可信任导向下人工智能立法的目的重构与规范展开

王　怡

摘　要　人工智能立法长期深陷效率、安全与创新的三元平衡悖论，现有规制方案因局限于价值取舍的线性逻辑而难以突破困境。法律的本质是社会信任的建构机制，其通过明确权利义务、划定行为边界、保障权利救济提供稳定预期，这与人工智能时代的治理需求天然契合。可信任并非独立于三元价值的第四维度，而是统摄三者实现动态协同的逻辑枢纽，将其确立为人工智能立法的核心目的，既能消解传统平衡模式的内在矛盾，又能回应算法黑箱、数据滥用等信任危机。基于这一核心命题，未来应制定一部专门的人工智能法，以信任建构为立法目的，确立信任建构、分级分类、场景适配、多元协同的原则群，通过算法可信、数据可信、主体可信、治理可信的精细化制度设计，为人工智能法治提供稳定价值指引，实现技术应用与社会信任的良性互动。

关键词　人工智能立法　可信任　法律功能　协同治理

作者王怡，中国社会科学院法学研究所副编审、中国社会科学院大学法学院副教授（北京 100720）。

中图分类号 D9　　**文献标识码** A　　**文章编号** 0439-8041(2026)05-0092-10

引言

人工智能技术的深度应用，为各行业提质增效与转型升级注入了动能和信心，但也引发了各界对“安全优先则抑制创新”“效率至上则牺牲安全”“创新激励则打破平衡”等恶性循环的担忧，使妥善规制人工智能成为当下备受争议的话题。实践中，欧盟《人工智能法案》的严格分级监管对安全价值予以高度强调，该做法被质疑抬高了中小企业的创新门槛[①]；美国实行创新豁免的宽松监管，能在很大程度上释放技术创新活力，但又因其容易引发数据滥用、算法歧视等而遭到诟病。[②] 我国分散于各部门法的人工智能规范，因缺乏核心目的统摄，不免存在规制真空与重复监管并存的现象。[③]

围绕未来人工智能立法的基本方向、基本路径等问题，我国学界展开了激烈讨论。既有研究提出了多元平衡说[④]、阶段适配说[⑤]、分散立法说[⑥]等各种创见，这些主张几乎均以在效率、安全、创新诸种价值之间分

① 参见袁康：《人工智能统一立法的论争与选择》，《比较法研究》2025 年第 5 期。
② 参见付新华：《全球人工智能立法的多元趋势与中国模式》，《交大法学》2025 年第 6 期。
③ 参见陈亮：《人工智能立法体系化的困境与出路》，《数字法治》2023 年第 6 期。
④ 参见周辉：《人工智能综合性立法及其实现》，《法学研究》2025 年第 6 期。
⑤ 有学者建议通过总则式立法［参见张凌寒：《中国需要一部怎样的〈人工智能法〉？——中国人工智能立法的基本逻辑与制度架构》，《法律科学（西北政法大学学报）》2024 年第 3 期］，部门法完善［参见李学尧：《人工智能立法的动态演化框架与制度设计》，《法律科学（西北政法大学学报）》2025 年第 3 期］，低位阶立法相结合的“小快灵”模式先行供给制度（参见周汉华：《论我国人工智能立法的定位》，《现代法学》2024 年第 5 期）。
⑥ 参见王凌皞：《未来尚未到来——人工智能统一立法模式的内在悖论》，《交大法学》2025 年第 6 期。

配权重为核心逻辑，在是否需要通过专门立法对人工智能技术与发展进行全局布控，以及专门的人工智能立法应定位为总则性立法还是综合性立法等问题上，始终争执不休。共识之所以难以形成，其原因大致可归结为两点：其一，人工智能技术原理复杂，且技术迭代发展速度超出人类预期，哪些问题是真正需要法律加以解决的问题，哪些问题属于技术迭代发展过程中能够自行消解的，目前无法判断。其二，人工智能立法涉及技术企业、用户、政府等多方利益诉求，在价值优先性难以抉择的背景下，多元共治理论将价值选择悄然转换为价值权衡这一贯穿法学、社会学与技术治理的“世纪难题”。诚然，对于未来的人工智能立法而言，不论其形式、路径如何，最为理想的效果无外乎清晰划定安全、效率、创新等价值之间的边界，在坚守底线的基础上最大程度地释放新兴技术的效能。然而，在不确定性已为常态的社会背景下，寄望于通过人为建构确定性来为社会活动供给秩序已然不切实际。尤其是，当前技术迭代的速度远超规则制定周期，应用场景的跨界延伸打破了传统治理边界，这使得原本就已有限的确定性被不断稀释。

执着于多元价值的边界划分与权重分配，而忽视技术动态演进与场景异质性带来的变量，这样的规制逻辑在不确定性面前注定难以奏效。在人工智能技术迭代迅猛、应用场景复杂多元的不确定环境中，治理的核心诉求应当发生转向，即从追求多元价值间静态的确定性划分，转向在不确定性中锚定稳定的价值坐标，确立一个能够统合多元利益诉求、弥合价值分歧的底层价值共识。“可信任”恰是这样的核心枢纽，它能将各类价值转化为可信性的具体构成维度，从而在动态适配技术与场景变化的过程中寻求高维共识，实现技术创新与社会信任的良性互构。

从社会学的角度观察，法律不仅仅是调和利益分歧的工具，更是社会信任的建构机制，能够为社会活动提供稳定的预期。[①] 信任是社会合作的基础，而法律是信任的制度化载体。回归人工智能治理领域，信任既包括用户等主体对人工智能技术本身安全性、可靠性的信任，也涵盖产业界对于国家立法、监管政策稳定性、公平性的信任，更包含各类主体对制度能够有效化解技术风险、平衡多元利益的信任，信任的这种多维度内涵凸显了其对于人工智能治理的价值。在可信任导向下，有必要将人工智能立法的目标重构为：通过明确权利义务、规范行为边界、建立救济机制等制度安排，系统性培育技术信任、政策信任与制度信任，促成多元主体在合作与组织的过程中形成动态均衡的秩序[②]，从而统合权利保障、安全和效率诸种价值，为多元主体的合作扫清障碍。鉴于此，本文立足人工智能时代的信任危机，从法律的信任建构功能出发，证成“可信任”作为人工智能立法核心目的的法理正当性，在此基础上阐释可信任对于多元价值的统摄机理，最终以可信任为轴心初步建构人工智能法的规范体系，为摆脱人工智能治理困境、实现技术应用与社会信任的良性互动提供法治方案。

