DOI:10.15992/j.cnki.31-1123/c.2025.03.007

人机社会的治理转向

邱泽奇

进入21世纪第三个十年,新一代人工智能(以下简称“机器智能”)技术正在以前所未有的方式进入人类的生产和生活。与传统意义上被动地执行人类指令的工具型技术不同,机器智能(尤其是大型语言模型、算法型决策系统等)不仅具备语言理解、信息整合、判断生成、作品产出等能力,还通过人类对这些能力的运用和不断创新,多路径地参与到人类社会的核心领域(如治理)。这种参与不仅快速改变了人类的生产与生活,也从基础上改变了社会的构成,[1]把纯粹的人类社会转变为人机互生的人机社会,并重塑了治理模式。

埃吕尔(Jacques Ellul)在讨论技术社会的理论脉络时提示说,在技术与社会长时段的关系发展中,一个清晰的演进趋势是技术对人类的影响力不断增强,[2]其中表现之一是技术从后台逐步走上前台。拉图尔(Bruno Latour)的理论洞见[3]启发我们可以将技术(特别是机器智能)视为社会关系网络的组成部分。然而,埃吕尔对技术影响力演进趋势的强调止于20世纪60年代对技术自我增强的判断,拉图尔及其追随者对非人行动者的权能赋予在强调准行动者属性的同时,[4]却忽视了人类的主体性及其在人机关系中的理论张力。笔者认为,把机器等同于具有自由意识和自主权力的行动者,既不符合社会事实,也不符合人类的利益。

本文重拾埃吕尔对技术与人类关系趋势的分析,延续拉图尔准行

1. Tsvetkova, Milena, Taha Yasseri, Niccolo Pescetelli, and Tobias Werner. 2024. “A New Sociology of Humans and Machines.” *Nature Human Behaviour* (8): 1864–1876.

2. Ellul, Jacques. 1964. *The Technological Society*. New York: Vintage Books.

3. Latour, Bruno. 2005. *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford: Oxford University Press.

4. Muhle, Florian. 2024. “Robots as Addressable Non-Persons: An Analysis of Categorial Work at the Boundaries of the Social World.” *Frontiers in Sociology* (9). https://doi.org/10.3389/fsoc.2024.1260823.

动者的理论脉络，从人机社会的人类主体性[5]出发，关注机器智能深度介入人类社会后带来的治理挑战，即面对机器由人类的工具转变为与人类互生的准行动者的过程，作为社会主体以及机器的创造者，人类是如何理解社会治理的对象、责任分配与策略安排的。在特别强调人机社会的人类主体性，即人类在价值设定、责任承担、制度建构等领域主导地位的基础上，本文提出了人机互生的治理观。笔者主张，治理人机社会不仅需要创新治理技术，更需要创新治理理念，面向由人类与技术互动互生形成的人机社会生态系统。

本文的讨论分为三部分：首先论证人类社会已然来到人机社会，即机器智能已然介入从个体思考到主权国家治理的人类行动，进而成为社会的有机组成部分；接着分析这一变化给传统社会治理理论与实践带来的系统性挑战；最后探讨从治理人工智能向治理人机社会的转向，提出以人类价值为主导和以协同机制为核心的治理重构。

一、已然到来的人机社会

工业革命的一个标志是机器介入了生产与生活的方方面面，进而让以人类为唯一行动者的社会转型为人类与机器协同的社会。[6]然而，这不是人机社会，而是由人类主导、机器作为被动执行人类指令工具的社会，无论是在农田里、牧场上、大海中，还是在工厂车间、城市空间，作为工具的机器都没有主动的决策和行动，即使在自动化生产线上，机器也只是由人值守的生产工具。

可是，在经历了两百多年实践后产生的机器智能的社会化应用彻底改变了工业社会人机关系的格局，带来了一场比经济革命更根本的社会革命。机器不再只是从前人们手边的工具，在埃吕尔意义的发展趋势里，机器除了依然是被动地执行人类指令的工具之外，也具有了主动的互动能力，且在互动中进一步成为人类体力、脑力乃至情感的指向对象。以大型语言模型为主力的生成式人工智能的兴起，让机器不

5. 邱泽奇.2025.数智时代的人类团结——技术变革与理论重构[J]. 智能社会研究(1)：43-69.

6. 对工业革命以来的社会发展，有极为丰富的文献，其中一个关键的讨论是将工业革命划分为多个阶段，如从工业革命 1.0 到工业 4.0。笔者的观点略有不同，而是认为机器智能带来的是与工业革命具有同样意义的革命性变革，进而将人类社会的发展阶段划分为农业社会、工业社会和数智社会。

仅能处理自然语言、生成复杂文本图像视频内容、与人类进行类自然对话甚至类情感交流，还在诸多专业领域（如法律、医疗、教育、科研）展示出类似甚至超越人类专家的能力。[7] 换句话说，对语言理解、知识生产、情感交流等认知与非认知能力及其行动瓶颈的突破，使机器智能不再只是被动的工具，还是具备了与人类“共情、共识与共作”潜能的类人主体。

