Advances in Philosophy 哲学进展, 2026, 15(8), 385-392
Published Online August 2026 in Hans. https://www.hanspub.org/journal/acpp
https://doi.org/10.12677/acpp.2026.158406

Hans汉斯

人工智能嵌入社会治理的算法偏见风险与规制路径探析

张晋翔[1]，李明珍[2]

[1]新疆师范大学政法学院，新疆 乌鲁木齐
[2]新疆师范大学历史与社会学院，新疆 乌鲁木齐

收稿日期：2026年7月24日；录用日期：2026年8月15日；发布日期：2026年8月25日

摘　要

人工智能嵌入社会治理的作用日益凸显，但其伴随的算法偏见问题对社会公平公正构成了挑战。算法偏见主要源于算法模型设计中融入的个人主观偏见、算法训练数据里内刻的社会隐性偏见以及算法实际使用后产生的机器学习偏见，可能引发法律与伦理的争议、侵犯大众的隐私与权利乃至损害政府形象与公信力。需要通过技术管控使算法数据透明化，通过制度规制使算法程序公正化，通过可问责的算法结果进行责任约束，并通过可升级的算法架构对算法整体进行迭代优化。在此基础上，持续推进全面深化改革，着力完善科技伦理治理体系，提升治理能力现代化水平，为实现治理现代化提供坚实的技术治理保障。

关键词

技术伦理，人工智能，社会治理，算法偏见

Algorithmic Bias Risks and Regulatory Pathways in the Integration of Artificial Intelligence into Social Governance

Jinxiang Zhang[1], Mingzhen Li[2]

[1]College of Political Science and Law, Xinjiang Normal University, Urumqi Xinjiang
[2]School of History and Social Sciences, Xinjiang Normal University, Urumqi Xinjiang

Received: July 24, 2026; accepted: August 15, 2026; published: August 25, 2026

文章引用: 张晋翔, 李明珍. 人工智能嵌入社会治理的算法偏见风险与规制路径探析[J]. 哲学进展, 2026, 15(8): 385-392. DOI: 10.12677/acpp.2026.158406

Abstract

The role of artificial intelligence (AI) embedded in social governance is increasingly prominent, yet the accompanying problem of algorithmic bias poses a challenge to social fairness and justice. Algorithmic bias mainly stems from personal subjective biases incorporated into algorithmic model design, implicit social biases embedded in algorithmic training data, and machine learning biases generated after the actual application of algorithms. These biases may trigger legal and ethical controversies, infringe upon public privacy and rights, and even undermine the image and credibility of the government. Therefore, it is necessary to achieve algorithmic data transparency through technical governance, ensure procedural fairness of algorithms through institutional regulation, enforce accountability through accountable algorithmic outcomes, and achieve iterative optimization of the overall algorithm through upgradeable algorithmic architectures. On this basis, we should continue to advance comprehensive deepening reform, focus on improving the science and technology ethics governance system, and elevate the modernization level of governance capacity, thereby providing solid technical governance safeguards for realizing governance modernization.

Keywords

Technology Ethics, Artificial Intelligence, Social Governance, Algorithmic Bias

1. 引言

人工智能作为当下新一轮科技革命的核心热点，在社会治理的应用中展现出了深远的影响力和潜力，正成为提升治理效率和决策精准性的重要工具。但其作为新兴技术，在嵌入社会治理的过程中仍存在一些亟待解决的问题：算法在处理大量数据并作出决策的过程中可能不自觉地嵌入并放大了既有的社会偏见，潜藏着算法偏见风险，乃至于可能引发一系列的科技伦理问题。加强科技伦理治理，既是科技不断发展的需要，更是国家战略层面的明确要求。2022 年 3 月 20 日，中共中央办公厅、国务院办公厅印发了《关于加强科技伦理治理的意见》[1]，要“加快构建中国特色科技伦理体系，健全多方参与、协同共治的科技伦理治理体制机制”[1]。在这样的背景下，对人工智能在社会治理中的应用进行深入的伦理审视，特别是对其可能带来的算法偏见及治理路径进行探析，助力推动科技伦理治理体系与治理能力现代化，这既是坚持进一步全面深化改革的应有之义，也是对推进治理现代化的积极探索。