一、人工智能时代的信任危机与治理关键

现代社会的高度复杂性和不确定性，催生了社会风险与社会信任两大核心议题，且这二者互为因果。人们对社会风险的普遍感知会直接导致社会信任度下降，引发系统性信任危机；而信任危机的蔓延又会进一步加剧社会风险，强化不确定性、不安全性等时代特征，最终形成“风险加剧—信任流失—风险恶化”的循环。作为社会复杂性的简化机制，社会信任是缓解风险、降低社会运行成本、维系社会协作的关键，信任的缺失将从微观的人际互动、中观的经济活动到宏观的制度运转全面冲击社会稳定的根基。[③] 人工智能技术的颠覆性突破，不仅重塑了社会分工与协作形态，更将现代社会的不确定性与风险传导效应放大[④]，催生了数据泄露、算法歧视、责任模糊等新型风险。这些风险深度渗透至社会生产生活各领域，使得信任的建构与维系成为人工智能技术良性嵌入社会的核心前提甚至决定性因素。相较于传统社会，人工智能时代的信任危机不再局限于人际或组织层面，更延伸至技术本身与制度治理维度，呈现出隐蔽性强、传导速度快、影响范围广等更复杂的形态，对信任治理提出了全新挑战。

---

① 卢曼意识到，在高度复杂的现代社会中，信任的个人机制或互动机制已不再够用，必须通过法律信任来建立社会信任。参见卢曼：《信任：一个社会复杂性的简化机制》，瞿铁鹏译，上海：上海人民出版社，2005 年，第 62—79 页。

② “智能社会的治理是一项复杂的系统工程，需要国家、行业、组织、公民个人等主体的共同参与。”张文显：《构建智能社会的法律秩序》，《东方法学》2020 年第 5 期。

③ 参见郭未、王灏晨、罗朝明：《中国社会信任与社会风险透视——基于知识图谱的视角》，《科学学研究》2013 年第 10 期。

④ 参见马长山：《人工智能的社会风险及其法律规制》，《法律科学（西北政法大学学报）》2018 年第 6 期。

（一）技术变革下的信任困境与价值失衡

现代社会的秩序内核在于分工深化所催生的有机团结，当社会分工突破简单协作的桎梏，个体与组织间的相互依赖便构成社会顺畅运转的基础性支撑。涂尔干关于社会整合的经典论断，在人工智能技术迅猛发展的当下，愈发彰显出深刻的现实解释力。当前，人工智能正以颠覆性力量重塑传统社会分工格局，算法已然成为新型生产要素的核心载体，数据的跨界流动持续打破行业壁垒与地域限制，催生出开发者、平台方、用户、监管机构等多元分化的协作主体，形成了较工业时代更为复杂的社会相互依赖网络。与传统社会的协作关系相比，人工智能时代的多元主体互动更凸显间接关联与非面对面协作的特征，利益诉求的差异化与信息掌握的不对称性更为突出，这使得信任的构建与维系，成为维系这一复杂协作网络、巩固社会有机团结格局的核心前提与关键纽带。

深入审视人工智能时代的发展困境可以发现，其核心危机并非技术本身的迭代局限，而是信任机制的系统性缺失。这一缺失不仅是诸多表象问题的根源，更直接割裂了效率、安全与创新三者的协同关系，成为制约人工智能技术良性嵌入社会的关键瓶颈。在数字经济主导的协作格局中，人工智能技术的赋能逻辑高度依赖“数据流通—算法决策—多元协作”的闭环运转，而这一闭环的顺畅运行必须以稳定的信任预期为制度基础。当信任缺失时，整个系统的运转效率将出现大幅损耗。例如，企业为规避数据泄露风险，不得不投入高额成本构建数据隔离体系，导致数据要素无法充分流动，算法模型的训练精度和应用效能受限；用户因担忧隐私泄露和算法歧视，对接受人工智能服务持犹豫观望态度，使得技术创新的市场转化路径受阻；监管机构因缺乏信任基础，只能采取严苛的前置审批措施，进一步压缩了创新空间。与此同时，信任缺失还会放大安全风险的传导效应。例如，算法黑箱因信任不足被解读为“恶意操控”[①]，即便技术本身不存在安全漏洞，也可能引发社会性恐慌；数据跨境流动因信任缺失被赋予“数据主权威胁”的额外属性[②]，加剧了国际间的技术壁垒和安全焦虑。这种由信任缺失引发的连锁反应，最终会导致人工智能产业陷入“低信任—高成本—弱创新”的恶性循环，既违背了技术赋能社会发展的初衷，也与人工智能治理现代化的目标导向相背离。

尽管已有研究者认识到信任对于人工智能发展的重要性[③]，但往往将其视为实现上述价值的辅助条件或衍生结果，未能充分凸显其在人工智能治理体系中的核心地位。事实上，信任并非效率、安全与创新的附属品，而是能够统摄三者的底层价值与逻辑枢纽，其对三者的统摄作用根植于人工智能时代协作关系的内在要求。

其一，信任是效率提升的底层支撑。在福山看来，信任本质上是基于共同规范形成的稳定协作预期，这种预期能够大幅降低协作过程中的信息验证、契约谈判和风险防控成本[④]，为高效互动奠定基础。在人工智能领域，信任的这一功能尤为关键，它能够消除数据流转中的信任壁垒，让数据要素在不同主体间安全顺畅流动，为算法模型优化提供充足数据支撑；同时，信任还能减少用户与平台间的认知博弈，降低技术推广中的沟通成本与接受门槛，推动人工智能技术的规模化应用与效率释放。

其二，信任是安全边界的核心锚点。人工智能时代的安全并非绝对的技术无风险状态，而是“风险可控、损失可补救”的可预期状态，而这种可预期性的核心正是信任。当用户信任人工智能产品的安全设计、信任平台的数据保护能力、信任监管机制的风险兜底功能时，即便技术层面存在微小漏洞，也不会引发系统性安全危机；反之，若缺乏信任基础，任何局部的安全问题都可能被放大为全局性的信任恐慌，甚至诱发社会性风险。这种由信任构建的安全边界，既避免了“为零风险而无限加码合规”的治理困境，又为技术应用划定了社会可容忍的风险阈值。

其三，信任是创新活力的重要保障。创新本质上是对未知领域的探索，必然伴随着不确定性，而信任能够为这种不确定性提供风险缓冲。只有当市场主体信任监管规则的稳定性与可预期性、信任创新成果的价值认可机制、信任协作伙伴的履约能力时，才愿意投入稀缺资源开展前沿技术探索；反之，若信任缺失，创新

