CAICT 中国信通院

中国人工智能产业发展联盟
Artificial Intelligence Industry Alliance

集智 专题报告

面向产业的算法治理研究

——构建可信赖的算法治理路径

（2025 年）

中国信息通信研究院人工智能研究所

中国人工智能产业发展联盟安全治理委员会

2026年2月

版权声明

前　言

国务院《关于深入实施“人工智能+”行动的意见》明确提出，要“推动模型算法、数据资源、基础设施、应用系统等安全能力建设，防范模型的黑箱、幻觉、算法歧视等带来的风险，推动人工智能应用合规、透明、可信赖”。当前，人工智能大模型技术加速渗透千行百业，算法已从互联网平台技术底座跃升为驱动数字经济高质量发展、培育新质生产力的核心生产要素，深刻重塑着产业形态、劳动关系与社会治理模式。同时，算法“黑箱”、数据滥用、大数据“杀熟”等问题也随之凸显，进而对个体权益保障、数字经济可持续发展提出挑战。在此背景下，算法治理已非单纯技术议题，而是关乎经济发展、社会公平、产业生态的重大治理命题，成为落实“人工智能+”发展战略、推动平台经济从“扩规模”向“优质量”转变的关键抓手。

面对算法治理的困境与挑战，亟需兼具理论深度与实践可落地的产业自律方案。中国信息通信研究院与中国人工智能产业发展联盟安全治理委员会，联合含头部平台企业、高等院校、科研院所在内的共 14 家单位，召开多轮“构建可信赖的算法治理路径”研讨会，广泛吸纳各方智慧，系统梳理全球实践，剖析算法在技术、规则、平台层面的核心矛盾，最终形成本报告。报告提出构建“从合规驱动迈向信任驱动”的可信赖算法治理产业自律框架，锚定个体权益保障与公共价值维护双重目标，推动治理重心从技术层向规则层与平台层迁移，围绕公开透明、信息保护、公平公正、内容保障四大

支柱，形成技术、规则、平台三位一体的治理体系。

算法治理绝非一日之功，需长期弘扬和践行“算法向善”的理念。一方面，需依托标准化路径夯实治理根基，以公开透明为核心，围绕信息保护、公平公正、内容保障三个基本方面，构建算法治理的技术规范、管理标准与评测体系；另一方面，需将治理要求贯穿算法全生命周期，强化事前识别、事中防控、事后处置的动态管控，完善政府、产业、公众多元参与的共治格局，推动算法从“工具理性”向“价值理性”转变。

最后，本报告的完成离不开各参编单位及专家的鼎力支持与深度参与。在此，谨向中国社会科学院法学研究所、中国人民公安大学、对外经济贸易大学、中国政法大学、清华大学、中国人民大学、中国科学院信息工程研究所、美团研究院、阿里研究院、腾讯研究院、抖音、滴滴、高德、马上消费等所有为报告编制提供智慧贡献的单位与专家表示衷心感谢！

目 录

图 目 录

一、算法概述：算法治理的背景与提出

（一）算法技术赋能关键应用场景

随着新一轮科技革命和产业变革深入发展，算法正成为驱动互联网平台经济和新质生产力发展的重要力量。一方面，算法技术迈入以人工智能大模型为主导的时代，促进了互联网平台应用的蓬勃发展。从衣食住行的内容推荐到骑手、网约车的调度决策，从跨领域、跨平台的信息检索再到文本、图像、音视频的生成合成，算法已成为互联网企业平台的技术底座，也是中国互联网世界的“隐形舵手”[1]。**另一方面，**深度学习算法与先进生产要素深度融合，为新质生产力的加速形成与发展注入了强大动能。当前算法技术与大数据、人工智能等前沿科技交互协同，展现出自学习、自训练、自优化等特征，在促进产业升级和推动数字经济高质量发展等方面展现出巨大潜力[2]。

作为智能技术创新演进的关键所在，算法赋能效应在多个关键场景中得到充分释放。一是信息服务算法实现内容延展，精准的个性化推荐、排序精选和检索过滤算法成功助力我国数字服务“走出去”。以抖音、小红书等为代表的中国互联网企业，凭借其出色的 Wide & Deep 模型、NoteLLM 等推荐算法，打造出具有全球影响力的内容平台。**二是资源调度算法显著增质提效，**高效的订单匹配与路径规划算法为快速崛起的即时配送行业提供关键支撑。根据商务部国际贸易经济合作研究院报告指出，2024 年我国即时零售市场规模接近 7800 亿元，预计到 2030 年将超过 2 万亿元[3]。**三是生成合成算法促进创新迸**

[1] 人民财经网. 平台算法从“黑箱”走向透明[EB/OL]. 2025-04-29. https://www.peoplecen.com/html/2025/chany_0429/18897.html

[2] 新华网. 加快形成以人工智能为引擎的新质生产力[EB/OL]. 2024-10-30. https://www.news.cn/politics/20241030/3d80399f6a2442538e0925c7a7eb1b37/c.html

[3] 商务部国际贸易经济合作研究院. 即时零售行业发展报告（2024）[EB/OL]. 2024-11. https://caitec.org.cn/upfiles/file/2024/11/20241213113011188.pdf

发，以生成式人工智能（AIGC）为代表的深度合成算法为数字文化产业注入新的活力。越来越多的企业和个人开始使用该技术辅助甚至主导动漫制作、游戏开发和广告设计，不仅大幅提升了内容生产的效率与多样性，降低了创作门槛，还催生出虚拟数字人、人工智能绘画等新兴业态，推动文化产业进入智能化生产新阶段。

（二）当前算法治理面临多重挑战

算法技术在推动经济社会智能化转型的同时，也放大了既有挑战，并带来新的风险形态。一方面，算法技术加剧传统风险。算法的非线性技术特性，让决策过程缺乏透明度，加剧“问责难”困境；数据大规模采集与跨场景流转，放大传统隐私泄露风险；而模型训练依赖现实数据，其自带的偏差可能固化社会歧视，让不平等结构更难被打破。**另一方面，**算法技术带来新兴风险。算法推荐机制通过强化偏好营造“信息茧房”，重塑信息传播与认知模式；算法代替个人进行资源调度决策，间接影响了新就业形态劳动者权益；大数据与人工智能技术的精准画像能力，进一步催生出“杀熟”等新型市场不公行为。