2. 问题的提出

1955 年，在一份题为《关于举办达特茅斯人工智能暑期研讨会的提议》的建议书中首次出现了“人工智能”这一术语，并将其定义为“使机器使用语言、形成抽象概念，解决目前仅由人类解决的问题，并且能自我改善”[2]在此基础上，对人工智能的定义在目的论维度上达成了相对一致的概念认同，即人工智能是“使机器能够模仿各种复杂的人类技能的技术”([3], p. 15)。而在其他层面，对人工智能的定义仍存在着许多争议。计算机科学家 Nils John Nilsson 将人工智能视为使机器能够在其环境中“适当地行事，

[1]https://www.gov.cn/zhengce/2022-03/20/content_5680105.htm

并具有远见”([3], p. 16)的一种技术。欧盟委员会下属的人工智能高级专家组也提出了相似的定义：“通过分析环境并采取行动——以某种程度的自主性——实现特定目标的系统”[4]。但有批评者认为这些定义太过宽泛，缺乏足够的精确度。并且随着技术的发展和研究的深入，人工智能的定义可能还会继续演化和扩展，因此难以准确完整地对人工智能的定义进行阐释。不过就其本质而言，人工智能是对人类智能的模仿，通过对人类智能的模拟、扩展或超越而实现特定的任务。

2.1. 人工智能嵌入社会治理的内在逻辑

人工智能作为新一轮科技革命和产业变革的重要驱动力量，是技术属性与社会属性的高度融合，其嵌入社会治理的内在逻辑主要体现在两方面：“一方面，人工智能为国家权力对个人和社会的监管提供了技术基础，借助人工智能可以提高国家权力能效和延伸权力范围；另一方面，人工智能也为建立公民权利保障机制创造了条件。”([5], p. 88)

技术的发展不可能超脱于现实的政治经济环境，相应地，技术的发展也会一定程度扩散到社会政治领域。这种扩散与转化的一个层级是政府层面主动、有意识地将技术吸收、融入自身运行机制中；另一个层级则是技术的快速发展倒逼政府进行社会治理转型以重新适应成为合理的行政机构身份。通过先进的算法、大数据分析和机器学习等技术，人工智能能够处理和分析大量信息，为社会治理提供深度洞察和决策支持，提高政府服务和决策的效率；通过建立数据模型、分析社会运行数据，人工智能在社会治理中可以用于预测和防范潜在风险，提升治理的前瞻性和科学性；通过智能服务和个性化推荐等技术，人工智能能够提供更加精准和便捷的公共服务，满足人民群众的多样化需求……社会治理在经由人工智能嵌入后展现出了更为简易高效、省时价廉的前景。

2.2. 人工智能嵌入社会治理的算法偏见风险

在技术层面，算法偏见通常源于数据收集和处理过程中的不完善，以及算法设计中的固有缺陷。刘友华对算法的类型做出了概括：算法偏见的类型包括损害公众基本权利的算法偏见、损害竞争性利益的算法偏见、损害个体民事权益的算法偏见等，并强调算法偏见萌芽于数据收集步骤，成熟于模型完善步骤，强化于模型应用阶段([6], p. 58)。Dietterich 则从技术角度提出通过统计算法的偏差与方差可以改善算法的基本设计从而减少算法偏见的出现[7]。伦理层面上，学术界普遍强调在人工智能伦理治理中应对算法偏见问题的重要性。中国信息通信研究院在《人工智能伦理治理研究报告(2023 年)》中提出，人工智能伦理治理是“多主体协作的全流程治理”，“应强化人类责任担当，提倡鼓励多方参与和合作”[8]。林爱珺和刘运红从人与算法的关系以及人与人的关系角度出发，认为通过预先主动承担道德责任的方式可以最大程度地避免算法技术的滥用，有效平衡人与算法的关系以及基于算法的人与人之间的关系[9]。法律层面上，学者们探讨了算法偏见与算法歧视的法律性质和规制路径。李学尧认为将人工智能伦理从道德原则转化为可操作的伦理合规实践，需要探究其法律性质，尤其是法律体系如何评价以及纳入人工智能伦理规范[10]。