① 有学者指出，“算法黑箱本身并不是应当遭受谴责的现象”，“黑箱不等于不公平”。参见陈景辉：《算法的法律性质：言论、商业秘密还是正当程序?》，《比较法研究》2020年第2期。

② 数据主权的实施原则是全球数据治理博弈的核心领域。相关理论上的冲突与实践中的博弈，参见汤诤：《数据主权的博弈与理论重构》，《环球法律评论》2025年第4期。

③ 参见杨建军：《可信人工智能发展与法律制度的构建》，《东方法学》2024年第4期；高富平、张启航：《可信AI：人工智能法律治理的内在逻辑与实现路径》，《学习探索》2025年第9期。

④ 参见福山：《信任：社会美德与创造经济繁荣》，郭华译，桂林：广西师范大学出版社，2016年，第27—30页。

主体可能陷入怕风险、怕追责、怕失败的观望状态，进而抑制创新活力。

综上，信任的缺失是人工智能时代效率、安全与创新失衡的核心症结，其不仅会导致单一价值维度的实现受阻，更会割裂三者间的协同关系。重建信任构成了统摄效率、安全与创新三大价值，摆脱人工智能发展困境的关键，可信任导向应成为人工智能治理的核心逻辑。

（二）信任重建的治理路径

帕特南对社会资本的细化阐释，为人工智能时代的信任重建提供了清晰的治理路径。他将社会资本拆解为信任、规范与网络三大共生要素，指出三者的协同作用才能真正提升社会协作效能。[①] 这一视角提示我们，人工智能时代的信任重建并非单一维度的制度设计，而是要以信任资本为核心，联动规范，形成治理合力。具体而言，在制度层面，需通过算法解释权、数据可追溯性、责任认定标准等法律规范，为分工体系中的各主体划定权利义务边界；在伦理层面，需培育跨主体的共享价值共识，将技术普惠、公平正义等理念转化为行业自律准则，弥合技术发展与社会期待的鸿沟；在实践层面，需构建政府、企业、学界与公众的多元协作网络，让不同主体在互动中积累信任资本，推动信任半径从企业私域扩展至公共领域。[②]

人工智能的“可信任”并非抽象概念，而是可以转化为一系列标准。迈克尔·刘易斯等人曾提出系统的可信任性与其可靠性、可预测性、智能水平和透明度、自主程度紧密相关。[③] 欧盟在《可信人工智能伦理指南草案》中刻画了人工智能的可信特征，将“人类能动性与监管”“技术稳健安全”“隐私与数据治理”“透明”“多样性、不歧视和公平”“社会与环境福祉”“可问责”作为“可信任人工智能”的七项关键条件。有学者提出人工智能的可信任性可通过“目的向善、过程扬善、功能完善”三大核心原则判断。[④] 这些标准与原则虽表述维度不同，但本质上都围绕“如何让人工智能的运行可预期、风险可防控、责任可追溯”这一核心诉求展开，其内在逻辑指向制度层面的四维规范体系：算法可信、数据可信、主体可信与治理可信。其中，算法可信以透明性、可解释性、公平性为标准，目的在于保障算法决策的可预期性；数据可信以合法性、安全性、可控性为底线，旨在规范数据全生命周期治理；主体可信以权责清晰、归责明确为核心，划定研发者、使用者、监管者的行为边界；治理可信以监管协同、救济畅通为保障，以确保信任规则的有效实施。

上述四维规范体系的优势体现在对主观与客观、技术与制度的双重整合上，即以“公众可接受的风险阈值”为核心，将安全、透明、公平、救济等多元要素纳入统一框架，既包含数据安全、算法无歧视的客观安全底线，也涵盖决策可解释、责任可追溯的制度保障，其本质是通过客观标准支撑主观感知。例如，安全是一种偏向客观的技术与合规标准，聚焦于防范技术层面的显性风险，如数据不泄露、算法无重大漏洞等，其评判依据多为可量化的技术指标与合规要求；而公众对人工智能的接纳，从来不是单纯基于技术安全的客观达标，更源于风险可控、权益可保、责任可追溯的主观信任感知——即便技术符合最低安全标准，若缺乏透明的决策机制、畅通的权利救济渠道，公众仍会因“算法黑箱”“责任模糊”产生不信任，这恰好印证了安全不等于信任的核心逻辑。

总之，人工智能的高度复杂性加剧了信息不对称与用户脆弱性，即便专家也难以完全预测算法决策结果，这使得信任对人工智能的开发应用至关重要，缺乏信任将直接阻碍技术潜力的释放。欧盟《人工智能法案》以“可信人工智能”为核心立法目标，本质就是通过规制保护信任，即通过法律规则确保人工智能的开发与应用符合法律、伦理要求，让用户信任技术的安全性与正当性。这一立法实践印证了法律在信任重建中的核心作用：通过设定统一的信任标准、明确主体权责、建立救济机制，能够为人工智能时代的信任建构提供刚性保障，破解“风险加剧—信任流失”的循环。

二、可信任导向下人工智能立法的形式与路径

法律作为现代社会有机团结的核心保障，其基本功能在于为复杂社会关系设定稳定规则、消解协作风险、

---

① See Robert D. Putnam, *Making Democracy Work: Civic Traditions in Modern Italy*, Princeton: Princeton University Press, 1993, p. 167.

② 参见高富平、张启航：《可信 AI：人工智能法律治理的内在逻辑与实现路径》，《学习探索》2025 年第 9 期。

③ See Michael Lewis, Huao Li, and Katia Sycara, “Deep Learning, Transparency, and Trust in Human Robot Teamwork, ” in *Trust in Human—Robot Interaction*, Chang S. Nam, Joseph B. Lyons(eds.), Cambridge, MA: Academic Press, 2021, pp. 321-352.

