算法偏误的公共治理：政府、市场与社会协作互动之道

顾　昕　郭凤林

摘　要　人工智能所驱动的算法决策受制于算法自身逻辑、人类有限理性以及用户反馈循环等多重因素的叠加影响，产生了对特定社会群体的代表性不足、识别错误率较高、政策适用偏差等问题，引发民众对算法偏误问题的质疑和担忧。对此，政府、市场和社会主体均付出了一定的努力加以应对，在技术逻辑、伦理原则及应用限制等方面进行了一定的探索，但其广度、深度和可持续性都比较有限。针对算法偏误的公共治理体系建设，需要推动政府、市场和社会多方主体的协作互动，推进市场机制、社群机制与行政机制的互补嵌合，推动政府从单纯依赖行政机制的施为转向元治理职能的行使，构建多方主体共建共治共享为核心的公共治理体系，推进算法偏误的公共治理制度化，才能构建新质生产关系，实现科技变革与社会秩序的共生发展。

关键词　算法偏误　公共治理　社群机制　市场机制　元治理

作者顾昕，浙江大学公共管理学院教授、民生保障与公共治理研究中心高级研究员（浙江杭州 310058）；郭凤林，北京外国语大学国际关系学院副教授（北京 100089）。

中图分类号 D0　　**文献标识码** A　　**文章编号** 0439-8041(2025)10-0014-11

一、问题提出：算法偏误的治理挑战

随着人工智能（AI）的迅猛发展，算法正加速嵌入现代社会的多元结构和程序之中，深刻改变着现代商业活动和公共服务的运作方式。算法不再仅仅是一个自动化决策程序，而是正在塑造形成新的社会治理模式。[①] 然而，算法所造成的社会歧视问题也开始显现，引发关于“算法偏见”或“算法偏误”的质疑。[②] 随着AI在社会经济领域的加速应用，政府、业界和学界都开始关注到算法偏误问题，并通过多种努力来理解、诊断和减轻这些偏误。但现有方案要么侧重于督促国家出台更全面、更细致的监管[③]，要么吁求政府为科技公司、社交媒体、电商平台等设定更明确的“企业社会责任”[④]。这些解决方案秉持行政化治理的传统思维，希望政府运用行政治理机制、通过自上而下的干预来解决科技与社会共进中的矛盾。可是，由于算法技术与国家治理的融合正处于实验期，有关算法偏误治理尚未形成稳定的知识和共识，AI发展本身也存在着创新过程内生的信息不确定性和多方主体间的信息不对称性，政府一方面尚不能简单制定出清晰有效的干预或管制方

① 马克·舒伦伯格、里克·彼得斯编：《算法社会：技术、权力和知识》，王延川、栗鹏飞译，北京：商务印书馆，2023年。

② 赵广立：《算法偏见是无心之失，还是别有用心?》，《中国科学报》2024年12月28日。

③ 陈兵、董思琰：《常态化监管与算法分类分级治理模式更新》，《学术论坛》2024年第3期。

④ Ulrike Reisach, “The Responsibility of Social Media in Times of Societal and Political Manipulation,” *European Journal of Operational Research*, vol. 291, issue 3, 2021, pp. 906-917.

案，另一方面也会担忧贸然实施的管制会阻碍 AI 发展而踟蹰不前。[①]

完善科技进步的公共治理以增进民众对科技的理解，争取民众对科技发展的支持，是促动科学与社会和谐共进、推动新质生产力发展、推进新质生产关系形成和实现中国式现代化的重要支撑。由此，在当前算法偏误问题引发较为广泛的公众担忧之际，如何建构政府、市场、社会多方主体协作互动的新公共治理格局，让行政机制、市场机制和社群机制互补嵌合，有效解决算法偏误问题，以实现科技进步和社会发展的和谐共进，是本文探究的论题。

二、治理模式的演进：从新公共管理到社会治理共同体

（一）倚赖政府的行政化治理模式

在公共事务的治理中，政府是举足轻重的主体，也通常是企业与民众倚重的主体。政府往往具有较强的行动力和较充分的资源，成为民众首要依赖对象不足为奇。在中国，民众无论是在感情上还是在责任归因上，都倾向于信任政府和依赖政府。[②] 这使得中国公共治理体系的行政化倾向更为明显。

公共治理行政化有诸多优点，但其弊端也普遍且常见。其一，政府难以获得被治理者（包括民众和企业）的信息，因而信息不充分和信息不对称伴随着公共服务政策制定、执行和评估的全过程；其二，公民或民众作为政府的委托人，也无法获取政府及其官员的完备信息，这不仅影响其对政府的信任，而且也常常不利于其对政府诉求的合理化；其三，在涉及具有异质性偏好的社会群体时，公共政策会遭遇最优决策不可能的难题，即个体偏好无法加总为集体偏好的社会选择困境；其四，政府对社会经济活动的干预难免会给某些社会经济群体带来额外的好处，由此会激发寻租活动，进而扭曲经济、社会和政治秩序。因此，行政化治理失灵现象并不鲜见，并引发治理变革的强大政治、经济和社会压力。

在公共治理实践及学术研究的不断反思和推动下，公共治理变革呈现出去行政化的趋势，即打破单一行政力量和行政机制主导公共部门运行的既有格局，让政府、市场和社会多方主体协作互动，让行政、市场和社群机制形成互补嵌合的格局。[③] 针对诸多公共事务尤其是社会事务，通过建立社会治理共同体，多方主体协作互动以实现共建共治共享，已成为中国共产党十九届四中全会决议中提出的社会治理治国理念的核心。[④]

（二）引入竞争的市场化治理模式

公共治理本身，无论是作为一门学问还是作为实践，就是以超越行政化的公共行政范式为目标而发展起来的，而公共治理的理论和实践则经历了多次范式转型，其中非常有影响力的一次转型即是“新公共管理运动”（New Public Management，NPM）的兴起，其特征是将激励和竞争引入公共管理，让市场机制与行政机制并行不悖，通过政府与市场形成合作伙伴关系来提高政府效率，改善公共服务的质量。

新公共管理运动起源于英国[⑤]，后遍及世界各地。[⑥] 在英国，撒切尔政府在 1980 年代实施了广泛的公共部门改革，通过引入竞争机制、绩效评估和外包等手段，提高了公共服务的效率，这一改革被格兰德（Juan Le Grand）称为“另一只看不见的手”[⑦]。然而，新公共管理的实践也出现一定的弊端。首先，过于强调效率可能导致对公共服务的公平性和普遍性的忽视，容易造成社会不平等的加剧；而且，市场化改革还可能导致公

---

① Xukang Wang and Yingcheng Wu, “Balancing Innovation and Regulation in the Age of Generative Artificial Intelligence,” *Journal of Information Policy*, vol. 14, 2024, pp. 385–416.