综上所述，国内外学者对算法偏见问题进行了多维度的分析和研究，提出了一系列具有建设性的意见和治理策略，旨在促进人工智能技术的健康发展，保障公民权利，维护社会公平正义。

3. 算法偏见的类型与成因

人工智能诞生的本意是作为提高工作效率、降低管理成本的技术工具，本应当公平公正和准确无偏；但为何在嵌入社会治理的过程中失去了其技术理性，陷入事实判断和价值判断，从而导致算法偏见的出现？要想对这一现象进行探析，就必须回到人工智能的运作机制进行深入分析。人工智能运作机理可以

简化为三个环节：输入环节→运作环节→输出环节，算法偏见就在此过程中产生([6], p. 60)。

3.1. 算法模型设计中融入的个人主观偏见

当下以 DeepSeek 和 ChatGPT 为代表的人工智能技术的平台都是基于深度学习的大规模自然语言生成模型，科学合理的算法模型设计是人工智能得以建立与运行的前提和基础。而正是在算法模型这一设计过程中，就可能融入设计者的个人主观偏见。

个人主观偏见可以分为无意识的偏见与刻意选择的偏见。每个不同的算法模型设计者都有着其自身独特的政治立场、意识形态、文化背景和学术研究背景，因此这些偏见可能在设计者的编码和模型构建过程中无意识地产生。首先，如果设计者在选择训练数据集时存在无意识的倾向性，就会导致算法模型因为训练数据集数据的缺失对某些群体或情境产生偏见。其次，当算法模型的假设基于设计者先入为主并理所当然的主观观念时，这便可能影响模型的预测和决策结果。再次，设计者的社会文化环境也可能影响他们对算法功能和目标的设定，从而在算法中嵌入特定文化或社会价值观的偏见。最后，如果算法模型的设计团队缺乏必要的多样性，团队成员共享相似的背景和观点，便可能在无意识中共同强化特定的偏见。

相对于个人主观偏见的无意识融入，基于刻意选择的偏见具有更恶劣的歧视与更严重的影响。算法模型设计者虽然能够意识到自身对某些群体存在认知偏见，却无法在设计过程中秉持足够的客观与公正原则，不仅未能有效避免偏见的引入，反而有意识地将个人主观偏见嵌入算法模型之中。这一现象主要是由于设计人员缺乏足够的道德伦理水平，并且未能建立合理的规制措施与制度对设计者进行足够的约束。在更深层次上则能归因于社会整体对于科技伦理的宣传教育不到位以及相关法律法规的不健全，难以从道德与法律上约束相关从业人员，因而由个人主观偏见引发算法偏见。

3.2. 算法训练数据里内刻的隐性社会偏见

在人工智能的算法模型建立起来后，就要选取训练数据对模型进行训练，训练数据中内刻的社会隐性偏见是导致算法偏见的一个关键成因。训练数据通常反映现实世界，其中可能包含着人类社会长期形成的各类偏见与刻板印象。因此，对训练数据加以筛选，是人工智能设计者必须完成的重要任务。在这一过程中，显性的社会偏见虽可被识别并剔除，从而避免对模型造成直接“污染”，然而隐性社会偏见往往难以察觉、不易消除，从而导致最终训练出的人工智能仍然可能带有潜移默化的歧视倾向。隐性社会偏见进入算法训练数据主要有两个原因。一方面在数据收集阶段，如果样本数据的选择存在偏差，某些群体被过度代表而其他群体被忽视，那么人工智能的算法就可能学习并放大这些群体特征的偏差。而数据收集的方法则容易受到设计者的主观性影响，导致在问题的设计、调查的选择和数据的筛选等环节中，将社会隐性偏见嵌入到数据中。