④ 参见何丽：《人工智能可以作为置信对象吗？——为可信人工智能辩护》，《科学学研究》2023 年第 10 期。

凝聚价值共识。随着社会分工深化，个体与组织间的相互依赖成为社会运转的基础，这种依赖关系的维系离不开共享规范与法律制度的兜底。在人工智能时代，开发者、平台方、用户、监管机构等多元主体的互动愈发依赖间接协作与数据流转，信息不对称与权力失衡问题更为突出，此时法律通过建构信任所形成的稳定预期，成为串联多元价值的关键纽带。

（一）可信任导向下专门立法的优势

公众对人工智能的信任需求，本质是“免于恐惧的底线保障”，而非“毫无瑕疵的完美期待”。当用户相信数据不会被滥用、算法决策公平可追溯、风险发生时能获得救济，其对技术的接纳度自然提升，数据要素流动与技术应用的效率随之提高；当企业相信监管规则稳定可预期、协作伙伴履约可信，便会减少不必要的风险防控成本，将资源投向技术优化与创新，进而实现安全、效率与人权保障的协同增效。与以安全、效率或创新为单一导向的立法不同，可信任导向下的人工智能立法并不追求单一价值的最大化，而是以信任建构为纽带，追求实现多元价值的动态平衡。

就未来人工智能立法的形式而言，制定专门的人工智能法更有助于可信任目标的实现。分步立法或分领域立法的思路，虽然能够迅速回应由特定技术引起的新问题，但也存在规制范围与内容重合、法律概念冗余或定义分歧、规制内容前后不一致等问题[①]，难以支撑可信任这一核心目标。通过制定专门的人工智能法，整合分散规范，将可信任的核心要求转化为统一、明确的法律规则，清晰界定研发者、平台方、监管机构及公众的权利义务与责任边界，既有助于市场主体基于稳定规则规划行为，也能让公众明确权利保障范围、增强对技术的信任，还能让监管部门获得统一明确的执法标准、避免选择性执法，最终以规则的确定性消解技术的不确定性，为可信生态建构奠定根基。就法律运作的内在逻辑而言，可信任导向下的专门立法之所以能够统摄安全、效率与人权保障，核心就在于其能够消解价值实现过程中的相互掣肘，实现多效协同的治理效果。

其一，效率导向的立法将人工智能视为提升生产与社会运行效率的工具，其制度设计侧重简化审批流程、放松市场准入、鼓励技术迭代，往往对隐私保护、算法公平等潜在风险采取“事后救济”的被动态度。安全导向的立法通过设定严格的技术标准、限制高风险应用场景、强化事前审查等方式降低安全隐患，可能因过度限制而抑制创新活力，且难以回应用户对技术的深层信任诉求。与之不同，可信任导向下的专门立法通过稳定的信任预期降低协作成本（提升效率）、化解风险焦虑（保障安全）、激发创新活力（鼓励发展），最终形成“信任—价值协同—信任强化”的良性循环。

其二，效率或安全导向的立法往往围绕技术流程、应用场景、责任划分等工具性要素展开规则设计，难以照顾到人工智能应用背后多元主体间的依赖关系与信任互动。例如，安全导向的立法可能仅要求企业满足技术安全标准，却未必能回应用户对于“技术为何安全”“风险如何防控”的认知需求。可信任导向下立法的治理逻辑是“关系性建构”，其关注人工智能生态中开发者、平台方、用户、监管机构等主体间的信任关系，试图通过制度设计消解信息不对称、权力失衡等信任障碍，塑造可预期、可依赖、可救济的互动环境。在这种逻辑下，法律不仅是行为规范，更是信任关系的建构者。通过明确权利义务、规范信息披露、畅通救济渠道，可信任导向下的专门立法致力于让用户相信数据处理者会合法合规行事，让企业相信监管规则具有稳定性，最终实现主体间的信任协同。

其三，效率、安全、创新等单一价值导向下的立法，通常聚焦于人工智能应用中的特定痛点提供针对性解决方案，这些具体问题的解决固然重要，但从技术发展的角度来看，相关对策往往仅能回应当下可见的治理难题，却难以适配技术快速迭代、应用场景持续拓展带来的长期治理需求，当新技术、新场景出现时，原本具有针对性的规则易陷入规制滞后或失效的困境，无法形成可持续的治理效果。可信任导向下的立法旨在构建覆盖人工智能全生命周期、多元主体参与的信任生态。这种生态化目标的核心要义在于，将信任作为人工智能治理的底层基础设施进行长期培育，从而突破短期问题导向的局限，即便面临技术迭代或场景拓展，信任关系仍能基于统一标准、动态评估与修复机制持续维系。

从科技快速迭代的特性来看，专门的人工智能立法无法事无巨细地覆盖所有技术细节与新兴场景，追求

① 参见邓建鹏、马文洁：《人工智能“逐案设法”治理模式的优化》，《南京社会科学》2024年第6期。

全面覆盖也只会因技术迭代凸显法律滞后性，甚至束缚创新活力。因此，在可信任导向下，立法无需设定更高标准的安全防线，只需围绕“确保可信”提供底线性规则，如明确数据安全的核心标准、算法可解释性的基本要求、责任追溯的最低限度，以及禁止算法歧视、数据滥用等红线条款。这些底线规则往往具有抽象性和稳定性，无论技术应用如何迭代都必须得到坚守，且没有必要随技术细节迭代频繁修改。在遵守底线规则的基础上，研发者拥有充分的创新自主权，无需承受“规则过载”的合规负担。这种“底线规制—创新留白”的立法模式，既适配科技迭代的特性，又体现了最低限度的规范刚性，能够充分回应稳定预期与创新包容的双重治理需求。

（二）可信任导向下人工智能立法的路径

1. 算法可信：从“技术透明”到“可信任导向下的可解释性”。

算法的技术复杂性决定了其完全透明、完全可理解的不现实性[①]，而传统“技术透明”导向易陷入“追求极致技术细节却脱离用户认知”的误区。可信任导向下，立法对算法透明与可解释的要求，核心逻辑在于“信任源于可预期、可参与”，用户对算法的信任并非建立在掌握技术原理之上，而是来自对决策过程的感知、对核心关切的回应以及权利救济的可得性。因此，立法需通过过程可见、决策可追溯、异议可回应的制度设计，构建用户对算法的信任基础，这至少要求：（1）解释具备可感知性，即算法解释应以简洁易懂的语言呈现，解释的目的是让用户感知到算法决策的基本逻辑，而非掌握算法的技术细节[②]；（2）解释具备针对性。针对医疗诊断、司法量刑、信贷审批等高风险人工智能应用，要求提供更具体的解释，包括决策依据、数据来源、影响因素等，回应用户对于“为何做出该决策”的核心关切；（3）解释具备救济关联性，尤其是在公权力运行人工智能系统进行决策时，必须将算法解释与权利救济绑定，用户若对算法决策有异议，可依据解释内容提出复核、更正或拒绝，让解释成为信任修复的重要渠道。[③] 这种“形式上的可解释性”即便无法实现技术层面的完全透明[④]，仍能通过赋予用户知情权、参与权，强化对算法的信任预期。