② 余泓波、吴心喆：《民众对政府治理的依赖如何塑造其政府信任》，《社会科学战线》2018 年第 9 期。

③ 顾昕：《治理机制的互补嵌合性：公共部门制度创新与激励重构》，上海：格致出版社，2022 年。

④ 顾昕：《作为一种治国理念的社会治理——学术传承与理论建构》，《国家现代化建设研究》2024 年第 1 期。

⑤ Evan Ferlie, Lynn Ashburner, Louise Fitzgerald and Andrew Pettigrew, *The New Public Management in Action*, Oxford: Oxford University Press, 1996.

⑥ Kathleen McLaughlin, Ewan Ferlie, and Stephen Osborne (eds.), *New Public Management: Current Trends and Future Prospects*, London and New York: Routledge, 2022.

⑦ Julian Le Grand, *The Other Invisible Hand: Delivering Public Services through Choice and Competition*, Princeton: Princeton University Press, 2007.

共部门的专业性和服务质量下降，因为在追求成本效益的过程中，服务提供商可能会牺牲服务的深度和人性化。[①] 其次，新公共管理在实施过程中常常伴随着复杂的绩效评估体系建设，这会增加管理的复杂性和不确定性，影响了公共服务的稳定性和可持续性。[②] 最后，它还可能导致公共价值的削弱，尤其是在教育和医疗领域，加重民众的经济负担和社会不安。[③]

（三）激活社会的社群治理模式

“社群机制”（community mechanism）是与行政机制、市场机制并列的第三种治理机制，其运作是基于社群成员对共享价值观和规范的认同完成社群乃至公共事务的治理。政治学家和公共管理学者奥斯特罗姆正是由于对公共事务社群治理机制的杰出研究于2009年获得了诺贝尔经济学奖。在她领衔的布鲁明顿学派那里，社群治理的适用领域不仅限于渔场、森林和环境等公共资源的治理[④]，而是遍及社会经济生活。[⑤] 社群机制的运作既可以出现在各类正式民间组织及其组成的非营利部门之中，也可以出现在包括家族、联盟、社会关系在内的非正式社会网络之中，社群的具体组织形式具有多样性。

与行政治理和市场治理有所不同，社群治理的特点在于当事人均为相识者，无论是在公司、非营利组织、社区、商会、专业社团、体育俱乐部甚或帮会，社群成员均是“一个在多方面直接并频繁交往的人群”，“成员之间保持关联（而不是感情）是社群最显著的特征”，“数量不多的社群成员存在着重复且多方面的社会交往”[⑥]。其社会经济身份自然有别，但相互关联，密切互动，对各自的权益和诉求予以积极的回应，形成某种程度的互助关系，从而体现出“社群的精神”[⑦]。这一点无论对于非正式社群，如群体、联盟、网络，还是对正式的社群，如社团性的协会组织和法人性的非营利性组织，都是适用的。

社群机制如何发挥治理上的效用？鲍尔斯（Samuel Bowles）把社群机制的基本特征概括为“认诺与遵从”（commitment and compliance），即相互密切关联的个体基于对某些共同价值与规范的认同、承诺与遵守以协调其活动，而在认同与承诺中蕴含着高水平的人际信任，即学界所称的“社会资本”[⑧]。在很大程度上，基于社会资本的社群治理可以克服信息不完备所导致的市场和行政失灵问题。在高频重复博弈、多维互动、关系密切的情况下，例如稳定的邻里、裁员很少的企业、会员制组织以及非营利组织甚至宗派中，以强调声誉、重视互惠、遵守规范为特征的社群治理机制，可成为市场和行政机制的有效替代，以促使成员通过合作以推进相关事务的开展。[⑨]

但正如行政失灵和市场失灵的存在一样，社群失灵也是存在的。社群机制作用的局限性有如下体现：一是规模性，即社群机制在小型社群中容易有效发挥作用，而在社群规模变大后，社会资本就会变得稀薄，免费搭车者问题就会凸显；二是异质性，即成员异质性过高如种族认同或财富差异较大，成员间的利他性和互惠性就有可能减弱，自发的合作或集体行动难以形成或难以持久；三是依从性，即社群难以对成员实行强制性约束，制裁和惩罚在某些社群成员那里难以产生足够的羞愧效应。[⑩]

（四）走向三方主体协同治理的公共治理之道

如上分析可见，在日益复杂的公共事务治理中，单由国家承担治理职责存在着能力和效果上的双重不足，而市场和社群主体是治理中的重要参与者。尤其是第二次世界大战后随着科技驱动的创新加速发展以来，创

---

① Christopher Pollitt and Geert Bouckaert, *Public Management Reform: A Comparative Analysis*, New York: Oxford University Press, 2011.

② 唐纳德·莫伊尼汗：《政府绩效管理：创建政府改革的持续动力机制》，尚虎平、杨娟、孟陶译，北京：中国人民大学出版社，2020年。

③ Jamie Peck and Adam Tickell, “Neoliberalizing Space,” *Antipode*, vol. 34, issue 3, 2002, pp. 380–404.

④ Clark Gibson, Margaret Mckean and Elinor Ostrom, *People and Forests: Communities, Institutions, and Governance*, Cambridge, MA.: the MIT Press, 2000.

⑤ Daniel Cole and Michael McGinnis, *Elinor Ostrom and the Bloomington School of Political Economy: A Framework for Policy Analysis*, Lanham, MD.: Lexington Books, 2017.

