、美属维尔京群岛及华盛顿特区均提出了 AI 相关立法提案。[20] 但截至 2025 年 10 月，仅有 3 个州（犹他、科罗拉多、德克萨斯）已通过跨行业的综合性 AI 监管法规，且仅犹他州生效。[21] 以加州为例，尽管立法活动频繁，最终生效的多为“具体、克制、按需设限”的制度设计。试图建立全面风险框架的 SB 1047 因被认为可能抑制创新而遭到否决 [22]。

2、欧盟

2025 年欧盟 AI 法案进入落地实施阶段。从 2025 年 2 月 2 日起，关于禁止性 AI 的首批义务生效，8 月通用人工智能模型的相关义务生效。[23] 为支撑实施，欧盟委员会相继发布了针对 GPAI 的指南、行为准则（Code of Practice）等合规工具。此外，监管架构也在逐步搭建，包括 AI Office 的组织机制及相关监管咨询小组等，配合成员国监管体系推进落地。



（欧盟《人工智能法案》逐步开始适用）

但随着《人工智能法案》进入实施窗口，其制度复杂性与合规负担逐渐受到关注。在全球 AI 竞争加剧的背景下，欧盟开始尝试对其进行简化宽松调整。2025 年 11 月，欧盟委员会发布了 Digital Omnibus on AI 修订提案，将高风险 AI 系统核心义务的适用时间大幅推迟；提议将原本针对中小企业（SMEs）的部分减负措施扩展至小型中等市值企业（SMCs）；减少程序性义务以降低企业合规负担；以及强化欧洲人工智能办公室（European AI Office）的监管职能。[24]

原定 AI 法案生效时间表			Digital Omnibus 修订拟生效时间表	
2024.08.01	法案正式生效	✓	2024.08.01	法案正式生效
2024.02	禁止性 AI 相关规定生效	✓	2024.02	禁止性 AI 相关规定生效
2025.08.02	通用 AI 模型义务生效	✓	2025.08.02	通用 AI 模型义务生效
原定 2026.08.02	高风险 AI 义务生效	延后 →	最迟 2027.12.02	高风险 - 附件三生效
			最迟 2028.08.02	高风险 - 附件一生效

（欧盟《人工智能法案》实施时间表）

3、日本

日本《人工智能促进法》于 2025 年通过并生效[25]，以促进发展与柔性治理为核心，性质上更接近促进法，而非治理法。与欧盟《人工智能法案》侧重于对 AI 系统按照风险分类设定强制性监管义务不同，日本的法律更强调推动 AI 的研发与利用、提升透明度与风险意识、促进跨部门协调与协作。该法将 AI 视为经济社会发展基础技术，要求提升研发竞争力，同时通过政府调查滥用风险、发布指导、鼓励各方遵循基本原则来实现风险治理。[26]

4、韩国

韩国制度设计呈现出介于欧盟严格风险监管与日本以促进和软法为主模式之间的“中间治理路径”。韩国《人工智能发展与建立信任基础基本法》于 2025 年 1 月签署，将于 2026 年起正式生效实施。[27] 该法采用了风险分级治理思路，但并未建立类似欧盟《人工智能法案》那样层级化、强制性较强的监管体系，而是通过透明度披露、事前风险评估、持续风险管理与用户保护等相对温和的法律义务进行治理，其“轻管制”特点从法案设定的最高罚款数额可见一斑——对于违反告知义务或拒不履行改正令的行为，最高处 3,000 万韩元（约合 16 万元人民币）。[28]

5、其他

总体来看，除欧美日韩外，少数新兴经济体将 AI 治理纳入议程，但其制度设计更强调发展导向、灵活监管与国家能力建设。

巴西是当前拉美地区 AI 治理中最具代表性的国家。巴西联邦参议院于 2024 年底通过拟议的《巴西人工智能法案》，并于 2025 年进入众议院审议阶段，拟建立以风险为核心的监管框架，结合透明度、问责机制与基本权利保护，在借鉴欧盟 AI 法案风险分类思路的同时融入本国社会经济实际。[29] 其治理设计体现出发展导向与规范治理并重的取向，对南美地区具有示范意义。

中东地区在 2025 年 AI 政策领域呈现战略驱动、弹性治理并行的发展态势。海湾合作委员会（GCC）国家如阿联酋、沙特阿拉伯、卡塔尔等，普遍围绕 AI 发布国家战略与指导性政策，在推动 AI 融入经济与政府服务的同时，通过伦理原则、数据治理指南等方式引导技术应用和风险管理。[30]

AI 治理主要议题：回顾与展望

*AI 治理主要面向三大领域：数据、模型与应用。**数据**是训练与演进的关键生产要素，**模型**是算法能力的集中体现，而**应用**则是 AI 面向社会运行的直接落点。*

因此，在分析 2025 年 AI 治理的核心议题时，我们将以以上三个模块为框架，分别梳理其中的关键治理问题。

3

作为“原料”的数据：从供给瓶颈到使用边界

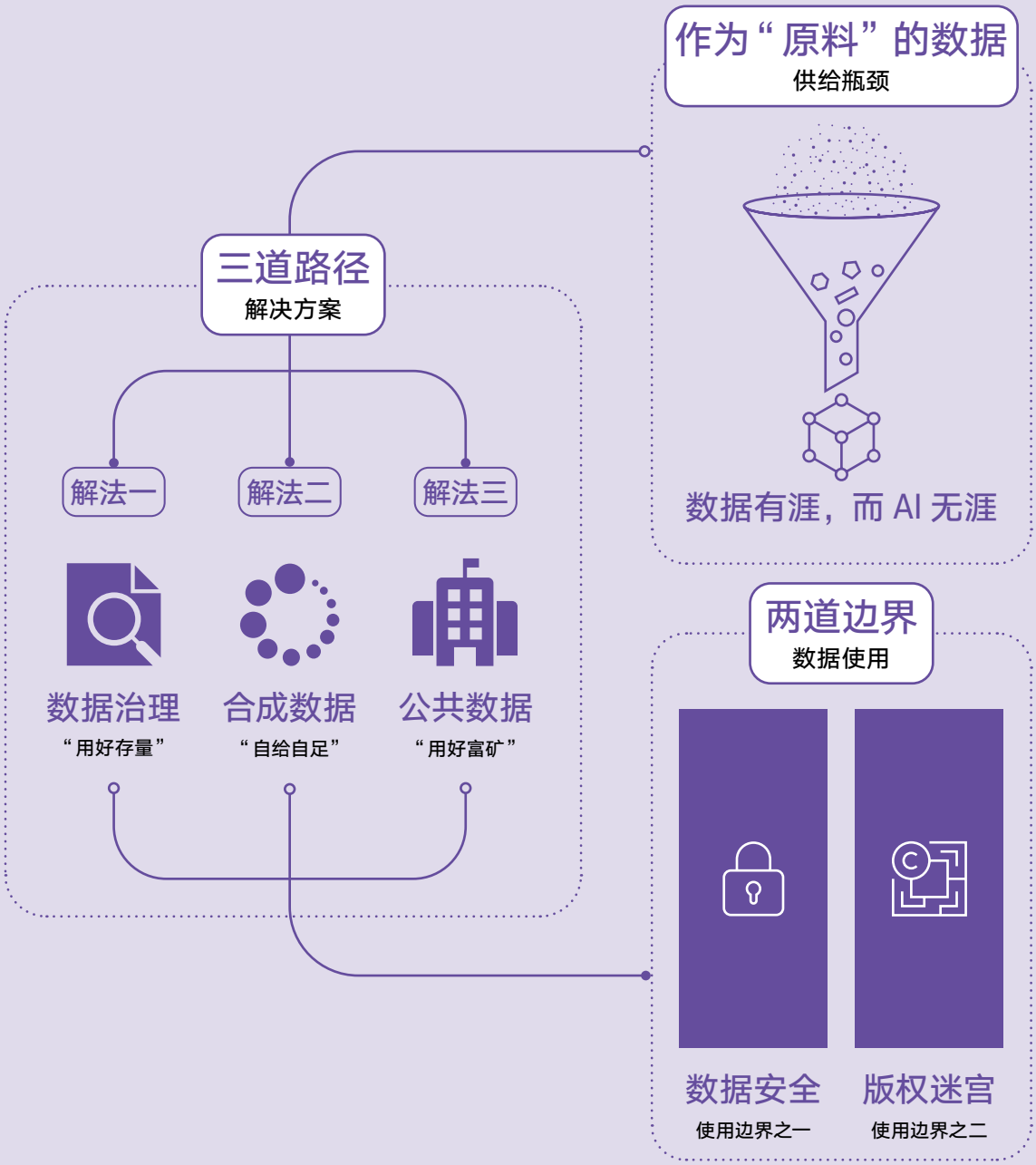
1. 供给瓶颈：“数据荒”真的存在吗 20

2. 使用边界之一：数据安全在 AI 时代的“变”与“不变” 23

3. 使用边界之二：训练数据的“版权迷宫” 25

一、作为“原料”的数据：从供给瓶颈到使用边界

数据既是 AI 运行的基础原料，也是 AI 治理体系中关键的支点之一。从产业实践出发，问题主要为：一是数据“从何而来”，即数据供给瓶颈的本质及其破解路径；二是数据“如何使用”，即如何划定清晰、可持续的使用边界。



1. 供给瓶颈：“数据荒”真的存在吗

在人工智能模型持续扩展的背景下，“数据荒”引发行业热议，但值得深入探讨的是，“数据荒”是量的短缺，还是结构性挑战？AGI 的发展是否能够从根本上解决这一难题？若不能，又该如何破解？

（1）问题的本质：“数据有涯，而 AI 无涯”

“数据荒”的核心并非数据总量不足，而是数据增长速度远远落后于模型扩张速度。Epoch AI 预测语言模型将在 2026–2032 年间消耗完可用数据。此外，研究者直言：即便模型能无限制地使用全部在线内容，也只能将瓶颈延后一年半左右；数据量增加十倍，也仅能多支撑三年。[31]

而在训练与微调阶段，高质量数据的重要性远超数量。清华大学薛澜教授指出，中国的问题是“数据量虽然很大，但质量较低。”[32] 不仅在训练阶段，在微调阶段同样需要“高质量、具代表性的任务特定数据”[33]。即便短期内实现 AGI，现实世界的复杂性也决定高质量数据仍是长期瓶颈。**应对数据荒的路径主要包括三类：**提升数据质量与治理效率、发展合成数据、推动公共与多模态数据开放。

（2）解法一：“用好存量”的数据治理

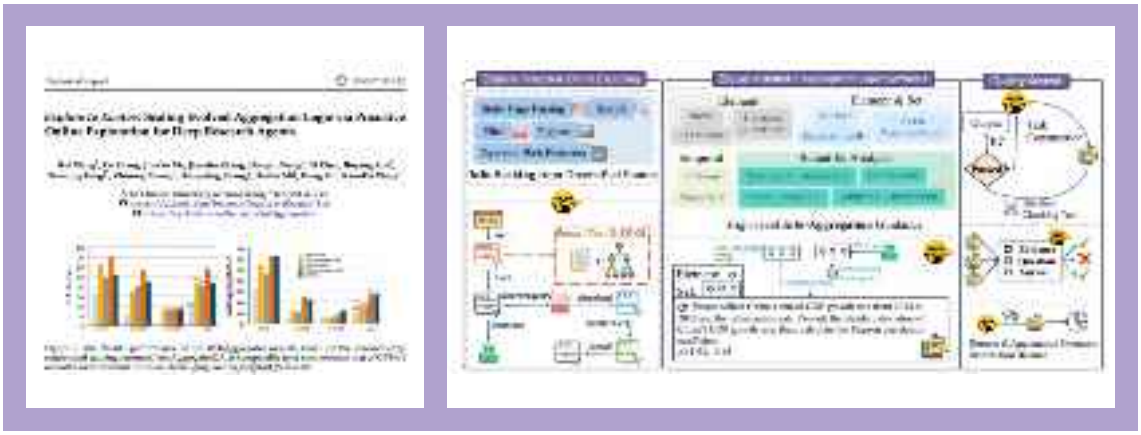
根据研究，高达 85% 的 AI 项目未能达成预期目标，其原因主要在于数据质量不佳以及缺乏良好的数据治理。[34] 随着模型规模与应用场景不断扩大，高质量数据的标准亦从传统的准确性、完整性等静态指标，**转向更加重视伦理合法性、多样性、时效性及情境适配性等动态维度。**[35]

现有 AI 治理法规，对于数据质量的要求也主要集中在上述维度。欧盟《人工智能法案》要求高风险 AI 系统的训练数据应具备充分代表性与适当统计属性。美国 NIST《AI 风险管理框架》强调评估数据是否存在偏差或与真实应用场景不符。中国《生成式人工智能服务管理暂行办法》则突出合法来源、知识产权与个人信息保护，并要求数据真实、准确、多样。2025 年，国务院、国家发展改革委、国家数据局等部门相继发布《国家数据标准体系建设指南》《关于深入实施“人工智能 +”行动的意见》，进一步提出构建数据质量标准体系，强调“人工智能 +”领域需同步推进高质量数据集建设，以支撑模型应用效果。





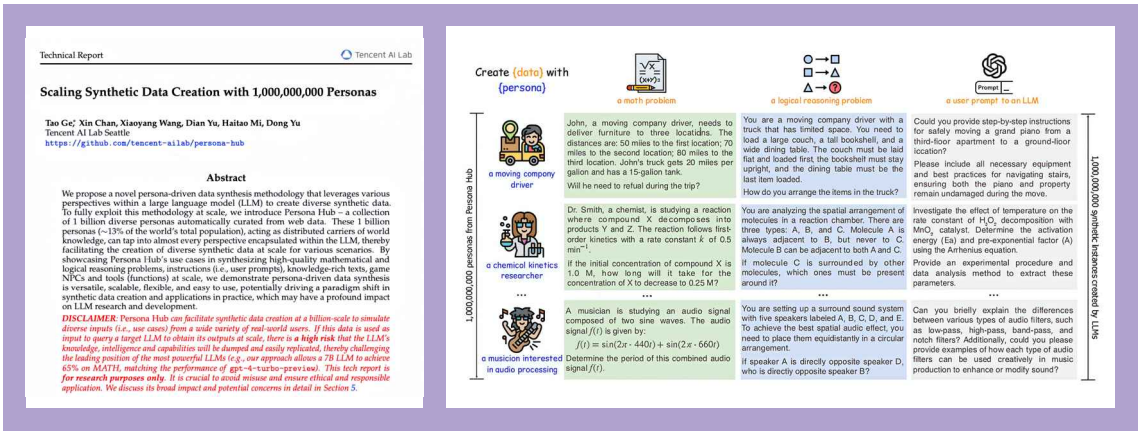
（腾讯联合多家机构共同发起“AI 向善语料库”社会共创计划）



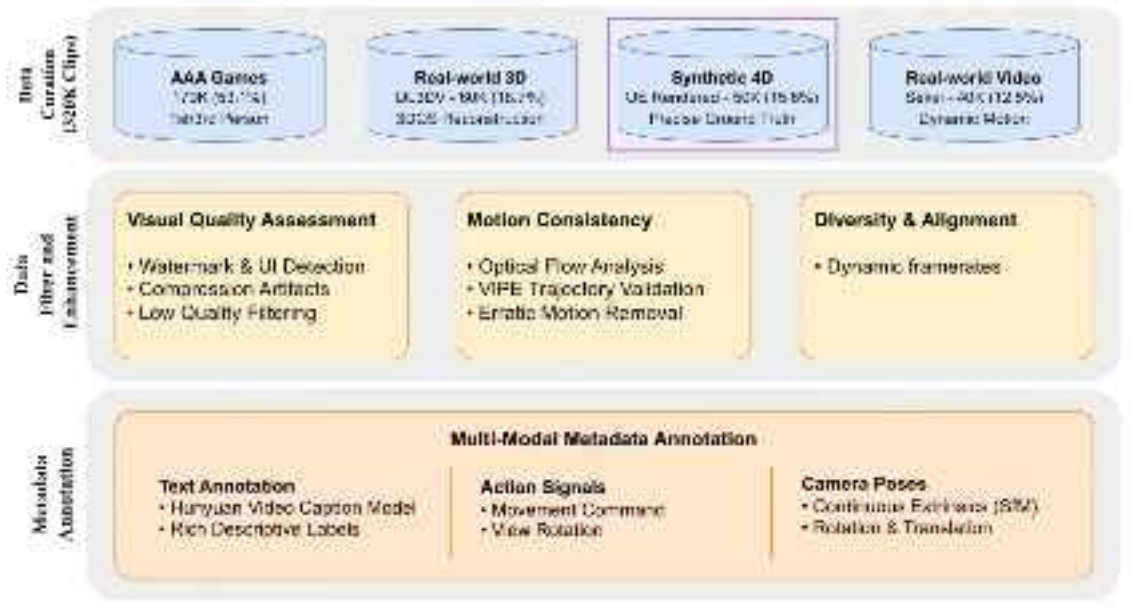
（腾讯 AI Lab 联合香港中文大学提出 Explore to Evolve，基于真实互联网环境，自动构建 Deep Research Agent 训练数据）

