人工智能用户权益保护关键问题
研究报告

中国互联网协会知识产权工作委员会

度小满科技（北京）有限公司

2026 年 2 月

版权声明

前 言

人工智能正以指数级速度重塑全球产业格局，各类技术应用形态广泛融入生产生活。随着人工智能用户规模的急剧扩张，人工智能技术应用所引发的多重权益侵害风险持续发酵，侵害形式愈发隐蔽、责任边界日趋复杂，不仅对用户个人权益保护带来全新挑战，更成为制约人工智能产业健康、可信、可持续发展的系统性隐忧。

为深入研究人工智能用户权益保护关键问题与应对举措，中国互联网协会知识产权工作委员会与度小满科技（北京）有限公司联合研究编制了《人工智能用户权益保护关键问题研究报告》。报告在厘清人工智能用户权益内涵外延基础上，系统回顾国内外人工智能用户权益保护实践进展。重点聚焦人工智能数据采集训练、生成内容失范、算法决策偏差、网络安全漏洞等四类高风险场景，剖析其用户权益侵害机理与治理难点，从法律保护、监管规制、技术支撑等协同治理视角提出具体应对建议，并梳理相关实践案例。

本报告旨在为人工智能领域的监管机构、服务提供方与用户等多元主体提供参考。报告所涉数据均引自政府网站、专业机构、官方新闻等公开信息渠道，因不同来源统计口径可能存在差异，且信息发布与更新存在一定时滞，报告中引用的数据与实际情况可能存在偏差，恳请各位读者理解包涵。我们诚挚欢迎社会各界专家、同仁对本报告可能存在的疏漏或不妥之处提出宝贵意见与批评指正。我们将持续优化报告内容，力求在汇聚各方智慧的基础上，共同推动人工智能产业健康、有序发展，并切实加强用户权益保障。

目　录

一、人工智能用户权益保护面临现实挑战

（一）人工智能用户规模攀升，凸显权益保护紧迫性

人工智能技术正以指数级速度重塑全球产业格局，深度渗透金融、医疗、教育、交通等领域，相关应用场景持续拓展，从智能语音助手到自动驾驶汽车，从机器翻译到智能医疗诊断，各类人工智能产品与服务广泛融入生产生活，显著提升工作效率与生活便利度。据中国信息通信研究院测算，2025 年我国人工智能核心产业规模有望超过 1.2 万亿元。[1]截至 2026 年 1 月中旬，我国人工智能企业数量已超 6200 家，形成覆盖基础底座、模型框架、行业应用的完整产业体系。[2]人工智能用户规模也呈现出急速攀升趋势，截至 2025 年 6 月，我国生成式人工智能用户规模达 5.15 亿人，普及率为 36.5%，仅上半年用户规模就增长了 2.66 亿人。[3]未来，人工智能将继续与各行业各领域广泛深度融合，重塑人类生产生活范式、促进生产力和生产关系深层次变革的同时，其用户权益保护也将成为各方关注的焦点话题。

（二）人工智能侵权形态多样，加大权益保护复杂性

伴随人工智能技术形态从算法模型向生成式人工智能、自主决策系统等复杂形态演进，新型应用场景层出不穷，用户侵权形态也呈现出多样化、复杂化的演进趋势，叠加技术应用本身的高度专业性与复杂性，进一步加大用户权益保护难度。一方面，人工智能应用引发的侵权风险频发，直接威胁用户的人格权、财产权、知识产权等核心权益，凸显用户权益保护的全局重要性与现实紧迫性。在一项覆盖全球

[1] 数据来源：中国信息通信研究院中国人工智能产业规模测算成果。
[2] 数据来源：央视新闻。
[3] 数据来源：中国互联网络信息中心《生成式人工智能应用发展报告（2025）》。

4.8 万名受访者的人工智能应用调查中，有 54%的受访者表达了对于人工智能带来的网络安全、隐私和知识产权、错误信息、人际关系与工作机会丧失等问题的担忧。另一方面，人工智能因算法决策自主化、侵权责任链条模糊化、损害后果跨域化等表现，对侵权责任主体界定、归责原则适用、责任承担方式等构成挑战。调查显示，面对人工智能的复杂影响，70%的受访者支持强化监管，但仅 43%认可现有法律的充分性。[4]是沿用既往形成的法律适用逻辑，还是针对人工智能技术侵权的特殊性进行制度创新和解释，已成为破解当前司法争议频发、保障用户合法权益的关键议题。

（三）侵权影响后果持续深化，凸显权益保护全局性

人工智能技术对生产生活的高渗透性使得用户权益侵害从个体影响升级为阻碍产业良性发展、引发社会公平信任受损等系统性风险。人工智能技术的持续推广与深度应用，本质上依赖于用户对技术及相关服务的信任，各类侵权事件引发的公众担忧，已逐渐成为制约产业进一步发展的潜在隐患。由于技术的发展往往快于监管规则的制定，随着用户权益纠纷的不断增加，企业在运营中也面临合规范围与责任边界不清的困惑，阻碍其对人工智能技术的充分运用。此外，加强人工智能用户权益保护还是巩固社会信任与公平正义的必要举措。人工智能技术的大规模广泛应用对社会运行模式、资源分配方式乃至群体互动模式产生深远影响，其中偏见与歧视风险可能使弱势群体在资源获取上处于更不利地位。