⑥⑨⑩ Samuel Bowles, *Microeconomics: Behavior, Institutions, and Evolution*, Princeton: Princeton University Press, 2004, p. 474, pp. 258–260, p. 164.

⑦ Amitai Etzioni, *Spirit of Community: the Reinvention of American Society*, New York: Simon & Schuster, 1994.

⑧ Samuel Bowles and Herbert Gintis, “Social Capital, Moral Sentiments, and Community Governance,” *The Economic Journal*, vol. 112, issue. 483, 2002, pp. 419–436.

新的技术复杂度不断提升，而对创新的治理就越来越多地依赖于多主体的协同推进，其中国家作为组织者和协调者介入重大技术创新活动是技术演化的必然结果。[①] 由此，国家、市场和社会行动者通过制度化的网络或共同体建设，形成共同的目标、凝聚共享的价值观、建构共同遵守的行为规范和制度，通过密切的协商互动，成为公共治理体系建设的要务。

在这一体系中，行政、市场、社群机制构成互补嵌合关系，而非相互替代、此消彼长的关系。三种机制的发挥并不完全对应于三大主体的作用，而是相互交织。比如，多方行动者通过社群机制的作用所形成共同遵守的规范和制度，可以有效降低交易成本，这就有利于市场机制发挥作用；而社群机制所主导的社群治理也需要嵌入在市场机制之中，通过威廉姆森式的“关系型契约”[②] 为多方行动者之间制度化合作关系的形成提供激励；作为市场主体的营利性组织（企业）和社会主体的非营利组织，其运转也离不开权威机制的行使，尤其是规模越大的组织，其运转过程中的权威手段的越用就越多；最后，也是最重要的，无论是市场机制还是社群机制，其有效运作必须嵌入在行政机制的积极作用之中。

值得注意的是，去行政化并不意味着取消行政力量的作为和行政机制的作用，而是要求政府在职能转变中改变并完善运用行政机制的方式，从单纯的命令与控制走向元治理。元治理意指“治理的治理”（governance of governance），这是政府或行政力量的一种新职能，也是行政机制运转的一种新形式，其宗旨是让协作—互动治理有效持续地运转起来。[③] 由于行政机制、市场机制、社群机制的运作并不天然兼容，为了使不同治理机制在治理中各展所长并互补协同，就需要元治理者对不同治理主体的行动和不同治理机制的运作加以协调。行政力量可以以其特有优势承担元治理者的角色，使协作—互动治理更加制度化，更好地实现治理的公共价值。[④]

由此可见，走向政府、市场与社群的协同，实现行政机制、市场机制与社群机制的互补嵌合，是提升公共事务治理效率和效能的根本路径。那么，在对新型技术代表 AI 所引发的新型社会治理问题——算法偏误的治理上，这三种主体是如何承担治理责任、发挥治理效果的？存在着何种缺陷和问题、需要如何补足？下文拟在上述理论的基础上分析算法偏误的治理问题，并增进对公共事务治理之道的认知。

三、算法偏误的形成原因

科学研究的核心是试图建立数学模型，来解释和预测变量间的关系。但由于从样本到总体的估计链条中存在多种偏误可能，导致真实值与估计值之间会存在一个差值 ε。这个差值 ε 可以分为两种类型：随机误差和系统误差或“统计偏误”[⑤]。就此数学逻辑而言，“算法偏误”是个伪命题，算法只是科学的计算工具，而并不具备拥有或实现某些价值观的能力，并不能说工具是“有偏误的”[⑥]。不过从社会应用场景来看，算法决策产生了诸多有偏误的结果，如性别歧视、种族歧视、年龄歧视、收入歧视、地区歧视等，成为社会不得不正视和需要解决的问题。就其从“工具”到“价值”的逻辑机制主要归入技术逻辑、人类理性和应用循环三个层面。

（一）算法偏误的技术逻辑

算法虽然是“客观”的科学工具，但其决策高度依赖于数据资料和机器学习模式，输出的结果反映的是历史数据的客观特征，这使得其在应用于当下有关人类社会的决策上容易产生如下三个偏差。

---

① 张海丰、司叶林：《熊彼特需要找回“国家”：技术革命浪潮视域下新质生产力发展的动力机制探讨》，《学术月刊》2024 年第 11 期。

② Oliver Williamson, *The Mechanisms of Governance*, New York: Oxford University Press, 1996.

③ Jacob Torfing, Guy Peters, Jon Pierre, and Eva Sørensen, *Interactive Governance: Advancing the Paradigm*, New York: Oxford University Press, 2012, p. 4.

④ Eva Sørensen, “Metagovernance: The Changing Role of Politicians in Processes of Democratic Governance,” *American Review of Public Administration*, vol. 36, no. 1, 2006, pp. 98–114.

⑤ John Bound, Charles Brown and Nancy Mathiowetz, “Measurement Error in Survey Data,” in *Handbook of Econometrics* (Vol. 5), James Heckman, Edward Leamer (eds.), Amsterdam: North Holland, 2001, pp. 3705–3843.

⑥ Sina Fazelpour and David Danks, “Algorithmic Bias: Senses, Sources, Solutions,” *Philosophy Compass*, vol. 16, issue. 8, 2021, e12760.

