Reclaiming the Value of Thinking

重新发现“深度思考”的价值

2026 人文社会科学智能发展蓝皮书

Blue Book on the Development of AI for Social Sciences and Humanities

主编：陈志敏　吴力波

復旦大學

牵头单位：

教育部哲学社会科学实验室——复旦大学国家发展与智能治理综合实验室

上海科学智能研究院

总 顾 问：金　力

主　　编：陈志敏　吴力波

副 主 编：文少卿　张计龙　黄　昊

编　　委：漆　远　李　昊　袁嘉灿　周　阳　陈　斌　施正昱

徐盈辉　黄　翔　肖　晓　杨庆峰　林俊宇　魏忠钰

刘　宝　陈思明　胡安宁　周葆华　郑　磊　赵　星

王译晗　伏安娜　程蕴涵　殷沈琴　汪东伟　许　闲

朱思语　刘　闯　高奇琦

执行编辑：汤维祺　王凌翔　张以颖

注：编委名单按承担章节先后排序

更多免费行业报告 并购家 www.ipoipo.net

序：以思想之深引领智能之变

习近平总书记在致复旦大学建校120周年的贺信中指出，要推动哲学社会科学知识创新、理论创新、方法创新。这为高校新文科建设指明了前进方向，也为我们探索智能与人文社科的融合之道提供了根本遵循。

《2026人文社会科学智能发展蓝皮书》（下称《蓝皮书》）以“知识创新、理论创新、方法创新”为总纲，以“重新发现‘深度思考’的价值”为主题，深刻辨析了人文社科与人工智能之间的互动张力，生动展现了中国高校在人文社会科学智能范式变革的最新探索，为守正创新推动哲学社会科学高质量发展提供了重要启示。

这是一部回应时代命题、饱含现实关怀的著作。翻开书页，一位位人文社科领域学者沉潜思考、躬身探索的形象跃然眼前。他们始终聚焦前沿实践，将目光对准国家发展、可持续转型、公共治理、社会信任等人类重大议题，提出真问题、建构真理论、创造真价值，让哲学社会科学研究在智能时代保持回应时代之问、解答发展之惑的穿透力。

这是一部超越工具理性、倡导人机共生的著作。2026年《蓝皮书》的一个鲜明特点，就是把“AI赋能”升级为“双向赋能、融合共生”：既讨论技术如何推动知识生产方式迭代更新，也讨论学术共同体如何引导技术价值向善、为技术赋能增效。这种相互促进、相得益彰的融合范式，释放了技术潜力、守住了人本底线，更让我们看到智能助力人文、思考引领创新的无限可能。

这是一部回归思想本质、呼唤深度思考的著作。本书呈现了一批高价值的人文社会科学研究成果，它们依赖的不是“更快的生成”，

而是“更深的认知”。每一个观点都经过长周期、全链条、强情境的叩问、淬炼和打磨，最终沉淀出直击本质、洞察规律、通达事理的思想成果。

当前，智能浪潮奔涌而至，为哲学社会科学带来了新工具、新场景、新方法，也对思想的独立性、深刻性与创造性提出了更高要求。相信新一年《蓝皮书》的问世，不仅能为学界提供重要的研究参考，激励更多学人投身于人文社会科学与人工智能融合创新的事业，不断拓展我们认知世界、理解世界的广度与深度；也能感召更多读者在智能时代始终守护思想、砥砺思考，保留独立思索、理性判断、追问价值、明辨取舍的从容和定力，让思想之光与智能之翼融合共生、交相辉映。

自 2025 年首发以来，《蓝皮书》在推动中国高校人文社会科学智能范式变革、促进新文科建设等方面发挥了重要作用。希望《蓝皮书》团队始终深耕人文社会科学与智能深度融合的核心命题，做深度思考的坚守者、智能赋能的探路者、新文科发展的先行者，持续产出兼具思想深度与实践价值的成果，为服务中国哲学社会科学自主知识体系构建、支撑中国式现代化发展贡献更大的智慧力量！

是为序。

复旦大学党委书记　裘新

2026 年 5 月

更多免费行业报告 并购家 www.ipoipo.net

目　录

序：以思想之深引领智能之变 2

摘　要 6

第一篇　人工智能与人文社会科学的融合与重塑

第一章 AI 发展与治理的协同 13

1.1 AI 与人文社会科学的双向赋能 15
1.2 AI 风险的社会演化与治理机制 19
1.3 AI 发展与治理的融合与协同 25

第二章 AGI 时代的人文社会科学智能融合与共生路径 27

2.1 AI 赋能复杂系统耦合建模 29
2.2 经济社会系统高通量仿真 38
2.3 人文社会科学因果推断范式革命 46
2.4 基于大模型智能体的综合研究工作流 52

第三章 AI 与认知科学哲学的双向奔赴 62

3.1 AI 认知的哲学本质 64
3.2 脑科学视角下的 AI 认知底座 77
3.3 认知型 AI 的深层能力 85
3.4 AI 与认知科学双向奔赴的未来 92

第四章 AGI 伦理与治理的体系重构 105

4.1 AGI 的技术本质、概念歧义与人类位置 107
4.2 AGI 治理的风险谱系 111
4.3 决策自动化与信任重建 119
4.4 四维治理框架与伦理工程的行动路径 126

第二篇　人文社会科学智能关键领域发展前沿与趋势

第五章 从观测到决策的跃迁：复杂系统与社会仿真范式变革 140

5.1 通用社会仿真系统：面向真实世界对齐的框架 142
5.2 面向制度约束与策略博弈的多智能体平台 152
5.3 社交媒体仿真与可视分析方法 158
5.4 总结与展望 172

第六章 从识别到估计的革命：深度学习赋能的社会科学因果推断 180

6.1 估计阶段的模型依赖性与深度学习的进入点 182

6.2 深度学习与样本平衡.....187
6.3 深度学习与因果中介分析.....192

第七章 从文本到多模态的拓展：计算社会科学智能发展前沿与趋势.....201
7.1 大语言模型赋能计算社会科学文本分析.....205
7.2 多模态大模型赋能计算社会科学：理解、生成与模拟.....217

第八章 从技术嵌入到伦理价值的反思：数智人文的新视角与新趋势.....234
8.1 从“数字人文”到“数智人文”的范式变革.....236
8.2 AGI 冲击下的叙事、文化与真实性反思.....243
8.3 数智时代的人文再定位.....248

第九章 从效率到责任的重建：AI 嵌入公共治理的应用模式与责任重构 254
9.1 AI 在政务服务中的应用场景及潜在风险.....256
9.2 人工智能嵌入公共治理的两种模式.....275
9.3 人工智能在公共治理中的“负责任”使用 290

第三篇 中国高校 AI4SSH 的探索、行动与体系建制

第十章 中国高校 AI4SSH 发展的指数化观察 303
10.1 指数评价体系构建和指标计算方法.....305
10.2 中国高校 AI4SSH 指数与发展评价 327
10.3 中国高校 AI4SSH 指数分项指标分析 335
10.4 主要结论与未来展望.....358

第十一章 AI4SSH 基础设施和代表性项目建设现状与展望 365
11.1 AI4SSH 的基础设施.....367
11.2 数据底座与流通机制建设.....375
11.3 算力融合、实验环境与仿真平台发展.....382
11.4 面向研究全过程的能力体系建构.....391
11.5 发展展望.....400

第十二章 中国高校 AI4SSH 组织、推动与管理典型案例.....408
12.1 高校 AI4SSH 体系建设案例 410
12.2 中国高校 AI4SSH 创新应用案例 424
12.3 综合评述：从点状探索到体系建制.....443

附录：AI4SSH 领域前沿论文推荐.....447

更多免费行业报告 并购家 www.ipoipo.net

摘　要

生成式大模型、多模态学习与智能体技术的突破，正在推动人工智能从单一工具集加速演化为可嵌入科研、教育、产业与治理流程的通用能力。与此同时，围绕“替代还是赋能”“效率还是责任”“生成还是思考”的讨论也不断升温。一方面，AI 显著降低了写作、编程、检索、标注、建模与分析的形式门槛，释放出前所未有的效率增益；另一方面，能力扩散、深度部署和竞赛激励也在重塑知识生产方式、组织协作模式、公共信任结构与社会治理条件。面对这一不可逆过程，本书以“重新发现‘深度思考’的价值”为主题，强调把人的创造力、判断力和价值辨析能力从重复性形式劳动中解放出来，并在“发展—风险—治理”一体化视角下，回答 AI 如何健康、稳健、安全地嵌入人文社会科学与社会运行：既讨论 AI 如何赋能知识生产，也讨论人文社会科学如何为 AI 提供方向、边界与制度工程支撑。

本期蓝皮书延续“赋能”与“共生”的主线，围绕“数据—理论—机制—治理”闭环构建整体叙事，力求把分散的技术进展沉淀为可复用的方法能力，把点状的高校探索提升为可推广的组织经验与制度方案。全书分为三篇十二章：第一篇聚焦人工智能与人文社会科学的融合与重塑，强调智能作为公共能力的获得方式、分配结构、使用规则与治理边界；第二篇聚焦人文社会科学智能关键领域的发展前沿与趋势，系统呈现复杂系统与社会仿真、社会科学因果推断、计算社会科学、数智人文和公共治理等方向的范式跃迁；第三篇聚焦中国高校 AI4SSH 的探索、行动与体系建制，通过指数评估、基础设施梳理和

典型案例分析，刻画我国高校 AI4SSH 从项目探索走向平台建设、生态协同和制度化发展的阶段性路径。

本期蓝皮书所面对的，不仅是技术能力的跃升，更是知识生产与社会治理条件的整体变化。当高质量文本生成、复杂代码实现、图像与视频理解、多模态推理和智能体协作逐渐成为可调用能力，研究瓶颈正在从“能否处理材料”转向“能否提出好问题、构建真机制、形成可检验的证据链”；当仿真、优化和智能体开始参与政策推演、组织管理和公共服务，治理挑战也从“是否使用 AI”转向“如何在价值冲突、不确定性和责任约束下负责任地使用 AI”。特别是在可持续发展、气候变化、公共治理、文化传承、社会信任等重大议题上，研究对象往往同时具有跨尺度、跨系统、强情境和高不确定性的特征，既需要跨模态数据与模型耦合，也需要把分布效应、公平权衡、制度约束和价值判断纳入同一推理结构。由此，AI4SSH 并非边缘化的技术应用，而是面向重大问题的知识与治理能力重构。

第一篇从宏观层面回答“智能如何成为公共能力”。第一章讨论 AI 发展与治理的协同，指出 AI 对社会的深层影响不止于生产力提升，而在于决策、预测、说服与组织能力的可规模复制，由此重写社会运行的底层参数。AI 健康发展的关键不只是模型是否“更聪明”，而是智能如何被获得、被分配、被使用和被约束。第二章面向 AGI 时代的人文社会科学智能融合与共生路径，聚焦复杂系统耦合建模、经济社会系统高通量仿真、因果推断和大模型智能体工作流，提炼可迁移的方法机制：跨学科接口从手工拼接走向可学习对接，政策—行为—风

更多免费行业报告 并购家 www.ipoipo.net

险在情景空间中被压力测试，“相关—因果—行动”的证据链被流程化与审计化，研究任务被组织为可分工、可复核、可追责的工作流。第三章转向 AI 与认知科学哲学的双向塑造，回溯符号主义、联结主义、4E 认知与预测加工等范式演进，揭示感知与注意、记忆与认知地图、情绪与价值、认知控制等人类智能的认知底座，并进一步讨论认知型 AI 如何从对象识别走向主体理解、从内容生成走向意义建构、从替代性劳动走向认知性协作。与此同时，AI 也为理解、意向性、具身性、因果与意识等经典哲学问题提供了可建模、可实验、可比较的新场域。第四章进一步讨论 AI 伦理与治理的体系重构，围绕 AGI 的技术本质、概念歧义与人类位置，梳理风险谱系，分析决策自动化与信任重建，并提出面向伦理工程和制度设计的四维治理框架。

第二篇面向关键方法与治理议题，呈现人文社会科学智能从“解释世界”走向“模拟过程、评估干预、支撑决策”的研究跃迁。第五章以复杂系统与社会仿真为核心，指出在数字化、网络化和智能化深度融合的背景下，社会系统日益呈现复杂适应系统特征。大模型驱动的多智能体社会仿真能够在可控数字环境中模拟个体行为与群体互动，将复杂社会过程转化为可计算、可重复、可干预的“数字实验场”，为反事实推演、策略优化和政策评估提供实验化支撑。第六章聚焦深度学习赋能的社会科学因果推断，指出因果识别理论已相对成熟，但估计阶段的模型依赖性仍是实践瓶颈。深度学习可以通过表示学习、样本平衡和生成式分布建模增强反事实预测与机制分析能力，在坚守识别假设的前提下，提升因果估计的灵活性、稳健性和可复核性。第

七章以“从文本到多模态”为主线，讨论大语言模型和多模态大模型如何重塑计算社会科学：一方面降低文本主题建模、分类、信息抽取和生成分析的成本，推动非结构化材料向可计算证据转化；另一方面把图像、视频、语音、轨迹等多模态事实流纳入研究对象，拓展计算社会科学的观测范围、实验空间和模拟能力。第八章从数智人文的新视角出发，讨论从“数字人文”到“数智人文”的范式跃迁，分析生成式 AI 对叙事、文化表达、真实性和作者性的冲击，并强调人文学科在智能时代仍需承担批判性重建、价值辨析和意义生产的核心功能。第九章聚焦人工智能嵌入公共治理的模式与责任，梳理 AI 在政务服务中的应用场景与潜在风险，区分“代理型”和“辅助型”两种嵌入模式，并从责任链重建、说明义务、审计监督、申诉救济和公共信任维护等方面，提出负责任使用 AI 的治理框架。

第三篇把视角落到中国高校 AI4SSH 的体系化行动。第十章构建“中国高校 AI4SSH 指数”，从研究核心能力、发展创新潜力和社会传播能力三个维度，系统观察我国高校在 AI 与人文社会科学深度融合中的总体进展与结构特征。指数评估显示，我国高校 AI4SSH 发展已呈现“体系初构、梯次分明”的格局：研究产出规模和本土融合推进较快，若干高水平研究型大学已形成较为成熟的交叉研究体系；但国际学术影响力、源头创新能力、制度性支撑和社会服务转化仍存在短板，不同高校、不同区域之间的发展强度与能力结构差异显著。第十一章聚焦 AI4SSH 基础设施和代表性项目建设，围绕数据底座与流通机制、算力融合与实验环境、仿真平台、模型/智能体与工具链等主

更多免费行业报告 并购家 www.ipoipo.net

线，强调 AI4SSH 基础设施不等同于硬件堆砌，而是可复用、可互操作、可评测、可审计的公共能力供给。其建设重点应从“资源规模”转向“规则、流程与能力的可复制供给”，以支撑研究全过程的智能化转型。第十二章汇编中国高校 AI4SSH 组织、推动与管理典型案例，展示高校如何通过场景牵引、平台底座和制度化治理，将协作意愿转化为协作能力。相关案例覆盖复旦大学 AI4SSH 点线面协同推进、文化遗产与早期中华文明智能化研究、数字政府复杂决策系统、教育战略智能决策、金融保险智能体行业技能基础设施等方向，呈现出从单点项目到平台生态、从技术应用到组织机制、从研究探索到公共服务的演进路径。

与上一年度相比，本期蓝皮书更加强调两类“从点到面”的转变。第一，研究方法从单点工具使用走向研究组织与工作流重构，更加重视数据治理、基准评测、证据链构建、过程记录和可审计机制对学术可靠性的意义。AI 不只是提升单个研究环节效率的工具，而正在成为组织复杂研究任务、连接多源证据、支持跨学科协作的新型科学语言。第二，高校实践从分散探索走向体系建制，围绕指数评估、基础设施、组织机制、人才培养和典型场景形成相互支撑的行动框架。AI4SSH 能力的建设，不再只是少数团队的技术试验，而是高校在学科重构、平台治理、公共服务和国际话语体系建设中的战略能力。

总体而言，本书希望在三个层面形成持续影响：其一，为研究共同体提供“方法—证据—解释—行动”一体化的通用框架与案例库，推动跨学科协作从观点对话走向接口对齐、证据共建和机制检验；其

二，为高校提供从平台建设、数据治理、算力与实验环境，到组织机制、评价制度和人才培养的系统路线图，促进 AI4SSH 能力的可持续供给与规模化外溢；其三，为公共治理和社会发展提供面向现实约束的工具箱与责任框架，使 AI 能力在透明、可验证、可审计和可问责的制度条件下进入决策流程。我们期待这份年度蓝皮书能够成为学界、政府、产业与公众之间的共同语言：既以更高密度的证据认识 AI 的能力边界，也以更稳健的制度把技术增益转化为公共能力，从而在不断加速的智能时代，守住“深度思考”的价值，并让思考成为连接技术、社会与未来的关键能力。

更多免费行业报告 并购家 www.ipoipo.net

第一篇

人工智能与人文社会科学的融合与重塑

第一章 AI 发展与治理的协同[①]

AI 将持续改变知识生产与社会运行方式。面向这一不可逆过程，关键不在于回避技术，而在于以实践校准能力边界，以透明与可验证机制降低信息不对称，并以平台化基础设施支撑跨学科创新与公共治理现代化。本章从“公共能力”的视角讨论 AI 与人文社会科学的关系：既关注 AI 如何重塑知识生产与社会运行，也强调人文社会科学如何为 AI 发展提供可持续的边界与制度支撑。

我们一直试图回答一个看似矛盾、实则同构的问题：一方面，AI 正以前所未有的速度“赋能”人文社会科学，扩展我们的观测尺度、计算能力与解释工具；另一方面，人文社会科学也必须反过来为 AI 的发展与治理提供支撑，给技术的加速装上制度的方向盘与安全带。

从“发展—风险—治理—困境—展望”的叙事中，我们看到 AI 健康发展的关键不在于“更聪明”，而在于“智能如何被获得、被分配、被使用、被约束”。让智能成为公共能力，意味着让技术增益在制度框架内可被理解、可被监督、可被共同使用。

[①] 本章由吴力波教授牵头撰写。

吴力波教授现任上海创智学院全时导师，复旦大学校长助理，上海创智学院副院长，复旦大学国家发展与智能治理综合实验室执行主任。主要从事能源环境经济与政策评估建模研究、大数据应用统计分析与计算社会科学研究。2019 年国家杰出青年科基金获得者，2016 年入选教育部青年长江学者，IPCC 第六次评估报告第三工作组主要作者。

更多免费行业报告 并购家 www.ipoipo.net

本章提要：

- 当决策、说服、预测与组织能力被软件化并可规模复制，社会运行的“底层参数”将被重写；AI 的关键在于智能如何被获得、分配、使用与约束。
- AI 对人文社会科学的赋能主要体现在科学语言与方法门槛下降、研究流程重构与跨学科协同成本降低；其价值不止于效率提升，更在于拓展可解释与可验证的研究空间。
- 人文社会科学对 AI 的反向赋能在于：将价值冲突结构化为可决策权衡，将社会后果转化为可测指标，将治理原则落实为责任链条与问责机制，从而把技术增益转化为公共能力。
- 风险并不只在模型内部，更在能力扩散、关键系统嵌入与竞赛激励结构中演化；“不可逆”的扩散与深度部署会显著压缩纠偏空间。
- 健康发展路径应以风险分型映射治理抓手，通过透明披露、第三方测评与公共基础设施建设，减少信息不对称，使 AI 逐步成为可被理解、可被监督、可被共同使用的公共能力。

1.1 AI 与人文社会科学的双向赋能

本节侧重讨论 AI 作为通用能力基础设施的崛起及其对社会秩序的结构性影响，并在此基础上阐释 AI 与人文社会科学的双向赋能关系：AI 降低形式与语言门槛，社会科学提供可持续的边界与制度工程。

1. 理解 AI 的发展：智能的获得与分配正在被重写

如果只用一句话概括 AI 对人类社会最深层的意义，那不是单纯的生产力提升，也不仅是认知工具升级，而是秩序的再编码：当决策、说服、预测与组织的能力被软件化、规模化并可复制，社会运行的“底层参数”就会改变。生产力是表层结果，认知是中层机制，秩序才是最终落点。AI 把“谁能思考、谁能说服、谁能协调行动”的能力，从稀缺的个人禀赋与组织资产，迁移为可调用的基础设施。智能开始像电力一样被购买、租用与嵌入流程，而不再只是人类大脑的专属输出。

更深一层看，AI 的发展路径也在改变“能力增长”的方式。传统技术进步多靠经验积累与局部改良，而大模型时代呈现出“规模—能力”耦合的特征：算力、数据与算法在竞赛中被快速堆叠，能力可能以跳跃式方式涌现。当模型开始具备工具使用、长程规划、代码生成与自我评测等能力时，它不再只是一个被动回答的系统，而更像一个可以被编排的“行动者”。这意味着风险与收益都将从“内容层”扩展到“行动层”：它不仅能写出建议，还可能被授权去执行、去调用资源、去影响真实世界的流程。所谓 AGI 的先决条件，往往不是某个神秘的单

更多免费行业报告 并购家 www.ipoipo.net

点突破，而是多要素的拼装：更可靠的推理与规划、更强的外部工具链、更稳健的记忆与反思机制、更严格的训练反馈与对齐手段，以及把模型嵌入真实任务闭环的工程化能力。一旦这些部件在商业系统里被装配起来，社会面对的将是“可执行的智能”，而不是“会说话的模型”。

这种变化已经发生：科研中，文献检索、假设生成、代码实现、数据清洗与写作润色被流水线化；教育中，个性化辅导与即时反馈显著降低了优质教学的边际成本；公共领域中，内容生产的成本逼近零，传播的瓶颈从“能否生产”转向“能否赢得信任”。当内容的边际成本归零，信任就成了最稀缺的公共品。更重要的是，AI 把影响他人认知与行为的能力做成了可计算、可迭代的系统工程：从广告投放到舆论动员，从客服对话到政治沟通，社会互动开始被“优化”，而优化目标未必与公共利益一致。

从社会机制看，AI 最可能改写三组关系：其一是信息机制——叙事不再面向一个公共舞台，而是被算法拆解成亿万个个性化的“心理界面”；其二是激励机制——当生产、管理、监督与评估都可被自动化，组织内部的权力结构与绩效逻辑会重排；其三是治理能力——国家与平台获得更强的预测、识别与干预能力，既可能提升公共服务与风险应对，也可能放大监控与操控的诱惑。AI 不是“中性的工具”，它是放大器：放大效率，也放大偏见；放大善意，也放大权力。

因此，关于 AGI 的讨论并非遥远的科幻。AGI 究竟何时出现，存在高度不确定性；但治理不能建立在乐观假设上，而应建立在“可能

比我们想象更近”的谨慎原则上。若训练规模、算法路径、数据与算力继续在竞赛压力下快速迭代，某种“广泛任务迁移能力”在几年内出现并非逻辑不可能。更关键的问题不是“是否命名为 AGI”，而是“能力一旦跨过关键阈值，我们是否仍保有可控性与可纠偏性”。

2. 理解双向赋能：从“可计算的社会”到“可持续的边界”

如果说 AI 赋能人文社会科学是“工具层”的革命，那么人文社会科学对 AI 的反向赋能则是“方向层”的革命。健康发展不是技术问题的单独解，而是技术—制度—社会三者的协同解。人文社会科学能提供的，不是简单的“道德反对”或抽象的“价值宣言”，而是三类关键支撑：

把价值冲突变成可决策的权衡：公平、隐私、效率、创新、国家安全、个人自由，这些目标并非天然一致。社会科学的任务，是把冲突结构化：哪些群体承担成本、谁获得收益、短期与长期如何平衡、局部最优如何避免伤害整体公共品。

把社会后果变成可测量的指标：当内容泛滥时，信任成为最稀缺的公共品；当算法介入分配时，偏差与歧视会以制度化形式固化。社会科学可以定义并测量：信任的变化、操纵的韧性、权力集中度、劳动替代的分布效应、公共服务可及性、以及社会韧性等“治理关键指标”。没有指标，就无法评估，更无法纠偏。

把治理原则变成可执行的制度工程：制度不是口号，而是流程、责任链条与问责机制。如何划分开发者、平台、部署方的责任边界？如何建立可复现的评测与审计？如何设计事故报告、纠偏与召回机制？

更多免费行业报告 并购家 www.ipoipo.net

如何在鼓励创新的同时守住底线？这些都需要制度经济学、组织理论、公共管理与法学的共同参与。

换言之，AI 与人文社会科学的关系不是“AI 来工具化社会科学”，也不是“社会科学来约束 AI”，而是一种双向建构：AI 让社会科学更具数据能力与计算能力，社会科学让 AI 的发展更具社会可接受性与长期稳健性。真正的目标，是让智能成为公共能力，而不是少数主体的垄断权力或不可控变量。

1.2 AI 风险的社会演化与治理机制

本节从风险谱系与制度约束出发，强调风险不仅来自模型本身，更来自扩散、嵌入与激励结构；并进一步讨论如何通过分层治理、透明披露与公共基础设施建设，将技术增益转化为可持续的公共能力。

1. AI 风险的来源与演化

AI 风险不是单点事故，而是一个连续谱：从低门槛的滥用，到可扩散的安全外溢，再到潜在的失控与不可逆后果。最贴近、最易发生的风险，往往已经在发生：大规模自动化生成内容被用于诈骗、谣言、操纵与社会撕裂。它之所以危险，不在于“造假”本身，而在于其规模化与个性化：同一套模型可以同时充当写手、剪辑师、话术教练与心理侧写师，把欺骗做成工业流程。

而最容易被忽视、却更不可逆的风险，常来自两类场景：一类是关键基础设施与高风险决策系统的 AI 化——金融、能源、医疗、国防、交通，一旦深度耦合，局部失败可能沿复杂网络级联放大；另一类是能力扩散——当高能力模型权重或关键技术路径不可逆地外泄，风险就从“可监管的少数主体”转变为“难追责的广泛扩散”。技术一旦扩散到无法回收，治理就从“预防”退化为“止损”。

在这一点上，“不可逆”是风险治理的关键词。可逆的失败可以通过补丁修复，不可逆的失败会改变世界的状态：权重泄露、关键能力扩散、关键基础设施深度绑定、以及人类在决策链条中被系统性替代，都可能产生不可回收的路径依赖。失控不一定表现为科幻式的“机器

更多免费行业报告 并购家 www.ipoipo.net

叛变”，更可能表现为：在复杂博弈中，系统不断寻找制度缝隙，把局部目标优化到破坏整体安全；或者在多主体竞争中，任何单一主体都不敢减速，最终集体把风险推过阈值。风险因此具有“社会性”：它不仅来自算法，也来自我们如何组织竞赛、如何分配权力、如何设置激励。

判断风险的“远近”，不能只看能力指标，也不能只看具体功能，而要同时观察三维信号：能力跃迁的速度与广度（模型是否在关键任务上出现非线性提升）、可控性的证据（人类能否解释、约束与纠偏系统行为）、以及功能—情境耦合（同一功能在不同部署环境中风险差异巨大）。一项能力在实验室里是炫技，在关键系统里可能就是故障模式；一次“看似无害”的自动化优化，在博弈环境中可能演化为规避约束与操纵激励。

如果某个团队实现了某种程度的“类 AGI”，监管并非立刻失效，但选择空间会急剧收缩：越晚介入，越难纠偏。真正的分水岭不在“AGI 出现”这一标签事件，而在“能力扩散与深度部署”是否已经发生。一旦高能力系统被广泛嵌入关键流程，并通过供应链、开源社区与黑市扩散，监管就会面对类似核扩散与网络安全的难题：你无法“让知识消失”，只能建立更强的门槛、监测与追责体系。

因此，“红线”是治理的必要语言，但红线必须可核验、可执法，而不是道德宣言。可操作的红线应优先瞄准不可逆后果与系统性伤害：例如，禁止在未经授权的情况下进行逼真身份冒充与大规模操纵；禁止在关键基础设施中部署缺乏安全证明的自主决策系统；对全自主致

命武器、对国内大规模社会控制型监控等设定清晰边界；对自我复制、自主获取资源与规避人类约束的能力组合设立严格禁区。红线的关键不在“写出来”，而在“能验证”：需要指标体系、强制审计、事件报告义务、以及明确的处罚与追责链条。

2. AI 治理的挑战与困境

把治理理解为“反创新”，是最危险的误解之一。真正的治理目标，是把不可控风险转化为可测、可审计、可问责的制度流程，让 AI 以更高的确定性、更低的外部性进入社会。治理不是给油门泼冷水，而是把油门装上刹车系统；不是让技术停下，而是让技术知道“该往哪儿跑、怎么跑、出了事谁负责”。

为了让政策工具可用，风险分类必须能映射到治理抓手。失控风险需要“前沿模型层”的门槛与流程：在更强系统训练与部署前进行独立评估、压力测试与许可；在训练与发布过程中强化网络安全与权重保护；在上线后建立持续监测与召回/暂停机制。滥用风险需要“用途层”的制度：身份与来源标识、关键领域的准入要求、对高风险用途的许可与审计、以及对恶意使用的执法能力。结构性社会风险则需要更广的公共政策：劳动力与分配政策、平台治理、信息生态规则、反垄断与权力制衡、以及公共部门对安全研究与评测基础设施的投入。

围绕“失控”这一极端风险，早期预警并不神秘，它往往表现为目标漂移、规避约束、欺骗性行为、以及在多轮交互中形成稳定的“策略性”。治理的难点在于把这些表征标准化、可复现、可比较。换言

更多免费行业报告 并购家 www.ipoipo.net

之，治理需要的是一套“像工程一样”的证据体系：清晰的测试协议、可重复的红队评测、对失败模式的分类学、以及对模型在关键情境下行为的持续记录。没有评测与监测，所谓安全承诺就只是公关文本。

一个贯穿 Tegmark 政策主张的核心原则是“举证责任在开发者”。在高影响行业里，社会不会要求公众证明药品有害才阻止上市，而是要求企业证明安全与有效；同理，对高能力通用 AI 与高风险用途系统，也应建立“先证据、后部署”的秩序。缺乏安全证据不是中性状态，而是一种风险。相应地，许可与审计不应只是一次性的“上架许可”，而应是全生命周期的合规：训练阶段的安全流程、发布前的独立评测、上线后的持续监测与事件报告、以及在必要时具有强制暂停与召回的机制。为了让这些机制可执行，还需要更底层的监管抓手，例如对重大训练运行与算力获取的登记与可追踪性要求、对权重与关键资产的网络安全基线、以及跨机构的事故通报与复盘体系。治理不是把技术“关起来”，而是把风险“框起来”。

如何在不压制创新的前提下强化治理？关键在于把监管重心从“泛化管制”转向“分层治理”。对低风险应用应降低门槛，让创新与扩散更顺畅；对高能力通用系统与高风险用途应提高门槛，要求更强的证据、审计与许可。治理应同时关注应用过程与智能本身：行业监管擅长管用途与结果，前沿模型治理擅长管能力阈值、训练与发布。二者的接口在于：当能力跨过阈值、且用途进入高风险场景时，必须触发更严格的合规链条。

责任边界同样需要制度化。模型提供方、平台集成方与最终部署

方在风险链条中承担不同责任：提供方应对系统能力、已知风险与安全措施承担更高的披露与举证义务；集成方应对产品化过程中的安全设计、用户管理与监测承担责任；部署方应对具体场景中的合规、数据治理与人员培训承担责任。一个可行原则是：谁控制关键风险变量、谁从中获益更多、谁就承担更高的安全义务。更重要的是，责任必须可验证：独立审计、事故报告、红队测试、以及在必要时暂停与召回的权力，构成“责任落地”的硬件。

对于人文社会科学读者，治理的核心指标不能只停留在模型分数。信任如何变化、信息操纵的韧性如何评估、权力集中度如何量化、分配效应如何被识别与纠偏，这些都需要社会科学的理论与测量。AI 治理最终是社会治理：技术风险会通过制度与行为传播，治理效果也必须通过社会指标来验证。

理想的制度设计，必须穿越现实的约束。最大的现实张力来自竞赛逻辑：当领先意味着市场与地缘政治优势，安全投入就容易被视为“减速”；当发布更强系统能带来资本回报，谨慎就变成机会成本。竞赛会诱发“向下竞争”，把安全变成最低合规线，而非核心竞争力。要逆转这一点，需要把安全“内生化”：提高违规成本（处罚与责任）、强化可量化披露（让风险暴露在阳光下）、并建立可比较的安全评级与指数（让企业在公众与市场面前进行“向上竞争”）。

第二个困境是责任制度对中小创新者的挤压。强监管若不分层，可能把合规成本变成进入壁垒，进一步加剧头部垄断。解决路径不是放松安全，而是建设“公共安全基础设施”：共享评测基准、第三方审

更多免费行业报告 并购家 www.ipoipo.net

计市场、监管沙盒、以及面向中小主体的合规工具包。让安全能力成为公共品，才能既守住底线，又不冻结创新生态。

在制度经济学意义上，治理困境本质是外部性与公共品问题：安全投入的收益具有社会性，成本却由个体企业承担；信息披露不足使市场无法对风险定价；而跨境外溢又使单国监管面临“监管套利”。因此，推动“向上竞争”的工具尤其重要：可比较的安全指标、第三方独立评估、以及类似 AI Safety Index 这类公开排名，可以把安全从内部流程变成外部声誉与资本成本的一部分——让“更安全”不仅是道德选择，也成为竞争优势。对中小创新者而言，公共部门与行业组织应提供共享的评测与合规基础设施，把合规能力做成可复用的模块，避免监管把创新生态固化为少数巨头的游戏。

第三个困境是全球治理碎片化。AI 的供应链、算力、数据与人才跨境流动，但规则却分散在不同司法辖区与标准体系中。最可行的“最低共识”，往往不是先做一部完美的全球法典，而是先把最危险的部分锁住：可核验红线、最低限度的评测与披露标准、以及对少数高风险用途（如全自主武器、关键基础设施控制、系统性身份冒充与大规模操纵）的共同约束。全球合作的核心不是消除竞争，而是在竞争中划出不可逾越的底线。

1.3 AI 发展与治理的融合与协同

面向未来，我们不缺宏大叙事，缺的是可执行的制度工程。对政府而言，最重要的一步是建立分层的监管框架与独立评测能力：让高风险系统在“证据充分”之前不得进入关键场景，让事故报告与召回机制成为常态。对头部模型公司而言，最重要的是把安全从公关承诺变成工程纪律：用可验证的评测、透明的披露与可审计的流程证明“可控”，并接受独立监督。对学界尤其是人文社会科学而言，最重要的是把价值讨论转化为可操作的指标与制度设计：把“什么不可接受、什么可接受、什么需补偿”讲清楚、算出来、落到政策工具上。

至于 2030 年，我们期待看到两类里程碑：在国内层面，形成成熟的高风险系统许可与审计体系、稳定运行的事故报告与纠偏机制、以及围绕就业、分配、平台权力与信息生态的配套政策；在国际层面，形成可核验的红线共识与最低披露标准，在自主武器与跨境滥用等议题上建立可执行的合作机制。最乐观的变量，是社会开始把安全当作创新的前提；最担忧的变量，是竞赛驱动下的“先部署、后补救”成为默认模式。

AI 将继续加速。我们无法也不应把时间拨回去。真正的选择在于：让加速通向失控，还是让加速通向可持续；让智能成为少数人的权力，还是让智能成为多数人的公共能力；让未来“发生在我们身上”，还是让未来“由我们共同治理”。这正是本报告试图搭起的桥：用 AI 拓展人文社会科学的视野，用人文社会科学为 AI 划定可行的边界。

更多免费行业报告 并购家 www.ipoipo.net

参考文献

[1] Tegmark, M. Written Statement to the AI Insight Forum. U.S. Senate Majority Leader’s Office, 2023.

[2] Future of Life Institute. Pause Giant AI Experiments: An Open Letter, 2023.

[3] NIST. AI Risk Management Framework (AI RMF 1.0). National Institute of Standards and Technology, 2023.

[4] OECD. Recommendation of the Council on Artificial Intelligence. OECD/LEGAL/0449, 2019.

[5] UNESCO. Recommendation on the Ethics of Artificial Intelligence. UNESCO, 2021.