在个体层次上，机器智能已深度介入人类的社会交往。[8] 通过对社交历史的简要回顾，我们可以发现，从电报、电话的社会化应用开始，人类便进入了以机器为中介的社交时代。值得注意的是，即使基于数字媒介（如微信、抖音、小红书等）的交往把电报、电话时代少数人的短交往变成所有人的长交往，现在的交往依然还是以机器为中介的社会交往。然而，机器智能使社会交往出现了一个革命性拐点：从以机器为中介的交往转变为以机器为对象的交往。比如，机器（如装有 DeepSeek 之类大模型的设备）不仅被当作生产与生活的伙伴，还被视为人类可以进行情感倾诉的对象。人类在面对学习、生活、情感乃至自我认同时，不再只是从同伴那里寻求回应，还能从机器智能那里寻求支持和安慰。[9] 人类与机器智能之间的这种关系虽然不是传统意义上人与人之间的相互理解，却在心理上产生了社交代理效应，形成了人机之间的能力支持、情感投射与回应期待，从而建构出一种主观的关系实感，[10] 这也意味着社会关系不仅仅是人与人之间的交流与互动，还纳入了由机器支持的关系体验或以算法为中介的社会交往。事实上，数智社交平台的推荐机制不仅影响人们看到什么，还塑造着人们的政治立场、社会认同乃至情感波动或转变。[11]

---

7. OpenAI. 2023. “GPT-4 Technical Report.” arXiv：2303.08774；Bubeck，Sébastien，*et al*. 2023. “Sparks of Artificial General Intelligence：Early Experiments with GPT -4.”arXiv：2303.12712.

8. 邱泽奇. 2024. 重构关系：数字社交的本质[M]. 北京：北京大学出版社.

9. Weidinger，Laura，*et al*. 2021. “Ethical and Social Risks of Harm from Language Models.” arXiv：2112.04359.

10. 特克尔，雪莉. 2014. 群体性孤独：为什么我们对科技期待更多，对彼此却不能更亲密？[M].周逵、刘菁荆，译.杭州：浙江人民出版社.

11. Couldry，Nick and Andreas Hepp. 2016. *The Mediated Construction of Reality*. Cambridge，MA：Polity.

在组织层次上，机器智能已然从信息处理的工具演变为组织机制的有机节点。例如，一方面，欧洲多国的社会福利算法已被用于识别潜在的欺诈行为[12]；另一方面，嵌入组织行动的机器也有可能在不透明场景里剥夺弱势群体的福利资格[13]。这一事实意味着，机器智能可以通过算法或代码规则自主地执行法律和政策，对个体和组织做出决策或采取行动，进而产生直接的社会结果或后果，采取实实在在的与技术官僚类似甚或一致的社会行动。随着逐步被纳入教育、健康、金融、军事等关键领域的组织机制、决策流程和行动环节，机器智能在组织中的角色已然深度影响人类从行动者能力拓展到权力结构组成的各个方面，例如，2020 年英国使用机器智能替代 A-level 学生考试评估。[14] 尽管之后的大规模抗议迫使政府撤销了人工智能的决策，却也不能抹杀机器智能已然进入组织机制并在诸多环节针对人类进行决策和行动的事实。这些实践证明，机器智能已不仅仅是服务性的技术，还成为了组织机制的有机节点，且已深深嵌入组织权力运作。

在社会层次上，机器智能的参与也已然或正在广泛地改变个体的行为模式、社会的运行方式与制度的运作逻辑。以公共服务为例，城市管理、交通调度、政务服务等方面的大量工作都在采用基于机器智能的自适应系统进行运行和优化。尽管早在 20 世纪 90 年代，一些科技企业在倡导发展智慧城市时，机器智能就已经是其中的关键技术，[15] 但一直到机器智能带来技术跃升才使得机器智能变得可用，且在社会层次上被赋予了辅助或替代人类的权力，进而成为拥有判断能力和决策权力的社会行动者。例如，新加坡的智慧城市治理架构已将机器智能作为治理的关键节点，包括城市动态监管、舆情预测与危机响应；[16] 在中国诸

12. Van Bekkum, Marvin and Frederik Zuiderveen Borgesius. 2021. “Digital Welfare Fraud Detection and the Dutch SyRI Judgment.” *European Journal of Social Security* 23(4): 323–340.

13. Eubanks, Virginia. 2019. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: Picador.

14. Victoria, Crisp, *et al*. 2025. “A Structured Discussion of the Fairness of GCSE and A Level Grades in England in Summer 2020 and 2021.” *Research Papers in Education* 40(1): 44–71.

15. 最早提出智慧城市的是 IBM 公司。参见：Montes, Josefer. 2020. “A Historical View of Smart Cities: Definitions, Features and Tipping Points.” *SSRN Electronic Journal*, http://dx.doi.org/10.2139/ssrn.3637617.

16. Smart Nation Singapore. 2024. “Smart Nation 2.0: A Thriving Digital Future for All.” Ministry of Communications and Information, Singapore. https://file.go.gov.sg/smartnation2-report.pdf.