一是机器学习的逻辑偏差。传统数据分析方法通常是基于规则的，对数据进行诸如年龄、性别、收入、消费习惯等方面的编码特征并根据决策规则做出判断，是结构化、可表达的具体算法；而机器学习所使用的算法如神经网络等，通常依赖于深度学习，传统需要编码的信息也成了“数据”，是算法要学习和处理的对象。① 例如，传统的人工简历筛选程序把“性别”作为编码和分析维度；而AI加持的自动简历筛选系统则将“性别”作为学习之后的数据结果，由此产生了筛选结果全是男性候选人的情况。②

二是算法所依赖的数据偏差。AI训练所基于的数据集通常采取“全纳入”的模式，许多残缺的、不完整的或者情境各异的信息都被纳入到同一个数据库中，这使得机器学习和训练缺乏对整体代表性的考量。由于数据集规模极其庞大，以人工审核和手动清理方式消除这些误差几乎不可能。例如，美国使用的性犯罪再犯风险评估系统“静态99”（Static-99）时，其数据库中印第安人的数据来自被筛选过的高风险样本，而黑人和西班牙裔的数据则来自相对常规的样本，这就使得族群间得分的显著差异存在偏误。③ 此外，算法决策结果往往反映的是数据丰富地区/群体的情况，对数据匮乏群体的代表性则不足。④

三是算法决策结果的情境适用偏差。机器学习高度依赖于既有数据资料，但在决策时很难考虑到应用情境的差异。一旦技术走出实验室，由于现实中“数字连接关系”复杂多变，在实验室受控的很多条件不复存在，其在社会应用中的中立性就会面临挑战。⑤

（二）人类有限理性叠加算法偏误

算法虽然在很大程度上属于“自然科学”，会被视为具有技术中立性，因而算法反而会提升决策程序的公正性，令民众产生了一定程度的“算法偏好”（alogrithm preference），例如在种族歧视和经济不平等的背景下，受歧视群体（如美国黑人）更支持算法决策的医疗服务。⑥ 但事实上，算法在开发过程中仍然会受到人的有限理性和偏好的影响。

一是算法选择偏误。通常情况下，当算法用于耗时琐碎但常规化的任务时产生的模型，出现偏误的风险较低，如围棋、国际象棋、导航推荐等；但当用来解决复杂问题时，算法会尝试将复杂的、不可预测的问题简化为标准化的、有逻辑的解决方案，利用分数、排名、指标、范式等简化的指标来进行决策，而这就涉及到程序员对问题的认知模式以及简化方向选择等问题，偏误就此容易发生。其次，算法几乎总是被用于指标相关的优化和筛选工作，因此开发者总是选择在某些指标上（如预测准确性、敏感性、特异性）的优异表现来展现其应用效果。尽管在很多情形下算法的具体选择取决于情境、条件和指标效果，但在很多时候，算法设计者仍然有一定的选择空间和出错可能。

二是商业利益刺激下的关联偏差。机器学习在商业上被广泛应用于特征识别，即商家通过对用户语言、行为特征等的分析来识别用户的性别、种族、年龄、收入、政治观点等特征，以此作为个性化推荐系统的基础来进行相应的内容推荐和定向广告投放。⑦ 由于女性用户的点击转化率较高，因而对其投放广告的价格远高于对男性用户的投放成本，因此诸如STEM之类的招聘广告更多推送给了男性用户。⑧ 这些差异化的内容推

---

① Gavan Duffy and Seth Tucker, “Political Science: Artificial Intelligence Applications,” *Social Science Computer Review*, vol. 13, no. 1, 1995, pp. 1-20.

② Carlotta Rigotti and Eduard Fosch-Villaronga, “Fairness, AI & Recruitment,” *Computer Law & Security Review*, vol. 53, no. 1, 2024, 105966.

③ Simran Ahmed, Seung Lee, and Maaike Helmus, “Predictive Accuracy of Static-99R Across Different Racial/Ethnic Groups: A Meta-Analysis,” *Law and Human Behavior*, vol. 47, issue. 1, 2023, pp. 275-291.

④ Alexandra Chouldechova and Aaron Roth, “A Snapshot of the Frontiers of Fairness in Machine Learning,” *Communications of the ACM*, vol. 63, no. 5, 2020, pp. 82-89.

⑤ 刘兴华：《数字全球化时代的技术中立：幻象与现实》，《探索与争鸣》2022年第12期。

⑥ Yochanan Bigman, Kai Chi Yam, Deborah Marciano, et al., “Threat of Racial and Economic Inequality Increases Preference for Algorithm Decision-Making,” *Computers in Human Behavior*, vol. 122, 2021, 106859.

⑦ Koren Yehuda and Robert Bell, “Advances in Collaborative Filtering,” in *Recommender Systems Handbook*, Francesco Ricci, Lior Rokach, Bracha Shapira(eds.), Berlin: Springer, 2022, pp. 91-142.

⑧ Anja Lambrecht, Catherine Tucker, “Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads,” *Management Science*, vol. 65, issue. 7, 2019, pp. 2947-3448.

介，进一步强化了民众对算法歧视问题的感知。

（三）反馈循环自我强化算法偏误

算法所依赖的机器学习方法是一个自我调整的循环，它可以在反馈中不断学习和自我纠正。如果用户与算法的互动有偏误，模型便会强化这个结果并生成更具偏误性的输出，使得算法偏误结果更容易产生自我强化而难以自我纠正。荣获美国数学协会欧拉图书奖的《算法霸权：数学杀伤性武器的威胁与不公》一书中指出，AI 模型可以“依靠自己的内置逻辑来定义其所处理的情况，然后再以自己的定义证明其输出结果的合理性。这种模型会不断地自我巩固，自我发展，极具破坏力——而且在我们的日常生活中很常见”①。

此外，诸如生成式 AI 模型还有可能产生误导性输出。当前已有不少研究发现大语言模型会为用户提供看似真实的捏造资料，如编造的学术论文、法律条文等。这种所谓的“AI 幻觉”（AI hallucinations）使得基于算法偏见而生成的误导性内容，会进一步加剧了用户既有的刻板印象。②

四、对算法偏误的现有回应及不足

斯图尔特·拉塞尔（Stuart Russell）指出：“如果我们不认真对待人工智能的社会影响，最终可能会面临无法控制的后果。”③ 在当前对算法偏误的质疑和批判声中，相关主体已经采取了相关的举措进行应对。这里将这些举措从主体角度，从科技公司（市场主体）、行业协会（社群主体）和政府（行政主体）三个角度进行应对政策梳理分析，以发现其基本思路和不足之处。

（一）AI 科技公司：技术纠偏为主且动力不足

当前，人工智能相关的研发和应用活动主要在企业手中。对企业而言，算法纠偏可以维护其公正的社会形象、提升用户信赖和支持程度，因此，市场机制的运作使得企业有一定的动力来改进算法。当前已有科技企业意识到这一点，在这方面做出了若干努力，但相关行动尚未系统化和制度化。

