人工智能治理准则中的文化价值差异与共识构建研究

刘祖琼
(仰恩大学,福建泉州　362014)

[摘　要]人工智能的迅速普及,推动着人工智能治理相关议题从学术研讨层面迈向实际操作阶段。这一分化的原因在于不同文明传统对“何为善治”存在根本性理解差异,集中体现在个人权利与社会福祉、对抗性透明与关系性信任、技术中立与价值嵌入三个方面,其成因涉及文化传统、制度路径与权力政治三重机制。在此基础上,提出“元准则、操作准则、执行协调”三层治理架构,即元准则提供了语言对话的合法性基础,操作准则处理实质分歧,执行机制保障落地与进化,三层架构环环相扣,共同构成一条既尊重文化多元又追求治理实效的共识构建路径。

[关键词]人工智能治理;文化价值差异;共识构建;等效互认;文化例外
[中图分类号] TP18; G0; B82-057　[文献标识码] A　[文章编号]2096-711X(2026)16-0126-03
doi:10.3969/j.issn.2096-711X.2026.16.043　[本刊网址] http://www.hbxb.net

引言

自 2023 年以来,生成式人工智能的迅速普及,推动着人工智能治理相关议题从学术研讨层面迈向实际操作阶段。如,2024 年欧盟《人工智能法案》正式通过;中国持续推进算法备案工作,在《互联网信息服务深度合成管理规定》对深度合成技术进行专项规制的基础上,进一步以《生成式人工智能服务管理暂行办法》将生成式 AI 纳入综合治理框架;2024 年,联合国大会通过首个人工智能决议,联合国教科文组织发布《人工智能伦理问题建议书》并启动全球实施进程;2025 年 AGILE 指数对 40 个国家的人工智能治理能力的评估,相关研究指出,全球人工智能治理不能采取一刀切模式,相关策略必须充分考虑文化多样性、社会规范和制度遗产等。

在这一背景下,“负责任的人工智能”“以人为本”“公平透明”等元原则在全球范围内获得了广泛共识,但落实到监管框架、技术标准、评估指标等操作层面时,仍然产生了分歧。这种分歧不是简单的利益博弈或技术路径差异,其根源在于不同文明传统和社会制度对“何为善治”有着迥然不同的理解。不同学者对此已有初步探索:蔡翠红、尹嘉慧(2025)从中西治理分歧的文化根源出发,指出儒家伦理与自由主义分别塑造了集体导向、政府主导与个体导向、多方共治的两种治理模式;邱嘉璇、程乐(2026)运用社会符号学方法分析了 47 份国际人工智能治理规范文本,发现安全(Safety)、以人为中心(Human-centric)与公平(Fairness)三个核心原则在文本层面趋于共识,但在符号协议与实际执行之间存在显著张力,使其可执行性受到削弱。从现有研究来看,多聚焦于对差异的“描述”和“解释”,对“如何在承认差异的前提下构建可操作的共识”问题回答得仍不充分。

基于此,对全球人工智能治理准则体系中的文化价值差异展开剖析,进而提出“元准则、操作准则、执行协调”的三层递进式治理架构,旨在探索兼顾文化多样性与治理协同性的全球人工智能治理共识构建路径,为推动全球人工智能治理的有序发展提供理论参考。

一、全球人工智能治理准则体系中的文化价值差异

要理解人工智能治理领域文化价值差异的深层逻辑,首要任务是明晰差异分布的结构性特征。以下将从治理范式的宏观分化与核心价值的微观张力两个维度展开剖析,并辅以典型案例进行阐释。

(一)治理范式的宏观分化

目前,全球人工智能治理过程中形成三种典型范式。比如,欧盟确立“基于权利保护的风险规制模式”,以《人工智能法案》为标志,将人工智能系统按风险等级分为不可接受风险、高风险、有限风险和最小风险四级,其核心价值是个人基本权利的优先保护,制度逻辑是通过事前规制预防风险的发生。美国则是通过联邦层面的“去监管化倾向”,以 2025 年特朗普政府系列行政令为代表,联邦层面转向激进放松监管以保持创新竞争力,这直接引发了与州级强监管立法的正面冲突——以加州为代表,多个州通过透明度披露、算法歧视风险评估等法案形成自下而上的补位格局。中国则逐步形成统筹发展和安全的“渐进式立法模式”,通过《生成式人工智能服务管理暂行办法》等专项立法,实施分类分级监管与算法备案制度,在促进产业创新与防范安全风险之间寻求动态平衡。

上述三种典型范式的差异,不仅体现在制度文本层面,更体现在治理目标的优先排序上。欧盟把权利保护置于发展速度之上,美国把创新领先置于统一规制之上,中国则把安全可控与产业发展置于同等重要位置。这种排序差异背后,正是不同文化传统对个人与国家、权利与发展、自由与安全之间关系的不同理解。