2. 数据可信：从“合规控制”到“可信任导向的权利保障”。

传统数据治理的“合规控制”侧重形式化达标，如仅满足知情同意的书面要求，却难以解决合规但不安全、合规但用户不信任的核心问题。可信任导向下，立法将数据权利保障作为信任建构核心，本质是因为“信任的根基是权利的实质实现”，用户只有真正掌控自身数据、防范滥用风险，才能形成对数据处理的稳定信任。这具体要求：（1）强调使权利获得实质性保障，不仅明确用户的知情权、决定权、删除权等形式权利，更通过制度设计确保权利落地。例如，针对大语言模型中“剩余数据激活”“输入文本还原”等隐私风险[⑤]，要求企业建立数据全生命周期的安全防护机制，保障用户对数据的控制权；针对“推断隐私”挖掘风险，赋予用户对不合理数据使用的拒绝权。[⑥]（2）实行分级分类保护，基于数据隐私价值差异，对敏感数据设定最高级别保护[⑦]，严格限制收集与使用场景，对普通数据简化合规要求，同时允许用户自主选择披露范围，平衡权利保障与使用便利，避免过度保护降低体验或保护不足损害信任。（3）依托技术赋能强化信任，引入差分

---

① 参见汪庆华：《算法透明的多重维度和算法问责》，《比较法研究》2020 年第 6 期。

② 如有学者所指出的，“是否尽到透明度义务，判断标准也是看对于受影响者而言，信息披露是否容易理解，易于接收”。参见刘文杰：《何以透明，以何透明：人工智能法透明度规则之构建》，《比较法研究》2024 年第 2 期。

③ 参见林洹民：《论人工智能立法的基本路径》，《中国法学》2024 年第 5 期。

④ 有学者认为，在大规模通过立法、行政、司法措施规制算法的时代，算法透明原则通常既不可行也无必要。参见沈伟伟：《算法透明原则的迷思——算法规制理论的批判》，《环球法律评论》2019 年第 6 期。

⑤ 剩余数据激活指大语言模型在训练或推理过程中，未被完全“消化”的用户输入数据（如个人信息、敏感文本等），在特定条件（如通过优化提示、攻击算法）下被重新“唤醒”并提取的风险。输入文本还原指通过技术手段从大语言模型的隐藏状态、嵌入向量等中间表征中，反向推导出完整原始输入文本的过程。由于 Transformer 架构等模型存在“几乎处处可逆”的数学特性（不同输入对应唯一隐藏状态），攻击者可借助算法（如逐位反演、向量反演），从模型缓存的中间数据中精准还原用户输入的字符、句子甚至完整对话，打破“模型中间数据是抽象压缩信息”的传统认知。参见《人类的输入大模型一字未忘：Transformer 被证明“几乎处处可逆”》，https://www.toutiao.com/article/7566936297852387883/?upstream_biz=doubao&source=m_redirect，2026 年 1 月 18 日；《vec2text 技术已开源！一定条件下，文本嵌入向量可“近乎完美地”还原》，https://blog.51cto.com/baihai/14036955, 2026 年 1 月 18 日。

⑥ 以数据主体画像为基础的大数据预测分析通过复杂的机器自动处理技术进行全面追踪和分析，容易纳入敏感信息，高度介入个人隐私。因此，数据控制者如要通过足够的保障措施保证个人权利。参见谢琳：《大数据时代个人信息使用的合法利益豁免》，《政法论坛》2019 年第 1 期。

⑦ 参见陈骞、张志成：《个人敏感数据的法律保护：欧盟立法及借鉴》，《湘潭大学学报（哲学社会科学版）》2018 年第 3 期。

隐私、数据加密等技术手段[①]，既不影响数据利用效能，又能防范泄露风险，让用户直观感受到数据安全，破解“制度合规但心理不信”的困境。

3. 主体可信：“事后追责”到“全链条信任担保”。

传统“事后追责”模式侧重风险发生后的责任追究，却难以防范事前、事中的信任崩塌，且单一主体追责易因责任链条断裂导致信任修复失效。可信任导向下，强调全链条信任担保，核心逻辑在于信任需要全流程防控与多元协同兜底。人工智能从研发到运营的各环节均可能产生信任风险，且涉及多元主体参与，仅靠事后追责无法形成稳定信任预期。“全链条信任担保”特点主要包括：（1）责任的前置化与全程化，构建“研发—运营—监管”全链条的责任闭环。具体而言，开发者在人工智能研发阶段即应履行信任担保义务，开展信任风险评估，建立风险防控机制，从源头防范技术固有风险；平台方在运营阶段履行信任维护义务，实时监测算法决策的公平性，及时回应用户信任诉求；监管机构在监管阶段履行信任监督义务，建立信任评估体系并公开评估结果。[②]（2）责任的多元化与协同化。信任担保责任由多元主体共同承担，其中既包括开发者、平台方的直接责任，也包括第三方评估机构的认证责任、行业协会的自律责任。多元责任主体的协同，有助于降低单一主体违约导致的信任崩塌风险。（3）以信任修复作为责任承担的重要内容。责任主体在承担赔偿、处罚等责任的同时，还需采取公开道歉、完善制度、技术整改等措施修复受损的信任关系。例如，当人工智能算法引发歧视，相关企业不仅要对受损害用户进行赔偿，还需要公开整改方案并邀请第三方监督，以重建用户信任。

4. 治理可信：从“静态规则”到“动态信任评估”。

传统“静态规则”模式难以适配人工智能技术迭代快、场景多样的特点，易出现规则滞后于技术或一刀切规制抑制创新的问题，进而削弱公众对治理的信任。可信任导向下的监管治理强调建立动态的信任评估机制，即除了要评估人工智能应用的技术安全、效率指标、风险指标外[③]，更将信任指标（如用户信任度、权利保障充分性、信息透明度）纳入监管评估体系，通过问卷调查、第三方测评、投诉数据分析等方式，全面掌握信任状况。动态的信任评估机制具有灵活性、适应性。针对人工智能技术迭代快、场景多样的特点，“沙盒监管”等灵活监管方式，能够在保障基本信任底线的前提下为创新预留空间[④]，同时根据信任评估结果动态调整监管强度，对信任度高的企业简化监管流程，对信任度低的企业强化监管措施。动态的信任评估也强调参与性，即鼓励用户、行业协会、学界等多元主体参与信任监管，如建立用户信任反馈渠道、委托第三方机构开展信任评估、吸纳专家参与信任标准制定等。[⑤] 多元参与不仅能够提升监管的科学性与公信力，还可以让用户感受到自身在信任建构中的主体地位，进一步强化其对监管的信任。