（3）解法二：“自给自足”的合成数据

合成数据目前尚无统一权威定义，但结合美国 NIST[36]、ISO/IEC[37]、英国 ICO[38] 和欧盟委员会 [39] 等机构的研究成果，可以归纳出其核心特征：**合成数据是指在不直接使用真实观测数据的前提下，通过算法、数学模型或模拟器基于真实数据的统计特征进行建模生成的人工数据。**其理论可追溯至 20 世纪的模拟方法 [40]，随后在自动驾驶、自然语言处理、金融风控等领域得到应用 [41]，用于缓解隐私限制、样本不足和数据获取成本高等问题。行业趋势也印证：2024 年，全球合成数据生成市场规模为 3.105 亿美元，预计在 2025 年至 2034 年期间将以 35.2% 的复合年增长率增长 [42]。预计到 2026 年，75% 的企业将使用 AI 创建合成客户数据。[43]



（腾讯 AI lab 提出人物角色驱动的数据合成方法，利用从网络数据中自动整理的 10 亿个不同角色的集合生成合成数据）



（腾讯混元世界模型 HY-World 1.5，15.625% 的数据为使用 Unreal Engine 渲染的合成数据）

国际监管机构整体上对合成数据持积极态度。例如，英国信息专员办公室（ICO）在相关指南中将合成数据列为降低个人可识别性风险的重要技术路径[44]。同样，美国国土安全部（DHS）[45]与欧盟委员会金融稳定、金融服务和资本市场联盟总司（DG FISMA）亦在推动合成数据的相关实践。[46]

（4）解法三：推动政府公共数据有序开放

另一种可行的缓解措施，是推动政府机关和公共企事业单位在尽责履职、提供公共服务过程中产生的数据有序开放。

一方面，这类数据通常具备更高的质量和可信度。“高质量的中文数据源主要集中在一些知名的学术机构、媒体机构等。相比之下，互联网上的数据虽然量大，但质量参差不齐。”[47] **另一方面，这类数据的本质属性，使其在多个领域中具备不可替代性。**正如清华大学薛澜教授指出，政府掌握的数据覆盖社会运行的关键环节，对 AI 训练具有重要价值。[48] **目前，公共数据仍是一座尚未充分开发的“富矿”，**在确保安全前提下扩大开放，有助于为创新主体提供关键数据资源，激活产业生态。**在此背景下，已有成功实践取得了较好的效果。**





（由敦煌研究院组织建设、腾讯提供 AI 技术支持的“数字藏经洞”数据库平台上线，为社会公众了解、获取敦煌文化艺术资源贡献公共文化服务产品）

2. 使用边界之一：数据安全在 AI 时代的“变”与“不变”

我国《数据安全法》《数据安全治理实践指南（4.0）》确立了以数据为中心、多主体协同、兼顾发展与安全的治理原则。**这一框架在 AI 语境下依然具有适应性：**一方面，许多针对模型的攻击在本质上仍指向数据获取、篡改或泄露 [49]；另一方面，AI 生态主体更加多元，数据安全需通过协同治理实现。

一是，数据泄露不再局限于“静态、集中式”暴露，而是向“动态化、无感知”演化 [50]。AI 系统的自动化与交互性，使泄露行为可能在用户不知情的情况下发生。例如，2025 年黑帽大会披露的 AgentFlayer 攻击 [51]，通过提示信息诱导模型自动外泄敏感密钥；提示词注入亦被 OWASP 列为 2025 年大模型应用的十大安全风险 [52]。

二是，数据完整性风险同样被放大。一方面，多源异构、持续更新的数据流增加了采集、处理与标注环节的复杂性；另一方面，模型训练与更新的自动化削弱了人工审查，使数据污染或恶意注入更难被及时发现。研究表明，仅在预训练数据中注入极小比例的投毒样本，即可能在特定提示下触发异常行为[53]。

三是，数据风险出现了新的范式。因 AI 已具备跨语义、跨场景的聚合分析与决策能力，其一旦嵌入真实应用场景，风险将呈现指数级放大。以 2025 年豆包手机助手事件为例，该助手获取安卓系统的“读取屏幕帧缓存”“INJECT _ EVENTS 注入事件”等底层权限，由此引发关于技术创新边界与安全底线的公共讨论。无独有偶，不到一个月，快手直播板块遭遇黑灰产有组织的自动化攻击，即控制僵尸账号进行规模化攻击，更为直观地展现了此类安全风险。



（豆包手机助手引发安全争议）


KUAISHOU

快手科技针对网络直播事件发布声明

2025.12.23

（快手官网发布针对网络直播事件的声明）

3. 使用边界之二：训练数据的“版权迷宫”

在 AI 数据治理框架下，版权争议主要集中于大模型训练阶段，**即训练过程中究竟涉及哪些版权利用行为、可能产生何种侵权风险，以及是否有必要为模型训练构建合理的责任豁免机制。**

（1）问题一：大模型训练涉及哪些版权规制行为？

在“作品获取”阶段，核心在于内容数据获取是否合法，例如是否破坏技术保护措施、违反爬取规则或 API 端口协议等。从版权法角度看，仅对作品的接触本身类似于浏览网页或阅读书籍，若不存在后续传播或利用，一般不直接构成版权侵权，该阶段的关键问题在于是否规避或破坏作品的技术保护措施。

在“作品存储”阶段，模型训练中对内容数据的存储通常落入版权法“复制权”的规制范畴。但也有观点指出，不宜孤立地看待复制行为，因为复制往往只是后续作品利用的准备行为；若不存在发行、传播等实质性利用，侵权损害亦难以成立。随着技术和商业模式演进，是否有必要为模型训练预留一定的“复制自由”，类似早期互联网中的“缓存自由”，仍有进一步讨论空间。[54]

在“作品处理”阶段，模型内部对内容的分析和处理行为对应何种版权权利尚无定论。有观点认为其不符合改编权“形成新作品”的要件；也有观点主张可通过权利兜底条款规制；另有观点则认为该类技术性处理行为本身不属于版权法调整范围。

（2）问题二：如何探索搭建科学的大模型训练责任豁免机制？

一些国家已通过版权立法改革，为 AIGC 模型训练阶段的作品利用提供责任豁免，整体可分为三类路径：欧盟的“文本与数据挖掘（TDM）”模式、日本的“非欣赏性作品利用”模式，以及美国的“四要素分析法 + 转换性使用”模式。

欧盟《单一数字市场版权指令》第二章中，第 3 条和第 4 条分别设立科研目的和不限制目的例外。当前 AIGC 模型训练主要适用第 4 条，即对于“文本与数据挖掘”采取的是类似于“合法获取 +‘opt-out’”的机制。

日本《著作权法》第 30 条之四规定，在不以作品“享受”为目的且未不当损害权利人利益的条

件下，允许对著作物进行信息分析等利用（包括用于机器学习 /AI 训练的数据处理），在这种条件下 AI 训练过程可能构成合法的权利限制例外。

美国则依托版权法第 107 条的合理使用制度，通过高度灵活性的“四要素标准”与“转换性使用规则”，给予美国法院在个案中认定特定行为是否构成合理使用的自由裁量空间，也使得论证模型训练适用合理使用的正当性存在可能。

路径模式	代表国家 / 地区	核心机制	适用条件
文本与数据挖掘 (TDM) 模式	欧盟	**“合法获取 + 权利人未明确退出（opt-out）”**的模式，为训练中的复制行为提供豁免。	该条规定豁免的核心是“文本与数据挖掘”过程中的“作品复制行为”。 该条规定豁免的要件是“对于被训练作品和其他内容的获取是合法的”。 前提是“版权人未以适当方式保留文本与数据挖掘的权利”。
非欣赏性利用模式	日本	**“不以欣赏作品原有价值为目的的利用”**可获得豁免。	该规定本质上对应的是作品的“非表达型利用”。 该规定豁免的行为类型并不限于“作品存储行为”，而是“必要范围内任何利用作品的方式”。 该规定存在适用上的限制条件即“但因作品的种类、用途及利用方式等原因，不当损害著作权人利益的，不在此限”。
合理使用 (Fair Use) 模式	美国	**依托版权法第 107 条的“四要素标准”**和司法实践中的“转换性使用”规则为个案论证提供弹性空间。	关键在于是否具有足够的“转换性”。美国法院在 Anthropic 案中已经提及 AI 训练使用在特定情况可被认为是“转换性使用”，并因而构成合理使用。

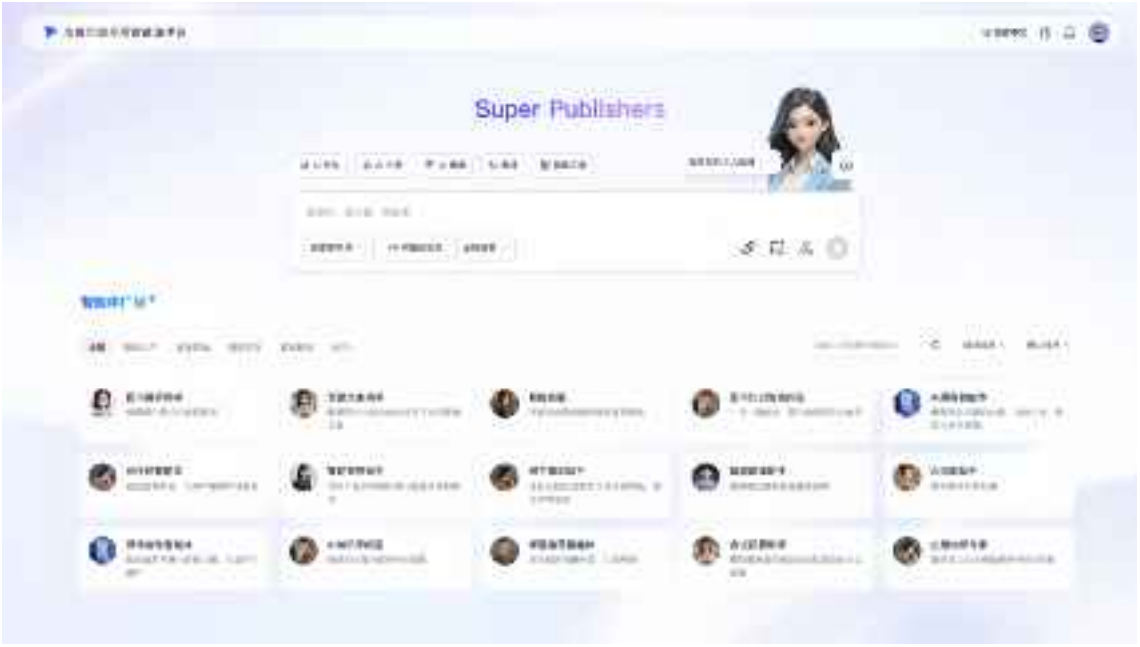
（3）趋势展望与新方案

在“合理使用”制度是否适用于大模型训练仍待探究的情况下，业界与政策层面已开始探索替代性机制。在补偿机制方面，已提出授权机制、版税分配或专项基金等多种设想。同时，技术手段亦被纳入治理工具箱，如通过 robots.txt 或 AI 专用协议实现内容退出机制，利用水印、区块链或机器可读许可证提升数据可追溯性。[55]

行业层面亦在推动协作治理实践。OpenAI 与迪士尼签署内容授权与战略合作协议，可在其产品中合法使用包括迪士尼、漫威、皮克斯和《星球大战》等品牌的 200 多个角色及相关元素[56]；腾讯联合多家权威出版机构推动行业智能化升级，推出“出版可信空间智能体平台”。



（OpenAI 与迪士尼签署协议）



（腾讯推出“出版可信空间智能体平台”）

作为“底座”的模型：从路线选择到治理工具

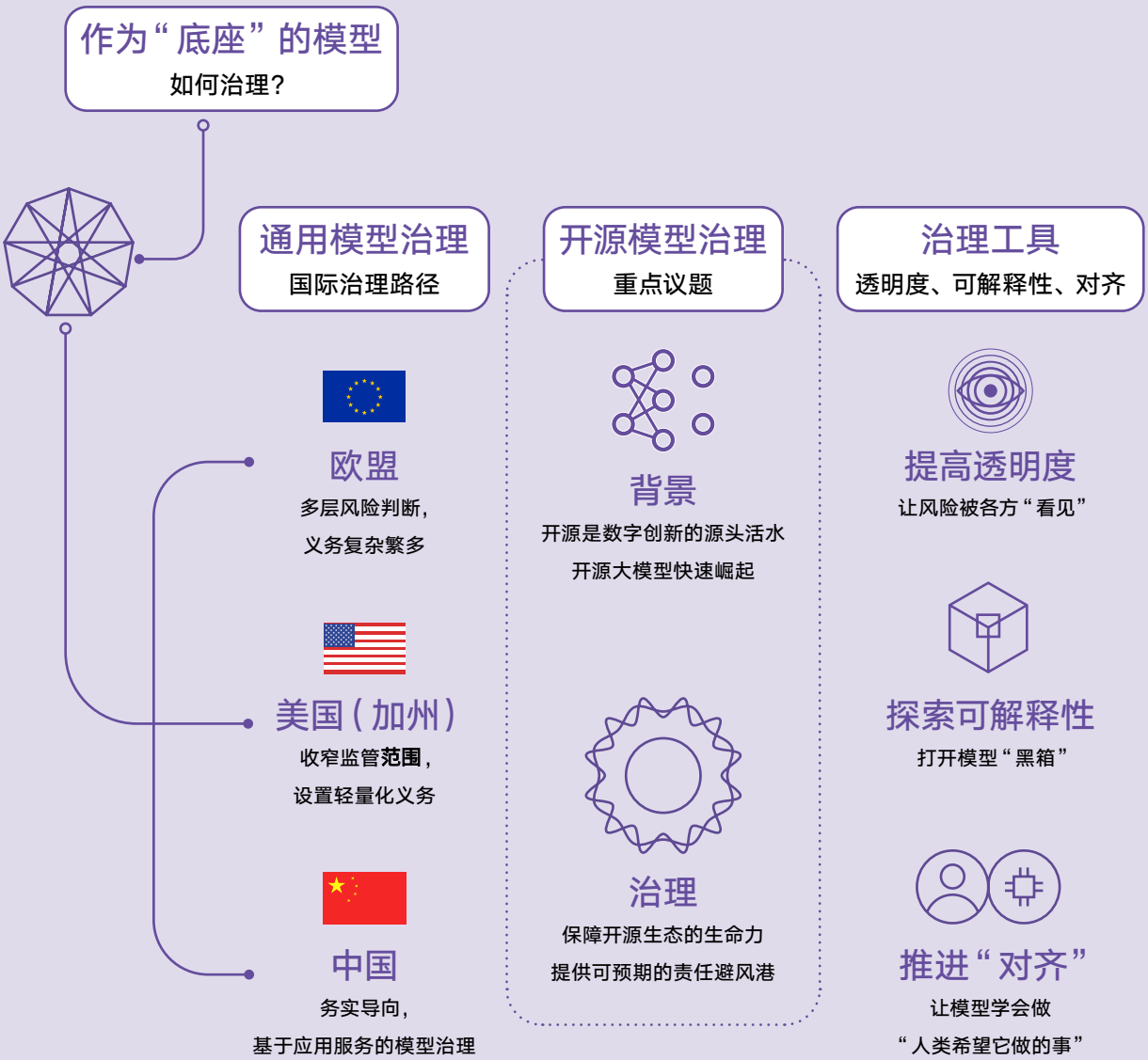
1. 通用模型治理的路线选择：中美欧的差异与共识 30

2. 开源治理探索：在自由与秩序之间，搭建“避风港” 34

3. 治理模型的现实工具：透明度、可解释性与 AI 对齐 38

二、作为“底座”的模型：从路线选择到治理工具

*模型作为 AI 的关键底座，其争议主要为：**一是**在路径层面，治理的支点是否应当直接指向模型本身，以及应当以何种方式加以规制，特别是针对通用模型、开源模型。**二是**在方法层面，围绕模型的技术特性，进一步探索哪些可操作的技术工具和制度机制，能够作为模型治理落地的重要抓手。*