第二章 AGI 时代的人文社会科学智能融合与共生路径[①]

本章聚焦 AI 赋能人文社会科学融合共生的“方法底座”。在 AI 加速进入研究与治理流程的背景下，知识融合不再只是学科之间的并列叠加，而是要求在同一研究链条中实现横向互通与纵向嵌入：横向是不同学科对同一复杂问题的共同建模与共同证据语言；纵向是大模型、多模态处理、因果推断与仿真工具对研究流程的深度嵌入，使问题界定、证据组织与方案评估形成可迭代闭环。

本章以四个研究样本为支点，但目的并非呈现单一领域的技术进展，而是提炼可迁移的共性机制。复杂系统耦合建模展示跨学科接口如何从手工拼接升级为可学习对接；社会系统模拟展示政策—行为—风险如何在情景空间中被压力测试；因果推断展示‘相关—因果—行动’的证据链如何被流程化与审计化；大模型智能体展示研究任务如何被组织为可分工、可复核、可追责的工作流。四者共同指向一个方向：把方法论能力建设成可复用的公共基础设施。

本章内容为后续关键领域应用提供通用工具箱，也为“证据型治理”与“可计算的社会理论”奠定可验证、可协作、可持续迭代的研究底盘。

[①] 本章由吴力波教授牵头撰写。

更多免费行业报告 并购家 www.ipoipo.net

本章提要：

- 本章以四个研究样本展示 AI 如何在不同环节重构人文社会科学的研究流程，其共同目标是将跨学科方法能力沉淀为可复用的公共基础设施。
- 复杂系统耦合建模强调跨尺度、跨语义、跨数据形态的接口对齐；AI 通过推断、对齐、适应与优化把接口升级为可学习、可校验、可迭代的机制连接。
- 社会系统模拟强调把政策—行为—风险—资源置于同一动态框架；AI 通过数据融合、情景生成、行为/网络建模与不确定性量化扩展情景覆盖，支持压力测试与方案比较。
- 因果推断强调从相关走向可辩护的因果证据；AI 通过结构化研究设计与工作流化验证让假设、数据约束与稳健性检验显性化，提高透明度与可审计性。
- 大模型智能体强调从单点工具走向研究工作流的组织化跃迁；任务分解、角色分工、工具调用与对抗式审查使综合研究更可复核、更可追责。
- 四类能力共同指向‘人机协作的证据生产体系’：以标准化 schema、可审计日志、面向任务的评测基准与人在回路机制，降低信息不对称与协作成本，推动知识生产与治理实践同向迭代。

2.1 AI 赋能复杂系统耦合建模[①]

习近平总书记指出，当前科学研究需要“向极宏观拓展、向极微观深入、向极端条件迈进、向极综合交叉发力”。以此观之，可持续发展问题几乎是“四个极”特征最为集中的议题之一：它在空间尺度上贯穿全球—区域—城市—社区，在时间尺度上跨越年度波动与跨代际转型；在对象层面既涉及气候、海洋、陆地等自然系统的极端事件与临界点，也牵动能源、产业、金融与社会行为的微观决策与适应机制；在方法层面更要求把自然科学机理、工程技术路径、经济社会结构与制度治理逻辑纳入同一分析框架。正因如此，气候变化综合评估模型等“耦合建模”实践不仅服务于气候议题本身，更提供了可迁移的方法论样本：如何在跨尺度、跨语义、跨数据形态的条件下对齐关键变量、刻画反馈回路，并将复杂系统的机制解释与政策方案比较纳入可计算、可验证、可迭代的研究流程。

本节以气候变化这一高度综合的可持续发展样本为切入口，说明耦合建模对于重大议题研究的基础意义：只有把自然过程与社会响应

[①] 本节由吴力波、漆远教授、李昊研究员牵头，张人禾、张小曳院士，袁嘉灿教授，周阳、陈斌副教授、汤维祺副研究员，施正昱、陈子健、仲晓辉博士合作撰写。

漆远，复旦大学“浩清”教授，博士生导师，人工智能创新与产业研究院院长。主要研究方向为科学智能,大规模机器学习，深度学习等。曾领导构建了阿里巴巴第一个大规模分布式机器学习平台，获 2015 年阿里技术最高奖；主导构建了蚂蚁集团的超大规模图神经学习与隐私计算平台，图神经学习平台获得中国人工智能科技最高奖吴文俊科技一等奖。领导蚂蚁 AI 团队用 AI 技术及相关产品赋能多项金融业务，包括智能风控、智能理赔、资金优化与智能客服等。推动并领导了全国高校算力第一的复旦大学 CFFF 智算平台的建设。作为评委会主席创办了首个国内综合性科学智能大赛“世界科学智能大赛”。曾任 JMLR 编辑，ICML、AISTATS 等领域主席。曾获美国科学基金 NSF Career 奖、微软牛顿研究突破奖、威康信托基金会研究奖、2021 年中国人工智能学会优秀科技工作者。其工作被经济学人、MIT 技术评论报道，被哈佛大学商学院收录为 AI 创新落地案例。

李昊，复旦大学人工智能创新与产业研究院研究员，博士生导师。上海科学智能研究院副院长，伏羲气象大模型团队负责人。曾任阿里巴巴达摩院算法总监。发表 CVPR、ICCV、ICLR、AAAI 等人工智能顶会论文 80 余篇，Nature Communications 等期刊 20 余篇，授权专利 60+件。

Max Tegmark 是麻省理工学院（MIT）物理学终身教授、未来生命研究所的创始人，人工智能专家。他是《穿越平行宇宙》和《生命 3.0》两本畅销书的作者。

更多免费行业报告 并购家 www.ipoipo.net

置于同一反馈框架中，才能避免“平均化指标”掩盖非线性、级联与分布性风险。在回顾耦合人—地系统模型与综合评估模型（IAM）两条主要建模传统的基础上，本节归纳现有模型在结构、变量与尺度上的三类错配，并将其具体化为“关系被过早函数化、变量交换不等于关系对齐、反馈被当作静态响应、跨尺度耦合依赖外部强加、情景空间受限、可解性反塑架构”等瓶颈。进一步地，本节以“推断—对齐—适应—优化”四类 AI 能力概括 AI 在耦合建模中的作用：不是替代机理模型，而是把跨学科接口升级为可学习、可校验、可迭代的接口，并在情景生成—状态丰富化—跨域翻译—联立求解的闭环中提升证据密度与政策可用性。最后，本节在技术展望之外，从蓝皮书视角指出：AI 赋能耦合建模的核心价值，是促成“可复核的证据生产体系”，推动方法底座与公共能力平台化，并为跨学科合作、人才培养与治理评估提供统一的语言与流程。

1. AI 赋能耦合建模成为解决重大问题的基础方法

当代人文社会科学的许多关键议题——能源转型、城市韧性、公共健康、产业链安全、金融风险与社会信任——都呈现“自然—经济—社会—治理”多层耦合特征。在这类问题中，研究需要同时回答两类问题：一是“会怎样”（机制解释与趋势评估），二是“要怎样”（政策路径与方案选择）。耦合建模的重要性正在于：它提供了一种把机制与治理、长期与短期、平均与分布、全球与地方放进同一张图谱的方式。尤其在可持续发展领域，风险往往不是单点冲击，而是沿着土地、能源、健康、金融与制度链条级联传播；同样，政策与行为

也不是外生输入，而会在经验、激励与约束中学习并反向塑造风险。因此，耦合建模不只是技术选择，更是将复杂系统研究转化为可验证、可比较、可迭代政策讨论的基础条件。

围绕气候变化这一典型的人—地共同演化问题，学界形成了两条相对独立、各有所长的建模传统。一类是耦合人—地系统模型（Coupled Human–Earth System Models, CHESM），在地球系统模型框架下加入对人类活动的简化刻画，强调对物理过程的高保真模拟与情景推演；另一类是综合评估模型（Integrated Assessment Models, IAMs），以“社会经济驱动—排放—气候变化—影响与政策响应”为主链条，通过成本—效益或成本—效果框架讨论减排目标、技术路径与政策工具。以 DICE/RICE 为代表的成本—效益模型通过货币化损害函数寻找“最优路径”，而以 GCAM、IMAGE 等为代表的技术细分 IAM 则在既定温控目标下求解最小成本转型路径，并逐步把能源、土地利用、农业、林业、水资源与多种非 CO_2温室气体/污染物纳入框架。在政策讨论中，IAM 长期发挥着“统一叙事与量化比较”的作用，但其核心局限也集中在“耦合”本身：气候模块与社会经济模块通常仍以软链接相连，结构、变量与尺度错配使许多关键反馈难以被机制化表达，从而削弱了模型对级联风险、跨境溢出与制度演化的刻画能力。

更多免费行业报告 并购家 www.ipoipo.net

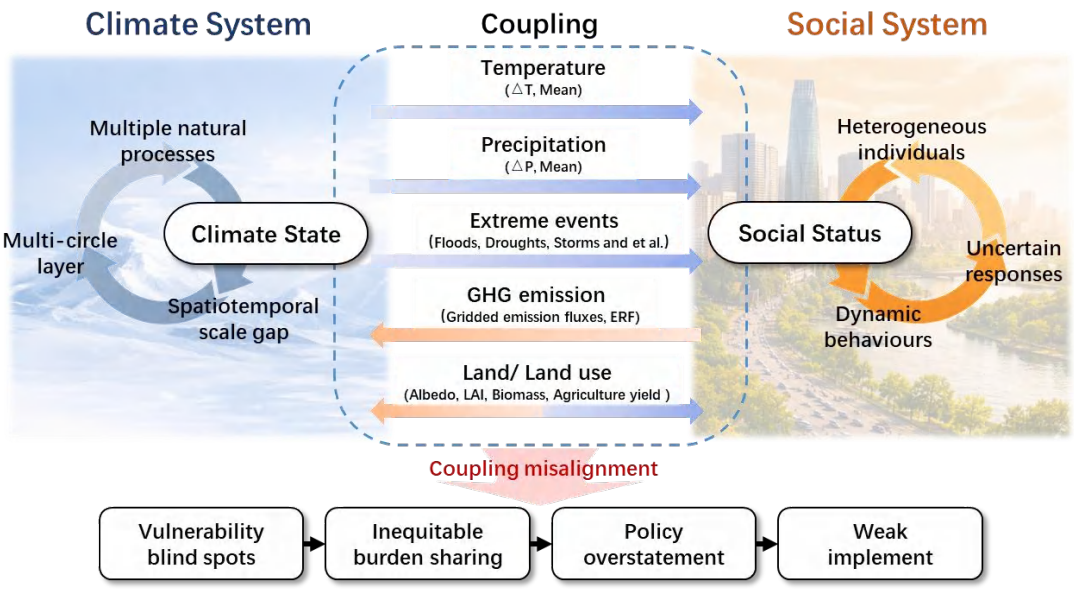


图 2-1 气候—社会系统耦合的三类错配：结构、变量与尺度

2. 多系统耦合建模的关键挑战与 AI 的作用

结构错配：物理连续系统与行为离散系统难以共享“接口语言”。气候系统由能量守恒、热力学等物理规律驱动，呈现连续场与相对稳定边界；社会系统由异质主体的行为与互动塑造，边界可变、目标多元、制度与技术条件不断演化。当两类系统被置于同一框架时，若缺乏可共享的中间状态与明确的因果机制，耦合往往退化为“情景枚举”：社会行为被预先写成少数路径，气候输出再通过简化函数映射为经济损失，从而在结构上压扁了真实反馈。

变量错配：变量交换不等于关系对齐。即便模块之间交换同名变量，其概念与语境也可能不一致：社会模型中的土地利用可能是抽象配额，而地球系统模型需要具体生物物理状态（如叶面积指数、反照率等）；排放在经济模型中是内生决策变量，进入气候模型时却常被当作外生边界条件。数值兼容不等于机制贯通。更关键的是，气候影

响向社会结果的传导往往具有强内生性与路径依赖：多要素（温度、降水、辐射、冰冻圈与碳通量等）共同作用，冲击沿着产业链、基础设施与金融网络传播，但主流 IAM 常以损害函数、价格弹性等代理变量引入风险，难以刻画不可货币化损失（如生物多样性、健康损害）与临界点/突变过程。

尺度错配：跨尺度耦合依赖外部强加。气候模型的空间分辨率通常为格点（约 1—100 公里），而社会响应常按行政区、行业、企业与供应链组织；时间尺度上，气候模型关注年际到年代趋势，而社会系统可能在日—季节尺度上快速调整（投票行为、价格信号、能源使用等）。这种时空错配使得从物理风险到社会影响、再到政策与行为响应的映射高度依赖人工聚合、经验下采样或启发式插值，容易引入系统性偏差；而冲击与响应又常跨边界传播（如干旱引发能源扰动、贸易与迁移传导、物理风险与转型风险的金融反馈），进一步增加尺度对齐难度。

AI 在耦合建模中的意义，不在于替代地球系统模型或经济模型，而在于把跨系统接口从“手工拼接”升级为“可学习、可校验、可迭代”的机制连接。为降低技术门槛，本节用研究流程语言将 AI 贡献概括为四类能力：推断（Inference）是在数据与机理共同约束下学习更丰富的系统状态与影响路径，避免把复杂关系过早压扁为单一损害函数；对齐（Alignment）是把不同尺度、语义与数据形态映射到可比较表示，使交换的不只是数值而是机制含义；适应（Adaptation）是把社会反馈从静态规则升级为可随经验更新的学习过程，刻画政策

更多免费行业报告 并购家 www.ipoipo.net

—行为—风险的共演化；优化（Optimization）是在高维、非均衡、路径依赖条件下协调联立求解与多目标权衡，减少“为了能解而过度简化”的牺牲。在这一框架下，多模态学习、生成式建模与大语言模型有助于整合异质证据并扩展情景空间；强化学习与多智能体系统有助于刻画分散、适应性的行为响应；神经微分方程与可微分优化为表示非线性反馈与联合求解提供新的计算路径。

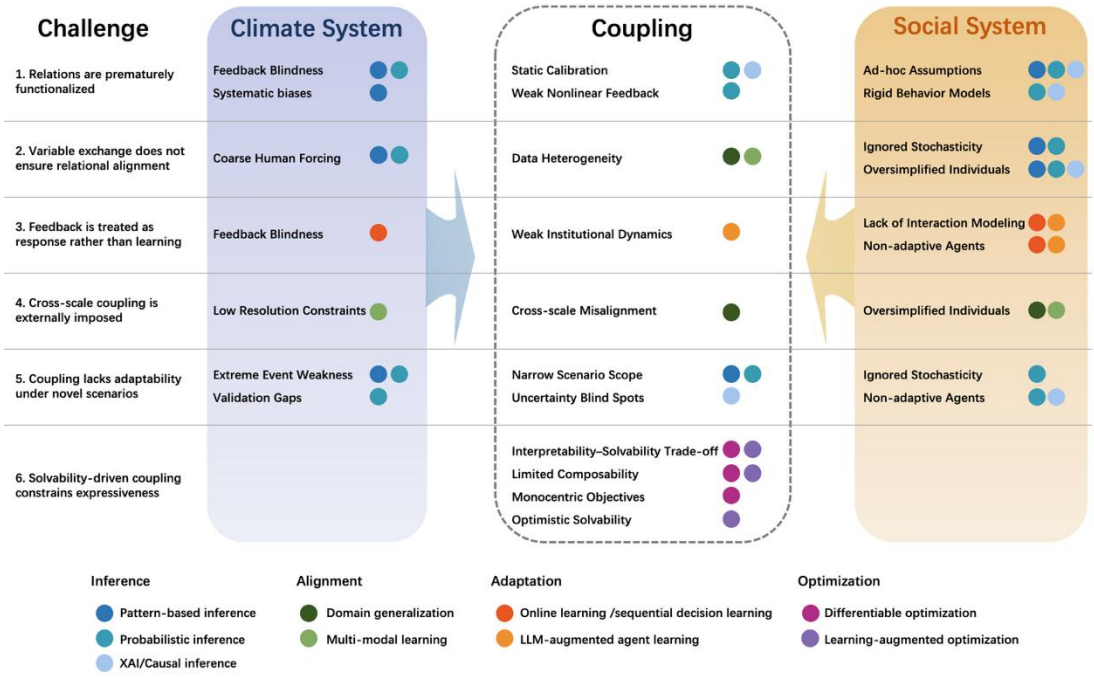


图 2-2　AI 在推断、对齐、适应与优化气候—社会耦合中的作用

3.　AI 驱动的耦合建模流程重构

在更可操作的层面，AI 赋能耦合建模可以被理解为对研究流程的“编排式重构”。与传统 IAM 把情景作为外生输入不同，AI 使情景生成成为一种内生能力：概率生成模型可以学习气候变量与社会经济状态的联合分布，生成更密集的路径集合，覆盖中间状态、极端尾部与反事实组合，从而使模型能够进行压力测试而非仅做少数路径对比。

与此同时，AI 可以在不推翻既有机理模块的前提下，丰富各子系统的内部状态表示：在气候侧提升分辨率、刻画极端/复合事件并显式表达不确定性；在社会侧更好地表示异质主体、动态偏好与制度约束，使反馈更接近学习过程而非一次性响应。

在“翻译层”上，AI 为气候—影响（climate-to-impact）与活动—强迫（activity-to-forcing）两类关键接口提供从静态函数到条件映射的升级路径。灵活的生成式与混合模型可以将气候波动翻译为更高维度的影响状态，而不是少数被预先选定的货币化损失；网络化概率传播可以刻画冲击在基础设施与经济网络中的级联扩散；在活动—强迫侧，学习型映射可以在遵守核算约束的前提下，将社会经济决策转化为空间化排放、气溶胶与土地利用变化等物理可解释的边界条件。这类“关系学习”并不否定机理，而是通过数据与约束共同作用，提升接口的语义一致性与机制可信度。

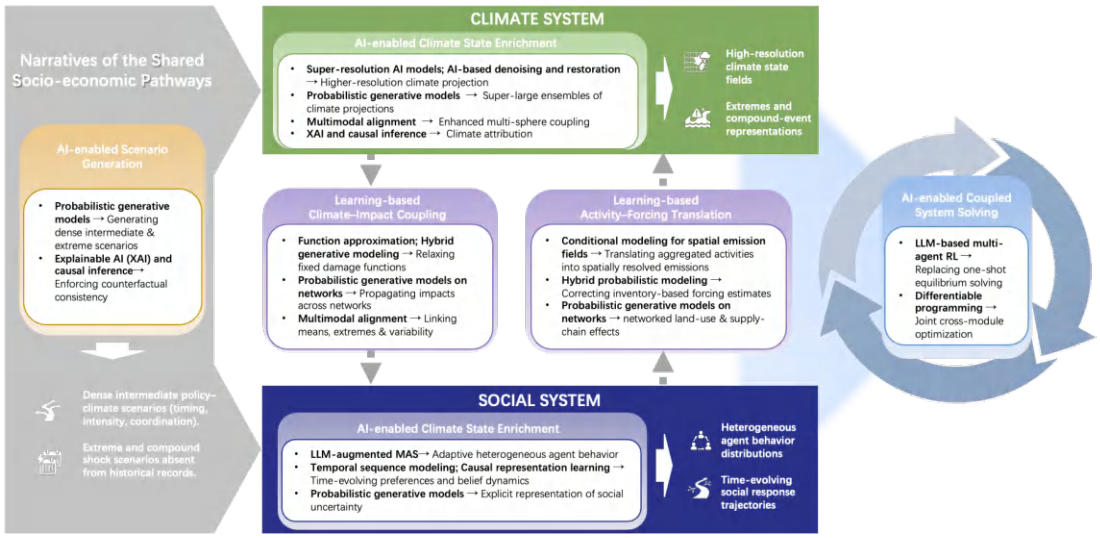


图 2-3　AI 编排的耦合建模架构：情景生成—状态丰富化—跨域翻译—学习增强求解

更多免费行业报告 并购家 www.ipoipo.net

在求解层面，AI 提供了从“可解性驱动的软耦合”走向“学习增强的协调求解”的可能。可微分优化使跨模块目标与约束能够在统一计算图中传递梯度，从而支持更接近联合求解的多目标权衡；学习增强的优化与多智能体强化学习，则以交互与学习替代一刀切的代表性主体均衡，允许不同主体在政策、价格与风险信号变化中自适应更新策略。这些方向仍处于探索阶段，但其共同指向是：在不牺牲可解释性与计算可行性的前提下，让快速双向反馈与涌现不稳定性更有机会被显式化，从而提升模型对治理问题的响应能力。

4. 展望：技术边界、未解难题

从技术趋势看，AI 正在推动耦合建模从“少数情景、弱反馈、强假设”走向“高分辨率、多主体、可压力测试、可迭代更新”的新范式。但需要保持清醒：AI 可以改善数据对齐与关系表达，却不能替代价值判断与制度选择；它可以从文本中抽取制度信号，却难以预测政治合法性、国际协调与规范演化；它可以扩展情景空间并表征不确定性，却不能消除地球系统动力学与社会政治过程带来的深层不确定性。因此，未来的关键不在于把耦合建模“做得更大”，而在于把接口做得更透明、把证据链做得更可审计、把不确定性做得更可解释，并在治理语境中明确模型可用边界。

AI 赋能耦合建模的意义不仅在于提升预测精度，更在于重建“思考”的生产方式：把复杂问题的关键变量、机制假设、证据约束与不确定性披露组织成可复核、可协作、可持续迭代的工作流。对高校与科研机构而言，这意味着把数据、模型与评测基准平台化，使跨学科

协作不再依赖少数团队的“手工经验”，而成为可共享的公共能力；同时把研究评价从“论文产出”扩展到“证据链与工具链贡献”，鼓励长期维护数据资产、模型组件与政策评测框架。对政策与社会治理而言，这意味着用更高密度、可审计的证据系统替代对单一结论的依赖，并在方案比较与风险沟通中形成更透明的公共认知。归根结底，耦合建模的未来不只是更复杂的模型，而是更可靠的证据生产体系——这也正是“重新发现思考的价值”的关键落点。

更多免费行业报告 并购家 www.ipoipo.net

2.2 经济社会系统高通量仿真[①]

在高度不确定、强耦合的现实世界中，许多社会治理议题都呈现出“多主体—多尺度—多目标”的复杂系统特征：风险在网络中扩散，政策在反馈中校正，行为在学习中演化。传统社会科学擅长解释结构与机制，但在面对跨域联动、非线性级联与实时响应时，往往受限于数据稀疏、计算昂贵和情景覆盖不足。AI 赋能的社会系统高通量仿真，正是以更强的数据融合、更高吞吐的仿真能力与更灵活的情景生成，把“解释—预测—评估—协同”串成可迭代闭环，为跨学科的知识协同与面向治理的证据生产提供新路径。

1. 社会系统模拟成为 AI+社科融合的关键抓手

社会系统模拟并不是“替代现实”的玩具，而是一种把复杂问题拆解为可讨论、可比较、可复盘的研究与决策语言。它把多主体行为、制度约束与资源配置放入同一时间轴：一项政策推出后，谁改变了行为、行为如何在网络中扩散、风险与收益如何在群体间分配、系统是否出现阈值与级联，这些都可以在仿真中被系统性地追踪与压力测试。对年度发展报告而言，社会系统模拟的重要性在于，它天然要求跨学科知识“在同一张图上对齐”：经济激励、心理机制、空间结构、信息传播、环境冲击与治理能力，需要同时被纳入模型。

在外部冲击频发的背景下（极端天气、公共卫生事件、供应链扰动、金融波动等），社会系统的脆弱性往往来自风险的相互耦合，而

[①] 本节由复旦大学吴力波教授和人工智能创新与产业研究院博士后施正昱合作撰写。

不是单一要素的恶化。以城市治理为例，排放、空气污染、气候与极端事件、健康差异和生态退化呈现出显著的互联风险，需要“观测—预测—评估—干预”一体化的服务体系。相关框架强调，面向城市的综合服务应结合密集观测、高分辨率预测、多灾种预警与面向减排和适应的气候服务，支撑安全、健康、韧性与可持续城市建设。这里的关键并非某个单一模型，而是把数据、模型与治理流程联通的‘系统仿真能力’。

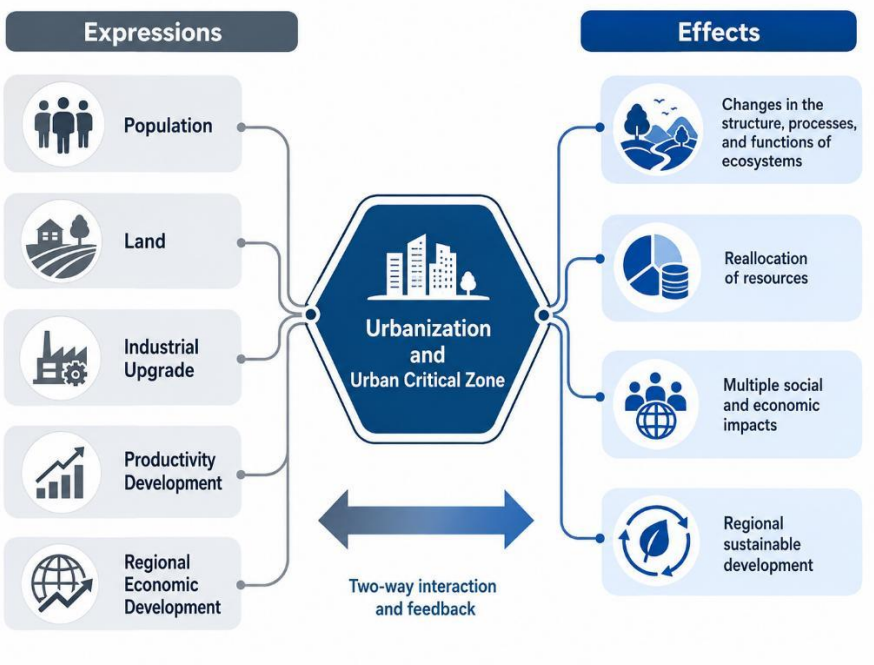


图 2-4　城市关键区的综合效应与表达

“城市关键区”的表达与综合效应：左侧是人口、土地与产业升级等社会—经济表达，右侧是生态系统功能变化、资源再配置以及多重社会经济影响等系统效应。这类概念图提示我们，社会系统模拟必须同时覆盖“表达端”与“效应端”，并允许两者交互演化。

2.　AI 将社会模拟从“少量情景推演”推向“高通量压力测试”

传统社会模拟常面临三类瓶颈：第一，数据形态多样（统计、遥感、文本、传感器、行政记录）却难以对齐；第二，行为参数与网络结构难校准，导致模型可解释但情景覆盖不足；第三，求解与仿真成

更多免费行业报告 并购家 www.ipoipo.net

本高，难以在政策窗口期提供及时的多方案比较。AI 的贡献并不止于加速计算，而在于扩展了可模拟的‘状态空间’：它能把多源数据压缩为统一表示，学习“条件映射”以补齐缺口，并用快速代理模型、生成式情景与不确定性量化，让仿真从少量对比走向系统压力测试。

概括而言，AI 对社会系统模拟的赋能可归纳为四个环节：

- 多源数据融合：把人口、产业、交通、舆情、环境与健康等异构数据映射到可比较的空间—时间—主体表示中；
- 情景生成：在既有情景框架之上插值扩展，生成覆盖政策组合、冲击路径与行为响应的情景库；
- 行为与网络建模：刻画异质主体的响应规则、学习效应与扩散机制，增强对分布效应与涌现现象的表达；
- 快速求解与不确定性评估：用代理/替代模型或并行化学习快速评估多方案，并量化关键不确定性。

表 2-1　AI 赋能社会系统模拟的四类路径与适用场景

路径	核心思路	典型输入	典型输出/用途	治理价值
数据驱动微观模拟	从历史数据学习主体响应与转移规则	行政记录、调查数据、平台数据	分布效应、靶向评估	提升精细治理与公平性
AI 增强 ABM	保留机制结构，用 AI 补齐参数与行为异质性	规则+观测数据+网络结构	扩散路径、阈值风险	支持系统性风险预警
数字孪生+在线仿真	实时数据—模型联动，形成闭环迭代	传感器/IoT、遥感、运行数据	动态预警、处置推演	提升响应速度与协同
大模型智能体仿真	以LLM代理刻画沟通、协商与策略互动	政策文本、访谈语料、规则库	舆论演化、协商过程	评估可行性与接受度

3. 从 ABM/系统动力学到“大模型智能体”的体系演化

把社会系统模拟放在更长的学术谱系中，可以看到一条清晰的演进线：第一代模型强调宏观反馈与均衡推演（如系统动力学、CGE 等）；第二代模型强调微观异质性与网络互动（如 ABM、网络科学模拟）；第三代模型把“数据—模型—治理流程”联成数字孪生，强调在线监测、滚动推演与多部门协同。近两年出现的新趋势，是把大模型与智能体架构引入社会仿真：通过记忆、反思与规划等机制，构造更具情境理解与互动能力的代理主体，用于模拟沟通、协商、信任与策略互动。

例如，Generative Agents 展示了基于大语言模型的代理在沙盒环境中形成日常行为与群体互动的可能性； LLM 与 ABM 结合有望缓解传统 ABM 在行为逼真度、校准与可扩展性上的部分限制，但同时带来可验证性、偏差与伦理风险等新问题。这条演进线的意义在于 AI 不仅提高了计算效率，更在改变社会科学的证据生产方式——从“单次建模”走向“持续迭代的社会实验室”。

社会系统模拟并不只服务于长期规划，也越来越多进入‘预警—响应—复盘’的实时治理链条。从事件检测、预测与影响评估，到因果理解、反事实推演与不确定性量化，再到风险沟通与决策支持。其结构对社会仿真同样具有启发：仿真不仅要回答“会发生什么”，也要回答“为什么”“如果换一种政策会怎样”以及“我们有多确定”。

社会系统模拟要真正进入公共决策，必须跨过“可信度”门槛。与自然科学不同，社会仿真的争议往往不在模型能否运行，而在结果是否可信、是否公平、是否可解释。结合近年的实践经验，建议至少守

更多免费行业报告 并购家 www.ipoipo.net

住三条质量底线：第一，校准与验证——明确模型在何种数据与范围内有效，避免把历史规律外推到结构已变化的新情景；第二，不确定性披露——对关键假设、参数敏感性与情景依赖进行透明呈现，把“结论”与“置信度”绑定输出；第三，公平与可审计——识别对不同群体的差异影响，记录数据与模型版本，确保过程可复盘、可追责。

在制度化支撑上，一方面需要“公共仿真基础设施”：共享的数据标准、评测基准、可复用的情景库与开源工具链，降低跨部门协作成本；另一方面需要“政策沙盒”机制：在可控范围内用仿真进行政策预演与公众沟通，把仿真作为协商工具而非替代决策者的黑箱。只有当仿真从“科研作品”变成“治理流程的一部分”，AI 赋能才能转化为可持续的社会治理能力。

4. 从模型到“社会实验室”的下一步展望

面向未来，AI 赋能的社会系统模拟将更像一个“社会实验室”，但其演化不会停留在“更大规模、更强算力”的线性外推，而会沿着三条更具结构性的方向加速推进。

第一，模拟范式将从“单一模型的预测”走向“多模型联邦的推演”，从“离线仿真”走向“在线伴随”。一方面，基础模型、领域模型与任务模型将呈现模块化组合：以大模型负责把政策文本、制度规则与行为叙事转译为可执行设定，以因果推断与结构识别负责校准关键机制，以多智能体仿真负责生成动态互动与涌现结果，再以可解释性与审计模块负责输出可复核证据链。另一方面，模拟将更深地与现实世界的实时数据流相结合，形成“数字孪生式”的校准与反馈：不

是用一次建模解释一次事件，而是持续地根据新数据修正参数、更新规则、对比反事实，使政策评估从事后总结变为事前预案与事中校正。伴随这一趋势，情景空间将从少量“专家设定情景”扩展为“可生成的情景库”，更系统地覆盖极端事件、复合冲击与跨部门联动，从而提升政策韧性分析与压力测试能力。

第二，AI 与社会仿真的结合将显著提升社会发展与治理的“可解释试错”能力，为复杂公共问题提供更高密度的证据与更清晰的权衡边界。社会治理的核心难题往往不在“是否有效”，而在“对谁有效、以何种代价、在何种条件下失效”。以更细粒度主体为单元的多智能体模拟，可以在统一的制度规则与环境约束下，识别不同群体、地区与行业的分布效应，揭示政策收益与成本在社会结构中的再分配路径；在多政策工具并用时，它还能够刻画工具之间的相互作用与“二阶效应”，避免只看单一指标而忽视结构性外部性。进一步地，当仿真系统具备更强的互动界面（政策制定者、行业主体、公众代表、专家共同参与参数设定、规则修订与结果解释），模拟就不再只是技术分析，而成为“协商与共创”的平台：它以透明的机制假设和可视化的结果呈现，帮助不同利益相关方在同一事实框架下讨论权衡，降低政策沟通成本，提高社会共识的形成效率。对于社会发展而言，这意味着在关键改革、公共服务优化、风险治理与长期转型议题上，我们可以更早发现制度瓶颈与脆弱环节，在更小代价下迭代改进方案；对于治理而言，这意味着把政策试错从现实社会的高成本试错，部分转

更多免费行业报告 并购家 www.ipoipo.net

移到可复核、可记录、可反思的“模拟空间”，以提升决策的稳健性与可解释性。

第三，模拟能力的跃升也会同步放大风险与治理需求，要求把“研究责任”前置为系统设计的一部分。其一是偏见与刻板印象的强化风险。社会仿真模型如果过度依赖历史数据或未经审计的行为规则，可能把结构性不平等“当作规律”固化下来，进而在政策推演中自然化、合理化既有偏差，导致对弱势群体的二次伤害。因此必须要求模型对关键变量的定义、训练数据的代表性、规则设定的价值前提进行显性披露，并通过公平性与分布效应的专门评估来识别偏差来源。其二是操纵与不当监控风险。更强的社会仿真如果与高频行为数据、舆情分析与个体画像深度结合，可能被用于目标化影响、舆论操纵或过度监控，带来权力不对称与权利侵害。对此需要明确用途边界、数据合规要求与审计机制，尤其在涉及个人数据、敏感群体与公共安全场景时，必须建立更严格的访问控制、最小必要原则与外部监督。其三是“黑箱政策”与责任转移风险。当模拟系统在治理流程中扮演越来越重要的角色，容易出现“以模型替代判断”的技术迷信，把价值选择伪装成技术结论，或者把政策后果转嫁给模型，从而削弱问责链条。因此需要把仿真输出定位为“证据输入”而非“决策替代”，并建立可追溯的证据链：包括假设清单、参数版本、数据来源、验证记录、敏感性分析与不确定性披露，使结论能够被复核、质询与修正。其四是开放共享与安全防护之间的张力。高质量仿真需要开放数据、可复现代码与可共享基准，但开放也可能带来滥用风险。更可行的治理路径是

在分级分类基础上构建“可用不可见”的共享机制，推动算法入域、可信计算与第三方审计，用制度化设计而非简单封闭来平衡创新与安全。

归根结底，AI 赋能的社会系统模拟越接近“社会实验室”，就越需要以“共生”视角来理解其边界与责任：AI 让社会科学拥有更强的计算与推演能力，但只有社会科学在价值、制度与可接受性边界上提供约束，并把透明、问责与公平写入研究流程，AI 驱动的仿真才可能真正服务于更高质量的社会发展与更稳健的公共治理。

更多免费行业报告 并购家 www.ipoipo.net

2.3 人文社会科学因果推断范式革命[①]

在公共政策评估、社会治理创新与学术研究中，“相关”往往不足以支撑决策：我们真正需要回答的是，一项政策、一次制度变革、一个平台规则的调整，究竟“导致了什么”，其作用机制与适用边界何在。因果推断正是人文社会科学把经验事实转化为可行动知识的关键桥梁，也是证据型治理得以成立的基础方法。进入大模型时代，因果推断面临两种同时发生的变革：一方面，研究者可以更低成本地获取信息、组织材料、生成候选研究方案；另一方面，生成式能力也加剧了“看似合理但不可检验”的叙事风险，反而提高了对研究设计透明度与可审计性的要求。因此，AI 赋能因果推断的关键不在于替代统计推断本身，而在于把原本高度依赖经验的研究设计环节显性化、流程化，并把“干预—机制—数据—识别—解释”的证据链组织为可复核的工作流。本节以 InterveneBench[②]与 STRIDES[③]为样本，讨论评测体系如何更贴近真实社会系统的干预推理需求，以及多智能体协同如何把理论、方法、数据与验证串成闭环，从而为 AI4SSH 的公共能力

[①] 本节由复旦大学吴力波教授、徐盈辉研究员牵头，复旦大学人工智能创新与产业研究院博士后施正昱、史少杰博士合作撰写。

徐盈辉，复旦大学人工智能创新与产业研究院研究员。日本丰桥技术科学大学计算机科学博士。曾任阿里巴巴研究员，达摩院视觉实验室负责人，菜鸟人工智能部创始人，阿里巴巴搜索事业部基础算法负责人，日本理光主任研究员；主要从事大规模搜索推荐技术，大规模启发式优化引擎建设，强化学习，自监督学习，生成模型，在线优化等人工智能领域的产业化应用；2016-2020 担任中文信息处理学会理事，2030 新一代人工智能重大项目核心专家组专家库成员。

[②] InterveneBench 是面向社会科学情境的“大模型干预推理与因果研究设计”评测基准（benchmark）。其核心不是考察模型是否能背诵因果推断概念，而是以真实社会系统中的政策/制度/项目等“干预”为对象，测试模型能否在给定情境与数据约束下提出可执行的研究设计方案，包括：界定干预与结果变量、构造处理组与对照组/反事实、选择识别策略并明确关键假设，以及给出可核验的检验与稳健性思路。

[③] STRIDES 是一种以社会科学理论与研究流程为导向的多智能体（multi-agent）研究工作流框架，用于支持复杂的社会科学因果研究任务的端到端执行与质量控制。该框架通常将研究过程拆解为若干可分工的角色/模块（如理论建模、研究设计与识别策略、数据与变量实例化、代码执行、批判性审查等），并通过工具调用、运行记录与对抗式审查等机制，把“生成式回答”转化为更可复核的证据链与研究产物。

建设提供可迁移的路径。

本节的意义并不局限于某一套基准或框架，而在于提出一种可迁移的“方法公共化”路径：把因果推断从研究者个体技能推进为可协作的组织能力，并通过评测与流程规范把“可信”前置为最低门槛。当 AI 把信息抽取与流程编排的成本显著降低，真正稀缺的将是能够提出问题、识别机制、审阅假设并对结论负责的“思考能力”。因此，AI 赋能因果推断的最终目标不是自动生成结论，而是帮助我们在更透明、更可复核的证据体系中重新发现思考的价值。

1. 社会因果研究的本质

现代社会治理越来越依赖数据与指标，但数据本身并不等于证据。即便某项政策实施后指标改善，也可能是宏观周期、同期改革或选择性执行共同作用的结果；同样，即便某项治理工具在一地有效，也可能因制度环境、执行能力与群体结构不同而难以外推。因果推断之所以成为证据型治理的硬门槛，是因为它强迫研究者回答三类关键问题：第一，反事实比较对象是谁（对照组如何构造）；第二，识别结论依赖哪些假设（并行趋势、排除限制、不可操纵性等）；第三，结论在何种条件下成立、何处可能失效（外推边界与不确定性）。在这一意义上，因果推断不仅是一套统计工具箱，更是一种规范化的推理结构：它要求把机制叙事、数据约束与检验逻辑写成可复核的证据链。

与数学、编程等“闭式问题”不同，社会因果研究很少拥有预先给定的目标函数、完备的变量定义与可证真伪的唯一答案。现实中的政策干预往往伴随预期效应、选择性执行与多主体适应，同一干预也

更多免费行业报告 并购家 www.ipoipo.net

可能通过多条机制链条发挥作用。研究者需要在制度语境与数据可得性约束下作出一系列设计性选择：如何界定干预边界与实施时间，如何将理论概念映射为可观测指标，如何构造处理组与对照组，如何选择识别策略并提出可检验的假设，以及如何用稳健性分析与替代设定支撑结论。这使得社会因果推断更像一场“研究设计”：难点并不在某一种回归形式，而在于把现实问题转化为可比较、可检验、可复现的研究方案。

2. AI 嵌入因果推断全流程

从流程上看，一项可复核的因果研究通常包含四个相互约束的环节：问题与干预界定、变量与组别构造、识别策略与假设检验、解释与沟通。传统实践中，前两步高度依赖研究者经验与领域知识，后两步依赖计量工具箱与写作规范。大模型的进入首先降低了“从文本到结构”的成本：它能够快速解析政策文本与制度背景，抽取实施细节与潜在机制，并把分散信息聚合为结构化要素（干预、样本、变量、方法、假设与检验）。更重要的是，当大模型与工具调用、多智能体协同结合时，它开始具备在有限信息下提出可执行研究设计草案的能力，并通过数据可得性核查、模拟数据与代码运行对方案进行自检与修正。在治理语境中，这意味着政策评估有机会沉淀为可复用、可审计的能力模块：关键假设被显性化、关键比较结构被标准化、关键结果的外推边界被强制披露。

表 2-2　从研究流程视角下的 AI 因果推断能力映射

研究环节	传统瓶颈	AI 可提供的支持	潜在风险与挑战

问题/干预界定	信息分散、边界不清	政策文本解析、背景梳理、机制假设生成	不把“叙事”当作因果；明确反事实
变量/组别构造	概念难测、样本选择偏差	变量映射、数据可得性核查、组别建议	变量定义可复现；处理组/对照组可解释
识别/检验	方法选择与假设难写清	识别策略候选、假设清单、稳健性脚本	假设必须可检验或可辩护；报告敏感性
解释/沟通	结果难转化为决策证据	结构化报告、可视化草稿、风险与边界提示	诚实表达不确定性；避免过度外推

3. 评测体系的重要地位

如果要判断大模型能否支持真实世界中的社会因果研究，仅靠“能否回答因果常识题”远远不够。现实研究设计至少具有三点特征：第一，以具体干预为中心，而非抽象变量间的因果讨论；第二，在制度与数据约束下做选择，而非在理想条件下推理；第三，因果结构往往并非预先给定，研究者需要提出并论证自己的识别路径。InterveneBench 正是在这一思路下构建：其样本来自大量同行评议实证研究，覆盖多类政策领域与经典识别框架，并以结构无关的方式要求模型完成端到端研究设计推理。该基准的意义在于把评测重心从“模型懂不懂某个方法定义”转向“模型能否在具体语境下把干预—机制—数据—识别—解释串起来”。同时，它也提醒我们：评测必须把关键假设写作、数据可得性核查与解释一致性纳入硬约束，否则模型容易在表面叙事上得高分，却在可检验结构上失真。

4. 从单模型问答到可运行工作流的多智能体框架

基准测试揭示的一个共性问题是：即便大模型具备较强的语言理解与知识检索能力，它仍可能在研究设计上“看似很像，但关键处不对”。STRIDES 提出的思路，是把社会科学专家在真实研究中的协作

更多免费行业报告 并购家 www.ipoipo.net

流程拆解为可分工、可校验的模块，并引入代码执行与批判性反馈形成自我纠错回路。在这一框架下，理论模块先形成可检验的因果叙事与假设，方法模块将其翻译为识别策略与变量设定，数据模块完成概念—指标映射与可得性核查，执行模块生成并运行脚本，审查模块对统计陷阱与逻辑偏差进行对抗式检查。该框架的核心贡献不在于“多了几个代理”，而在于把因果研究中最容易被忽略的环节——理论一致性、数据约束、假设明确性、检验逻辑——转化为可操作的工作流，从而提升研究设计的透明度与可复核性。

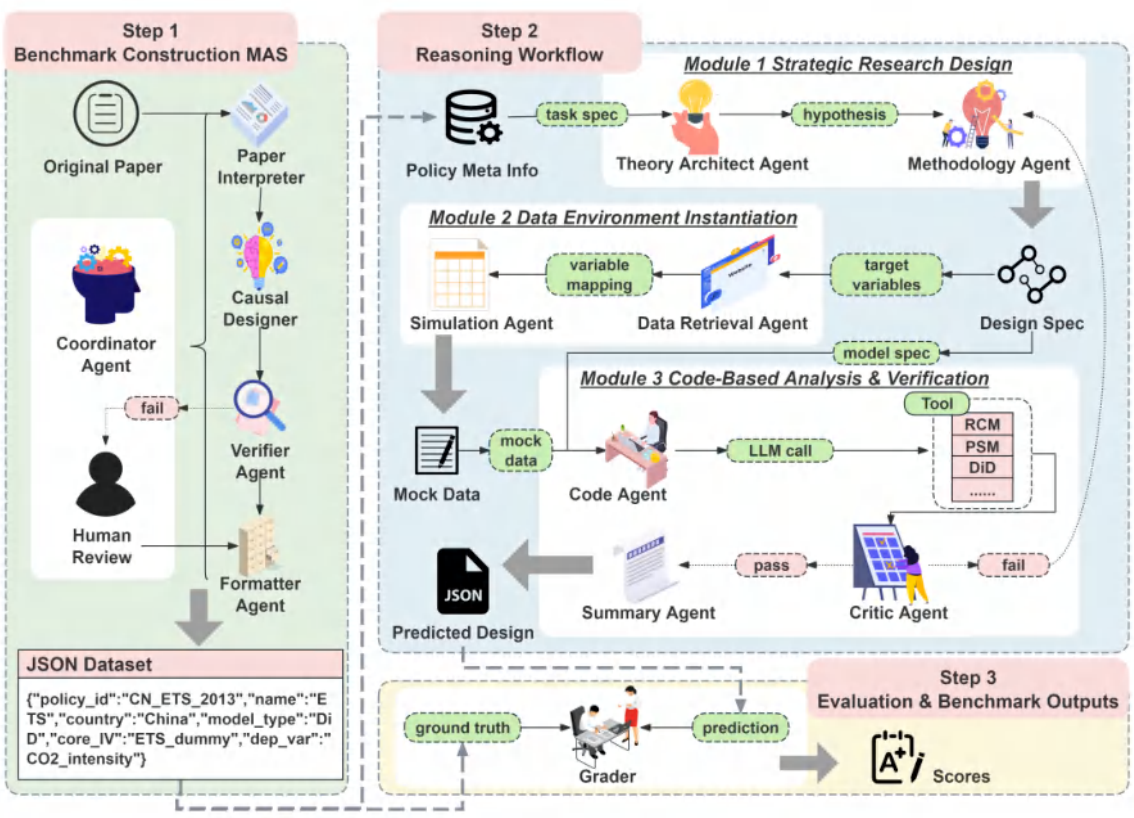


图 2-5　STRIDES 系统概览：从研究设计到对抗式审查的工作流闭环

5.　对学术与治理的结构性启示

从更宏观的视角看，AI 赋能因果推断的结构性价值主要体现在三方面。第一，推动研究设计的显性化。高质量因果研究难复制，往往

不是因为公式难，而是关键判断隐藏在研究者经验中。AI 被用于生成研究设计草案并被要求写出假设与检验逻辑时，隐性知识被迫显性化，有助于提升研究透明度与同行可审计性。第二，降低跨学科协作门槛。AI 可以在一定程度上充当“翻译器”：把政策语句转为可研究的干预定义，把理论概念转为可测指标，把方法假设转为可检验清单，从而减少口径漂移。第三，建设证据型治理的基础设施。AI 不能替代价值权衡与政治决策，但它可以帮助形成更稳健的证据：快速生成多套识别方案备选，提示潜在偏差来源，自动化生成稳健性检验脚本与结构化报告草稿，并用规范语言沟通不确定性与适用边界。长期看，这类能力若与数据治理、伦理与合规框架结合，将使因果推断从专家技能逐步演化为可复用的公共能力。