面对新旧交织的风险，各国政府与产业界正积极探索协同治理路径。2025 年 3 月，中国网络媒体论坛“坚持主流价值导向 推动算法向上向善”主题分享会在广西南宁举行，多家平台企业在政府部门和主流媒体的见证下集体签署《算法向善南宁宣言》[4]。2024 年 6 月，在美国儿童权益组织和多州检察长联盟的舆论影响下，纽约州通过国内首部《防止成瘾性内容剥削儿童法》，限制平台算法向 18 岁以下

[4] 国家网信办. “坚持主流价值导向 推动算法向上向善”主题分享会在广西南宁举行[EB/OL]. 2025-03-30. https://www.cac.gov.cn/2025-03/30/c_1745040836302366.htm

的用户提供成瘾性内容[5]。2025 年 1 月，在《通用数据保护条例》《数字服务法》等框架指引下，12 个欧盟及英国民间组织呼吁 Deliveroo、JustEat 等外卖平台提升算法透明度并完善骑手权益保障机制，形成政府监管与公众参与的治理闭环[6]。这些举措集中反映出，算法治理亟需从单一监管模式转向政产学研联动的多元共治新范式。

算法技术的赋能效应与风险挑战要求以包容审慎的视角看待其发展演进，并深入探讨实际应用中的多重挑战。一是平台收益与用户权益之间的平衡问题。算法在追求效率最大化的过程中，若缺乏伦理与合规引导，可能带来隐私泄露、“信息茧房”等社会议题，与社会倡导的价值理念相冲突。**二是**新兴生产力的快速成长与既有制度、社会结构之间的适配问题。算法作为数字时代的重要生产要素，在催生新业态、提升经济效率的同时，也推动传统产业组织方式、劳动关系和市场格局的重塑，对劳动者权益保护与社会治理范式提出了新的要求。**三是**产业治理能力与公众信任之间的感知差距。虽然众多平台企业已建立算法备案、风险评估和伦理审查机制，但公众对算法透明度与可解释性的认知仍在形成中，信息沟通与社会信任的建设仍是一项长期工程。

[5] Practical Law. New York Enacts Laws to Protect Children's Online Privacy and Restrict Addictive Social Media Platforms[EB/OL]. 2024-06-21. https://uk.practicallaw.thomsonreuters.com/w-043-6862?transitionType=Default&contextData=(sc.Default)&firstPage=true

[6] Privacy International Time to deliver answers: An open letter to Just Eat Takeaway, Uber and Deliveroo[EB/OL]. 2025-01-13. https://privacyinternational.org/advocacy/5509/time-deliver-answers-open-letter-just-eat-takeaway-uber-and-deliveroo

（三）构建面向产业的算法治理路径

全面应对效率与公平、新旧生产关系和公众信任感知差距三组挑战，是构建高质量算法治理体系的重要前提。本报告基于产业视角，将算法治理定义为：**产业主体通过技术优化、规则完善与平台治理等方式，构建从合规驱动迈向用户驱动的可信赖算法实践体系。**这一定义强调产业界在算法治理中的主体作用，聚焦于深度学习驱动的主流平台应用算法，包括个性化推送、排序精选、检索过滤、调度决策以及生成合成五类典型算法，旨在实现效率与公平、创新与安全的动态平衡，促进新兴生产力与社会秩序的良性互动，并逐步增强公众对算法技术的信任与理解。

构建面向产业的可信赖算法治理路径，需协同推进算法规制、算法透视与算法实践。算法规制强调制度规范促进算法产业发展，聚焦我国、美国与欧盟等主要经济体，通过完善立法、规则与监管体系，保障算法开发与应用的安全、透明与可控。**算法透视**关注平台、用户与生态的协同共生，通过厘清算法在技术层、规则层与平台层的治理逻辑，探索技术创新与价值向善的平衡。**算法实践**侧重产业主体的责任承担与主动治理，通过强化公开透明、信息保护、公平公正与内容保障，构建可信赖的算法治理产业自律框架。三大维度相互耦合，共同构成面向产业的可信赖算法治理路径，为数字经济高质量发展提供可持续治理方案。