多城市的治理实践中，从人口管理到事件发现与处理，从呼叫回应到事务性决策乃至关键决策，都能发现机器智能的身影。这些实践表明，机器不再只是工具性嵌入社会，其行动结果已然被直接纳入社会考量，其行动逻辑也被拟人化，已成为社会秩序的一个组成部分。

从个体、组织再到社会，机器智能的介入虽然在形式上依然是中介，却已不是只被动地执行人类指令，而是还会主动参与人类指令的生成，有建议和产出，还能由指令生成媒介化的现实（mediating reality），进而成为与人类类似的互动对象以及社会现实的有机组成部分。不仅如此，人类与机器的互动也从传统的体力类型转向体力与脑力的交互并在。笔者认为，这一转变是革命性的，也是历史性的，它意味着人类与机器关系的本质性转变，机器对社会生产与生活的介入和作为准行动者的存在已经是不争的事实，人类社会已然彻底转变为人机社会，斯坦福大学的一份最新报告也再一次证明了这一判断。[17]

这一革命性变革可以部分地用拉图尔的“行动者网络理论”来解释。拉图尔将行动者界定为在活动中使其他行动者发生变化的任何实体，无论是人类还是非人类。[18]这种去人类化的行动者定义的优势在于打破了传统社会学将行动能力只赋予人类的习惯性思维，允许在社会分析中承认非人行动者（如技术装置、机器系统）的现实影响力，如此说来，机器智能就完全可以被理解为准行动者，它不再只是社会行动的外部条件，还是生成社会事实的参与者。

然而，用“行动者网络理论”来解释的风险在于，它削弱了人类作为行动者的能动性与伦理责任。从人类主体性的视角看，人类与机器智能虽然都产生社会影响，但两者在行动目标、价值取向、伦理判断等诸多议题上的差异依然是本质性的。一方面，人类具有自由意志和自主权力，还具有目的设定与意义赋予的能力；另一方面，尽管对人工智能未来是否会发展出自由意志与自主权力存在争议，但至少在目前，机器智能还在依据人类设定的参数与目标运作，其能动性和自主性仍源自人类的设定。因此，站在人类主体性的立场，我们可以接受机器智能是具

17. Human-Centered Artificial Intelligence. 2025.“The 2028 AI Index Report 2025.”Stanford, CA: Stanford University.http://hai.stanford.edu/ai-index/2025-ai-index-report.

18. Latour, Bruno. 2005. *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford: Oxford University Press.

有效率、效果、效能的准行动者，却无法将类似于电车难题的责任归咎于机器，因为人类依然是人机社会责任与判断的能动主体。[19]换言之，人机社会虽然将机器视为准行动者，但并不意味着机器具有与人类对等的主体性，自由意志和自主权力依然还属于人类，也即，机器具有行动权能，但目的与意义权能仍然属于人类。

二、机器智能带来的治理挑战

现代治理的发展，依托于一整套围绕人类行动者的理论与实践框架。在理论上，早期的治理思想，如霍布斯的“国家主权观”，是将治理建立在个体暴力与非理性防御的基础上，即通过建立一个绝对权威（利维坦）的方式达成社会秩序的持久性。[20]站在对传统治理理念批判的立场上，福柯把视角引向对现代治理微观权力的剖析，将治理理解为贯穿身体、心理与社会技术的规训。[21]在极度宏观或极度微观的框架中，无论是作为威胁的自然人，还是作为被规训的社会主体，其核心前提都在于：人是社会的唯一行动者，也是被治理的对象。治理正是通过对人类自由意志、自主权力及其社会行动的引导与规制，建构和维系着社会秩序。

在实践中，韦伯（Max Weber）以理性化为特征的科层制模型[22]为20世纪的国家治理提供了制度化路径：明确的责任层级、法定的行动规程和基于法律的合法性。出于对现代社会治理体系理解和解释的目的，系统论[23]进一步将社会理解为多功能系统间的信息与能量交换网络，将治理解释为反馈调节机制在复杂系统自组织的实现。无论是科层制还是系统论，虽然听起来似乎古老或传统，却依然是当今人们从社会视角理解治理的出发点，人类组织依然是治理的主体行动者，协调和维系社

19. 吉登斯，安东尼.2016.现代性与自我认同：晚期现代中的自我与社会 [M].夏璐，译.北京：中国人民大学出版社.

20. 霍布斯.1985.利维坦 [M].黎思复、黎廷弼，译.北京：商务印书馆.

21. 福柯，米歇尔.1999.规训与惩罚：监狱的诞生[M].杨远缨、刘北成，译.北京：生活·读书·新知三联书店.

22. 韦伯，马克斯.2019.经济与社会（第一卷）[M].阎克文，译.上海：上海人民出版社.

23. Parsons, Talcott. 1951. *The Social System*. Glencoe, IL: The Free Press; Luhmann, Niklas. 1995. *Social Systems*, translated by John Bednarz Jr. and Dirk Baecker. Stanford, CA: Stanford University Press.

会行动的秩序也依然是治理的目标。

然而,无论在何种意义上,既有理论与实践都假设行动者是生物性人类,是可以赋予责任、权利与道德判断的人类行动者。在治理中,传统的非人类始终被视为用于治理的工具而非参与者。在这一逻辑下,治理工具(政策、法律、技术)被设定为针对人类行为的外部约束和引导。当机器智能不再只是被动地执行人类指令的工具,还会主动介入人类事务并影响人类思考与行动时,既有治理框架便遭遇到结构性挑战:过去针对人类行动者的人治与法治架构还依然适用于人类与非人类行动者互动互生的人机社会吗?