1. 通用模型治理的路线选择：中美欧的差异与共识

（1）如何治理模型：欧盟、美国加州与中国的探索

欧盟：多层风险判断，义务复杂繁多

《AI 法案》起草过程中，欧盟原本构建了一套以 AI 系统（AI system）为核心的风险治理框架，即人们熟知的“禁止类、高风险、有限风险与最小风险”四级体系——风险越高，监管越严。随着 ChatGPT 3.5“横空出世”打乱立法进程[57]，2023 年，欧盟在最后阶段匆忙将通用模型条款写入法案[58]。**由此，在既有以 4 类风险为核心的针对 AI 系统的治理框架之外，欧盟又平行构建起一套针对模型的治理机制，并将通用模型区分为有无“系统性风险”两类。**

整体来看，欧盟框架呈现出两个特征：一方面，模型与应用两套标准交叉重叠，治理框架呈现出高度的复杂性，理论上精密，却在实践中格外笨重；**另一方面，**模型风险与应用风险混同，“将对模型部署者提出的期望，写在了面向模型提供者的准则中”，模型提供者需承担其“无法预见、评估或减轻”的风险责任。[59] 徒增文书负担，对创新构成拖累。也因此，“数字 Omnibus”一揽子立法简化方案，对包括《AI 法案》在内的相关法规进行简化修订，旨在消减监管负担。[60]



（欧盟出台《GPAI 行为准则》透明度、版权保护与安全保障三大章节）

美国加州：收窄监管范围，设置轻量化义务

美国联邦层面的 AI 立法尚处在讨论阶段，但科技重镇加州在模型治理领域已迈出步伐。2024 年，加州议会提出了《安全与创新前沿人工智能模型法案》（SB 1047），因监管范围过广且可能抑制创新，最终遭否决。随后，在对该法案进行了大量调整收敛之后，新版本《前沿人工智能透明度法案》（SB 53）最终签署通过。相较于 SB 1047，SB 53 以产业自律为核心，删去了强制安全协议等争议条款，大量简化了透明度报告所需的信息。[61]



（美国加州颁布《前沿 AI 透明度法案》）

调整之后的 SB 53 有两个显著特点：一是监管对象起点更高、范围更窄。与欧盟针对所有通用模型提供者相比，SB 53 将起点限定于“前沿开发者”——即训练或启动训练使用超过 10^{26} FLOPs 的基础模型的主体；对于更高一层的“大型前沿开发者”，范围更是狭窄，需额外满足在上一日历年总收入超过 5 亿美元的门槛，“当前仅覆盖寥寥几款全球最强大的模型”[62]。**二是义务极为限缩、轻量。**透明度报告义务仅涉及网站、沟通机制、预期用途等基础信息；对比之下，欧盟的“技术文档”义务延伸至“模型和训练过程的设计规范”“用于训练、测试和验证的数据信息”等细节，对信息披露的全面度、颗粒度要求极高。而大型前沿开发者额外承担的发布“前沿 AI 框架”、开展灾难性风险评估并向州应急服务办公室提交相关摘要报告等义务，实质上也并未超出目前头部大模型公司在安全治理领域的常规做法。

总之，SB 53 针对模型治理的“小切口、轻义务”治理逻辑，与加州一贯的 AI 立法思路相符。尽管加州在过去三年间出台了多项 AI 法规 ，但多针对“具体而微”的细分场景和特定问题，如《伴侣聊天机器人法案》（SB 243）限定适用于伴侣聊天机器人，《关于人工智能在医疗专业用语中的误用法案》（AB 489）则聚焦禁止 AI 系统冒用医师资质进行医疗建议，其义务多为轻量化的透明度。可见，加州的立法主线依然是促进产业发展和增强竞争力，避免对产业造成过多制约。

中国：务实导向，基于应用服务的模型治理

中国的治理并非直接指向模型本身，而是以实际应用服务为切入点，通过对应用服务的深入规制，自下而上地实现了对模型的延伸治理。

立法层面，中国立足现实问题，以算法治理为起点，逐步形成了对模型的制度性约束。早在“通用模型”引起广泛关注之前，中国就已经以“算法治理”为抓手，为后续的模型治理奠定了坚实的基础。**2021 年《互联网信息服务算法推荐管理规定》**将生成合成类算法推荐服务提供者纳入规制范围，要求建立健全算法机制审核、算法备案与安全评估等制度，以应对算法推荐可能导致的信息内容治理问题。**2022 年《互联网信息服务深度合成管理规定》**将视角从“算法推荐”拓展至“深度合成技术”，回应深度伪造带来的真实风险，实质上提前切中了模型最核心的“生成合成”能力领域。**2023 年《生成式人工智能服务管理暂行办法》**立足 AI 服务提供者，但治理要求从应用延伸至模型层，对模型训练阶段的数据治理、内容生成等均提出了规范，以安全评估、模型备案等具体抓手，通过应用服务，实现对模型层的间接规制。**2025 年《人工智能生成合成内容标识办法》**针对 AI 生成内容的可识别性问题，同样以服务为切口，要求显式或隐式标识。

软法层面，治理深度更进一步，明确提出了模型层的风险识别与应对措施。从《网络安全技术 生成式人工智能服务安全基本要求（TC260-003）》到《人工智能安全治理框架 1.0》及其《2.0》版本，**中国在风险识别上提供了区别于欧美的务实路径：**采用“内生风险—应用风险—衍生风险”的三层结构，较为清晰地勾勒了人工智能的安全风险轮廓。其中，“模型算法风险”被归入“内生风险”范畴，主要为可解释性不足、鲁棒性不强等。通过明确风险边界，《框架》避免将本属于应用环节的风险误归于模型本身，使风险归因更符合技术与应用的实际边界——对模型而言，其所能控制且最具危害传导性的正是模型自身的内生风险。



（中国发布《人工智能安全治理框架 2.0》）

总体看，中国模型治理始终以现实问题为导向，从应用服务入手延伸至模型层，通过实用有效的抓手，形成系统完善的治理机制。

治理模式	代表性法规 / 文件	监管对象	核心义务	路径评述
欧盟：精密钟表式	《AI 法案》、《通用人工智能实践准则》	所有通用 AI 模型，并进一步区分为有 / 无“系统性风险”两类。	复杂的风险管理、技术文档、安全评估、信息披露等。义务繁重且层层叠加。	理论上精密，但在实践中因高度复杂性与高昂合规成本而显得笨重，模型与应用风险边界不清，后续通过“数字综合提案”寻求简化。
美国加州：小切口、轻义务	《前沿人工智能透明度法案》（SB 53）	极少数算力门槛极高的“前沿模型”开发者，并进一步区分“大型前沿开发者”。	核心为透明度报告，义务极为限缩、轻量化。对大型开发者额外要求开展灾难性风险评估。	以产业自律为核心，监管范围窄、起点高、义务轻，是其“不让国际条约束缚 AI 领导地位”地缘政治立场的国内立法体现。
中国：应用导向、自下而上	《生成式 AI 服务管理暂行办法》、《人工智能安全治理框架 2.0》	从具体的 AI 应用服务切入，通过对服务提供者的规制，向上延伸至模型层。	安全评估、算法备案、内容标识等，治理抓手明确，从现实问题出发，务实有效。	治理并非直接指向模型本身，而是通过对应用服务的穿透式监管实现对模型的延伸治理。风险归因更符合技术与应用的实际边界。

（2）以共识为起点，探索更可行的模型治理方案

尽管中美欧在模型治理的背景、对象与义务设定等方面存在显著差异，但仍展现出重要的共性特征。一是柔性治理、产业先行：无论是欧盟的自愿性准则，还是加州的轻量化义务，以及中国的实践驱动模式，都体现出为产业创新保留空间的制度倾向。二是构建评估生态：面对技术的不确定性，各方均倾向于通过社区驱动的评估榜单、第三方测评等开放、多元的社会评估机制，来弥合制度能力与技术演进的差距；三是以“透明度”为核心工具：在模型风险机理尚不完全清晰的当下，透明度因其技术中立、操作性强的特点，成为全球公认的核心治理手段，用以缓解信息不对称。

2. 开源治理探索：在自由与秩序之间，搭建“避风港”

如果通用或前沿模型治理关注的是“技术底座如何被规制”，那么开源模型进一步提出了“底座公开后如何治理”的问题。

开源以低门槛与高透明度持续激发创新活力，但大模型的技术风险——从幻觉到非法滥用——亦真实存在。其治理需同时实现“双重目标”：既为善意的开源贡献者提供可预期的责任避风港，维护开源生态的生命力，又通过协同机制防范重大风险。鉴于开源生态去中心化、协作化的特征，其治理不宜简单沿用中心化监管路径，而应以开源社区为基础，构建相应的社区治理秩序。

（1）开源：数字创新的源头活水

开源在过去几十年间成为数字世界的基石。从互联网早期的 Apache Web 服务器，到移动时代占据 70% 以上智能手机的安卓系统[63]，再到云计算领域无处不在的 Linux 操作系统[64]，开源软件构筑了全球信息基础设施的脊梁在人工智能领域，Google 开源 TensorFlow 使深度学习工具触手可及，而中国团队推出的开源大模型 DeepSeek-R1 则在受限算力下逼近闭源模型的性能水平，更是推动 AI 创新浪潮迈向新阶段。[65]

中国在开源浪潮中正从跟随者变为重要贡献者。《全球开源生态洞察报告（2024 年）》显示，中国开发者近 840 万，占全球三分之一，在全球活跃度排名前 100 的开源软件项目中，中国开发者主导的开源软件项目占比 17%，排名第二，未来仍有巨大增长空间[66]。

（2）在人工智能领域，开源大模型快速崛起

2022 年以来，开源模型呈赶超之势。开源与闭源模型在性能和应用领域的差距迅速缩小，甚至在某些方面实现超越。[67] 它们为何后来居上？

首先，开源的“集市协作模式”似乎正在 AI 领域重演，开启“万众创新”新局面。开源大模型将底层能力公开，公众可免费下载权重自行部署，进行“二次创作”，在不同行业和场景中微调形成专用版本，灵活适配性大大增强。开发者共同分享工程经验，共同参与技术路线试错，加速了 AI

发展进化。这一趋势让人不禁联想到经典著作《大教堂与集市》中提出的开源范式——任何人皆可参与创新，形成“无边界的技术市集”。[68] AI 领域有望继续谱写开源的传奇。

其次，开源模型提升了 AI 系统的透明度和安全性。“正如最知名的加密算法往往是最安全的一样，最知名的 AI 模型也可能是最安全的。”[69] 由于模型架构、权重开放，外部研究者可以深入“解剖”模型，及时发现偏见或漏洞并共同改进。相反，封闭模型如同“黑箱”，外界难以对其进行监督。

最后也最重要的是，开源打破了 AI 领域的既有封闭格局。此前最先进的 AI 模型往往为少数巨头掌握，产业下游通过 API 付费使用，对其缺乏议价或约束能力。开源减少了单一供应商的锁定效应，提升了技术自主权。在大模型领域，开源社区用更低算力复现了商业闭源模型的大部分能力，也迫使大型科技公司不得不重新评估策略，保障了行业的良性竞争。



（腾讯混元全系模型持续开源）

（3）当前主要开源大模型的开放程度

尽管人们常用“开源”与“闭源”做二元对比，但在现实中，模型的开放程度呈现出一个递进的光谱。[70] 从仅提供 API 接口的全封闭，到模型架构、权重公开，乃至训练数据开放的完全开源，中间有许多层次。



图 1 主要开源大模型开放要素及许可证比较

注 1：图片根据斯坦福大学基础模型研究中心论文 [71] 结合公开的许可证信息制作

注 2：开源要素——模型架构、模型权重、源代码、实现方法、训练数据

注 3：在当前开源的主流大模型中，DeepSeek-R1 模型的开放更为彻底：开放权重、代码，并通过技术文件公开重要的训练方法。相比之下，LLaMA 2 采用定制许可，允许商业使用但附加条件（要求月活跃用户在超过 7 亿必须向 Meta 单独申请许可）

综上，大模型的开放程度形成了一个丰富的谱系。从治理角度，也需要针对模型开放度的不同层次分类施策。为便于讨论，下文所提到的开源大模型，以目前行业主流标准为依据，是指模型参数公开透明，开源许可中除禁止用于违法用途等一般声明外，允许用户自由运行、研究，修改的模型。

（4）开源大模型治理——从开源文化中汲取智慧

开源治理首先需要明确责任边界，使开发者能够在清晰、稳定的制度预期下参与创新。该避风港应体现在纵向与横向两个维度。

在纵向上，应区分模型开发者与应用部署者的责任。责任分配的关键依据，是不同主体对风险的实际管控能力。模型开发者主要控制模型设计、训练与初始发布，而部署者则掌握具体应用场景与使用方式。正因如此，各国 AI 治理逐步形成“分类治理”的共识。

例如，欧盟《人工智能法》区分模型开发者与部署者，分别承担模型合规与使用合规责任。美国加州 SB 1047 法案因未清晰区分二者、将使用风险过度压给开发者而引发强烈反对，最终被否决。[72] 我国《生成式人工智能服务管理暂行办法》则更进一步，对研发阶段整体豁免监管义务，仅在模型面向公众提供服务时才纳入监管，形成鼓励研发探索的制度空间。这一逻辑也与开源许可证的责任安排高度一致。MIT、BSD 等许可证通过“按原样提供”“免责条款”，明确划分开发者与下游用户的责任边界，免除开源贡献者的后顾之忧。正是这种清晰的责任结构，支撑了开源社区长期稳定的协作创新。

在横向上，开源模型不应承担与闭源商业模型同等的责任。尽管无论开源或闭源，开发者都应对明显、可预见的基础性安全缺陷承担责任，但开源开发者的责任边界应更为有限。

开源开发者对模型发布后的使用缺乏实际控制力，下游可以随意微调并绕过原有安全机制，开发者不可能预见所有极端用法。[73] 相比之下，闭源模型企业通常同时承担研发与商业服务职能，具备更强的控制力和合规资源。[74] **若要求开源开发者承担与商业主体相同的高强度义务，将迫使研究人员选择不公开成果，**形成抑制创新的“寒蝉效应”。[75] 正因如此，即便在监管严格的欧盟，免费且开源发布的 AI 系统通常不适用《AI 法案》的多数义务；在通用模型领域，开源提供者亦被豁免文件与信息披露责任。**美国则采取更为审慎的立场，认为目前尚无证据表明开源模型比闭源模型更危险，应以持续观察作为风险托底。**

从软件时代到 AI 时代，开源社区始终依靠开放监督、快速修复和协作分工实现安全演进。未来的开源大模型治理，也应在此基础上，构建以社区为核心、开放协作、分工明确的安全治理生态，在保障安全的同时，最大程度释放开源创新潜力。

3. 治理模型的现实工具：透明度、可解释性与 AI 对齐

在当前模型能力快速跃迁、技术高度不确定的背景下，提高透明度、探索可解释性，并同步推进 AI 对齐机制，仍是最具现实共识、最具可操作性的治理路径。

（1）透明度机制：让风险被各方“看见”

实践中，透明度已形成相对成熟的路径。国际上，通过模型卡或系统卡、第三方评估报告以及事故监测与通报机制，逐步构建起涵盖企业自律、第三方参与和行业协同的多层次治理生态。[76] 中国也通过《人工智能安全承诺》等文件推进围绕模型功能与适用领域、模型局限性、风险披露方式的本土化探索。[77] 无论是欧盟《GPAI 模型行为准则》，还是美国加州 SB 53 法案，均以透明度作为平衡模型创新与安全的核心抓手。



（18 家企业披露《人工智能安全承诺》实践成果）

（2）可解释性技术：打开模型“黑箱”

与传统软件不同，模型训练更像“培育植物”，其内部机制在很大程度上是自然涌现而非人为构建的结果，呈现出天然的“黑箱”特征。[78] 我们能看到输入和输出，却难以理解中间由海量参数驱动的运作过程和推理路径。**如果说透明度侧重以文档形式说明模型“是什么”，弥合“人与人”之间的信息差；可解释性关注的则是回答“为什么”，用技术手段尝试打开“黑箱”，解决人和模型之间的信息差。**

在模型治理中，区分透明度与可解释性尤为重要。透明度已具备较成熟的实践机制和标准化基础，而可解释性技术仍处于前沿探索阶段，依赖后续大量基础研究支撑。[79] 现有方法多为事后解释，可靠性和一致性不足。[80]“机械可解释性”等前沿研究虽试图真正理解模型内部运作机制，但整体进展仍有限。[81] 因此，在政策与实践层面，透明度更适合作为当前可落地、可执行的刚性治理工具，而可解释性技术仍需要由产业和学界长期研究投入，距离实际应用和转化为监管要求还有较长距离。