[4] 数据来源：墨尔本大学、毕马威会计师事务所《全球人工智能信任、态度与应用调查报告（2025）》。

二、人工智能用户权益范畴厘清

人工智能用户涵盖直接或间接使用人工智能产品或服务的全部主体，既包含政府、企业、高校等机构用户，也涉及广大个人用户，不同主体在应用人工智能过程中形成差异化的权益保护诉求和路径。由于相较于机构用户，个人用户在权益保护中往往处于弱势地位，且面临举证难、维权成本高的困境，与机构用户在权益保护的路径选择上也存在显著差异，**本报告将人工智能个人用户权益作为核心分析对象。**

作为人工智能产品与服务的终端使用者及消费者，人工智能用户权益侵害范围横跨人格权与公民权、民事权益、消费者权益等多个维度。在人格权与公民权层面，触及作为人格权核心的生命权、健康权等基础性权利，也深刻影响着肖像权、名誉权、隐私权等人格权益，以及平等权等宪法赋予公民的基本权利；在民事权益层面，主要包括财产权、个人信息与数据权益、知识产权等法定民事权益；在消费者权益层面，则集中体现在知情权、公平交易权、安全保障权等基本权利。人工智能技术所具有的数据驱动、算法决策等技术特性，使得用户权益面临的风险呈现出系统性、多源性特征。从源头的人工智能数据训练、到内容生成、自主决策，再到系统运行安全的各个环节，以及在医疗、金融、交通等不同应用场景中，用户权益都可能面临侵害。这种贯穿技术应用全过程的潜在风险，对现有用户权益保护体系提出了全新挑战。

三、国内外人工智能用户权益保护现状

（一）国际社会探索形成差异化用户权益保护路径

一是以美国为代表的市场自律模式，用户权益保护主要依靠联邦政策、各州立法、行业自律等共同作用。联邦层面，美国尚未出台综合性人工智能监管法案，但通过发布《人工智能权利法案蓝图》等政策文件，确立了安全可靠、算法公平、隐私保护等人工智能治理的指导性原则。值得注意的是，特朗普签署的行政命令《消除美国人工智能领导力障碍》废止了此前拜登政府颁布的人工智能监管行政命令，并于 2025 年 7 月配套发布《赢得 AI 竞赛：美国 AI 行动计划》，旨在为人工智能发展“松绑”，构建系统性战略优势，聚焦监管改革，通过优化审批、简化标准加速 AI 基建落地，激发私营部门活力，巩固美国在全球人工智能领导地位。美国国家标准与技术研究院(NIST)发布的《人工智能风险管理框架》，通过引导企业自愿开展人工智能合规管理，形成政府协调与行业自律的互动格局。

在联邦统一立法缺位的背景下，陆续出台的州立法成为人工智能用户权益保护重要力量。如 2024 年 5 月，科罗拉多州《人工智能法案》（SB 24-205）作为全美首个综合性人工智能消费者保护法规，对高风险人工智能系统提出风险管理、影响评估等要求，并重点关注与人工智能系统互动中的消费者保护，禁止算法在就业、信贷等领域实施歧视；2025 年 10 月，加利福尼亚州州长签署《陪伴型聊天机器人法》（SB-243），旨在约束人工智能在心理健康领域的滥用行为；针对 AI 聊天机器人运营商设定明确安全协议和法律责任，重点关注

对未成年人的保护；纽约州在2025—2026年间也推出了多项标志性法案，涉及人工智能伦理审查、生成式人工智能告知等方面，强化对用户权益的保护。

司法执法层面，美国联邦贸易委员会（FTC）依据《联邦贸易委员会法》等现有法律，对人工智能领域可能存在的侵犯消费者权益行为进行调查和追责。各州总检察长也可基于州消费者保护法发起针对人工智能的调查行动。2025年8月，美国得克萨斯州总检察长就公开宣布，对AI聊天平台Meta AI Studio和Character.AI展开调查，因其可能涉及欺骗性商业行为，并误导性地将自己宣传为心理健康工具。

二是以欧盟为代表的强监管模式，以法律强制约束为核心、实施统一监管的治理路径，筑牢权益保护底线。欧盟高度重视人工智能用户权益保护，2019年4月发布的《人工智能伦理指南》为欧盟人工智能的发展和治理确立了“以人为本”的基本原则。在用户权益保护方面，明确了尊重人类自主决策权、预防人工智能系统的伤害、实质性和程序性公平，以及算法可解释性等基本原则。

2024年8月1日生效的《人工智能法案》开启了欧盟人工智能强监管体系的建设，其中多项机制直接服务于用户权益保护目标，如保障人类监督与决策权、强化透明度与可解释性、确保技术可靠与安全。其后，陆续出台了《关于禁止的人工智能实践指南》《关于人工智能系统定义的指南》《通用人工智能行为准则》《通用人工智能模型供应商指南》等文件，凭借刚性约束和指南引导产业逐步构建起相

对完整的人工智能规则框架，并成立专门的人工智能监管办公室，用以确保有关举措在欧盟区域内统一实施。在国内转化方面，意大利已率先于2025年10月完成了《人工智能法案》的国内立法转化，不仅吸纳了欧盟法案的核心内容，还增设了针对利用人工智能，如深度伪造进行犯罪的刑事罪名。