另一方面，数据的代表性不足也会导致隐性社会偏见。如果一个群体因为各种原因未能充分参与到数据生成过程中，在算法训练时无法全面学习和理解这些群体的特征和需求，便可能导致人工智能对这些群体产生算法偏见。一个主要基于互联网用户的行为数据进行训练的人工智能模型，就有可能无法准确反映那些不经常使用互联网或者无法使用互联网的群体的特征。

3.3. 算法实际使用后产生的机器学习偏见

在接受训练并能输出符合要求的结果后，人工智能在此过程中会试图通过不断地进行学习以改进自身，这便是机器学习。机器学习是一项使计算机系统可以通过经验进行自动改进的技术([5], p. 90)。这一过程主要分为人工智能的自主优化和设计者的主动选择两部分。

人工智能算法模型的自主优化主要体现为其在训练和应用过程中对数据的依赖和处理方式，以及算法自身的设计和优化策略。如果在算法模型设计时和算法数据训练时存在偏见或者歧视，那么人工智能在数据预处理、特征选择和模型训练的每一步都可能加剧偏见。

设计者的主动选择则体现为对人工智能算法模型的主动优化与纠偏。如果设计者过分重视与预期结果相关的特征而忽略其他可能同样重要的特征，这种选择性偏差会导致算法对某些特征过度敏感而忽视其他特征。如果设计者在优化模型时只关注了某些特定的性能指标，那么模型的公平性和适用性就难以得到保证。除此之外，人工智能的输出结果可能会作为输入数据再次进入系统，形成反馈循环。如果未能在人工智能实际使用后的机器学习过程中对这一问题进行纠正，那么就可能导致偏见融入进人工智能的数据审查与结果评估过程中，形成根深蒂固的偏见倾向。

4. 算法偏见的潜在风险与危害

人工智能在嵌入社会治理方面表现出了显著潜力，但与之相对，其在应用过程中也普遍暴露出算法偏见乃至算法歧视等问题。这一系列偏离技术理性的现象，如算法偏见、算法歧视等不仅造成了对社会公平正义的挑战，也可能侵害公民隐私与合法权益，并引发一系列法律与伦理争议。更为深远的影响是，若此类偏见引发重大社会影响，将进一步造成对政府公信力的损害，偏见将逐步侵蚀公众对算法驱动服务及相关机构的信任，造成对政府形象的破坏，最终对社会整体的数字化转型进程形成阻碍。

4.1. 可能侵犯大众的隐私与权利

侵犯大众的隐私与权利是算法偏见带来的直接后果之一。在数据驱动的算法决策过程中，个人数据的收集、处理和分析可能会无意中侵犯用户的隐私权。

在社会治理层面中，人工智能具有赋权与约束的非对称性，“拥有公共权力和公共资源的国家政府在掌握和运用人工智能技术时拥有巨大优势，可以在国家控制体系中利用技术优势来收集和存储公民信息，对公民的行为进行监管”[11]。当权力无法关入制度的笼子时，人工智能技术作为政府公权力的现实显现便落入不受规制的境地，政府对特定群体或特定个人的偏见便一定会体现在算法中，侵犯大众的隐私与权利。

倘若人工智能的使用者出于私利，在缺乏监督与制约的情况下，有意将其个人主观偏见嵌入人工智能系统，便可能使算法沦为偏见的传导工具，进而侵害特定群体的正当权利与利益。此外，若政府在数据管理中存在疏漏，未能充分保障个人信息安全，便可能导致数据库遭受污染——原本中立的个人数据被添加唯有算法能够识别的隐性标记，从而诱发系统性偏见，导致对公众隐私与权利侵犯。