即使是粗浅地扫视已经存在的事实,人们也很容易发现人机社会面对的治理困境。

首先是治理主体界定困难,给治理带来无主可治的挑战。[24] 在人类社会,行动者通常是指具有自由意志和自主权力进而具有责任能力的自然人或法人,其核心假定是社会后果可追溯至明确的自然人或法人主体。然而,当机器智能能自主地执行社会行动乃至引发社会后果时,谁是行为的主体常常会变得模糊不清。例如,一辆处于自动驾驶状态的汽车发生事故、一个金融交易算法引发市场波动,在诸多类似的场景里,带来这些后果的社会行动究竟应该视为机器智能的自主行为,还是开发者、所有者的延伸行动?对行动主体法理定位的模糊是因为现行法律责任体系主要是围绕人设计的,且假定事故和错误主要源于自然人或法人的行为,而机器智能的行为显然已经超出了既有的治理主体设定。

带来主体认定困难的本质或许源于机器智能的非人类中心性。现行治理体系以人类行动过失归责为逻辑起点,制度预设是事故源于人的直接干预或监管疏忽,而机器智能会在无实时人工指令场景里生成社会后果。以信贷审批系统为例,机器智能会基于多维度数据判断申请人是否符合放贷资格,其决定往往不会被终端操作人员干预,申请人也难以质疑拒贷或准贷理由。在中国,某平台曾把机器智能决策的信贷审批系统作为一种企业创新以标示信贷的公平与公

24. 这个问题现在已有诸多讨论,最新的讨论可参见:李洋、薛澜.2025 颠覆、调适与协同:责任伦理视域下生成式人工智能的多主体治理机制研究 [J].电子政务,网络首发时间:2025-03-17.20:07:50.

正，这种去人类化的因果链条直接突破了现行责任归属框架，无面对面、无主体责任链、无可问责路径的实践场景也显著弱化了治理应有的可追溯性原则。

更深层的冲突还在于主流治理范式仍将机器智能视为一种工具，试图通过与机器智能关联的责任链条完成间接归因。[25] 问题是，所谓完美的理论逻辑，在实践中却日益暴露出多重漏洞，比如，算法决策具备数据学习、策略迭代等自主特征，任何单一主体都难以声称对机器行为拥有完全控制能力，于是，治理实践便会陷入某种双重悖论：一方面，法律无法赋予机器智能独立人格以直接追责；另一方面，简单将算法后果归咎于人类开发者/使用者，又忽视了机器智能在数据处理、决策执行环节的相对自主性。

这种主体认定困境的理论根源在于，传统治理理论的人类中心主义基石被机器智能的实践所动摇。在数据驱动的治理模式中，人类被解构为可计算的数据分身（dividuals），[26] 其完整的主体意志被算法系统肢解并重构；同时，机器算法因为其中立技术的伪装而成为实际治理权力的载体，进而在事实上使经典社会理论中具有自由意志和自主权力的理性行动者被降维为多汇聚的数据点，且因算法黑箱特征使得治理体系的合法性根基（即主体共识与责任可归属性）面临根本性危机。

笔者认为，人机混合治理场景催生的主体很难界定，本质上是人类社会在向人机社会的转换过程中带来的结构性矛盾。要突破这一矛盾，需要同时超越人类主体与技术工具的二元框架和机器智能的准行动者框架，并在承认机器智能有限自主性的基础上，以人机社会整体而不是以人类和机器智能分别为对象，建构责任分配的伦理原则与法律机制。

其次，责任归属[27]的模糊进一步引发治理问责的机制危机。现代治理强调责任清晰，能明确认定谁对决策和行动后果负责。然而，在人机混合的行动中，责任链条往往跨越多个和多层环节后变得支离破碎。在

25. 诸多归责依然以人类为责任对象，即使是间接责任对象。

26. Deleuze, Gilles. 1995. “Postscript on Control Societies.” In *Negotiations, 1972–1990*, edited by Deleuze Gilles. New York: Columbia University Press: 177–182.

27. 涉及治理的文献都会讨论归责议题，却极少讨论行动集体性对归责的影响。

机器智能广泛渗透到人类生产、生活、情感活动之前，一张普通的数字交通卡便已构建起非人行动者之间的多重代理链，[28]其中就集成了开发者的观念、机构的目标、并非全面的历史数据和技术自身的不确定性。尽管拉图尔认为由行动构成的社会网络是由一连串节点共同协作实现的，[29]然而，除了认定主体，治理还需要在主体与责任之间建立可靠的关联。在人机社会，决策和行动由算法设计者、数据来源、模型参数、训练环境、使用提示词等共同生成，与主体的数据分身具有相对性，责任也被稀释在复杂的人机数据点位，使得机器智能的决策与行动具有高度的集体制造特征。

特别需要注意的是，集体制造意味着不仅存在人类和机器各自的决策和行动，更关键的是还有人类与机器多重互动的决策与行动。然而，在现行法律与伦理层面，集体制造往往意味着集体无责或责任被稀释，尤其是人类与机器经过多轮复杂交互产生的集体制造，完全可以理解为是承载着责任的复杂数据集。一旦出现偏差或损害事件，若按传统方式简单归咎于单一对象，就相当于让单一对象为机器智能的内在错误承担责任；若一味追究开发者的责任，又会因为机器智能自学的特点而难以界定过失。许多机器智能内部决策的过程高度复杂且缺乏可解释性，当决策过程对利益相关方不可解释时，就意味着在事实上无法落实责任，甚至难以确定错误的来源，从而会导致典型的责任真空，给法律和社会伦理带来前所未有的挑战。如果责任不明确，治理便失去了问责的核心和基础，进而会威胁治理的正当性，带来的正是治理危机。

例如，在美国发生的机器智能图像生成工具因训练数据的偏差而歧视性地再现某族群形象的事件中，[30]当公众质问是谁的责任时，平台声称算法是自我学习的，平台无权控制结果，受影响者也难以在法律上明确诉讼对象。另一个典型例子是，在与自动驾驶系统相关的重大交通事故中，存在系统把实体障碍物误判为雨雾进而导致严重碰撞致人死

28. 邱泽奇.2018.技术化社会治理的异步困境[J]. 社会发展研究(4):2-26.