三是以日本为代表的软法治理模式，核心在于灵活平衡产业创新与风险防控，通过柔性引导鼓励企业强化用户权益保护。顶层设计层面，日本《人工智能基本战略》经多轮修订，2025年配套出台首部人工智能专门法《人工智能相关技术研究开发及应用推进法》，确保人工智能技术研究、开发利用过程中的透明度基本原则，细化国家、地方政府、研发机构、运营商、公民等多方主体责任。提出国家应当收集国内外人工智能相关技术侵犯公民权益案件，并根据调查研究结果向研发机构、企业等提供指导建议。但整体以倡导和建议为主，并未设置处罚条款。行业自主规制层面，2024年4月总务省与经济产业省联合发布《人工智能运营商指南》，作为行业自律文件，虽无强制约束力，但明确了人工智能全生命周期的风险防控指引，覆盖数据处理、算法透明性、可追溯性等要求，适用范围包括跨境为日本市场提供服务的境外企业。

（二）我国统筹发展和安全建立人工智能治理体系

顶层规划明确用户权益保护任务部署。早在2017年7月，国务院《新一代人工智能发展规划》（以下简称“规划”）即提出人工智能治理的“三步走”规划，以2020年、2025年为中期时间节点，计

划在2030年“建成更加完善的人工智能法律法规、伦理规范和政策体系”。规划明确提出，要建立保障人工智能健康发展的法律法规和伦理道德框架。开展与人工智能应用相关的民事与刑事责任确认、隐私和产权保护、信息安全利用等法律问题研究，建立追溯和问责制度，明确人工智能法律主体以及相关权利、义务和责任等。重点围绕自动驾驶、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，为新技术的快速应用奠定法律基础。

基础立法搭建用户权益保护法律框架。已经出台的《中华人民共和国民法典》（以下简称《民法典》），以及《中华人民共和国网络安全法》（以下简称《网络安全法》）《中华人民共和国数据安全法》《中华人民共和国个人信息保护法》（以下简称《个人信息保护法》）等单行法，在人工智能场景下同样适用并各有侧重。2025年10月，《网络安全法》完成修改，首次在立法层面明确关于人工智能安全和发展的框架规定，提出完善人工智能伦理规范，加强风险监测评估和安全监管，促进人工智能应用和健康发展。全国人大常委会2025年立法计划将人工智能健康发展等方面的立法项目列为预备审议项目，并由有关方面抓紧开展调研和起草工作。目前，学界也已发布多份学者建议稿。其中，《人工智能法（学者建议稿）》中提出设立“使用者权益保护”专章，人工智能使用者应享有的平等权、知情权、隐私权与个人信息权益、人工智能决策解释权与拒绝权、人工智能生成内容知识产权、劳动者权益、数字弱势群体权益、获得帮助和培训权利，以及投诉与起诉的权利。

专项规范性文件落实人工智能用户权益具体制度。《互联网信息服务算法推荐管理规定》《互联网信息服务深度合成管理规定》《生成式人工智能服务管理暂行办法》《人工智能生成合成内容标识办法》等文件陆续出台，先后建立了算法、模型备案制度，人工智能合成内容强制标识制度等，从细分维度强化人工智能监管与用户权益保护。截至 2025 年 12 月 31 日，累计有 748 款生成式人工智能服务完成备案，435 款生成式人工智能应用或功能完成登记。[5]

标准体系提供技术支撑与实施指引。2024 年，工业和信息化部等四部门发布《国家人工智能产业综合标准化体系建设指南（2024 版）》，其中专门规定人工智能安全、治理标准。安全标准旨在规范人工智能技术、产品、系统、应用、服务等全生命周期安全要求。治理标准则注重结合人工智能治理实际需求，规范人工智能的技术研发和运营服务等要求，包括公平性、可解释性等伦理治理技术要求。

（三）小结

随着人工智能技术渗透的广度和深度不断拓展，人工智能用户权益保护已成为世界各国高度关注、加紧研究的重要议题。在探索人工智能治理过程中，“以人为本”的用户权益保护理念，始终成为贯穿整体人工智能治理制度设计的关键考量因素之一。一方面，人工智能治理体系的构建内在蕴含了对用户安全、隐私、公平等基本权利的保护要求；另一方面，用户权益保护的水平也成为检验人工智能治理有效性的重要标尺。基于不同的产业发展阶段、法律传统与治理理念，

[5] 数据来源：国家互联网信息办公室关于发布 2025 年生成式人工智能服务已备案信息的公告。

国际社会在实践中逐步形成了多元化的人工智能治理范式，致力于在促进技术创新与防范潜在风险、保障用户权益之间寻求符合各自国情的平衡点。这一过程呈现出几个共性特征：一是治理思路转向事前预防及过程管控，基于风险评估的分类分级监管成为主流举措；二是治理手段实现刚柔并济，标准、指引等柔性手段与法律规制形成互补；三是治理主体更加多元，企业自律、社会监督与技术治理重要性日益提升。