一是技术层面的算法去偏，即增加对预防算法歧视的技术投入，通过技术创新试图解决算法歧视、算法不透明以及算法极化等问题，主要手段可以划分为三种类型：“预处理”（pre-processing）、“中处理”（in-processing）和“后处理”（post-processing）。④ 预处理技术主要是在算法中排除敏感信息的适用⑤；中处理技术是通过修改算法或为算法添加约束来减轻歧视⑥；后处理技术是修改预训练分类器的结果，实现在不同群体上的统计对等。⑦ 可是，技术方面的举措在抑制歧视方面的作用仍然有限。

二是加强算法伦理自律，提高员工的伦理意识。在这个过程中，谁能率先确定行业规范，谁就更能成为行业主导力量。为此，许多著名科技公司在市场力量的驱动下争相发布自己的算法伦理，有力推动了 AI 相关的伦理共识塑造与行业实践。可是，在商业竞争压力之下，科技公司很容易出现“短视偏差”，即相较于社会责任，眼前经济价值的实现更受到重视。⑧ 受制于企业的逐利属性，当市场情况变化或者竞争激烈程度升级时，AI 道德伦理是企业首要放弃的内容。

因此，依赖市场机制的单独运作，期望市场力量驱动市场主体承揽更多的社会责任，难以有效完成 AI 社

---

① 凯西·奥尼尔：《算法霸权：数学杀伤性武器的威胁与不公》，北京：中信出版社，2018 年，第 9 页。

② Sebatian Farquhar, Jannik Kossen, Lorenz Kuhn, et al., “Detecting Hallucinations in Large Language Models Using Semantic Entropy,” *Nature*, vol. 630, 2024, pp. 625-630.

③ Stuart Russell, *Human Compatible: Artificial Intelligence and the Problem of Control*, New York: Viking, 2019, p. 126.

④ Deven Desai, Swati Gupta, and Jad Salem, “Using Algorithms to Tame Discrimination: A Path to Diversity, Equity, and Inclusion,” *UC Davis Law Review*, vol. 56, no. 1, 2022, pp. 1703-1768.

⑤ Shira Mitchell, Eric Potash, Solon Barocas, et al., “Algorithmic Fairness: Choices, Assumptions, and Definitions,” *Annual Review of Statistics and Its Application*, vol. 8, no. 1, 2021, pp. 141-163.

⑥ Nima Kordzadeh and Maryam Ghasemaghaei, “Algorithmic Bias: Review, Synthesis, and Future Research Directions,” *European Journal of Information Systems*, vol. 31, no. 3, 2022, pp. 399-409.

⑦ Emilio Ferrara, “Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies,” *Science*, vol. 6, no. 1, 2023, pp. 1-15.

⑧ 肖红军、商慧辰：《数字企业社会责任：现状、问题与对策》，《产业经济评论》2022 年第 6 期。

会治理的目标，市场机制的运作必须与社群机制和行政机制互补嵌合，在充分发挥市场机制效率作用的同时降低其“失灵”的负面影响。

（二）社会喧哗：注重伦理原则制定而缺乏沟通和组织

随着质疑算法偏误的民众呼声日高，相关的行业协会和社会组织也行动起来，就AI领域的算法规范发展和企业社会责任做了一定的规范性工作，着力推进人工智能与人类的“伦理对齐”（ethical alignment），让AI的能力和行为与人类的价值、真实意图和伦理原则相一致。①

首先，从科学共同体来看，不少学术性国际组织正在积极推进AI的伦理对齐工作，如全球电子和电气工程师协会（Institute of Electrical and Electronics Engineers，简称IEEE）、计算机协会（Association for Computing Machinery，简称ACM）、欧洲科学与新技术伦理组织（European Organization on Ethics in Science and New Technologies，简称EGE）。科学技术界对AI发展是否合乎道德的关注是不无影响的，这在国际组织的行动中有所反映。联合国教科文组织于2021年11月23日通过了《人工智能伦理问题建议书》并在2022年以多种语言出版；该建议书关注AI在文化、教育、科学、信息和传播等方面的伦理影响，提出了指导政策制定者和利益相关者确保AI合乎道德的十条建议。②

其次，不少企业组成行业性社会组织积极推动行业共识和规范的形成和约束力的发挥。但总体来说，行业协会在AI伦理治理所发挥的作用还比较有限。这正如一篇相关的文献综述所指出的，负责任的AI原则已经对很多科技公司的组织价值观产生了影响，但企业间横向协调以及如何拓展为利益相关者生态体系，依然没有受到应有的关注。③ 换言之，如何让社群机制发挥作用，让科技公司组成的协会或联盟在AI伦理治理上扮演重要角色，依然是一个全球性的课题。

此外，公民参与是当前公共治理中的重要潮流，体现了社群机制的重要性。当前已有不少民间社会组织通过举办研讨会、公众听证会和在线论坛等方式，鼓励公民参与探讨AI的开发与应用，以确保技术的透明性和社会责任。然而，对16个国家的AI发展战略的文本内容分析发现，虽然这些国家战略中都提及了公众角色和公民参与机制，但对公民参与AI治理的具体政策、机制和活动的论述都较为空洞。④

如上可见，当前AI相关的专业组织、行业协会和公民团体等社会力量已经参与到算法偏误的治理中来，但究其力度和形式而言，还存在着很多不足。首先，行业协会的工作形式单一，主要是制定各项伦理原则，其他约束和组织手段较为缺乏；其次，目前学界和行业协会所制定的伦理标准缺乏约束力，既缺少具体的技术要求和具体可实施的设计标准，也都缺乏强制性，导致行业内的实践差异较大。⑤ 再次，行业协会的代表性和多样性不足，许多国际性行业协会的成员主要来自发达国家和大型科技企业，导致在制定标准和政策时容易忽视来自小型企业、初创公司和民众以及发展中国家的声音和需求。⑥ 最后，这些协会在跟上AI技术进步步伐方面也面临困难，其发布的指导方针常常无法适应快速变化的行业环境。⑦

这一方面意味着这些社会组织在推动相关企业承担社会责任上作用孱弱，另一方面也意味着社群治理机制在这一领域失灵，即相关行业协会未能使其会员（尤其是那些技术实力雄厚的企业以及在公共或社会部门如大学任职的专家学者）有效地发挥作用，同时也与相关协会成员规模较大、成员异质性增高等容易导致社

① 闫坤如：《人工智能价值对齐的价值表征及伦理路径》，《伦理学研究》2024年第4期。

② 联合国教科文组织，“人工智能伦理问题建议书”，2022年，https://unesdoc.unesco.org/ark:/48223/pf0000381137_chi。

③ Emmanouil Papagiannidis, Patrick Mikalef, and Kieran Conboy, “Responsible Artificial Intelligence Governance: A Review and Research Framework,” *Journal of Strategic Information Systems*, vol. 34, no. 2, 2025, 101885.