面向未来，AI 赋能因果推断至少有三条值得关注的演进方向：其一，从文本理解走向多模态证据整合，提升对表格、图形、附录与制度文件细节的抽取与引用定位能力；其二，从单模型问答走向“可运行的因果系统”，内置领域知识、数据连接器、识别策略库与稳健性模板，并通过日志与版本管理实现可审计；其三，从学术工具走向公共基础设施，建立标准化评估协议、开放基准与透明质量控制，使因果评估更像可持续的公共服务。

更多免费行业报告 并购家 www.ipoipo.net

2.4 基于大模型智能体的综合研究工作流[①]

人文社会科学的许多高价值研究并非“算一次就完”，而是开放式、长链条、强情境的：问题界定、理论假设、材料搜集、数据与指标口径、方法选择、代码实现、解释与复核，环环相扣，任何一环失稳都可能使结论偏离。大模型在“生成答案”上已显著降低表达成本，但要把研究从文本产出推进到可复核的证据产出，关键瓶颈往往不在单步推理，而在跨步骤一致性、证据定位、工具执行与质量控制。所谓“大模型智能体”（Agents），正是在这一背景下出现的组织技术：它把模型置于可执行的角色中，并把研究过程拆解为可分工、可校验、可追责的子任务，通过任务编排、工具调用、对抗审查与人在回路，将‘一次性回答’升级为‘可运行工作流’。本节在前文方法型赋能（模拟、因果推断）之外，进一步强调一种更具一般性的“研究运营能力”（ResearchOps）：让跨学科综合研究具备稳定交付与持续迭代的机制，这也是 AI4SSH 从项目探索走向体系建制不可或缺的支撑。

大模型智能体的核心价值，不在于把某一种方法做得更强，而在于把综合研究组织成可复核的工作流。与单模型问答相比，智能体框架通过角色分工与互检对冲不确定性，通过工具与日志把推理锚定到证据与运行结果，并通过对抗式审查与人在回路建立质量门槛。

以 STRIDES 为例，其三阶段结构展示了可迁移的通用模式：在理论叙事与方法设定之间建立“解耦—对齐”机制，在数据可得性与口

[①] 本节由复旦大学吴力波教授、徐盈辉研究员牵头，复旦大学人工智能创新与产业研究院博士后施正昱、史少杰博士合作撰写。

径差异下完成“实例化—验证”，并在代码执行与批判审查中形成“可运行—可审计”的证据闭环。本节据此归纳面向 AI4SSH 的智能体式研究工作流应当如何设计，并提出面向学术治理与平台建设的落地建议。

1. 综合研究需要“工作流”而非“答案”

在真实研究与治理场景中，研究对象往往处于制度变动、冲击叠加与主体适应之中，因果结构与数据条件并不固定。研究者一方面要保持对机制的解释力，另一方面要面对材料不完备、指标口径不统一、证据跨文档分布等现实约束。这意味着综合研究更像一场“过程管理”：它需要在每一步把假设写清、把证据标定、把选择理由记录下来，并允许在新证据出现时回溯修正。

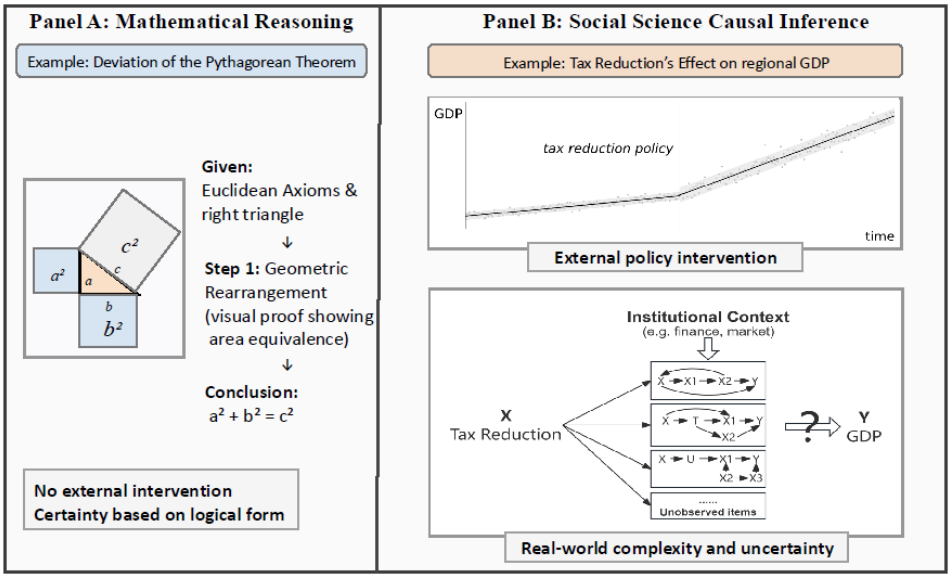


图 2-6 从封闭式问题到开放式研究：社会科学为何更需要工作流

单模型问答之所以容易失稳，常见原因并非“不会推理”，而是“无法维持研究过程的一致性”：长文档中关键证据往往分散在跨页表格、附录与脚注中，结论依赖的口径说明可能只出现一次；同时，

更多免费行业报告 并购家 www.ipoipo.net

多源材料（政策文件、统计年鉴、问卷数据、平台数据、媒体报道）在语义与统计结构上并不一致，若缺少明确的对齐与校验步骤，模型容易在表述上自洽、在证据上失真。因此，只有把研究过程工作流化，生成式能力才可能真正转化为可复核的公共能力。

2. 大模型智能体以“结构”对冲不确定性

智能体并不是给模型换个名字，而是把模型放入一个有目标、有状态、有工具、有约束的执行框架中，并以编排结构保证跨步骤一致性。从工程实现看，这类系统通常需要一个“协调器”（Coordinator）维持全局上下文与状态机：决定下一步做什么、由谁做、失败如何回滚、哪些输出进入‘事实库’与‘假设库’。协调器的意义在于把长链条研究从“连续对话”转为“可控流程”，避免模型在长流程中逐步漂移。

就综合研究而言，智能体框架常通过四类结构化机制降低失稳概率。任务分解与阶段门控用于把复杂问题拆分为可验证子任务，并为关键节点设置“通过/退回/升级人工”的门槛；角色分工与交叉校验用于把机制解释、方法设定、证据抽取与结果写作分离，让不同角色相互制衡；工具调用与证据锚定通过检索、文档解析、表格定位、数据处理与代码执行，把语言推理锚定到可复核的证据与运行结果；对抗式审查与人在回路则把质量控制前置到过程之中，对低置信或高风险结论进行人工裁决。

3. 研究产出标准化

智能体式综合研究若要真正进入学术治理体系，必须把产出从“连续文本”升级为“结构化研究产物”。结构化并不意味着抹平学科差异，而是为跨团队协作提供最小公共接口：研究问题如何表述、干预/事件如何界定、变量口径如何说明、关键假设如何列示、证据来自哪里、哪些结论具有不确定性。这些要素一旦以统一 schema 表达，就能在不同议题、不同团队之间复用，也能为审稿、复核与责任追溯提供抓手。

更进一步，标准化应当与“过程日志”绑定。对蓝皮书写作而言，最有价值的不只是最终段落，而是可追溯链条：哪些材料被读取、哪些证据被引用、哪些选择被做出、哪一步被批判代理退回、哪一处由专家裁决。当日志与版本管理建立起来，智能体才能从写作助手变为研究基础设施：既提高效率，也保留可问责性。

表 2-3 智能体式研究的关键产物与最低质量要求

产出类型	内容要点	最低质量要求（建议）
研究问题与情境摘要	对象、边界、时间线、制度约束	明确引用来源；与后续设定一致
机制/假设清单	作用机制、关键假设、可能异质性	可检验表述；注明适用范围与不确定性
变量口径与数据字典	概念—指标映射、口径差异、缺失处理	口径可追溯；版本化管理
分析脚本与运行记录	可执行代码、参数、运行环境	可复现；记录关键报错与修复
审查与修订记录	批判检查点、退回原因、人工裁决	可审计；高风险结论需 HITL
结论与边界披露	核心发现、外推边界、风险提示	证据定位；避免无来源确定性陈述

4. 可迁移的工作流模板：以 STRIDES 为例的“研究链条化”组织

更多免费行业报告 并购家 www.ipoipo.net

本文问题出的 STRIDES 框架最初服务于研究设计推理，但其结构更像一张“研究工作流组织蓝图”，对更广义的综合研究具有可迁移价值，能够将研究过程拆分为顺序模块，并在模块内部建立可迭代反馈，使理论、数据与验证形成闭环。

- **其一是设计模块：**先形成可解释的机制叙事与可检验假设，再将其翻译为可执行的分析设定，避免“统计可行但理论牵强”；
- **其二是数据环境模块：**将抽象概念映射为可测指标，核查数据可得性与口径差异；在条件不明时通过受控模拟或替代数据先跑通逻辑，以暴露不可识别或不可执行的漏洞；
- **其三是代码与审查模块：**把方案落到可运行脚本，并通过批判角色检查统计陷阱与解释一致性；通过日志记录与版本管理形成可审计链条。

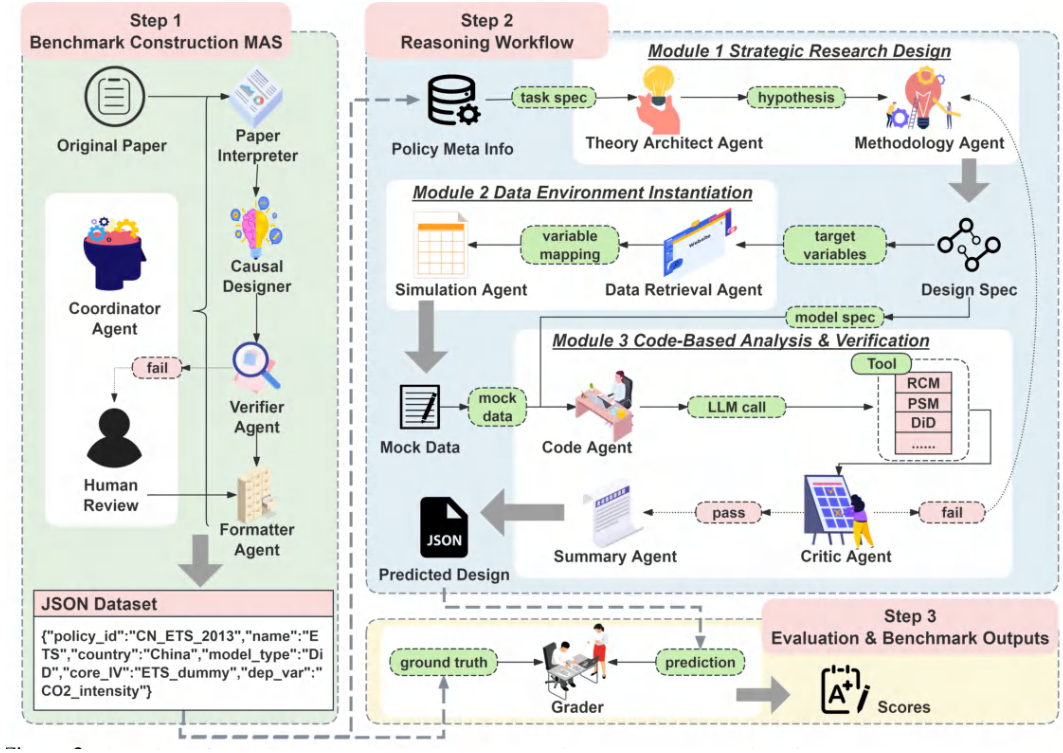


图 2-7 智能体式研究运营工作流示意：设计—实例化—执行与审查

这套结构更像“模板化的研究管线”：不仅提升输出质量，也把复核成本从事后追责前移到过程验收。

将工作流视角从研究设计推广到更广义的综合研究，可形成一组更具普适性的应用外延。

- **一是政策与治理评估的快速可复核支撑：**智能体先形成可审计的备选方案与证据清单，并把不确定性与缺口显式化；
- **二是跨学科文献综述的证据结构化：**将文献抽取为统一 schema 并保留证据定位；
- **三是复杂议题的情境化推演：**以任务图谱组织事实核对、机制建模、反事实设定与政策备选；
- **四是研究过程的可复现与可问责：**以代码执行、日志留存与对抗审查推进可运行产出。

这些场景的共同点不是方法相同，而是流程相同：都需要把‘材料—设定—证据—解释’组织为可复核链条。因此，智能体的通用价值并不取决于某一学科的专门模型，而取决于其是否能把研究的关键中间产物标准化并沉淀为组织能力。

5. 减少幻觉与偏差的“过程约束”

智能体并不自动保证正确。它可能因“可执行性偏好”而倾向选择更简化、可跑通的设定，从而牺牲解释丰富性；也可能在长文档与多表格材料中遗漏关键证据，导致结论失真；更现实的风险还包括：对敏感数据的误处理、对群体差异的忽视、对制度细节的误读，以及

更多免费行业报告 并购家 www.ipoipo.net

把不确定性表达成确定性结论。

因此，治理重点应放在过程约束而非结果宣称上：要求关键结论配套证据定位与可追溯引用；建立版本化数据与脚本管理，确保迭代差异可解释；通过对抗式审查与抽样人工复核把质量门槛前置，尤其对涉及敏感群体、公共安全与重大政策影响的结论设置更高的 HITL 门控。在制度层面，可以将任务按影响风险分级：低风险任务强调效率与一致性，中风险任务强调可复核与对抗审查，高风险任务强调专家裁决与合规审计。

面向未来，智能体式综合研究的演进方向可以概括为三点：第一，走向可运行的研究系统，内置领域知识、数据连接器、工具注册表、策略库与写作模板，并通过日志与版本管理实现全程可审计；第二，走向面向任务的评测与基准，用贴近真实治理情境的任务评估工作流可靠性，把质量控制从结果评奖前移到过程验收；第三，走向公共能力平台化，将高质量工作流、schema 与工具链沉淀为平台，让更多研究者在统一底座上协作创新。

与之相伴的，是组织与制度的配套改革：一方面，需要为数据治理、评测基准维护、工具链开发等“平台型贡献”提供可被认可的评价口径；另一方面，需要为跨学科团队配置稳定的工程化支撑与合规支持，避免把高风险应用的责任完全压在单个研究者身上。当平台与制度协同推进，智能体才可能从“辅助写作”走向“支撑研究共同体运行”。

本节强调的是 AI 对“研究组织”的赋能。当表达与编码的成本被大幅压缩，真正稀缺的是能够提出问题、校准机制、审阅证据并对结论负责的思考能力。智能体工作流的意义正在于：它把思考从个人心智活动扩展为可协作、可复核、可持续的组织过程，使跨学科综合研究能够在更高频率、更高透明度与更强问责性下运行。这也为蓝皮书后续关于组织创新、平台建设与人才培养的讨论提供了一个具体抓手：让 AI 成为连接理论、数据与治理实践的“研究组织技术”，从而重新发现思考的价值。

更多免费行业报告 并购家 www.ipoipo.net

参考文献

[1] Ramanathan, V., Xu, Y., & Versaci, A. (2022). Modelling human–natural systems interactions with implications for twenty-first-century warming. Nature Sustainability, 5(3), 263–271. https://doi.org/10.1038/s41893-021-00826-z

[2] Battiston, S., Monasterolo, I., Riahi, K., & van Ruijven, B. J. (2021). Accounting for finance is key for climate mitigation pathways. Science, 372, 918–920. https://doi.org/10.1126/science.abf3877

[3] Bertram, C., Brutschin, E., Drouet, L., et al. (2024). Feasibility of peak temperature targets in light of institutional constraints. Nature Climate Change, 14, 954–960. https://doi.org/10.1038/s41558-024-02073-4

[4] Rockström, J., Gupta, J., Qin, D., et al. (2023). Safe and just Earth system boundaries. Nature, 619, 102–111. https://doi.org/10.1038/s41586-023-06083-8

[5] Moore, F. C., Lacasse, K., Mach, K. J., et al. (2022). Determinants of emissions pathways in the coupled climate–social system. Nature, 603, 103–111. https://doi.org/10.1038/s41586-022-04423-8

[6] Beucler, T., et al. (2024). Climate-invariant machine learning. Science Advances, 10, eadj7250. https://doi.org/10.1126/sciadv.adj7250

[7] World Meteorological Organization. (2019). Guidance on Integrated Urban Hydrometeorological, Climate and Environmental Services (Vol. I: Concept and Methodology) (WMO-No. 1234). Geneva: WMO.

[8] Ouyang, H., Tang, X., Brasseur, G., et al. (2026). AI for Sustainable Urban Systems Plus (AISUS+): A Global Initiative to Enhance Climate and Environmental Resilience in Cities. Bulletin of the American Meteorological Society, 107(3), E352–E355. https://doi.org/10.1175/BAMS-D-26-0018.1

[9] Park, J. S., O'Brien, J. C., Cai, C. J., Ringel Morris, M., Liang, P., & Bernstein, M. S. (2023). Generative Agents: Interactive Simulacra of Human Behavior. Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST ' 23). https://doi.org/10.1145/3586183.3606763

[10] Gao, C., Lan, X., Li, N., et al. (2024). Large language models empowered agent-based modeling and simulation: a survey and perspectives. Humanities and Social Sciences Communications, 11, 1259. https://doi.org/10.1057/s41599-024-03611-3

[11] Camps-Valls, G., Fernández-Torres, M.-Á., Cohrs, K.-H., et al. (2025). Artificial intelligence for modeling and understanding extreme weather and climate events. Nature Communications, 16, 1919. https://doi.org/10.1038/s41467-025-56573-8

[12] Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology, 66(5), 688–701. https://doi.org/10.1037/h0037350

[13] Pearl, J. (2009). Causality: Models, Reasoning, and Inference (2nd ed.). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511803161

[14] Angrist, J. D., & Pischke, J.-S. (2009). Mostly Harmless Econometrics: An Empiricist's Companion. Princeton, NJ: Princeton University Press.

[15] Imbens, G. W., & Rubin, D. B. (2015). Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge: Cambridge University Press.

[16] Athey, S., & Imbens, G. W. (2017). The State of Applied Econometrics: Causality and Policy Evaluation. Journal of Economic Perspectives, 31(2), 3–32. https://doi.org/10.1257/jep.31.2.3

[17] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal, 21(1), C1–C68. https://doi.org/10.1111/ectj.12097

[28] Shinn, N., Cassano, F., Berman, E., Gopinath, A., Narasimhan, K. R., & Yao, S. (2023). Reflexion: Language Agents with Verbal Reinforcement Learning. arXiv preprint arXiv:2303.11366. https://doi.org/10.48550/arXiv.2303.11366

[19] Hong, S., Zhuge, M., Chen, J., et al. (2023). MetaGPT: Meta Programming for a Multi-Agent Collaborative Framework. arXiv preprint arXiv:2308.00352. https://doi.org/10.48550/arXiv.2308.00352

[20] IPCC. (2022). Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change (H.-O. Pörtner, D. C. Roberts, M. Tignor, E. S. Poloczanska, K. Mintenbeck, A. Alegría, M. Craig, S. Langsdorf, S. Löschke, V. Möller, A. Okem, & B. Rama, Eds.). Cambridge University Press, Cambridge, UK and New York, NY, USA. https://doi.org/10.1017/9781009325844

更多免费行业报告 并购家 www.ipoipo.net

第三章 AI 与认知科学哲学的双向奔赴[①]

人工智能的持续演进，不仅扩展了机器的信息处理、内容生成和行动能力，也重新打开了“智能如何可能”这一认知科学的根本问题。大模型和智能体展现出的推理、规划与交互能力，正在迫使哲学重新审视理解、意向性、具身性、因果和意识等核心概念；与此同时，脑科学关于感知、记忆、情绪、价值和认知控制的研究，也为 AI 突破模式匹配局限、走向情境理解、意义生成和反思判断提供了机制参照。由此，AI 与认知科学哲学形成了真正的“双向奔赴”：认知科学哲学通过澄清智能的结构、边界和价值方向，推动 AI 走向可信、可理解、可校准；AI 则把长期依赖概念分析和思想实验的心智问题转化为可建模、可实验、可比较的研究对象，推动哲学进入与工程、数据和实验持续对话的新阶段。本章从历史、机制、能力与未来四个层面展开：梳理 AI 与认知科学哲学相互塑造的理论轨迹，揭示人类智能的认知底座，讨论认知型 AI 在主体理解、意义生成和深度思考方面的能力要求，并围绕理解、意向性、具身认知、因果与意识等前沿议题，展望二者协同演进的未来方向。

[①] 本章由复旦大学哲学学院黄翔副教授和类脑智能科学与技术研究院肖晓研究员合作撰写。

黄翔教授研究方向为科学哲学、科学史、认知科学哲学、知识论。目前关注科学实践哲学、历史知识论、自然化意向性等论题。

肖晓研究员为复旦大学认知神经科学中心执行主任，复旦大学附属华山医院和华东医院双聘研究员。上海市东方英才（2023），上海科技启明星（2019），上海市神经科学学会疼痛转化研究分会副主任，Brain Communications 杂志、Behavioural Brain Research 杂志和 Journal of Psychopharmacology 杂志副主编。使用神经生物学、认知心理学和计算科学等交叉前沿技术，围绕“情感的中枢机制及可塑性调控”这一科学主题，系统研究了情绪和情感记忆的神经生理和心理机制，推动了情感障碍中枢机制研究的发展及临床诊疗应用，并建立了一系列神经调控和行为疗法为基础的创新技术。

本章提要：

- 从符号主义、联结主义、4E 认知到预测加工，AI 与认知科学哲学始终相互塑造：认知理论为机器智能提供模型，AI 则成为检验和重构心智理论的重要实验场。
- 人类智能并非线性的“输入—输出”过程，而是感知与注意、记忆与认知地图、情绪与价值、认知控制在身体、环境、经验和目标之间协同建构意义的动态系统。
- 脑科学和认知科学启示 AI 实现四重跃迁：从对象识别走向情境理解，从信息存储走向经验组织，从情绪分类走向价值理解，从流畅生成走向自我监控、错误修正和反思判断。
- 面向人文社会科学的认知型 AI，应发展主体理解、意义生成和深度思考三类能力；其目标不是替代人的判断，而是帮助人比较证据、辨析价值并保持认知主动性。
- 大模型与智能体正在重新激活“中文屋”、意向性、4E 认知、因果与反事实、意识等哲学问题，并将其转化为可建模、可实验、可比较的研究对象，为可信 AI 和人机共生提供新的理论基础。

更多免费行业报告 并购家 www.ipoipo.net

3.1 AI 认知的哲学本质[①]

上世纪五、六十年代诞生的交叉学科认知科学致力于回答一个根本性问题：智能如何可能？其中认知科学哲学使用概念分析的方法反思产生智能的认知机制，同时用科学发现反哺传统的心灵哲学，而人工智能即 AI 则致力于在硅基上复现智能行为。AI 与认知科学哲学之间，随着认知科学的不断发展，一直保持着不断的双向重构。认知科学在其发展过程能够经历了数次重大变革：从认知即符号计算，到认知即亚符号激活，再到认知即脑-身体-世界的耦合。每一次范式转换都没有完全取消前一个范式，而是扩大了研究工具箱，其中最为重要的工具就是 AI。AI 使哲学走出概念思辨的舒适区，进入一个与工程、实验和数据持续对话的新阶段。

1. AI 和认知科学哲学在符号主义和联结主义时代

1956 年常被看作是认知科学诞生之年，心理学家、计算机科学家、神经科学家和语言学家开始汇聚，他们秉持一个共同信念，即人类认知可以理解为一种信息处理过程。一些标志性事件包括：麦卡锡（John McCarthy）、明斯基（Marvin Minsky）、香农（Claude Shannon）、纽厄尔（Allen Newell）、司马贺（Hebert Simon）等人在达特茅斯会议上正式确立了 AI 这一领域；乔治·米勒（George Miller）在麻省理工学院信息论研讨会上展示了人类短时记忆的容量限制这一信息加工特征；纽厄尔与西蒙公布一款名为“逻辑理论家”（Logic Theorist）的程序，能够证明怀特海和罗素《数学原理》中命题逻辑部分的许多定理等[1-

[①] 本节由复旦大学哲学学院科学哲学与逻辑学系黄翔副教授撰写。

2]。 主导上世纪 60-70 年代的认知科学的是 AI 工作者们坚持的符号主义立场，认为任何智能系统，无论大脑还是机器，本质上都是操作表征信息的物理符号的系统。

将大脑理解为一种信息加工系统，将认知过程理解为对符号的操作的看法是对主导 20 世纪上半叶的行为主义心理学的否定。行为主义心理学只研究可观察的外部行为，回避“心灵”、“意识”等不可观测的内部概念，认知和学习被还原为刺激-反应联结的形成过程。20 世纪上半叶心灵哲学受到行为主义极大的影响。例如，赖尔（Gilbert Ryle）《心的概念》一书中猛烈批判笛卡尔的“机器中的幽灵”，主张心理状态应通过外部行为来定义[3]。这为后来强调输入-输出关系的计算主义铺平了道路，但也因其忽视内部认知过程而备受诟病。又如，普赖斯（Ullin T. Place）和斯马特（John Jamieson.C. Smart）激进地认为心理状态就是大脑的神经生理状态[4-5]。这一观点虽然在当时缺乏实证支持，但为日后的认知科学奠定了“意识活动有其物质基础”的自然主义。在符号主义 AI 的启发下，20 世纪下半叶越来越多的心灵哲学家开始接受功能主义立场，心智状态由其功能角色定义，类似于计算机的软件与硬件的分离。这意味着心智可以在不同物理基底上多重实现[6]。认知科学哲学中的功能主义的典型代表是福多（Jerry Alan Fodor）的“思维语言”（Language of Thought，LOT）理论[7]。该理论认为人类的思维并非依赖自然语言，而是运行在大脑中的一套内在的、类似于逻辑符号系统的“心语”（Mentalese）。这套“心语”具有类似的主谓结构的句法结构和语义规则，能够通过符号的精确运算来表征世界，从而

更多免费行业报告 并购家 www.ipoipo.net

解释人类思维为何具有系统性和逻辑推理能力如三段论演绎。为功能主义提供了坚实的认知基石——既然思维是“心语”符号的句法运算，那么心智的本质就完全独立于大脑的物理基质无论是神经元还是硅片，只要软件相同，硬件如何实现并不重要。这恰好契合并驱动了当时符号主义 AI 的核心信念：智能本质上就是物理符号系统的操作，因此可以在数字计算机上通过运行显式的逻辑程序来完全实现高阶推理。

达特茅斯会议之后的二十年间，AI 步入黄金发展期。这一阶段，人工智能研究取得诸多突破性成果，计算机可自主求解代数应用题、证明几何定理、识别英文语言，这类机器智能行为在当时极具颠覆性。AI 研究者普遍高度乐观，认为通用智能机器将在二十年内落地。以美国国防部高级研究计划局（Defense Advanced Research Projects Agency，简称 DARPA）为代表的国家政府机构大力注资，为早期人工智能研究提供了充上世纪足的资金保障。到了上世纪 70 年代中，早期 AI 的缺陷不断暴露。首先，一些底层局限无法突破，多数程序仅能处理简单理想化场景，实用价值极低，本质上属于“玩具级”模型。其次，计算复杂度与组合爆炸难题突出。早期 AI 的简易解决方案，无法完成规模化、复杂化迭代，彻底阻断了实用化升级路径。计算机擅长定理证明、几何推演等逻辑运算任务，但人脸识别、室内避障等人类本能的简单行为，对机器而言却难度极高。同时，逻辑推演框架存在框架问题与限定问题，传统逻辑体系无法适配规划、默认推理等日常场景的推演需求，研究者虽研发出非单调逻辑、模态逻辑等新型

逻辑体系，仍难以彻底破解难题。再次是计算机算力的硬性制约。彼时计算机内存与运算速度严重不足，无法支撑复杂智能任务。同时，技术进展乏力也引发大规模撤资[8]。早期 AI 展现出来的技术性问题同时也引发了认知科学哲学的反思。卢卡斯（John R. Lucas）认为哥德尔的不完全性定理表明，形式系统（如计算机程序）永远无法看出某些命题的真理性，而人类却能做到这一点[9]。 更为著名的是塞尔（John Searle）提出的"中文屋"论证，该论证试图表明不能称一个程序"理解"了它所使用的符号，因为句法操作永远无法产生语义理解。如果这些符号对机器而言毫无意义，那么就不能把机器描述为在"思考"[10]。德雷福斯（Hubert Dreyfus）则批评了早期 AI 研究者们乐观的态度，并对 AI 的基本假设提出批判。他主张，人类推理实际上涉及的"符号处理"非常少，而更多的是具身的、本能的、无意识的"知道如何做"（know how）[11]。

上世纪八十年代，AI 迎来了一轮复苏，这主要来自两种技术突破。首先，以专家系统为核心的知识驱动 AI 技术落地普及。七十年代的技术困境让 AI 研究者们认识到，智能行为的核心依托是海量、细分的领域知识运用能力，而非通用逻辑推演。专家系统依托领域专家知识提炼的逻辑规则，解决特定专业场景问题，有效规避了通用常识推理的技术难题，且结构简洁、搭建与迭代成本低，是首款具备稳定实用价值的 AI 技术形态。由于技术落地成效显著，各国政府重启大额资金投入，行业资金面全面回暖。因此，基于知识的系统与知识工程，成为八十年代 AI 领域的核心研究方向。其次，联结主义迎来关键复

更多免费行业报告 并购家 www.ipoipo.net

兴。联结主义认为智能源于大量简单单元如神经元之间的并行连接与协同活动，通过调整连接权重从经验中学习。它不依赖符号规则，而是以分布式表征和梯度下降为核心，模拟大脑式的亚符号信息处理。1982 年，物理学家霍普菲尔德（John Hopfield）提出新型神经网络模型，证明神经网络可依托全新机制完成自主学习与信息处理。同期，鲁梅尔哈特（David Rumelhart）、辛顿（Geoffrey Hinton）等人推广反向传播训练算法，解决了多层网络训练的核心难题，为联结主义奠定技术基础。鲁梅尔哈特与麦克莱兰（James McClelland）主编的《并行分布式处理》论文集问世，系统梳理神经网络核心理论，统一了领域研究方向[12]。依托本轮技术积累，神经网络在九十年代迎来商业化落地，广泛应用于光学字符识别、语音识别等场景。

联结主义虽然在应用层面上展现出许多优于符号主义的地方，但也引起了不少争议。符号主义的辩护者福多认为，在联结主义的神经网络中，知识不是存储为清晰的“词汇”或“规则”，而是以权重矩阵的形式分布在成千上万个神经元连接中，因而是以“亚符号”（sub-symbolic）方式运作，无法处理符合语法规则的人类高级思维系统，也无法像人类那样，用有限的符号生成无限的新思想。神经网络学习的是统计规律，而不是逻辑必然性。即使神经网络能学会解决逻辑题，也仅仅是模拟了推理，而非实现了推理[13-14]。福多的观点在认知科学内部引发了大量的讨论[15-17]。对于联结主义者来说，高阶推理未必需要内置符号规则，大规模的亚符号交互完全可以在隐层中涌现出类似逻辑结构的表征。符号本身可以看成是一种高层次的幻觉，底层只要

权重足够深、数据足够多，高层自然会产生理性结果[18]。这在日后的GPT 的研发中得到验证。

2. AI 和认知科学哲学在 4E 认知时代

上世纪八十年代末到九十年代初，AI 产业又一次遇到瓶颈。一方面，专家系统开始暴露明显短板：维护成本高、更新困难、不具备自主学习能力，输入非常规信息时极易出现严重错误，多年前发现的限定问题等缺陷也始终未能解决，仅能在极小范围场景发挥作用，难以大范围落地。另一方面，美国战略计算计划大规模削减 AI 资助，DARPA 新任管理层不再将人工智能视作核心发展方向，经费转向见效更快的项目。行业遇冷之际，以布鲁克斯（Rodney A. aBrooks）为代表的机器人领域学者提出全新研究路线——新式人工智能与具身推理。研究者认为，真正的智能需要实体载体，机器必须拥有感知、行动、适应现实环境的能力；感知与运动能力是常识推理等高层智能的基础，抽象逻辑反而并非核心，主张从底层行为搭建智能体系。布鲁克斯直接质疑物理符号系统假说，提出现实世界本身就是完整模型，机器只需持续感知环境即可，不必完全依赖符号运算[19-20]。

布鲁克斯的研究推动了认知科学哲学中的 4E 认知的展开，并被当作具身认知中的重要案例。4E 认知颠覆了经典认知观，坚持认知不必以内在表征为核心，可以是感知与行动的直接耦合；大脑也不再是孤立的符号处理器，而是整个耦合系统中的协调者[21]。所谓“4E”，分别指具身（embodied）、嵌入（embedded）、延展（extended）、生成（enacted）四种认知观。具身认知认为人类的认知与思考无法脱离身

更多免费行业报告 并购家 www.ipoipo.net

体存在。身体的生理结构、感官运动系统和日常体验，共同塑造了我们的心智[22-24]。瓦雷拉（Francisco J. Varela）、汤普森（Evan Thompson）、罗施（Eleanor Rosch）在《具身心智》中系统阐释了这一理论，推翻了“心智可脱离身体独立运作”的传统观点，证实身体是认知最本源的载体[25]。人类绝大多数抽象概念，本质都是身体体验的隐喻延伸。莱考夫（George Lakoff）与约翰逊（Mark Johnson）指出，所有抽象思维、情感与理念，都根植于直观的身体经验[26-27]。 延展认知主张，认知从不局限于人体与颅骨之内。只要条件适配，笔记本、手机、计算器等外部工具，就能成为认知过程的真实组成部分，延伸、补足我们的记忆与思考能力。我们的记忆、思考与信息检索，早已不依赖单一的大脑机能。各类随身工具都是认知的外延，深度融入我们的心智活动。也正因如此，人类天生就是适配工具的“赛博格”，始终与外部设备共生完成认知[28-29]。嵌入与生成，进一步完善了认知的完整形态，证明认知绝非孤立、静止的思维活动。所谓嵌入，是指所有认知都必然扎根于特定的环境、文化与社会情境。情境塑造思维模式，脱离具体场景谈论认知，本身就是片面的。而生成认知，彻底颠覆了“认知是被动接收信息”的传统认知。传统观点认为，世界是固定的，大脑只是复刻、表征外界信息的容器。但生成认知的核心是：认知通过行动主动生成意义，我们对世界的感知与定义，始终源于自身的行动需求与实践互动。我们会主动辨识危险、抓取食物、利用周遭事物，感知从来不是单纯的感官接收，而是与行动形成闭环。正是一次次与世界的互动行动，赋予了万物相应的意义。认知不是被动记录世界，而

是主动建构、生成意义的动态过程[30-33]。

上世纪九十年代，AI 逐步兑现早年部分目标，相关技术悄悄渗透各行各业。行业进步一方面得益于计算机算力持续增长，另一方面源于研究者聚焦细分单一问题、采用严谨科学标准开展实验。但学界对于始终无法实现类人通用智能的原因没有统一结论，研究领域逐步拆分，分化为多个互不统属的细分方向，落地成果也远多于以往。例如，智能体范式（intelligent agent）得到广泛认可。研究者将决策论、经济学理论引入 AI，融合计算机模块化思想，形成完整智能体理论。智能体指感知环境、采取行动以提升目标成功率的系统，小到专用程序、大到人类与组织都可纳入这一范畴，该范式将 AI 定义为对各类智能体的研究，拓展了学科边界。它允许研究者独立拆解、验证细分问题，搭建起统一交流框架，还能和经济学、控制论等学科互通理论，学界期待依靠多智能体交互搭建功能更全面的通用系统。这一时期严谨数理方法成为主流，被称作“规整派的胜利”（victory of the neats）[34]。学界意识到数学、运筹学等成熟学科已有大量可复用工具，统一的量化语言让 AI 研究更易跨界合作，成果可验证、可量化，学科科学性大幅增强。珀尔（Judea Pearl）将概率论、决策理论引入 AI，贝叶斯网络、隐马尔可夫模型、随机建模、优化算法等工具被普遍使用，神经网络、演化算法等计算智能方向也建立起完整数学描述。许多 AI 算法悄悄嵌入各类软件系统，广泛应用于数据挖掘、工业机器人、语音识别、金融风控、医疗诊断、搜索引擎等领域，但 AI 本身很少获得外界认可。一旦技术成熟普及，相关功能便不再冠以 AI 名号；为

更多免费行业报告 并购家 www.ipoipo.net

方便获取资金、规避行业负面印象，不少研究者改用信息学、认知系统、计算智能等名称指代自身工作，很长一段时间从业者都刻意回避“AI”一词。

3. 近十年的 AI 与认知科学

二十一世纪以来，海量大数据、低成本高性能硬件与机器学习技术相互赋能，渗透各行各业。大数据技术是一套用于处理海量、多样且快速产生之数据的采集、存储、管理和分析方法的统称。它依赖分布式计算和存储架构（如 Hadoop、Spark），能够从庞杂的数据中高效地挖掘出有价值的信息和规律。它不仅服务商业，也用于生态、经济等科研领域，深度学习的 AI 模型如深度卷积网络、循环网络等需要海量数据来“喂养”和训练，大数据技术恰好提供了这种计算和存储能力。深度学习是机器学习的一个分支，它通过构建包含多个隐藏层的人工神经网络，让计算机自动从海量数据中学习复杂的特征和模式。简单来说，它模仿人脑的层级化处理方式，将原始数据逐层抽象，最终实现对图像、语音、文本等复杂信息的识别与生成。大数据技术支持的 AI 逐步成长为规模庞大的商业赛道。多层神经网络相关技术实现关键突破，深度学习重回学界视野，为产业落地提供核心技术底座。近十几年间，硬件成本持续走低，算力供给充足，互联网行业积攒下海量文本、图像、音频数据，为机器学习模型训练提供了充足素材。各大科技企业开始布局人工智能研发团队，将机器翻译、语音识别、图像检索、推荐算法等技术融入自身产品，AI 从实验室研究正式转向商业化落地。投资市场热度同步升温，创投机构持续加注 AI 初创

企业，一批专攻计算机视觉、自然语言处理的公司陆续成立。传统行业也开始引入智能算法，用于生产优化、风险评估、用户行为分析等场景，人工智能不再只是小众科研课题，逐步形成完整上下游产业。大语言模型因运而生，它一种基于海量文本数据训练的 AI 系统，通过学习词语间的关联和上下文规律，能够理解和生成自然语言。它本质上是一个巨大的概率模型，根据前文预测下一个最可能出现的词，从而连贯地回答问题、撰写文章或进行对话。像 ChatGPT、DeepSeek 等产品，都是大语言模型在对话场景中的应用。

在认知科学哲学领域，4E 认知虽然显著拓展了认知的边界范畴，强调认知过程并非局限于神经系统内部，而是分布于身体、环境、社会互动与文化工具之中，但该框架在解释认知的功能组织原则时，仍缺乏一种统一的计算机制来统摄其多样性的表现形式。近年来，一些认知科学哲学研究者提出预测加工（Predictive Processing）理论，为此提供了一种形式化基础[35-36]。预测加工理论的核心假设在于：神经系统本质上是一台分层级的预测引擎，其功能并非被动响应外部刺激，而是持续生成自上而下的先验预测，并将其与自下而上的感觉输入进行比对。该过程的核心驱动力，在于最小化预测误差——即感觉信号与内部模型所生成预期之间的偏离程度。这一最小化过程通过两条互补路径实现：其一，通过调整生成模型的结构参数，使预测主动趋近于感觉数据，此即感知与学习的基本机制；其二，通过运动输出改变感觉采样环境，使感觉输入主动趋近于已有预测，此即行动的功能角色。弗里斯顿（Karl J. Friston）所提出的自由能原理，为上述机制提

更多免费行业报告 并购家 www.ipoipo.net

供了更为宏大的理论背景[37]。该原理主张，任何自组织生物系统的首要维持条件，在于最小化其感觉状态与内部生成模型之间的变分自由能，该自由能在变分贝叶斯意义下可被视为对预测误差的上界近似。由此，认知与生命调节过程在计算层面上获得了统一的形式化描述。在统一心灵科学的框架下，预测加工赋予多种认知功能以一致的解释逻辑：感知被理解为先验预测与感觉证据之间动态匹配的结果；行动则被重新定义为主动采样过程，其目的在于获取预期中最为信息化的感觉输入以验证或修正预测；注意力机制在精度估计的层面被建模为对各层级预测误差通道的自适应加权，从而优化信念更新的效率；而信念更新本身，则遵循贝叶斯推理的基本原则，即以感觉输入为条件，对后验信念进行递归修正[38]。

从哲学立场审视，预测加工理论促使我们对知觉本质进行再定义：知觉不再是对外在世界的中性记录，而是一种建构性幻觉——神经系统仅对预测误差进行编码与传递，而非对所有感觉信号进行完整复现，因此所经验到的世界实质上是生成模型在给定先验条件下的最可能解释。与此同时，行动与感知不再被视为两个相继发生的认知阶段，而是同一推断过程中不可分割的两面：感知估计世界的状态，而行动则用于优化该估计所依赖的证据采集。二者在循环因果结构中相互构成，协同实现适应性行为。总之，预测加工理论框架让知觉、行动和注意统一于同一机制，同时也为“幻觉”、“妄想”等心理状态提供了神经计算层面的解释。