4.2. 可能引发法律与伦理的争议

当人工智能的决策带有偏见色彩，不仅可能侵犯大众的隐私与权利，也会引发民众对公平性和正义性的质疑。如果人工智能的算法决策过程不透明或难以理解，用户和监管机构难以评估其公正性，也难以追究责任。这种“黑箱”特性不仅增加了法律监管的难度，也对社会伦理构成了挑战。

在法律层面，算法偏见违背了法律的公平性原则，引发合规性争议。实现全面推进依法治国总目标的基本原则之一就是坚持法律面前人人平等，而算法偏见造成的不平等就是对这一原则的违背。除此之外，为了确保决策的公正性和伦理性，人工智能的决策过程需要是透明的和可解释的。算法偏见显然对其透明性与可解释性造成了致命性的危机，也导致可追责性问题。如果人工智能的“黑箱”未能破除，那么在进行追责时便可能引发对责任归属问题的争议，倒逼着相关法律法规的补充与完善。

在伦理层面，算法偏见是人类社会的道德标准与价值取向的问题，它会让人们重新审视技术发展与人类基本权利的关系，如何才能实现尊重个体权利与增进社会整体福祉的平衡。科学技术是第一生产力，

科技进步可以推动社会发展，但是并不是所有的技术进步都是积极的。治理现代化在实施过程中，保障和改善民生是一项重大任务，科技发展应该为这个目标服务。然而，算法偏见又会加剧社会不平等，固化社会的分层与歧视，与公平、平等的价值取向相违背。同时，“数字鸿沟”的客观存在也会让残障人士、老年人等人群在数字化进程中失语，更无法融入数字社会。他们不应该被忽视或排斥于社会发展的洪流之外。其实，这种系统性的边缘化本身就是算法偏见的另一种表现形式，也是技术普惠性的现实缺失可能损害政府形象与公信力。

如果政府没有对算法偏见进行有效的规制和治理，那么算法偏见便可能导致决策不公，在就业、教育、司法等领域的应用可能会因为算法的设计或数据集的偏差而导致对特定人群的不公平待遇。这种不公平性一旦造成公共影响，被公众意识到，便会引发对政府决策透明度和公正性的质疑，从而损害政府的形象与公信力。人工智能系统在处理大量个人数据时，一旦发生数据泄露，不仅是对个人隐私的侵犯，也会让公众对政府的数据保护能力产生怀疑。在此基础上，技术失误和责任归属的不明确，也会加剧公众对政府的科技治理能力感到不安。

5. 算法偏见的规制路径

2019 年 6 月 17 日，国家新一代人工智能治理专业委员会发布《新一代人工智能治理原则——发展负责任的人工智能》，强调要“促进新一代人工智能健康发展，更好协调发展与治理的关系，确保人工智能安全可靠可控，推动经济、社会及生态可持续发展”[12]。在面对人工智能嵌入社会治理所带来的算法偏见风险时，对其进行相应规制既是政府达成“善治”的必然转向，也是政府回应公民期望、维护民众权利与利益的必须之举。对算法偏见的规制归于本质，是要保持算法的透明性、可解释性、公正性、可问责性与可升级性。

5.1. 技术管控：算法数据透明化

技术管控作为规制算法偏见的关键环节之一，算法数据透明化则是技术管控的核心措施。算法透明化作为对算法“黑箱”的打破与超越，要求算法的设计者和使用者向利益相关方公开算法的基本原理、优化目标、决策标准等关键信息，以消除社会疑虑并推动算法的健康发展。这当然不是全部的公开透明，而是“依赖于行为规范和私权保护双重范式的协同作用，进行事前预防、事中约束和事后救济等全过程控制，并结合对主体、客体、程度、条件等要素的场景化思考，实现智能科技创新、整体经济效益和社会公共利益之间的平衡与协调”[13]。