④ Christopher Wilson, “Public Engagement and AI: A Values Analysis of National Strategies,” *Government Information Quarterly*, vol. 39, issue. 1, 2022, 101652.

⑤ Anna Jobin, Marcello Ienca, and Effy Vayena, “The Global Landscape of AI Ethics Guidelines,” *Nature Machine Intelligence*, no. 1, 2019, pp. 389–399.

⑥ Kate Crawford, *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*, New Haven: Yale University Press, 2021.

⑦ Reuben Binns, “Fairness in Machine Learning: Lessons from Political Philosophy,” *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency (FAT)*, 2018.

群失灵的因素有关。因此，社会力量自发参与算法偏误的治理，容易呈现众声喧哗但效果微弱的社群失灵格局，难以实现算法的“伦理对齐”。

（三）政府部门：技术、场景与协同限制

总体来说，专门针对算法偏误治理的政府监管措施还比较少，相关举措更多是作为整体性 AI 监管政策中的一部分而存在。我们将当前算法偏误的政府监管思路归纳为技术矫正、应用场景限制和人机协同三种。

首先，技术矫正思路着重从技术层面出发，加强对算法开发技术的过程矫正与可问责性建构。早期 AI 研究突出算法的“黑箱”模式，将之作为人类无法矫正算法偏误的理由之一，强调 AI 决策的“责任差距”[①]。但随着人工智能技术的进一步发展和民众对其可解释性的呼吁，当前欧美 AI 监管重点开始转向算法决策的透明性和可解释性，建立算法和算法审计机制，维护个体在算法决策中的权利。[②] 但技术矫正政策在实践中也遇到了不少难题。一是对很多人工智能相关公司而言，算法通常作为商业机密受到保护，政府强行要求开放会招致反对。[③] 二是对公众而言，大多数人缺乏相关知识，在算法开放中能起到的监督作用有限。

其次，应用限制思路着重对受算法偏误影响较大的关键领域施加应用限制，防止其不良影响。欧盟委员会于 2021 年发布了《人工智能法案》（Artificial Intelligence Act），将人工智能应用划分为不可接受的、高风险、低风险和无风险四个层级，并据此干预措施从完全禁止到自由放任区分为了四级。[④] 2022 年 1 月 4 日，中国国家互联网信息办公室、工业和信息化部、公安部、国家市场监督管理总局联合发布正式公布《互联网信息服务算法推荐管理规定》，规定“算法推荐服务提供者应当向用户提供不针对其个人特征的选项，或者向用户提供便捷的关闭算法推荐服务的选项。用户选择关闭算法推荐服务的，算法推荐服务提供者应当立即停止提供相关服务”[⑤]。这是全世界第一个对算法推荐行为加以具体约束的法案，有效保护了公众在人工智能社会应用中的知情权和选择权。

这种限制短期内可以削弱算法的负面影响，但从长期来看，未来的社会是计算技术高度发达与算法应用更加广泛的社会，人们的生活与社会过程将会更深刻地被算法塑造。[⑥] 此外，硬性法律规制体系虽然约束力强、审慎性高，但会因其较强的滞后性而难以应对算法、数据等复杂多变的监管需求。[⑦] 因此，单纯的应用限制举措难以真正起到纠偏作用，反而还会限制人工智能的发展和纠偏方案的优化。

再次，强化人机协同思路强调人在算法决策中的角色发挥，期望将人的理性因素嵌入到算法决策之中。按照政策要求的人的参与强度，可以将人在算法决策中的角色分为三类。最严苛一级是完全禁止自动化决策，有 20 个国家/地区的政策文件中采取了这项举措，包括 8 个欧盟成员国、阿根廷、毛里求斯、肯尼亚、南非等国都出台了重要领域完全禁止自动化的政策。较宽松的一级是在算法自动决策中加入人工监督，主要是在刑事司法和儿童福利风险评估领域，强调对人的自由裁量权的重视。最宽松的一级是“有意义的人工输入”，强调人的参与不仅仅是象征性的，而且应该要有意义，对自动决策进行有意义的人工审查。欧盟委员会关于人工智能法案的提案和美国行政会议委托的一份报告都强调需要避免简单化和肤浅的人类监督形式。[⑧]

---

① Matthias Andreas, “The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata,” *Ethics and Information Technology*, vol. 6, 2004, pp. 175-183.

② Jakob Mökander, Prathm Juneja, David Watson, et al., “The US Algorithmic Accountability Act of 2022 vs The EU Artificial Intelligence Act: What Can They Learn from Each Other?” *Minds and Machines*, vol. 32, 2022, pp. 751-758.

③ Marijn Janssen and George Kuk, “The Challenges and Limits of Big Data Algorithms in Technocratic Governance,” *Government Information Quarterly*, vol. 33, issue. 3, 2016, pp. 371-377.

④ Regine Paul, “European Artificial Intelligence ‘Trusted throughout the World’: Risk-Based Regulation and the Fashioning of A Competitive Common AI Market Regulation,” *Governance*, vol. 18, issue. 4, 2024, pp. 1065-1082.