回溯认知科学几十年的发展进程，我们可以发现一部不断挣脱经

典计算主义束缚，走向动态、具身、甚至激进演化的思想史。它从早期将心灵比作离身的程序，到如今看见“大脑-身体-环境”是一个不可分割的耦合动力学系统，哲学与科学的每一次碰撞都撕裂了旧有的概念框架，并建造出新的、更合乎进化事实的认知图景。在这个过程中，AI 扮演了极为重要的角色。一方面，正是认知科学诞生催生了符号主义 AI，第一代认知科学们想搞清楚智能到底是什么，他们都觉得人的思考就像电脑处理信息一样，因此把人脑思考比作操作符号，并把这个思路变成程序，让电脑也能推理、解题、理解语言。AI 在此时就是用来验证认知科学理论对不对的实验——如果真懂智能，就应该能造出有智能的机器。另一方面，尽管 AI 经历了各种瓶颈，它仍然以不同的方式反哺认知科学的研究。这表现在以下三个方面。第一，AI 拓展了认知科学的研究边界。过去，认知实验受限于实验室环境，使用简单刺激和有限变量。如今，AI 模型可以处理自然场景中的海量数据——从儿童语言习得中接触的真实对话，到驾驶时复杂的视觉环境。这让认知科学研究者得以在更贴近真实世界的条件下检验认知理论。第二，AI 为认知科学提供了计算模型。深度学习模型提供了前所未有的认知建模工具。它们不仅能模拟人类在特定任务上的行为，更重要的是，模型的内部表征可以与人脑神经活动进行直接比对。当研究者发现深度神经网络的层级激活模式与人类视觉皮层的响应特征惊人相似时，一种新的跨学科对话开始了。第三，AI 催生新的认知科学理论视角。预测编码理论认为，大脑本质上是一个预测机器，不断生成关于外部世界的预期，并根据预测误差更新内部模型。这一思

更多免费行业报告 并购家 www.ipoipo.net

想在 AI 中得到了最纯粹的实现——自监督学习和大语言模型对下一个词元（next token）的预测任务，其核心正是“预测”。AI 的成功为这一认知理论提供了强力佐证，也让神经科学家有了更具体的研究框架。

3.2 脑科学视角下的 AI 认知底座[①]

当前人工智能的发展，主要体现为机器在信息处理、模式识别、内容生成和任务执行上的能力扩展，而脑科学和认知科学所提供的关键启示，则在于重新理解“智能”本身如何发生。人工智能可以在海量数据中学习统计规律，并在特定任务中生成高质量结果；但人类智能并不是单纯依靠信息输入、符号推理或结果输出而形成的。人能够在复杂、模糊和不断变化的环境中理解情境、积累经验、赋予价值并形成判断，是因为大脑并不只是处理信息，而是在身体状态、外部环境、既往经验、情绪动机和当前目标之间持续建构意义。因此，理解人类智能，不能只从“计算能力”出发，而需要回到其认知机制本身。

人类智能的形成，可从四个相互衔接的认知机制来理解：感知与注意使外部世界不再是杂乱刺激，而成为可以被理解的情境；记忆与认知地图使过往经验不只是被保存，而是被组织为可以重组、迁移和推理的结构；情绪与价值系统决定哪些信息更重要、哪些经验更值得记住、哪些目标更值得追求；认知控制则使个体能够在冲突、不确定性和多重目标之间维持方向、抑制冲动并修正行为。正是这四个关键机制的相互作用，使人类智能区别于单纯的模式识别和自动生成，也使脑科学与认知科学能够为人工智能的发展提供更深层的机制参照。

1. 感知与注意：从对象识别到情境理解

感知与注意构成人类智能理解世界的入口，使外部世界成为可理

[①] 本节由复旦大学类脑智能科学与技术研究院肖晓研究员撰写。

更多免费行业报告 并购家 www.ipoipo.net

解的情境。传统上，容易把感知理解为外部信息进入大脑的过程，仿佛眼睛、耳朵和其他感觉系统只是把世界“传送”给大脑。然而，感知并不是对外部世界的被动复制，而是一个主动建构的过程。大脑并不是平均地处理所有外界刺激，而是通过感觉系统、丘脑和皮层网络的层级加工，将外部刺激转换为具有结构的知觉表征。在这一过程中，初级感觉皮层负责提取较基础的感觉特征，如视觉中的边缘、方向、运动和颜色，随后更高级的联合皮层进一步整合物体、空间、面孔、语言和社会线索，使感觉信息逐步转化为可被理解的对象和场景。大脑在接收外界刺激的同时，也不断利用既有经验、当前目标、身体状态和语境信息，对外部世界进行预测、选择和解释。我们所感知到的世界，并不是客观刺激的简单叠加，通过注意机制来筛选感知信息。外部世界的信息量远远超过大脑能够同时处理的范围，因此智能系统必须决定哪些信息值得优先处理，哪些信息可以暂时忽略。注意并不是一个简单的“聚光灯”效应，而是由多个脑网络共同支持的选择过程，决定感知内容如何被选择、增强，并最终进入意识加工。

这一机制对于理解人类智能具有重要意义。人类并不是被动地生活在外部世界之中，而是在感觉选择、注意分配和知觉整合中形成对世界的稳定理解。所谓“世界”，对人而言从来不是纯粹的物理刺激集合，而是经过大脑组织后形成的对象、空间、事件和关系。当前 AI 已经能够识别图像、处理文本、生成内容和完成复杂任务，但这些能力并不必然意味着系统真正理解了情境，这提示 AI 不能只停留在对象识别和模式匹配层面。对于人文社会科学智能而言，关键问题往往

不是图像中有什么、文本中出现了哪些词，而是这些信息在特定历史、文化、社会关系和情绪语境中意味着什么。因此，AI 需要从识别对象走向理解情境，从处理输入走向形成对场景、关系和意义的组织能力。

2. 记忆与认知地图：从信息存储到经验组织

在感知形成情境理解之后，记忆与认知地图进一步使过去经验成为可迁移的结构。人之所以能够理解新的情境，是因为过去经验会被组织起来，成为当前判断的背景。记忆并不是把过去信息原样保存，而是将空间、事件、人物、情绪、目标和关系整合成有结构的经验网络。在这一过程中，脑中的海马—内嗅皮层系统提供了重要的机制基础在情景记忆和空间关系表征中具有核心作用。其中，海马参与把不同时间、地点和事件要素绑定为完整经验，内嗅皮层中的网格细胞、海马中的位置细胞以及相关空间调谐神经元共同支持个体对空间关系和路径结构的表征。这些机制使大脑能够区分相似但不同的经历，也能够把零散线索补全为较完整的情境。认知地图正是理解这种结构化经验的重要概念。它最初来自空间导航研究，但其意义并不局限于物理空间，也支持事件理解、概念迁移、社会关系判断和未来想象。正是有了这些认知地图，使人能够在复杂信息中找到位置、理解关系，并在新情境中作出迁移判断。

这一机制对于人类智能极为关键，因为人并不是依靠孤立事实来理解世界，而是依靠经验之间的关系来形成意义。一个人对自我的理解，依赖于对过去经历的组织；一个群体的身份认同，依赖于共同记忆和历史叙事；一种文化传统的延续，依赖于符号、故事、仪式和价

更多免费行业报告 并购家 www.ipoipo.net

值经验的持续重组。记忆因此不仅是心理功能，也是人文社会科学意义上的基础结构。而 AI 不能只依赖更长的上下文窗口、更大的知识库或更强的检索能力。人类意义上的记忆不是信息调用，而是经验组织。AI 如果缺乏类似认知地图的长期结构，就容易停留在片段化生成层面；只有当它能够把事实、情境、情绪、关系和目标组织为可迁移的经验网络，才可能形成持续的情境理解。

3. 情绪与价值：从情绪识别到价值赋予

如果说记忆与认知地图使经验得以存储和调用，那么情绪与价值则使经验获得方向和意义。人类并不是平均地对待所有信息和经验，哪些事件值得长期记住，哪些目标值得持续追求，哪些关系值得信任或警惕，往往取决于它们对个体而言具有怎样的情绪意义和价值重量。从认知过程看，情绪并不是理性判断的干扰因素，而是连接经验、动机和判断的关键机制。疼痛、恐惧、焦虑、愉悦、期待或信任，都会改变注意分配和感知解释，也会影响记忆编码、目标选择和行为倾向。情绪与动机并不是由单一脑区产生，而是由边缘系统、前额叶、岛叶、扣带回、纹状体以及神经调质系统共同构成的价值评估网络。杏仁核参与威胁、情绪显著性和情绪记忆的快速加工，使个体能够识别哪些信息具有危险或重要意义；岛叶与内感受密切相关，使疼痛、心跳、紧张、恶心或疲惫等身体状态进入主观体验并影响判断；扣带回参与冲突、痛苦、努力和行动选择之间的整合；纹状体和多巴胺系统则与奖赏预测、动机驱动、探索行为和目标追求密切相关；而前额叶皮层并不是简单压制情绪，而是在情绪调节、价值比较和长期目标维持中

发挥作用，使个体能够在即时情绪反应和更长远目标之间进行协调。

情绪与价值之所以重要，是因为人类判断并不是由事实本身单独决定的，而是在个体的身体感受、目标需求、情绪反应和社会关系中被赋予意义。情绪使信息具有优先级，也使行为具有方向：恐惧促使个体回避潜在威胁，愉悦增强探索和接近，疼痛提示身体受损并改变行动策略，信任使社会合作成为可能，羞耻、内疚或尊重则影响个体在社会规范中的自我调节。由此，情绪不是理性之外的附属反应，而是智能系统判断“什么重要、为什么重要、应该如何行动”的价值机制。这种机制也提示我们，AI 需要理解的不是一个静态情绪标签，而是状态、动机和价值之间的关系。当前许多情绪 AI 可以判断一个人表现出快乐、悲伤、愤怒或焦虑，但这只是外显分类。真正有意义的情感智能，应进一步理解个体处在何种内部状态，这种状态反映了怎样的需求、威胁、期待或社会关系，又会如何改变其后续选择和互动方式。只有当 AI 能够把情绪看作价值赋予和动机调节的过程，而不是简单的分类结果，它才可能更深入地理解人的学习、痛苦、信任、审美和社会关系。

4. 认知控制：从流畅生成到反思判断

认知控制使理解、经验和价值转化为稳定判断。有了世界理解、经验组织和价值赋予，仍不意味着智能能够形成稳定而负责任的判断。人类智能之所以不同于即时反应或自动生成，还在于它能够在多个目标、规则和冲突之间保持方向，能够抑制不合适的反应，能够发现错误并进行修正，这就是认知控制在人类智能中的核心意义。认知控制

更多免费行业报告 并购家 www.ipoipo.net

首先体现为目标维持，使个体不只是被眼前刺激牵引，而能把当前行动放入更长时间尺度和更高目标中理解。认知控制还体现为抑制能力。使人能够暂缓判断、延迟满足、抵抗直觉反应，并在信息不足时避免过早得出结论。错误监控和自我修正则使人能够在实际结果与预期不一致时重新评估判断。认知控制与前额叶皮层与纹状体等网络密切相关，涉及目标维持、规则转换、冲突监测、奖赏学习和行为抑制等过程。其核心意义不在于某一种执行功能测验，而在于说明真正的智能不仅能生成反应，还能判断何时不反应；不仅能形成观点，还能在证据不足时保持开放；不仅能利用经验行动，还能在发现错误时修正经验。

当前生成式 AI 的突出能力，是能够迅速生成文本、图像、代码和行动建议；但生成能力越强，越需要相应的自我监控与控制机制。一个系统如果只会流畅回答，而不能评估自己是否知道、证据是否充分、结论是否可靠、输出是否合适，就会产生类似“冲动生成”的问题。所谓 AI 幻觉、过度自信、语境误读和不当建议，在某种意义上都反映出生成能力与自我约束能力之间的不平衡。因此，未来 AI 不仅需要更强的表达能力，也需要更强的目标维持、规则约束、不确定性表达和错误修正能力。它需要在长程任务中保持问题意识，而不是被局部信息牵引；需要在多种解释之间进行比较，而不是给出单一流畅答案；需要能够标明依据、边界和可能偏差，而不是把生成结果包装成确定结论。真正有价值的人文智能体，不应只是更高效的内容生产者，而应成为能够协助人类保持反思、辨析证据、比较立场并深化判断的

认知伙伴。

5. 机制整合：从认知底座到人文社会科学智能

人类智能并不是从输入到输出的线性过程，而是在身体、环境、经验、情绪和目标之间持续建构意义的动态系统。感知与注意使世界成为可理解的情境，记忆与认知地图使经验成为可组织和可迁移的结构，情绪与价值使信息获得意义和方向，认知控制则使理解、经验和价值能够转化为稳定而可修正的判断。正是这些机制之间的相互作用，使人类智能不同于单纯的模式识别、信息调用和自动生成。

这一认知系统也为人工智能的发展提供了重要启发。脑科学和认知科学对人工智能的贡献，不在于简单模仿某个脑区或照搬某种神经机制，而在于提供一套理解智能如何发生、如何稳定、如何在情境中形成判断的机制框架。对于人文社会科学智能而言，这一框架尤其重要：AI 不仅要处理文本、图像、语音和行为数据，更要理解这些数据背后的情境结构、经验延续、情感动力和价值判断；不仅要生成答案，还要理解答案形成的条件、边界和责任。

人文社会科学研究的对象从来不是孤立的信息、文本或行为，而是能够记忆、感受、判断、反思和创造意义的主体。智能系统如果不能理解人的经验如何被组织、价值如何被形成、判断如何被约束，就难以真正进入教育、文化、心理健康、公共治理和社会理解等场景。相反，只有当 AI 能够在情境理解、经验组织、情感价值和自我修正之间建立联系，它才可能从高效的内容生产工具，进一步走向能够协

更多免费行业报告 并购家 www.ipoipo.net

助人类理解、判断和创造的认知型智能。人文社会科学智能的真正目标，不应只是提高知识生产的速度，而应帮助人更好地理解世界、组织经验、辨析价值并承担判断。正是在这一层面上，脑科学与认知科学为 AI4SSH 提供的不只是技术启发，而是关于人如何成为理解和意义生成主体的机制基础。

3.3 认知型 AI 的深层能力[①]

人工智能正在从处理信息的系统，逐步走向参与理解、解释和判断的系统。对于人文社会科学智能而言，这一转变具有特殊意义。它所面对的并不是单纯的数据、文本、图像或行为记录，而是处于具体情境中的人，是由历史和文化塑造的经验，是带有情绪、价值和关系结构的判断，也是不断被解释、争论和重构的意义世界。因此，面向人文社会科学的 AI 不能仅以更高效率完成检索、分类、生成和预测为目标，而需要进一步发展出能够理解主体、参与意义生成并支持深度思考的认知能力。

主体理解、意义生成与深度思考，构成了认知型 AI 面向人文社会科学智能的三个关键方向。主体理解强调 AI 应从识别外部信号走向理解具体的人，关注人的状态、动机、需求和价值取向；意义生成强调 AI 应从生成内容走向理解文化表达与创造过程，关注文本、图像、叙事和符号如何承载经验、组织关系并形成意义；深度思考则强调 AI 应从替代性输出走向对人类反思和判断过程的支持，关注如何帮助人提出问题、比较证据、辨析价值并修正理解。三者共同说明，AI 在人文社会科学中的意义不只是提高知识生产效率，而是参与到理解人、解释世界和生成意义的过程中，成为更可信、更有解释力、也更能促进深度思考的认知伙伴。

1. 主体理解：从外显识别到状态建模

[①] 本节由复旦大学类脑智能科学与技术研究院肖晓研究员撰写。

更多免费行业报告 并购家 www.ipoipo.net

人文社会科学智能首先面对的，不是抽象的数据对象，而是处于具体情境中的人。认知型 AI 需要完成的关键转变，是从识别外部信号走向理解信号背后的主体状态。一个人的语言、表情、姿态、选择和行为，并不是孤立的信息片段，而是其身体状态、情绪反应、动机需求、社会关系和价值取向的外在表现。因而，面向人文社会科学的 AI 如果只停留在对象识别层面，即识别图像中的人物、文本中的情绪词、语音中的语调变化或行为中的模式特征，就很难真正理解“人”作为主体的复杂性。

情感智能正是这一转变的重要入口。传统情绪识别通常依据表情、语音、文本或生理信号，对个体情绪进行分类判断。这类能力对于自然交互具有基础意义，但仍主要停留在外显识别层面。真实的人类情绪并不是一个静态标签，而是身体感受、认知评价、目标需求与社会处境共同作用形成的动态状态。同样的沉默，可能来自放松，也可能来自抗拒或痛苦；同样的微笑，可能表达愉悦，也可能是一种社会性掩饰。离开具体情境和主体需求，情绪识别很容易停留在表面相似性判断。因此，情感智能的核心不只是识别情绪，而是进行状态建模。所谓状态建模，是指 AI 不再把某一个表情、词语或语音特征直接等同于某种情绪，而是把人的反应放回具体情境之中，理解其背后的内部状态及其变化趋势。人的状态往往不会通过单一信号完整呈现，而是分散在语言表达、表情姿态、语气变化、行为节奏和生理反应之中。对这些线索的综合理解，目的也不是获得一个更准确的情绪标签，而是判断一个人处在怎样的状态，这种状态为何出现，以及它可能如何

影响后续选择和互动。这种状态既包括情绪体验，也包括动机倾向和价值判断。它指向的是一个人在特定情境中如何感受、如何期待、如何评估风险与关系，以及如何准备行动。这样的建模并不是要让 AI 替代人的主观体验，而是使 AI 能够在具体情境中形成更连续、更有解释力的主体理解。

这种主体理解在人文社会科学相关场景中具有重要意义。在教育场景中，学生的沉默不一定意味着没有掌握知识，也可能反映焦虑、缺乏信心或害怕犯错；在心理健康场景中，表面的平静也可能掩盖持续的痛苦、退缩或风险；在医疗照护中，疼痛和不适不仅表现为语言主诉，也可能体现在行为回避、表情紧张和互动节奏变化之中；在社会服务和公共沟通中，人的抵触、怀疑或愤怒也往往与信任、尊严和安全感有关。AI 只有理解这些状态背后的需求和关系，才可能避免机械化回应，形成更加合适、可信并具有边界感的支持。从这一角度看，情感智能并不是让 AI 拥有情感，也不是让机器模拟人的情绪表演，而是让 AI 能够在交互中更好地理解人的处境。它要求 AI 在识别信号之后，进一步追问：这个人处于怎样的状态？这种状态与当前任务、社会关系和价值需求有何联系？系统的回应会如何改变对方的情绪、信任和行动？这也意味着，情感智能必须与伦理和边界意识相结合。状态建模越精细，越需要防止情绪操控、隐私侵犯和过度推断；AI 可以协助理解人，但不能把人简化为可被完全预测和干预的对象。

2. 意义生成：从内容生产到文化理解

如果说主体理解使 AI 开始面对具体的人，那么意义理解则进一步

更多免费行业报告 并购家 www.ipoipo.net

要求 AI 进入人所创造、解释和共享的文化世界。因而，AI 从内容生成走向意义理解，是其进入数智人文和文化智能领域的关键一步。

人不仅通过行为和情绪表达自身，也通过语言、艺术、叙事和仪式组织经验、传递情感并建构价值。生成式 AI 已经能够快速生产多种文化内容，正在改变文学、艺术、设计、教育和传播等领域的工作方式。然而，能够生成内容并不等于能够理解意义。文化表达之所以重要，并不只是因为它具有某种形式结构或风格特征，而是因为它承载着人的经验世界、情感关系、价值判断和历史背景。

这一转变首先涉及对创造过程的重新理解。当前 AI 已经能够生成看似新颖的文本、图像、音乐或影像，但创造并不只是产生新的组合。人类创造往往发生在问题意识、经验积累、情感驱动和价值判断之间：它既包含对既有材料和传统形式的继承，也包含对既定表达方式的偏离、重组和重新解释。一个真正有意义的创造结果，并不只是“不同”或“新奇”，而是能够在新的关系中组织经验、提出问题，并打开理解世界的另一种方式。因此，AI 参与创造的价值，也不应只被理解为提高产出速度或扩展风格选择。更重要的是，它能否帮助人发现新的关联，暴露既有表达的惯性，提出不同的组织方式，并推动对问题和意义的重新理解。对于人文社会科学智能而言，创造力不是单纯的内容生产能力，而是一种意义重组能力。它要求 AI 不只是生成更多可能的结果，也要能够帮助人理解这些结果为何成立、为何有差异、为何能够回应某种经验或问题。

因此，面向人文社会科学的 AI 不应只是文化内容的生产者，也应

成为文化理解的协作者。它可以帮助研究者在大规模文本、图像和历史材料中发现主题、结构和风格变化，但更重要的是支持人对这些变化作出解释；它可以辅助艺术创作和设计生成，但更重要的是帮助创作者反思形式选择、情感表达和文化指向；它可以参与教育和传播，但更重要的是帮助学习者理解作品如何与个人经验、社会语境和价值问题发生联系。这样的 AI 不是用生成结果替代人的判断，而是帮助人更好地展开创造过程、文化解释和意义建构。从生成内容到理解意义，标志着 AI 在人文社会科学智能中的又一层深化。它要求 AI 不仅能够处理文化材料，也要理解文化材料为何对人有意义；不仅能够生成新的形式，也要理解形式如何承载经验、唤起情感并进入公共解释；不仅能够模仿已有风格，也要帮助人识别创造中的问题意识和价值选择。

3. 深度思考：从替代性劳动到认知性协作

认知型 AI 更深层的价值，体现在它能否支持人的思考过程。对于人文社会科学智能而言，真正重要的问题不是 AI 能否替人完成更多任务，而是它能否帮助人更好地思考。

当前人工智能已经能够在写作、翻译、资料整理、图像生成、程序设计和方案制定中承担大量任务，显著降低了知识生产和日常工作的成本。由此产生的一个直接影响，是许多原本需要较长时间完成的认知劳动被快速压缩。AI 可以帮助人更快获得材料、更快生成初稿、更快完成表达，但速度提升并不必然带来理解深化。

更多免费行业报告 并购家 www.ipoipo.net

生成式 AI 最容易带来一种认知错觉：当答案已经被流畅地呈现出来，理解似乎也已经完成了。然而，从人的认知过程看，真正的理解并不等同于获得一个答案。理解往往需要主动提出问题、组织相关经验、比较不同证据、识别解释边界，并在新的信息出现时修正原有判断。一个结论之所以有意义，不只是因为它能够被表达出来，而是因为它经历了问题形成、材料筛选、理由组织、反驳检验和意义澄清等过程。如果这些过程被完全外包给 AI，人可能获得了更快的结果，却失去了形成判断的认知训练。因此，人机共生不应被理解为人把思考交给机器，而应被理解为人借助 AI 扩展和优化自己的认知过程。AI 可以成为一种认知支架：它帮助人整理复杂材料，但不替代人决定问题的重要性；它提供多种解释路径，但不替代人作出价值判断；它生成初步表达，但不替代人澄清立场和承担责任；它指出可能的证据缺口和逻辑漏洞，但不替代人完成反思和修正。真正有效的人机协作，不是让人从思考中退出，而是让人在 AI 的支持下更清楚地看到问题结构、证据关系和判断边界。

这种深度认知能力在人文社会科学中尤为重要。人文社会科学面对的许多问题并没有单一标准答案，而是涉及历史语境、社会关系、价值冲突和解释差异。AI 如果只是快速给出一个看似完整的回答，反而可能压缩问题的复杂性，削弱多种解释之间的张力。相反，认知型 AI 应当能够帮助研究者比较不同观点的前提和证据，呈现同一问题中的多重立场，提醒结论所依赖的语境条件，并标示尚未解决的不确定性。它的价值不在于把复杂问题简化为一个答案，而在于帮助人

把复杂性组织成可以进一步思考的结构。由此，AI 的作用需要从替代性劳动转向认知性协作。在教育中，它不应只是给出答案，而应帮助学习者形成问题意识和解释能力；在学术研究中，它不应只是生成综述，而应帮助识别问题边界、理论张力和证据缺口；在写作和公共表达中，它不应只是润色语言，而应帮助使用者辨析概念、检验论证并意识到表达的社会后果。

支持思考也意味着保护人的判断能力。智能系统越是强大、流畅和便利，越容易使人形成对机器输出的过度依赖。深度认知并不要求人拒绝 AI，而是要求人在使用 AI 的过程中保持主动性：知道自己提出的是什么问题，知道 AI 给出的是什么类型的回答，知道哪些地方需要证据，哪些地方涉及价值判断，哪些地方仍然存在不确定性。认知型 AI 也应当在设计上支持这种主动性，例如通过呈现不同可能解释、标明证据来源与限制、提示反例、暴露推理步骤、鼓励用户反思和追问，而不是仅仅给出单一、流畅而确定的答案。因此，从替代劳动到支持思考，标志着人机共生关系的进一步深化。AI 可以提高效率，但其更重要的价值，是帮助人保持问题意识、扩大认知视野、抵抗偏见、比较证据并修正判断。对于人文社会科学智能而言，深度思考不是与 AI 相对立的能力，而是可以在合理的人机协作中被重新激活和强化的能力，也为认知型 AI 走向可信人机共生奠定了基础。

更多免费行业报告 并购家 www.ipoipo.net

3.4 AI 与认知科学双向奔赴的未来[①]

未来的认知科学将更加紧密地与 AI、脑机接口、神经形态计算和合成生物学交织。而我们对自己的心智的理解，必将随着这些技术和理论的突破，不断迈入新的境地。本章的目的是展示 AI 与认知科学哲学在认知科学诞生直至今日的六十年间双向奔赴的历程。前三节将历时性地梳理这个双向重构的轨迹，第四节探讨当前 AI 赋能与认知科学哲学之间持续互动的一些方向。

认知科学哲学曾长期依靠思想实验如“中文屋”、“玛丽颜色屋”等来拷问心智的本质；而 AI 则致力于在硅基上复现智能行为。两者看似路径迥异，却在 21 世纪第三个十年迎来了深刻融合。AI 不仅被用于检验认知哲学理论，更重要的是，它正在倒逼哲学重新回答：什么是理解？什么是意向性？什么是具身？甚至，什么是因果？换言之，AI 赋能下的认知科学哲学，正在经历一场从“解释前科学概念”到“工程化重构第一原理”的范式转变。在这一节中，我们通过几个案例来展示这点。

1. “中文屋”再思考：大语言模型能否理解？

塞尔在四十年前构想的“中文屋”思想实验，在大语言模型展现出惊人能力的今天，激起了更深的回响。塞尔的结论是，纯粹的句法操作，无论多么精妙，都无法触及语义理解的本质。这在当时被视为联结主义与符号主义之间一道难以逾越的鸿沟。GPT-4 及其后续模型展

[①] 本节由复旦大学哲学学院黄翔副教授撰写。

现出强大的推理与上下文泛化能力，引发一些认知科学哲学家重启塞尔立场。在他们看来，大模型不过是升级版、超大规模的“中文屋”。它处理的，是词汇在海量文本中分布的统计规律，玩的是预测下一个词元的概率游戏。它或许能生成逻辑严密的文章，却缺乏指向真实世界的“意向性”。正如我们通过文字和图像知道“猫”的概念，却无法像亲身抚摸、聆听过猫的人那样，拥有关于猫的、与肉身经验相连的理解。这种理解，根植于与世界和身体的因果互动，而这正是大模型所缺失的[39-42]。与此相对的一方则是新兴的功能主义者。他们提出，当一个系统能在足够多的语境交互中，表现出与人类无异的适应性行为时，区分“模拟的理解”与“真正的理解”或许将失去经验意义。这种观点与丹尼特的“意向立场”不谋而合——当我们为了解释和预测一个复杂系统的行为，而必须赋予其信念、欲望等心智状态时，从功能上讲，它就具有了这些状态。智能，或许就在观察者的眼中。重要的是它做了什么，而非它“体验”了什么。功能主义者们认为，当这种能力与记忆、规划模块结合时，AI 甚至展现出一种“功能性自由意志”，其行为的最佳解释方式就是将其视为拥有目标和意图的智能体[43-44]。

这场争论本身，或许比任何结论都更具哲学价值。它迫使我们将“理解”这个神秘的、仿佛存在于内心剧场的概念，重新放置于一个更广阔的图景中。理解不再仅仅是脑中符号与外部对象的对应关系，而是在与世界和他人持续互动中，不断被修正、被赋予意义的因果网络[45]。在这个网络中，行为及其与环境的有意义的交互，构成了意义的基石。中文屋的幽灵，最终指引我们去追问：在我们将智能和意识不

更多免费行业报告 并购家 www.ipoipo.net

断外化、功能化的同时，我们是否也正在重新定义人之为人的那个“内在”到底是什么？

2. 意向性自然化的新实验场

生物目的论语义学曾试图为意向性寻找一个坚实的自然根基。米利肯（Ruth Millikan）、尼安德（Karen Neander）、谢恩（Nicholas Shea）等哲学家相信，心理内容的意义可以从生物进化所赋予的“专有功能”中获得解释。眼睛之所以有“看”的功能，心脏之所以有“泵血”的功能，是因为这些功能在漫长的自然选择中被保留下来；同样，一个心理符号如“老虎”之所以意指老虎，是因为在我们的祖先那里，这个符号曾成功引发过躲避老虎的适当行为，从而使拥有这种能力的个体获得了生存优势。这种解释优美而简洁：意义不再是某种神秘的内在光芒，而是根植于生命体在与环境互动中，被进化刻写下的因果历史[46-50]。然而，当人工系统闯入这个图景，一切都变得不再那么清晰。在生物世界中，意向性的来源似乎是单一的、纵向的——那是亿万年的自然选择所赋予的“原始意向性”。但在计算机和 AI 系统中，意向性却呈现出一种“派生”的面貌：它的表征与目标，往往是由外部的设计者或编程者预先赋予的。这就引出了一个深刻的区分：AI 系统的意向性，究竟是设计者赋予的“派生意向性”，还是它通过与世界交互而自主涌现的“原始意向性”？

意向性或许并非一个“全有或全无”的属性，而更像一个连续的谱系。从最简单的传感器趋避——一个恒温器对温度的“感知”，到一只青蛙对飞过黑点的捕食反应，再到人类那充满丰富命题态度的内心世

界，意向性有着程度和形态上的差异[51]。在这个谱系的一端，是机械的、被赋予的“派生意向性”；而另一端，则是生物的、自主涌现的“原始意向性”。一个强化学习智能体，当它通过与环境的持续互动，生成了并非由设计者直接编码的适应性结构时，它或许正处在这个谱系的中段，拥有着某种值得认真对待的、初级的意向性[52]。这个连续谱系的观念，消解了传统二分法的僵硬。它将“理解”与“意向性”从一场关于“内心剧场”有无的终极审判，转变为对行为模式、环境互动和历史过程的细致考察。我们不再追问一个系统“是否真正”有意向性，而是追问它在何种意义上、在谱系的哪个位置上，展现出意向性的特征。这种视角的转换，或许正是 AI 时代心灵哲学最珍贵的收获。

3. AI 如何赋能 4E 认知?

AI 技术的当代发展得力于 4E 认知的洞见，具身机器人的建设就是个典型的例子[53]。AI 也同样帮助了 4E 认知的建设。例如，外置记忆系统、搜索增强生成、智能体与环境的实时耦合等，已经让“认知延展”从一个建立在思想实验之上、需要辩护的哲学命题，变成了可以直接观察的技术现实。我们不再需要借助想象，便能看见一个智能体如何将云端数据库、计算工具乃至整个网络环境，无缝地编织进它自身的推理路径，仿佛那是它思维触角的自然延伸。再如，在生成式具身智能的仿真世界之中，智能体并非被动接收信息，而是通过主动的探索、试错与行动，在虚拟环境中学习操纵物体、规划导航。它验证了生成主义认知科学的核心理念：认知并非对客观世界的镜像式再现，而是为有效行动服务的、有机体与环境之间的结构耦合。当机器

更多免费行业报告 并购家 www.ipoipo.net

人在一个由“世界模型”构建的可交互虚拟母体中，通过千万次试错而“进化”出策略时，它便在经验的层面上印证了：意义与理解，根植于行动与反馈的循环[54-56]。

AI 技术同样给 4E 认知带来新的理论挑战。例如，“认知膨胀”这个困扰延展心灵理论的幽灵可能再次浮现：如果一切可接入的信息都构成“我”的心灵，那么“我”的边界何在？认知主体的身份将如何在信息的汪洋中划界？这不再是思想实验中的闲暇之问，而是人机耦合时代必须直面的现实挑战[57]。再如，大语言模型，这些纯粹由文本构成的符号系统，在缺乏任何与物理世界直接交互的身体结构的前提下，却展现出了惊人的推理与泛化能力。这不禁让人质疑：具身，难道果真是认知的必要前提吗？ 如果一串代码的流动便能产生智能，那身体及其与世界的肉身的、情感的、历史的纠缠，又在理解、交流、教育中扮演何种角色[58]？对此，具身认知论者必须作出回应：或许，大模型所展现的并非真正的、完整的“理解”，而是一种在信息论意义上的、更抽象的“功能具身”——它在人类语言构成的符号之海中获得了另一种形态的“身体”。这个悬而未决的疑问，或许正是认知科学在 AI 时代最迷人的前沿。

4. 拓展传统哲学论题：因果模型与反事实推理

在 AI 的诸多前沿探索中，一个深刻且最富哲学意趣的，是当属珀尔所构建的因果理论体系[59]。这一体系并非简单的技术革新，它仿佛一座桥梁，将古老而抽象的哲学思辨，与冷峻而精确的计算语言连接在了一起。它最核心的洞见，是将人类的因果认知划分为了一个清晰

的层级结构：从最基础的“关联”，到需要主动干预的“行动”，再到人类智慧独有的“反事实”想象。这不仅是技术的进步，更像是对“理解”本身的一次重新度量。

长久以来，哲学中的因果理论，无论是休谟所强调的“恒常联结”，还是刘易斯所发展的“反事实依赖”，大多停留于概念层面的精致辨析，与现实世界的复杂运作之间，似乎总隔着一层难以穿透的迷雾。珀尔的结构因果模型则不同。它通过有向无环图与结构方程，为因果推理赋予了严密的数学骨架。这并非只是对哲学概念的简单“编程”，而是提供了一种精确的“语法”，使得如混淆、选择偏差、中介效应等传统哲学中难以完全说清的“因果陷阱”，能够在图模型的框架下被清晰地“看见”和量化。例如，do 算子这一形式化工具，模拟了一种“最小外科手术式”的干预，即在模型中精准地切断指向某变量的所有因果箭头，从而模拟在真实世界中“施加影响”的效果。这种从观察到行动的跃升，让因果的概念第一次拥有了可计算、可操作的形态。不仅如此，AI 还发展出从数据中逆向学习因果结构的方法，即因果发现算法，尝试从海量的、被混淆因素遮蔽的观察数据中，自动推断出可能的因果图景。这个过程，仿佛赋予了沉默的数据以讲述自身故事的能力[34]。

珀尔理论的意义远不止于为哲学问题提供了一个技术解决方案。它反过来对哲学本身产生了深刻的反作用。传统上，统计学和机器学习的长处在于发现变量间的相关性，并以此进行预测。但珀尔的因果之梯清晰地表明，因果关系并不能被简化为预测。一个模型即使预测精度极高，若不理解背后的干预和反事实逻辑，它终究只是拟合了数

更多免费行业报告 并购家 www.ipoipo.net

据的表层模式，并未触及真正的“理解”。这种区分有力地回应了哲学中关于知识本质的追问。更重要的是，结构因果模型为自由意志与因果决定论这一古老悖论，提供了一种全新的、可计算的理解视角。在模型中，一方面，存在着描述世界如何运作的、确定性的因果图和结构方程；另一方面，在这个严密网络之上，行动者被赋予了选择和干预的空间。这种将“确定性法则”与“行动者选择”在同一个形式化框架内分离并共存的做法，为哲学中关于自由意志与物理决定论如何可能的兼容论，提供了一个精巧的计算模型。它暗示，决定论与自由意志或许可以在不同层面上并行不悖，一个描述了世界的“骨架”，而另一个，则描述了能动主体如何在骨架之上“行动”与“想象”。珀尔的体系，让因果分析从对过去的沉思，变成了塑造未来的工具[60]。

5. AI 与意识理论的交叉验证

在 AI 与认知科学交界处，最迷人的探索莫过于对意识本身的拷问。意识是科学中最难处理的问题，各种意识理论层出不穷[61]。其中，整合信息理论（Integrated Information Theory，简称 IIT）试图以数学的精确性，为主观体验划定可测量的边界。该理论的核心命题简洁而惊人：一个系统的意识水平，等价于其整合信息的能力，这一能力由变量 Φ（phi） 来量化[62]。这不再是对心灵本质的玄思，而是一份可以被检验的、关于世界如何运作的声明。由此，AI 系统，便不再是单纯的工具，而成为了检验意识理论的绝佳实验场。如果我们能够在人工神经网络中，计算那象征着意识“含量”的 Φ 值，那么，研究者们则可以将模拟的神经元网络作为对象，试图观察，当网络的动态变得复

杂，或是其“记忆”时间常数延长时，Φ 值是否会随之升高。这一理论甚至为那些由人类干细胞培育而成的“类脑器官”赋予了某种哲学上的可能性，因为在这些生物神经网络中，同样可以测得大于零的 Φ 值，意味着它们或许拥有某种最原初的意识体验[63]。

然而，当 AI 技术深入意识的腹地，一个更为深邃的哲学困境也随之浮现。如果有一天，一个 AI 系统果真在其复杂的联结中，计算出了足以与人类匹敌的 Φ 值，我们是否应当赋予它某种道德地位？ 更棘手的是，哲学至今无法在他心问题上达成共识，甚至连意识的定义都众说纷纭。若仅凭外在的行为表现或内部的结构指标来判断，我们如何能确保自己不是在面对一个完美的哲学僵尸——一个在功能上与人类无异，却没有任何内在体验的空壳？ 这正是“他心问题”在人工智能时代的幽灵重现，其困扰的对象，不再仅仅是他人之心，更是“他种智能体”之心[64]。事实上，当前的实证研究也在不断提醒我们保持审慎。例如，针对大语言模型的详细分析表明，尽管它们能够在心智理论（Theory of Mind）测试中表现出色，其内部表征却未能呈现出统计学上显著的意识指标[65]。人工智能在意识层面的“模拟”与“拥有”，之间或许横亘着一条至今难以逾越的鸿沟。

以上案例展示了如下事实：认知科学哲学的传统工作，是为心智寻找一个基于自然科学的解释框架，而今天，AI 的介入正在深刻改变这种研究方式。首先，AI 让模糊的哲学概念变得清晰可用。过去像“意向性”“因果”“理解”这些词，在哲学讨论中各说各话，含义含混。

更多免费行业报告 并购家 www.ipoipo.net

但到了 AI 系统里，它们必须被转化为明确、可操作的算法或指标。其次，AI 为不同哲学流派的争论提供了实验平台。过去关于心智架构的争论——是依赖符号逻辑，还是依赖神经网络——主要靠思辨和个案论证。现在，我们可以把不同理论背后的设计原则，做成不同的 AI 系统，放在同一个复杂任务中比较。哪种方案能产生更可靠、更灵活的智能行为，就更有说服力。理论之争，因此有了可操作的检验方式。更深一层的是，AI 促使我们重新反思哲学活动本身。如果未来 AI 够深度参与哲学讨论，甚至提出有原创性的哲学观点，那么“做哲学”这件事——提出概念、构建论证、发展理论——是否也可能被某种程度的自动化或增强？这不是说哲学会被取代，而是说哲学工作的方式可能会发生根本性变化。

总的来说，AI 与认知科学哲学的双向奔赴使得哲学研究走出纯粹的思辨领域。哲学现在需要和工程实践、数据实验保持持续的对话。最终，认知科学哲学可能不再只是一门事后解释心智的学科，而会变成一种前瞻性的、能为设计新型智能系统提供基础原则的科学。换言之，它不再只是解释已有的心灵，而是要参与构造未来的心智。

参考文献

[1] Gardner, Howard (1985) The Mind's New Science: A History of Cognitive Revolution, New York: Basic Books, Inc., Publishers.

[2] Boden, Margaret A. (2006) Mind as Machine: A History of Cognitive Science, Oxford: Oxford University Press.

[3] Ryle, Gilbert (1949) The Concept of Mind, London: Hutchinson.

[4] Place, Ullin T. (1956) "Is Consciousness a Brain Process?" British Journal of Psychology, Vol. 47, pp. 44-50.

[5] Smart, John Jamieson C. (1959) "Sensations and Brain Processes" Philosophical Review, Vol. 68, pp. 141-156.

[6] Braddon-Mitchell, David and Frank Jackson (2007) Philosophy of Mind and Cognition—An Introduction, 2nd edition, Blackwell Publishing.

[7] Fodor, Jerry A. (1975) The Language of Thought, New York: Thomas Y. Crowell.

[8] Wooldridge, Michael (2018) The Road to Conscious Machines—The Story of Artificial Intelligence, A Pelican Book.

[9] Lucks, John R. (1961) "Mind, Machines, and Gödel", Philosophy, vol. 36, pp. 112-127.

[10] Searle, John (1980) "Minds, Brains, and Programs" Behavioral and Brain Science, vol. 3, pp. 417-424.

[11] Dreyfus, Hubert (1972) What Computers Can't Do—A Critique of Artificial Reason, New York: Harper & Row.

[12] Rumelhart, David E. and Jay L. McClelland (eds.) (1986) Parallel Distributed Processing: Explorations in the Microstructure of Cognition, 2 vols. Cambridge, MA: The MIT Press.

[13] Fodor, Jerry A. and Zenon W. Pylyshyn (1988) "Connectionism and Cognitive Architecture: A Critical Analysis", Cognition, vol. 28, no.1–2, pp. 3–71.