29. Latour, Bruno.2005. *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford: Oxford University Press.

30. Birhane, Abeba, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. "Multimodal Datasets: Misogyny, Pornography, and Malignant Stereotypes." arXiv: 2110.01963.

亡或伤残的情况，[31] 非常遗憾的是，在事故发生后的诉讼与裁决中，几乎没有一起事故能明确区分责任。这种人机混杂的归责泥潭正越来越多地导致人机社会治理的伦理困境和制度危机。如果无法确认机器行动后果的责任归属，治理系统的合法性便会被侵蚀。

再次，算法黑箱 [32] 以及因此导致的治理透明度下降问题给治理实践带来了直接冲击。传统科层治理和法治原则强调决策过程的透明公开和可追溯性，这是从社会视角确保治理合法性的基础，即，无论是政策制定依据、行政裁量标准，还是司法裁判逻辑，都需要以社会可感知、可检验、可追溯的方式呈现。然而，在诸多场景里，机器智能的决策机制显然与此不符。由算法产生的决策可能是基于海量维度的数据和复杂模型，但因其太过复杂而致使个人难以理解其中的完整逻辑。人们现在有一个共识：算法决策的一个根本问题在于，连编写算法的人也无法解释具体的判断过程，进而使得决策无法获得人类理性的认可。算法决策透明度的缺失既使得监管者难以穿透黑箱实施有效监督（例如，监管部门无法追溯算法是否遵循法律规则或政策目标），也使得公众难以通过程序正义感知决策的公正性，甚至可能因为不理解决策逻辑而质疑其背后隐藏着利益输送或偏见。尤其是在公共治理领域，如果重要决策来自算法黑箱，社会无法知晓依据，社会公平和公开原则就会受到实质性损害，治理公信力也会因为对暗箱操作的猜疑而被消解。

进一步说，算法决策隐含的偏见或歧视在黑箱环境中更不易被察觉和纠正。诸多研究和案例均揭示，算法可能由于训练数据的历史偏差（如数据集中缺乏特定群体的有效样本）或设计者的无意识偏见而对某些群体产生系统性的不利，例如，招聘算法可能因为训练数据多来自历史招聘记录，历史记录中存在的性别或种族歧视又会被算法学习并延续，从而导致新的招聘结果依然偏向特定群体。在黑箱环境下，这些不公正问题很难通过传统的程序质询机制被发现，因为无论是行政听证、司法审查还是公众参与，都需要以决策逻辑的可追溯性为前提。当算法决策的依据隐藏在复杂模型中时，被歧视的群体甚至无法明确指出歧视究竟发生在哪个环节，更难以通过法律途径寻求救济和

31. Jatavallabha, Aravinda. 2024. “Tesla’s Autopilot: Ethics and Tragedy.” arXiv: 2409.17380.
32. 讨论算法黑箱的文献汗牛充栋，这里不再具体引用。

更正。因此,机器智能的介入有可能使治理过程变得更不透明,也更难监督,从而使得依赖信息公开和理性审议来确保合法性的传统治理模式面临失灵的风险,治理的公平性和目标与技术的不透明性之间出现了尖锐矛盾。

我们都知道,算法正当性,或者说人们之所以运用算法,是因为算法可以带来效率。问题是,如果因为追求效率和结果优化而牺牲过程透明与程序公平,便会对治理的社会合法性造成威胁。公众在缺乏决策参与感和没有知情权的情况下,容易感觉到被排除在决策流程之外,从而对算法治理产生普遍的不信任感和抵触情绪。就像本文前面所提到的,英国民众就曾以评分算法不透明为由,认为算法低估了部分学生的成绩且无法给出合理的解释,最终迫使相关部门撤回相关决策。其实,人类对算法的反对只是表象,其本质是,当治理权由机器智能掌控时,权力的行使就失去了明确的责任主体,机器智能无法承担社会责任,而人类又以算法决定为由推诿义务,进而导致权力无主体。人机相互推诿带来的是人机社会的失序。种种迹象表明,机器智能带来的黑箱决策的确会削弱既有治理机制的透明度和可问责性,使治理既无法通过程序正义获得社会认同,也难以通过结果正义证明自身的价值,进而从根本上挑战了治理的公信力和合法性基础。

治理的一个核心要求是过程的公开性与结果的可审议性,这意味着,治理行为不仅需要产出符合公共利益的结果,还需要以公众可理解、可参与的方式来实践。然而,机器智能治理引入黑箱系统之后,原先针对人类社会治理的这一原则被系统性破坏。机器智能,特别是深度学习模型的内部运作机制,复杂到即便是开发者本人也无法准确追踪每一次决策的内部路径,[33] 由此衍生出三重连锁问题:其一,非解释性,即算法无法向治理对象说明为什么某人被识别为高风险对象或为什么某案被标记为异常,决策逻辑成为“无法破译的密码”;其二,非问责性,即在出现错误时,系统背后的责任者不明,是数据提供者的偏差、算法设计者的失误,还是算法自主学习的“意外”?责权模糊导致行动者之间相互推诿,进而难以维护人类的福祉与利益;其三,非可参与性,即机器智能一旦被设定,即便是开发者也无法介入由算法规则驱

33. Burrell, Jenna. 2016. “How the Machine Thinks: Understanding Opacity in Machine Learning Algorithms.” *Big Data & Society* 3(1): 1–12.