(二)核心价值的微观张力

治理范式的分化,根源在于核心价值维度的结构性张力,可将其归纳为三个方面。

第一,个人权利优先与社会福祉优先。欧盟 GDPR 根植于人格尊严传统,将隐私权作为基本权利以制度性保障加以维护。美国算法公平性框架侧重限制政府干预,保障个人自主性。两者共享自由主义传统但路径不同,隐私权、知情权和不受算法歧视的权利均被视为不可妥协的底线。中国生成式人工智能管理办法则将维护国家安全和社会公共利益置于优先地位,反映集体福祉优先的价值取向。

第二,对抗性透明与关系性信任。美国人工智能治理以制度化的不信任为驱动力,通过信息披露、算法审计和独立评估等方式构建对抗性问责机制,将透明度建立在事后追责之上。中国则更倾向于通过政企持续对话、行业自律公约和沙盒试点等柔性机制,以协调配合替代对抗制衡。

第三,技术中立预设与价值嵌入自觉。在美国自由主义

收稿日期:2026-6-5
基金项目:本文系 2025 年度福建省中青年教师教育科研项目(社科类)一般项目(项目编号:JAS25180)。
作者简介:刘祖琼(1983—),女,福建泉州人,副研究员,主要从事高等教育管理与科技伦理研究。

传统中,技术常被预设为中立工具,治理重心落在事后纠偏上,发现并修正算法运行中的歧视性结果。中国则将人工智能系统视为价值载体,要求在设计阶段就将伦理标准嵌入技术架构。欧盟选择了一条中间道路:以可信人工智能为核心,通过前置合规与伦理评估,在技术中立与价值嵌入之间寻求制度性平衡。三种路径的分歧,归根结底是价值进入技术的时机问题,是事先内置、事后校准,还是两端兼顾。

这三种张力并非简单的政策工具差异,而是不同文明传统对权利与善治关系的根本预设不同。

(三)差异的生动体现

各国人工智能治理道路的分歧,本质上不是技术标准之争,而是对"何为善治"这一根本问题的不同回答。比如,人脸识别在公共空间的应用,折射的远不是技术标准之争。欧盟《人工智能法案》将其列为不可接受风险直接禁止,着眼点在个人隐私与自由。中国在公共安全领域广泛应用,法律框架的重心是规范使用边界而非禁用技术本身。即便在共享自由民主话语的欧美之间,也存在差异:美国多个城市立法禁止政府使用监控,英国却因悠久的公共监控传统而对此习以为常。制度相近不足以推导治理趋同,文化惯性和历史路径往往更具决定作用。这些分歧,说到底不是规则差异,而是不同文明传统对公共秩序与个人自由等基本价值给出了不可通约的权重。当我们在争论AI治理准则该统一到什么程度时,实则是在追问一个更根本的问题:谁的善治理念,有资格被写入全球规则。

二、差异形成的原因

(一)文化传统的深层塑造

文化传统是塑造人工智能治理差异的深层力量。中国治理中集体福祉优先的倾向,扎根于儒家伦理对社群关系的强调。在儒家视野中,个体始终嵌于关系网络,这就是为何中国监管框架在个人数据保护之外,更关注算法对公共舆论的宏观效应。欧美自由主义传统则从先验权利主体出发,将划定技术不可逾越的个人权利边界视为治理的首要关切。两种进路的分歧不在政策工具层面,而在各自文明对"人"的基本预设不同。一个从关系中理解人,一个从自主个体定义人。正是这一根本差异,决定了两种治理框架在价值排序和方法选择上的系统性分野。

(二)制度路径的锁入效应

文化传统塑造的价值取向一旦进入操作层面,便不只是观念偏好的问题,而是被制度路径稳稳地"锁入"了治理选择。付新华(2025)指出,中国逐步形成了"统筹发展与安全、靶向治理与机制协同的渐进式立法模式"。在具体实践中,该模式通过《生成式人工智能服务管理暂行办法》等专项立法,实施分类分级监管与算法备案制度,力求在促进产业创新与防范安全风险之间寻求动态平衡。美国联邦层面长期去监管化,根源在于其普通法司法能动与市场自治的历史惯性。制度一旦成形便倾向于自我强化,治理差异由此从价值偏好转化为难以逆转的路径锁定。更隐蔽的风险在于,这种锁定会催生治理盲视,决策者不自觉地将自身路径视为唯一合理选项,将他国的差异化选择径直判定为治理滞后或越位。