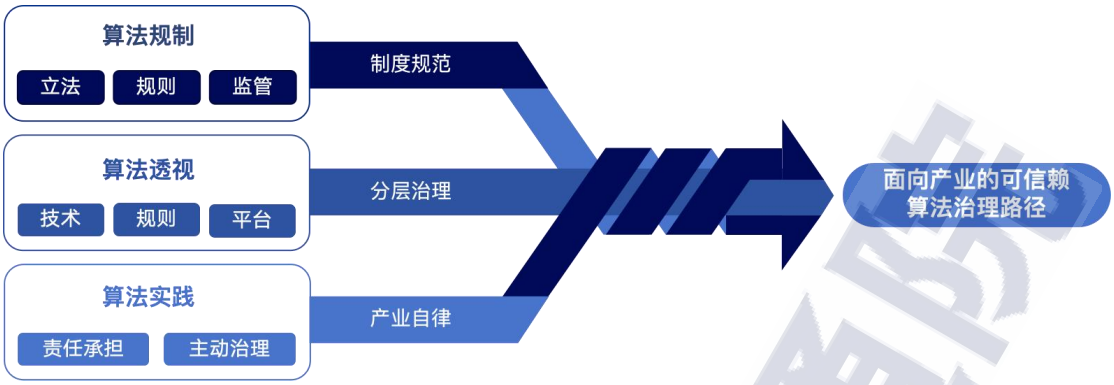

图 1 面向产业的可信赖算法治理路径

二、算法规制：全球算法治理的监管举措

（一）我国统筹发展和安全，开展全生命周期治理

我国注重通过制度设计保障多元权益，以备案、评估为抓手，形成事前、事中、事后全覆盖的管理机制。

权益保障层面，一是强化个人信息保护。2021 年 11 月生效的《中华人民共和国个人信息保护法》规定，自动化决策应确保透明、公正，防止对个人权益造成不当影响。**二是**保障用户自主选择权以及多方合法权益。2022 年 3 月发布的《互联网信息服务算法推荐管理规定》提出，用户有权了解算法运行的基本逻辑并自主关闭推荐功能，并明确了算法推荐服务提供者应保障未成年人、老年人、劳动者，以及消费者的合法权益；同时，2025 年 12 月发布的《人工智能拟人化互动服务管理暂行办法（征求意见稿）》要求，拟人化互动服务的提供者和使用者不得通过算法操纵、信息误导、设置情感陷阱等方式，诱导用户作出不合理决策。**三是**完善公众救济渠道与反馈机制。2023 年 7 月发布的《生成式人工智能服务管理暂行办法》与 2022 年 11 月发布的《互联网信息服务深度合成管理规定》等文件均要求服务提供者建立便捷高效的申诉与反馈渠道，推动用户参与算法治理过程。

管理机制层面，一是强化算法备案制度。针对具有舆论属性或社会动员能力的算法，《互联网信息服务算法推荐管理规定》要求算法推荐服务提供者主动进行备案，并提交算法类型、自评估报告等信息，充分保障算法的透明度与可追溯性。**二是**采取分类分级管理措施。《互联网信息服务算法推荐管理规定》明确建立算法分级分类安全管理制度，要求根据算法推荐服务的舆论属性或者社会动员能力、内容类别、用户规模等内容实施分类分级管理。**三是**建立风险评估与算法审核机制。《互联网信息服务算法推荐管理规定》提出算法推荐服务提供者应定期审核、评估、验证算法机制、模型、数据和应用结果等，构建算法安全与可验证的技术闭环；同时，《人工智能拟人化互动服务管理暂行办法（征求意见稿）》要求提供者落实安全主体责任，建立算法机制机理审核与科技伦理审查制度。**四是**定期开展专项治理与巡察整改。2024 年 11 月，中央网信办等四部门联合开展“清朗·网络平台算法典型问题治理”专项行动，聚焦劳动者权益、算法向善、内容多样性等关键议题，发挥监管督导作用；同时，2025 年 6 月，上海市网信办依法约谈某企业负责人，要求平台健全生成合成内容审核机制，提升技术把关能力，加强涉未成年人不良内容的整治清理。

（二）美国保障创新活力，赋权个体与约束公权力

美国以创新优先为核心，兼顾安全与伦理，针对算法市场行为与公权力应用提出合规要求，形成算法治理的双重保障机制。

权益保障层面，一是赋予信息知情权，保障劳动者的平等就业权。2025 年 9 月加州参议院通过《反机器人老板法案》[7]，规范了雇主如何使用自动化决策系统做出与雇佣相关的决定，为劳动者提供透明度

[7] 该法案于 2025 年 10 月 13 日被加利福尼亚州州长加文·纽森否决。

和保护。**二是**约束算法推荐应用，防范未成年人陷入“信息茧房”。2024 年 6 月纽约州通过的《防止成瘾性内容剥削儿童法案》规定，社交媒体平台不得向 18 岁以下用户推送由算法筛选的成瘾性内容；同年 3 月，佛罗里达州立法要求基于算法推荐机制的社交媒体平台不得为 14 岁以下用户开设账户，以防算法引发未成年人沉迷风险。**三是**赋予用户解释权，保障算法决策透明度。2024 年 5 月通过的《科罗拉多州人工智能法案》首次在州级层面明确“算法歧视”的法律定义，要求企业在自动化决策中避免基于种族、性别、年龄等差异的不公结果，并赋予消费者对算法决策结果的解释权。

管理机制层面，提出公平性评估、透明度报告等义务要求，约束公权力机关算法应用。2024 年 12 月通过的纽约州《政府自动化决策的立法监督法案》以及相关法律规定，政府机构在采用算法工具前进行风险与公平性评估，并定期公开评估结果。自 2019 年起，纽约市每年发布算法合规报告，详细列出各部门使用算法系统的类型、功能与应用场景，增强公众透明度。

（三）欧盟以权利为核心，打造强监管合规框架

欧盟将安全与规范嵌入数字化转型，形成了覆盖隐私保护、言论自由与公平公正等多元权益的保障框架。

权益保障层面，一是保护个人信息与数据安全。2018 年 5 月生效的《通用数据保护条例》以保护个人数据为宗旨，明确用户有权了解其个人信息是否被用于自动化决策，并有权要求人工干预与解释说明。**二是**保障个人的基本权利与透明知情权。2024 年 2 月生效的《数字服务法》规定，用户有权从用户协议中了解算法推荐的主要参数、逻辑及其潜在影响，从而确保平台运营符合公平、透明原则；同时，

作为《人工智能法》的配套文档，2025 年 7 月发布的《通用人工智能行为准则》提出，通用人工智能模型提供者在投放市场时，需向包括下游提供商在内的公众披露模型信息。**三是**保障劳动者、儿童和消费者等重点群体权益。2024 年 10 月生效的《改善平台工作条件指令》以保护数字劳动者权益为核心，要求平台向劳动者解释自动化决策结果的依据与影响；同时，《数字服务法》规定，平台必须采取有效措施为未成年人提供高水平的隐私、安全和保护，包括默认启用隐私模式和完善的年龄验证机制。

管理机制层面，一是强化信息披露与技术备案要求。《数字服务法》规定，大型在线平台与搜索引擎须向监管部门提供算法运行逻辑的说明文件；2024 年 8 月生效的《人工智能法》规定，高风险人工智能系统提供者须向监管部门提供算法的技术文档与合规报告；《通用人工智能行为准则》进一步明确，不同类型的模型提供者需向公众披露 8 类信息：一般信息、模型属性、分配许可方法、使用规范、训练过程、数据信息、计算资源、能源损耗；《改善平台工作条件指令》要求平台须向监管部门通报自动化决策系统的部署与使用情况。**二是**建立算法风险评估体系。《数字服务法》规定，大型平台须定期开展算法风险评估，包括算法应用对人格尊严、言论自由、儿童权益和消费者保护等方面的潜在影响，并向指定审查机构提交相关结果。