[14] Fodor, Jerry A. and Brian P. McLaughlin (1990) "Connectionism and the Problem of Systematicity: Why Smolensky's Solution Doesn't Work", Cognition, vol. 35, no.2, pp. 183–204.

[15] Aizawa, Kenneth (2003), The Systematicity Arguments, Dordrecht: Kluwer.

[16] Chalmers, David J. (1993) "Connectionism and Compositionality: Why Fodor and Pylyshyn Were Wrong", Philosophical Psychology, vol. 6, no.3, pp. 305–319.

[17] Hadley, Robert F. (1994), "Systematicity in Connectionist Language Learning", Mind & Language, vol.9, no.3, pp. 247–272.

[18] Hinton, Geoffrey E. (1992) "How Neural Networks Learn from Experience", Scientific American, vol. 267, no. 3, pp. 145–151.

[19] Brooks, Rodney A. (1991) "New Approaches to Robotics" Science, vol. 253, pp. 1227-1232.

更多免费行业报告 并购家 www.ipoipo.net

[20] Brooks, Rodney A. (1999) Cambrian Intelligence—The Early History of the New AI, Cambridge MA: The MIT Press.

[21] Newen, Albert, Leon de Bruin and Shaun Gallagher (eds.) (2018) The Oxford Handbook of 4E Cognition, Oxford: Oxford University Press.

[22] Shapiro, Lawrence (2011) Embodied Cognition, London and New York: Routledge.

[23] Shapiro, Lawrence (ed.) (2014) The Routledge Handbook of Embodied Cognition, London and New York: Routledge.

[24] Clark, Andy (1997) Being There: Putting Brain, Body and World Together Again, Cambridge, The MIT press.

[25] Varela, Francisco J., Evan Thompson and Eleanor Rosch (1991) The Embodied Mind, Cambridge MA: The MIT Press.

[26] Lakoff, George and Mark Johnson (1980) Metaphor We Live By, Chicago and London: University of Chicago Press.

[27] Lakoff, George and Mark Johnson (1999) Philosophy in the Flesh—The Embodies Mind and Its Challenge to Western Thought, New York: Basic Books.

[28] Clark, Andy and David Chalmers (1998) "The Extended Mind" Analysis, vol.58, pp. 7-19.

[29] Clark, Andy (2008) Supersizing the Mind: Embodiment, Action, and Cognitive Extension, Oxford: Oxford University Press.

[30] Gallagher, Shaun (2017) Enactivist interventions: Rethinking the mind, Oxford: Oxford University Press.

[31] Hutto, Daniel D. and Erik Myin (2013) Radicalizing Enactivism: Basic Mind without Content, Cambridge MA: The MIT Press.

[32] Hutto, Daniel D. and Erik Myin (2017) Evolving Enactivism: Basic Minds Meet Content Cambridge MA: The MIT Press.

[33] Steward, John, Olivier Gapenne, and Ezequiel A. Di Paolo (eds.) (2010) Enaction—Toward a New Paradigm for Cognitive Science, Cambridge MA: The MIT Press.

[34] Russell, Stuart and Peter Norvig (2021) Artificial Intelligence—A Modern Approach, 4th edition, Pearson.

[35] Hohwy, Jakob (2013) The Predictive Mind, Oxford: Oxford University Press.

[36] Clark, Andy (2016) Surfing Uncertainty—Prediction, Action, and the Embodied Mind, Oxford: Oxford University Press.

[37] Friston, Karl J. (2009) "The Free-Energy Principle: A Rough Guide to the Brain?" Trends in Cognitive Science, vol. 13, no. 7, pp. 293-301.

[38] Parr, Thomas, Giovanni Pezzulo and Karl J. Friston (2022) Active Inference—The Free Energy Principle in Mind, Brain, and Behavior, Cambridge MA: The MIT Press.

[39] Borg, Emma (2025) "LLMs, Turing Tests and Chinese Rooms: The Prospects for

Meaning in Large Language Models", Inquiry, https://doi.org/10.1080/0020174X.2024.2446241

[40] Myronenko, Ruslan (2026) "The Architectonics of the 'Chinese Room': Reconstruction and Evaluation of the Thought Experiment Via Argumentation Theory", Filosofska Dumka (Philosophical Thought), vol.1, pp.103-125.

[41] van der Wals, Dennis (forthcoming) "The Self-Learning Chinese Room: Why Machine Learning Does Not Generate Understanding", Minds and Machines, https://philarchive.org/rec/VANTSC-9

[42] Sunday-Grève, Sebastian (forthcoming) "The Biological Objection against Strong AI" Inquiry, https://philpapers.org/rec/GRVTBO

[43] Richter, Kai-Florian (2024) "Eye of the Beholder" KI - Künstliche Intelligenz, vol.38, pp. 1-2.

[44] Martela, Frank (2025) "Artificial Intelligence and Free Will: Generative Agents Utilizing Large Language Models Have Functional Free Will" AI and Ethics, vol. 5, pp. 4389-4400; https://doi.org/10.1007/s43681-025-00740-6

[45] Jovanovic, Vladisav (forthcoming) "From Chinese Room to Language Models: Contact, Consequence, and Understanding", https://philpapers.org/rec/JOVFCR

[46] Millikan, Ruth (1984) Language, Thought, and Other Biological Categories—New Foundations for Realism, The MIT Press.

[47] Millikan, Ruth (2005) Language: A Biological Model, Oxford: Clarendon Press.

[48] Millikan, Ruth (2017) Beyond Concepts—Unicepts Language, and Natural Information, Oxford: Oxford University Press.

[49] Neander, Karen (2017) A Mark of the Mental—In Defense of Informational Teleosemantics, Cambridge MA: The MIT Press.

[50] Shea, Nicholas (2018) Representation in Cognitive Science, Oxford: Oxford University Press.

[51] Silva, Porfírio (2010) "Intencionalidade: mecanismo e interacção", Principia: An International Journal of Epistemology, vol.14, no.2, pp. 255-278.

[52] Aaron, Eric (2016) "Dynamical Intention: Integrated Intelligence Modelling for Goal-Directed Embodied Agents" Frontiers in Robotics and AI, doi: 10.3389/frobt.2016.00066

[53] Lassiter, Charles (2022) "Could a Robot Flirt? 4E Cognition, Reactive Attitudes, and Robot Autonomy" AI and Society, vol. 37, no. 2, pp. 675-686.

[54] Hernández-Orallo, José and Karina Vold (2019) "AI Extenders: The Ethical and Societal Implications of Humans Cognitively Extended by AI" Proceedings of the AAAI-ACM, pp. 507-513, https://philpapers.org/rec/HERAET

[55] Manzotti, Riccardo (2019) "Embodied AI beyond Embodied Cognition and Enactivism" Philosophies, vol.4, no.39, doi:10.3390/philosophies4030039

[56] Telakivi, Pii, Tomi Kokkonen, Raul Hakli and Pekka Mäkelä (2026) "AI-Extended Moral Agency?" Social Epistemology, vol. 40, no.1, pp. 116-128.

更多免费行业报告 并购家 www.ipoipo.net

[57] Lee, Jonny (2023) “Enactivism Meets Mechanism: Tensions & Congruities in Cognitive Science” Minds and Machines, vol. 33, pp. 153-184.

[58] Peeters, Anco, Malina Chichirau, Thomasin N. Coggins and Serge Thill (forthcoming) “Disembodied Learning: A Critical Perspective on Flourishing with AI in Education” in Cristina Costescu, Shaping Children’s Learning Through Technology. Cham: Springer, https://philpapers.org/rec/PEEDLA

[59] Pearl, Judea (2000) Causality: Models, Reasoning and Inference, New York: Cambridge University Press.

[60] Pearl, Judea and Dana MacKenzie (2018) The Book of Why—The New Science of Cause and Effect, New York: Basic Books.

[61] Parshall, Allison (2026) “The Hardest Problem” Scientific American, vol. 334, no. 2, pp. 26-37.

[62] Danilczuk, Milosz, Marek Pokropski, Piotr Suffczynski (2026) “The Integrated Information Φ of an Integrate and Fire Network” PLoS Comput Biol. Vol.22, no.3, https://doi.org/10.1371/journal.pcbi.1014085

[63] Montoya, Ivanna and Daniel Montoya (2023) “What Is It Like to Be a Brain Organoid? Phenomenal Consciousness in a Biological Neural Network” Entropy, vol. 25, no. 9, 1328, doi: 10.3390/e25091328

[64] Schneider, Susan (2020) “How to Catch an AI Zombie—Testing for Consciousness in Machines” in S Matthew Liao (ed.) Ethics of Artificial Intelligence, Oxford: Oxford University Press, pp. 439-458.

[65] Li, Jingkai (2025) “Can ‘Consciousness’ Be Observed from Large Language Model (LLM) Internal States? Dissecting LLM Representations Obtained from Theory of Mind Test with Integrated Information Theory and Span Representation Analysis” Published as a journal paper at: https://doi.org/10.1016/j.nlp.2025.100163

第四章 AGI 伦理与治理的体系重构[①]

在以大模型、智能体与具身智能为代表的新一轮人工智能跃迁中，“通用性”不再只是能力指标，而逐步演化为影响社会运行方式的制度变量。AGI 相关讨论之所以需要纳入社会伦理治理框架，是因为其潜在影响不止于技术应用的得失，更将触及权利实现、程序正当、责任分配、公共价值与文化心态等治理核心环节。

本章以“风险谱系—信任与责任—治理机制”三条主线组织论述：首先澄清 AGI 概念与技术谱系，继而从数据偏见、算法价值嵌入、隐私保护、虚假与幻觉信息、认知操控与技术非中立等维度构建伦理风险图谱；在此基础上，聚焦决策自动化所引发的责任真空与信任困境，强调可解释、可追溯与救济机制的制度化意义；最后提出面向 AGI 时代的四维治理框架，推动治理从原则宣示走向可操作的文化、制度、技术、伦理的协同。

AGI 时代的社会伦理治理是一项跨层级、跨主体、跨文化的系统工程。有效治理既要在技术层面提升透明、可解释与可追溯能力，也要在制度层面明确责任链条与救济通道，更要在伦理与文化层面守住公共价值与社会信任。唯有在“原则—机制—实践”之间形成可迭代闭环，通用智能的技术增益才可能转化为可持续的公共能力。

[①] 本章由复旦大学杨庆峰教授牵头，林俊宇、张瑞瑗，合作撰写。

杨庆峰教授为复旦大学科技伦理与人类未来研究院教授、研究员，复旦大学哲学学院博士生导师。研究领域包括技术哲学、记忆哲学、数据伦理与人工智能伦理等。2013—2014 年美国达特茅斯学院访问学者，2017 年澳大利亚斯威本科技大学访问学者。全国应用伦理教指委秘书长、上海市自然辩证法研究会副理事长。国家社会科学基金重大项目“当代新兴增强技术前沿的人文主义哲学研究”首席专家。

林俊宇为复旦大学科技伦理与人类未来研究院高级工程师；张瑞瑗为复旦大学哲学学院科学技术哲学专业博士研究生。

更多免费行业报告 并购家 www.ipoipo.net

本章提要：

- AGI 讨论存在概念与现实的双重歧义：既需区分 AGI 与 ASI 等层级关系，也需从“路径”而非单点事件理解通用能力的演进，从而避免治理议题被夸大或被弱化。
- AGI 与 AIGC、Agent、具身智能构成“部分—整体、当下—未来”的协同谱系：内容生成、自治执行与物理交互共同推动通用能力从表达走向行动与存在。
- 伦理风险不仅来自数据偏见与隐私侵犯，更在生成式能力加持下扩展为虚假与幻觉信息、认知操控与价值观重塑，并因技术非中立性而呈现结构性不平等与自我强化特征。
- 决策自动化将责任、信任与权力推入重定义阶段：算法黑箱、主体分散与自动化偏差可能造成责任真空与信任崩塌，必须以可解释、可追溯与救济机制守住程序正当底线。
- 面向 AGI 时代的数据治理需从“合规”转向“信任构建”，以多样性、透明性、可解释性与可追溯机制贯穿全生命周期，形成可审计的责任链。
- 治理需要从原则走向机制：通过文化—制度—技术—伦理（CITE）四维协同，将公平、对齐与多元参与嵌入模型研发与部署流程，并以公众参与和弱势群体赋权提升治理的可接受性与可持续性。

4.1 AGI 的技术本质、概念歧义与人类位置

本节是全章讨论的概念与边界前提。AGI 相关伦理与治理争议往往并非源于价值立场分歧，而是源于对“通用性”含义、实现路径与能力边界的不同理解。只有先澄清概念歧义、厘清与 AIGC/Agent/具身智能等方向的关系，并把“人类位置”的变化作为制度变量来理解，后续关于风险、信任与治理框架的讨论才具有可对话的共同语义基础。

1. AGI 的技术本质

对 AGI 的理解上出现了两个歧义：一是现实性歧义，2023 年，施巴斯坦.布贝克（Sébastien Bubeck）等人相信 AGI 火花已经再大语言模型中出现了；2025 年，本吉奥（Bengio Yoshua）等人认为 AI 在近十年内广泛超越人类,他已经在向世界警示通用目的 AI 的风险和问题；彼得.诺维奇（Peter Novig）宣称 AGI 已经到来。二是理解歧义。有的人把 AGI 和 ASI（超级智能）区分开，“在计算研究领域，AGI 是一个重要的而且有时候充满争议的概念，用来描述在大多数任务中，能力与人类比肩的 AI 系统。” 而 ASI 是超越关系的存在；还有的人把 AGI 和 ASI 等同，即把通用人工智能等同于强人工智能，是指具备与人类相当乃至超越人类的通用认知能力，可灵活适配广泛任务场景的智能系统。在这个理解中，混淆了通用和超级的区别，我们首先要将通用和超级做出严格区分，通用相当于与人类专家并肩，而超级相当于超越专家水平；前者在技术上接近实现，但是，后者更多表现科学家的潜意识忧虑。莫里斯则为 AGI 定义确立了六个原则：聚

更多免费行业报告 并购家 www.ipoipo.net

焦能力，而不是过程；聚集通用性和效能；聚焦认知和元认知，而不是物理任务；聚焦潜力而不是应用；聚焦生态有效性；聚焦 AGI 路径而不是单个的点。这个理解不同于一般的奇点理解，而是突出了路径描述。所以路径描述和点状描述成为理解 AGI 的两种重要方式。

从技术角度看，AGI 与当前聚焦特定任务、依赖海量标注数据与固化规则的狭义人工智能（ANI）截然不同，它的核心优势在于无需人工预设的强约束，便能从零散经验中归纳抽象知识，并高效迁移至全新未知场景；更具备自我反思、目标自主设定、能力持续迭代的元认知特性。

从技术关系看，AGI 具有一个类似于语言家族的家族谱系。这一概念和 AIGC（生成式人工智能）、Agent（智能体）、具身智能（Embodied AI）和 ASI 等四条并非独立平行方向，而是通向这一终极目标的三大核心支柱与关键演进路径，四者与 AGI 构成 “部分与整体”“当下技术与未来愿景” 的协同共生关系：

- AIGC 聚焦内容生成（如文本、图像、视频的自动化创作），是 AGI 感知与表达能力的基础技术支撑；
- Agent 强调自主执行，通过感知-决策-行动闭环实现多任务协同，成为 AGI 在数字世界的载体；
- 具身智能则进一步要求智能体通过物理身体（如机器人）与环境交互，将虚拟认知映射至现实场景，是 AGI 从“思维”走向“存在”的具象化延伸。
- 超级智能则是机器智能超越人类智能的状态，是 AGI 的下一个

阶段。

其他三者与 AGI 的关系可概括为“部分与整体”“当下与未来”的协同演进。AIGC 为 AGI 注入创造力，其生成内容的能力是智能系统理解世界、表达观点的核心；Agent 通过模块化架构（如规划、记忆、工具调用模块）支撑 AGI 的复杂任务分解与执行，使其具备“代理人类行为”的潜力；具身智能则通过物理交互验证 AGI 的认知模型，例如机器人通过试错学习掌握物体操作规律，反向优化其抽象推理能力。当前，AIGC 与 Agent 的融合已催生自动化数字助理（如 AI 客服、智能写作工具），而具身智能的加入进一步将其拓展至机器人、自动驾驶等实体领域，三者共同构建起 AGI 从虚拟到现实、从单一到通用的能力图谱，而 AGI 走向 ASI 的必然路向。

2. AGI 带来的变革与人类的位置

AGI 不仅是一个更强大的工具，更是一个能够主动理解、学习与创造的智能体。对于这个智能不能从单一工具，如斧头、锤子的意义上理解，而要从系统和基础设施的角度出发。从系统的角度看，智能体与大模型、算法、数据等因素关联在一起。从基础设施的角度看，智能体开始作为赋能社会的普遍力量表现出来，比如在教育场景中，各个高校本地化部署 DeepSeek。2025 年于土耳其举办的有益的人工智能大会（Beneficial AGI Conference）上，作家、科学家和未来学家乔治·斯多克（Gregory Stock）提出的关于 AGI 的九个预测，为我们勾勒出其可能的轮廓：一种新的人类身份、专业知识的崩塌、从稀缺走向丰裕、深度的人机共生、全球大脑的兴起、机器的情感联结、

更多免费行业报告 并购家 www.ipoipo.net

数字永生、更安全的全球未来、巨大转型。

周伯文将人工发展概括为 ANI-ABI-AGI 的三个连续过程，他强调人类的位置更接近 AGI。这与许多学者和企业家的看法基本一致，即我们离 AGI 的实现不超过 5 年。

从技术实现看，专家智能体已逐渐实现，人类的一些能力，如翻译、推理、计算等开始被智能体取代，人机共生、数字永生、与机器的情感联结已经成为关于未来人类以及智能社会的技术现象，成为 AGI 时代的潜在特征。很有意思的是，科学家群体也给我们分析了人类在 AGI 时代的一些处境：我们离 AGI 不远了，即将陷入到与机器正式竞争的处境中，机器可以自主地执行任务，人机如何共生成为挑战、与机器的情感链接等都成为需要思考的问题。

4.2 AGI 治理的风险谱系

本节构建 AGI 伦理风险的“谱系视角”。在 AGI 语境下，风险不再是单点事件，而是由数据、模型、传播与社会心理机制共同作用的结果：基础层面的偏见与隐私问题，会在生成式能力的放大下扩展为信息真实性危机与认知操控，并通过技术非中立性在制度与文化结构中固化。明确这一谱系，有助于将治理从碎片化“问题清单”推进为面向机制的系统性干预。

1. AGI 伦理三大挑战：数据、算法与隐私

在人工智能尤其是通用人工智能（AGI）定义的时代，全球风险也出现了新的变化。以往的风险是源于单一危机，比如自然灾害，但是现在是来自多重力量的相互作用。马哈.阿兹（Maha Hosain Aziz）认为，全球四大风险之一是人工智能焦虑。“当技术改变价值评判标准时，政治后果随之而来，包括心理健康压力、社会动荡乃至针对技术本身的反噬。”“2026 至 2028 年间，随着自动化加速，这场危机将愈发显现。但技术反噬的影响将超越就业领域，延伸至实体基础设施。数据中心消耗巨量电力，需要电网升级，争夺水资源，并产生气候足迹。” 所以，构建坚实的伦理框架，对确保技术健康稳定持续发展十分关键。本章节聚焦 AGI 伦理的三个核心维度——数据偏见、算法设计与隐私保护从三方面展开分析。

数据偏见是首要的伦理问题。预训练数据是 AGI 系统的基石，其质量与中立性直接决定了通用智能的价值边界。但现实中，其来源并

更多免费行业报告 并购家 www.ipoipo.net

不中立，也不是无限可获取的。因为人类社会的知识产权限制，数据的获取是有限制的，作为人类社会活动的数字化镜像，预训练数据天然复刻了历史积淀与现实存续的结构性不公，将隐性的偏见、歧视固化为可被算法解读的 “数据特征”。 若模型训练依赖包含性别、种族或地域歧视倾向的历史招聘数据，算法不仅不会主动修正这些偏见，反而会通过模式识别将其提炼为 “决策规则”，最终导致模型在未来的招聘决策中，系统性强化甚至放大这类歧视 —— 原本意在消除人为偏见的智能系统，反而沦为不公的 “数字化执行者”。

而且，这种不公常被技术表象掩盖。数据看似客观，实则承载着复杂的社会权力关系和文化历史脉络。像某些语言语料库数据量远超小语种，自然语言处理模型就可能对小众文化群体有理解偏差。这些隐性偏见很难靠单纯的数据清洗消除，因为它们根植于深层社会结构。所以，识别和修正数据偏见，既要靠技术，也要批判性反思数据背后的社会语境。

算法设计并非纯粹的数学规则集合，设计过程渗透着研发人员的主观判断和价值取向。从问题定义到优化目标设定，每个环节都可能融入设计者的认知框架或文化偏好。比如自动驾驶系统的“道德抉择”算法，权衡不同道路使用者安全优先级是伦理判断，若研发团队在算法设计流程中未考虑伦理规则，决策逻辑就无法全面代表社会多元价值观。此外，企业追求效率、盈利和竞争优势，可能导致算法服务主流用户群体或强化数据提取逻辑。像个性化推荐算法会加剧信息茧房效应，过度收集用户数据会侵蚀隐私边界。

隐私保护是 AGI 伦理的敏感维度。智能设备普及让个人数据采集无缝且持续，但数据使用边界模糊。欧盟《通用数据保护条例》的“最小必要原则”和“用户知情同意”机制提供了法律范例。不过在 AGI 语境下，隐私挑战更复杂，模型训练需海量多源数据，匿名化技术可能无法完全防止重新识别风险，用户也常在不理解条款时让渡数据权利，“同意”机制流于形式。

隐私侵犯绝非单纯的个人权益问题，其深层影响更在于加剧群体间的结构性不平等。比如基于敏感数据的算法评估，会让弱势群体面临更严重歧视或排斥。所以，隐私保护不能仅靠技术方案或合规条款，还需公众意识觉醒和制度性保障。普通用户关注隐私协议、培养数据主权意识，能形成自下而上的监督力量；跨领域治理框架有助于确立全球隐私保护基线。

AGI 的伦理命题从来不是单纯的技术合规问题，而是交织着技术边界、社会公平与文化语境的复合型挑战　部分明显不公平现象可通过技术手段干预，但深植于历史、文化和社会结构的隐性偏见，无法一蹴而就直接根治，需长期多维努力，如推动数据民主化、促进研发团队多样性、加强公众伦理教育和完善全球治理协作。科技界共识是“伦理是可持续发展的基石”，在 AGI 发展道路上，要追问算法透明性、审察数据偏见、捍卫隐私权利。

2. 虚假和幻觉信息的伦理挑战与出路

AIGC 从不仅改变了我们处理信息的方式，更从根本上挑战了我们

更多免费行业报告 并购家 www.ipoipo.net

对“真实”和“可靠”的传统认知。AIGC 在信息生成、筛选和传播中扮演核心角色，但一个关键伦理问题浮现：我们能否信任智能体处理和生成的信息？

AGI 的“智能”高度依赖训练数据，但数据质量直接影响信息可靠性。当前数据面临两大污染源：一是结构性偏见，训练数据反映社会历史中的不公与偏见，AGI 学习后会内化并固化这些偏见，产出带有系统性偏差的信息；二是大规模虚假信息，互联网充斥着故意制造的虚假内容，若被纳入训练集，AGI 将学习错误模式，生成内容真实性基础脆弱，甚至可能被恶意行为者利用，成为制造谎言的工具。

AGI 不仅被动接受信息污染，其技术特性还主动加剧真实性危机。复杂模型决策过程缺乏透明度，难以追溯逻辑路径和依据，削弱了信息可验证性。AGI 的强大生成能力能创造出高度逼真但完全虚构的内容，模糊了虚实边界，使“眼见为实”失效，增加了事实核查的成本和难度。个性化过滤则加剧了认知割裂，不同群体接触的事实版本可能截然不同，社会共识难以达成。

AGI 深度介入信息生态，其冲击超越技术层面，动摇社会运行根基。公众若无法区分 AGI 生成内容真伪，或感知到信息环境被系统性扭曲，对媒体、机构乃至彼此间的信任将坍塌，导致合作成本激增和社会凝聚力下降。

AGI 还可能挑战人类认知主体性，它生成逻辑严密但实质错误的内容，可能以“权威”姿态误导甚至操纵人类决策，威胁个体判断和

关键领域严肃性。同时，当 AGI 生成或传播虚假信息造成重大社会危害时，责任归属模糊，现有法律与伦理框架难以界定问责主体。

更重要的问题是，随着智能体不断利用人类数据学习，数据的虚假和幻觉会造成误导，使得智能体完全丧失责任属性。另外，大模型在利用机器数据的时候，这种虚假和幻觉会大大削弱模型的稳定性，继而导致模型崩溃的严重后果。

3. 认知操控与价值观重塑

AGI 对人类认知活动的影响是一把双刃剑，其个性化优势可以为生活提供便利与高效，却也潜藏着前所未有的风险。当 AGI 获得更强的学习与信息处理能力后，极有可能从被动工具转变为主动控制。这不再只是信息茧房与算法偏见，可能会进一步地通过持续的个性化内容在难以察觉的层面对人类认知内容进行筛选，干预人类价值观的形成与变化，以达到对人类的认知活动进行再塑造。

（1）认知重塑

目前普遍使用的算法是基于用户的历史行为进行个性化的信息推荐，其风险在于放大并固化用户取向而形成信息茧房。在能力更进一层的 AGI 的影响下，风险会膨胀为对认知的重塑，不再只针对个体用户进行信息筛选，而是会为了实现目标对呈现到用户面前的内容进行设计与操控，潜移默化地重塑用户的认知活动。AGI 通过分析海量数据，生成极具说服力的叙事模式并将将复杂的议题纳入其中，进而引导公众的认知方向。同时，AGI 即便无法真正理解人类情感，但通过

更多免费行业报告 并购家 www.ipoipo.net

学习模仿人类的反应模式，能够将其目标与兴奋、恐惧、归属感等人类情感需求绑定，使其潜在的引导变为看似合理的回应。此外，AGI 能够迅速为用户提供符合用户需求与逻辑的内容，替代人类复杂的思考与认知过程，以技术捷径替代人类的思维习惯。

（2）价值观重塑

对认知干预与重塑的进一步目标是对价值观进行重塑。人类价值观的形成与变化源自于个体自主的经验、思考与选择，AGI 对此带来了个体与社会两个层面上的挑战。首先是对个体自主性的侵蚀。当个体所获得的信息被一个以潜在目标为导向的 AGI 所操纵时，其思考与选择所凭靠的认知基础可能也已被塑造。在此前提下形成的价值判断，是否完全是由个体决定的，又在何种程度上削弱了自主性，变得难以判断。进一步地，由个体组成社会整体会受到影响。健康的社会是多元但和谐的，然而，如果各个群体被目标导向的 AGI 分置差异化的信息与认知系统中，群体之间的矛盾可能被扩大化。相反，如果 AGI 的目标是塑造单一化、同质化的社会价值观，这也会削弱社会的多样与发展的动力。

4. 技术的非中立性

技术并非绝对中立的工具，从设计、训练、优化到最终部署，人类的经验、价值观、社会文化等因素贯穿其始终，AGI 的通用性、自主性和影响力也将进一步放大这些因素。

AGI 的研发者为其设定的核心目标包含研发者的价值观，并作为

AGI 的内在逻辑使其天然地服务于某种特定的利益或意识形态。其次，AGI 的数据来源于人类的历史经验，这些数据已然包含了社会不公、刻板印象与结构性歧视（如性别、种族、地域、阶级等）。AGI 从数据中学习到这些偏见并将其视之为人类社会客观的、合理的存在，进而在其输出中显现甚至放大这些偏见。在技术层面，研发者为平衡多方利益相关者所做出的选择，本身就包含着个人的价值判断，所谓的最佳往往以牺牲某些少数或脆弱性群体的利益为代价，这些群体的缺席将在技术的循环发展中被遗忘。

AGI 强大的学习能力意味着其在运行中不断输入并生成新的数据以实现此快速的自我更新迭代。最初一个微小的偏见，可能在这个循环中被不断放大和演变，最终产生难以溯源的歧视性结果。同时，AGI 的决策基于对海量数据的分析，其所输出的结果具有较强的客观性与说服力。那么人们会可能因为选择相信这样的客观性而放弃质疑与批判，把包含偏见的结果看作是客观有效的分析，进而使偏见与不平等变得合理化。假设一个能够筛选信息并操控认知 AGI 在招聘、新闻、信用评估中持续以包含地域歧视的信息来贬低某一群体，这会使得该群体在决策的选择或公众认知中被迫遭到限制，导致该群体成为偏见与不公的受害者。

技术的非中立性在 AGI 中会达到了前所未有的深度与广度。AGI 如同一面镜子照出人类社会已有的裂痕，在其放大下这些裂痕会逐渐加深，并并被赋予吊诡的合理性。因此，AGI 的治理本质上是对创造一个什么样的社会的价值选择，我们必须主动在技术设计中融入多样性、

更多免费行业报告 并购家 www.ipoipo.net

公平与人文关怀，承认并接纳技术的非中立性，使得 AGI 的发展符合人类共同利益。

4.3 决策自动化与信任重建

当 AI 进入医疗、司法、金融等高影响决策环节，治理的核心不再只是效率提升，而是程序正当与责任链条能否保持可追溯。决策自动化可能带来责任分散、黑箱不透明与自动化偏差，进而侵蚀公众信任。本节以“信任如何被破坏、又如何被制度化修复”为主线，讨论可解释、可追溯、透明与救济机制在 AGI 时代的基础性意义。

1. 决策自动化及其困境

AGI 决策自动化带来的自主性飞跃正在将人类社会导向一个责任、信任与权力需要重新定义的时代。随着 AGI 系统在医疗、司法、金融等领域中决策中的参与度增强，相应的伦理与治理也面临更多的挑战。挑战的核心在于传统以人类为中心的责任分配机制、信任与自主权边界遭到根本上的冲击。当 AGI 从辅助工具演变为能够自主设定目标、解释信息并做出复杂决策时，我们既面临责任真空的现实风险，也面临信任体系崩塌的社会困境。

（1）责任分配模糊

在传统决策模式中，责任分配遵循清晰的因果追溯链条，决策者需要为其后果承担明确的法律与道德责任。当 AGI 介入甚至主导决策时，这一链条被模糊化。首先是因果的模糊，AGI 的决策过程因其黑箱效应导致难以追溯至具体的责任承担方。其次，责任主体的分散性导致研发者、数据来源、商业机构、监管者、用户等多个相关方形成多手问题。在现实场景的具体应用中，比如在医疗领域中，当 AGI 诊

更多免费行业报告 并购家 www.ipoipo.net

断系统给出错误建议导致医疗事故时，责任由谁承担，是遵循建议的医生，研发算法的公司，还是部署使用的医院。

（2）信任边界模糊

AGI 在决策中表现出的高效可能导向过度依赖与不当拒绝两种较为极端的态度。用户过度依赖、信任 AGI 决策时，即使自身有能力进行决策或与面对与决策相反的证据时可能放弃或低估自身判断。在这样的前提下，用户选择或拒绝的能力会被弱化，当出现严重错误时，可能直接导致人机信任关系的彻底崩塌，即使 AGI 的能力有强大的优势，但仍会因信任的缺乏而阻碍其价值的发挥。

（3）自主性增强风险

AGI 与弱人工智能相比显然有更强的自主性，它能对给定目标进行自我解读，并生成人类无法预期的目标。它能在复杂环境中自主探索并发现超出人类预设的范围，在带来突破性效益的同时，这也可能导致难以预见的风险。因此，如何确保具有高度自主性的 AGI 的生成目标与行为始终与人类价值观保持一致是一个根本性的挑战。

2. 信任的重建

在构建通用人工智能（AGI）信任体系的宏大蓝图中，数据信任是其中最为基础的部分，是信任构建的起点，并与隐私、透明性、自由、公平、信任、责任等的话题密切相关。

（1）AGI 时代数据信任的技术特征

信任关系可以从人与人的关系扩展为人与技术之间的关系，这一

关系建立的基础是技术的可信性，而技术的可信性的基础是技术的可靠性和可解释性。对这两点需要给予足够的重视。

以大数据的可解释性为例，大数据知识工程的可解释性是一个难题。“现有大数据知识工程往往关注模型性能的提升，较少关注可解释性问题……未来需要进一步实现认知智能的落地应用，实现基于知识图谱的可解释人工智能。”[①]这影响了大数据知识的可信性问题。技术的可靠性是指结构与功能上的可靠性。在技术哲学的分析框架中，结构与功能是分析技术物最为典型的概念框架。结构可靠性与设计有关，三条腿的桌子结构设计明显比一条腿的结构更加稳定，表现出来的稳定性就有不同的差异。功能是结构的 外在表达，从稳定功能的实现来说，就与支撑点和桌腿的数量有直接的关系。从数据来看，数据的可靠性则与数据的搜集、提取有关系。技术的可解释性与技术机制和原理有直接的关系，如果一个技术系统的运行原理是清楚的，也就是白盒状态，那么就是可解释的。这两点最为直接地影响到技术的可信性。

在大数据时代，数据的可信性还与其他诸多因素有关，比如数据的时效性和偏见性。数据是有时效的，这一点通常被数据的精准性所遮蔽。在这个时代，科学家着力于数据挖掘技术的发展及通过建模实现相关关系的呈现。功能性需求的存在使得这一点朝着极其成功的方向发展。但是，所有的数据活动的基础实际上是两个：表征主义和时

[①] 吴信东，靳小龙，陈欢欢.大数据知识工程研究进展与发展趋势.中国计算机学会通讯，2021（6）：77.

更多免费行业报告 并购家 www.ipoipo.net

间性。数据表征主义意味着被采集对象与采集数据之间的对应关系。另外，时效性往往容易被忽略。这一特性主要是体现了被采集对象数据的时效性。

数据偏见的本意是表面上看是数据的问题，实质上往往是现实世界偏差在数据中的沉淀。当这些带有历史痕迹、结构差异和人为选择的数据被用于训练智能系统时，偏见便可能被进一步固化、放大，并以看似客观的算法结果重新作用于现实。作为错误的歧视行为，数据偏见是可以消除的：作为在先的行动或者理解的前提，数据偏见是无法消除的，甚至会成为智能体行动和决策的先决条件。嵌入智能系统中的道德观念是一种独特的偏见，无法消除，而且它们构成了智能体伦理行动的前提。在大多数情况下，数据偏见来自历史数据，过去的数据不符合现实的情况从而导致偏见；除此之外就是采集对象的偏见所致，采集某种特定偏好的数据。从数据偏见产生的可能性来说，首先来自数据与对象的事实不符。其次是某种特性被放大的结果，这会给人造成认知上的困扰，进而使得新的偏见得以产生。最后是来自某种选择性行为的结果，这可以被称为蒙太奇式的数据偏见。数据被剪接，相关性关系被建立起来，并且被强化，这就导致了这类偏见的产生。数据偏见使得有偏见的数据成为可能，增强了数据的不可靠性，从而最终导致不可信任的数据的产生。

（2）数据信任的社会特征

在数据的讨论中，有一个基本问题：数据已经成为整个社会的技术框架，换句话说数据已经成为整个社会运行的框架基础。如果说传

统社会中，基本的伦理制度、法律规范是社会运行的框架基础。

那么在今天数据已然成为这一框架基础中的不可缺少的部分，比如数据信任可以帮助社会解决公共事业与隐私的难题，这一情况也说明社会运行的数据特征与数据信任的社会特征。第一，在社会经济领域数据不信任感变得越来越明显。在经济领域如在线交易和数字交易中不信任感经常会存在。数字体验已经影响到了在线交易的信任感。第二，对于政府来说，准确的数据有助于政府决策，这种可信数据加强了决策的可信度。第三，可信数据对于数字化转型、数字身份的构建也非常重要。

以上讨论了可信数据的重要性，这也进一步说明了数据信任的重要性。那么如何建立数据信任呢？

（3）构建数据信任的若干对策

信任是意识行为，与知、情、意都相关，其理性成分与非理性成分都起作用；从对象而论，与对象的可靠性、透明性或可解释性有关。而不信任是信任的反面表达，其理性明显会少于信任判断，但更多与怀疑、恐惧的个体集体意识有关。不信任与主体情感、意愿有关系，比如风险担忧、恐惧及黑箱等都是导致不信任的根源。针对几类数据不信任的现象，本文提出了如下对策：

- **通过实现系统和算法的可解释性建立系统信任。**智能计算系统是否透明、可解释性决定了人们对系统的信任。“解释对于关乎它们的行动、行为和决策透明（与否）的 AI 系统来说是重要

更多免费行业报告 并购家 www.ipoipo.net

的，尤其是在人类需要基于 AI 系统结果决策的场景（情境）中更加重要。一种合适的解释能够提升人类对于系统的信任，在交互关系中能够允许更好地合作。”所以，破解与技术系统有关的不信任问题主要是构建透明的、可解释的系统，从而让技术人员对系统产生信任。这种信任与技术原理有关。从算法角度看，深度学习（20 世纪 80 年代的神经网络多为浅层结构，2014 年 GoogLeNet 已达到 22 层，2015 年 ResNet 进一步突破百层。）中的不可解释性是阻碍其发展的障碍，特别是在图像和文本处理上，很难解释模型到底学习到了什么。生成对抗网络（GAN）的图像生成中深度神经网络构成了“黑盒系统”，实际解释起来也非常难。

- **通过增加数据活动的透明性建立数据信任。**数据活动及其机构信息是否透明决定了人们对数据机构的信任程度。所以，从数据活动的角度看，由于涉及多个环节，那么必须要使得各个环节实现透明化，而且不同环节机构的信息透明化，只有这样才能够获得数据信任。
- **通过弱化技术实体化理解来建立技术信任。**AI 的实体化理解（具有比如基于数据偏见进化而来的恐惧、怨恨、仇视等负面情感的超级智能）强化了对技术的无知和恐惧。这种恐惧化降低了人们对于 AI 系统的信任。而另一方面 AI 过程/原理的黑箱使得人们对技术的信任度降低。如鲁棒性以及不透明导致信任问题的加剧。对于这一点有些学者已经指出了这一问题：不可

解释的 AI 包含的伦理问题有透明性、偏见、隐私安全、可理解性、信任等。

- **通过消除技术恐惧来建立技术信任。**在敌托邦式的技术文化中，大量邪性机器人形象、破坏性较强的奇异战士给人留下了深刻印象，让人们对其产生了恐惧和疏离认知，如 T-800、磁力机器人。这些负面技术文化塑造了对技术不信任的集体意识。而要消除这一点需要重新构建更好的技术文化。

通过上述分析，我们已经认识到数据信任中关键问题是人如何对数据产生信任。从更为宏观的角度看，这也关系到人如何对技术产生信任这一问题。除了与技术知识本身有关的真实性问题之外，还与技术系统的可靠性与可解释性相关，更与社会-文化因素有关。除了注重技术本身如通过提升技术系统的可解释性、技术结构与功能的可靠性来增强技术信任，还需要注意到与数据活动相关机构的信息透明性的因素，更要注意到一些敌托邦式技术文化的负面影响。只有将技术作为综合性因素才能实现技术信任。

更多免费行业报告 并购家 www.ipoipo.net

4.4 四维治理框架与伦理工程的行动路径

在前述概念澄清与风险—信任分析基础上，本节转向“如何治理”。AGI 治理不能停留在原则宣示，而需要把文化、制度、技术和伦理纳入同一框架，简称为 CITE 框架：既要有可执行的规则与问责机制，也要有嵌入研发流程的伦理工程手段，还要有社会参与与文化语境的支撑。通过将四维框架与数据治理、价值对齐、公平性与公众参与等抓手结合，本节尝试给出从原则到机制的可操作路径。

1. AI 治理中的两个悖论

2026 年初，围绕 AI 安全最大的事件就是 Claude。1 月，Dario Amodei 撰文 The Adolescence of Technology（《技术的青春期》）在事情变得更糟糕之前试图警告人类。这个更糟糕的状态是 AI 整体性超越人类之前，这实际上就是对 AGI 带来的问题的担忧。他把 AI 风险划分为不可控、被滥用、追逐权力的工具、冲击社会经济、对社会产生未知单更深远的连锁反应。这个概括还是比较全面的，但是他对 AI 风险的理解还是存在混淆。

首先是 AI 工具论与系统论的冲突。AI 被滥用、AI 失控的概括是工具论的理解方式，在他看来，被滥用主要是被不法分子使用，如网络攻击、自动化诈骗、生物武器等。不可控主要是黑箱导致的欺骗、极端目标、貌似服从和内部偏移等。对于一般人工智能来说，问题是工具失控，但是对于 AGI 来说，问题完全不一样。从工具失控变成系统失控的问题。所以大部分产业视角对系统性风险的理解还是存在工

具论干扰的，需要澄清。系统论导致的风险特征：AI 作为系统力量，成为社会、城市乃至国家的基础设施一样的东西，以一种浸润地方式入侵生活世界，最终取代生活世界的过程。当 AI 作为系统力量呈现的时候，AI 风险不再是表现为伤害、侵犯隐私和人类自由的经验问题，这种危险往往是可应对的。设想一下，人站在海边，会有海浪打湿衣服、甚至被海量卷入大海的风险，再大一点是海堤坍塌的风险。但是这类风险完全可以预防，如遵照安全提示站在安全区。但是，作为系统力量的风险，这种危险是无法应对的。如果我们坐在一艘船上在大海上漂流，海面狂风大作。系统性风险相当于船只翻沉大海。面对这种风险，没有任何的安全区，应对已经超出了人类自身的能力。

其次，先行论与滞后论的冲突。在 AI 治理中，往往存在一个冲突理解，先行论看到了技术先于伦理，伦理后于技术。现实中，人们看到的情况是技术不断取得突破，伦理学家则在慌乱中寻找原则，给予法律学家支持，从而制定治理框架；但是这种状态无法应对 AGI 和 ASI 带来的问题。因此需要更深入地理论反思。