相对于算法方的主动管控，有序推广算法备案工作、建设算法透明标准体系也是实现算法数据透明化、规制算法偏见的重要途径。通过有序开展算法备案，可以持续扩大算法备案覆盖的行业和领域，加强算法管理。这不仅有助于监管机构了解和监督算法应用情况，也为算法使用者提供了一个展示其算法透明度的机会与场景。而制定和完善算法透明领域的标准，如各行业各场景的算法透明标准、算法影响评估、算法透明共性技术等，则是实现算法数据透明化的重要支撑。在此基础上，建立健全的算法透明标准体系，进一步明确算法公开透明要求，将引导算法透明治理落地实施，最终实现科技伦理治理水平的高质量高水平发展。但是，数据透明化在实践中面临多重挑战与约束。首先，算法的核心技术往往构成企业知识产权与商业秘密的核心组成部分，完全的透明化要求可能引发知识产权保护与算法治理之间的紧张关系。对此，可借鉴分级分类的透明原则：对涉及公民基本权利与重大公共利益的算法，要求较高程度的透明度，包括公开算法设计逻辑与影响评估报告；对仅涉及一般性公共服务的算法，则可要求相对有限的透明度，如仅公开算法的优化目标与决策标准，而不必公开全部技术细节。其次，公共安全领域的算法透明度面临着特殊约束：过度透明可能导致恶意主体利用公开信息进行规避，从而削弱治理

工具的效能。因此，需要建立与国家安全部门的协作机制，在保障公共安全底线的前提下实现有限但有效的透明治理。

5.2. 制度规制：算法程序公正化

算法偏见作为一种科技伦理问题，昭示出科技发展与法律规制之间存在的不匹配性，是对完善的制度建构的迫切现实需求。算法权力作为政府公权力的一部分，对社会治理中出现算法偏见进行规制同样需要构建一个全面完善的制度框架，以制度化方式约束算法权力。一个完善的制度需要制定明确的法规标准来指导算法的设计和应用，确保算法的公正性和透明度。但与常规的法律治理逻辑不完全相同，科技伦理治理模式特征体现为适应性治理与软法规制，即坚持“审慎使用实体权利——义务规范方式规定伦理问题，推动法律通过规范程序方式促进伦理反思；建立法律规则、行政指引与自律规范相结合的多层次规范体系，完善风险交流机制与伦理共识达成机制，增强伦理原则执行力”[14]这两条原则，可以通过适度法治化路径提升伦理治理效率。

建立健全完善的算法偏见规制制度需要建立完善的审查机制与持续的监控机制，对算法的整体运行过程进行监督与定期评估，并根据反馈进行必要的调整和更新，以实现人工智能的公正化，防止算法偏见的产生。除此之外，也要建立起配套人才培训体系与培养制度，对人工智能的开发者和使用者进行必要的伦理教育和技能培训，不仅要提升他们对算法偏见的认识和处理能力，更要培养他们的职业道德感和责任心。上述制度性建设在实际推进中亦面临显著挑战。一方面，建立完善的算法审查机制需要大量具备人工智能、法律与公共管理交叉知识背景的专业审查人员，而当前此类复合型人才严重短缺，可能制约审查机制的实质化运行。另一方面，持续的算法监控机制在技术层面面临动态性问题：算法系统——尤其是基于在线学习的系统会在使用过程中不断自我更新，参数与行为随之变化，这意味着一次性的审查结论可能在较短时间内即告失效，监管机构需要建立具备持续跟踪能力的动态监控制度，而非依赖静态的节点式审查。

5.3. 责任约束：算法结果可问责

算法结果可问责实质上就是对算法权力的强制性责任约束。相较于强调在人工智能运行过程中进行管理控制的制度规制，一套完善的责任机制强调的是在算法偏见导致算法决策产生不良影响时，能够明确责任主体并追究其责任，并以此约束算法设计者与算法使用者的行为。可问责的算法是以可解释性为基础的，具体而言则是“组织使用算法的正当理由以及向提出问题、做出判断和施加影响的问责机制解释其结果”[15]。