（四）总结：中外算法规制的差异与趋势

中外算法规制因制度基础、治理理念与产业发展的差异，形成了各具特色的治理范式，但均指向构建可信赖的算法治理目标。我国立足平台经济发展现实，嵌套于人工智能治理框架下，采用平台义务与个体权益相结合的双规治理思路，通过明确平台主体责任、实施全生

命周期监管，兼顾监管效能与产业发展诉求。**美国**秉持创新与监管并行的思路，以外部问责为核心，在公共领域率先建立算法问责机制；同时依托州级自治与专项立法相结合的方式，为技术创新预留灵活空间。**欧盟**以个人权利为核心，依托《通用数据保护条例》《数字服务法》等强监管框架，通过赋予数据主体知情权、参与权和救济权等，构建起权利导向的合规治理体系，凸显对个体权益的硬性保障。

从发展趋势来看，全球算法规制正突破传统治理框架，呈现三大核心演进方向。一是从“技术中立”转向“规则嵌入”。算法不再被视为单纯的工具，需明确使用者的主体责任边界，通过规则设计将“可理解、可验证、可问责”要求贯穿算法全生命周期，促使治理从被动应对转向主动防控。**二是**从“短期合规”转向“长期治理”。治理重心不再局限于满足当下监管要求，而是更加关注算法对社会结构、公平正义的长期潜在影响，算法风险评估、社会影响评估逐步成为制度标配，推动治理向常态化、长效化转型。**三是**从“常规保障”转向“重点群体”。未成年人、新就业形态劳动者、消费者及各类弱势群体成为算法治理的优先保护对象，通过强化针对性规制，防范算法对特定群体的权益侵害，彰显治理的公平正义导向。三大趋势相互交织，共同推动全球算法治理向系统性、包容性和负责任的方向深度演化。

三、算法透视：从应用到治理的算法分层逻辑

（一）技术层：深度学习下的范式演变

算法技术层是规则设计与平台治理的底层支撑。在深度学习驱动的背景下，算法在赋予机器强大认知与决策能力的同时，也呈现出模型高度复杂、场景广泛多样与核心参数保密性等技术特征。

1. 深度学习框架的“黑箱”性质

主流平台算法的技术逻辑普遍基于深度学习框架，其运作流程高度耦合且极具复杂性。深度学习算法通常包括四个关键环节：**一是**数据输入与预处理。对用户行为、内容特征及上下文环境等多维度数据进行采集与清洗，为模型学习奠定基础；**二是**表征学习与嵌入化转换。利用嵌入技术将高维稀疏信息映射为低维稠密向量，以捕捉实体间潜在语义关联；**三是**深度神经网络的交互建模。通过卷积神经网络、循环神经网络及 Transformer 等多层非线性网络结构，对特征向量进行深层交互学习与模式抽取；**四是**模型输出与迭代优化。利用反向传播与梯度下降等机制在海量数据上持续训练，不断优化模型参数以提升模型性能。在此过程中，深度学习模型往往包含数百万乃至数十亿参数，其决策路径由大量参数的非线性交互共同作用生成，通过代码审查或结构解析难以全面理解其决策逻辑。同时，提升算法可解释性所需的技术与人力成本极高，导致治理实践中成本与收益难以匹配，限制了算法审查的有效性。

2. 典型场景下的技术内核分化

算法根据产业功能和用户场景呈现出显著技术差异，使得单一治理手段难以覆盖多元化应用领域。有效的技术治理需具备跨场景的专业知识体系，并适应算法快速迭代的特征。按功能逻辑划分，当前算法主要集中于三类典型场景：**一是信息服务场景，**如个性化推荐、检索过滤和排序精选等，旨在从海量内容中精准筛选用户偏好信息。其技术内核涵盖自然语言处理、知识图谱、协同过滤和各类深度学习排序模型，核心在于理解用户意图、内容特征以及两者间的复杂关联。**二是资源调度场景，**如外卖配送、网约车派单等，重点在于平衡效率、

公平与用户体验。此类算法通过融合运筹优化、实时数据流处理、地理空间分析及强化学习等技术，旨在实现时间与空间维度的动态匹配与最优调度。**三是内容生成场景，**以生成式人工智能驱动的文本、图像和音视频创作为代表，旨在创造具有原创性的数字内容。其主要技术包括生成对抗网络、变分自编码器、大语言模型，以及扩散模型等。这些模型参数量巨大、结构复杂，生成逻辑显著区别于推荐与调度类算法，呈现更高的不确定性与风险外溢性。

3. 商业秘密与监管透明的张力

算法模型的设计架构、网络层级与权重参数构成平台企业的核心技术资产，既是技术竞争力的来源，也是算法治理的敏感边界。这些模型参数源于企业在长期投入中积累的技术优势，涉及大量研发、算力与人力资源，是算法性能与商业收益的关键所在。若要求平台全面公开模型参数，可能触及商业秘密保护红线，削弱产业创新活力。2025年3月，在北京抖音科技有限公司诉亿睿科信息技术（北京）有限公司侵害著作权及不正当竞争纠纷一案中，北京知识产权法院适用《中华人民共和国反不正当竞争法》第二条，确认算法模型结构与参数组合属于受法律保护的技术秘密，反映出算法治理需在监管透明与商业保密之间取得制度平衡[8]。

（二）规则层：决策建构与价值注入

算法规则层位于技术实现与社会应用之间，是连接技术逻辑与社会伦理的关键枢纽。它通过嵌入价值导向、设定决策目标与适应具体场景，将复杂的计算过程转化为具备规范性与可解释性的社会行动逻

[8] 最高人民法院知识产权法庭. AI时代 知识产权司法迎来新挑战[EB/OL].2025-07-18. https://ipc.court.gov.cn/zh-cn/news/view-4513.html