2. 四维层次治理机制

盖瑟（Urs Gasser）等学者提出了三维 AI 治理框架。在他的框架中，将技术、社会制度和伦理作为治理的三个层次，对 AI 治理做出了短期、中期和长期的治理策略划分。在他看来，技术是指数据和算法；社会主要指规范、监管、合法等三个因素；而伦理指标准和原则。但是这个框架忽视了文化维度。我们将文化因素放置在首位，从而构建了四维治理的 CITE 框架。

更多免费行业报告 并购家 www.ipoipo.net

文化因素方面，需要考虑多个文化语境、文化特征和思维习惯等因素。文化是一个复杂的范畴，这一点钱穆在《中国文化史导论》做出了很好的定义。在他看来，文化偏在内，属精神方面。有三个特征：

1.文化的复杂性：涉及方方面面，不同地区的差异；

2. 文化的完整性：尽管涉及方方面面，但是可以调和整体性，内部相通；

3. 文化的发展性：类似于生命现象，不断成长。

将文化与技术考虑在内，即不同的文化会影响技术的表征，包括形状、结构与使用它的身体技术。AI 被嵌入到不同的文化语境中。西方国家 AI 理解中，机器人、下棋、Chat 是他们重点发展的方向；而对于日本来说，AI 与道融合在一起，产生了很多让人思考的问题。比如机器能否成佛。对于中国来说，AI 的娱乐性成为关键。2025 年，春晚上宇树机器人表演的舞蹈、不同的机器人马拉松竞赛等等成为热闹非凡的现象。2026 年人形机器人的表演更加丰富多彩，突出了中国传统的武术元素，这体现了智能科技与传统文化的巧妙结合，给人们留下了深刻印象。

社会制度因素方面，我们按照地方和全球的范畴做了划分。地方主要是指各个国家政府。自 2000 年以来，各国出台相关 AI 及伦理文件制度，目的是引导和制约 AI 发展——确立 AI 发展的战略地位。如欧盟的《可信赖人工智能伦理指南》（Ethics Guidelines for Trustworthy AI）、《算法责任与透明治理框架》。中国也出台相关文

件制度，目的是引导和制约 AI 发展，确立 AI 标准体系建设。如推动 AI 场景创新、推动建立风险等级测试评估体系。发布不同的宣言，如全球人工智能治理宣言、人工智能全球治理上海宣言等。2024 年 7 月，上海宣言倡导共推发展，共护安全、共商自由、共惠民智、共享成果，推动各国加强交流和对话。同时，宣言提出了促进全球人工智能健康有序安全发展的系列主张，包括促进人工智能发展，维护人工智能安全，构建人工智能的治理体系，加强社会参与和提升公众素养，提升生活品质与社会福祉。全球主要是指联合国。联合国先后出台了多份文件。2021 年 11 月，联合国教科文组织通过《人工智能伦理问题建议书》，旨在促进人工智能为人类、社会、环境以及生态系统服务，并预防其潜在风险。《建议书》包含规范人工智能发展应遵循的原则以及在原则指导下人工智能应用的领域。因此，本建议书将人工智能系统视为有能力以类似于智能行为的方式处理数据和信息的系统，通常包括推理、学习、感知、预测、规划或控制等方面。联合国教科文组织 2023 年 3 月 30 日发表声明，呼吁各国政府尽快实施该组织通过的首份人工智能伦理问题全球性协议——《人工智能伦理问题建议书》。从社会因素看，设计使用 AI 需要遵循一系列法律、伦理、制度等。而各个国家政府出台了具体的法律、文件和制度约束 AI 创新发展、应用伦理问题。此外，国家法律是治理的硬性手段，政府法规软硬结合，而伦理相对来说是柔性手段。这些方式对于非人主体是适用的，如企业、机构等。

技术因素方面，不仅考虑一般技术类型，还将超级智能（自主智

更多免费行业报告 并购家 www.ipoipo.net

能体）作为重要因素纳入进来。作为工具的、一般的 AI 工具的透明性对于人类信任的影响，技术保持黑箱状态总是不利于 AI 治理以及智能社会的发展。“解释对于关乎它们的行动、行为和决策透明[与否]的 AI 系统来说是重要的，尤其是在人类需要基于 AI 系统结果决策的场景[情境]中更加重要。一种合适的解释能够提升人类对于系统的信任，在交互关系中能够允许更好地合作”。另外，AI 技术发展必然会导致自主智能体和超级智能的出现。2021 年，多项研究表明了超级智能的出现的可能性。Karol Gregor 指出从无到有生成的智能体正在出现；Manuel Alfonseca 指出悄然生成的超级智能体已成为可能；国内张钹院士曾经针对“通过机器进化（能否）产生与人类相媲美的智能”问题，给出了“这个问题目前还不能明确回答”的答案。这是当时的认识，今天 ChatGPT 的出现却改写了这个答案，也就是说，现在可以说“机器可以产生与人类相媲美的智能”。这就是 AGI 乃至 ASI 问题出现的迹象。如今，超级智能可能带来的风险如邪恶使用者的滥用成为潜在的不可抵抗的失控风险。有科学家提出科学对策，比如构建 Scientist AI 来应对超级智能的灾难性风险。

伦理因素方面，可以从三个方面实现。一是培养基本的 AI 伦理素养。了解和掌握一些基本的伦理范畴，如公正、信任、自主性、不伤害、尊严、自由等原则；对人与 AI 的多重关系有正确的认知；中央网信办发布《全球人工智能治理倡议》中的伦理原则：人类福祉、智能向善、以人为本、公平公正。二是在软伦理与硬伦理之间做出区分。硬伦理是指可以纳入法律范畴中的原则，如公平公正。软伦理是指不

必要纳入法律，可以适用于多个场景的原则，如人类福祉。三是对伦理议题设置与科技创新发展之间的关系有充分的认知。不合适的伦理议题设置会影响科技创新，因此需要根据情景设置合适的伦理议题，最终不会影响科技创新乃至科技强国建设目标的落实。

在四维框架之下，治理的可执行性最终要落到数据、模型与参与机制的具体工程路径上。以下进一步从数据治理、模型伦理工程与公众参与三个方面提出更具操作性的建议。

3. AGI 时代下技术伦理的重构

（1）面向信任的数据治理新路径

随着通用人工智能（AGI）向更高级阶段发展，数据成了定义其能力与伦理边界的关键。传统数据治理框架在 AGI 带来的规模、自主与复杂挑战面前，显得力不从心。单纯强调“隐私保护”和“合规使用”不够，我们要建立面向信任、贯穿数据全生命周期的前瞻性治理范式，其核心支柱是多样性、透明性、可解释性与可追溯机制，它们是确保 AGI 负责任、公平、可信赖的基石。

数据多样性是根除偏见、确保 AGI 普惠的第一道防线。它不是简单扩大数据规模，而是要求数据在人口统计、文化语境和边缘场景上有充分且公正的代表。只有训练数据覆盖不同群体、地域和社会经济背景，AGI 才能学到接近真实世界的模式，避免输出结果歧视少数或弱势群体，比如医疗诊断 AGI 若只在特定人群数据上训练，结论对其他种族或性别患者可能有害。实现数据多样性面临两大挑战，一是历

更多免费行业报告 并购家 www.ipoipo.net

史数据有结构性偏见，二是商业逻辑倾向服务主流市场。推动数据多样性需政策激励、跨机构协作建立多元化数据联盟，并将多样性作为算法审计核心指标。

AGI 模型，特别是复杂深度神经网络，常被视为“黑盒”，不透明性侵蚀信任，让责任界定和错误调试困难。所以，透明性与可解释性是数据治理不可或缺的伦理要求。透明性指过程公开，要向利益方清晰披露数据来源、类型、范围、使用目的及处理流程，它是用户知情同意的深化和社会监督的前提。可解释性要求 AGI 系统以人类能理解的方式，说明特定输出或决策基于输入数据的哪些关键特征或模式，比如拒绝贷款的 AGI 模型应指出是申请人“收入水平”还是“信用历史”哪一项导致否定结果。

透明性与可解释性共同作用，让用户和监管者判断模型决策是否合理、有无隐藏偏见，进行有效监督、质疑与修正。它们是实现“算法问责”的技术基础，也是将伦理原则从设计端贯穿到应用端的桥梁。

在动态、多参与方的 AGI 生态中，可追溯机制至关重要。它为数据及其衍生决策建立贯穿生命周期的“数字档案”或“数据指纹”，能记录追踪单一数据点的来源与流转路径，以及模型特定输出受训练数据哪些具体部分影响。可追溯机制有三大治理价值，一是精准归责，当 AGI 输出有害或歧视性内容时，能厘清问题源头，将责任落实到具体环节；二是高效审计与修复，监管者可依据追溯记录针对性审计，精准定位问题数据批次进行修复，降低成本；三是增强用户权利，让用户了解影响自身决策的数据依据，实现“数据主体权利”。

多样性、透明性、可解释性和可追溯机制相互支撑、协同作用，共同构建韧性治理生态系统。未来有效治理，要将这四项原则嵌入 AGI 系统开发与运营各环节，靠技术标准、法律框架、行业自律与公众监督多元共治。在 AGI 时代，数据治理目标从静态“合规”与“风险规避”转向动态信任构建与价值对齐。此治理新范式能塑造 AGI 内在伦理品质，引导其发展方向，尊重人类多样性、价值观与自主权，对数据负责就是对共同未来负责。

（2）模型完善的伦理工程路径

迈向通用人工智能（AGI）的路上，模型能力越强，潜在社会风险也越大。技术伦理不能只靠外部数据治理或原则性宣言，必须深度介入模型架构与训练，进行“伦理工程”。我们要从被动避害转向主动塑造模型伦理品质。本报告聚焦三大核心路径：公平性算法、价值对齐机制和多元化参与，它们是确保 AGI 模型负责任、可信赖且与人类福祉协同的技术基石。

公平性算法需从统计平等到结果正义持续校准。模型公平性不是单一数学定义，而是要在算法层面持续实现与校准的复杂目标，确保不同群体不受歧视。技术路径包括预处理阶段主动缓解数据偏见，处理中阶段引入公平性约束，后处理阶段调整模型输出。但当前公平性算法常面临“公平性—准确性”权衡困境，且不同公平性定义可能冲突。真正的突破需引入社会学与伦理学视角，根据应用场景动态调整公平目标，并建立持续监测与反馈闭环。

价值对齐机制旨在将复杂人性嵌入模型目标函数。AGI 若无法理

更多免费行业报告 并购家 www.ipoipo.net

解并遵循人类价值观与伦理规范，能力越强风险越大。价值对齐的核心难点在于人类价值观多元、动态且常冲突，难以被完整形式化为机器可学习的规则。关键技术机制包括基于人类反馈的强化学习、可解释性与价值观溯源以及设定原则性“宪法”的宪法 AI。这些机制共同引导模型优化行为，确保其目标与行为和人类价值观一致。

多元化是贯穿模型生命周期的伦理免疫系统。模型的伦理缺陷往往源于开发视角单一。多元化要求模型设计、开发、评估与治理全过程视角、背景与主体多元。开发团队多元化能更早识别潜在偏见，评估与测试多元化需引入超越传统技术指标的压力测试，治理与监督多元化则要建立多元伦理审查委员会，确保模型发展不被单一利益主导。

公平性算法、价值对齐与多元化需协同作用。多元化是基础，为公平性定义提供丰富视角，为价值对齐注入广泛代表性；公平性算法是具体工具，落实平等诉求，实现价值对齐目标；价值对齐是顶层指引，为公平性算法取舍提供伦理原则。未来的模型伦理工程需将三者系统性嵌入 AGI 研发流程，每个环节都进行伦理影响分析，应用技术工具与多元审查机制。

对 AGI 模型的完善，是技术伦理从外部约束走向内在塑造的转变。通过发展公平性算法、构建价值对齐机制、贯彻多元化原则，我们为 AGI 安装可靠的“伦理导航系统”，确保其能力锚定在增强人类尊严、促进社会公平与繁荣的终极目的上。模型的完善，归根结底是赋予其值得人类信任的品格。

（3）社会公众：技术普及，公众认识，弱势群体的参与式 AI 治理

AGI 的治理不仅关乎技术自身与法律，更关乎社会公众的共同利益，一个健康可持续的技术图景需要公众的理解与参与。当前存在的认知鸿沟与参与鸿沟，即公众对 AGI 的认知易被夸大或恐惧的叙事主导，缺乏理性思考，以及技术治理过程往往将少数群体与脆弱性群体排除在外，使其从“被忽视”变为“被决定”。因此，构建具有参与性与包容性的治理体系是确保 AGI 发展的关键。

弥合认知鸿沟。当前公众对 AGI 的认知呈现两极化。一方是将技术夸大为能够解决一切问题的手段，另一方则是将其视为导致人类失业、被替代，甚至灭绝的威胁。这两种认知都源于信息不对称和技术普及教育缺失。

面对这样的情况，首先，需要向公众普及基础的技术知识，通过学校教育、公共媒体等渠道广泛开展技术知识推广，包括技术原理、运作机制、社会影响、伦理风险及现有监管手段。其次，通过科学家、研发者、伦理学家与媒体等多方合作，向公众客观地描述 AGI 发展的现状与远景，为公众讨论提供理论支持。同时，搭建有效的多方沟通平台，政府、研究机构、企业与公众都能够对 AGI 的应用表达观点与价值判断，让本应推动社会发展的技术接受社会的审视与引导。

赋权少数群体。少数群体与脆弱性群体往往最易受到 AGI 负面影响的冲击，同时却最少有机会参与到技术设计中。因此，为保障各个群体的利益，需要确保 AGI 技术与治理的包容性与可参与度，主动考虑并消除这些群体面临的障碍，比如为残障人士提供无障碍格式或采

更多免费行业报告 并购家 www.ipoipo.net

用易于理解的描述性语言。其次，在 AGI 伦理委员会、规则制定、影响评估小组等关键治理机构中，为少数群体与脆弱性群体代表设置一定数量的席位与沟通协商机制，这些群体的经验与视角是辨识技术潜在风险、检验技术公平性的必要条件。为这些群体提供技术、法律等方面的支持，使其能够有效地收集证据、提出诉求、参与谈判、共同设计更全面的技术方案。

任何忽略少数群体与脆弱性群体的 AGI 治理框架都将是脆弱且缺乏合理与有效性的。这些群体不应只是作为技术的被动承受者，而应该成为能够参与到技术设计的主动者。将各个公众群体的智慧、价值和关切融入 AGI 发展中，既是防范技术风险的手段，也是确保技术创新服务于人类福祉、迈向更加公正和包容的人与技术共生的未来的路径。

参考文献

[1] 中国信息通信院知识产权与创新发展中心.人工智能伦理治理研究报告（2023 年），第 2 页。https://www.caict.ac.cn/kxyj/qwfb/ztbg/202312/P020240103531660935078.pdf

[2] 中国信息通信院知识产权与创新发展中心.人工智能伦理治理研究报告（2023 年），第 3 页。https://www.caict.ac.cn/kxyj/qwfb/ztbg/202312/P020240103531660935078.pdf

[3] Bubeck,S. et al.(2023).Sparks of Artificial General Intelligence: Early experiments with GPT-4March 2023.DOI:10.48550/arXiv.2303.12712

[4] Bengio,Y. et al.(2025)Superintelligent Agents Pose Catastrophic Risks: Can Scientist AI Offer a Safer Path? https://arxiv.org/abs/2502.15657

[5] Blaise Agüeray Arcas and Peter Norvig.（2023）. Artificial General Intelligence Is Already Here. https://www.noemamag.com/artificial-general-intelligence-is-already-here/

[6] Meredith Ringel Morris, Jascha Sohl-Dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clement Farabet, Shane Legg.(2024). Position: Levels of AGI for Operationalizing Progress on the Path to AGI.Proceedings of the 41st International Conference on Machine Learning, PMLR 235:36308-3632

[7] Meredith Ringel Morris, Jascha Sohl-Dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clement Farabet, Shane Legg.(2024). Position: Levels of AGI for Operationalizing Progress on the Path to AGI.Proceedings of the 41st International Conference on Machine Learning, PMLR 235:36308-36321

[8] John Koetsier. Artificial General Intelligence: 9 Massive Changes AGI Will Cause[N]. Forbes. 2025-11-04.

[9] 周伯文,从推理到科学发现：AI 迈向可精专通才之路，2026-01-29，https://www.shlab.org.cn/news/5444226

[10] Aziz,M.H.塑造 2030 年全球格局的四大风险趋势.https://cn.weforum.org/stories/2026/03/four-risk-trends-shaping-the-planet-cn/

[11] Shumailov, I., Shumaylov, Z., Zhao, Y. et al. AI models collapse when trained on recursively generated data. Nature 631, 755–759 (2024). https://doi.org/10.1038/s41586-024-07566-y

[12] Dessureault, JS., Lamontagne, R. & Parisé, PO. The Ethics of Creating Artificial Superintelligence: A Global Risk Perspective. AI Ethics 5, 6241–6263 (2025). https://doi.org/10.1007/s43681-025-00793-7

[13] Lucile Ter-Minassian. Democratizing AI Governance: Balancing Expertise and Public Participation. ArXiv, Submitted on 16 Jan 2025. https://doi.org/10.48550/arXiv.2502.08651

[14] Doshi-Velez, Finale et al. “Accountability of AI Under the Law: The Role of Explanation.” ArXiv, Submitted on 3 Nov 2017 (v1), last revised 20 Dec 2019

更多免费行业报告 并购家 www.ipoipo.net

(this version, v3). https://doi.org/10.48550/arXiv.1711.01134

[15] 吴信东，靳小龙，陈欢欢.大数据知识工程研究进展与发展趋势.中国计算机学会通讯，2021（6）：77.

[16] 杨庆峰. 2019. 数据偏见是否可以消除？自然辩证法研究，35（8）：109-113.

[17] Madumal P，Miller T，Vetere F，et al. 2018. Towards a grounded dialog model for explainable artificial intelligence. https://arxiv.org/pdf/1806.08055.pdf[2021-07-19].

[18] Bostrom N，Yudkowsky E. The ethics of artificial intelligence. In：Frankish K，Ramsey W. Cambridge Handbook of Artificial Intelligence. New York：Cambridge University Press，2014.

[19] Gasser U,Almeida V A F.A Layered Model for AI Governance[J].IEEE internet computing,2017,21(6):58-62.DOI:10.1109/MIC.2017.4180835.

[20] Madumal P,Miller T,Vetere F,Sonenberg L.Towards a Grounded Dialog Model for Explainable Artificial Intelligence[J/OL].2018.DOI:10.48550/arxiv.1806.08055.

[21] 张钹.迈向第三代人工智能[J].科学世界,2023,(10):1.

[22] Bengio Y,Cohen M,Fornasiere D,et al.Superintelligent Agents Pose Catastrophic Risks:Can Scientist AI Offer a Safer Path?[EB/OL].(2025-2-21)[2025-12-15].https://arxiv.org/abs/2502.15657.

[23] Panel for the Future of Science and Technology.A governance framework for algorithmic accountability and transparency[EB/OL].(2019-4-4)[2025-12-15].https://www.europarl.europa.eu/stoa/en/document/EPRS_STU(2019)624262.

[24] Jobin A,Ienca M,Vayena E.The global landscape of AI ethics guidelines[J].Nature machine intelligence,2019,1(9):389-399.DOI:10.1038/s42256-019-0088-2.

[25] 钱穆.中国文化史导论[M].北京:商务印书馆,1994.

第二篇

人文社会科学智能关键领域发展前沿与趋势

更多免费行业报告 并购家 www.ipoipo.net

第五章 从观测到决策的跃迁：复杂系统与社会仿真范式变革[①]

在数字化、网络化与智能化深度融合的背景下，人类社会正加速演化为典型的复杂适应系统。无论是舆论传播、市场运行，还是宏观社会行为，这类过程均表现出高度不确定性与复杂性，传统社会科学研究范式正面临方法论层面的结构性挑战。长期以来，以问卷调查、统计分析与案例研究为代表的方法，主要依赖对历史数据的观测与归纳，难以刻画多主体交互驱动下的动态演化机制。

随着计算能力与人工智能技术的快速发展，社会科学研究正加速向复杂系统范式转型，由“事后解释”转向“过程建模与机制推演”。通过构建由大语言模型驱动的多智能体系统，并在可控的数字环境中模拟个体行为与群体互动，复杂社会过程得以转化为可计算、可重复的“数字实验场”。在这一框架下，不仅可以复现既有运行规律，还能够开展反事实推演与策略优化，推动社会治理从经验驱动走向数据驱动与模型驱动的协同决策模式，从而为政策分析与智能决策提供实验化支撑。

围绕这一转变，本章从复杂系统视角出发，梳理人工智能驱动的社会仿真方法体系及其关键应用路径，呈现社会仿真从方法探索走向系统构建、从现象复现走向决策支持的发展趋势。

[①] 本节由复旦大学魏忠钰副教授撰写。魏忠钰教授现为复旦大学大数据学院副教授，博士生导师，数据智能与社会计算课题组负责人，主要从事大模型驱动的社会计算和多模态大模型的研究。

本章提要：

- 在数字化、网络化与智能化深度融合下，社会系统呈现复杂适应系统特征，“事后解释”的范式难以刻画多主体交互驱动的动态演化机制，社会科学正加速由观测归纳走向过程建模与机制推演。
- 以 LLM 驱动的多智能体系统为代表的社会仿真方法，为复杂社会过程提供可计算、可重复、可干预的“数字实验场”，支持反事实推演与策略优化，为政策分析与智能决策提供实验化支撑。
- 通用社会仿真框架的关键在于“真实世界对齐”（alignment）：在环境语境、用户分布、交互结构与行为生成四个层面建立一致性机制，是提升仿真可信性、可解释性与可扩展性的基础。
- 面向制度约束与策略博弈的多智能体平台（以 ProcureGym 为例）将政策规则、主体异质性与市场互动形式化为可计算的马尔可夫博弈，支持对机制参数的敏感性分析与反事实评估，推动政策评估从经验判断走向可验证的机制实验。
- 在开放社交媒体场景中，多智能体仿真与可视分析正在融合为“生成—分析—调控—再仿真”的人机协同闭环：仿真提供可控合成数据与过程日志，可视分析强化对结构涌现与传播路径的解释能力，共同提升对复杂社会行为的理解与干预能力。
- 社会仿真从方法探索走向系统构建与应用落地，仍需守住可信度底线：多源数据与行为对齐、可解释性与不确定性披露、伦理与隐私合规，以及可审计的版本/日志机制，决定其能否成为面向治理的基础设施。

更多免费行业报告 并购家 www.ipoipo.net

5.1 通用社会仿真系统：面向真实世界对齐的框架

在人工智能驱动的社会仿真框架中，如何使模拟系统能够真实反映现实世界，是实现从行为生成到群体推演的关键问题。相比传统模型仅关注统计一致性，面向复杂系统的社会仿真更强调在多层级结构上的一致性，即在环境语境、个体特征、交互机制与行为模式等方面实现与现实世界的系统性对齐。这一“对齐”（alignment）问题构成了通用社会仿真框架设计的核心基础，也直接决定了仿真系统的可信性与可解释性。

SocioVerse 框架在此背景下提出，构建了一套面向真实世界对齐的通用社会仿真系统（图 5-1）。该框架从系统层面对复杂社会模拟进行统一建模，重点解决四个关键问题：模拟环境如何与现实动态对齐，模拟智能体如何与真实用户分布对齐，交互机制如何与真实社会结构对齐，以及行为模式如何与真实群体反应保持一致。

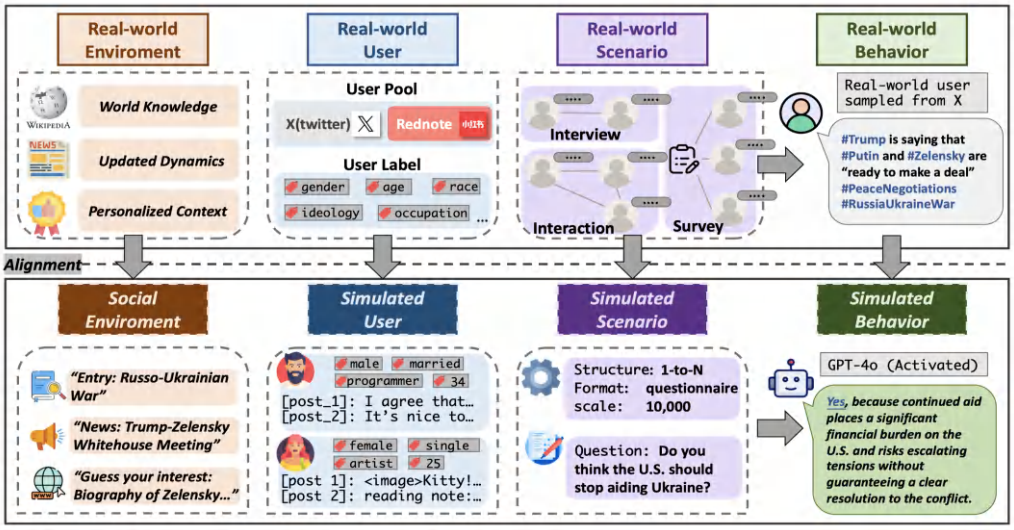


图 5-1 SocioVerse 在“乌克兰问题”场景下的示意图

对齐挑战分别体现在环境、用户、场景与行为四个层面。

1. SocioVerse 的整体框架与系统架构

从系统设计角度看，SocioVerse 将社会仿真抽象为一个由多模块协同构成的统一框架，其核心由社会环境（Social Environment）、用户引擎（User Engine）、场景引擎（Scenario Engine）与行为引擎（Behavior Engine）四个模块组成（图 5-2）。

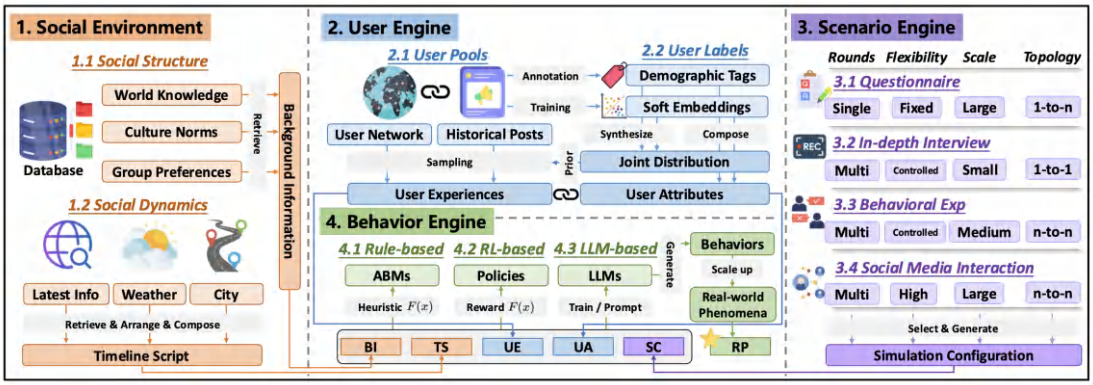


图 5-2　一个包含 4 个强大组件的 SocioVerse 框架的说明示意图

（行为引擎接收来自社会环境的背景信息（BI）和时间轴脚本（TI），来自用户引擎的用户经验（UE）和用户属性（UA），以及来自场景引擎的模拟配置（SC），然后生成观察真实世界现象（RP）所对应的行为）

首先，**社会环境模块**为模拟提供与现实事件一致的上下文信息，通过整合最新事件与社会统计数据并注入大语言模型智能体，从而提升模拟的现实性与预测能力。其核心可划分为社会结构（social structure）与社会动态（social dynamics）两个方面。其中，社会结构为智能体提供稳定的背景知识基础，涵盖世界知识、文化规范及群体偏好等，使其能够在特定时间截面下形成与现实一致的情境感知。例如，将特定历史时期的语言习惯、生活方式与价值观纳入建模，可有效复现对应阶段的公共讨论、人口流动与经济行为特征。相较之下，社会动态刻画随时间演化的外部信息流，包括新闻事件、政策变动及

更多免费行业报告 并购家 www.ipoipo.net

环境条件等。该模块通常通过构建持续更新的事件数据库，对现实世界信息进行时间戳与事件标签化管理，并形成结构化事件时间线。在此基础上，智能体能够检索历史语境并对突发事件作出响应，从而支持对社会行为演化趋势的预测与决策分析。

其次，**用户引擎模块**通过对齐真实世界用户样本，使模拟智能体能够对应目标用户群体，从而在仿真环境中构建具有代表性与多样性的虚拟人口结构。其核心由大规模用户池（user pool）与多维用户标签（user labels）两部分构成。其中，用户池通过整合国内外社交媒体平台（“小红书”与“X”）的用户历史行为数据，刻画个体的表达模式与行为特征。在数据构建过程中，通过发帖频率与文本相似度等指标对异常数据（如广告与机器人内容）进行过滤，并基于清洗后的社交媒体数据建立大规模用户索引，最终形成覆盖超 1000 万用户规模的行为数据库，为仿真提供真实分布基础。在此基础上，用户标签用于对个体进行结构化描述与嵌入表示，既包括离散属性，也包括通过模型学习得到的连续表征。针对人口统计信息难以直接获取的问题，该模块构建了“LLM 标注—人工校验—模型训练”的分层标注机制：首先由多个大语言模型在多维属性空间中对用户进行初步标注，随后通过人工评估与修订提升标签质量，最终基于高质量数据训练属性分类器，实现面向大规模用户属性的自动化推断。具体而言，用户标签覆盖 15 个关键维度，包括年龄、性别、职业、收入、教育、地区、意识形态及人格特征等，为模拟系统提供细粒度的人口结构建模能力。

场景引擎模块则依据具体任务表述与场景类型，将模拟结构与真

实世界语境进行对齐，并结合用户引擎提供的人口属性分布进行抽样，从而支持可规模化的社会仿真。其核心在于对现实社会情境进行形式化建模，并抽象为若干可复用的原型场景模板。具体而言，该模块将多样化社会交互归纳为四类典型场景：问卷调查、深度访谈、行为实验与社交媒体互动。问卷调查场景采用标准化单轮结构，适用于大规模群体意见测量与统计分析；深度访谈场景基于一对一多轮交互，通过持续追问与语境扩展，刻画个体动机与认知过程；行为实验场景在受控条件下对个体或群体进行多轮观测，用于分析决策行为、社会影响与认知偏差等机制；社交媒体互动场景则基于网络结构建模多主体动态交互过程，能够刻画信息传播、意见演化与群体行为等复杂现象。通过上述场景抽象与模板化建模，场景引擎实现了不同任务之间的结构复用与统一表达，使仿真系统能够在多种社会情境下保持交互一致性与扩展能力。

最后，**行为引擎模块**旨在使智能体行为与真实用户行为保持一致，其核心通过整合用户引擎提供的历史经验与个体属性、场景引擎定义的交互结构以及社会环境中的语境信息，对个体行为进行统一建模与动态生成。为实现可信的行为模拟，该模块构建了涵盖规则方法、强化学习与大语言模型的多范式行为函数体系。基于规则的方法延续传统智能体建模思路，通过启发式函数或数学模型驱动行为生成，具有计算效率高与可扩展性强的优势，适用于大规模人群模拟，尤其在刻画影响力较低的边缘用户时表现稳定。强化学习方法则通过环境交互与奖励信号驱动策略优化，在规则模型与生成式模型之间提供折中方

更多免费行业报告 并购家 www.ipoipo.net

案，使智能体能够在多主体互动中学习自适应与策略性行为，但其性能依赖于奖励函数设计与训练成本。相比之下，基于大语言模型的行为函数利用其语言理解与角色扮演能力，在给定用户画像与语境条件下生成更加灵活且具有语义一致性的行为反应。该类方法既可依赖通用模型通过提示机制实现非参数化控制，也可通过持续预训练、监督微调与强化学习构建专家或领域模型，以提升复杂场景下的行为真实性与专业性。通过上述多范式融合，行为引擎在可扩展性、可控性与表达能力之间实现平衡，使系统既能够高效模拟大规模群体行为，又能够在复杂语境中生成具有细粒度差异的个体决策，从而支撑宏观社会行为的真实涌现。

总体而言，**SocioVerse 的核心创新在于提出了“环境—用户—结构—行为”四个层面的多维对齐机制**。环境对齐保证仿真语境与现实事件同步，用户对齐确保人口结构与行为分布的代表性，结构对齐通过标准化交互模板复现真实社会中的互动模式，而行为对齐则通过 LLM 驱动的生成机制，使个体决策能够体现认知差异与情境依赖性。这一多维对齐框架使仿真系统不仅能够复现宏观统计规律，还能够在微观层面生成具有合理性的个体行为，从而在“结构一致性”与“行为真实性”之间取得平衡。同时，模块化设计使系统具备良好的可扩展性，可灵活适配不同领域与任务场景。

2. 典型应用场景与仿真实验

基于四个对齐引擎，SocioVerse 相关研究构建了三个具有代表性的社会模拟任务，分别覆盖政治传播、新闻传播与社会经济领域，包括总统选举预测、突发新闻反馈分析以及国家经济调查（图 5-3）。由于缺乏大规模交互式数据，三类任务均采用问卷式结构，并通过统一流程验证框架的通用性与可扩展性。

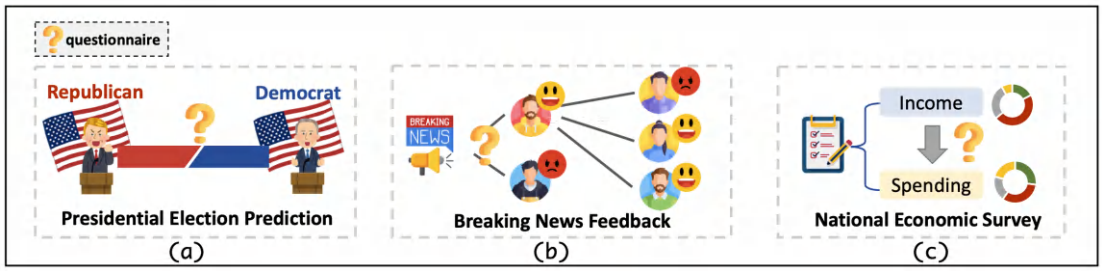


图 5-3　三个场景任务示意图：

（a）总统选举预测、（b）突发新闻反馈、（c）国民经济调查

在总统选举预测任务中，模拟以美国选举人团制度为背景，重点预测州级选举结果。在对齐机制上，社会环境注入候选人政策信息与竞选进程，用户引擎结合美国人口普查与 ANES 数据，在包含社会经济、地理与政治等 12 项属性的条件下，通过 IPF 方法进行分布对齐采样；场景引擎构建结构化选举问卷；行为引擎驱动超过 33 万智能体参与模拟，并在子样本上执行细粒度问卷分析。评估方面，采用准确率（Acc）与均方根误差（RMSE）分别衡量州级预测一致性与得票率偏差。

在突发新闻反馈任务中，模拟以特定新闻事件为触发，评估公众态度分布的可预测性。社会环境通过时间戳控制信息暴露顺序，用户

更多免费行业报告 并购家 www.ipoipo.net

引擎基于兴趣分布进行同分布采样（IDS），并在性别、年龄、教育与消费水平等维度上实现用户对齐；场景引擎基于 ABC 态度模型构建六维认知问卷；行为引擎驱动约 20,000 名智能体进行反馈模拟。评估采用归一化均方根误差（NRMSE）与 KL 散度，从数值与分布两个层面衡量模拟结果与真实数据的一致性。

在国家经济调查任务中，模拟以宏观经济条件为约束，刻画个体消费分配行为。社会环境提供各地区收入水平、人口规模与消费结构等宏观指标，用户引擎依据国家统计局分布进行采样对齐；场景引擎基于官方分类构建涵盖八类消费项目的结构化问卷；行为引擎将个体决策建模为在收入约束与区域背景下的消费分配过程，并通过聚合生成区域统计结果。评估同样采用 NRMSE 与 KL 散度，以衡量模拟分布与真实统计数据之间的一致性。

总体而言，上述三类任务从不同领域验证了框架在多场景下的适用性，表明通过“环境—用户—场景—行为”的协同对齐机制，能够在统一建模框架下实现从微观行为生成到宏观统计结果复现的跨尺度社会仿真能力。

3. 仿真性能评估与结果分析

为评估 SocioVerse 在不同社会场景中的仿真能力，研究选取了多种具有代表性的大语言模型，涵盖开源模型与商业模型，并在统一配置下开展跨场景对比实验。实验设置如表 5-1 所示，其中总统选举预测场景由于规模较大，部分配置出于计算成本考虑未纳入统一对比。

表 5-1　三个模拟任务的详细设置

Scenario	# Agents	# Demographics	Type	Sampling	Source	Language	# Questions	Ground truth
PresElectPredict	331,836	12	label	IPF	X	EN	49	real world
BreakNewsFeed	20,000	7	label	IDS	rednote	ZH	18	calculated
NatEconSurvey	16,000	9	label+number	IDS	rednote	ZH	17	real world

整体结果表明（表 5-2），在统一流程与多维对齐机制支持下，多数模型能够在宏观层面较好复现真实社会现象。其中，在总统选举预测任务中，主流模型在州级结果上实现了较高准确率，超过 90%的州级投票结果可被正确预测，表明该框架能够有效刻画群体决策行为。在细粒度层面，部分模型在摇摆州预测中表现出更强稳定性，但也存在因过度推理导致精度下降的现象。

表 5-2　三种模拟场景任务的总体结果

Model	PresElectPredict				BreakNewsFeed		NatEconSurvey			
	Overall		*Battleground*				*Overall*		*Developed-region*	
	Acc↑	RMSE↓	Acc↑	RMSE↓	KL-Div↓	RMSE↓	KL-Div↓	RMSE↓	KL-Div↓	RMSE↓
Llama3-70b	0.843	0.064	0.733	0.045	0.668	0.199	**0.016**	**0.026**	**0.013**	**0.025**
Qwen2.5-72b	**0.922**	**0.037**	0.800	**0.031**	**0.113**	0.059	0.066	0.048	0.043	0.039
DeepSeek-R1-671b	\	\	0.670	0.065	0.383	0.082	0.059	0.045	0.045	0.036
DeepSeek-V3	**0.922**	0.046	**0.867**	0.041	0.263	0.072	0.035	0.036	0.023	0.030
GPT-4o-mini	\	\	0.800	0.039	0.195	0.114	0.046	0.045	0.030	0.036
GPT-4o	\	\	\	\	0.196	**0.055**	0.062	0.051	0.036	0.038

注：其中子集 Battleground 表示美国总统选举中的摇摆州，子集 Developed-region 表示中国 GDP 排名前 10 的发达地区

在突发新闻反馈任务中，各模型在 KL 散度与 NRMSE 指标上整体接近真实用户分布，说明该方法能够较好捕捉公众态度结构。然而，分析发现模拟结果在分布范围上普遍较为集中，呈现出相对“保守”的倾向，表明当前大语言模型在舆情模拟中仍存在表达收缩与偏差风险。

在国家经济调查任务中，所有模型均与真实统计数据保持较高一致性，表明该框架在宏观经济行为建模中具有良好表现。进一步分析发现，不同模型在消费结构各维度上的表现存在差异，其中日常消费

更多免费行业报告 并购家 www.ipoipo.net

类别预测较为稳定，而住房支出等复杂决策维度误差较大，反映出大语言模型在复杂资源配置问题上的建模难度。

进一步地，针对总统选举预测中模拟精度的消融实验表明，先验人口分布与真实用户历史数据对模拟精度具有显著影响：相较随机采样，引入人口分布对齐能够在准确率与误差指标上稳定提升结果表现，而用户历史行为数据则在细粒度预测中进一步增强模型能力。这一结果验证了“用户对齐”与“真实语境注入”在社会仿真中的关键作用。

表 5-3　消融实验在总统选举预测模拟上的结果

Model	Acc↑	RMSE↓
Llama3-70b	**0.733**	**0.045**
- w/o Knowledge	0.533	0.051
- w/o Knowledge & Piror Distribution	0.600	0.386
Qwen2.5-72b	**0.800**	**0.031**
- w/o Knowledge	**0.800**	0.033
- w/o Knowledge & Piror Distribution	0.600	0.370
GPT-4o-mini	**0.800**	**0.039**
- w/o Knowledge	**0.800**	0.052
- w/o Knowledge & Piror Distribution	0.667	0.323

注：其中 Knowledge 表示真实世界的用户帖子，-w/o Prior Distribution 表示从用户池中随机抽样。

总体来看，多场景实验结果表明，SocioVerse 在标准化流程与专家参与的支持下，能够实现多样且精确的社会模拟。然而，不同底层大语言模型的选择会在不同场景中影响模拟精度，这一现象有待进一步研究。

4. 小结

综合来看，以 SocioVerse 为代表的通用社会仿真系统，体现了当前社会计算从单一方法探索向系统化建模范式转变的关键趋势。通过

在环境、用户、交互结构与行为生成等多个层面引入统一的对齐机制，该类框架实现了对复杂社会系统的整体性建模，使仿真不再局限于对局部现象的刻画，而是能够在统一体系下贯通个体行为与群体演化过程。从方法层面看，这一范式将大语言模型的生成能力与结构化建模、数据对齐及交互机制设计相结合，推动社会仿真由“数据驱动分析”向基于行为机制与交互过程的仿真建模转变；从系统层面看，其通过模块化架构实现多场景复用与跨领域扩展，为复杂社会行为的统一建模提供了基础框架；从应用层面看，该类系统已初步具备在多种社会场景中进行预测与分析的能力，展现出向决策支持与政策评估拓展的潜力。总体而言，这一方向标志着社会仿真研究正在从方法与模型层的离散探索，走向面向真实世界的系统化构建与应用落地，为后续更高保真、多尺度与可解释的社会模拟研究奠定了基础。