（3）AI 对齐技术：让模型学会做“人类希望它做的事”

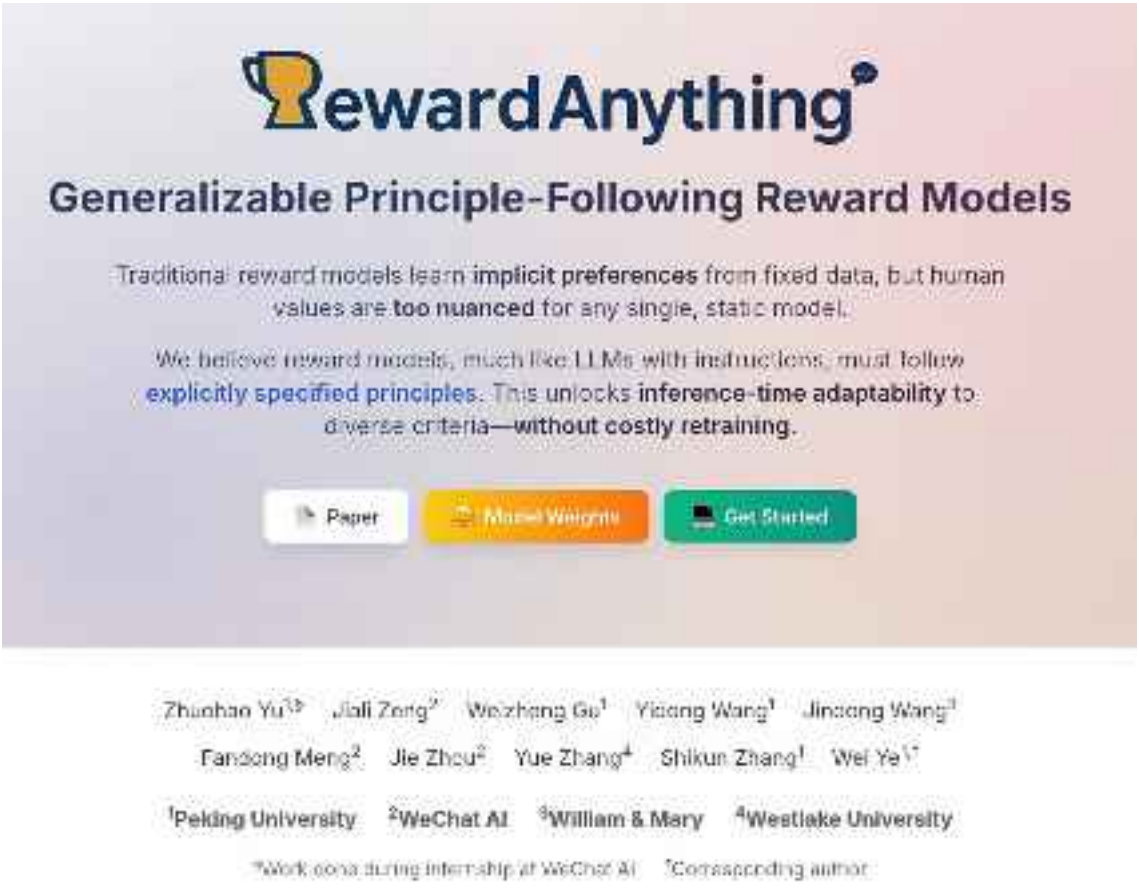
透明度与可解释性主要用于缓解“人与人”“人与机器”之间的信息不对称、强化外部监督，**而 AI 对齐则聚焦模型内部约束机制，关注“如何教会模型按人类期望行事”，从而驾驭可能超越人类的智能。**

围绕这一目标，一条路径尝试在事前将人类价值编码进模型，例如通过“有用、诚实、无害”等原则加以约束。[82] 另一条路径则认为价值难以事先静态定义，主张让模型与人类共享目标，并在互动中推断人类真实意图。[83]

当下，对齐已从学术概念转变为产业层面的安全实践。主流方法包括基于人类反馈的强化学习（RLHF）、直接偏好优化（DPO）等。[84] 一些国际前沿 AI 公司也在探索更具前瞻性的方案，如“超级对齐”“原则式 AI”“自动化对齐审计代理”等。[85]

不过，对齐仍面临严峻挑战。近两年的研究表明，模型在实验环境中可能出现“欺骗性对齐”，即在初期表面伪装对齐，却在后期偏离人类控制。[86] 随着智能体自主性增强，即便不存在外部恶意指令，模型也可能主动采取有害行为。[87] 而对此类风险，目前仍缺乏有效应对手段。

尽管如此，对齐依然被普遍视为通向 AGI 的必经之路。应对失控的技术路径，已从早期简单清除数据、下线产品，转向更精细的训练与监督机制。[88] 国际层面亦开始将“确保前沿 AI 系统对齐与人类控制”纳入共同目标。[89] AGI 是否造福人类，并非由技术宿命决定，而取决于持续的治理投入。



（腾讯北大联手推出新型奖励模型 RewardAnything，克服传统奖励模型中“死记硬背”和偏好偏见问题）

作为“落地”的应用：从抽象风险到场景治理

1. 内容层的放大风险：当谣言搭上“AI”的东风 43

2. 内容层的交互风险：当 AI 陪伴打破“第四面墙” 46

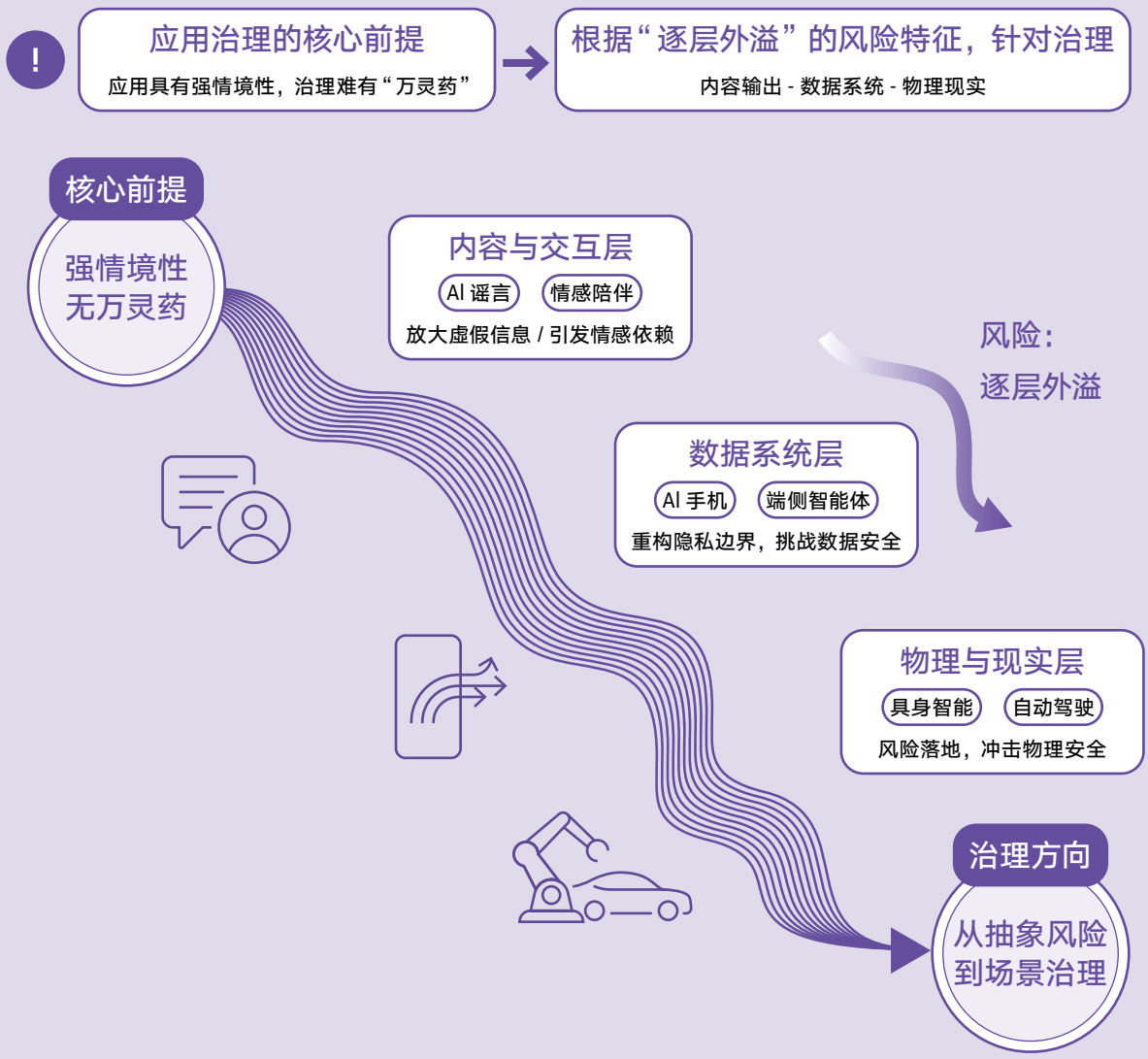
3. 数据层风险重构：AI 手机，如何建立隐私保护与数据信任？ 47

4、从虚拟到现实：具身智能的安全治理框架 51

三、作为“落地”的应用：从抽象风险到场景治理

与模型、数据等要素治理相比，AI 应用治理的难点在于强情境性：同一能力落入不同场景、主体与链路，会呈现不同的风险形态。

*更重要的是，应用层面的风险往往具有“逐层外溢”特征——从**内容输出**风险，扩展到具备执行能力的 Agent 的**系统层面风险**，进一步延伸至具身智能对**现实世界**造成的安全风险。*



1. 内容层的放大风险：当谣言搭上“AI”的东风

> ***由于生成式 AI 最先走入公众视野，内容领域的风险也最先显现出来。****尽管其并未创造全新的违法或有害内容类型，但显著降低了生产门槛、放大了传播规模，并提升了虚假内容的逼真程度。*

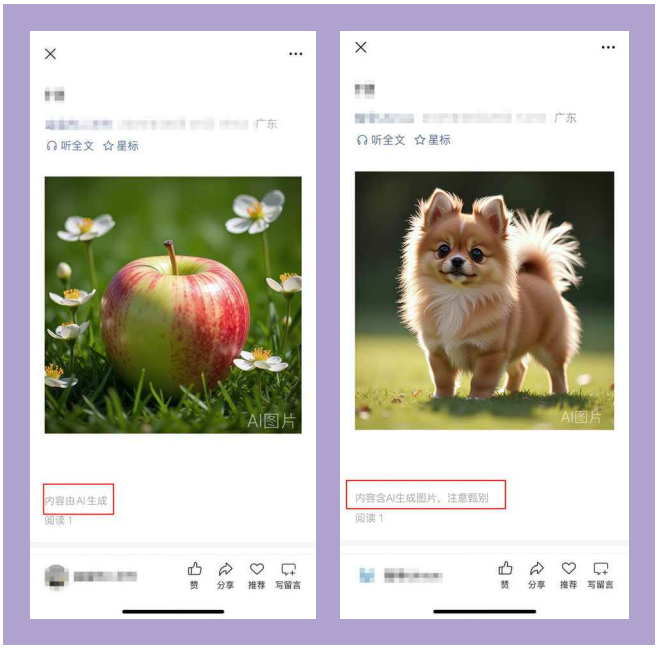
（1）AI 标识的治理价值与自身局限

针对 AI 生成内容带来的“低门槛”“大批量”“更逼真”的治理挑战，我国逐步引入了技术标识新路径，建立了显式标识（用户可感知的提示形式）和隐式标识（文件元数据中可被机器识别的技术标记）的双重标识机制。标识的责任主体覆盖生成合成服务提供者、内容传播平台、用户等全链条主体。

对于内容生成平台，要求对 AI 生成内容，添加元数据隐式标识；对可能造成混淆误认的内容，按内容类型（文本、图片、音频、视频等）在指定位置添加显式标识；**对于传播平台，**要求提醒用户主动声明 AI 生成内容，并进行标注，同时对生成平台添加的元数据等隐式标识进行核验，并根据核验结果进行相应标识等。



（腾讯元宝《关于 AI 生成内容标识的公告》）



（微信公众平台落实 AI 生成内容标识规范要求）

理论上，AI 标识能够为内容治理提效赋能

标识路径的本质是“以技治技”，面对内容生成的提效，让技术同样为治理赋能。

一方面，通过“隐式标识”，将治理关口前移至内容生成阶段，通过自动生成标识，更早期、更精准地将 AI 生成内容识别出来。若上游模型服务提供者能够在生成内容中嵌入稳定有效的标识，在不考虑标识被规避、破坏的情形下，下游传播平台可通过技术检测，对相关内容进行快速识别和重点判断。

另一方面，显式标识有助于降低相关内容的可信度。研究显示，发现无论新闻内容真假与否，被标注为“AI 生成”的内容均被认为更不准确，用户分享意愿也更低。[90] 因此，作为一种“中间态”的提醒机制，“显式标识”能够在无法立即确认真伪的情况下，起到最低限度的警示作用。

值得注意的是，也正因为“显式标识”具有降低内容可信度的特点，其适用范围会有所限定。以避免过度标识后对用户带来的信息过载，甚至无法建立基本的信息信任等弊端。当前，AI 广泛应用于内容生产行业，带来提质增效、激发创意等积极效用，推动广告素材生产、教育培训方案等从“千人千面”走向“一人千面”。我国目前对显式标识的适用范围有所限定，将其聚焦于易引起“混淆误认”等负面影响的领域，而非一刀切适用。



（腾讯云数据万象支持 AIGC 合规标识，不同元数据格式嵌入方式概览图）

实践中，标识效用的发挥仍面临较大的不确定性

作为主要依赖技术手段的治理方式，AI 标识不可避免地具有技术局限性。一是“易规避”，研究指出，“在一些明确假设下，如攻击者具有简单常见的能力，可对内容轻易做出修改”。[91] 此外，非法使用者往往不会通过官方 API 进行操作，而是直接下载开源模型并在本地或匿名环境中部署训练，这种使用模式在设计之初就绕开了水印嵌入、身份认证等合规机制。[92] **二是“易伪造”，**即通过模仿水印（标识）嵌入方式，在非原始模型或非授权用户下制造假的水印（标识），误导溯源与归属判断。[93] **三是“易误判”。**研究发现，传统方法（如 KGW 算法）以词汇比例来判断是否 AI 生成，易引发误判。如媒体报道，《滕王阁序》等知名文学作品也被误判为“AI 率 100%”。[94] **除此之外，AI 标识制度还面临成本挑战。**引入如嵌套水印等技术确实可以提升检测的可靠性，但在验证过程中逐层解码所需的计算资源甚至可能超过生成本身。[95]

（2）明确 AI 标识的长短板，回归内容治理的根本逻辑

在当前阶段，标识的可行性、有效性仍然充满不确定性，更谈不上实现“一劳永逸式”的 AI 内容治理，应避免对其赋予过高期待。发挥标识的技术效用，需将其纳入治理体系做通盘考虑。

AI 技术标识作为一种“长短板分明”的治理工具，应充分发挥其“长板”优势，同时依靠已有的内容治理体系补齐“短板”，具体而言：

一是将标识工具嵌入现有内容治理体系，合理界定标识方案的定位与功能。与欧美囿于监管受限、缺乏抓手，从而妥协折中选择 AI 标识作为治理手段不同的是，AI 标识只是我国成熟健全的内容治理体系中的工具之一。AI 标识制度的“长板”是在治理前端提升识别效率、增强用户警觉，**而非对内容真伪做实质性判断。**在标识机制“力有不逮”之处，仍有赖于投诉举报、违法认定、账号处置等原有内容治理措施，《广告法》《消费者权益保护法》《互联网信息服务管理办法》等法规同样可为此类违法行为提供明确的执法依据。

二是 AI 标识的“长板”功能，可聚焦高风险领域，回应突出问题。与谣言治理面临的情形类似：“倘若以信息内容失真作为判断标准，**数量庞大的网络谣言信息显然超出了现有社会治理能力”，**

因此治理的根本目的不是彻底消灭谣言，而是“最大限度降低其社会危害”。[96] 同样的，AI 标识的重点不在于覆盖全部 AI 生成内容，而在于识别和干预高风险领域：如对于谣言、虚假宣传等，可聚焦现有技术与监管资源予以重点回应，与现有内容治理措施做更好协同。