(三)权力政治的结构性驱动

当前国际人工智能标准制定中的代表性问题尤为突出,事实上治理差异不仅是文化"表达"的结果,更是权力政治"建构"的产物。薛澜、梁正(2025)的研究明确指出,当前"发达国家在人工智能技术研发和规则制定中占据主导地位,发展中国家面临被边缘化风险",并且"不平等的既有国际格局也在根本上阻碍了包容、普惠的全球治理体系的塑造"。这种结构性不平等的后果是双向的:一方面,主流人工智能治理话语体系由欧美发达国家主导,包括中国在内的"全球南方"国家的治理理念在元准则层面表达尚不充分;另一方面,发展中国家因缺乏实质性参与,其治理选择往往被主流话语简单归类为"特殊主义"或"例外",而主流话语自身的特殊性则被包装为"普世标准"。

三、跨文化共识的构建路径

通过上述分析,以"承认差异的不可消解性"为前提,"在差异中达成可操作的共识"为目标,构建治理的核心框架。即不强求单一的全球标准,而是在不同治理层面上采取不同的共识策略,通过制度设计实现差异的"互操作"。

(一)元准则层:巩固既有共识与包容性修订

UNESCO《人工智能伦理问题建议书》经193个成员国协商一致,确立了相称性、公平与非歧视、可持续性、人类监督等核心价值,是目前全球最具文化代表性的人工智能治理基础文本。治理路径的第一步,是将这一软法文本推向可操作框架,各国在制定或修订国内人工智能治理准则时,以建议书为参照文本,保持核心术语的语义通约性,避免因话语体系断裂而加剧互信赤字。

然而,现有国际伦理文件多由欧美学术传统主导起草,儒家伦理传统中的治理理念在其中表达尚不充分。未来的修订程序应以文明对话为基础,通过平等的概念互译与意义协商,使儒家的和谐观、非洲的关系性人格概念、伊斯兰的公共利益原则等治理理念在准则中获得对等表述权,确保准则在不同文化圈具有同等的感召力。

(二)操作准则层:等效互认与文化例外

其一,等效互认机制。指不同治理体系即使路径不同,只要能够实现实质等同的保护效果,即应获得相互承认。以当前全球并行的三大监管范式,欧盟"风险分级"模式、中国"分类分级监管"逻辑、新加坡"自愿性框架+沙盒试点"路径为例,三者均指向对高风险人工智能应用的审慎控制,强求规则文本的统一既不现实也无必要。

等效互认的操作方案,由国际标准化组织或IEEE标准协会等机构牵头,制定一套涵盖安全底线、透明度、可问责性、人权影响评估等关键维度的"治理等效性评估指标"。一国的监管框架经第三方评估确认达到与另一国"实质等同"的保护水平后,双方签署互认协议,降低因制度差异造成的技术合作壁垒与合规成本。这一机制将差异转化为可比较、可评估的技术参数,在尊重文化多元性的前提下保障治理的有效衔接。其法理基础在于国际私法中"等效原则"的扩展运用。与欧盟现行的单边"充分性认定"不同,此处主张的等效互认应由多方参与的国际机构主导,指标经多方协商确立,评估结论定期复审,且互认双向对称。

其二,文化例外条款与负面清单。并非所有治理差异都可以通过"等效"逻辑调和。某些分歧根植于深层文化价值的不同排序,属于"合理的分歧"。为此,在操作准则层面引入"文化例外"条款与"全球红线"负面清单,形成正反互补的弹性治理结构。

在正面,"文化例外"允许成员方就特定治理条款提出基于文化价值的差异化安排,但须满足三项前提条件:公开声明确认所依据的文化传统与价值立场;论证该安排的必要性与相称性;接受定期审查以验证其合理性。不同社会对公共空间人脸识别技术的接受阈值差异,可经由这一机制获得有约束的表达空间。在反面,负面清单划定"无论何种文化传统均须恪守"的全球底线,至少应涵盖人工智能武器化系统的研发部署、基于敏感特征的系统性歧视、大规模无意识行为操纵、深度伪造对民主程序的颠覆性干预。这一正一反的制度设计,既为文化差异保留了合法表达渠道,又以刚性红

线防止“文化相对主义”滑向“价值虚无主义”。

(三)执行协调层:代表性与争议解决

第一,多利益相关方治理理事会的代表性重构。现有国际人工智能治理平台存在结构性失衡,欧美发达国家和东亚技术大国主导议程,全球南方的话语权明显不足。这种失衡直接导致治理规则难以反映不同发展阶段国家的差异化需求。薛澜、梁正(2025)指出,全球人工智能治理面临的深层挑战之一,正是“如何协调不同发展阶段、不同文化背景国家的治理需求”,并呼吁建立“更具包容性和代表性的多边治理机制”。这一判断为制度设计提供了基本方向。拟设立的治理理事会应在代表权分配上遵循文明圈层均衡性、利益相关方多元性和知识生产地域正义性三项原则,特别要确保原住民社群的文化传统和宗教文化传统不被系统性地忽视。