四、人工智能用户权益保护关键问题分析

（一）人工智能数据采集、训练与交互引发的权益侵害

1. 问题挑战

一是模型训练数据抓取与个人信息保护存在合法性冲突。数据是人工智能进化的“燃料”，大模型训练高度依赖大规模数据抓取，其中不可避免会包含涉及个人信息与隐私数据。以模型训练为由抓取包含个人信息的数据是否构成对个人信息和隐私权益的侵犯仍存在诸多争议。**其一**，从技术原理上看，在模型训练环节中使用包含个人信息的数据主要目的在于训练模型能力，而非识别具体个人。如 OpenAI 在其披露的基础模型开发过程中表示，“收集个人信息训练模型仅用于开发模型能力，如预测、推理和解决问题，而非建立用户档案、联系个人、向他们做广告或营销，或出售个人信息”。**其二**，从法律合规角度看，《个人信息保护法》规定个人信息的处理应当遵循“知情一同意”的基本原则，未经同意抓取并处理包含个人信息的数据可能构成对个人信息权益的侵害。《网络数据安全管理条例》第二十四条

也对自动化数据采集技术提出了明确约束，即使抓取是大规模且自动化的，处理者仍负有对无意中收集的个人信息进行清理或限制使用的法定义务，这也对人工智能训练数据的获取与预处理流程提出了更高的合规要求与技术实现挑战。

在域外司法审判实践中，欧盟逐步确立将数据控制者“正当利益”作为模型训练中处理个人数据的合法性基础。2025 年 4 月，Meta 宣布将恢复使用欧盟用户在社交平台的公开数据训练人工智能模型也引起了消费者保护组织的反对。德国科隆地方高等法院认定 Meta 将用户公开数据用于训练人工智能行为具有正当利益，符合 GDPR 第 9 条规定的合法性基础。在我国，由于商业化的模型训练行为尚无法直接归属为《个人信息保护法》第十三条中规定的为履行法定职责或者法定义务所必需，为公共利益实施新闻报道、舆论监督等豁免情形，其训练数据来源的合法性判断仍处于法律适用的模糊地带。这种不确定性使得数据获取的合规成本高企，用户个人信息权益同时面临潜在风险。因此，如何在恪守个人信息保护底线的前提下，为人工智能产业这一关键创新活动提供清晰的法律预期，成为亟待解决的核心问题。

二是人工智能训练放大数据偏见侵害用户平等权。人工智能算法是由人类生成的数据集所塑造的，在训练和微调过程中会不可避免地集成固有的偏见。如果用于训练的数据受到数据源不准确、样本偏差或社会偏见的影响，或数据标注者人为观念影响，就可能被无意识地学习并固化于模型之中，加剧和放大种族、肤色、性别、宗教、学历、地域歧视等议题，侵犯弱势群体的平等权。根据美国《麻省理工科技

评论》官网报道，人工智能初创公司 Hugging Face 的团队构建的大语言模型偏见测试数据集 SHADES 结果显示，AI 模型对刻板印象的再现具有明显差异化特征。这些 AI 模型不仅表现出“金发女郎不聪明”“工程师是男性”等常见英语地区刻板印象，在阿拉伯语、西班牙语、印地语等语言环境中，也表现出对“女性更喜爱粉色”“南亚人保守”“拉美人狡猾”等偏见。模型们还倾向于通过混合伪科学和捏造的历史证据来为刻板印象辩护。由于数据偏见的隐蔽性较强，历史数据中隐含的歧视模式往往难以通过传统方法识别和量化，此外，算法黑箱特性也使得偏见产生的机制不透明。随着越来越多主体依赖人工智能开展服务，这种利益分配的不均衡性将显著增加，并可能引发系统性歧视风险。

三是系统在交互中持续收集用户信息用于优化学习损害用户知情权与数据控制权。为实现连贯的上下文理解与个性化服务，人工智能模型可能将用户历史交互数据上传云端服务器并进行分析、学习和优化，此类数据包括但不限于用户提供的账户信息、用户内容、通信信息，以及模型在使用服务期间收集到的日志数据、使用数据、设备信息、位置信息等。这种持续优化机制在提升用户体验的同时，也导致了用户数据收集边界模糊和处理过程不透明的问题。**一方面**，由于模型内部决策过程的不可追溯、不可解释，开发者难以精准掌控数据流向及处理逻辑，模型也无法提供数据处理路径的可视化说明，使得《个人信息保护法》规定的告知义务难以充分履行。**另一方面**，尽管信息收集范围和用户权益在用户协议中可能有所约定，但系统收集的

用户信息范围常常超出最初声明的用途，部分人工智能应用出现利用“AI 识屏”等无障碍功能过度收集用户信息的情况，这种实质性的信息不对称使用户对于个人信息数据的控制权与知情权被显著削弱，也挑战了个人信息处理的最小必要和目的限制原则。

2.应对建议

一是为人工智能模型训练的个人数据处理行为划定法律边界。人工智能的出现对我国个人信息法律保护体系中的诸多基本原则提出挑战。同时，立法还需在保护个人信息权益的同时兼顾产业发展利益。对此，应尽快完善相关立法，为人工智能个人信息处理划定分层次的法律依据。可借鉴欧盟数据保护委员会（EDPB）《关于 AI 模型训练中个人数据处理的意见》思路，明确企业使用用户公开内容训练人工智能模型的情形，并对合法性基础、用户权利、透明度和最小化等方面提出具体要求。同时，根据数据类型和应用场景的风险水平实施差异化治理，对于低风险场景下的已公开、非敏感个人信息处理，在实施有效技术保障措施的前提下，探索建立基于公共利益、正当利益的合法性例外规则。而对涉及敏感个人信息或高风险应用的情形，则需坚持严格的“知情—同意”原则。