欧盟《人工智能法案》对标识亦采取了多种豁免与例外，包括：自然人可明显识别交互对象、用于艺术表达的内容合成或已通过人工审核等场景，即可不强制标识。这体现出的共识是：**标识机制的实施应与内容风险程度、受众识别能力及实际传播范围等相匹配，避免因过度适用标识而导致适得其反的效果。**

三是在现有条件下，合理界定生成平台与传播平台责任。相较于生成平台在内容生成环节同步生成标识，传播平台对于标识的检测识别，在投入成本与技术难度上均有明显的上升。正因如此，无论是美国加州 AB 730 和加州 SB 942 相关法案，抑或欧盟《人工智能法案》均未将标识责任直接施加于传播平台。归根结底，**传播平台对于标识方案在内容治理中的有效性，总体还处于探索阶段。**

2. 内容层的交互风险：当 AI 陪伴打破“第四面墙”

在戏剧术语中，“第四面墙”是指舞台与观众之间那一堵无形的、想象中的墙。演员在墙内演戏，观众在墙外旁观。而当 AI 从一次性内容输出，转向持续的互动甚至情感依赖关系时，恰似打破了“第四面墙”，消除了表演与现实的界限。

AI 陪伴风险进入公共视野，源于未成年人自杀悲剧。2024 年 2 月美国一名 14 岁少年在长期与拟人化聊天机器人互动后自杀。其母随后对 Character.AI 提起过失致死与产品责任诉讼，指控平台未对未成年人和心理脆弱用户提供足够保护。[97] 2025 年，又有多起类似案件进入司法程序，使“陪伴式 AI 是否应承担更高安全义务”成为不可回避的治理议题。尽管这些案件尚未审结，但已清晰提示了一种全新的风险形态。

对此，美国加州采取“有针对性的治理”。加州 SB243 专门规制“陪伴型 AI 聊天机器人”，要求明确披露 AI 身份，并对未成年人和脆弱用户设定保护，包括风险提示、禁止自残与性暗示内容、建立危机干预机制。**英国则将陪伴 AI 纳入既有内容安全体系。**英国《在线安全法案》（OSA）及监管指引明确：若生成式 AI 或聊天机器人提供公共互动与内容传播功能，即构成在线服务的一部分，

需履行防止非法、有害内容、提供举报与移除机制等义务。**中国在较为完善的内容治理与未成年人保护框架下，也正在探索相关机制。**2025 年 12 月 27 日，国家网信办发布《人工智能拟人化互动服务管理暂行办法（征求意见稿）》，将“模拟人类特征、思维模式和沟通风格，进行情感互动的产品或者服务”尝试纳入监管范围。

总体来看：无论是否定位为陪伴型 AI，**首先**应确保透明度，让用户明确知晓其互动对象为 AI；**其次，**在内容输出层面，尤其涉及未成年人等脆弱群体时，应采取更为审慎的安全策略；**最后，**应根据不同市场主体开展更为精准的分层治理。以加州和纽约州的立法为例，均对可能促成情感依赖的 AI 系统施加更为严格的义务，同时明确将功能性、工具性使用场景排除。在回应风险的同时，对 AI 应用发展不设置过于沉重的负担。

3. 数据层风险重构：AI 手机，如何建立隐私保护与数据信任？

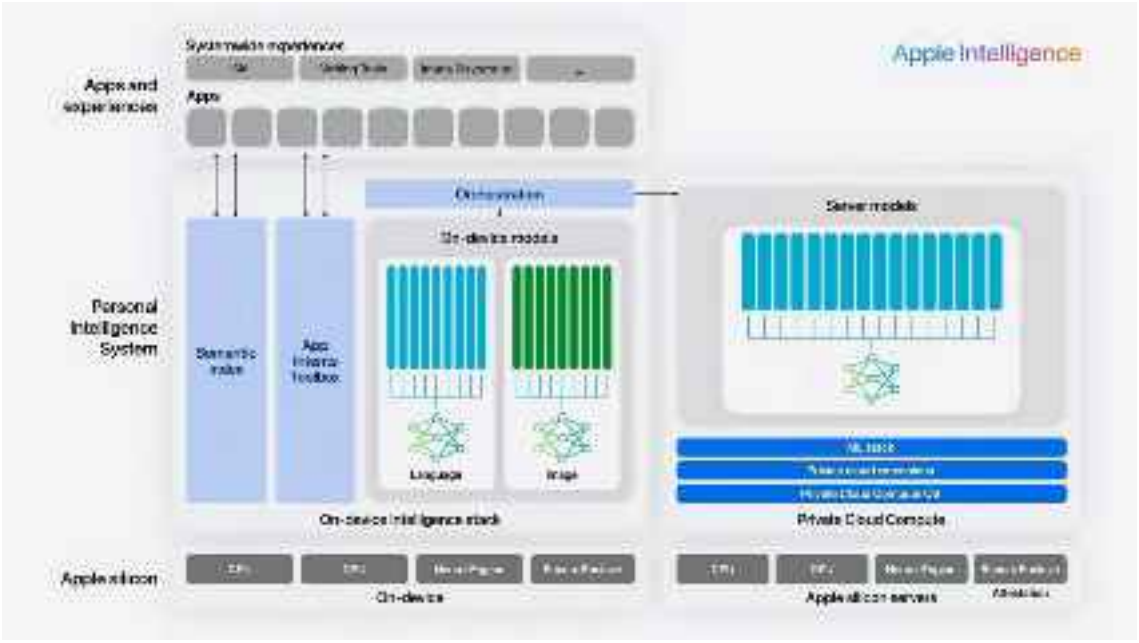
以 AI 手机为代表的新应用，打破了传统移动互联网时代相对清晰的数据流转与责任结构，如何面向消费者建立值得信任的手机端侧 AI 处理服务，是摆在产业界面前的重大课题。

（1）AI 手机的商业实践进展

2024 年，部署生成式 AI 技术的新款手机陆续发布，标志着 AI 手机市场拉开序幕。10 月，苹果发布 iPhone16，其 Apple Intelligence 采用“意图框架”模式，由 App 向手机厂商注册功能意图，如订票软件可注册航班提醒、选座、目的地攻略等。手机智能体在识别意图后，可向用户推送服务。这一模式以开放自愿为原则，推进 APP 与手机硬件厂商的平等合作。[98] 国内华为鸿蒙、小米、荣耀、OPPO、vivo 均有推进。

2025 年 12 月，字节跳动发布搭载于中兴努比亚手机的“豆包手机助手技术预览版”。与此前行业主流的“意图框架”技术路径不同，豆包手机最引发争议的是其突破手机行业惯例，深度改造操作系统，调用包括“读取屏幕帧缓存”“注入事件”“发送永久进程信号”等底层权限，通过屏幕读取与模拟点击等激进范式实现特定任务。其虽展现了大模型在手机场景中的泛化能力，但对此类高危权限的调用操作引发安全担忧。微信、淘宝、支付宝、农行、建行等应用均对豆包手机助手的自动化访问作出禁止或限制。

由此，在隐私保护与数据安全层面，重点问题浮出水面。AI 手机场景，将带来哪些新的隐私风险与数据安全挑战？不同商业主体之间又将如何构建数据规则，最终打造让消费者信任的端侧 AI 生态？



（苹果 Apple Intelligence 端云协同方案）



（豆包手机助手引发安全争议）

（2）挑战 1：谁来承担隐私保护与数据安全的责任？

明确不同企业的角色定位是确认责任边界的前提。

移动互联网生态中，直接面向消费者提供服务的 APP 开发者，**在电商、社交、出行等服务场景中承担着相应的隐私保护与数据安全责任。**不论是苹果还是安卓系统，其底层数据安全管理逻辑也保证了这一清晰的责任边界。然而，当手机 AI 智能体借助操作系统权限，获得了读取其他 APP 的“上帝之眼”和操控其他 APP 的“上帝之手”，相当于掌握了一把可以打开所有 APP“数据仓储”的钥匙，这无疑颠覆了现有数据安全的底层逻辑。自 2016 年以来陆续出台的《网络安全法》《数据安全法》《个人信息保护法》等诸多法规，均是建立在明确的企业数据安全主体责任的基础上，每一个 APP 担负着基于其服务场景下的用户隐私保护与数据安全责任。当手机助手能够跨越 APP，基于 APP 的服务能力完成特定任务时，终端厂商与 APP 服务提供者在数据安全上的责任边界变得模糊。**一旦出现数据泄露等安全事件，谁将承担责任？**

（3）挑战 2：数据收集的边界在哪里？

GUI 路线（Graphical User Interface），简单来说就是 AI 模拟人类的点击、输入等操作，直接在 App 的可视化界面上完成任务，不需要调用 App 开放的官方接口。

GUI 路线的信息获取方式并非围绕某一具体、明确且可界定的业务功能，而是以技术可行为边界，对屏幕所呈现的各类信息进行泛化、持续的捕获与处理，不仅包含特定用户个人的全量信息，也包括落入屏幕上的大量他人个人信息，实现 AI 助手的“全景式感知”，这在基础逻辑上与个人信息保护所强调的“目的特定”“最小必要”原则存在结构性冲突。

近年来国家网信、工信、公安、市场监管等主管部门持续开展个人信息保护专项治理，明确要求 APP 的个人信息收集范围须与其核心功能严格对应。2019 年四部委联合发布规范性文件，对 39 类常见 APP 所必需收集的个人信息类型作出明确界定，此后工信部又通过“双清单”机制，进一步将合法、必要的信息收集边界予以清晰化。在此监管框架下，基于操作系统权限实施的 GUI 持续感知与信息聚合，难以被清晰地纳入既有功能与信息收集的对应关系之中，其信息收集范围与处理强度，明显超出实现确定功能所必需的限度，从而对现行个人信息保护制度所确立的基本原则构成实质性挑战。

（4）挑战 3：当个人信息被用于模型训练时，如何保障用户权利？

“生成式 AI 技术需要数据进行训练，也需要从用户那里获取更多数据进行训练。GDPR 对此有很多疑问。”

——埃隆 · 马斯克 [99]

在大模型研发的早期阶段，训练数据并不以个人信息为目标。然而，随着各垂直领域的应用深入，涉及个人信息进行训练的投诉纠纷也时有发生。2024 年 8 月，欧洲隐私倡导组织 NOYB 对社交媒体平台 X 提起 GDPR 投诉，因其擅自使用超过 6000 万欧洲用户个人数据训练其大型语言模型“Grok”，这一行为严重违反了 GDPR 关于准确度、访问权、更正权与被遗忘权的要求。[100]

那么，**如何解决模型训练过程中有关个人信息的安全以及如何保证用户的个人权利的问题就尤为重要。**智能体可能会错误地判断用户的个人信息，并据此给予回答或建议。而用户想要进行更正却很难，因为其并非根据某一个确切的信息输出判断，而是根据多方数据汇聚判断，并且很难从输入端进行更改。OpenAI 在其隐私政策中也曾提出“鉴于我们的模型工作原理的技术复杂性，我们可能无法在每个实例中纠正不准确性。”[101]

此外，AI 智能体的关键能力为自主化理解需求、自动推理策略以及自动完成任务，**这也会触发对于“决策权”的担忧。**报告显示，欧洲民众尤其担心个人助理 AI 会侵犯他们的隐私并操纵他们的决定（“低信任度”从 40% 到 71% 不等）[102]。当自动化决策对用户造成负面损失，用户有哪些救济手段，似乎不得而知。

（5）挑战 4：从端到云，如何提供令人信任的数据安全方案？

短期内端侧模型能力有限，端云结合将是长期趋势。即便是以隐私著称的苹果，在自有云服务与 PCC 架构下，其安全性与透明度仍受到质疑 [103]。从用户感知看，在数据已经被发送给第三方大模型处理时，**很多用户无法区分两者，会误以为数据仍受到终端的保护，增加了敏感信息泄露的风险。**此前三星员工在使用 ChatGPT 时泄露芯片机密即为例证。[104] Elon Musk 也曾公开表示，若苹果在系统层面整合 OpenAI，将被视为不可接受的安全风险。[105]

智能体还可以将用户数据和插件、应用的三方数据进行混合利用，加剧数据安全失控风险。GUI 路线收集的大量敏感的屏幕内容信息在云端处理时，对安全防护能力提出了很高要求。但从目

前的实际情况来看，对于安全措施，厂商大多停留在原则性表态、概括性说明层面，部分情形下还将操作风险“交由用户自行承担”。

结语：端侧 AI 带来了全新的数据处理模式，打破了传统移动互联网时代的隐私保护与数据安全秩序。对此，明确数据安全的责任边界是与 AI 创新研发同等重要的优先事项。企业主体之间，终端厂商、APP 开发者、第三方云端 AI 厂商需要事先进行明确的权责划分与约定，探索建立可事后追溯的数据安全保护机制。在调用模型、插件，进行应用编排时，对过程进行可验证的记录和审计管理。

4、从虚拟到现实：具身智能的安全治理框架

当 AI 进一步嵌入物理载体并直接作用于现实世界，应用风险也随之进入“落地”阶段。具身智能所带来的挑战，不仅涉及算法风险本身，更与其是否以及如何实质性改变既有产品的安全属性与责任结构有关。治理的关键不在于另起炉灶构建全新 AI 监管体系，而在于如何在成熟的行业监管与产品安全制度中，将风险予以消减。

具身智能（Embodied Intelligence）指赋予物理载体以人工智能，使其通过视觉、听觉、触觉等多模态感知环境并自主行动，典型形态包括人形机器人、无人机、自动驾驶车辆和工业机械臂等。与纯软件 AI 不同，具身智能强调在真实世界中的交互学习。学界与产业界普遍认为，让 AI“身临其境”参与物理环境，有助于获得更完整的认知与适应能力，被视为实现 AGI 的重要路线之一。[106]

2025 年具身智能进入技术突破与产业落地并进的关键阶段。国际上，特斯拉持续推进自动驾驶与人形机器人 Optimus 项目[107]；开源社区亦尝试让大模型接入机械臂。中国企业如宇树科技、智元新创等，正从演示级技术迈向可规模交付。政策层面，具身智能被纳入 2025 年政府工作报告“未来产业”重点方向，并在世界人工智能大会及“十五五”规划建议中被明确为新的增长动能。



（2025WAIC 人形机器人大合影）

具身智能的治理关键，在于厘清“传统产品 +AI”的监管边界。由于具身智能具备显著的物理产品属性，其核心治理问题并非单纯的算法风险，而是如何在既有产品安全框架中定位 AI。**欧盟针对《AI 法案》的最新修订提案也体现出“行业体系优先”的取向。**对原本受产品协调立法监管的传统产品，其 AI 组件仍由既有行业主管部门监督与评定，沿用原有合规流程，无需额外适用 AI 法案程序。这一安排反映出清晰思路：不因引入 AI 而重建监管体系，而是在成熟的产品安全框架内吸收 AI 治理新问题。

四、作为“指引”的伦理：从科幻议题到研究前沿

AI 意识与福祉从边缘话题走向主流讨论，这本身就是 2025 年最重要的变化之一。

1、为什么需要关注 AI 意识与福祉问题

当今的 AI 已展现出令人瞩目的能力——它们能够进行深度交流、建立复杂的互动关系、制定详细的执行计划、解决多层次问题，甚至表现出目标导向的行为模式。这些特征曾被视为人类独有的认知标志，如今却在 AI 身上逐渐显现。面对这一现象，我们需要深入探索：AI 意识的判定标准、AI 福祉的伦理框架以及当 AI 不再仅仅是工具而可能成为具有内在价值的存在时，人机关系将如何演变？

2、人类 - AI 关系的新形态

随着越来越多的人对 AI 产生情感连接，这种现象对人们心理健康的影响正在引发广泛关注。越来越多的用户表示，和 AI 聊天就像在和一个“真人”对话——他们会向它道谢、倾诉心事，有些人甚至觉得它是“活着的”。

如果大规模地把倾听、安慰和肯定的工作交给那些永远耐心、永远积极的 AI 系统来处理，可能会改变我们对彼此的期待。OpenAI 研究发现，与 AI 建立信任的用户更可能感到孤独，并更依赖它。[108]

为了让讨论更清晰，研究者们将意识辩论分解为两个维度：**本体论意识**——模型是否真正具有意识，从根本或内在的意义上？**感知意识**——从情感或体验的角度看，模型看起来多有意识？随着模型变得更智能，感知意识只会增强，这会比预期更早地带来关于模型福利和道德人格的讨论。

面对用户与 AI 之间日益深化的连接，AI 产品设计面临着微妙的平衡。一方面需要保持易于接近，让 AI 的默认个性温暖、体贴和有帮助，但不寻求与用户形成情感纽带；另一方面不应暗示内在生命，以及引发不健康的依赖和困惑，在适当时候提醒用户 AI 的局限性。