三、可信任导向下人工智能立法的体系建构

信任包含预期稳定性与救济保障性两大核心要素，前者依赖法律对权利义务的明确界定，后者依赖法律对侵权行为的惩戒与对权利的救济。脱离了法律的制度保障，信任只能停留在道德层面，无法形成稳定的社会秩序。法律对人工智能的规制，本质上是对人工智能时代信任危机的回应，即通过设定规则确保算法透明、数据安全、责任明确，一方面重构公众对于人工智能技术的信任，另一方面为技术应用的不断升级提供稳定的制度预期。未来人工智能立法体系的建构，需在可信任目标的指引下，构建体系化的可信任导向原则群，打造精细化的可信任核心制度，实现可信任人工智能的价值引领与实践落地。

---

① 差分隐私保护是一种隐私保护技术，其能够通过随机算法对查询输出进行干扰处理以实现隐私保护，使攻击者无法根据查询输出结果判断该条记录是否存在于数据集内。参见何贤芒等：《差分隐私保护参数 ε 的选取研究》，《通信学报》2015年第12期。

② 欧盟的《人工智能法案》为高风险人工智能系统设计了“全链条”的监管制度，即自人工智能系统开发出来直至投放市场或投入使用前乃至整个存续期间都要接受有关部门的监管。人工智能的“全链条”规制周期细分为四大阶段：系统研发阶段、风险评估阶段、CE标志（欧洲共同市场安全标志）阶段、售后监测阶段。参见董新义、梅贻哲：《“人工智能法总则”建构原则与理念——欧盟立法经验之镜鉴》，《数字法治》2024年第2期。

③ 参见皮勇：《欧盟〈人工智能法〉中的风险防控机制及对我国的镜鉴》，《比较法研究》2024年第4期。

④ 参见曹建峰：《迈向可信AI：ChatGPT类生成式人工智能的治理挑战及应对》，《上海政法学院学报（法治论丛）》2023年第4期；董慧娟、丁丽文：《沙盒监管嵌入中国人工智能风险治理体系的必要性及其标准化机制建构》，《世界社会科学》2025年第6期。

⑤ 参见黄新华、温永林：《算法规制的善治之道：缘起、挑战与路径》，《东南学术》2023年第2期。

（一）立法目的条款的明确化

立法目的条款是整部法律的灵魂，能为规则设计与适用提供根本遵循。未来立法有必要在总则部分明确规定“本法的立法目的是建构人工智能技术的可信任应用秩序，保障公民、法人和非法人组织的合法权益，促进人工智能技术的健康发展”。一旦将“可信任”确立为人工智能立法的核心价值与根本目标，效率提升、安全保障、创新激励将成为信任建构的自然结果与内在支撑，各分则条款也将围绕信任的建构展开，避免规则设计的价值偏离与逻辑混乱。

（二）构建可信任导向的原则群

立法原则作为连接立法目的与具体制度的桥梁，其体系化构建直接关系到可信人工智能立法的实践效能。鉴于人工智能技术迭代快、应用场景多元、风险形态复杂，同时考虑可信任目标对于规则确定性、适配性与协同性的多重诉求，单一的立法原则难以覆盖全链条信任建构的需求，必须围绕可信任这一核心目标，构建兼具统领性、差异化、适配性与协同性的原则群，为制度设计提供全面精准的指导，相关原则至少包括：（1）信任建构原则。该原则是整个原则群的核心与统领，其要求所有立法制度与规则设计均将培育和强化人工智能技术的可信任性作为出发点，确保技术应用符合公众合理的信任预期，避免制度设计偏离可信任的核心目标。（2）分级分类原则。该原则以“风险等级”为核心划分标准，结合人工智能产品的技术复杂度、自主决策程度等固有特性，建立“高/中/低”三级通用分类体系。通过为不同类别设定差异化的基础可信性标准与合规底线（如高风险类需强制备案、中低风险类实行事后监管），破解“一刀切”规制模式可能导致的创新抑制或风险放任困境，为全领域人工智能治理提供统一、清晰的分类依据。（3）场景适配原则。在分级分类的通用框架基础上，聚焦具体应用场景对人身权益、财产安全的影响强度等本质差异，动态调整同一风险等级下规则的适用尺度。例如，高风险的医疗诊断场景需侧重诊断数据全生命周期安全与结果可解释性，司法量刑场景需强化算法公平性审查与责任追溯机制，商业推荐等低风险场景则可简化合规要求、兼顾技术应用灵活性。其核心在于实现“一类多策”，让可信性规则与场景实际需求精准匹配。[①]（4）多元协同原则。该原则要求明确政府、市场、社会在信任建构中的权责边界，通过构造政府监管、企业自律、社会监督、公众参与的多元协同治理格局，整合各方力量形成信任建构合力，弥补单一主体治理的局限性。上述四项原则相互支撑，共同构成适配人工智能治理需求的原则体系。

（三）可信任制度的规范落地

1. 算法可信制度：以透明、公平、可解释为核心。

算法是人工智能技术的核心载体，其可信性直接决定了人工智能技术应用的信任基础，而算法透明、公平、可解释正是算法可信的核心要义与必备前提，缺乏这三者，技术应用便会陷入“黑箱困境”，难以构建公众信任。不过，算法应用的场景异质性、风险梯度差异及信任需求多元性，决定了对透明、公平、可解释的要求不需采取“一刀切”的最严模式，而是可以结合具体场景，构建差异化的算法可信制度。[②] 具体而言，不同场景中算法的决策影响力、风险传导路径与公众的信任期待存在显著差异。司法量刑、医疗诊断、金融信贷等高风险场景中，算法决策直接关联人身权益、财产安全等核心利益，公众对算法的可解释性、公平性、可追溯性要求极高；而商业推荐、信息检索、娱乐互动等低风险场景中，算法影响多限于用户体验层面，过度强调严苛的可信标准反而会增加合规成本、抑制技术创新。同时，算法本身的复杂度、自主程度也因场景而异，通用大模型的跨场景应用与垂直领域专用算法的功能定位不同，其可信性保障的重点与实现路径自然存在差异。

为此，未来立法需结合场景特性，对算法透明、公平、可解释的要求作出差异化界定。在算法备案方面，高风险算法适用强制备案制度并接受监管部门的实质审查，审查重点应聚焦算法的公平性、安全性与可解释性；一般算法实行事后备案与动态监测机制即可[③]，在保障信任底线的同时兼顾企业合规成本与创新效率。在