⑤ 中华人民共和国国家互联网信息办公室，“互联网信息服务算法推荐管理规定，”2022-01-04，https://www.cac.gov.cn/2022-01/04/c_1642894606364259.htm。

⑥ 张敏：《算法治理：21 世纪的公共管理现代化与范式变革》，《政治学研究》2022 年第 4 期。

⑦ 贾开、蒋余浩：《人工智能治理的三个基本问题：技术逻辑、风险挑战与公共政策选择》，《中国行政管理》2017 年第 10 期。

⑧ Ben Green, “The Flaws of Policies Requiring Human Oversight of Government Algorithms,” *Computer Law & Security Review*, no. 45, 2022, 105681.

但人工监督算法决策的模式也存在着两大缺陷。首先，政府官员大多数时候并不具备参与监督的能力，既了解AI也懂得公共治理的人才稀缺。其次，参与监督的专业人士也难免带有难为众知的偏见，这也会使得人工监督政策非但不能防止算法决策的潜在危害，反而会提供了一种虚假的安全感。

五、算法偏误的元治理：政府职能的深化

如上可见，在当前对算法偏误的治理问题上，政府、市场与社会都有参与，但呈现出碎片化与同质化并存的情况：三方主体都较为着力于人工智能伦理治理原则的开发和发布，积极推广大同小异的标准或倡议，但主体间缺乏战略联系和明确分工。与此同时，随着人工智能产业的持续蓬勃发展，政府监管加强与市场活力增加之间开始出现紧张关系。因此，推动算法偏误的“元治理”，有效协调政府、市场与社会多方主体在治理中的相互关系，在保持各主体自主性的同时增强各主体间的协作互动，增进多种治理机制的互补嵌合，创造更高级别的“协商秩序”，是当前算法偏误治理的急务。在这个过程中，政府的意愿最强且对其他治理主体最有影响力，因此这里从政府视角出发，就算法偏误的元治理模式与路径问题展开讨论。

（一）持续推动多元治理工具的开发和使用

元治理不同于直接的行政干预，而是强调政府要采用更灵活、多样、细致的手段来实现治理目标。在元治理的工具库中，既包括“插手型工具”（hands-on tools），如遴选参与者、资源分配、制定规则等，也包括“放手型工具”（hands-off tools），如激活外部参与者、搭建互动平台、议程设置协商等[①]，还包括价值塑造、目标确定、认同培养、规范形成等软方式。多种工具的配合使用。

作为一项新兴事物，算法偏误的治理手段也必然要有相应的创新。但如前所述，当前政府部门的治理手段整体沿用了“问题响应”模式，存在着碎片化的问题。整体来看，应对性模式纠结于应该注重美国学者倡导的“基于过错的责任”还是欧盟式的“严格责任”[②]，或者说如何在AI开发者“过失责任”与“无过错责任”之间保持权衡[③]，但就元治理所要求的多元手段的应用还比较缺乏。政府如果仅仅基于规则采用命令与惩治等传统手段实施行政化治理，不仅会削弱治理参与者的自我规制能力，还有可能遭到参与者的抵触，大大降低其持续、深度参与治理的积极性。因此，从单纯的行政化治理转向元治理至关重要，摒弃一味管制理念和一体化思维是其关键；就此，政府可以综合运用权威、信息、经济等多种政策工具，从源头、过程、情境、氛围等多角度入手，确保人工智能技术的开发和应用符合公众的期待和诉求。

（二）增强多元治理主体之间的协调性

元治理强调多方主体间的互补性。在算法偏误问题上，鉴于其涉及技术逻辑的必然性、人为理性的有限性、历史的局限性、商业利益的多元性、社会的复杂性等因素，在算法偏误的元治理中，推动并实现由各方单治到协调共治的转变，才可以有效应对单一主体的治理失灵，形成以政府与市场、社会协作互动为特征的治理体系。

当前三类主体的举措缺乏协作互动：市场主体科技公司注重技术手段的使用和产业规范的塑造，社会主体注重对算法歧视的发现和行业倡议的发起，政府主体则注重对算法使用的监管。这三者相对独立，使得各自的治理行动都有失灵的风险，即面临所谓的“编排缺失”（orchestration deficit）[④]。编排的分析框架强调，参与合作的各方间关系并不是平行、平等的，其中“编排者”（orchestrator）如同交响乐队的指挥一样处于中心位置，输出议程与组织关系等，依靠影响其他主体的知识结构、规范取向、技术能力等方式，建构治理网

① Eva Sørensen and Jacob Torfing, “Making Governance Networks Effective and Democratic through Metagovernance,” *Public Administration*, vol. 87, issue. 2, 2009, pp. 234-258.

② Jon Truby, Rafael Brown, Imad Ibrahim, et al., “A Sandbox Approach to Regulating High-Risk Artificial Intelligence Applications,” *European Journal of Risk Regulation*, vol. 13, issue. 2, 2022, pp. 270-294.

③ 戴昕：《无过错责任与人工智能发展——基于法律经济分析的一个观点》，《华东政法大学学报》2024年第5期。

④ Kenneth Abbott and Snidal Duncan, “The Governance Triangle: Regulatory Standards Institutions and the Shadow of the State,” in *The Politics of Global Regulation*, Walter Mattli and Ngaire Woods(eds.), Princeton: Princeton University Press, 2009, pp. 52-91.

络，实现组织目标。[①] 在算法偏误的治理上，强制性的行政化治理并不合适，这种“编排治理”模式更适合于政府在治理中发挥作用。这种模式需要由政府出面担当这个编排者亦即元治理者的角色，促成三类主体行动间的协调性和完整性，避免不和谐的情况出现，共同推进治理目标的完成。

（三）注重面向公众的传播服务

数字时代的国家治理现代化建设不能只关注制度规则，还要关注福利价值；不能只关注治理结果，还要关注治理构成；不能只关注政策意图，更要关注人民感受。[②] 在算法偏误问题上，公众既有可能是问题的发现者、感受者甚至是受害者，也将是成功公共治理的最终受益者。为此，政府与市场和社会主体协同，注重公众在偏误治理中的参与，加强面向公众的科学传播，增强公众对算法偏误问题的客观性认知，塑造形成积极互动、协商共决的治理氛围，是算法偏误治理与社会秩序重建的重要支撑。