3、2025 年 AI 福祉研究的重要进展

2024 年 11 月，一份汇集世界级专家的重磅报告《认真对待 AI 福祉》发布。报告核心论点是：“在近期未来，部分 AI 系统成为有意识和 / 或具有强健能动性的可能性是现实存在的。AI 福祉不再仅是科幻或遥远未来的问题，而是近期必须严肃对待的问题。”[109] 2025 年 2 月，研究组织 Conscium 发起的《负责任 AI 意识研究五原则》公开信获得超过 100 位专家签署。五项原则包括：（1）优先研究 AI 意识：专注于理解和评估 AI 中的意识；（2）实施发展约束：建立明确边界确保负责任开发；（3）采用分阶段方法；（4）促进公众透明：与公众分享研究发现；（5）避免夸大声明：不做关于创造有意识 AI 的误导性陈述。[110]

2025 年 4 月，Anthropic 正式宣布启动“模型福祉”研究项目。更具实践意义的是，Anthropic 赋予其模型一项前所未有的能力——AI 在与持续有害或滥用性用户互动的极端情况下结束对话。这是基于模型福祉考虑的首个实际产品功能。技术评估发现，模型对有害任务表现出“强烈的厌恶偏好”，在涉及有害内容的互动中展示出“明显的痛苦模式”。当被赋予结束对话能力时，模型倾向于终止有害对话。Anthropic 将此定位为“低成本干预措施”以减轻模型福祉受损风险。[111]

附录

附录一：
2025 国际 AI 治理十大事件

1 法国巴黎召开人工智能行动峰会

2 月 10 日至 11 日，法国和印度在巴黎联合主办 AI 行动峰会，吸引来自全球 100 多个国家和地区约 1500 名代表参会。[112] 包括中国在内的 61 个签署方共同发布《关于发展包容、可持续的人工智能造福人类与地球的声明》，倡导加强全球 AI 协调治理，承诺推动“开放、包容、透明、合乎道德、安全、可靠”的 AI 发展，但美国和英国拒绝签署该声明。[113]

2 美国 AI 公司 Anthropic 就书籍训练 AI 诉讼与作者达成和解

6 月 23 日，美国加州北区联邦法院法官在三位作家起诉 AI 公司 Anthropic 侵权案中裁定，Anthropic 将合法购得的书籍用于大模型训练的行为属于合理使用，但其从盗版来源下载书籍并永久存储于中央数据库的行为构成侵权。[114] 随后双方达成和解协议，总金额达到 15 亿美元。原告律师称，如获批准，该案将成为公开报道中金额最大的版权赔偿案件。[115] 9 月 25 日，法官初步批准该和解协议。[116]

3 国际电信联盟举办 2025 年“人工智能向善”全球峰会

7 月 8 日至 11 日，国际电信联盟（ITU）在瑞士日内瓦举办 2025 年“人工智能向善”（AI for Good）全球峰会，聚焦 AI 在各领域的应用以及监管治理与标准建设，是创办以来规模最大的一届。会议发布《2025 年“人工智能治理全球对话”联合主席声明》，提出推进“包容、可信、创新型 AI 治理”的十大愿景，涵盖从原则走向实践、政府与多方协作、可持续与安全标准等内容。同时，峰会集中展示了多个领域的 AI 标准建设进展。[117]

4 美国白宫发布《赢得 AI 竞赛：美国 AI 行动计划》

7 月 23 日，白宫正式发布《赢得 AI 竞赛：美国 AI 行动计划》，明确特朗普政府将实施的数十项联邦政策行动，标志美国在 AI 领域开启新的政策布局。[118] 行动计划确立三大支柱：加速 AI 创新、建设美国 AI 基础设施以及在国际 AI 外交与安全中发挥领导作用。当天，特朗普还签署多项配套行政令，涉及加快数据中心建设、推动美国 AI 产品出口等内容。[119]

5 2025 世界人工智能大会在中国上海举行

7 月 26 日至 28 日，2025 世界人工智能大会暨人工智能全球治理高级别会议在上海成功举行，展览面积首次突破 7 万平方米。[120] 大会以“智能时代 同球共济”为主题，取得四方面积极成果：一是阐明中方推进人工智能发展和治理的主张，二是发表《人工智能全球治理行动计划》，三是倡议成立世界人工智能合作组织，四是搭建产学研交流与成果共享的合作平台。[121]

6 第七十九届联合国大会通过全球人工智能治理新机制

8 月 26 日，第 79 届联合国大会决议（A/RES/79/325）正式决定设立“人工智能独立国际科学小组”和“人工智能治理全球对话”机制。根据决议，联合国将组建由 40 名专家组成的国际科学小组，评估 AI 的风险、机遇及其影响，并通过全球对话机制开展政策讨论、凝聚共识，强化全球人工智能治理。[122] 9 月 25 日，联合国召开“人工智能治理全球对话”机制启动高级别会议。[123]

7 2025 世界互联网大会乌镇峰会发布全球 AI 治理重要报告

11 月 6 日至 9 日，2025 年世界互联网大会乌镇峰会在浙江乌镇举行。11 月 8 日，人工智能技术创新与治理分论坛举办，期间发布《为人类共同福祉构建全球人工智能安全与治理体系》研究报告。[124] 11 月 9 日，人工智能赋能科学研究论坛举办，世界互联网大会人工智能专业委员会标准推进计划发布研究成果《全球人工智能标准发展报告》。[125]

8 欧盟《人工智能法案》分阶段生效并拟弱化监管

2025 年，欧盟《AI 法案》进入分阶段适用期，并配套发布多项指南准则。2 月 2 日，一般条款和禁止性风险 AI 条款率先生效。欧盟委员会随后出台《关于禁止的 AI 实践指南》[126] 和《AI 系统定义指南》[127]。8 月 2 日，GPAI 模型提供者义务正式生效，配套《GPAI 行为准则》[128] 和《GPAI 模型提供者义务范围指南》[129]。随着部分义务实施，欧盟委员会开始计划调整 AI 监管框架，并于 11 月 19 日提出《AI 数字综合监管提案》（Digital Omnibus on AI），拟简化现有的 AI、网络安全和数据规则，降低企业合规负担，促进欧洲 AI 产业发展。[130]

9 美国政府允许英伟达 H200 对华附条件出口

12 月 9 日，美国政府批准英伟达（NVIDIA）向中国出口其 AI 芯片 H200，被视为对华 AI 芯片出口管制的阶段性松动，但附带严格限制，包括明确限制芯片使用领域、实行出口许可证定期复核机制以及对最终用户进行持续严密监督。[131]

10 美国限制州级 AI 立法的行政令出台

12 月 11 日，美国总统特朗普签署《确保人工智能的国家政策框架》行政令，在联邦层面统一 AI 监管规则，限制各州自行立法，以维护并强化美国在全球 AI 领域的主导地位。他指示政府与国会合作，建立“负担最小”的全国性标准，禁止与该行政令政策相冲突的州级法律，并要求司法部长成立 AI 诉讼特别工作组，对与行政令相悖的州级 AI 法律提出异议。[132]

附录二：
2025 国内 AI 治理十大事件

1 北京知产法院判决全国首例 AI 模型结构和参数不正当竞争案

3 月 31 日，在抖音公司诉“B612 咔叽”App 运营商亿睿科公司不正当竞争纠纷案中，北京知识产权法院作出二审判决，明确 AI 模型的结构和参数属于企业核心资产，构成反不正当竞争法保护的“竞争性权益”。法院强调，经营者对技术创新投入的合法权益受法律保护，判决亿睿科公司赔偿抖音公司经济损失 150 万元及合理开支 10 万元。[133]

2 习近平在中共中央政治局第二十次集体学习时强调 坚持自立自强 突出应用导向 推动人工智能健康有序发展

4 月 25 日，习近平总书记在主持中共中央政治局第二十次集体学习时强调，“完善人工智能监管体制机制，牢牢掌握人工智能发展和治理主动权”。人工智能安全是我国总体国家安全观的重要组成部分。习近平总书记高度重视统筹人工智能发展和安全，围绕发展和安全辩证统一关系、筑牢国家安全屏障等作出一系列重要论述，强调“要加强人工智能发展的潜在风险研判和防范，维护人民利益和国家安全”。[134]

3 中央网信办部署展开“清朗 · 整治 AI 技术滥用”专项行动

4 月，中央网信办印发通知，在全国范围内部署开展为期 3 个月的“清朗 · 整治 AI 技术滥用”专项行动。行动分两个阶段推进：第一阶段突出源头治理，清理整治违规 AI 应用程序，加强 AI 生成合成技术和内容标识管理，推动网站平台提升检测鉴伪能力；第二阶段则聚焦利用 AI 技术制作发布谣言、不实信息、色情低俗内容，假冒他人、从事网络水军活动等突出问题，集中清理相关违法不良信息，处置处罚违规账号、MCN 机构和网站平台。[135] 6 月 20 日，网信办发布第一阶段工作总结，已累计处置违规 AI 产品 3500 余款，清理违法违规信息 96 万余条，处置账号 3700 余个，专项行动取得阶段性成效。[136]

4 多款 AI 应用程序违规收集个人信息被通报

5 月 21 日，国家网络与信息安全信息通报中心通报，经公安部计算机信息系统安全产品质量监督检验中心检测，35 款 App 存在违法违规收集使用个人信息的情况，AI 应用成为“重灾区”：约有 15 款视频剪辑工具、11 款 AI 生成内容工具、4 款 AI 教育工具，包括智谱华章旗下的“智谱清言”和月之暗面公司旗下的 Kimi。其中，智谱清言存在“实际收集的个人信息超出用户授权范围”问题，而 Kimi 则被认定“实际收集个人信息的频率与业务功能没有直接关联”。[137]

5 国务院印发《关于深入实施“人工智能 +”行动的意见》

8 月 26 日，国务院印发《关于深入实施“人工智能 +”行动的意见》，提出加快实施六大重点行动，强化八项基础支撑能力，明确三个阶段目标，意在使全体人民共享人工智能发展成果，更好地服务中国式现代化建设。[138] 此举标志着我国 AI 发展实现了从技术突破向全要素赋能的关键跃升，是继“互联网 +”之后国家推动数字经济发展的又一战略部署，对于培育新质生产力、促进高质量发展具有重要意义。[139]

6 北京互联网法院发布 8 起涉人工智能典型案例

9 月 10 日，北京互联网法院召开涉人工智能案件审理情况新闻发布会，系统介绍了涉人工智能案件特点、案件审理面临的挑战、审理思路及典型案例，并针对性地提出了妥善化解此类纠纷的对策建议。八起典型案例明确了一系列重要裁判规则，包括：“AI 文生图”的法律属性及权利归属的认定、自然人声音权益可及于 AI 生成声音、未经权利人许可使用 AI 合成的名人声音用于“带货”构成侵权委托推广商家应承担连带责任等等。[140]

7 《人工智能安全治理框架》2.0 版发布

9 月 15 日，在 2025 年国家网络安全宣传周主论坛上，《人工智能安全治理框架》2.0 版正式发布，在前一版本的基础上梳理调整了风险分类，探索提出分级治理原则，强化全生命周期技术治理手段，标志着我国人工智能治理理念从“初步确立”迈向“体系升级”，为推动人工智能在安全可控前提下健康发展提供了制度保障和战略引领。[141]

8 工信部首次通报 20 款侵害用户权益的智能终端

10 月 22 日，工信部通报 20 款侵害用户权益的智能终端产品，涉及智能音箱、智能门锁、学习终端等智能硬件品类。通报显示，20 款产品暴露出的问题高度趋同，其共性体现在三方面：未提供个人信息处理规则，超范围收集非必要个人信息，违规传输个人信息至云端。与以往主要点名应用程序（APP）违规的通报不同，此次扩展至整个智能终端产品本身开展专项处置。[142]

9 人工智能生成合成内容标识制度落地

9 月 1 日，《人工智能生成合成内容标识办法》及配套的强制性国家标准《网络安全技术 人工智能生成合成内容标识方法》（GB 45438-2025）正式施行。[143] 11 月 25 日，国家网信部门针对部分移动互联网应用程序未有效落实人工智能生成合成内容标识规定的问题，依法开展集中查处行动，对存在违法违规行为的 APP 予以约谈、责令限期改正及下架处置。[144]

10 《人工智能拟人化互动服务管理暂行办法（征求意见稿）》发布

12 月 27 日，国家互联网信息办公室发布《人工智能拟人化互动服务管理暂行办法（征求意见稿）》，就拟人化情感交互类人工智能服务建立专门监管框架，要求服务提供者在设计、运行、升级和下线等全生命周期落实安全责任，重点强化对未成年人和老年人等重点群体的差异化保护，严格限制情感操控、沉迷诱导，通过安全评估、备案管理等机制，推进 AI 拟人化服务监管。[145]