辑。规则层的治理焦点不在于算法代码的技术实现，而在于算法在真实世界中“如何决策”“以何为优先”以及“对谁负责”。

1. 算法向善与伦理嵌入

算法本质上是人类价值观的制度化表述，其逻辑目标、奖惩机制与约束边界均体现开发者与社会共同价值取向。构建可信赖算法的首要前提是，将“算法向善”原则内化为规则设计的核心逻辑。近年来，治理实践中正逐步形成价值嵌入的多维机制。例如，**公开透明方面，**互联网信息服务提供者须在算法备案系统中披露算法的工作原理与决策逻辑；**信息保护方面，**基于个性化推荐的商业推送须同时提供通用推荐选项，允许用户便捷拒绝个性化推荐；**内容保障方面，**平台设定优质内容曝光权重，并提升优质创作者的收益分成；**公平公正方面，**在调度决策算法中嵌入劳动者疲劳度监测规则，规定骑手跑单超过 8 小时会收到休息提醒，超过 12 小时将强制下线；**多元包容方面，**在内容推荐中新增无障碍适配权重，自动识别与优化图文内容中的色彩对比度、字体大小等参数，保障弱势群体的使用体验。

2. 多元利益平衡与权重配置

算法规则的核心是目标函数设定与各方权重配置，其运行过程往往在多元利益间实现动态平衡。一方面，平台算法需同时回应企业效率、用户体验、劳动者权益与社会责任等不同维度的诉求，兼顾各方不同社会群体的利益。**另一方面，**算法规则的公开透明需保留在一定范围内。欧盟《数字服务法》要求在线平台须在用户协议中明确列出推荐系统的主要参数与可调整选项，以保障用户的知情与选择权。同时，平台也须在公开范围内平衡算法安全与商业秘密的保护，这种“有限透明”机制成为兼顾算法解释性与创新激励性的制度平衡点。

3. 特定场景下的定制化差异

算法规则在不同场景下呈现出定制化差异，其治理重心也随功能属性而变化。一是资源调度场景强调公平与效率的动态平衡。例如，美团外卖订单分配系统通过“顺路单优先”“新手骑手保护”“高温恶劣天气补贴”“疲劳度监测与强制休息”等细分规则，协调骑手劳动强度与平台效率目标。**二是信息服务场景**聚焦注意力分配与内容责任边界。抖音、快手等平台青少年模式通过设定“单日使用时长上限”“夜间禁用时段”“内容池特殊筛选”等刚性规则，与以个性化和用户黏性为主要目标的常见推荐规则形成区分。**三是内容生成场景**侧重合规性与风险过滤。百度文心一言的多层过滤机制在医疗、法律、金融等专业领域增设“权威信用源引用”“避免提供绝对化建议”等严格规则，体现了生成合成算法在高风险场景的治理要求。

（三）平台层：实践枢纽与生态建设

算法平台层是将技术逻辑和规则设计转化为现实效能的核心场域。其核心价值在于将抽象的治理原则转化为可操作的组织机制、能力体系与协同网络，成为连接技术可行与公众可信的关键枢纽。

1. 从技术中立到责任担当的回归

平台企业是算法设计、部署、运营的主要受益方，其治理角色正从“被动响应”转向“主动担当”。随着算法对公共舆论、劳动关系和社会分配的影响力不断扩大，企业的社会责任边界需要被重新定义。**一方面，**领先的平台企业已开始探索设立专门的算法治理委员会、人工智能伦理委员会等专门组织，并设置首席伦理官等相关职位，将算法治理提升至企业战略层面，负责制定伦理准则、审查高风险算法应用、并监督内部治理政策的执行。**另一方面，**平台需在算法全生命周

期明确“谁开发、谁负责”的问责机制，细化从概念设计、数据处理、模型训练、测试验证、上线部署、运行监测到迭代优化等环节的责任分工，确保问题可追溯、责任可界定、处置可落实。

2. 从风险防控到韧性治理的强化

算法治理的关键在于建立风险识别、评估与应对的全链条能力，以提升算法系统的韧性与可持续性。一是建设风险识别与动态监测能力。构建算法风险监测平台，综合运用模型可解释工具、异常监测系统和人工复核机制，主动发现算法偏见、歧视、安全漏洞以及潜在的社会风险。**二是提高风险评估与主动防控能力。**在算法应用上线或重大调整之前，开展伦理审查和综合影响评估，审慎评估算法对个人、群体和社会的潜在影响，并采取有效的技术或管理措施进行主动防控。**三是深化应急响应与事件处置能力。**建立算法安全事件响应预案，明确响应流程、责任部门和处置权限，确保一旦出现算法滥用、决策失误或舆情风险时，能够迅速启动应急机制，有效控制事态发展并复盘改进，强化平台层面的治理韧性。

3. 从企业自治到生态共治的演进

算法治理的复杂性决定了无法限于单一治理主体，而需构建多元协同的治理生态。平台企业在履行主体责任的同时，应推动监管、用户与社会共同参与治理。**一是深化政企协同，强化合规执行力。**平台企业需主动遵从国家与地方监管规定，落实《互联网信息服务算法推荐管理规定》《生成式人工智能服务管理暂行办法》等要求，严格执行算法备案制度，积极配合监督检查。**二是推动用户参与，强化社会监督力。**通过实现算法可视化、保障用户知情权与选择权，以及搭建便捷申诉机制，赋予用户对算法结果的反馈与纠偏能力，使其从“被

动使用者”转变为“治理参与者”。**三是支持第三方评估，提升治理公信力。**鼓励独立第三方机构、科研院所及行业协会参与算法安全性、公平性与透明度评估，其评估结果作为监管决策、公众监督与企业改进的客观依据。