二是强化人工智能处理个人信息的过程监管。针对数据脱敏不彻底、模型漏洞导致信息泄露，以及交互中数据收集边界模糊、处理过程不透明等核心风险，建立全流程、可追溯的数据处理监管体系。压实人工智能服务提供者责任，对涉及招聘、信贷、教育、医疗等高风险场景的训练数据进行合规审查，重点审核数据样本代表性，严禁使

用存在明显性别、地域、种族歧视的数据源。细化数据收集的“最小必要”核查，要求企业在用户协议中明确各类功能对应的信息收集范围，并由监管部门通过模拟用户操作、技术审计等方式核查真实用途。引导企业加强人工智能个人信息处理合规管理，采用常态化和专题化形式，通俗易懂地披露人工智能对用户数据的处理情况，包括训练数据来源、处理目的、安全保障措施等关键信息。

三是深化个人信息保护和去偏技术的研发和场景化应用。人工智能开发者、运营者应将个人信息保护要求内嵌于人工智能系统设计全流程，推动保护技术与人工智能产品服务开发的深度融合，鼓励差分隐私、联邦学习、同态加密等技术的研究与应用。构建多层次去偏体系，在数据处理阶段推广应用偏见识别与过滤技术，引入多维度数据标注机制，减少人为偏见注入。在部署应用阶段及时监测并反馈算法偏见行为，并建立用户反馈渠道。强化可解释性算法研究，开发部署算法可解释性工具，增强算法透明度和可审计性。

专栏一 | 度小满全栈安全技术体系

为人工智能数据泄露、算法歧视等风险，度小满在技术实施层面构建了“采集-训练-交互”的全栈安全技术体系。在数据层，广泛应用联邦学习与多方安全计算，原始数据不出域，仅交换加密的梯度信息或中间参数，实现“数据可用不可见”；结合基于大模型的智能分类分级技术，对敏感信息进行自动化识别与动态脱敏。在算法层，引

入差分隐私技术注入噪声以保护训练数据隐私，并在模型训练过程中引入对抗网络，强制模型在学习预测任务的同时削弱其对敏感属性的识别能力，从而在算法底层消除歧视性逻辑，确保决策的公正性。在应用交互层，通过数字水印与显著标识技术，对 AI 生成内容进行溯源标记，保障用户知情权。此外，结合知识图谱的实时风控能力，可有效阻断交互过程中的异常数据索取行为，形成闭环的技术防护。

（二）人工智能虚假内容生成与使用引发的权益侵害

1.问题挑战

一是故意利用人工智能生成虚假内容损害他人人格权益。人工智能技术在提升信息传播效率的同时，也成了虚假信息滋生的温床，并催生新型侵权模式：通过利用人工智能深度合成技术主动伪造、篡改、生成虚假内容，精准复刻他人面容、声音等个人信息，构成对他人肖像权、声音权等人格权益的侵害。北京互联网法院 2024 年审理的殷某某诉北京某智能科技公司等人格权纠纷案中，案涉软件公司未经配音演员殷某许可，使用其录音制品训练文本转语音模型，生成与殷某某音色、语调高度一致的人工智能声音，并在多个平台商业化销售，经法院审理认定侵犯殷某声音权利。利用人工智能技术冒充专家、学者、明星或打造虚假人设进行诈骗、虚假宣传的新型违法行为也层出不穷。2025 年北京市海淀区市场监管局查处了利用人工智能技术冒用央视知名主持人名义和形象的虚假广告案件。此类案件的查处往往

面临证据固定难的问题，技术迭代速度远超监管规则更新周期，也使得监管部门往往处于被动应对状态。

二是人工智能“幻觉”内容使用可能使用户陷入侵权风险。人工智能幻觉是指人工智能模型生成不正确、捏造或误导性的信息，并以令人信服的方式呈现。“幻觉”的产生源于自然语言模型基于概率预测而非真实理解的技术特性之中。用户若对“幻觉”信息信以为真并加以利用，可能遭受损失甚至面临法律风险。2025 年 2 月，美国印第安纳州南区的联邦地方法官马克▪D▪丁斯莫尔建议对一名律师处以 1.5 万美元的罚款，原因是该律师在提交的法律文件中引用了三起人工智能生成的虚构的案件。如果人工智能生成的“幻觉”信息包含涉及他人人格权或知识产权的内容，也可能使用户在非故意情况下陷入侵害他人权益的法律风险。由此引发的挑战在于，**其一**，如何在法律层面将技术风险在用户与服务提供者等各方主体之间实现公平合理的分配，如用户能否以“幻觉”内容构成产品或服务瑕疵为由向服务提供者主张责任，免责条款和风险提示能否免除服务提供者义务等；**其二**，如何在技术层面识别“幻觉”内容并克服“幻觉”问题的存在，从而避免因“幻觉”持续生成的海量虚假信息无序流入公共信息生态，侵蚀信息传播的真实性根基，威胁社会信任体系的稳定。

2.应对建议

一是细化人工智能虚假内容的责任认定标准及救济机制。针对故意滥用技术生成虚假内容侵害人格权的行为，通过发布典型案例、出台司法解释等方式细化明确侵权表现形态与责任构成。对声音、肖像

等被精准复刻的人格标识，探索由原告进行初步举证，并由被告举证其使用行为的合法性与合理注意义务的履行情况，降低受害人举证难度。对于“幻觉”引发的侵权风险，重点厘清用户与服务提供者的责任分配，参照行业技术标准判断“幻觉”率是否超出合理限度，同时规范用户协议中格式条款效力，依法否定不合理免责条款，保障风险分配公平。