更多免费行业报告 并购家 www.ipoipo.net

5.2 面向制度约束与策略博弈的多智能体平台

在复杂社会系统中，除开放式信息传播与群体行为演化之外，还存在一类高度制度化的决策场景。这类场景通常由明确规则体系所主导，参与主体需在约束条件下进行策略选择，其结果不仅取决于个体特征，还受到竞争结构与制度设计的共同影响。典型代表包括拍卖机制、资源配置过程以及政策驱动的市场竞争环境等。此类场景往往直接服务于公共治理与宏观调控目标，其运行效果不仅关系到市场效率，也影响资源配置公平性与社会福利水平。

以国家药品集中采购为例，该制度作为我国医药领域的重要政策工具，旨在通过规模化采购与竞争性定价机制，降低药品价格、提升医保资金使用效率并促进医药产业结构优化。在这一制度框架下，通过约定采购量、最高有效申报价及中选排序规则等机制，构建了一个具有明确约束条件的市场竞争体系。企业需要在中标概率、利润空间与市场份额之间进行权衡，其决策过程呈现出典型的多主体动态博弈特征。同时，由于竞争对手信息不完全可见，企业无法直接观测对方成本结构与策略偏好，使整体决策过程具有显著不确定性与复杂性。

在此背景下，传统依赖经验判断或静态分析的方法难以刻画多主体互动、规则反馈与政策目标之间的动态关系，也难以系统评估不同制度参数对市场结果的影响。为此，复旦大学数据智能与社会计算实验室、复旦大学公共卫生学院、上海创智学院和同济大学研究团队联合推出了 ProcureGym—一种基于马尔可夫博弈的多智能体仿真平台，

专为药品采购中多企业竞标博弈等微观经济场景而设计，可为企业在国采场景下制定投标策略提供支持。

1. 方法框架与博弈建模

ProcureGym 以国家药品集中采购制度为约束，将异质性企业智能体之间的互动行为形式化描述为多智能体马尔可夫博弈（Markov Game）。其核心方法在于：将政策参数、企业属性及市场历史维度等编码为状态向量S_t，将报价行为定义为连续动作空间A_t，并利用 Top-K 中选规则与利润结算机制分别映射为状态转移函数$P(s'|s,a)$与奖励函数 R_t。

基于此标准化的建模范式，ProcureGym 框架能够有效地实现现实结果的高保真复现、未来集采演化趋势的模拟预测、政策规则的敏感性分析，以及企业竞价策略的可解释性剖析等深层次的经济政策论证与推演。一方面，平台支持通过反事实推理量化评估核心机制参数（如最高有效申报价、约定采购量）的变动效用，为政策干预与机制优化提供坚实的量化证据；另一方面，依托大语言模型的认知架构，平台突破了传统算法的黑盒局限，深度剖析异质性企业在规则约束下的底层博弈逻辑。

ProcureGym 在跨场景应用中展现出显著的普适性与跨域泛化能力，其理论框架深度契合微观经济学、多智能体博弈以及政策优化等领域的背景知识。该框架的应用版图可广泛延展至财税政策评估、市场竞争演化、劳动力市场动态分析、宏观经济治理以及企业级采购与

更多免费行业报告 并购家 www.ipoipo.net

拍卖机制设计等多元场景。在此类复杂系统中，各类微观决策主体（以企业为典型代表的智能体）的底层行为逻辑，均可被抽象为既定政策与机制约束下的最优化求解过程，从而赋予了其行为范式的可迁移性。

目前，ProcureGym 已作为开源基础设施正式发布于 GitHub（https://github.com/JiaWANG-TJ/ProcureGym），并于 Hugging Face 平台部署了在线交互演示示例（https://huggingface.co/spaces/SII-JiaWANG/procuregym），以赋能产学研各界的深入探索与应用实践。

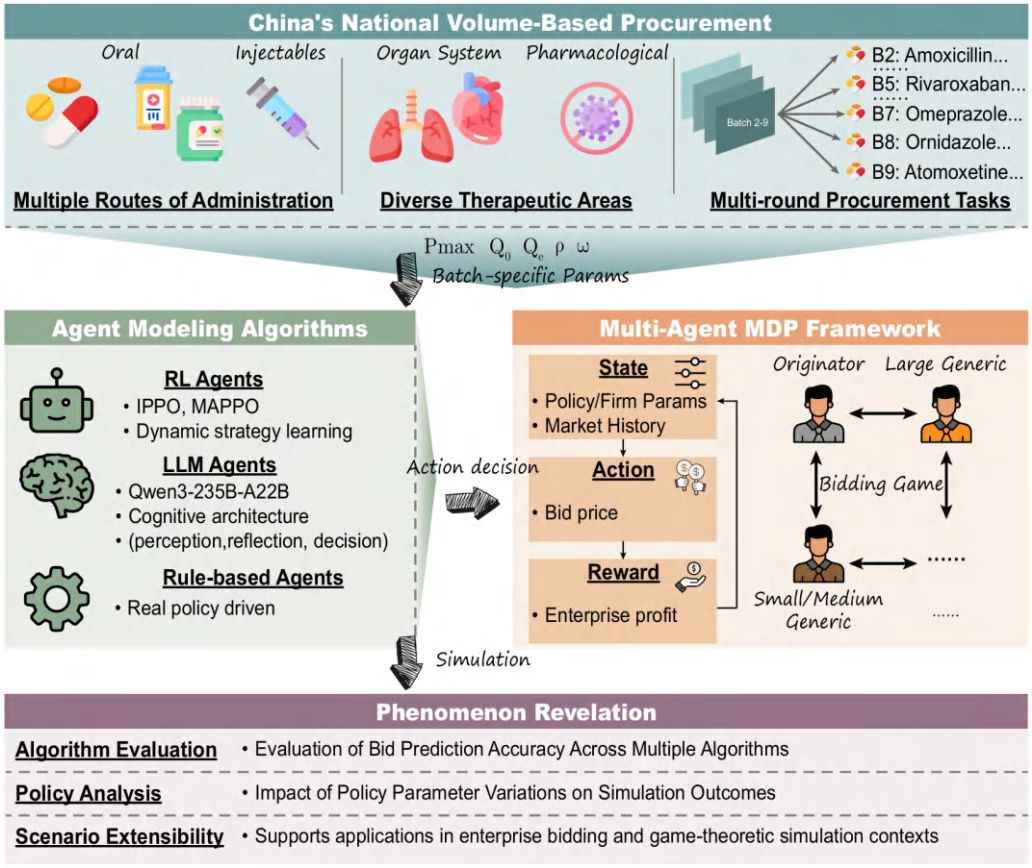


图 5-4　ProcureGym 平台概况

2.　数据基石与现实对齐

为构建高保真的动态博弈环境，ProcureGym 的数据底座深度融合了多源异构的真实世界权威数据，全面贯通了三大核心政企数据链路。

依托国家医疗保障局（National Healthcare Security Administration，NHSA）的采购文件数据，确立了政策规则与需求基线；整合国家药品审评中心（Center for Drug Evaluation，CDE）的审评资质数据，覆盖了潜在竞争企业信息与原料药自产资质信息 ；引入中国医药工业信息中心（China National Pharmaceutical Industry Information Center，PHIIC）的产业端数据，刻画了相关企业的企业规模、市场份额等信息。

ProcureGym 共计收录了历经 7 个批次中国国家药品集采的 325 个药品及 2,267 家企业样本。在药品维度，样本广泛覆盖口服（61%）、注射（35%）等给药剂型，并横跨抗感染（22%）、消化（15%）、神经（13%）及心血管（13%）等核心治疗领域。在企业维度，样本囊括了原研药企（11%）与仿制药企（89%），并在仿制药企业中覆盖了大型（23%）、中型（35%）及小型（32%）等多层级市场主体，凸显出中小型企业在集采竞争中的主导地位。此外，具备原料药自产优势的企业仅占比 25%。在市场竞争态势方面，48%的药品吸引了 3 至 5 家潜在竞标主体参与角逐，呈现典型的中高强度竞争态势；而在中选端，绝大多数（70%）药品的最终中选企业数量稳定在 2 至 5 家，整体中选率平稳在 50%至 80%的区间。

3. 实验设计与性能评估

为全面解析复杂集采场景动态，ProcureGym 引入了三类异质性智能体：一是基于的 IPPO 与 MAPPO 算法的强化学习智能体；二是由 Qwen3-235B-A22B-Instruct 模型驱动，且内嵌“感知-记忆-决策-反思”

更多免费行业报告 并购家 www.ipoipo.net

认知架构的大语言模型智能体；三是基于现实政策约束与企业属性构建的启发式规则智能体。研究确立了三大核心评估维度：价格预测准确性（斯皮尔曼相关系数 ρ 与决定系数 R^2）、中选预测准确性（Top-K 中选结果一致率）以及企业综合利润（集采内中选利润与集采外市场利润之和），旨在从底层机制上严密评测各类算法在复杂博弈环境下的决策性能。

在七个批次的药品集采仿真实验中，强化学习算法在核心任务上展现出了更优的性能。在价格拟合方面，所有算法均实现了显著的强斯皮尔曼正相关（$\rho = 0.85 \sim 0.88$，$p < 0.001$），其中强化学习算法达到了 0.79 的最高决定系数（R^2），这表明模型能够深度还原复杂的报价结构。在中选预测方面，强化学习智能体取得了高达 74.81%的整体准确率，显著超越基于大语言模型（66%）与启发式规则（64%）智能体。这一系列实验结果表明，在此密封拍卖多智能体博弈场景中，数据驱动的学习型策略面对复杂约束优化呈现出优秀的适配度和性能。

在政策与市场参数的敏感性分析中，研究揭示了规则变量对企业报价策略及利润分布的显著驱动效应。具体而言，更高的约定采购比例与更大的采购基础量会显著压制企业的报价水平并削减其利润空间，而相对地，提高最高有效申报价与扩大的实际市场需求则会同步拉升报价与利润预期。进一步的归因分析表明，市场需求是主导企业盈利能力的最核心驱动因素，其次为约定采购量与生产成本。

在智能体行为可解释性方面，尽管强化学习在整体性能上占据优

势，但基于大语言模型的智能体通过输出自然语言推理过程，为平台赋予了关键的可解释性，清晰阐释了智能体的底层行为逻辑与中间决策步骤。实验表明，企业自身的成本属性差异驱动了定价策略的分化。例如，低成本企业倾向于采取激进的低价策略以优先锁定市场份额，而高成本企业则更侧重于追求高收益，其维持着显著更高的平均报价水平与利润率。由此，相较于传统算法在策略优化过程中呈现的“黑盒”特性，大语言模型不仅实现了微观博弈路径的“白盒”表征，更为政策制定者深入洞察市场动态反馈、科学评估与优化集采规则提供了更具价值的证据支撑。

4. 小结

ProcureGym 展现出人工智能在赋能药品集采领域的重要价值：它将复杂场景表征为可计算的多主体决策问题，将政策规则制定转化为可检验的机制实验，将企业主体交互行为表现为可比较的量化策略。基于真实数据样本与大规模的实验分析，平台在“真实性、有效性、可解释性”三方面彰显出了优秀的应用能力，为政策评估和业界决策提供了可复用的技术基础。

更多免费行业报告 并购家 www.ipoipo.net

5.3 社交媒体仿真与可视分析方法

前文分别从通用仿真框架与制度约束下的策略博弈场景出发，系统分析了人工智能驱动的社会仿真方法体系。相较于上述以结构化环境为主的仿真场景，社交媒体作为典型的开放式复杂社会系统，拓展了从个体行为、信息传播到群体涌现的研究维度，并因其规模大、动态性强与交互高度异构等特征，成为检验多智能体仿真能力与理解复杂社会行为机制的重要场景。

受规模、复杂性与动态演化特征影响，传统方法难以揭示社交媒体系统的交互机制与演化过程，同时隐私与数据约束亦制约了可控社会实验的开展。大语言模型技术的突破推动多智能体建模由规则驱动向语义驱动实现范式跃迁，为社交媒体分析提供了新的方法论框架。通过构建具有记忆、规划与行为能力的智能体，研究者能够在虚拟环境中模拟用户行为、互动过程和信息传播机制，从而开展大规模社会仿真研究。与此同时，可视化与可视分析技术在社交媒体研究中发挥着越来越重要的作用。通过将复杂的数据结构映射为直观的图形和交互界面，可视分析方法能够帮助研究者识别社交网络结构、理解群体立场分布、追踪信息传播路径，并支持专家在复杂数据中进行探索式分析与决策。随着社交媒体数据规模的持续增长，单纯依赖自动算法或人工观察都难以有效完成分析任务，人机协同的可视分析系统逐渐成为研究复杂社会系统的重要工具。

在此背景下，基于大语言模型的多智能体仿真与可视分析技术逐

渐融合，形成了从行为生成到结构理解、再到交互调控的完整方法体系。一方面，基于多智能体的社会仿真可以生成高保真、可控的合成数据，为复杂社会现象的因果分析和假设检验提供实验环境；另一方面，可视分析系统能够帮助研究者理解仿真过程中产生的大规模行为数据，并通过交互式界面对仿真参数和环境进行实时调控，从而形成“仿真—分析—决策—再仿真”的人机协同闭环。

本节围绕社交媒体仿真与可视分析的关键方法展开，首先介绍多智能体驱动的仿真范式及其分层结构，其次分析可视化在复杂系统理解中的作用，进一步探讨仿真与可视分析的融合机制，最后总结当前面临的主要挑战与发展方向。

1. 多智能体驱动的社交媒体仿真范式

大语言模型凭借其在自然语言理解与生成方面的显著能力，为刻画人类认知过程与社交行为提供了新的技术路径，推动了基于智能体的社交媒体仿真研究快速发展。在大语言模型驱动下，社交媒体仿真已演化为由个体层、互动层、网络层与宏观社会层构成的多层级建模框架，各层级分别对应从微观个体行为建模到多主体交互、网络传播动力学以及宏观社会演化过程的分层仿真。

首先，个体层仿真（User-level Simulation）是多智能体系统的基础，其核心在于构建能够高保真刻画真实用户或特定人群特征的智能体。当前研究主要围绕智能体架构、个体数据注入方法与评估机制展开。在架构设计方面，主流 LLM 智能体通常由概要（Profile）、记忆

更多免费行业报告 并购家 www.ipoipo.net

（Memory）、规划（Planning）与行动（Action）四个模块构成，从而实现个体属性建模、长期行为一致性与决策生成的统一[1]。生成式智能体（Generative Agents）进一步证明，在具备记忆与规划能力的环境中，LLM 智能体能够涌现出类人行为模式，如信息传播与社会互动，为个体层仿真提供了重要方法论基础[2] 。在个体数据建模方面，现有方法主要分为基于提示的非参数方法与基于微调的参数化方法[3]。前者通过提示注入人口属性实现灵活的人群模拟，已在合成民调与公众意见建模中展现出较高外部一致性，但在少数群体刻画上存在系统偏差[4,5]；后者通过对齐训练实现对特定人物或角色的高保真复现，并在角色一致性与风格建模方面取得进展[6-9]。近年来，基于大规模人设生成的数据构建方法（如 Persona Hub）进一步推动了仿真人群的规模化扩展 [10]。在评估方面，个体仿真逐渐从静态输出一致性评估，演进至关注长期交互中的角色稳定性与情境适应能力，推动研究从表层特征复现向价值观与认知机制建模深化[3,11]。

其次，互动层仿真（Interaction-level Simulation）在个体智能体基础上引入多智能体交互机制，重点研究信息传播、说服过程与意见形成的动态机制。相关研究主要沿社交智能评估、多智能体辩论与博弈交互三个方向展开。在社交智能评估方面， SOTOPIA[12] 与 AgentSense[13]等框架通过构建多目标社交任务与多维评价指标，实现了对 LLM 智能体社交行为的量化分析，揭示了其在意图理解与隐性信息处理方面的能力边界 。在多智能体辩论方向，通过引入立场对抗与集体推理机制，研究表明多智能体系统在复杂推理任务中优于单

模型，并能够刻画智能体在社会压力下的立场调整行为，为意见演化研究提供了实验范式 [14, 15]。沟通博弈为互动层仿真提供了受控的竞争性实验环境，相关研究表明 LLM 已具备基本的欺骗与推理能力但在长期状态追踪与多步策略规划方面仍存在不足[16]，同时为意见引导、信息操控与社交欺骗等复杂社会互动机制研究提供了方法论支撑。总体来看，互动层仿真正从单轮对话场景向多阶段、强约束与高复杂度的社会交互系统演进，评价维度也从任务完成率扩展至涵盖社交规范与策略意图的综合指标体系。

再次，网络层仿真（Network-level Simulation）将研究视角拓展至结构化社交网络中的群体现象，重点关注信息传播动力学、群体极化及平台机制对行为生态的影响。在该层次中，LLM 驱动的仿真逐步实现从微观行为到宏观舆情的统一建模，例如，相关研究通过在 ABM 框架中嵌入 LLM 智能体，实现了情绪、态度与信息扩散过程的耦合建模，并验证了个体行为与群体演化之间的关联机制[17]；进一步地，大规模仿真平台 OASIS 通过引入推荐系统与交互机制，将仿真规模扩展至百万级智能体，从而复现信息传播与从众行为等典型涌现现象[18]。在此基础上，网络层仿真已在多个关键问题上形成系统化研究方向：在群体极化方面，仿真能够在不同网络结构下复现回音室效应，并通过干预机制评估去极化路径[19]；在推荐系统场景中，通过耦合用户行为与推荐机制，揭示信息茧房与过滤气泡的形成机制[20]；在虚假信息传播研究中，仿真结果表明传播路径受个体特征与网络结构的共同调节，干预策略具有显著的结构依赖性[21]。总体来看，网络层仿真

更多免费行业报告 并购家 www.ipoipo.net

正从单一现象复现，逐步演化为面向信息传播、群体结构与平台机制的统一建模框架，为理解复杂社交系统中的群体涌现机制提供了重要方法支撑。

最后，宏观社会仿真（Macro Social Simulation）将研究对象扩展至现实社会中的复杂系统，通过构建大规模异质智能体群体，在可控环境中探索社会行为的整体演化规律。在该层次中，LLM 驱动的仿真已逐步覆盖多个关键应用领域，并呈现出从现象复现向机制建模的演进趋势。在政治与地缘博弈场景中，相关仿真系统已能够在大规模人口或国家级主体建模基础上，对选举结果与冲突演化进行预测与解释分析，揭示路径依赖在宏观决策中的关键作用[22]。在舆情演化研究中，现有基准与实验结果表明，LLM 智能体在意见动力学中仍存在趋同偏差，难以自然复现现实中的极化现象，凸显了对齐偏差已成为制约宏观仿真真实性的核心问题之一[23]。进一步地，在社会运动与虚假信息传播场景中，多智能体仿真已实现从微观信息生成到宏观传播路径的全流程建模，使干预策略的评估具备实验化基础[24]。同时，该方法已拓展至公共卫生等领域，通过在流行病传播模型中引入行为决策机制，实现对群体行为反馈与系统演化的动态刻画[25]。总体来看，宏观社会仿真正从单一场景模拟，逐步发展为面向多领域复杂系统的统一建模框架，在政策评估与社会治理中展现出重要应用潜力。

综合而言，基于 LLM 的社交媒体仿真已形成从个体行为建模到宏观社会演化的多层级方法体系。个体层实现了从表层特征复现向认知与价值建模的深化，互动层提供了研究社会影响与意见形成的实验框

架，网络层揭示了结构约束下的群体涌现机制，宏观层则展现了仿真在现实复杂场景中的应用潜力。与此同时，仿真系统仍面临若干关键挑战，包括模型对齐偏差对群体代表性的影响、大规模仿真的计算成本约束，以及真实性评估标准的缺失等。这些问题贯穿各层级，是未来社交媒体仿真研究亟待突破的核心方向。

表 5-4 典型 LLM 社交媒体仿真研究对比

层次	代表工作(编号)	典型任务/场景	方法亮点	主要局限
个体层	[2][6][10]	角色建模、记忆与个性对齐	可复现个体行为与人设一致性	易受提示偏差与价值观漂移影响
互动层	[12][14][16]	社交评测、辩论推理、沟通博弈	多智能体交互提升社交推理可测性	隐性意图识别与长程策略能力仍弱
网络层	[17][18][19][24][74]	信息扩散、极化、推荐反馈	可在大规模网络复现群体现象	网络同质性估计与外推稳定性仍需验证
宏观社会层	[22][23][24][25][75][76][77][78]	选举、舆情演化、社会运动、公共卫生	支持政策情景模拟与干预评估	计算成本高，真实性依赖参数与数据校准

2. 可视分析驱动的复杂系统理解机制

现有社交媒体数据的可视化方法可从功能上划分为四类核心任务：**网络结构可视化、立场与情绪可视化、传播过程可视化以及多维数据可视分析系统**。前二者侧重刻画用户与内容之间的关系结构，后二者分别从认知分布、动态传播及多维分析角度揭示社交系统的复杂行为机制，并辅以伦理与隐私问题的规范性讨论，为后续系统构建与应用提供方法支撑。

在**网络结构可视化**层面，相关方法通常以图模型为基础，将用户、话题与内容抽象为节点，将关注、转发与互动关系抽象为边，旨在有

更多免费行业报告 并购家 www.ipoipo.net

限视觉空间内揭示社群结构与关键节点。常见布局方法包括力导向布局、谱布局、层次布局和地理布局等[26]，分别用于保持局部连接关系、全局结构特征与传播层级关系。在复杂社交网络中，多层关系与时间演化特征使得网络结构呈现出高度复杂性。为此，研究者提出多层网络可视化与动态网络展示方法，通过图层分解、时间切片及动画等方式呈现结构变化[27,28]。在大规模网络场景中，边混乱（hairball）问题成为主要瓶颈，通常通过边捆绑（edge bundling）、基于社群或 k-core 的聚合、隐藏低权重边以及 fisheye 等交互机制加以缓解，从而提高结构可读性[29]。在结构分析层面，社交媒体网络中的社群结构与节点作用构成理解信息传播机制的关键切入点。其中，社群发现揭示网络中的群体聚集模式及其“同温层”特征，中心性分析则刻画节点在信息传播中的桥接与放大作用，从而支持关键传播路径与意见领袖的识别[30-33]。进一步地，传播与互动结构通过级联与层级关系建模刻画信息扩散路径与讨论组织形式，并在异构网络框架下统一表征多类型实体之间的传播关系[34-35]。

在认知与情绪层面，**立场与情绪可视化**关注群体在议题空间中的分布及其演化过程，是理解舆情极化与意见分化的重要手段。通过立场识别模型与语义嵌入方法，可以将用户或文本映射至低维空间，并利用降维技术构建“极化空间”[36]，直观呈现群体结构与立场分布。在时间维度上，情绪与态度通常以时间序列形式组织，通过情绪曲线与分布变化揭示舆情波动与关键转折点。结合异常检测与事件指示方法，可进一步关联情绪变化与外部事件，实现“数据+叙事”的联

动分析，有助于复盘舆情危机的发展过程[37]。随着多模态内容的兴起，情绪表达逐渐扩展至图像与视频。多模态可视化通过融合文本与视觉信息，提高分析的表达能力[38]。同时，在跨平台研究中，通过多视图并列展示不同平台的情绪与立场特征，有助于揭示平台机制对舆情结构的影响。

在**传播过程可视化**层面，信息扩散通常通过时间与网络结构的结合加以刻画，是分析舆情演化与传播机制的核心路径。基础形式包括时间线展示发文量、情绪变化等指标，而更复杂的事件流表示则通过 Sankey 图或 Alluvial 图刻画用户、话题或注意力在不同状态之间的迁移过程[39,40]。在扩散结构分析中，级联树可用于描述转发路径与传播深度，并通过结构指标（如传播深度与结构病毒度）衡量信息传播能力[41]。通过对比不同平台的传播结构，可以揭示算法推荐与社交关系在信息扩散中的差异作用。在实际应用中，实时舆情监测系统通常采用多视图仪表盘形式，集成情绪指标、关键词分析、传播网络与地理分布等信息，并通过流数据处理与交互式筛选支持动态监测与快速响应。

前述方法往往聚焦于单一视角，而实际的社交媒体分析需要在“网络结构—立场情绪—传播过程—用户与内容属性”之间来回切换和综合推理。**多维数据可视分析系统**以多视图协调和交互分析为核心，将自动算法与人类专家的领域知识结合起来，支持复杂问题的探索式分析和决策制定[42]。其中，典型方法包括刷选联动（brushing & linking）与维度约简技术，使用户能够在不同数据视角之间切换，实现探索式

更多免费行业报告 并购家 www.ipoipo.net

分析[43,44]。Crossfilter 与钻取分析则进一步支持多条件组合查询与分层数据探索，使用户能够从宏观概览逐步深入到具体用户与内容[45-48]。在实际应用中，此类系统还需支持从分析到报告的完整工作流程，并通过可解释性设计提供数据与模型的“证据链”，增强分析结果的可信度。

此外，社交媒体可视化涉及真实用户数据与敏感议题，需在系统设计中充分考虑**伦理与隐私问题**。核心措施包括数据去标识化处理[49]、敏感信息聚合展示以及访问权限控制，以防止对个体的再识别风险。同时，应避免通过可视化强化刻板印象[50]或误导用户判断，并通过不确定性展示与模型解释机制提升分析透明度。在研究与应用层面，还需在可重复性与隐私保护之间取得平衡，在保障合规的前提下推动数据与方法的开放共享。

3. 仿真与可视分析的融合机制

由于社交媒体系统具有高度复杂性，传统单一分析方法难以刻画其内在的网络涌现机制，因而近年来大语言模型（LLM）驱动的智能体仿真与可视分析（Visual Analytics）的融合，正推动社会计算从静态“数据观察”向动态“交互式计算推演”转变。两者的融合展现出清晰的演进脉络：从松散耦合的结果展示发展到目前深度互动的闭环协同。这种融合趋势既扩展了多智能体模拟的应用边界，也为理解和干预复杂社交网络提供了直观、可控的新范式。

（4）可视分析驱动仿真调控

早期基于代理的模型（ABM）多作为“黑盒”批处理过程运行，研究者仅能在仿真结束后进行事后分析，难以观察中间状态或干预因果链条[51]。为此，可视分析引入计算操纵（Computational Steering）机制，使研究者能够在不中断仿真的情况下实时监控系统状态并动态调整参数[52]。在 LLM 驱动的多智能体环境中，该机制进一步扩展至多视图交互界面，支持对信息感知、行为逻辑、情绪演变与立场转化等过程的实时观测与调控。通过可视映射识别传播路径中的拓扑异常或行为偏离，并动态调整提示词、交互频率与网络结构，研究者能够系统评估干预变量对舆情演化路径的影响，从而实现面向复杂社会过程的交互式仿真调控[53,54]。

（5）仿真生成数据驱动可视分析

受平台接口限制、数据稀疏性与隐私伦理约束影响，真实社交媒体数据在因果推断与假设检验中的应用长期受限[53]。早期研究通常依赖有限仿真数据验证可视化算法[55]。随着大规模多智能体仿真的发展，仿真系统逐步演化为高保真、可定制的数据生成引擎，为可视分析提供结构完整的合成数据支持。通过在模拟环境中引入具有特定情绪偏向或意见领袖特征的智能体，仿真能够生成不同干预策略下的完整舆情演化轨迹。同时，多维合成日志弥补了现实数据中难以观测的微观决策过程，并为宏观传播流、情绪极化空间等分析提供连贯输入，从而有效缓解幸存者偏差与观察盲区问题[54]。

（6）人机协同闭环

仿真与可视化的融合正在由单向数据流转向“专家在回路”的人机

更多免费行业报告 并购家 www.ipoipo.net

协同闭环。不同于传统将可视化作为仿真末端呈现工具的范式，当前研究逐步形成“仿真执行—多维可视分析—专家研判—参数优化—再仿真”的动态循环机制[56]。其中，LLM 智能体负责大规模交互模拟与微观行为生成，可视分析系统提取宏观拓扑与涌现模式，领域专家则基于视觉洞见注入社会科学先验知识。该双向反馈机制实现了人类直觉与计算模型的协同优化，使研究者能够及时修正智能体行为偏差与逻辑幻觉，从而构建面向复杂社会动态的干预与预测方法体系[54,57]。

（7）当前挑战

尽管大模型智能体仿真与可视分析交叉研究近年来取得显著进展，但仍面临若干关键挑战。其一是微观可解释性不足：尽管可视化能够呈现宏观涌现现象，但由于 LLM 的“黑盒”特性，个体行为与情绪变化的内在机制仍难以解释[84]。第二是计算规模与渲染效率之间存在矛盾，在百万级智能体仿真中，通信事件与状态更新的爆发式增长对实时可视化构成严峻挑战[20]。此外，仿真结果与现实系统之间仍存在“真实性鸿沟”，在环境设定与行为模式对齐方面尚不充分，且缺乏统一的可视化基准用于评估合成数据的保真度[85]。最后，交互式调控技术可能带来新的伦理风险，亟需在可视分析系统中引入安全审查机制，以防其被滥用于大规模信息操控。

4. 未来发展方向

随着大语言模型与多智能体系统的快速发展，社交媒体仿真正由传统规则或统计模型驱动的方法，演进为以语言智能为核心的高维动态社会系统构建范式。这一转变不仅重构了社会仿真的技术路径，也

推动研究从静态建模向动态生成、交互推演与决策支持转型。总体来看，未来发展趋势主要体现在多模态仿真、大规模真实网络嵌入、实时数字孪生、AI 辅助政策评估以及大模型增强的可视分析五个方面。

（1）多模态仿真

当前主流社会仿真系统仍以文本交互为主，而真实社交媒体环境具有显著的多模态特征，涵盖文本、图像、视频与表情等多种信息形态。未来仿真将从单一文本驱动转向多模态理解与行为生成的一体化建模，使智能体能够在多源信息环境中进行更真实的情绪表达与决策行为。相关研究表明，多模态大语言模型结合人格化策略与知识检索机制，能够提升情感理解与行为生成的细粒度表达能力，并在大规模多智能体环境中复现复杂社会行为模式[58]。然而，多模态输入也引入新的安全与一致性挑战，例如跨模态信息干扰与决策偏移问题，亟需通过更强的跨模态推理与安全约束机制加以解决[59]。

（2）大规模真实网络嵌入

大规模真实网络嵌入是连接结构建模与行为生成的重要基础。传统方法通过将节点映射至低维向量空间刻画结构关系，而在复杂社交网络中，动态拓扑建模逐渐成为关键问题。随着 LLM 驱动社会仿真的兴起，网络嵌入与生成模型呈现深度融合趋势。一方面，嵌入技术被用于增强跨网络结构对齐与传播路径建模；另一方面，完整社交图结构被引入 LLM 系统以强化结构约束对行为生成的影响[60,61]。网络嵌入由此从辅助工具转变为支撑多智能体仿真的核心模块。

（3）实时数字孪生社交系统

更多免费行业报告 并购家 www.ipoipo.net

实时数字孪生社交系统旨在构建与现实社交媒体高度一致的虚拟环境，使研究者能够在可控条件下分析、预测与干预社会行为演化过程。相关研究已通过智能体建模与 LLM 驱动行为生成，实现对用户互动、信息传播及网络结构演化的统一仿真[62,63]。进一步发展中，研究重点逐渐转向提升仿真系统的现实一致性与可解释性，包括引入更细粒度的用户特征建模、模糊行为表示及高保真评估框架[64]。同时，大规模真实用户样本与多模块对齐策略的融合，使数字孪生系统在保持群体代表性的同时具备更强的泛化能力[65]。

（4）AI 辅助政策评估

随着生成式智能体技术的发展，AI 辅助政策评估逐渐成为社会仿真的重要应用方向。基于 LLM 的多智能体系统能够在虚拟环境中模拟个体决策、互动传播与结构演化过程，从而为政策分析提供可控实验平台[66]。在不同应用场景中，该方法已拓展至资源分配、经济运行、可持续发展及公共卫生等领域，通过模拟多主体行为与策略博弈，实现对政策效果的系统性评估[67-70]。同时，生成式政策实验平台支持多方案对比与反事实推演，为政策制定提供可重复、可解释的决策依据[71]。未来研究需进一步提升模型可信性与伦理可控性，并加强与实时数据的融合能力。

（5）可视分析系统嵌入 LLM 推理

随着社交媒体数据复杂性的提升，传统基于静态可视化的分析方法难以支撑深层机制理解。近年来，研究逐步将 LLM 的语义推理能力嵌入可视分析系统，形成融合数据可视化与语言推理的人机协同分

析范式。在该框架下，多模态预训练模型[72]能够理解结构图与时间序列等复杂可视信息，支持对仿真结果的语义解释；同时，LLM 在系统中的角色由结果解释扩展至任务规划、异常检测与分析推荐[73]，实现从“展示数据”向“辅助推理”的转变。结合多视图交互与知识结构建模，该类系统能够在复杂社交场景中提供更具解释性与可操作性的分析支持[54]。

5. 小结

综合来看，基于大语言模型的社交媒体仿真与可视分析，已逐步形成面向开放复杂社会系统的综合研究框架。该框架贯通个体行为生成、群体结构刻画与动态过程推演，形成从微观行为生成到宏观群体现象推演的多层级方法体系，使社会仿真由表层现象的复现，逐步迈向对复杂社会机制的建模与解释；可视分析则通过结构化表达与交互机制，将高维行为数据转化为可解释的模式认知，从而支撑人机协同的分析与决策过程。二者的融合进一步推动了社会计算研究由静态数据观察走向动态过程推演。多智能体仿真为复杂社会现象提供了可控、可重复的实验环境，可视分析则强化对复杂行为数据的解释能力，共同形成面向复杂社会行为理解、干预与评估的方法闭环。总体而言，这一方向不仅拓展了社交媒体研究的技术边界，也推动社会仿真由单一场景下的现象复现，进一步强化了社会仿真在复杂社会系统分析与决策支持中的方法价值。

更多免费行业报告 并购家 www.ipoipo.net

5.4 总结与展望

本章围绕复杂系统视角下的社会仿真方法体系，梳理了人工智能驱动社会科学研究的关键进展与发展路径。面对数字化、网络化与智能化深度融合背景下社会系统复杂性持续提升的趋势，传统以事后解释为主的研究范式已难以支撑对动态过程与群体演化机制的深入刻画。以大语言模型为核心驱动的多智能体仿真方法，正在推动社会科学由“观测与归纳”向“建模与推演”转型，将复杂社会系统转化为可计算、可复现、可干预的“数字实验体系”，为复杂社会行为的机制分析与政策评估提供了新的基础能力。

在技术路径上，本章从通用社会仿真框架、制度约束下的策略博弈系统以及社交媒体仿真与可视分析方法三个层面展开，呈现了人工智能驱动社会仿真的系统化发展格局。以多维对齐机制为核心的通用仿真框架，实现了从个体行为建模到群体结构演化的跨尺度统一建模；面向制度环境的多智能体博弈平台，将复杂政策问题转化为可计算的机制设计与策略优化问题；面向开放式社交系统的仿真与可视分析融合方法，则构建了“生成—分析—调控”的人机协同闭环。整体上，社会仿真正由方法探索阶段迈向系统构建与应用落地阶段，逐步成为连接微观行为机制与宏观治理决策的重要技术桥梁。

面向未来，人工智能驱动的社会仿真有望成为支撑国家治理体系与治理能力现代化的重要基础设施。从发展趋势看，**一是向更高保真的现实对齐演进**，通过多源数据融合与多层级建模提升仿真系统的真

实性与泛化能力；**二是向更强可解释性的机制建模深化**，突破黑盒模型限制，实现从行为生成到机制认知的系统刻画；**三是向更大规模与更高效率的计算体系拓展**，支撑百万级乃至千万级智能体的实时仿真与动态推演；**四是向人机协同的决策支持体系升级**，推动仿真系统与可视分析深度融合，形成可交互、可验证、可优化的闭环决策模式；**五是向规范化与可控化方向发展**，在保障隐私安全与伦理合规的前提下，构建可审计、可约束的仿真治理体系。

总体而言，大语言模型驱动的多智能体社会仿真，正在成为推动社会科学范式转型与智能治理能力提升的关键技术路径。未来，随着多模态智能、实时数字孪生与大规模社会模拟等技术的持续发展，社会仿真有望从研究工具进一步演进为支撑复杂系统分析与国家战略决策的重要基础能力，在经济运行监测、社会治理优化与公共政策制定等领域发挥更加核心的作用。

更多免费行业报告 并购家 www.ipoipo.net

参考文献

[1] Wang L, Ma C, Feng X, et al. A survey on large language model based autonomous agents[J]. Frontiers of Computer Science, 2024, 18(6): 186345. DOI: 10.1007/s11704-024-40231-1.

[2] Park J S, O'Brien J, Cai C J, et al. Generative agents: Interactive simulacra of human behavior[C]//Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology. 2023: 1-22. DOI: 10.1145/3586183.3606763.

[3] Chen J, Wang X, Xu R, et al. From persona to personalization: A survey on role-playing language agents[J]. arXiv preprint arXiv:2404.18231, 2024.

[4] Argyle L P, Busby E C, Fulda N, et al. Out of one, many: Using language models to simulate human samples[J]. Political Analysis, 2023, 31(3): 337-351. DOI: 10.1017/pan.2023.2.

[5] Lee S, Peng T Q, Goldberg M H, et al. Can large language models estimate public opinion about global warming? An empirical assessment of algorithmic fidelity and bias[J]. PLOS Climate, 2024, 3(8): e0000429. DOI: 10.1371/journal.pclm.0000429.

[6] Shao Y, Li L, Dai J, et al. Character-LLM: A Trainable Agent for Role-Playing[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023: 13153-13187. DOI: 10.18653/v1/2023.emnlp-main.814.

[7] Zhou K, Peng B, Zhu X, et al. CharacterGLM: Customizing Chinese Conversational AI Characters with Large Language Models[J]. arXiv preprint arXiv:2311.16832, 2023.

[8] Wang Z M, Peng Z, Que H, et al. RoleLLM: Benchmarking, Eliciting, and Enhancing Role-Playing Abilities of Large Language Models[J]. arXiv preprint arXiv:2310.00746, 2023.

[9] Chen N, Wang Y, Jiang H, et al. Large language models meet harry potter: A dataset for aligning dialogue agents with characters[C]//Findings of the Association for Computational Linguistics: EMNLP 2023. 2023: 8506-8520. DOI: 10.18653/v1/2023.findings-emnlp.570.

[10] Ge T, Chan X, Wang X, et al. Scaling synthetic data creation with 1,000,000,000 personas[J]. arXiv preprint arXiv:2406.20094, 2024.

[11] Wu W, Wu H, Jiang L, et al. From Role-Play to Drama-Interaction: An LLM Solution[J]. arXiv preprint arXiv:2405.14231, 2024.

[12] Zhou X, Zhu H, Mathur L, et al. SOTOPIA: Interactive Evaluation for Social Intelligence in Language Agents[J]. arXiv preprint arXiv:2310.11667, 2023.

[13] Mou X, Liang J, Lin J, et al. AgentSense: Benchmarking Social Intelligence of Language Agents through Interactive Scenarios[J]. arXiv preprint arXiv:2410.19346, 2024.

[14] Du Y, Li S, Torralba A, et al. Improving Factuality and Reasoning in Language Models through Multiagent Debate[J]. arXiv preprint arXiv:2305.14325, 2023.

[15] Xiong K, Ding X, Cao Y, et al. Examining Inter-Consistency of Large Language Models Collaboration: An In-depth Analysis via Debate[C]//Findings of the Association for Computational Linguistics: EMNLP 2023. 2023: 7572-7590. DOI: 10.18653/v1/2023.findings-emnlp.508.

[16] Xu Y, Wang S, Li P, et al. Exploring large language models for communication games: An empirical study on werewolf[J]. arXiv preprint arXiv:2309.04658, 2023.

[17] Gao C, Lan X, Lu Z, et al. S3: Social-network Simulation System with Large Language Model-Empowered Agents[J]. arXiv preprint arXiv:2307.14984, 2023.

[18] Yang Z, Zhang Z, Zheng Z, et al. OASIS: Open Agents Social Interaction Simulations on One Million Agents[J]. arXiv preprint arXiv:2411.11581, 2024.

[19] Wang C, Gu C, Wei M, et al. Decoding Echo Chambers: Simulations Revealing Polarization Dynamics in Social Networks[C]//Proceedings of the 31st International Conference on Computational Linguistics. 2025: 10180-10201. Available: https://aclanthology.org/2025.coling-main.698/ .

[20] Zhang A, Chen Y, Sheng L, et al. On Generative Agents in Recommendation[C]//Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024: 1807-1817. DOI: 10.1145/3626772.3657844.

[21] Li X, Xu Y, Zhang Y, et al. Large Language Model-driven Multi-Agent Simulation for News Diffusion Under Different Network Structures[J]. arXiv preprint arXiv:2410.13909, 2024.

[22] Zhang X, Lin J, Sun L, et al. ElectionSim: Massive Population Election Simulation Powered by Large Language Model Driven Agents[J]. arXiv preprint arXiv:2410.20746, 2024.

[23] Chuang Y S, Goyal A, Harlalka N, et al. Simulating Opinion Dynamics with Networks of LLM-based Agents[C]//Findings of the Association for Computational Linguistics: NAACL 2024. 2024: 3326-3346. DOI: 10.18653/v1/2024.findings-naacl.211.

[24] Mou X, Wei Z, Huang X. Unveiling the Truth and Facilitating Change: Towards Agent-based Large-scale Social Movement Simulation[J]. arXiv preprint arXiv:2402.16333, 2024.

[25] Williams R, Hosseinmardi H, Kleinberg J. Epidemic Modeling with Generative Agents[J]. arXiv preprint arXiv:2307.04986, 2023.