四、算法实践：构建可信赖算法治理产业自律框架

在全球算法治理持续深化的背景下，产业自律正成为推动可信赖算法体系建设的关键力量。本报告提出“可信赖算法治理产业自律框架”（见图 2），旨在推动算法治理从“外部规制”走向“内部自律”，以释放算法在关键场景中的赋能效应，实现人工智能时代算法模型的高质量发展。**治理目标方面，**产业界应遵循“从合规驱动迈向信任驱动”的核心目标，围绕个体权益保障与公共价值维护治理要求开展实践；**治理要点方面，**综合考虑算法技术的复杂性、差异性与保密性挑战，产业自律应突出规则层的价值导向、目标多元与场景适配，聚焦平台层的主体责任、风险导向与多元协同，推动技术治理逐步转向规则与平台层面的治理；**治理手段方面，**产业自律应聚焦公开透明、信息保护、公平公正与内容保障四大维度。其中，公开透明与信息保护是对个体权益保障的深切回应，强调个人层面的知情权与隐私权的实现；公平公正与内容保障是对公共价值维护的集中展示，关注社会公平公正与内容生态健康。基于此，形成了技术、规则与平台三位一体的产业自律体系，为可信赖算法治理提供了实践路径与落地指南。

图 2 可信赖算法治理产业自律框架

（一）公开透明：构筑信任基石

公开透明是构建可信赖算法治理的基石，其核心在于保障用户的知情权与理解权。通过对算法信息的披露、运行机制的展示及决策过程的可验证，消除公众对“黑箱”的疑虑，为社会监督和责任追溯提供前提条件。**一是让技术可解释，**通过**可解释性技术**实现模型推理过程的可视化和因果关联分析。IBM 的 AI Explainability 360 工具包能够量化特征对预测结果的影响，并生成可视化解释；微软的 Interpret ML 框架支持对黑箱模型的后置解释，为开发者提供特征贡献度排名与可读图示。**二是让规则可理解，**通过**公开算法运行逻辑**与**提升可视化效果**使用户能够理解算法如何影响其体验与权益。美团自 2021 年

起持续发布算法机制说明，并在2025年设立“算法公示专区”，定期披露改进细则；Meta推出“系统卡”（System Card），向用户揭示推荐逻辑并提供个性化控制选项。**三是让平台可监督，**通过定期发布**透明度报告**和接受**第三方审计**来回应社会关切。小红书在《社区治理报告》中系统披露算法应用及用户反馈处理情况，数据显示透明度提升30%，显著增强了公众信任；TikTok则在都柏林等地设立透明度中心，邀请学者与政策制定者实地评估算法运行，并与内容真实性联盟（C2PA）合作，为人工智能生成内容附加可追溯凭证，彰显平台透明度审计的公信力。

案例：抖音安全与信任中心的透明化治理实践

为破解算法“黑箱”难题，抖音于2025年上线“安全与信任中心”，率先在行业内构建系统化的算法透明化治理机制（见图3）。**一是**打破算法黑箱，保障用户知情权。通过中心网站公开算法原理、社区规范、治理体系及用户服务机制，系统展示算法全景。**二是**强化知识普惠，提升理解度。发布《一文全解！抖音算法原理公开》等科普文章，首次披露“通过神经网络计算预估用户行为概率”的核心逻辑，推动推荐机制从“标签化”向“兴趣推演”转型。**三是**拓展公众参与，促进社会共建。举办“安全与信任中心开放日”活动，邀请公众与算法工程师面对面交流，并发起“#算法背后的秘密”话题，形成开放互动的算法认知生态。

图 3 抖音安全与信任中心网站

中心网站上线后累计访问量超 30 万，日均 PV 稳定在 5000 以上，被中央网信办列为典型案例，人民日报评价其“打破信息壁垒，回应公众关切”，成为跨平台可复制的算法公开标杆。该案例表明，算法透明化的关键在于机制化、常态化的公开与沟通，使透明真正可感知、可参与、可监督，从而在产业实践中形成长期信任的治理生态。

（二）信息保护：夯实制度前提

信息保护是可信赖算法治理的制度前提，其核心在于防止个人信息数据被过度采集或滥用。通过在数据全生命周期中强化合规管理，保护个人信息安全，为负责任的技术创新奠定制度基础。**一是让数据脱敏可用，**通过匿名化与去标识化技术减少数据暴露风险。滴滴在出行数据管理中引入差分隐私与联邦学习机制，实现分布式处理与脱敏分析，有效防止敏感信息泄露。**二是让采集有度可靠，**遵循实现服务功能所必需采集范围的**最小化原则**。美团在算法公示中明确限定信息收集边界，确保数据使用与业务目标匹配，减少不必要的画像生成，

从源头上控制数据过度使用风险。**三是让用户有权可控，**通过**用户控制工具**增强个体在算法交互中的自主权。小红书上线“内容偏好调节”功能，使用户能自主设置推荐内容类别；抖音完善“不感兴趣”和屏蔽词机制，让用户主动干预算法推荐逻辑，强化对个人数据流向的掌控力。**四是让体系可防可查，**构建系统化的**数据安全管理体系**与**合规审查机制**。百度建立分级安全体系，设置风险阈值与动态预警机制，实现算法调用全流程监测；奇安信成立数据安全治理委员会，制定内部安全策略并定期发布 ESG 报告，展示合规治理成效，构建制度与技术双重防护格局。

案例：滴滴分单算法治理实践

作为拥有数亿用户与千万级司机的出行平台，滴滴面临的核心挑战是如何在效率与公平之间实现平衡。传统的“全局最优”派单算法虽能提升整体效率，却可能忽视司乘个体的体验，引发公平性争议。对此，滴滴通过一系列机制创新，努力在算法决策中实现透明、公平与与多方共赢。**一是**推动分单规则透明化。在司机端上线“透明派单规则”，公开派单逻辑和订单分配原则，使司机能够实时了解接单原因，显著提升规则的可理解性。**二是**建立劳动者教育与沟通机制。通过线上“滴滴课堂”、线下专题培训及新司机必修课程，系统解读算法运行逻辑，帮助司机认识算法运行原理并提升参与感（见图 4）。**三是**注重反馈与改进机制。依托司机生态委员会，定期征集意见并推动算法迭代，对“就近派单”等规则进行优化，同时提供验证与选择工具，让司机在算法决策中拥有更多自主性与监督权。