第二,争议解决的双轨制设计。治理体系运转中不可避免会出现争议,争议解决机制的设计本身也是文化价值的前沿阵地。建议设立双向争议解决程序,首先处理技术性争议,由专业技术专家组依据科学证据,并在此基础上通过工程共识形成判断,裁定涉及技术标准、风险评估、算法测试等可实证化的问题;其次处理价值性争议,当争议实质触及不同文化价值排序时,不追求判定某一方“正确”或“错误”,而是进入由跨文化伦理委员会主持的对话程序,将分歧转化为“在特定条件下可接受的差异化安排”,并以书面备忘录形式固定,为未来类似争议提供参照。邱嘉璇等(2026)的跨文化规范分析也表明,全球人工智能治理准则的“合法性”本身是一个动态建构过程,需要通过持续的跨文化对话来维持和再生产。因此,双轨制设计的要义在于:不让技术问题被政治化,也不让价值问题被技术化遮蔽。

第三,制度废止与衔接。随着技术不断更新迭代,制度跟不上技术是常态。因此,给规则装上定期更新的引擎,初衷是防止制度僵化。但需要警惕两种风险。其一,并非所有准则都因技术突破而过时,公平、透明等基本原则需审慎稳定,只有技术性规范才需灵活响应。其二,失效后的规则真空,往往比僵化更危险。迭代机制不仅要解决“何时废”,更要设计“废后如何无缝衔接”,否则更新本身就可能成为治理中断的源头。换言之,制度弹性与程序刚性之间的平衡,才是关键所在。

四、结论与展望

人工智能治理准则的分歧,表面上是技术标准之争,实际上是不同文明传统对“何为善治”给出了不同的回答。儒家伦理将人视为关系网络的节点,治理的重心落在社群和谐;自由主义传统将个人自主视为不可逾越的底线。这类价值预设并不悬浮于观念层面,而是经由制度路径的锁定效应,转化为自我强化的治理轨道。因此,可得出结论,差异不会因技术演进而自动消弭,正视差异是共识构建的真正起点,而非障碍。

展望未来,跨文化人工智能治理共识的构建,需要在两个方面同时协作。向下,守护底线准则——公平、安全、透明,以元规则的形式获得更广泛的文明背书,为操作层面的灵活性提供不可退让的伦理锚点。向上,操作规则需要为文化差异留出制度化的合法空间,等效互认与文化例外并非策略性妥协,而是让不同治理传统得以在同一个框架内对话的机制前提。当然,不排除因地缘政治的张力,随时可能将制度善意误读为规则博弈。毕竟,技术演进的速度始终快于制度协商的节奏,但治理的成熟不在于消除差异,而在于让差异从对抗的源头,转变为彼此理解与持续对话的资源。

参考文献:

[1] Yi Zeng, Enmeng Lu, Xiaoyang Guo, et al. AI Governance International Evaluation Index 2025 (AGILE Index 2025)[J]. Societal Impacts, 2025,6: 100154.

[2]Cai Cuihong, Yin J. Cultural and ethical foundations of AI governance divergence: A comparative analysis of China and the West[J]. Política Internacional,2025, VII(1):215-233.

[3]Jiaxuan Qiu, Le Cheng. Consensus and legitimation in global AI regulations: A sociosemiotic perspective [J]. International Journal of Law in Context, 2026:1-22.

[4]付新华. 全球人工智能立法的多元趋势与中国模式[J]. 交大法学,2025(6):60-73.

[5]薛澜,梁正. 构建包容、普惠的全球人工智能治理体系:现实挑战与对策建议[J]. 人民论坛·学术前沿,2025(9):23-32.

Research on the Cultural Value Differences and Consensus Building in AI Governance Principles

LIU Zu-qiong

(Yang-En University, Quanzhou Fujian 362014, China)

Abstract: The rapid popularization of artificial intelligence has pushed the related issues of AI governance from the academic discussion stage to the practical operation stage. The reason for this differentiation lies in the fundamental differences in understanding of "what constitutes good governance" among different civilizations, which are mainly reflected in three aspects: individual rights and social welfare, adversarial transparency and relational trust, and technological neutrality and value embedding. The causes involve three mechanisms: cultural traditions, institutional paths, and power politics. On this basis, a three-layer governance framework of "meta-principles, operational principles, and implementation coordination" is proposed. The meta-principles provide the legitimacy basis for language dialogue, the operational principles handle substantive differences, and the implementation mechanism ensures implementation and evolution. The three-layer framework is interlinked and together constitutes a consensus-building path that respects cultural diversity while pursuing governance effectiveness.

Key words: AI governance; cultural value differences; consensus building; equivalent mutual recognition; cultural exception

(责任编辑:桂杉杉)