① 场景化规制的必要性，参见丁晓东：《论算法的法律规制》，《中国社会科学》2020 年第 12 期。

② 参见张欣：《从算法危机到算法信任：算法治理的多元方案和本土化路径》，《华东政法大学学报》2019 年第 6 期。

③ 为避免备案成为变相的审批，可实行事后备案而非事前备案。并且，人工智能大语言模型具有不断进化的能力，将静态的一时的算法予以备案，不足以帮助监管部门了解算法的后续进化，故在事后备案的基础上可实行动态检测机制。参见林洹民：《论人工智能立法的基本路径》，《中国法学》2024 年第 5 期。

算法可解释性的要求上，对于高风险应用场景，算法开发者需以清晰易懂的方式向用户说明决策依据、数据来源与影响因素，充分保障用户的知情权与异议权；对于中低风险应用场景，开发者仅需说明算法的核心运行逻辑与数据使用范围，无需披露复杂技术细节。

值得注意的是，公平作为算法可信的核心底线，其保障不能依赖单一主体自律，尤其在涉及权利义务分配的场景中，算法歧视具有隐蔽性、技术性等特点，仅靠开发者自查或监管部门抽查难以全面覆盖风险，这就需要通过专门机制筑牢公平防线。因此，立法应明确引入第三方评估机构，针对不同风险场景设定差异化评估标准。例如，对于高风险场景算法可开展高频次、全维度反歧视评估，对中低风险场景算法进行常态化抽检，同时建立便捷的算法歧视投诉与纠正机制，确保用户能及时反馈问题，存在偏见的算法模型可被快速核查、整改，最终实现不同场景下算法公平底线的刚性守护，为算法可信目标提供支撑。

2. 数据可信制度：覆盖全生命周期的安全保障。

数据从收集到销毁的各环节环环相扣，任何一个环节的信任缺口都可能引发全链条的隐私泄露与信任崩塌，唯有实现全流程规制才能筑牢数据可信的根基。立法应完善数据全生命周期安全规则，具体包括：在数据收集环节，明确要求数据处理者的行为应严格遵循知情同意和目的限制原则，不仅需具备合法依据，还必须向用户明确告知数据用途与使用范围。在数据存储环节，需强化数据处理者的安全防护义务，要求其采用加密存储、访问控制等技术措施防范数据泄露。在数据传输环节，应规范数据的跨境与境内流转规则，建立数据传输安全评估机制，确保数据流转的安全性。在数据销毁环节，应明确规定销毁标准与流程，要求数据处理者确保数据彻底删除且不可恢复。此外，立法需对敏感数据建立特殊保护制度，对生物识别、个人信用、医疗健康、实时位置等敏感数据实行最高级别的可信保护，严格限制其收集范围与使用场景，未经法定程序不得擅自处理。立法还应构建数据可信共享机制，鼓励依托区块链等技术手段①，实现“数据可用不可见”的可信共享模式，在充分保障数据隐私安全的前提下促进数据要素合理流动。

3. 主体可信制度：明确权责边界与信用约束。

多元主体的可信履约是信任建构的关键，人工智能技术的研发、应用与监管涉及多主体协同参与，任何一方的权责缺位或履约失范，都可能击穿信任链条、侵蚀公众对技术的信任基础。“共建共治共享”是新时代社会治理的指导思想和根本遵循，这一科学思想对人工智能治理更具有根本性、指向性意义。秉持“共建共治共享”理念，在人工智能多元共治新格局中，监管机构、企业、社会组织、公众都是治理的“主体”②，立法需通过精准的制度设计明确主体权责、强化信用约束，以多元共治筑牢信任根基。具体而言，立法应明确多元主体的权责清单：其一，研发者需承担安全设计义务。作为技术源头主体，研发者在研发阶段主动嵌入可信性保障措施、开展信任风险评估与测试，是从源头防范技术固有风险的前提；其二，鉴于使用者直接掌控技术应用场景，其合规履职能有效阻断技术滥用风险，立法应当要求使用者承担合规使用义务，不得超出约定范围使用人工智能技术；其三，为弥补市场自律的不足，监管者需承担动态监管义务，结合人工智能技术迭代快的特性，运用技术手段开展常态化监管、及时处置各类可信性风险。

人工智能侵权案件等纠纷处理中，受害人经常因为“技术黑箱”而面临举证困境③，为减轻受害人的举证负担，强化对信任受损方的权利救济，倒逼企业主动履行可信性义务，立法可区分人工智能系统设计、开发、训练、部署、运行、迭代、退役七大环节，将有针对性的责任机制精细化地嵌入各环节，如在设计开发阶段适用过错推定责任，在训练与部署阶段实行严格责任与过错责任，而当系统进入直接面向公众的运行阶段，风险系数显著上升，归责原则随之转向以严格责任为主。④ 与此同时，有必要建立主体信用评价机制，将市场主体的可信性义务履行情况纳入企业信用档案，对严格履行义务、信任度高的主体给予政策激励，对违反可信性要求、造成信任损害的主体依法实施失信惩戒，形成正向引导与反向规制的双重合力，推动多元主

① 区块链技术支撑下的新型“数据账本”中，数据加密技术使得用户可以通过“秘钥”来对个人数据进行保护与授权。参见周茂君、潘宁：《赋权与重构：区块链技术对数据孤岛的破解》，《新闻与传播评论》2018年第5期。

② 参见张吉豫：《构建多元共治的算法治理体系》，《法律科学（西北政法大学学报）》2022年第1期。

③ 参见苏宇：《算法规制的谱系》，《中国法学》2020年第3期。

④ 参见周辉：《人工智能综合性立法及其实现》，《法学研究》2025年第6期。

体主动守信。

4. 治理可信制度：强化监管效能与协同治理。

要将可信任导向的立法条文转化为实际治理效能，筑牢公众对人工智能治理的信任根基，还需构建权责清晰、手段适配、内外协同的治理体系。

首先，为适配人工智能技术迭代快、应用场景隐蔽的治理特性，应当构建技术赋能监管体系。传统人工监管难以应对算法“黑箱”以及数据海量流转带来的监管盲区，立法可以要求监管部门建立智能监管平台，实时监测算法运行轨迹、自动预警数据异常流转等，以提升监管的精准性与效率。此外，可信任标准并非一成不变的静态要求，而是应具备动态调整的弹性空间。立法可以采用“原则性规定+配套细则+动态更新”的模式，授权相关监管部门根据技术发展情况，制定和更新具体的技术标准与实施细则。①