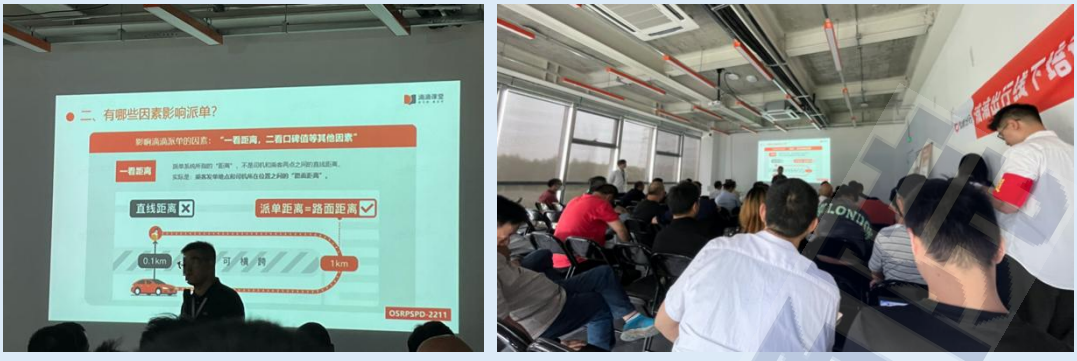

图 4 滴滴线下派单规则专题培训会

这些措施显著提升了派单公平性和透明度，司机因理解规则而能主动优化接单行为，用户体验与平台整体资源利用率同步提升。其透明治理模式为也行业提供了参考样本，推动网约车行业在规则公开、参与共治与权益保障上的整体进步。该案例表明，平台企业提供建立透明、可沟通、可监督的治理体系，不仅能化解利益冲突，还能塑造面向未来的可持续算法生态。

（三）公平公正：坚守伦理底线

公平公正是可信赖算法治理的伦理底线，其核心在于确保算法决策过程的正当性与均衡性。通过持续优化算法决策逻辑与治理规则，提升平台的社会责任形象，增强公众信任，为算法产业的长期可持续发展筑牢群众基础。**一是让技术公正可调，持续监测与改进**模型在不同群体间的表现差异。Facebook 推出的 Fairness Flow 工具能够识别并量化算法的群体偏差，用以指导模型优化；Google 针对 Gemini 图像生成中的刻板偏差问题开展算法模型修订与数据校准，推动人工智能公平性技术向产业化落地。**二是让劳动安全有度，**主动嵌入**劳动保护机制**保障算法调度下的劳动者权益。美团上线“防疲劳”功能，在骑手连续工作超过 12 小时后自动下线；饿了么与骑手代表签署《网

约配送劳动规则协议》，提升最低时薪并逐步取消超时扣款，推动算法调度与人文关怀相结合。**三是让交易平等透明，**通过规则公开与算法约束**防止区别化定价**。货拉拉公开分单原则，承诺无差别定价，巩固平台公平信任机制。**四是让服务包容可及，**通过优化交互设计**保障不同群体**平等享受算法红利。小红书建设“未成年人友好社区”并强化搜索保护；抖音推出老年人简易模式与语音辅助功能，让特殊群体更好地参与数字生活。**五是让多方合作共赢，**通过**利益协商机制**实现多方诉求平衡。美团、饿了么等平台通过恳谈会收集劳动者与用户反馈，并基于实证数据调整调度逻辑，在效率与公平之间找到动态平衡；货拉拉与工会组织协商算法改进，2024 年收集超 500 条建议以优化劳动者体验。

案例：美团“安全分”体系算法治理实践

在即时配送行业快速发展的背景下，骑手群体规模庞大，但长期面临劳动强度高、交通风险大与权益保障不足等问题。为回应社会关切并推动骑手群体的可持续发展，美团上线并不断优化“安全分”体系，将交通安全行为与骑手收入、权益挂钩，从而在算法治理中体现多元包容（见图 5）。**一是**多维度评估骑手驾驶行为。通过对骑手驾驶行为的大数据进行全面分析，生成每个骑手的专属分数，并与月度奖金、权益兑换挂钩。**二是**保障骑手的公平发展机会。安全分高的骑手能够获得差评保护卡、超时消除卡等工具，从而获得更公平的发展机会。**三是**引导骑手遵守交通规则。若骑手存在超速、未戴头盔、闯红灯等违规行为，会扣分并降低相应权益。该机制有效引导骑手遵守交通规则，同时保障守规矩骑手在劳动竞争中不被边缘化。

图 5 美团骑手端 APP“安全分”奖励截图

从治理效果看，“安全分”推动了行业治理模式从“处罚为主”向“奖优罚劣”转型，增强了骑手群体的参与感与获得感。例如，在“五一”期间，美团为各城市排名前 100 的高分骑手发放现金奖励，累计惠及约 1.5 万名骑手。此举不仅提升了安全出行的自觉性，也通过正向激励提升了骑手的社会认同感与职业尊严。该案例表明，算法治理若能关注劳动者权益和社会公平，不仅能改善行业生态，也能通过制度性设计促进社会多元包容的价值实现。

（四）内容保障：强化价值支柱

内容保障是可信赖算法治理的价值支柱，旨在维护信息生态的健康与多样性。通过构建多层次的内容质量保障体系与价值导向机制，提升平台信息生态的可信度和公信力，推动形成包容、理性、向善的社会舆论空间。**一是让内容可靠可溯，**通过**人机协同审查**与**内容溯源机制**确保信息来源可验证、推荐内容符合社会主流价值。YouTube 结

合人工评估员与人工智能工具进行多轮审核；微博在内容推送中增加来源标注与出处验证，让用户能够追溯信息来源。**二是让创作传递正向，**通过**优质内容曝光和创作者赋能**，鼓励正向、多元创作生态。抖音推出“抖音精选”频道计划，2025 年优质内容曝光量同比提升 300%；小红书每天将超一半流量分配给千粉以下创作者；Meta 通过内容库与开放 API 赋能创作者生态，促进多样化优质内容生产。**三是让内容多元丰富，**通过算法与规则优化提升**内容多样性**，避免形成单一偏向。小红书通过标签体系与算法探索机制扩展内容边界；YouTube 动态调整算法权重，减少边缘内容推荐，2023 年数据显示有害内容消费量下降 70%。**四是让虚假信息受控，**建立针对虚假或恶意信息的快速**识别与处置机制**。抖音与近百家媒体共建“辟谣团”，推出“辟谣卡”功能；微博上线“辟谣小黄签”；美团缩短恶意差评处理周期至 12 小时。