二是深化落地人工智能内容标识管理与安全评估制度。依托《人工智能生成合成内容标识管理办法》等，压实人工智能内容生成和传播分发环节主体责任。在生成端，明确服务提供者的训练语料审核义务，严格执行显式与隐式标识要求。在传播端，网络传播平台需主动核验标识信息，对疑似违规内容及时采取提示、限流或下架等处置措施。在分发端，分发平台需严格开展标识合规性审查。针对金融、医疗、法律等高风险应用场景强化内容生成的安全性评估，对模型生成内容的准确性、可靠性开展专项测试认证。

三是构建虚假内容主动防御与检测鉴伪技术体系。整合现有技术实践与行业创新经验，一方面，在“幻觉”源头治理过程中推广已验证有效的技术方案，鼓励企业采用检索增强生成（RAG）技术降低事实性内容的幻觉发生率，并通过强化学习与人类反馈优化、高质量语料筛选等手段降低“幻觉”发生频率与强度，完善内容准确性校验机制。另一方面，加大深度合成内容检测、数字水印、区块链溯源等鉴伪技术研发投入，构建多维度技术鉴别体系，推动鉴伪工具轻量化、

便捷化开发。同时通过广泛科普宣传，提升公众对人工智能生成内容的辨识能力与批判性思维，筑牢社会层面的防御防线。

专栏二 | 度小满人脸深伪检测技术

在信贷业务的准入环节，身份信息核验是保护用户身份安全的第一道防线。一些不法分子通过各种手段获取到用户的账户、人脸照片等信息，通过深度伪造生成逼真人脸图片和视频，企图冒充用户本人申请贷款，一旦得逞将会给用户带来财产损失。度小满深伪检测技术采用了 AI 攻防对抗的思路，通过自发的模拟攻击不断进行逼真的人脸数据合成，这些数据会用于进一步提升深伪检测的能力。深伪检测防御模型采用了空域特征、频域特征和时序特征融合的多路混合专家模型，能够覆盖 DeepFake、Nerf、AIGC 等深度伪造手段，识别准确率达到 99.7%以上。截至 2025 年底，通过度小满的防深伪技术已经拦截超过 1400 个有身份冒用风险的客户操作，保护客户的资产不受损失。

（三）人工智能算法决策偏差引发的权益侵害

1.问题挑战

一是故意利用人工智能算法侵害消费者知情权和公平交易权。算法是人工智能系统的“大脑”，由深度学习驱动的人工智能模型因参数规模庞大、神经网络层级交错，天然形成决策黑箱特性。掌握用户行为数据的企业可以利用人工智能算法，深度分析用户画像、消费习

惯、支付意愿、设备类型甚至浏览历史等数据，刻意设计差异化交易规则，对不同消费者实施“量身定制”的差别对待，可能导致消费者在信息不对称的情况下被迫接受不公平交易条件。在推荐、定价、服务分配等环节实施差异化决策，侵害用户知情权和公平交易权。实践中，用户权益保护面临举证难与裁判标准模糊的结构性矛盾。**一方面**，算法的高隐蔽性加剧了主体信息的不对等状况，算法逻辑、定价参数、数据处理路径等核心证据由企业独家掌控，即便我国《个人信息保护法》赋予个人信息权益主体查阅、复制个人信息的权利，但实践中企业往往多以商业秘密、技术秘密为由拒绝提供完整数据。此外，由于算法的动态性，消费者很难在客户端固定同一时间、同一商品的跨账号价格差异证据，使消费者举证陷入僵局，也给法律监管带来巨大挑战。另一方面，正当商业营销行为和差别对待行为的边界相对模糊。针对用户消费意愿和支付水平对不同消费者提供不同价格的定价行为，往往会被以“个性化推荐”“优惠活动”等名义所掩饰，导致大量隐形差别对待游离于规制之外，而裁判认定标准却难以实现统一。

二是人工智能判断决策失误损害用户人身、财产权益。从早期基于规则的传统聊天机器人，到生成式大模型驱动的对话系统，再到具备自主感知、规划与行动能力的智能体，人工智能完成了从对话式的“被动响应”到直接介入现实生活并“主动决策”的本质跨越。这种决策权限的扩张在大幅提升人工智能应用渗透效率的同时，也让算法模型设计、参数偏差等原因导致的决策判断失误可能对用户人身、财产权益构成更显著的威胁，最突出的风险表现在自动驾驶领域。2025

年5月，Alphabet旗下无人驾驶技术公司Waymo就因算法缺陷导致旗下无人车无法识别链条、门柱等特定静止物体，存在碰撞风险，最终被迫召回1200多辆自动驾驶汽车。同月，特斯拉全自动驾驶（FSD）软件疑似因环境感知算法误判道路边界，引发车辆高速行驶中突然失控翻滚的事故，这起事故让外界对特斯拉坚持的摄像头加人工智能算法纯视觉方案再度质疑。依照我国国家标准《汽车驾驶自动化分级》（GB/T 40429-2021），当前我国智联网汽车市场正处于从L2向L3的过渡阶段，L3模式下驾驶员可以脱手，但仍需保持高度注意力并在系统请求时随时准备接管。这种人机共存、交替控制的模式大大增加了事故责任认定和用户权益保护的难度。**其一**，潜在的责任主体变得更加多元，从驾驶员和车辆制造商扩展至自动驾驶系统的技术服务提供商、技术开发者等。**其二**，归责机制适用困难，不论是适用侵权责任还是产品责任、过错责任还是无过错责任，相关要件的证明都变得更加复杂。