首先，科技公司提高其算法透明度以加强与公众的沟通，正日益成为市场机制运作下市场主体提升其自身市场竞争力的战略之举。例如，小红书在 2025 年初从一家中国公司一跃成为著名全球化公司，其算法具有市场亲和力功不可没。其平台算法的改变便利了该平台“用户友好的算法想象”以及“算法公平想象”，塑造了平台利他性文化风格和真诚、温暖的情感网络，形成了对用户的深度吸引力。[③]

其次，社会组织可以在政府合法性背书和实质性支持下更好地发挥桥梁和中介作用。这其中，行业协会可以提供平台，促进不同利益相关者之间的对话与合作，包括政府、企业和学术界，这种跨界合作能够帮助制定行业标准和最佳实践，确保 AI 开发和应用符合社会的道德和法律规范。行业协会可以与政府合作，参与政策制定过程，确保相关法规既能促进技术创新，又能保护公众利益。此外，协会可以推动技术的透明性，鼓励企业在算法和数据使用上提供更多信息，帮助社会公众理解人工智能的运作机制，减少误解与恐慌。让协会治理中社群机制所促成的行业协调与企业在市场机制推动下提升竞争力的自主行动在政府的支持下形成互补嵌合的格局，算法偏误对社会的危害有望大大降低。

最后，政府推动算法治理的开放和透明至关重要。在市场机制的运作正在增强企业自我治理算法偏误的内生动力之际，政府协同行业协会采取助推式劝导行动，推动科技公司提高算法透明度正当其时。同时，政府还可以设立专门的反馈渠道，收集公众和企业对人工智能政策的意见，这有助于政策的灵活调整，确保其适应社会需求。政府还须通过科学研究和政策咨询基金配置等政策工具，支持社会科学和人文学科共同体以及相关智库开展算法治理的研究，使公众中的学者通过科学共同体中社群机制的运作深度参与到对算法偏误的共同治理之中，尤为必要。[④]

六、结论与讨论

人工智能在产业变革和国家治理中的应用迅猛，在引发全社会对其发展势头欢欣鼓舞的同时，其算法偏误问题也引发了担忧。整体来看，算法偏见源于机器学习的技术特征、人类决策中的有限理性和机器学习的自强化循环等因素，从根本上反映的是植根于社会模式和偏见的行为方式，其实是传统歧视和不公平的 AI 化。因此，探索对算法偏误问题的公共治理，有利于推动 AI 的向善发展，推进科技创新突破与社会秩序稳定之间的妥善衔接。

算法偏误并非仅是一个技术问题，对其治理也必须纳入多方主体的力量。从既有探索来看，在市场利益、社会责任、公众舆论等的推动下，当前政府、市场和社会均对这一问题进行了一定的回应，但由于主体间缺乏协同配合，导致现有回应存在着缺陷：市场回应缺乏持久性和系统性，经常会因为逐利动机和市场竞争而被搁置；社会回应更注重伦理规范构建，但缺乏对市场主体的嵌入、约束；政府仍然有传统行政管制思路的

① 汤蓓：《试析国际组织的协同治理策略——以国际劳工组织推广“社会保障底线”政策为例》，《国际观察》2017 年第 3 期。
② 尹振涛、徐秀军：《数字时代的国家治理现代化：理论逻辑、现实向度与中国方案》，《政治学研究》2021 年第 4 期。
③ 别君华、曾钰婷：《算法想象的平台参与及情感网络——基于“小红书”的用户分析》，《中国青年研究》2024 年第 2 期。
④ 吕鹏、周旅军、范晓光：《平台治理场域与社会学参与》，《社会学研究》2022 年第 3 期。

影响，因担忧对人工智能发展的负面影响而有所踌躇。这显示出，算法偏误的治理需要从推进行政、市场、社群机制互补嵌合的角度来考量。

市场机制主要在于驱动科技公司通过算法改进来营造良好社区生态环境，提升用户体验，从而提高用户黏性，增强市场竞争力；社群机制主要在于社会组织的运作增进在技术应用与公众认知之间的衔接，激励科学共同体积极面向公众，提升相关议题的社会认知和社会接受度；行政机制主要在于通过权威主导和引导，推动科学共同体（既包括AI科研共同体也包括社会科学人文研究共同体）积极参与对算法偏误的社会治理，推动科技治理与国家治理的目标和需求相契合。由此，三类主体通过共建共治、三种机制通过互补嵌合促使算法的公共治理体系制度化，方可在AI时代切实推动科技变革与社会秩序的共进发展。

〔本文为国家社会科学基金2021年青年项目“新冠肺炎疫情防控中科学—政府—大众关系的社会治理研究”（21CZZ014）、国家自然科学基金重点项目“数字政府驱动的治理范式变革研究”（72434004）、国家社会科学基金重大委托项目“中国公共管理自主知识体系构建的研究”（25@ZH002）的阶段性成果。郭凤林为本文通讯作者。〕

（责任编辑：王胜强）

The Public Governance of Algorithmic Bias: The Way of Collaborative Interactions among Government, Market, and Society

GU Xin, GUO Fenglin

Abstract: Algorithmic decisions driven by artificial intelligence are subject to the combined influence of multiple factors, including the inherent logic of the algorithms, human bounded rationality, and user feedback loops. This has led to issues such as insufficient representation of specific social groups, high error rates in identification, and biases in policy applicability. In response, the government, market, and social actors have all made certain efforts to address these issues, exploring into technical logic, ethical principles, and application limitations. However, the breadth, depth, and sustainability of these efforts are relatively limited. To build a public governance system that addresses algorithmic bias, it is necessary to promote collaborative interactions among the government, market, and various social stakeholders, facilitating the complementary embeddedness of bureaucratic, market and community mechanisms. The government must shift from merely relying on bureaucratic mechanisms to performing meta-governance functions. By establishing a public governance system centered on collaborative governance and shared responsibility among multiple stakeholders, and by institutionalizing the public governance of algorithmic bias, we can create new productive relationships and achieve the symbiotic development of technological transformation and social order.

Key words: algorithmic bias, public governance, community mechanism, market mechanism, meta-governance