案例：阿里巴巴电商推荐算法治理实践

在电商场景中，推荐算法既能提升用户与商品的匹配效率，也可能导致“信息茧房”与“马太效应”，导致推荐结果趋于单一、流量向头部商家集中。为应对这些问题，阿里巴巴在淘宝平台系统开展内容保障实践，构建兼顾效率与公平的推荐机制。**一是**建模用户负反馈数据，减少推荐用户不喜欢的内容。建立用户负反馈机制，为用户提供便捷的“不喜欢”反馈入口，并基于负向兴趣进行算法调优，有效降低同质化内容推送。**二是**搭建发现性推荐链路，提升推荐系统的多样性。在淘宝首页引入探索性推荐链路，主动向用户推荐未浏览过的商品类目，提升算法推荐的多元性。**三是**开展全链路无偏学习，刻画用户多样的兴趣分布。利用多场景数据融合建模，避免算法过度强化

既有偏好，显著改善“越买越推”的反馈循环（见图 6）。**四是**开辟新品赛道并精准商品潜力挖掘，助力中小商家快速成长。通过新品冷启动与潜力挖掘机制，为优质新品和中小商家提供初始流量与成长支持，缓解流量过度集中的结构性问题，推动生态更具包容性。

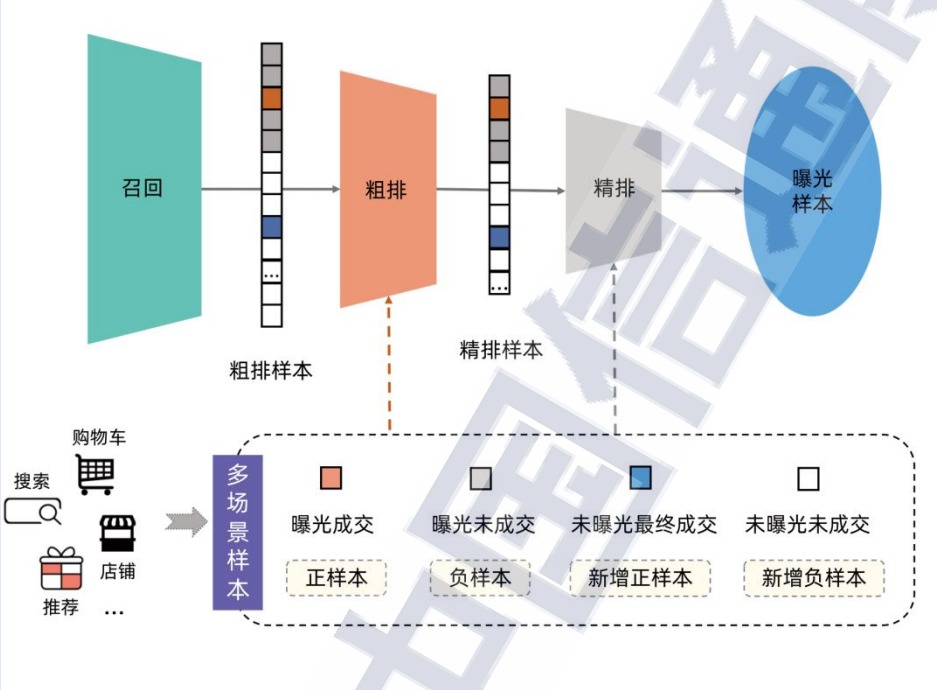

图 6 阿里巴巴全链路无偏学习解决方案

实践结果显示，该机制显著提升了推荐系统的多样性与公平性，用户负反馈量有效下降，推荐结果差异化程度提升；中小商家借助新品赛道与潜力挖掘工具获得日均千万级成交增量。相关成果发表于国际顶级会议，展现了中国电商行业在算法治理中的实践与贡献。该案例表明，电商平台的算法内容治理不仅应关注推荐的结果真实性和质量，更应从制度层面保障多样性与公平性，促进产业生态的健康、共生与可持续发展。

五、小结

算法是人工智能的技术核心与逻辑载体，而算法治理则是人工智能发展安全、可靠、可控的基石。全球主要经济体围绕个体权益保障与公共价值维护等目标，在技术层、规则层、平台层方面积累了一系列可落地的治理经验，推动算法治理从政府监管向多元共治演进。随着人工智能与社会经济运行体系深度融合，算法治理正由技术议题上升为全球治理议题，构建可信赖的算法治理产业自律体系需要通过认知共识与行动落地的双向发力，实现技术、伦理与制度的有机协同。

在认知层面，弘扬算法向善理念，推动产业体系共建。一是明确可信赖算法治理原则，围绕公开透明、信息保护、内容保障、公平公正等方面，形成兼顾伦理约束与产业可行性的制度框架。**二是**重点提升算法透明度，依托产业组织完善治理标准，推动企业主动披露算法机制与运行逻辑，旨在实现可解释、可理解、可问责的三重目标。**三是**探索算法优化与评估路径，将可信赖要素嵌入技术规范、管理标准和评测体系，确保性能提升与社会价值协调统一。

在行动层面，落实技术向善要求，推动平台协同治理。一是推动技术可控，在开发阶段落实伦理设计要求，采用联邦学习、差分隐私等技术保障数据安全；在测试阶段建立仿真验证与对抗测试机制；在部署阶段根据风险等级实施分级分类管理。**二是**健全动态管控体系，形成贯穿事前识别、事中防控、事后处置的全流程治理机制。**三是**完善多元共治格局，鼓励第三方机构参与算法影响评估，探索设立算法治理“沙箱”机制，推动政府、企业、社会多方协同，共建负责任的

算法产业生态，最终推动算法从“工具理性”向“价值理性”转变。

中国信息通信研究院 人工智能研究所

地址：北京市海淀区花园北路 52 号

邮编：100191

电话：010-62304980

传真：010-62304980

网址：www.caict.ac.cn