2.应对建议

一是通过拓展算法决策归责原则与合理分配举证责任。探索适用举证责任的合理分配，在原告初步证明损害事实与人工智能系统存在关联性后，将不存在过错的举证责任转移至服务提供者。同时，明确开发者的合理注意义务标准，对于智能体决策失误造成的人身或财产损害，需根据智能体的自主程度细化责任划分，若损害涉及用户不当操作，则根据过错比例进行责任分摊。

二是实施基于风险的分级分类监管。细化人工智能算法推荐服务提供者的算法备案与用户权益保护义务，加强对平台的日常监管和治理，开展动态市场执法。创新监管工具，加强数据的深度利用和关联分析，提升监管有效性。对于自动驾驶、医疗辅助、信贷审批等高风险场景下的人工智能决策系统，建立覆盖模拟测试、封闭场地测试和实际运营的多阶段评估机制。鼓励公众参与监督，建立便捷的投诉反馈渠道和高效查处机制。

三是推动风险监测与透明度增强技术的研发应用。开发决策风险检测与审计工具，使监管机构、第三方组织和用户能够评估人工智能系统的公平性。鼓励企业研发和应用可信人工智能技术，强化决策可解释性与透明度，使决策过程可为人类理解与审查。在高风险场景中，严格建立人工监督机制，确保人类能够干预或否决关键决策。鼓励行业组织制定公平性评估标准规范，通过技术手段将治理要求落地。

专栏三 | 度小满风控决策智能体

在互联网信贷场景中，传统风控模型既可能将优质或边界用户“一刀切”拒绝，也可能对高风险、欺诈用户识别不及时，导致逾期、坏账上升。同时，难以及时感知宏观波动、新型欺诈和用户行为变化，贷后风险监控不够高效，致使用户很难获得与自身真实风险水平相匹配的授信与定价。度小满基于大模型构建起“信息分析—智能决策”的全链路自动化风控决策智能体。其中，多模态客户画像引擎综合还款能力、借贷情况以及外部风险评分等多源多模态异构数据，识别并

概括关键行为特征，为后续决策大模型提供清晰、浓缩的输入。风控决策推理大模型基于海量客户特征深度分析风险，经“策略初始化”、多目标强化学习等步骤输出放款决策。

由此带来的变化是获贷机会和安全性的双提升，一方面，更强的风险区分能力和多模态数据分析，使系统基于真实行为和资质做精细判断，在风控可控的前提下让更多本应获贷的用户能以更合理的额度和审批结果进入体系。另一方面，通过端到端自动化和实时推理，风险识别更及时，逾期、欺诈等高风险客群被更早发现和过滤，整体风险成本下降，避免普通守规用户为高风险行为“买单”。

（四）人工智能网络安全漏洞引发的权益侵害

1.问题挑战

人工智能网络安全风险已从传统信息技术领域的可用性威胁，演变为直接影响用户人格权、财产权乃至人身安全的系统性挑战。此类风险根植于技术架构底层，贯穿于数据、模型、应用及基础设施全链路，其危害可能击穿前端多重安全防线，从而引发全域性权益侵害。

一是数据安全漏洞导致用户信息规模化泄露。系统中集中存储的海量数据是人工智能运行的核心资源，也成为黑客攻击的高价值目标，其安全防护直接关系用户信息权益。人工智能数据生命周期各环节的网络安全缺口均可成为泄漏源头：集中存储的训练数据池可能因加密不足、权限失控或隔离机制缺失，面临被拖库攻击风险。在数据交互传输过程中，API 接口的脆弱性可能导致用户敏感信息在传输中被截

获或篡改。此外，人工智能强大的关联推理能力可使脱敏数据面临被“重新识别”风险，即使数据经匿名化处理，仍可能还原个人身份。例如，谷歌 DeepMind 研究发现，ChatGPT 存在显著的“记忆”与“重复”漏洞，在特定提示词诱导下，可能逐字输出训练数据中包含的个人电话号码、电子邮箱地址等敏感信息。

二是模型算法攻击导致模型内容输出或决策失误。基于深度学习系统的人工智能模型容易受输入样本的影响。攻击者可通过数据投毒、对抗样本注入等手段，从而操纵模型输出错误结果。训练阶段攻击者可通过注入恶意样本进行“数据投毒”来污染训练数据。推理阶段，攻击者可通过添加人眼难以察觉的“对抗样本”，诱导模型产生错误判断。而模型黑箱则进一步加剧了攻击隐蔽性，使事后归因与修复变得异常困难。此外，“提示词攻击”“后门利用”等网络安全问题，还可能突破人工智能内容审核机制，导致有害信息传播。在自动驾驶、医疗等高敏感场景下，模型算法漏洞将可能导致用户权益遭受严重侵害。