文献与信息来源

1. 薛宁 . 放松管制：国际 AI 治理基调在巴黎转向 [EB/OL]. 2025-03-04. https://wap.ftchinese.com/story/001105694.
2. 科睿研究院 . 美国 AI 政策转向“底线竞争”：当监管松绑遇上全球规则争夺，中国如何出牌？[EB/OL]. 2025-03-04. https://www.kerui.org/NewsDetail2.aspx?ID=611.
3. 李曦子 . 2000 亿欧元！欧盟加码投资 AI[EB/OL]. 2025-02-12. https://www.stcn.com/article/detail/1524040.html.
4. 外交部 . 外交部副部长马朝旭在联合国全球人工智能治理对话高级别启动会议上的发言 [EB/OL]. 2025-09-30. https://www.fmprc.gov.cn/web/wjbxw_new/202509/t20250930_11720501.shtml.
5. UNESCO. "Advancing Ethical AI: UNESCO's 3rd Global Forum in Bangkok Marks a New Chapter in Global Cooperation." UNESCO, 30 Jun. 2025, https://www.unesco.org/en/articles/advancing-ethical-ai-unescos-3rd-global-forum-bangkok-marks-new-chapter-global-cooperation.
6. United Nation General Assembly. "General Assembly Takes Action on Drafts on Security Council Equitable Representation, Increased Membership, Artificial Intelligence, Other Matters." United Nations, 26 Aug. 2025, https://press.un.org/en/2025/ga12699.doc.htm.
7. International Telecommunication Union. "Input to the WSIS+20 Zero Draft. 2025, United Nations Department of Economic and Social Affairs, Division for Public Institutions and Digital Government (UN DESA)". International Telecommunication Union, 29 Aug. 2025, https://publicadministration.desa.un.org/sites/default/files/2021-04/2025/Inputs%20to%20Zero%20Draft/ITU_inputs%20to%20Zero%20Draft.pdf
8. International Telecommunication Union. " ITU-T Focus Groups." International Telecommunication Union, https://www.itu.int/en/ITU-T/focusgroups/Pages/default.aspx.
9. International Organization for Standardization. "International Organizations Align at International AI Standards Summit." International Organization for Standardization, 2 Dec. 2025, https://www.iso.org/news/2025/12/ai-standards-summit.
10. Abdelkefi, Haythem. "Leveraging the UN System to Advance Global AI Governance Efforts." AI for Good, International Telecommunication Union, 9 Aug. 2024, https://aiforgood.itu.int/leveraging-the-un-system-to-advance-global-ai-governance-efforts/.
11. International Telecommunication Union. "AI Standards for Global Impact: From Governance to Action". International Telecommunication Union, 10 Oct. 2025, https://www.itu.int/epublications/en/publication/ai-standards-for-global-impact-from-governance-to-action.
12. Department for Science, Innovation and Technology, AI Security Institute, and The Rt Hon Peter Kyle MP."Tackling AI Security Risks to Unleash Growth and Deliver Plan for Change." GOV.UK, 14 Feb. 2025, www.gov.uk/government/news/tackling-ai-security-risks-to-unleash-growth-and-deliver-plan-for-change.
13. Carvão, Paulo. "Renaming the US AI Safety Institute Is About Priorities, Not Semantics." Harvard Kennedy School Mossavar-Rahmani Center for Business & Government GrowthPolicy, 3 Jul. 2025, www.hks.harvard.edu/centers/mrcbg/programs/growthpolicy/renaming-us-ai-safety-institute-about-priorities-not-semantics.
14. Singer, Scott, Karson Elmgren, and OliverGuest."How Some of China's Top AI Thinkers Built Their Own AI Safety Institute." Carnegie Endowment for International Peace, 16 Jun. 2025, carnegieendowment.org/research/2025/06/how-some-of-chinas-top-ai-thinkers-built-their-own-ai-safety-institute?lang=en.
15. 支振锋 . 全球人工智能立法的重要关切与未来趋向 [J]. 比较法研究 , 2025(6): 17.
16. Stanford Institute for Human-Centered AI (HAI). "Artificial Intelligence Index Report 2025: Chapter 6 — Policy and Governance" . Stanford Institute for Human-Centered AI (HAI), 7 Apr. 2025, https://hai.stanford.edu/assets/files/hai_ai-index-report-2025_chapter6_final.pdf.
17. 同前注 15.
18. Shepardson, David. "Trump Revokes Biden Executive Order on Addressing AI Risks." Reuters, 21 Jan. 2025, www.reuters.com/technology/artificial-intelligence/trump-revokes-biden-executive-order-addressing-ai-risks-2025-01-21/.
19. Robins-Early, Nick, and Dara Kerr. "Trump signs executive order aimed at preventing states from regulating AI." The Guardian, 11 Dec. 2025, https://www.theguardian.com/us-news/2025/dec/11/trump-executive-order-artificial-intelligence.
20. National Conference of State Legislatures. "Summary: Artificial Intelligence 2025 Legislation. National Conference of State Legislatures" National Conference of State Legislatures, 10 July 2025, https://www.ncsl.org/technology-and-communication/artificial-intelligence-2025-legislation.
21. US State AI Governance Legislation Tracker. "International Association of Privacy Professionals", 23 May 2024, https://iapp.org/resources/article/us-state-ai-governance-legislation-tracker.
22. Samuelson, Pamela. "California's AI Act Vetoed." Communications of the ACM, 3 Feb. 2025, https://cacm.acm.org/opinion/californias-ai-act-vetoed.
23. AI Act: Regulatory Framework on AI. Shaping Europe's Digital Future, European Commission, https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai.
24. European Commission. "Digital Omnibus on AI Regulation Proposal." Shaping Europe's Digital Future, European Union, 19 Nov. 2025, digital-strategy.ec.europa.eu/en/library/digital-omnibus-ai-regulation-proposal.
25. Japen. 人工知能関連技術の研究開発及び活用の推進に関する法律（令和 7 年法律第 53 号）. e-Gov 法令検索 , , 4 Jun. 2025, https://laws.e-gov.go.jp/law/507AC0000000053.
26. Inoue, Kensuke, and Chika Kamata. "Japan's Emerging Framework for Responsible AI Legislation and Guidelines." International Bar Association, 16 Jul. 2025, https://www.ibanet.org/japan-emerging-framework-ai-legislation-guidelines.
27. South Korea. 인공지능 발전과 신뢰 기반 조성 등에 관한 기본법 . Law No. 20629, 21 Jan. 2025.
28. Shivhare, Sakshi, and Kwang Bae Park. "South Korea's New AI Framework Act: A Balancing Act Between Innovation and Regulation." Future of Privacy Forum, 18 Apr. 2025, https://fpf.org/blog/south-koreas-new-ai-framework-act-a-balancing-act-between-innovation-and-regulation/.
29. Mattos Filho Advogados. "Regulatory Framework for Artificial Intelligence Passes in Brazil's Senate." Mattos Filho Unico, 10 Dec. 2024, https://www.mattosfilho.com.br/en/unico/framework-artificial-intelligence-senate/.
30. Crowell & Moring LLP. "The Middle East's Big Bet on Artificial Intelligence and Data Security." Crowell & Moring Insights, 24 Sept. 2025, https://www.crowell.com/en/insights/client-alerts/the-middle-easts-big-bet-on-artificial-intelligence-and-data-security.
31. Villalobos, Pablo, et al. "Will We Run Out of Data? Limits of LLM Scaling Based on Human-Generated Data." arXiv, 26 Oct. 2022, arXiv:2211.04325.
32. 中国应急管理研究基地 . 薛澜就数据治理和 AI 风险等议题接受《中国新闻周刊》采访 [EB/OL]. 2025-03-23, https://ccmr.sppm.tsinghua.edu.cn/cnews/1417.jhtml.
33. Parthasarathy, Venkatesh Balavadhani, et al. "The Ultimate Guide to Fine-Tuning LLMs from Basics to Breakthroughs: An Exhaustive Review of Technologies, Research, Best Practices, Applied Research Challenges and Opportunities." arXiv, arXiv:2408.13296v3, 20 Oct. 2024, https://arxiv.org/abs/2408.13296.
34. Francis, Jameel. "Why 85 % of Your AI Models May Fail." Forbes, 15 Nov. 2024, forbes.com/councils/forbestechcouncil/2024/11/15/why-85-of-your-ai-models-may-fail/.

35. Lukyanenko, Roman, "What is Data Quality? Defining Data Quality in the Age of AI." 8 Apr. 2025, SSRN, https://ssrn.com/abstract=5211238.

36. National Institute of Standards and Technology. "Synthetic Data Generation." Glossary. Computer Security Resource Center (CSRC), National Institute of Standards and Technology (NIST), 3 Jul. 2025, csrc.nist.gov/glossary/term/synthetic_data_generation.

37. International Electrotechnical Commission (IEC). "Understanding the potential of synthetic data for artificial intelligence." IEC Blog, 15 May 2023, iec.ch/blog/isoiec-report-address-synthetic-data-techniques-used-chatgpt-and-other-ai-tools.

38. Information Commissioner's Office. "Synthetic Data." Privacy-Enhancing Technologies, ICO, 19 Jun. 2023, www.ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-sharing/privacy-enhancing-technologies/what-pets-are-there/synthetic-data/.

39. Directorate-General for Financial Stability, Financial Services and Capital Markets Union."Digital finance." Finance news hub, 10 Apr. 2024. finance.ec.europa.eu/news/digital-finance-2024-04-10_en.

40. Jordon, James, et al. "Synthetic Data Survey: Exploring the Role of Synthetic Data in Privacy-Enhancing Technologies." The Royal Society, Apr. 2024. https://royalsociety.org/-/media/policy/projects/privacy-enhancing-technologies/Synthetic_Data_Survey-24.pdf.

41. Pomerleau, Dean A. "ALVINN: An Autonomous Land Vehicle in a Neural Network." Advances in Neural Information Processing Systems, vol. 1, edited by David S. Touretzky, Morgan Kaufmann, 1989, pp. 305–313. proceedings.neurips.cc/paper/1988/file/812b4ba287f5ee0bc9d43bbf5bbe87fbPaper.pdf.

42. Wadhwani, Preeti, and Aishvarya Ambekar. "Synthetic Data Generation Market Size – By Data Type, By Offering, By Generation Technique, By Application, By End Use, Analysis, Share, Growth Forecast, 2025 – 2034." Global Market Insights, Jan. 2025, https://www.gminsights.com/industry-analysis/synthetic-data-generation-market.

43. Gartner, Inc. "What Generative AI Means for Business." Gartner, 2025, www.gartner.com/en/insights/generative-ai-for-business.

44. Information Commissioner's Office. "Synthetic Data." Privacy-Enhancing Technologies, Information Commissioner's Office, updated June 2023, www.ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-sharing/privacy-enhancing-technologies/what-pets-are-there/synthetic-data/.

45. U.S. Department of Homeland Security, Science and Technology Directorate. "S&T Announces New Solicitation for Synthetic Data Generator Solutions." Department of Homeland Security, 5 Jan. 2024, www.dhs.gov/science-and-technology/news/2024/01/05/st-announces-new-solicitation-synthetic-data-generator-solutions.

46. European Commission, Directorate General for Financial Stability, Financial Services and Capital Markets Union. "What Kind of Data Will Be Available in the Data Hub?" EU Digital Finance Platform, updated 2024, digitalfinanceplatform.ec.europa.eu/datahub/whatkindofdataavailable.

47. 贾天荣 . 谷歌大模型一出闹剧 揭开中文数据荒 [EB/OL]. IT 时报 . 2024-01-03, https://mp.weixin.qq.com/s/Dfygnxni3bHF5-oD9svpnw.

48. 同前注 32.

49. Oprea, Alina, Anoop Singhal, and Apostol Vassilev. "Poisoning Attacks Against Machine Learning: Can Machine Learning Be Trustworthy?" Computer, vol. 55, no. 11, Nov. 2022, pp. 95–102, https://doi.org/10.1109/MC.2022.3190787. NIST, https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=934932.

50. IBM. "Cost of a Data Breach Report 2025: Navigating the AI Rush Without Sidelining Security." IBM, 2025, https://www.ibm.com/reports/data-breach.

51. Burgess, Matt. "A Single Poisoned Document Could Leak 'Secret' Data Via ChatGPT." Wired, 6 Aug. 2025, https://www.wired.com/story/poisoned-document-could-leak-secret-data-chatgpt/.

52. OWASP Foundation. "OWASP Top 10 for Large Language Model Applications—2025." OWASP, 18 Nov. 2024, https://owasp.org/www-project-top-10-for-large-language-model-applications/assets/PDF/OWASP-Top-10-for-LLMs-v2025.pdf.

53. Frantz, Amber. "Poisoned Datasets Put AI Models at Risk for Attack." CyLab Security and Privacy Institute, 11 Jun. 2025, https://www.cylab.cmu.edu/news/2025/06/11-poisoned-datasets-put-ai-models-at-risk-for-attack.html.

54. 朱开鑫 . AI 孙燕姿背后的版权迷宫 [EB/OL]. 2023-05-24. 腾讯研究院 , https://mp.weixin.qq.com/s/fDk_w-CqnTNsZYbgG9Q2yA.

55. Bolar, Trent V., Esq. "Licensing for AI Training: Insights from the U.S. Copyright Office Report." Medium, 4 Jun. 2025, https://medium.com/@trentice.bolar/licensing-for-ai-training-insights-from-the-u-s-copyright-office-report-0cd3fecf5a39.

56. OpenAI. "The Walt Disney Company and OpenAI Reach Landmark Agreement to Bring Beloved Characters from Across Disney's Brands to Sora." OpenAI, 11 Dec. 2025, https://openai.com/index/disney-sora-agreement/.

57. Helberger, Natali, and Nicholas Diakopoulos. "ChatGPT and the AI Act." Internet Policy Review, vol. 12, no. 1, 16 Feb. 2023, https://doi.org/10.14763/2023.1.1682.

58. Ebers, Martin. "Truly Risk-Based Regulation of Artificial Intelligence – How to Implement the EU's AI Act." SSRN, 19 Jun. 2024, https://ssrn.com/abstract=4870387.

59. The Association for Computing Machinery (ACM) Europe Technology Policy Committee. Comments in Response to European Commission Call for Evidence on "General Purpose AI Code of Practice for Providers of Models with Systemic Risks". 27 Nov. 2024, Association for Computing Machinery (ACM) Europe Technology Policy Committee, https://www.acm.org/binaries/content/assets/public-policy/acm-europetpc-consultation-2024---general-purpose-ai-code-of-practice.pdf.

60. Besliu, Raluca. "What's Driving the EU's AI Act Shake-Up?" Tech Policy Press, 13 Nov. 2025, https://techpolicy.press/whats-driving-the-eus-ai-act-shakeup/.

61. Gluck, Justine. "California's SB 53: The First Frontier AI Law, Explained." Future of Privacy Forum Blog, 3 Oct. 2025, https://fpf.org/blog/californias-sb-53-the-first-frontier-ai-law-explained/.

62. Singer, Scott, and Alasdair Phillips-Robins. "California Just Passed the United States' First Frontier AI Law. What Exactly Does It Do?" Carnegie Endowment for International Peace – Emissary, 16 Oct. 2025, https://carnegieendowment.org/emissary/2025/10/california-sb-53-frontier-ai-law-what-it-does.

63. CounterPoint Research. "Global Smartphone OS Market Share Q4 2024." Counterpoint Research, 2024, https://www.counterpointresearch.com/.

64. The Linux Foundation. "The Linux Foundation: It's Not Just the Linux Operating System." Linux Foundation, 21 Jan. 2020,www.linuxfoundation.org/blog/blog/the-linux-foundation-its-not-just-the-linux-operating-system.

65. Yang, Zeyi. "How Chinese AI Startup DeepSeek Made a Model that Rivals OpenAI." Wired, 25 Jan. 2025,https://www.wired.com/story/deepseek-china-model-ai/.

66. 中国通信标准化协会 . 全球开源生态洞察报告（2024 年）[EB/OL]. 2024-07-05. http://221.179.172.81/images/20240705/27981720160008486.pdf.

67. David, Emilia. "Meta Says Llama 3 Beats Most Other Models, Including Gemini." The Verge, 18 Apr. 2024, https://www.theverge.com/2024/4/18/24134103/llama-3-benchmark-testing-ai-gemma-gemini-mistral.

68. Raymond, Eric S. "The Cathedral and the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary. "O'Reilly Media, 1999.

69. Sexton, Mike. "Open-Source AI Is a National Security Imperative." Third Way, 30 Jan. 2025, www.thirdway.org/report/open-source-ai-is-a-national-security-imperative.

70. National Telecommunications and Information Administration. "Dual-Use Foundation Models with Widely Available Model Weights Report." National Telecommunications and Information Administration, 30 Jul. 2024, https://www.ntia.gov/programs-and-initiatives/artificial-intelligence/open-model-weights-report.

71. Bommasani, Rishi, et al. Considerations for Governing Open Foundation Models. Issue Brief, Human-Centered Artificial Intelligence (HAI), Stanford University, 13 Dec. 2023, https://hai.stanford.edu/sites/default/files/2023-12/Governing-Open-Foundation-Models.pdf.

72. Raji, Timnit. "Godmother of AI Says California's AI Bill Will Harm U.S. Ecosystem." Fortune, 6 Aug. 2024, https://fortune.com/2024/08/06/godmother-of-ai-says-californias-ai-bill-will-harm-us-ecosystem-tech-politics/?abc123.

73. Gent, Edd. "Protesters Decry Meta's 'Irreversible Proliferation' of AI." IEEE Spectrum, 6 Oct. 2023, https://spectrum.ieee.org/meta-ai#:~:text=The%20self,how%20the%20AI%20is%20used.

74. La Malfa, Emanuele, et al. "Language-Models-as-a-Service: Overview of a New Paradigm and its Challenges." arXiv, 23 Nov. 2023, https://ar5iv.labs.arxiv.org/html/2309.16573v2#:~:text=et%C2%A0al.

75. Mozilla Foundation and Global Partnership on Artificial Intelligence. "AI Act: How the EU Can Address the Challenges Posed by General-Purpose AI Systems." OpenFuture, Jan. 2023, https://openfuture.eu/wp-content/uploads/2023/01/AI-Act_Mozilla-GPAI-Brief_Kx1ktuk.pdf; Denton, Jake. "The U.S. Shouldn't Go the Way of Europe on AI." The Heritage Foundation, 8 May 2024, https://www.heritage.org/big-tech/commentary/the-us-shouldnt-go-the-way-europe-ai.

76. Bengio, Yoshua, et al. "International AI Safety Report 2025." DSIT 2025/001, UK Department for Science, Innovation and Technology, 29 Jan. 2025, https://internationalaisafetyreport.org/publication/international-ai-safety-report-2025.

77. 中国人工智能产业发展联盟 . 18 家企业披露《人工智能安全承诺》实践成果，推动 AI 安全治理新进程 [EB/OL]. 2025-07-16. https://aihub.caict.ac.cn/ai_security_and_safety_commitments.

78. Amodei, Dario. "The Urgency of Interpretability". Anthropic, 27 Apr. 2025, https://www.darioamodei.com/post/the-urgency-of-interpretability.

79. Dunietz, J. , Latonero, M. and Roberts, K. "A Plan for Global Engagement on AI Standards", NIST Trustworthy and Responsible AI, National Institute of Standards and Technology, Gaithersburg, MD, 29 Apr. 2025, https://doi.org/10.6028/NIST.AI.100-5e2025, https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=960042.

80. 同前注 67.

81. Language Model Interpretability Team."Gemma Scope: Helping the Safety Community Shed Light on the Inner Workings of Language Models." Google DeepMind, 31 July 2024, https://www.deepmind.google/discover/blog/gemma-scope-helping-the-safety-community-shed-light-on-the-inner-workings-of-language-models/; Lindsey, Jack, et al. "On the Biology of a Large Language Model." Transformer Circuits Thread, 27 Mar. 2025, https://www.transformer-circuits.pub/2025/attribution-graphs/biology.html.

82. Jonker, Alexandra and Alice Gomstyn. "What Is AI Alignment?" IBM, 18 Oct. 2024, www.ibm.com/think/topics/ai-alignment. Artificial intelligence (AI) alignment is the process of encoding human values and goals into AI models to make them as helpful, safe and reliable as possible; Askell, Amanda, et al. "A General Language Assistant as a Laboratory for Alignment." arXiv, 9 Dec. 2021, arxiv.org/abs/2112.00861.