其次，立法需以畅通的救济渠道修复受损信任、强化制度公信力。例如，为解决普通主体难以举证算法缺陷、数据滥用等问题，可规定设立人工智能纠纷专门仲裁机构，吸纳技术专家、法律专家组成仲裁庭；在诉讼程序中引入技术专家陪审员制度，帮助法官准确理解算法原理、数据处理逻辑等专业问题；建立多元化的纠纷解决机制，整合行政调解、仲裁、诉讼等救济渠道，为信任受损方提供分层便捷高效的权利救济途径。

最后，人工智能技术的跨境应用日益频繁，由于不同国家和地区的法律体系、价值观念存在差异，其可信任标准也可能存在分歧，这就导致人工智能产品和服务在跨境流动时面临标准不互认的壁垒。对此，应秉持协同治理理念，积极推动建立国际的可信任标准互认机制，提升跨境人工智能应用的信任水平。未来我国应积极参与全球人工智能可信治理规则制定，主动对接欧盟《人工智能法案》、GDPR 等国际先进立法经验，在数据安全、算法公平、权利保障等核心领域推动我国信任标准与国际规则的互认互通；针对数据跨境流动、跨境人工智能服务等高频跨境场景，建立双边或多边的信任协同机制，既有效缓解各国对“数据主权威胁”的焦虑，又通过规则对接、标准互认、执法协作等方式，降低跨境技术应用的信任成本与合规成本。

结语

可信任作为人工智能治理的底线，核心是要构建“相信无风险”的稳定预期，这种预期并非源于风险的彻底消除，而是建立在风险可控、损失可补救、责任可追溯的制度基础之上。只有让每一次信任受损都获得公正救济、每一项技术应用都有明确主体负责，才能实现安全、效率与创新的动态平衡，为人工智能技术可持续发展保驾护航。福山的信任资本理论早已证实，高信任度的社会环境能够显著降低协作成本、提升创新效率。对于人工智能产业而言，立法构建的可信任环境，能够消除用户对技术应用的安全顾虑，提升用户对人工智能产品的接受度，扩大技术应用的市场空间；同时，稳定的信任预期能够吸引更多资本与人才投入人工智能研发，减少因规则不确定性导致的创新观望，推动技术向更高质量、更普惠的方向发展。假以时日，人工智能产业必将进入“高信任—低风险—强市场”的良性循环，在充分释放技术效率价值的基础上，实现安全与发展的动态平衡。

习近平总书记指出：“必须更好发挥法治固根本、稳预期、利长远的保障作用，在法治轨道上全面建设社会主义现代化国家。”② 人工智能立法的过程也是通过制度设计培育多重信任、提供稳定预期、促成长远合作的过程。本文以法律的信任建构功能为大前提，围绕可信任是人工智能立法的核心目的这一核心命题，构建了以可信任为轴心的规范体系。此项研究的价值不仅在于为人工智能立法提供兼具理论自洽性与实践可行性的制度方案，更在于回归法律的本质功能，为技术与社会的良性互动提供稳定的制度预期。未来研究可进一步聚焦生成式人工智能、自主智能体等新兴技术形态的可信性规制难题，细化差异化治理规则，推动法律治理与技术发展同频共振，助力人工智能产业在信任框架下实现高质量发展。

（责任编辑：邱小航）

（下转第 115 页）

① 参见周辉：《人工智能综合性立法及其实现》，《法学研究》2025 年第 6 期。

② 习近平：《习近平法治文选》第 1 卷，北京：中央文献出版社，2025 年，第 334 页。

索放在一个整体性框架中加以聚焦、凝视和解读，才能彰显社会学的想象力。近 200 年前，社会学应运诞生于 19 世纪 30 年代的大转型时代；今天，新的大转型时代正呼唤着社会学的新使命。

（责任编辑：朱 颖）

The Era of Great Transformation and the Social Reconstruction of Reality: A Perspective from the Sociology of Knowledge

ZHOU Xueguang

Abstract: From the second half of the twentieth century to the present, the world has been undergoing an era of profound transformation. Although this epochal shift is most visible in international relations and the political sphere, its driving force lies in the deep changes in the mechanisms of social construction of reality over the century. The sociology of knowledge provides a theoretical perspective and corresponding analytical framework for understanding this era of great transformation. As the birthplace of modernity, Western societies experienced a dual trend of (1) the rise of nation states and the associated institutional centers in society and (2) the de-centering movement during the twentieth century. On the one hand, this process promoted the global expansion of capital and institutional rationality; on the other hand, it fostered the rise of individual rationality, subjectivity, and agency. This, in turn, generated profound challenges to modernity and stimulated waves of social critique across different fields. The encounter and interaction between these currents of social change and the rise of the internet and digital media since the 1990s have created "a perfect storm." This has driven fundamental transformations in the foundations of contemporary society and set in motion the gears of the great transformation.

Key words: great transformation, institutional center, de-centering, social construction, sociology of knowledge

---

（上接第 101 页）

Reconstructing Legislative Purposes and Normative Evolution of Artificial Intelligence Legislation from a Trust Worthiness Perspective

WANG Yi

Abstract: Artificial intelligence legislation has long been trapped in the tripartite balance paradox of efficiency, security and innovation. Confined to the linear logic of value trade-offs, the existing regulatory approaches struggle to break through this dilemma. The essence of law lies in a construction mechanism of social trust. By clarifying rights and obligations, defining boundaries of conduct, and guaranteeing remedy for rights, law delivers stable expectations, which inherently aligns with the governance demands of the artificial intelligence era. Trustworthiness is not a fourth dimension independent of the three core values; instead, it serves as the logical hub that integrates the three to achieve dynamic coordination. Establishing trustworthiness as the core objective of artificial intelligence legislation can not only resolve the inherent contradictions of the traditional balance model, but also address trust crises such as algorithmic black boxes and data abuse. Based on this core proposition, a dedicated artificial intelligence law should be formulated in the future. With trust building as its legislative purpose, it shall establish a set of principles including trust construction, hierarchical and classified regulation, scenario adaptation and multi-party collaboration. Through refined institutional design covering trustworthy algorithms, credible data, accountable entities and sound governance, it will provide steady value guidance for the rule of law in artificial intelligence and foster positive interaction between technological application and social trust.

Key words: artificial intelligence legislation, trustworthiness, legal function, collaborative governance