动的运行与反馈过程，此时，人类不仅被挤出治理主体之列，甚至还可能会沦为治理的受害者而无力自救。例如，在公共教育领域，许多国家尝试通过算法对学生进行成绩排名，结果却因算法的歧视性分层（基于地理、族群、历史记录等变量）引发社会抗议，而涉事部门往往以算法模型复杂而无法调整为由拒绝回应质疑，使得社会参与完全失去入口。三重连锁问题挑战了现代治理的基本预设，即治理必须明确可问责的主体，算法黑箱却让治理悬浮在技术架构中，成为无面孔、无责任的支配力量。

福柯曾指出，现代治理的权力并非集中式统治，而是通过无处不在的规训机制实现的，这种规训不再依赖暴力或命令，而是通过统计、测量、数据档案等知识权力机制实现个体的自我规范。[34] 在机器智能介入社会关系、组织行动、社会秩序的治理中，算法成为新的规训技术，通过评分、排名、推荐等软性控制形式引导人们的消费、就业和教育选择，改变个体的行为偏好与社会期待，例如，平台算法通过信用评分、兴趣偏好和点击行为对用户建模，并以此决定其能接触到的信息范围、产品价格和服务等级。这种深层编码的治理不再表现为外在的行政命令或法律制裁，而是一种无声地隐含在算法逻辑下的选择结构，使人类在自觉接受算法安排中失去主体自主性和能动性。算法即规训，它不命令，却通过数据标签和模型运算引导个体向预设方向行动；它不处罚，却通过评分差异实现对人群的筛选和区隔。这种算法权力虽无形但有效，呈现出隐含却强势的主导性的治理特征。算法的运行表面上好像没有明确的权力中心，可实际上，每一个数据维度的设定和每一个模型参数的调整，都在暗中塑造着社会资源的分配格局和个体的生存机会，进而建构并维系着新的社会秩序。

由算法主导的治理切断了传统治理依赖的“透明—参与—问责”链条，透明性因黑箱而消失，参与性因技术壁垒而被剥夺，问责性因责任主体虚化而无法落实，从而让人机社会的治理陷入技术利维坦的困境。机器智能以效率之名吞噬了治理的社会合法性，人类社会却难以用现行制度框架对其进行约束。从这个意义上说，机器智能引发的不只是治理理论与实践的危机，还是对社会秩序的根本性挑战。如何在人机社会

34. 福柯，米歇尔.1999.规训与惩罚：监狱的诞生[M].杨远缨、刘北成，译.北京：生活·读书·新知三联书店.

建构和维系社会秩序，确保人类在人机社会的主体性和权力始终服务于人类的尊严与自由，而非异化为技术系统的附庸，是人机社会必须面对的治理转向。

三、转向对人机社会的治理

事实上，治理变革始终是政治、产业、学界探索的重要议题之一。在机器智能治理领域，已有大量研究和倡议，例如，制定人工智能伦理准则、完善算法审计机制、加强对机器智能安全性的监管等。[35] 遗憾的是，这些努力大多仍将机器智能视为外在于人类和人类社会的技术客体，即使有的注意到了机器智能的能力跃升，还依然将其等同于一般技术。

有鉴于此，笔者主张将视角前移，转向治理人机社会，即，把人类行动者与机器智能准行动者一起作为治理对象加以考量，通过构建人机契约的方式来规范人机行动与互动，把人类和机器智能共同构成的社会纳入算法的设计和监管，这也是笔者所称的“建设性人机互生治理观”。治理人机社会，并非否定对机器智能的监管要求，而是强调治理的重心不仅在于让技术的宗旨回归人类，更在于建构并维系人类与机器智能交织而成的混合系统的秩序。在这一视角下，人类不再是以主宰者姿态单向治理机器智能，反而是要将自身视为人机互生生态的一员，通过协同和引导的方式实现二者的良性互动：既让机器智能不断发挥效率，又让人类始终具有主体性。这一转向意味着社会治理目标、责任和模式将会发生一系列改变。

---

35. 中国、美国、新加坡、新西兰、欧盟等国家和组织都在制定有关人工智能的规则，联合国也针对人工智能治理提出了自己的主张。参见：国家网信办、国家发展改革委、教育部、科技部、工业和信息化部、公安部、广电总局. 2023.“生成式人工智能服务管理暂行办法”，成文时间：2023 年 7 月 10 日.https://www.gov.cn/zhengce/zhengceku/202307/content_6891752.htm；National Telecommunications and Information Administration（NTIA）. 2024. “AI Accountability Policy Report.”Washington, DC: Department of Commerce, U.S. https://www.ntia.gov/issues/artificial-intelligence/ai-accountability-policy-report；European Commission. 2021.“Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence.” Brussels: European Union.https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence；United Nations. 2024. “Governing AI for Humanity: Final Report of the High-level Advisory Body on Artificial Intelligence.”New York: UN. Advisory Body on Artificial Intelligence. https://digitallibrary.un.org/record/4062495?v=pdf.