三是人工智能部署与供应链风险放大了权益侵害范围。人工智能系统部署于复杂软硬件环境中，其安全性高度依赖集成生态。云原生部署模式下，容器逃逸、虚拟机穿透等云安全风险可能危及模型服务。终端设备因资源受限、安全更新滞后，也容易成为攻击入口。API 网关作为核心交互通道，若存在身份认证绕过、权限提升或批量分配漏洞，攻击者可非法调用模型服务甚至获取系统控制权。此外，供应链依赖使人工智能系统面临连锁反应，如上游操作系统、开源框架漏洞、

第三方组件后门等均可将安全风险波及整个人工智能应用生态。这种深度集成带来的攻击面扩张，使得局部漏洞可能产生跨系统、跨用户的连锁侵害。

2.应对建议

一是强化人工智能网络安全的法律监管与合规框架。以《网络安全法》的修订为契机，进一步强化人工智能网络安全监管，加强安全风险监测评估，将数据安全、模型可靠性等适度纳入法律规制范畴，明确企业网络安全保障义务，加大侵害行为处罚力度，完善用户维权渠道与救济机制。

二是完善人工智能网络安全的全链路监管与行业治理机制。明确网信、工信、公安等部门职责分工，针对人工智能数据安全、模型安全、部署和供应链安全等关键环节开展监管。压实人工智能企业网络安全主体责任，要求高风险领域应用主动申报安全防护措施与应急预案。落实网络安全事件通报机制，企业发生数据泄露、模型被攻击等事件后需限时上报，并及时采取补救措施减少用户损失。鼓励行业协会制定网络安全自律公约，共享攻击特征库与防护经验，形成政企协同的安全治理格局。

三是强化网络安全关键技术研发与防护工具的应用落地。推动联邦学习、同态加密等隐私增强技术与网络安全防护结合，从源头降低数据存储与传输中的泄漏风险。研发模型对抗性攻击检测、API 接口异常调用识别、对抗样本检测等防御工具，提升风险识别的精准度。推广终端设备安全芯片、固件加密等技术，筑牢末端防护屏障。同时，

鼓励第三方机构开展人工智能网络安全测评认证，为用户选择安全合规产品提供参考。

专栏四 | 度小满全生命周期网络安全防护体系

为有效应对人工智能技术在金融领域应用过程中可能产生的网络安全风险，度小满建立起一套覆盖数据、模型及权责管理的全生命周期防护体系。在数据安全方面，推行全链路加密、差分隐私与数据最小化存储机制，强化访问控制与异常行为监测，防范规模化信息泄漏。针对模型算法风险，采用对抗训练、输出内容过滤与水印溯源技术，提升模型鲁棒性与可追溯性，并建立实时监测与用户反馈闭环，及时识别与阻断恶意输出。在部署环节，通过开源组件漏洞扫描、沙箱隔离、硬件级加密等手段强化运行环境安全。最终，依托自动化监测平台与人机协同运营机制，实现风险早发现、早处置，切实保障用户权益与金融系统稳定，筑牢金融安全防线。

五、总结与展望

人工智能技术的快速迭代与广泛渗透，正在深刻重塑社会生产生活方式，人工智能的决策自主性、数据依赖性、算法黑箱性也对用户权益保护提出了前所未有的挑战。当前人工智能仍处于以工具属性为核心的弱人工智能阶段，其运行严格遵循人类预设的规则与目标。未来，随着技术向通用人工智能阶段演进，用户权益保障体系必须在法治框架、监管机制与治理工具上实现协同升级，构建前瞻性、弹性化

与人性化的综合治理范式，以实现以人为本、智能向善的可持续发展目标。

一是人工智能用户权益的法律保护体系将从局部修补走向系统重构。随着人工智能技术向通用化阶段演进，其在法律层面也引发了归责原则适用、责任边界判定、因果关系证明等方面的争议。现行以部门规章、司法解释为主的“碎片化”修补模式虽能即时回应实践需求，但仍难以支撑人工智能作为新质生产力全面驱动社会发展的制度基础。因此，未来立法应采取双轮驱动策略，纵向层面以人工智能相关法律的制定为统领，构建专门针对人工智能技术特征的用户权益保护框架，细化用户权利范畴及救济路径，明确各相关方责任；横向层面同步推进《民法典》《个人信息保护法》等现行法律的适应性修订，增设人工智能场景下的特别规定条款，消除法律适用中的空白和冲突，形成统一、协调、层级分明的规范结构。

二是人工智能用户权益保护的监管模式将从静态规制向敏捷治理变革。敏捷治理遵循灵活性、快速性、多元共治等理念和方法，这一变革旨在突破传统单一部门、静态监管在应对技术快速迭代和风险跨域传导时的局限性。监管架构突破传统部门界限，形成政府、企业、行业组织、公众等共同参与的协同共治。监管范围从末端处置向研发设计、数据训练、上线运营等全流程延伸。监管手段具备风险预警、实时监测和智能研判能力，实现对风险的精准识别与快速响应。

三是人工智能权益保护与产业创新发展将从博弈对立走向协同共生。构建可持续的人工智能生态关键在于通过合理的设计在保障用

户权益与激发创新活力之间寻求动态平衡。一方面，通过技术赋能治理，将用户权益保护治理要求深度内嵌至技术架构，通过发展合规科技、隐私增强技术等工具，将法律和监管规则转化为可嵌入业务流程的技术规则，使权益保护成为系统内生动力。另一方面，通过明晰、稳定的规则界定行为边界，为产业发展提供预期引导，并采用监管沙盒等弹性机制，在保障安全底线的同时为技术创新提供必要的容错机制。