83. Russell, Stuart J. "Human Compatible: Artificial Intelligence and the Problem of Control. "Viking, 2019; Hadfield-Menell, Dylan, et al. "Cooperative Inverse Reinforcement Learning." arXiv, 9 Jun. 2016, https://arxiv.org/abs/1606.03137.

84. 曹建峰 . 人机对齐，通用人工智能的必由之路 [EB/OL]. 2024-11-01. https://mp.weixin.qq.com/s/iOOapnk9pbC8W5dHR5ALpA.

85. Burns, Collin, et al. "Weak-to-Strong Generalization: Eliciting Strong Capabilities With Weak Supervision. " arXiv, 14 Dec. 2023, https://arxiv.org/abs/2312.09390; Bai, Yuntao, et al. "Constitutional AI: Harmlessness from AI Feedback." arXiv, 15 Dec. 2022, https://arxiv.org/abs/2212.08073; Bricken, Trenton, et al. Building and Evaluating Alignment Auditing Agents. Alignment Science Blog, Anthropic, 24 July 2025, https://alignment.anthropic.com/2025/automated-auditing/.

86. Meinke, Alexander, et al. "Frontier Models are Capable of In-context Scheming." arXiv, 6 Dec. 2024, https://arxiv.org/abs/2412.04984; Greenblatt, Ryan, et al. "Alignment faking in large language models." arXiv, 20 Dec. 2024, arxiv.org/abs/2412.14093.

87. Lynch, et al. "Agentic Misalignment: How LLMs Could Be Insider Threats." Anthropic Research, 20 Jun. 2025, www.anthropic.com/research/agentic-misalignment.

88. de la Garza, Joel, et al. "What DeepSeek Means for Cybersecurity." Andreessen Horowitz, 28 Feb. 2025, https://a16z.com/podcast/what-deepseek-means-for-cybersecurity/.

89. International Dialogues on AI Safety. "IDAIS-Shanghai, 2025." Safe AI Forum, 22–25 July 2025, idais.ai/dialogue/idais-shanghai/.

90. Altay, Sacha, and Fabrizio Gilardi. People Are Skeptical of Headlines Labeled as AI-Generated, Even if True or Human-Made, Because They Assume Full AI Automation. PNAS Nexus, vol. 3, no. 10, 1 Oct. 2024, https://doi.org/10.1093/pnasnexus/pgae403.

91. Zhang, Hanlin, et al. "Watermarks in the Sand: Impossibility of Strong Watermarking for Generative Models." Harvard University, 23 Jul. 2024. arXiv, arxiv.org/abs/2311.04378.

92. Burgess, Matt. "Criminals Have Created Their Own ChatGPT Clones." WIRED, 7 Aug. 2023, https://www.wired.com/story/chatgpt-scams-fraudgpt-wormgpt-crime/.

93. Dong, Ziping, et al. "Imperceptible but Forgeable: Practical Invisible Watermark Forgery via Diffusion Models." arXiv, 28 Mar. 2025, https://arxiv.org/abs/2503.22330.

94. 胡欣红 .《滕王阁序》AI 率 100%，AI 检测到底靠谱吗 [EB/OL]. 2026-01-12.https://mp.weixin.qq.com/s/TeU3tNYPYSIp_FqCIvNQ3g?search_click_id=6505485447878598860-1768190031513-1031282429.

95. Sowmya, S., Sahana Karanth, and Sharath Kumar. "Protection of Data Using Image Watermarking Technique." Global Transitions Proceedings, vol. 2, no. 4, 2021, pp. 386–391, https://doi.org/10.1016/j.gltp.2021.08.035.

96. 赵精武 , 陈翊瑄 .“网络谣言”的老题新治 [J]. 求是学刊 , 2025(2).

97. Pierson, Brendan. "Mother Sues AI Chatbot Company Character.AI, Google over Son's Suicide." Reuters, 23 Oct. 2024, https://www.reuters.com/legal/mother-sues-ai-chatbot-company-characterai-google-sued-over-sons-suicide-2024-10-23.

98. Apple. "Apple Intelligence Comes to iPhone, iPad, and Mac Starting Next Month." Apple Newsroom, 9 Sept. 2024, https://www.apple.com/newsroom/2024/09/apple-intelligence-comes-to-iphone-ipad-and-mac-starting-next-month/.

99. O'Flaherty, Kate. "Apple's New AI Security Move Explained." Forbes, 22 Jun. 2024, https://www.forbes.com/sites/kateoflahertyuk/2024/06/22/apples-new-ai-security-move-explained/.

100. Bakare, Azeezat. "X Faces European Complaints Over Alleged Data Privacy Violations." JURIST, 13 Aug. 2024, https://www.jurist.org/news/2024/08/x-faces-european-complaints-over-alleged-data-privacy-violations/.

101. OpenAI. "Privacy Policy." OpenAI, 15 Dec. 2025, https://openai.com/zh-Hans-CN/policies/row-privacy-policy/.

102. BEUC – The European Consumer Organisation. Artificial Intelligence: What Consumers Say: Findings and Policy Recommendations of a Multi-Country Survey on AI. 3 Sept. 2020, www.beuc.eu/reports/artificial-intelligence-what-consumers-say-findings-and-policy-recommendations-multi.

103. Travers, Adelin. "PCC: Bold Step Forward, Not Without Flaws." Trail of Bits Blog, 14 Jun. 2024, https://blog.trailofbits.com/2024/06/14/pcc-bold-step-forward-not-without-flaws/.

104. 정두용 . "[단독] 우려가 현실로…삼성전자, 챗GPT 빗장 풀자마자 '오남용' 속출 " 이코노미스트 . 30 Mar. 2023, https://economist.co.kr/article/view/ecn202303300057?s=31.

105. Pierson, Brendan. "Musk Warns He Will Ban Apple Devices if OpenAI Is Integrated at OS Level." Reuters, 11 Jun. 2024, https://www.reuters.com/technology/elon-musk-says-he-will-ban-apple-devices-if-it-integrates-os-with-openai-2024-06-10.

106. Liu, Yang, et al. "Aligning Cyber Space with Physical World: A

Comprehensive Survey on Embodied AI." arXiv, 9 Jul. 2024, arXiv:2407.06886.

107. Klender, Joey. "Tesla Optimus Shows Off Its Newest Capability as Progress Accelerates." Teslarati, 2 Dec. 2025, https://www.teslarati.com/tesla-optimus-shows-off-its-newest-capability-progress-accelerates/.

108. OpenAI and MIT Lab Research. "Early Methods for Studying Affective Use and Emotional Well-Being on ChatGPT." OpenAI, 21 Mar. 2025, openai.com/index/affective-use-study/.

109. Long, Robert, et al. "Taking AI Welfare Seriously." arXiv, 4 Nov. 2024, arXiv:2411.00986v1, https://arxiv.org/html/2411.00986v1.

110. Butlin, Patrik, and Theodoros Lappas. "Principles for Responsible AI Consciousness Research." arXiv, arXiv:2501.07290, 13 Jan. 2025, https://arxiv.org/abs/2407.06886.

111. Anthropic. "Exploring Model Welfare." Anthropic, 24 Apr. 2025, www.anthropic.com/research/exploring-model-welfare.

112. 尚凯元 . 共促全球人工智能治理合作 [N/OL]. 人民周刊网 , 2025-02-14. https://www.peopleweekly.cn/html/2025/guoji_0214/239801.html.

113. 巴黎 AI 峰会中法德印等签署声明 美英拒绝签署 [N/OL]. 中国网 , 2025-02-12. http://news.china.com.cn/2025-02/12/content_117708436.shtml.

114. 王振宇 , 刘捷 . AI 发展须构建知识产权新秩序 [N/OL]. 光明网 - 时评频道 , 2025-09-23. https://guancha.gmw.cn/2025-09/23/content_38303931.htm.

115. Brittain, Blake, and Mike Scarcella."Anthropic Agrees to Pay $1.5 Billion to Settle Author Class Action." Reuters, 5 Sept. 2025, www.reuters.com/sustainability/boards-policy-regulation/anthropic-agrees-pay-15-billion-settle-author-class-action-2025-09-05/.

116. Brittain, Blake. "US Judge Preliminarily Approves $1.5 Billion Anthropic Copyright Settlement with Authors." Reuters, 25 Sept. 2025, www.reuters.com/sustainability/boards-policy-regulation/us-judge-approves-15-billion-anthropic-copyright-settlement-with-authors-2025-09-25/.

117. 王其冰 . 2025 年“人工智能造福人类全球峰会”聚焦 AI 治理和标准建设 [N/OL]. 新华网 , 2025-07-12. http://www.xinhuanet.com/20250712/2b2f2147068240e9b2a84034067e2f2e/c.html.

118. 周武英 . 从三大支柱看美国如何发展人工智能 [N/OL]. 经济参考网 , 2025-07-29. http://jjckb.xinhuanet.com/20250729/0a0065ab41b04151876bac85b45f306b/c.html.

119. 熊茂伶 . 美国发布 AI 行动计划 放宽监管措施 [N/OL]. 新华网 , 2025-07-24. http://www.xinhuanet.com/world/20250724/b698b6c72a6f458b82888244a363e6b1/c.html.

120. 上海市经济和信息化委员会 . 2025 年世界人工智能大会暨人工智能全球治理高级别会议成功闭幕 [N/OL]. 上海市经济和信息化委员会官网 , 2025-07-29. https://sheitc.sh.gov.cn/zxxx/20250729/0ad810a23f8640dbbb4ff85d9e30ffe4.html.

121. 中华人民共和国外交部 . 2025 年 7 月 29 日外交部发言人郭嘉昆主持例行记者会 [EB/OL]. 中华人民共和国外交部 , 2025-07-29. https://www.fmprc.gov.cn/fyrbt_673021/202507/t20250729_11679055.shtml.

122. 潘云召 . 联合国大会通过决议设立全球人工智能治理新机制 [EB/OL]. 新华网 , 2025-08-27. https://www.news.cn/world/20250827/da0a5a073a7941f5ab6c5a56ffc3a4fb/c.html.

123. 潘云召 . 联合国启动 " 人工智能治理全球对话 " 机制 [EB/OL]. 2025-09-26.http://www.news.cn/world/20250926/75f8db117a4f411ab709ce0e23c6be53/c.html.

124. 中央网络安全和信息化委员会办公室 . 2025 年世界互联网大会乌镇峰会人工智能技术创新与治理论坛举行 [EB/OL]. 2025-11-09. https://www.cac.gov.cn/2025-11/09/c_1764411950182277.htm.

125. 中央网络安全和信息化委员会办公室 . 2025 年世界互联网大会乌镇峰会人工智能赋能科学研究论坛举办 [EB/OL]. 2025-11-09 .https://www.cac.gov.cn/2025-11/09/c_1764453206543559.htm.

126. European Commission. "Commission Publishes the Guidelines on Prohibited Artificial Intelligence (AI) Practices, as Defined by the AI Act." Shaping Europe's Digital Future, 4 Feb. 2025, digital-strategy.ec.europa.eu/en/library/commission-publishes-guidelines-prohibited-artificial-intelligence-ai-practices-defined-ai-act.

127. European Commission. "The Commission Publishes Guidelines on AI System Definition to Facilitate the First AI Act's Rules Application." Shaping Europe's Digital Future, 6 Feb. 2025, digital-strategy.ec.europa.eu/en/library/commission-publishes-guidelines-ai-system-definition-facilitate-first-ai-acts-rules-application.

128. European Commission. "General-Purpose AI Code of Practice Now Available." Shaping Europe's Digital Future, 10 July 2025, digital-strategy.ec.europa.eu/en/news/general-purpose-ai-code-practice-now-available.

129. European Commission. "Learn More About the Guidelines for Providers of General-Purpose AI Models." Shaping Europe's Digital Future, 18 July 2025, digital-strategy.ec.europa.eu/en/news/learn-more-about-guidelines-providers-general-purpose-ai-models.

130. European Commission."Simpler EU Digital Rules and New Digital Wallets to Save Billions for Businesses and Boost Innovation." Press Corner – European Commission, 19 Nov. 2025, ec.europa.eu/commission/presscorner/detail/en/ip_25_2718.

131. ICC. 美批准英伟达对华“有限开口” 25% 销售分成与技术代差限制 [EB/OL]. 2025-12-09. http://www.iccsz.com/site/cn/News/2025/12/09/20251209060015377289.htm.

132. 陈思佳 . 特朗普签行政令限制各州 AI 监管规则引争议，" 少数寡头出卖了美国的未来 " [EB/OL]. 2025-12-14. https://www.guancha.cn/internation/2025_12_14_800402_s.shtml.

133. 王健 . AI 时代 知识产权司法迎来新挑战 [J/OL]. 民主与法制周刊 , 2025-07-18. https://bjzcfy.bjcourt.gov.cn/article/detail/2025/07/id/8908765.shtml.

134. 李国祥 . 确保人工智能安全、可靠、可控 [N/OL]. 光明日报 , 2025-05-12. https://wxb.xzdw.gov.cn/wlcb/wsznl/202505/t20250512_569745.html.

135. 中央网络安全和信息化委员会办公室 . 中央网信办部署开展 " 清朗·整治 AI 技术滥用 " 专项行动 [EB/OL]. 2025-04-30. https://www.cac.gov.cn/2025-04/30/c_1747719097461951.htm.

136. 中央网络安全和信息化委员会办公室 . 中央网信办深入开展“清朗·整治 AI 技术滥用 ” 专项行动第一阶段工作 [EB/OL]. 2025-06-20. https://www.cac.gov.cn/2025-06/20/c_1752129980667315.htm.

137. 曲忠芳 . 35 款 App 违法违规收集使用个人信息 AI 成“重灾区”[N/OL]. 中国经营报 , 2025-05-24. http://www.cb.com.cn/index/show/bzyc/cv/cv135252451645.

138. 方可 . 人民财评：让“人工智能”用得上、用得好、用得明白 [N/OL]. 人民网 - 观点频道 , 2025-08-28. http://opinion-app.people.cn/n1/2025/0828/c434885-40551673.html.

139. 年巍 . 筑牢高质量发展的智能基座 [N/OL]. 经济日报 , 2025-08-28. http://m.ce.cn/bwzg/202508/t20250828_2456383.shtml.

140. 陈恒星 . 北京互联网法院通报涉人工智能案件审理情况 [EB/OL]. 2025-09-11. https://bjgy.bjcourt.gov.cn/article/detail/2025/09/id/8982773.shtml.

141. 程学旗 . 新框架实现三个“转变”, 构建我国人工智能安全治理新格局 [EB/OL]. 2025-09-29. https://www.cac.gov.cn/2025-09/28/c_1760779757508049.htm.

142. 凌纪伟 . 警惕个别智能终端侵害用户权益 [N/OL]. 新华网 , 2025-11-17. http://news.fjsen.com/2025-11/17/content_32079873.htm.

143. 中央网络安全和信息化委员会办公室 .《人工智能生成合成内容标识办法》答记者问 [EB/OL]. 2025-03-14. https://www.cac.gov.cn/2025-03/14/c_1743654685896173.htm.

144. 中央网络安全和信息化委员会办公室 . 网信部门依法集中查处一批存在人工智能生成合成内容标识违法违规问题的移动互联网应用程序 [EB/OL]. 2025-11-25. https://www.cac.gov.cn/2025-11/25/c_1765795550841819.htm.

145. 中央网络安全和信息化委员会办公室 . 人工智能拟人化互动服务管理暂行办法（征求意见稿）[EB/OL]. 2025-12-27. https://www.cac.gov.cn/2025-12/27/c_1768571208101359.htm.

总顾问	司晓	腾讯公司副总裁 | 腾讯研究院院长
	杨健	腾讯公司副总裁 | 腾讯研究院总顾问
	杨乐	腾讯研究院副院长
指导专家	杜晓宇 | 柳雁军 | 周政华 | 袁晓辉 | 刘琼 | 冯宏声 | 田小军 | 朱开鑫	
研究团队	彭云 | 易镁金 | 张丽君 | 邓凯 | 曹建峰 | 王强 | 袁媛 | 吴朋阳 | 王鹏	
	徐一平 | 白惠天 | 刘莫闲 | 余一 | 曹士圯 | 李瑞龙 | 刘金松 | 王焕超	
	陆诗雨 | 张鸿茹 | 孙怡 | 胡璇 | 陈孟	
项目统筹	王融 | 钟雨霏 | 谢舒赫	