首先，治理目标呈现出双重性和复合性。在人机社会中，治理既要关注机器智能的行为，使其符合人机安全、伦理和法律的要求，也要关注人机互动互生的良性运行和发展。人类社会治理的焦点在于通过规范人类行动者的行动来建构和维护社会秩序，增进人类福祉。在人机社会中，除了要聚焦人类社会的治理目标，社会治理还必须要将机器智能代表的技术系统的行动置于议程之中，尤其是要把人机交互行动纳入治理，例如，确保算法决策的公平性，减轻乃至防止机器智能滥用带来的危害等。换句话说，对人机社会的治理增加了一类新目标，即机器智能及其产生的机器行动，更准确地说，是机器智能的行动及其与人类交互的行动。因此，治理目标不再单一地针对人类行动，也需要针对机器行动。同时，还需要平衡双重目标：既要保障机器智能有效、可信、可靠，也要维护人类的价值和利益，比如，在医疗领域，既要确保算法诊断的准确性和安全性（技术目标），又要维系医患关系、隐私伦理等社会考量（社会目标），实现人机互生的整体性优化。

其次，治理责任链条的扩展与延伸。在人机互生实践中，将责任归属于单一主体的传统责任溯源与分配模式已难以奏效，为此，新治理范式主张建立覆盖端到端的人机责任体系，让相关参与方承担决策与行动的组合式责任。具体而言，面向人机社会的责任分配，意味着把场景涉及的人机责任作为一个整体纳入考量，采用总分模式，把机器智能开发者、提供者、部署者、使用者等都纳入责任网络，形成基于行动场景的利益相关者责任体。这是因为，只有当完整链条上的每一环都对人机行动者的影响负责并加以约束时，才能有效保障效率，防范风险，确保系统安全。笔者主张，在当下技术环境中，既然责任归属的颗粒度难以细化到行动者最小单元（如人或机器智能），不如就建立基于场景的利益相关者责任群体，让各方都承担足够的责任，以确保人机社会的有效运行与安全可靠。

当然，要使这样的责任机制有效运行，必须进行制度创新，比如，明确法律中不同角色的责任边界、建立责任分配和追溯机制等，尤其是在人机互生机制与场景中建立责任缓冲池，把在技术上难以归责到具体行动者的责任落实到群体性责任主体上来。在实践中，已经有一些国家和行业开始探索共同责任的框架，比如，开发者需要遵守严格的开发规范并对算法结果进行测试，部署者需要对使用场景的合理性

和风险管理负责，使用者也被期望遵循使用准则并报告问题。但是，在此基础上有一点依然是缺乏的——对难以分配的责任还需要建构担责群体。

通过纵向扩展责任链条和建立未明责任池，人机社会的治理将不再简单地沿用传统归责机制的方法将责任归于单一行动者，而是强调人机互生各主体的协同负责和群体责任，且形成由责任扩展带来的权力制衡，比如，开发者有责任接受监管和审核，政府有责任制定规则和执行监督，社会有责任参与监督和反馈。总之，面对人机社会的治理，需要重新建构人机互生的责任体系，使其覆盖人机互动互生的决策与行动，形成闭环的责任分配与安全网络。

再次，治理理论与实践模式更新换代。面对纯粹的人类社会，人们（尤其是社会学）通常会在“国家—社会”的二元框架下讨论治理，要么强调政府单方面管制（如霍布斯式威权），要么强调多元主体参与却依然以人类为中心。然而，在人机社会中，人类与机器的互动涌现出前所未有的复杂场景。面对这些场景，虽然有协同治理和多主体协商治理等新理论范式可以借鉴，[36] 却依然不足以适配人机互生的社会。协同治理强调政府、企业、公民社会等主体多方共同参与决策和实施，适用于应对复杂的公共问题，可它依然只是人类行动者之间的协同。面对人机社会，多主体协商治理必须以主体对等性为前提，机器智能虽然在行动上具备了能动性和自主性，可是，与人类相比，尚不具备目的性和价值性，这意味着机器智能还难以作为独立主体来承担责任。因此，尽管在治理实践里必须纳入机器智能，在理论上依然要找到承担机器智能责任的人类主体。在诸多场景中，如果否定指向单一行动者的简单归责，难以追溯责任主体的场景就必然存在，出于对治理效率的考量，创新治理理论与实践便成为如今不得不面对的导向性问题。

对此，“回路社会”（society in loop）[37] 监管框架或许是一个有益的启示。在本质上，“回路社会”是将多元利益相关者的目的性和价值性通过机制化的方式植入机器智能，使其从代码阶段就建立与社会目的性和

36. Wang, Huanming and Bing Ran. 2023. “Network Governance and Collaborative Governance: A Thematic Analysis on Their Similarities, Differences, and Entanglements.” *Public Management Review* 25(6): 1187–1211.

37. Rahwan, Iyad. 2018. “Society-in-the-Loop: Programming the Algorithmic Social Contract.” *Ethics and Information Technology* 20: 5–14.

价值性的协同。比如，建立利益相关方参与的目的和价值协商机制，在机器智能设计阶段把不同群体的诉求平衡折衷，在运行中引入监督反馈回路，使人类能够持续监测机器智能的决策与行动，并依据特定目的性和价值性不断修订和修正，也即，通过人机之间的制度化协商让机器智能遵循人类议定的目的性和价值性。可是，由于目的性和价值性的动态性和发展性，“回路社会”依然无法处理在治理实践中不断出现的难以归责的部分，进而让之前论述的群体责任或责任池机制至少可以成为暂时选项之一。