算法的可问责主要针对算法权力的主体，即算法设计者与算法使用者。对于算法设计者的责任约束机制一般通过算法备案制度，将其作为事前固定责任的重要手段，“以信息收集为核心、以柔性治理为特点，有助于形成便利多元主体参与的协同性治理、符合科技发展规律的动态性治理、多重途径有机结合的复合型治理模式”[16]。算法备案制度通过要求算法设计者或算法推荐服务提供者向监管部门备案其算法的相关信息，可以在算法运行之前就明确责任主体和责任范围。这不仅有助于监管部门进行风险评估和监督，也为事后的责任追究与优化升级提供了依据。

对于算法使用者而言，确保算法结果可问责的重要措施是算法影响评估程序与独立的算法解释环节。算法影响评估程序是对算法决策可能带来的社会影响和风险进行评估，以评估影响与评估结果去约束算法使用者。通过算法影响评估程序可以强化算法使用者的义务与责任，预防和减少因算法使用者不当使用而产生的算法偏见和其他不良后果。独立的算法解释环节是指算法使用者应算法服务接受者要求依法予以说明并承担相应责任。当算法决策实施后并产生重大影响时，独立的算法解释环节是平衡算法使用

者与接受者之间不平等地位的重要途径。在这一环节中，不仅要保证解释的公正性与公开性，还要做到对算法服务接收者而言的可理解性。

参考文献

[1] 关于加强科技伦理治理的意见[N]. 人民日报, 2022-03-21(001).

[2] McCarthy, J., Minsky, M., Rochester, N., et al. (2006) A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August, 31, 1955. *AI Magazine*, **27**, 12-14.

[3] Sheikh, H., Prins, C. and Schrijvers, E. (2023) Artificial Intelligence: Definition and Background. In: Sheikh, H., Prins, C. and Schrijvers, E., Eds., *Mission AI*, Springer, 15-41. https://doi.org/10.1007/978-3-031-21448-6_2

[4] High-Level Expert Group on Artificial Intelligence (2019) A Definition of AI: Main Capabilities and Scientific Disciplines. European Commission. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=56341

[5] 王小芳, 王磊. “技术利维坦”: 人工智能嵌入社会治理的潜在风险与政府应对[J]. 电子政务, 2019(5): 86-93.

[6] 刘友华. 算法偏见及其规制路径研究[J]. 法学杂志, 2019, 40(6): 55-66.

[7] Dietterich, T.G. and Kong, E.B. (1995) Machine Learning Bias, Statistical Bias, and Statistical Variance of Decision Tree Al-Gorithms. https://web.engr.oregonstate.edu/~tgd/publications/tr-bias.pdf

[8] 中国信息通信研究院. 人工智能伦理治理研究报告(2023 年) [R/OL]. https://www.caict.ac.cn/kxyj/qwfb/ztbg/202312/t20231226_468983.htm, 2026-07-26.

[9] 林爱珺, 刘运红. 智能新闻信息分发中的算法偏见与伦理规制[J]. 新闻大学, 2020(1): 29-39, 125-126.

[10] 李学尧. 人工智能伦理的法律性质[J]. 中外法学, 2024, 36(4): 884-898.

[11] Jordan, M.I. and Mitchell, T.M. (2015) Machine Learning: Trends, Perspectives, and Prospects. *Science*, **349**, 255-260. https://doi.org/10.1126/science.aaa8415

[12] 发展负责任的人工智能: 新一代人工智能治理原则发布[EB/OL]. https://www.most.gov.cn/kjbgz/201906/t20190617_147107.html, 2019-06-17.

[13] 季冬梅. 人工智能透明性原则的制度构建: 范式选择与要素分析[J]. 科学学研究, 2022, 40(4): 611-618, 757.

[14] 谢尧雯, 赵鹏. 科技伦理治理机制及适度法制化发展[J]. 科技进步与对策, 2021, 38(16): 109-116.

[15] 马克·舒伦伯格, 里克·彼得斯. 算法社会[M]. 王延川, 粟鹏飞, 译. 北京: 商务印书馆, 2023: 91.

[16] 张吉豫. 论算法备案制度[J]. 东方法学, 2023(2): 86-98.