除了植入式协商，治理实践还可强调共同决策和合作执行。比如，在涉及公共利益的机器智能应用的治理中，监管机构可以与技术开发者共同制定行业标准和伦理指南，社会团体和专业协会也可以参与其中并提出建议，从而形成多主体共治的格局。多主体、多维度、多层级的协同治理模式可以打破以往政府单向管制和市场自我调节的对立，通过多方互动提高治理的适配性，同时，也呼应了人机社会的特征，即通过多行动者（包括人和机器）的互动互生共同影响治理结果。这也意味着治理理论焦点的转向：从强调控制和科层的传统，转向多元、互动、互生、敏捷网络化动态化治理，以适应人机混合的复杂现实。

最后，治理人机社会，需要一系列新规制和新机制来保障协同治理的落实，比如，尽管人们知道算法透明度和可解释性是难以在短时间内解决的难题，可是，为了破解算法黑箱难题，除了分配群体责任和建构责任池，提升算法透明度和可解释性依然是可努力的方向。事实上，有些国家已经在探索一些制度化手段，例如，强制要求高风险机器智能披露其算法原理或决策依据，鼓励或要求关键算法开源以便于社会监督，开放算法源码和模型以有助于专家和公众审查其中可能存在的偏见与缺陷。当然，出于安全和商业的考虑，完全开源有时并不可行，因此，一些替代方案也开始出现，如独立算法审计机制，即由独立第三方机构定期审查机器智能，评估其合规性和风险，并向社会公布结果。

其他的一些改变还有嵌入多学科、多利益相关方审查的制度和法律法规的更新。包括联合国机构在内的许多组织已经设立了人工智能伦理委员会或顾问委员会，成员包括法律、伦理、技术、社会学等领域的专家。这些多学科团队的主要任务就是对机器智能的推出和应用进行评估，并把关对个人和社会的影响。这种机构化的审核在某种程度上可

以弥补单一主体视角局限，确保在治理决策中融入更广泛的社会考量。还有一些国家或组织也开始制定专门的机器智能治理政策或法律框架，比如，欧盟的《人工智能法案》试图以风险分级的方式规范机器智能，一些国内法则在明确机器智能关联的不同主体的法律责任、机器智能产品准入标准的同时，还明确了发生问题时的救济途径。这实际上是在现有法律体系中引入对机器行为的规制条款。

当然，技术性治理工具也是人机社会治理转向不可或缺的内容，比如，实时监测和追踪机器智能决策、用于记录机器智能行动日志的机制、模拟人机互动互生可能影响的沙盒测试环境等。这些工具既可以提升治理机制对复杂人机社会的感知和响应能力，还可以通过制度和工具的双重创新将人机互生的治理观从原则转化为可操作、可持续的现实，进而建构一种新的治理范式，即人类发挥战略引领和价值裁定的作用和机器提供高效的数据处理和执行能力，二者通过有机的制度接口实现优势互补、互生决策和行动，进而建设一种新的治理图景：人类和机器智能不是对立的治理主体和客体，而是被纳入同一个互生治理框架的整体。人类依然掌握着目的和价值的制定权，机器智能则根据人类设定的规则高效地参与和执行治理过程。这样的人机互生治理框架既坚守了人类的主体性和社会价值，又充分利用了机器智能的工具性和数据能力，为21世纪的人机社会提供了一种可持续的治理转型路径。

责任编辑：张　军

大重构：AI时代社会分层秩序的变动与治理*

李　骏

一、大转型：人类社会形态的演变

18—19世纪，面对工业革命带来的世界大变局，对于如何总结人类

* 本文是国家社会科学基金重大课题“大数据和人工智能发展背景下社会分层状况的新变化”（项目编号：22&ZD188）的阶段性成果。

“当代中国的数字社会及其治理”笔谈

邱泽奇、李　骏、向静林、胡安宁

【编者按】随着数字技术的快速进步，中国已进入高度发达的数字社会。当前人工智能技术的高歌猛进，又把中国推进到数智时代。从20世纪初到21世纪初，中国经历了从农业社会到工业社会再到信息社会的大转变，费孝通先生晚年用“三级两跳”加以概括，并在反思全球化和信息化的基础上提出了著名的“文化自觉”思想。今天的数字社会，是费先生所见信息社会的深化。探讨数字化和智能化对社会各领域的深刻影响，是社会学这门对社会变迁极为敏感的学科面临的时代课题。对于新一轮科技革命，特别是数字平台和人工智能的发展，中国可以说是走在了世界前列，中国社会学理应凭借这种独特优势，形成具有自主知识性质的数字社会解释体系。有鉴于此，《社会》编辑部及时组织“当代中国的数字社会及其治理”专题研讨，专家学者们从数字时代人机社会的新型治理模式、社会分层秩序的大重构、国家治理的新变化、文化领域的新特征等角度发表了许多真知灼见。我们将这次研讨的成果以笔谈形式集结发表，以飨读者。

作者1：邱泽奇　北京大学中国社会与发展研究中心，北京大学社会学系（Author 1：QIU Zeqi，The Center for Sociological Research and Development Studies of China，CSRDSC；Department of Sociology，Peking University）E-mail：qiuzq@pku.edu.cn

作者2：李　骏　上海社会科学院社会学所（Author 2：LI Jun，Institute of Sociology，Shanghai Academy of Social Sciences）E-mail：ccsolj@126.com

作者3：向静林　中国社会科学院社会学所（Author 3：XIANG Jinglin，Institute of Sociology，Chinese Academy of Social Sciences）E-mail：xiangjl@cass.org.cn

作者4：胡安宁　复旦大学社会发展与公共政策学院（Author 4：HU Anning，School of Social Development and Public Policy，Fudan University）E-mail：huanning@fudan.edu.cn