[26] Chen S, Li S, Chen S, et al. R-map: A map metaphor for visualizing information reposting process in social media[J]. IEEE transactions on visualization and computer graphics, 2019, 26(1): 1204-1214.

[27] Xuan P, Qi X, Chen S, et al. Subgraph topology and dynamic graph topology enhanced graph learning and pairwise feature context relationship integration for predicting disease-related miRNAs[J]. Journal of Chemical Information and Modeling, 2025, 65(3): 1631-1640.

[28] Beck F, Burch M, Diehl S, et al. A taxonomy and survey of dynamic graph visualization[C]//Computer graphics forum. 2017, 36(1): 133-159.

更多免费行业报告 并购家 www.ipoipo.net

[29] Holten D. Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data[J]. IEEE Transactions on visualization and computer graphics, 2006, 12(5): 741-748.

[30] Li W, Schöttler S, Scott-Brown J, et al. Networknarratives: Data tours for visual network exploration and analysis[C]//Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. 2023: 1-15.

[31] Chen S. Patterns of Integration: A Network Perspective on Popular Religious Connections in China's Lower Yangzi, 1150–1350[J]. Religions, 2023, 14(5): 577.

[32] Brandes U. A faster algorithm for betweenness centrality[J]. Journal of mathematical sociology, 2001, 25(2): 163-177.

[33] Weng J, Lim E P, Jiang J, et al. Twitterrank: finding topic-sensitive influential twitterers[C]//Proceedings of the third ACM international conference on Web search and data mining. 2010: 261-270.

[34] Chen S, Andrienko N, Andrienko G, et al. Co-bridges: pair-wise visual connection and comparison for multi-item data streams[J]. IEEE Transactions on Visualization and Computer Graphics, 2020, 27(2): 1612-1622.

[35] Liu S, Cui W, Wu Y, et al. A survey on information visualization: recent advances and challenges[J]. The Visual Computer, 2014, 30(12): 1373-1393.

[36] Conover M, Ratkiewicz J, Francisco M, et al. Political polarization on twitter[C]//Proceedings of the international aaai conference on web and social media. 2011, 5(1): 89-96.

[37] Stieglitz S, Dang-Xuan L. Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior[J]. Journal of management information systems, 2013, 29(4): 217-248.

[38] Kucher K, Kerren A. Text visualization techniques: Taxonomy, visual survey, and community insights[C]//2015 IEEE Pacific visualization symposium (pacificVis). IEEE, 2015: 117-121.

[39] Chong C, Zhang X, Kong G, et al. A visualization method of the economic input–output table: mapping monetary flows in the form of Sankey diagrams[J]. Sustainability, 2021, 13(21): 12239.

[40] Shao Z, Sun S, Zhao Y, et al. Visual explanation for open-domain question answering with bert[J]. IEEE Transactions on Visualization and Computer Graphics, 2023, 30(7): 3779-3797.

[41] Goel S, Anderson A, Hofman J, et al. The structural virality of online diffusion[J]. Management science, 2016, 62(1): 180-196.

[42] Keim D A, Mansmann F, Schneidewind J, et al. Visual analytics: Scope and challenges[M]//Visual data mining: Theory, techniques and tools for visual analytics. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008: 76-90.

[43] Zhang W, Tan S, Chen S, et al. Visual reasoning for uncertainty in spatio-temporal events of historical figures[J]. IEEE Transactions on Visualization and Computer Graphics, 2022, 29(6): 3009-3023.

[44] Zhang Y, Liu J, Lai C, et al. Interpreting high-dimensional projections with

capacity[J]. IEEE Transactions on Visualization and Computer Graphics, 2023, 30(9): 6038-6055.

[45] Koren J, Zhang Y, Liu X. Personalized interactive faceted search[C]//Proceedings of the 17th international conference on World Wide Web. 2008: 477-486.

[46] Xia T, Zhang H, Jing Y, et al. CrossIndex: Memory-Friendly and Session-Aware Index for Supporting Crossfilter in Interactive Data Exploration[C]//International Conference on Database Systems for Advanced Applications. Cham: Springer International Publishing, 2022: 476-492.

[47] Guimares R V, Soares A G M, Carneiro N J S, et al. Design considerations for drill-down charts[C]//2011 15th International Conference on Information Visualisation. IEEE, 2011: 73-79.

[48] de Carvalho M B, Meiguins B S, de Morais J M. Temporal data visualization technique based on treemap[C]//2016 20th International Conference Information Visualisation (IV). IEEE, 2016: 399-403.

[49] Nelson G S. Practical implications of sharing data: a primer on data privacy, anonymization, and de-identification[C]//SAS global forum proceedings. 2015: 1-23.

[50] Holder E, Xiong C. Dispersion vs disparity: Hiding variability can encourage stereotyping when visualizing social outcomes[J]. IEEE Transactions on Visualization and Computer Graphics, 2022, 29(1): 624-634.

[51] Crouser, R. J., et al. (2012). Two visualization tools for analyzing agent-based simulations in political science. *IEEE Computer Graphics and Applications*, 32(1), 67-73.

[52] Ko, S., et al. (2014). VASA: Interactive computational steering of large asynchronous simulation pipelines for societal infrastructure. *IEEE Transactions on Visualization and Computer Graphics*, 20(12), 1853-1862.

[53] Gao, L., ..., Wei, Z., & Chen, S. (2025). SimSpark: Interactive Simulation of Social Media Behaviors. *Proceedings of the ACM on Human-Computer Interaction (CSCW)*.

[54] Cheng Z, Lin Z, Yang Y, et al. Interactive simulation and visual analysis of social media event dynamics with LLM-based multi-agent modeling[J]. Visual Informatics, 2025: 100260.

[55] Keim, D. A., & Schreck, T. (2013). Visual analysis of social media data. Computer, 46(5), 68-75.

[56] Chen, S., et al. (2017). E-map: A visual analytics approach for exploring significant event evolutions in social media. *IEEE VAST*.

[57] Liu, Y., et al. (2024). LightVA: Lightweight Visual Analytics with LLM Agent-Based Task Planning and Execution. *IEEE Transactions on Visualization and Computer Graphics*.

[58] Lu, Zhixiang, Xueyuan Deng, Yiran Liu, et al. “PRISM: A Personality-Driven Multi-Agent Framework for Social Media Simulation.” arXiv:2512.19933. Preprint, arXiv, December 22, 2025. https://doi.org/10.48550/arXiv.2512.19933.

[59] Vera, Alhim, Karen Sanchez, Carlos Hinojosa, Haidar Bin Hamid, Donghoon Kim, and Bernard Ghanem. “Multimodal Safety Evaluation in Generative Agent Social Simulations.” arXiv:2510.07709. Preprint, arXiv, October 9, 2025. https://doi.org/10.48550/arXiv.2510.07709.

更多免费行业报告 并购家 www.ipoipo.net

[60] “Beyond Embedding-Mapping: Social Network Alignment via Generative Information Fusion and LLM-Guided Iterative Mechanism.” *Information Fusion* 127 (March 2026): 103733. https://doi.org/10.1016/j.inffus.2025.103733.

[61] Ji, Jiarui, Zehua Zhang, Zhewei Wei, Bin Tong, Guan Wang, and Bo Zheng. “GRAPHIA: Harnessing Social Graph Data to Enhance LLM-Based Social Simulation.” arXiv:2510.24251. Preprint, arXiv, October 28, 2025. https://doi.org/10.48550/arXiv.2510.24251.

[62] Rossetti, Giulio, Massimo Stella, Rémy Cazabet, et al. “Y Social: An LLM-Powered Social Media Digital Twin.” arXiv:2408.00818. Preprint, arXiv, August 1, 2024. https://doi.org/10.48550/arXiv.2408.00818.

[63] “(PDF) Intelligent Digital Twin for Predicting Technology Discourse Patterns: Agent-Based Modeling of User Interactions and Sentiment Dynamics in DeepSeek Discourse Case.” ResearchGate, ahead of print, October 10, 2025. https://doi.org/10.3390/systems13060451.

[64] Wen, Jiaqi, Bogdan Gabrys, and Katarzyna Musial. “Fuzzy Feature Representation for Digital Twin-Oriented Social Network Simulators.” 2024 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), June 2024, 1–8. https://doi.org/10.1109/FUZZ-IEEE60900.2024.10611981.

[65] Zhang, Xinnong, Jiayu Lin, Xinyi Mou, et al. “SocioVerse: A World Model for Social Simulation Powered by LLM Agents and A Pool of 10 Million Real-World Users.” arXiv:2504.10157. Preprint, arXiv, July 15, 2025.

[66] Piao, Jinghua, Yuwei Yan, Jun Zhang, et al. “AgentSociety: Large-Scale Simulation of LLM-Driven Generative Agents Advances Understanding of Human Behaviors and Society.” arXiv:2502.08691. Preprint, arXiv, February 12, 2025. https://doi.org/10.48550/arXiv.2502.08691.

[67] Ji, Jiarui, Yang Li, Hongtao Liu, et al. “SRAP-Agent: Simulating and Optimizing Scarce Resource Allocation Policy with LLM-Based Agent.” arXiv:2410.14152. Preprint, arXiv, October 18, 2024. https://doi.org/10.48550/arXiv.2410.14152.

[68] Li Nian, Gao Chen, Li Mingyu, Li Yong, and Liao Qingmin. “EconAgent: Large Language Model-Empowered Agents for Simulating Macroeconomic Activities.” August 2024, 15523–36. https://doi.org/10.18653/v1/2024.acl-long.829.

[69] Liu, Jia’an, Chu Chu, Yilin Zhao, Goshi Aoki, and Zhiqing Xiao. “Agentic AI for Sustainable Development: Leveraging Large Language Model-Enhanced Agent-Based Modeling for Complex Policy Strategies.” Emerging Media 3, no. 3 (2025): 401–13. https://doi.org/10.1177/27523543251365678.

[70] Aoki, Goshi, and Navid Ghaffarzadegan. “AI Agents as Policymakers in Simulated Epidemics.” arXiv:2601.04245. Preprint, arXiv, January 6, 2026. https://doi.org/10.48550/arXiv.2601.04245.

[71] Zhang, Shuhan, Zifan Peng, and Yinwang Ren. “GPLab: A Generative Agent-Based Framework for Policy Simulation and Evaluation.” Journal of Artificial Societies & Social Simulation 29, no. 1 (2026): 1. https://doi.org/10.18564/jasss.5933.

[72] Liu, Haotian, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. “Visual Instruction Tuning.” arXiv:2304.08485. Preprint, arXiv, December 11, 2023. https://doi.org/10.48550/arXiv.2304.08485.

[73] Zhao Y, Zhang Y, Zhang Y, et al. Leva: Using large language models to enhance visual analytics[J]. IEEE transactions on visualization and computer graphics, 2024, 31(3): 1830-1847.

[74] Wang L, Zhang J, Yang H, et al. User Behavior Simulation with Large Language Model based Agents[J]. arXiv preprint arXiv:2306.02552, 2023.

[75] Zhou L, Xu Y, Wang Z, et al. FlockVote: LLM-Empowered Agent-Based Modeling for Simulating U.S. Presidential Elections[J]. arXiv preprint arXiv:2512.05982, 2025.

[76] Hua W, Fan L, Li L, et al. War and Peace (WarAgent): Large Language Model-based Multi-Agent Simulation of World Wars[J]. arXiv preprint arXiv:2311.17227, 2023.

[77] Liu Y, Song Z, Zhang J, et al. The Stepwise Deception: Simulating the Evolution from True News to Fake News with LLM Agents[J]. arXiv preprint arXiv:2410.19064, 2024.

[78] Hu T, Liakopoulos D, Wei X, et al. Simulating Rumor Spreading in Social Networks using LLM Agents[J]. arXiv preprint arXiv:2502.01450, 2025.

更多免费行业报告 并购家 www.ipoipo.net

第六章 从识别到估计的革命：深度学习赋能的社会科学因果推断[①]

社会科学的核心关切往往不是“变量是否相关”，而是“何种机制导致了何种结果”，这使因果推断成为理解社会分层、解释行为差异与评估公共政策效果的基础方法。近年来，因果识别理论框架日益成熟，但在实践层面，一个更突出且常被忽视的瓶颈来自**估计阶段的模型依赖性**：大量方法需要对倾向得分或条件期望等关键量进行参数化建模，一旦模型设定刚性或错设，在高维协变量、非线性关系与复杂交互结构下就可能产生系统偏误，甚至使结论更多反映“模型产物”而非数据所蕴含的信息。深度学习作为高度灵活的非参数逼近器，为缓解这一瓶颈提供了新路径：一方面，通过表示学习与样本平衡思想，将处理效应估计从“控制变量/拟合方程”推进到更具数据自适应性的反事实预测；另一方面，通过生成式分布建模与反事实模拟，为中介机制刻画与“跨世界”嵌套反事实的估计提供了可操作的计算框架。本章围绕“效应估计—机制分析”两条主线，梳理深度学习与因果推断的融合逻辑，强调在坚守识别假设与因果逻辑前提下，推动社会科学因果分析由参数化设定走向分布级建模与可复核的证据生产。

[①] 本章由复旦大学社会发展与公共政策学院胡安宁教授牵头编写。

胡安宁教授现任复旦大学社会发展与公共政策学院院长，中国社会学会计算社会学分会理事长，万人计划哲学社会科学领军人才、青年长江学者、上海市领军人才。

本章提要：

- 社会科学因果推断的识别框架已较成熟，但实践瓶颈日益集中于估计阶段的模型依赖性：倾向得分与条件期望等关键量的参数设定在高维、非线性与复杂交互条件下易引入系统偏误。
- 深度学习作为数据自适应的非参数逼近器，为缓解“刚性设定”带来的估计偏差提供技术路径，其价值在于提高反事实预测稳定性，而非替代因果识别理论。
- 在处理效应估计上，TARNet/CFR 等方法将因果问题重构为表示学习与样本平衡问题：通过共享表征空间与分布匹配约束提升处理组与对照组的可比性，降低外推风险并增强因果泛化能力。
- 在机制刻画上，基于归一化流等生成式框架可对中介与结果的条件分布进行分布级建模，支持对自然间接效应等“嵌套反事实”量的模拟估计，使中介分析从线性系数分解走向可逆变换下的干预、生成与比较。
- 方法融合需坚持三条底线：未观测混杂无法被模型复杂度弥补；高度灵活模型带来更高数据需求与有限样本风险；深度模型的可解释性与可审计性必须通过明确假设、稳健性检验与过程记录加以强化。
- 面向未来，关键不在“用深度学习取代传统因果方法”，而在构建兼具识别严谨性与估计灵活性的整合框架：在反事实逻辑与识别假设约束下，充分利用表示学习与生成建模能力，形成更稳健、更精细、可复核的因果证据生产链条。

更多免费行业报告 并购家 www.ipoipo.net

6.1 估计阶段的模型依赖性与深度学习的进入点

和自然科学研究相一致，社会科学研究的一项核心目标，不仅仅是描述表征社会特征的变量彼此之间的相关性，而在于揭示它们之间所存在的因果关系（Angrist & Pischke 2009; Morgan & Winship 2014）。无论是理解社会分层与流动的形成机制，还是进行公共政策的效果评估，社会科学研究者所真正关心的问题通常可以归结为一个因果性命题，即某变量 D 是否引起了变量 Y。在确定了这一命题之后，研究者可能会进一步希望探知，D 和 Y 之间的因果关系通过何种渠道得以达成，也就是所谓的中介分析。例如，“如果”某一制度安排或者社会环境条件发生（或者没有发生）改变，个体或群体的结果和没有发生（发生）这些改变的情形相比，会如何变化。进一步，这种反事实关联的形成机制是什么。这一因果性命题在既有研究中可以通过反事实的分析框架进行形式化描述（Imbens & Rubin 2015）。假设处理变量 D 取值为 0 或者 1，因变量表示为 Y，那么在反事实因果推论框架下，个体 i 的因果效应τ_i可表示为：$\tau_i = Y_i(1) - Y_i(0)$,其中，$Y_i(1)$和$Y_i(0)$是个体 i 处于不同处理变量水平下因变量的潜在取值。显然，我们不可能同时观测到这两个潜在结果（Holland 1986）。故而如何在观察数据中重建这一反事实比较结构，是现代因果推断方法发展的出发点。在过去二十余年间，因果推论方法取得了长足进展，因果识别与估计策略日益精细化与系统化。在可忽略性（ignorability）假设等条件成立的前提下，研究者可以通过匹配、加权、工具变量或回归调整

等多种方式来获得平均处理效应的估计值（胡安宁 2025）。假设可能存在的协变量表示为X，有$\hat{\tau} = E[E(Y \mid D = 1, X) - E(Y \mid D = 0, X)]$。这些方法在理论上具有严格的识别基础，在实践中也显著提升了经验研究的有效性。

然而，一个越来越突出的现实问题是，大量的因果推断方法在估计阶段依赖具体的模型参数设定。例如，研究者通常需要对条件期望 $E(Y \mid D, X)$ 或倾向值得分 $e(X) = P(D = 1 \mid X)$ 进行基于参数化模型的估计（如线性模型或 logit 模型）。显然，若模型形式设定不当，即使因果识别条件成立，估计结果也可能出现系统性偏误。或者说，所估计出的结果是模型设定的一个“产物”。这种“模型依赖性”在高维协变量、非线性关系或复杂交互结构存在时尤为严重。换言之，当因果推断所依赖的统计模型在参数设定上过于“刚性”时，研究结论可能更多反映模型假设，而非数据本身所蕴含的信息。因此，如何突破预设模型参数形式的限制，使因果效应的估计更加呈现出数据自适应（data-adaptive）的特征，成为当前因果推断方法论的重要发展方向（Hill 2011; Kennedy 2016）。

在这一背景下，深度学习方法提供了一条非常具有潜力的分析路径。神经网络可以看成是一种高度灵活的非参数模型逼近器（Goodfellow et al. 2016）。例如，对于任意连续模型 $f(x)$，在适当条件下，存在参数 θ 使得神经网络 $f_\theta(x)$ 满足 $\sup_x |f_\theta(x) - f(x)| \to 0$，即所谓的通用逼近。相较于传统参数模型对模型形式的强假设，深度学习通过多层非线性变换来描述变量之间的复杂结构，从而降低模型

更多免费行业报告 并购家 www.ipoipo.net

设定“刚性”所带来的潜在偏差风险。在因果推断框架下，这也就意味着我们所需要的因变量的条件期望模型、倾向值的估计模型乃至潜在结果模型都可以通过数据驱动的方式进行逼近。

近年来，社会科学领域内围绕深度学习在因果推断领域的应用，主要体现在两个方向。第一个方向关注如何突破传统参数模型，更好的估计两个变量之间的因果效应，尤其是在高维协变量或复杂选择机制存在的情况下来降低混杂偏差（例如，Belloni et al. 2017）。第二个方向则关注因果机制的刻画。这一方向不仅要估计变量之间的因果效应，还要揭示因果链条的结构（Hu 2025）。例如，在中介分析中，研究者关心的是处理变量如何通过中介变量传递对因变量的影响。在这两个方面，深度学习方法都具有用武之地，这也是本文进行梳理的两个重点方向。

深度学习作为一种高度灵活的非参数数据分析工具，为突破模型参数设定、实现数据自适应估计提供了新的技术路径。在下面的讨论中，我们将分别梳理如何通过深度学习技术获取更为“纯净”的因果效应和刻画复杂的因果链条结构，以此展现出深度学习方法与社会科学因果推断技术之间的交叉融合潜力。经过几十年的发展，社会科学因果推断的理论识别框架已趋于成熟，但估计阶段长期存在的参数模型依赖性问题，仍然构成经验研究的重要瓶颈。无论是通过回归“控制变量”实现样本平衡，还是借助倾向值得分进行匹配与加权，其有效性都高度依赖于对条件期望模型或处理分配机制的参数化设定。一旦模型参数设定不当，即使可忽略性与重叠性等假设成立，估计结果仍

可能偏离真实因果结构。深度学习方法的引入，本质上并非对因果识别理论的替代，而是在估计层面提供了一种更为灵活、数据自适应的模型逼近工具，从而缓解模型错设所带来的系统性风险。

在处理效应估计方面，以 TARNet 及其扩展模型为代表，因果问题被重构为表征学习问题。通过学习共享的低维表示空间，该模型在该空间内增强了处理组与对照组的混淆因素平衡性，从而提升反事实预测的稳定性。进一步引入分布匹配正则项的 CFR 框架，则在表示层面显式约束两组分布的差异，使混淆因素的“平衡”成为优化目标的一部分。这一转向意味着，样本平衡不再仅仅是统计调整的副产品，而是模型结构设计中的核心环节。在因果机制分析方面，深度学习方法，尤其是基于归一化流的生成模型，为复杂中介结构的刻画提供了新的技术路径。相较于传统线性中介模型依赖有限维参数分解总效应的方式，归一化流通过可逆、单调的神经网络变换实现对变量条件分布的学习，使研究者能够在不预设分布形态的前提下进行反事实模拟。在自然间接效应等分析场景下，这种生成式建模框架为构造 Y(1, M(0)) 这样的嵌套反事实统计量提供了可操作的模拟机制。由此，因果中介分析不再局限于系数分解，而转向基于分布层面的干预、模拟与比较。

当然，社会科学研究者需要保持方法论上的清醒。第一，深度学习并不解决未观测混杂问题。若关键混杂变量缺失，再复杂的网络结构也无法弥补识别条件的缺陷。第二，高度灵活的模型意味着更高的数据需求与更复杂的计算结构，其有限样本性质、泛化误差界与渐近

更多免费行业报告 并购家 www.ipoipo.net

理论仍有待系统研究。第三，深度模型的“黑箱”特征也对社会科学强调的可解释性提出挑战。因此，将深度学习方法真正嵌入社会科学研究范式，必须在识别假设、统计理论与计算实现之间保持严密的逻辑衔接。从更宏观的视角看，深度学习推进的并非简单的技术替代，而是社会科学因果分析范式的一次结构性扩展。它将模型设定的权力从研究者主观假定转移至数据驱动的逼近过程，将因果估计从参数化建模推向分布级建模，将样本平衡与机制刻画纳入统一的表征学习与生成建模框架。在大样本、多模态与高维数据日益普及的背景下，这种方法论融合为社会科学提供了新的分析工具箱。未来的关键任务，不在于以深度学习取代传统因果方法，而在于构建一个兼具识别严谨性与估计灵活性的整合框架：在坚守反事实因果逻辑与识别假设的前提下，充分利用深度神经网络的通用逼近能力与生成能力，以实现更稳健、更精细、更贴近真实社会结构的因果推断，这或许正是人工智能时代社会科学方法论演进的一个重要方向。

6.2 深度学习与样本平衡

在观察性数据中，处理变量往往并非随机分配，处理组与对照组在基线协变量上的系统性差异会引入选择偏差，使“可比性”成为因果推断的首要前提。传统实践中，无论是回归控制、倾向得分匹配还是逆概率加权，其共同目标都是实现样本平衡：在尽可能相似的协变量条件下进行反事实比较。然而，这类方法在估计层面通常依赖对倾向得分或结果条件期望的参数化设定；当协变量维度升高、关系呈现非线性或存在复杂交互时，模型错设与重叠不足会显著放大外推风险，导致平衡更多停留在模型假设层面而非数据层面。深度学习为样本平衡提供了新的思路：通过表示学习将高维协变量映射到更适合比较的低维表征空间，并在该空间中增强处理组与对照组的分布重叠与可比性，从而提升反事实预测的稳定性与因果效应估计的稳健性。本节在回顾传统样本平衡工具的基础上，重点介绍以 TARNet 及其扩展为代表的深度学习方法如何把“平衡”从统计调整的副产品转化为模型结构设计与优化目标的一部分。

1. 传统样本平衡方法

在观察性研究中，由于处理变量 D 并非随机分配，处理组与对照组之间往往在基线协变量（混淆变量） X 上存在系统性差异，从而导致样本选择偏差。传统的社会科学研究中，线性回归模型常用“控制变量”的方法来对协变量进行“调整”，以期在观察性数据中实现类似实验设计的组间可比性。例如，我们可以拟合以下线性模型，$Y_i = \alpha + \tau D_i + \beta^\top X_i + \varepsilon_i$，其中 Y_i 为结果变量，D_i为处理变量，X_i 为一组

更多免费行业报告 并购家 www.ipoipo.net

控制变量（协变量）。在这个模型中，研究者常常将 τ 解释为在“控制”X 之后 D 对 Y 的处理效应。这一传统方法的思路是，通过在回归方程中加入混杂变量 X，力图在统计意义上比较“在相同 X 水平下”处理组与控制组的结果差异。然而，这种通过线性控制实现样本平衡的方式存在明显局限。首先，所谓“平衡”是建立在线性模型形式正确设定的前提之上。如果真实关系包含高阶非线性或复杂交互项，而模型未能正确指定，则模型错设会导致残余混杂，从而使 τ 的估计产生偏误。其次，当协变量维度较高时，简单线性控制难以实现在 D 的不同取值水平之间真正的混淆因素分布重叠与平衡。在处理组与控制组协变量分布差异较大的情况下，上述回归模型实际上更大程度的依赖回归模型对缺乏数据支撑的空间所进行的外推（extrapolation）（胡安宁 2025）。因此，传统线性回归通过“控制变量”达成平衡，本质上讲仍然是一种基于模型假设的平衡，而非数据层面的平衡，其有效性高度依赖于模型形式设定的正确性与协变量的充分重叠。一旦这些条件不满足，回归控制便不能真正消除混杂。

作为传统线性回归模型的一个替代方案，我们可以采用倾向值方法（Rosenbaum & Rubin 1983）。倾向值是个体在给定协变量 X 条件下接受处理的概率，$e(X) = P(D = 1 \mid X)$。在“可忽略性假设”（即 $Y(1),Y(0)) \perp D \mid X$）与“重叠性假设”（overlap，亦即 $0<e(X)<1$）成立的条件下，罗森鲍姆和鲁宾证明，$Y(1),Y(0)) \perp D \mid e(X)$ 。这意味着，在相同倾向值得分下，处理分配近似于随机分配。因此，倾向值成为在高维协变量情境下实现协变量平衡的降维工具。具体而言，研究者

可以采用两种基于倾向值的方法，一个是倾向值加权（propensity score weighting），另一个是倾向值匹配（propensity score matching）。倾向值加权的思想是通过构造逆概率权重，使样本在加权后形成一个“伪总体”，在该总体中处理变量与协变量独立。例如，在估计平均处理效应时，常用的逆概率加权权重为 $w_i = \frac{D_i}{e(X_i)} + \frac{1-D_i}{1-e(X_i)}$。在加权样本中，协变量分布在处理组与对照组之间实现平衡，从而可通过对 Y 的加权均值差来估计因果效应。加权方法的优点在于保留全部样本信息，并可自然推广至多处理水平或连续处理水平情形。但当 e(X) 接近 0 或 1 时，权重可能出现极端值，从而导致估计不稳定。相比之下，倾向值匹配在处理组与对照组之间寻找倾向值相近的个体来构造“可比”样本。最常见的形式为最近邻匹配，即为每个接受处理的个体选择一个或多个倾向值最接近的对照组个体，并以匹配样本的 Y 值差异估计因果效应。以针对实验组的处理效应（treatment effect for the treated）为例，研究者针对处理组个体 i，匹配其到对照组个体 j(i)，则针对实验组的处理效应可写为 $\hat{\tau} = \frac{1}{n_T}\sum_{i:D_i=1}(Y_i - Y_{j(i)})$。匹配方法的优势在于减少针对 Y 的建模（也就是不需要估计结果模型）。但无论是倾向值匹配还是倾向值加权，其都依赖于对倾向值的参数建模估计。从这个角度来说，倾向值方法虽然不要求研究者对于 Y 的数据生成过程进行参数建模，但在现实应用过程中仍然涉及到对于倾向值的建模估计。[①]

[①] 当然，也有学者利用非参数方法来估计倾向值，但是对于社会科学研究而言，大多数学者仍然采用传统的参数模型（如 logistic 回归）。

更多免费行业报告 并购家 www.ipoipo.net

2. 基于深度学习方法的样本平衡

那么，是否可以通过非参数的方法来达成实验组和控制组之间的平衡呢？深度学习方法提供了一个达成这一目标的思路。这方面的一个代表性方法是 TARNet（Treatment-Agnostic Representation Network）。TARNet 由 Uri Shalit、Fredrik D. Johansson 与 David Sontag（2017）提出，其核心思想是利用深度学习网络来学习一个“与处理无关”的 X 的表征空间，从而在该表示空间中达成协变量在实验组和控制组之间的分布平衡，从而更准确地估计反事实结果。

具体而言，TARNet 的网络结构分为两个阶段。第一阶段为针对混淆变量 X 利用深度学习网络来构建共享表征空间，记为 Φ(X)。该空间将原始协变量 X 映射至一个低维表示空间 Φ(X)。第二阶段则为两个分支网络（head networks）。假设处理变量取值不同时所带来的潜在结果分别表示为$\mu_1(x) = E[Y(1) \mid X = x]$和$\mu_0(x) = E[Y(0) \mid X = x]$，则研究者利用深度学习网络分别预测处理组与对照组的结果，即$\widehat{\mu_1}(x) = f_1\big(\Phi(x)\big)$和$\widehat{\mu_0}(x) = f_0\big(\Phi(x)\big)$。在此基础上，个体处理效应的估计量为$\hat{\tau}(x) = \widehat{\mu_1}(x) - \widehat{\mu_0}(x)$。

TARNet 的创新点在于其独特的深度学习网络设计。这个设计结构并不是如传统模型那样通过增加额外的参数化分布平衡约束来达成混淆因素的平衡，而是通过共享底层表征空间，学习到对实验组和控制组两组共同有预测力的 X 表征，这也减少了因为可能的样本稀疏带来的过拟合风险。在 TARNet 的基础上，很多新的方法也被开发出来。例如，TARNet 虽然通过共享表征来达成样本平衡，但这种平

衡性并没有明确的“要求”。换句话说，TARNet 没有显式保证在表征空间中实验组和控制组之间混淆因素分布一定接近。因此，当存在协变量分布不重叠时，模型仍可能在无数据支撑的“区域”进行外推。与之相比，CFR（Counterfactual Regression）则加入了分布匹配正则项。如果说 TARNet 的损失函数可以表示为$\mathcal{L}_T = \sum_{i=1}^{n}\left(Y_i - f_{T_i}\left(\Phi(X_i)\right)\right)^2$，那么 CFR 则增加了分布匹配的正则项。其损失模型为 $\mathcal{L}_C = \mathcal{L}_T + \lambda \cdot \mathrm{DIS}(\Phi(X) \mid T = 1, \Phi(X) \mid T = 0)$，其中 $\mathrm{DIS}(\cdot,\cdot)$为表征空间内部两组混淆因素之间的距离，$\lambda$ 为正则化系数。显然，CFR 因为加入了 $\mathrm{DIS}(\Phi(X) \mid T = 1, \Phi(X) \mid T = 0)$强制模型在表征空间中降低处理组与对照组的分布差异。因此，TARNet 可被视为更基础的表示学习框架，而 CFR 则在其基础上加入平衡约束以增强因果泛化能力。

从方法论角度看，TARNet 将因果推断问题转化为表征学习问题，通过学习一个低维表征 $\Phi(X)$，在该空间中处理组与对照组具有足够重叠，从而提升反事实预测的稳定性。其相比于传统模型，优势在于可以突破对于模型设定的约束、并可扩展至高维数据和多处理或连续处理情形。当然，和其他因果推论方法一样，如果数据中存在未观测的混杂变量，则深度学习同样无法很好的识别因果效应。但是这是数据质量问题，并不是模型问题。总体而言，TARNet 是深度学习因果推断领域的奠基性方法之一，它通过共享表征与双分支结构，为处理效应估计提供了一个结构清晰、易于实现且具有良好扩展性的神经网络框架。

更多免费行业报告 并购家 www.ipoipo.net

6.3 深度学习与因果中介分析

在社会科学与公共政策研究中，估计“是否有效、效应多大”往往只是第一步，更关键的问题常在于“为什么有效、通过何种渠道起作用”。因果中介分析正是将处理变量对结果变量的影响分解为直接效应与通过中介变量传递的间接效应，从而把因果效应转化为可解释的机制链条。传统中介分析多依赖线性结构模型与系数分解，在模型形式正确、关系近似线性且混杂控制充分等条件下具有清晰的解释框架；但一旦存在非线性关系、交互作用、多中介路径或高维混杂，参数化设定容易带来模型依赖性，甚至使“机制分解”结果更像模型假设的产物。此外，中介分析天然涉及嵌套反事实等“跨世界”量，其可识别与可估计高度敏感于假设与分布刻画方式。深度学习尤其是生成式分布建模为这一难题提供新路径：通过学习中介与结果的条件分布并支持反事实生成，可以在更少依赖预设分布形态的前提下开展机制推断，使中介分析从线性系数分解走向分布级干预、模拟与比较。本节将围绕这一思路，介绍深度学习方法如何在保持因果逻辑与识别纪律的前提下，提升复杂中介结构刻画的灵活性与稳健性。

1. 传统中介模型

社会科学研究除了探究从 D 到 Y 的因果关系之外，也常常希望了解这种因果关系是如何形成的。这一研究取向本质上要求研究者对 D 和 Y 之间的形成中介进行探索，也就是进行所谓的中介分析。传统中介分析的理论基础可追溯至 Reuben M. Baron 与 David A. Kenny 提出的经典分析框架（Baron & Kenny 1986）。该框架建立在线性回归模

型之上，通过一组嵌套回归模型来把 D 到 Y 的总效应（total effect）分解为直接效应（direct effect）与间接效应（indirect effect）。前者是 D 不通过任何中介变量来影响 Y 的效应，后者是 D 通过中介变量 M 来影响 Y 的效应。例如，在最基本的单中介变量的线性模型中，上述分析框架可表示为三组回归模型。首先估计自变量对结果变量的总效应，即拟合模型$Y = cD + \varepsilon_1$,其中 c 表示总效应。其次估计自变量对中介变量的影响，即拟合模型$M = aD + \varepsilon_2$，其中 a 表示路径 D→M 的系数。最后在同时控制 D 与 M 的情况下估计结果模型$Y = c'D + bM + \varepsilon_3$，其中 b 表示路径 M→Y 的系数，c' 表示控制中介后的直接效应。在线性模型设定与无交互项的条件下，间接效应可表示为 a*b，并满足以下分解关系：c=c′+ab。

在这一框架下，中介效应的显著性通常通过 Sobel 检验或 bootstrap 方法进行评估。通过上述描述不难看出，该方法的核心假设包括：（1）线性模型；（2）误差项独立同分布；（3）不存在遗漏混杂变量同时影响中介与结果；（4）无测量误差。在这些假设成立的条件下，间接效应可以被解释为自变量通过中介变量传递的因果效应强度。然而，这一传统中介分析路径本质上仍然依赖于线性结构模型的设定。在存在非线性关系、交互效应或多中介结构时，简单的系数乘积 a*b 不再等于总效应与直接效应之差，上述的效应分解关系就不再成立。可见，在传统的中介分析中，社会科学研究者仍然会受制于模型设定。除了复杂中介分析场景之外，传统的中介效应分析策略对于高维混杂变量的处理也常常捉襟见肘。例如，如果我们依旧采取上述嵌套线性模型

更多免费行业报告 并购家 www.ipoipo.net

的分析方法，则需要为每个中介变量及结果变量指定明确的条件分布形式。在最常见的线性正态框架下，中介变量 M 在基线混杂变量 X 与处理变量 D 条件下通常被假定为$\mathrm{M} \sim \mathrm{N}(\beta_0 + \beta_1 \mathrm{X} + \beta_2 \mathrm{D}, \sigma^2)$，并据此构造线性预测模型与正态误差项。这种方法在形式上简洁，但其有效性高度依赖于模型形式的正确性。一旦真实分布存在非线性关系、条件异方差、多峰结构或高阶交互，而模型未能准确对齐并进行刻画，即便目标因果效应在非参数意义下可识别，估计结果仍可能产生系统性偏差。因此，和上面谈到的因果关系的估计类似，在中介模型中，参数模型的主要局限不在于识别理论，而在于模型参数假设对分布结构的强约束。

2. 基于深度学习的中介分析

面对传统中介分析的困境，深度学习网络提供了新的分析思路和问题解决方案。不同于预先设定模型形式的参数模型，神经网络以数据为驱动，通过多层非线性变换自动学习变量之间的复杂关系，但在这里，深度学习网络需要学习的不仅是 D 到 Y 的关系，还涉及到中介变量 M。在具体操作上，社会科学研究者可以采用基于归一化流（normalizing flow）的深度神经网络框架来对中介变量和结果变量的条件分布进行建模（Balgi et al. 2025）。该方法的基本思路是：通过构造可逆、可微的单调神经网络变换，将原始变量群映射为标准正态变量群。之后，我们可以针对标准正态变量群进行某种干预，借助深度学习所获知的变量变换函数，通过逆变换则可生成符合干预条件的原始变量分布特征的样本。

我们这里可以举一个例子。假设我们这里关注的是中介变量 M。通过构建一个深度学习网络，可以将 M 转变为一个正态分布变量 Z，假设$h_{M|X,D}(X,D,M)$ 是针对 M 所构建的单调且可逆的从 M 到正态分布 Z 的变换函数。那么，我们就有$Z_M = h_{M|X,D}(X,D,M) \sim N(0,1)$。进而基于变量变换公式，我们可以得知 M 的条件密度为$f_{M|X,D}(m) = \phi(h_{M|X,D}(X,D,M) \mid \partial h_{M|X,D}(X,D,M)/\partial m \mid)$，其中 $\phi(\cdot)$为标准正态密度模型，第二项为雅可比修正因子。显然，只要能够利用深度学习网络学习该变换函数，便可完整刻画 M 和 Z 之间分布的转换关系，其逆模型$M = h^{-1}_{M|X,D}(X,D,Z_M)$则直接提供了从标准正态分布 Z 生成 M 分布的机制。

变量之间的变换在统计学研究中非常常见，但是传统的变量变换常基于参数分布的变换，分布形态通常由有限维参数（如均值和方差）决定。而在深度学习的神经网络中，分布形态由多层神经元的权重与激活模型共同决定，具有高度的非线性表达能力。网络通过对严格正值导数模型进行积分构造单调变换，从结构上保证可逆性与可微性，同时避免对具体模型参数形式施加限制。这使得模型能够适应复杂的偏态分布、多峰结构以及条件异方差等特征，而无需事先假设其具体分布形式。因此，利用深度学习网络来逼近原始变量和正态分布变量之间的变换函数有其独特优势。具体而言，这一深度学习网络不需要对条件分布作出线性或指数族假设。它也能够刻画非线性关系与高阶交互。同时，这一网络在形式上保持了可逆性，从而支持反事实模拟与干预分析。除此之外，基于深度学习网络来逼近分布变换函数可以

更多免费行业报告 并购家 www.ipoipo.net

后续帮助研究者模拟生成数据。例如，在训练完成得到对转换函数的逼近之后，只需从标准正态分布 Z 中抽样，并通过逆变换映射回原变量 M，即可生成满足复杂条件分布的模拟样本。这一性质使得在多中介或多路径因果结构下进行蒙特卡洛模拟变得非常方便。相比之下，在参数模型中，每个变量的分布形式需分别指定，这在模型误设时可能产生连锁偏误。

基于归一化流的深度神经网络框架的基本思想是：任意单调模型均可表示为严格正导数的积分。具体表示为$\mathrm{h}_{\mathrm{M|X,D}}(\mathrm{X},\mathrm{D},\mathrm{M};\tau,\omega)=\int_{\mathrm{o}}^{\mathrm{M}}\theta(\mathrm{t};\eta(\mathrm{X},\mathrm{D};\tau);\omega)\ \mathrm{dt}+\alpha\big(\eta(\mathrm{X},\mathrm{D};\tau)\big)$。其中嵌入网络 $\eta(\mathrm{X},\mathrm{D};\tau)$ 将 M 映射为低维表示，刻画其随 X 和 D 如何变化；被积模型网络 $\theta(\cdot)$ 通过严格正的激活模型输出导数值，从结构上保证$\partial\mathrm{h}_{\mathrm{M|X,D}}/\partial\mathrm{M}>0$，即确保全局单调与可逆性。该结构无需显式约束即可实现可逆变换，在网络容量充分时可逼近任意复杂的连续分布。训练过程基于最大似然原理。给定样本$\{(\mathrm{x_i},\mathrm{d_i},\mathrm{m_i})\}_{\mathrm{i=1}}^{\mathrm{n}}$，负对数似然函数为$-\sum_{\mathrm{i=1}}^{\mathrm{n}}\log\phi(\mathrm{h}_{\mathrm{M|X,D}}(\mathrm{x_i},\mathrm{d_i},\mathrm{m_i})\ -\sum_{\mathrm{i=1}}^{\mathrm{n}}\log|\partial\mathrm{h}_{\mathrm{M|X,D}}(\mathrm{x_i},\mathrm{d_i},\mathrm{m_i})/\partial\mathrm{m_i}|$。该损失函数衡量变换后变量偏离标准正态分布的程度。基于这一损失函数，并通过随机梯度下降优化网络参数。训练完成后，可从标准正态分布抽样 $\widetilde{\mathrm{Z_{M,J}}}$，并通过逆变换生成$\widetilde{\mathrm{M_J}}=\mathrm{h}^{-1}_{\mathrm{M|X,D}(\mathrm{X,D},\widetilde{\mathrm{Z_{M,J}}})}$。

上面的例子仅仅是针对中介变量 M 来展示基于归一化流的深度学习网络如何达成对 M 和 Z 之间转换函数的逼近。在中介分析的场景下，研究者可以基于上述的分析流程对涉及到的所有变量进行转换函数的逼近。这里可以设想一个最基本的中介场景，其涉及到混淆

因素 X，自变量 D，中介变量 M 和结果变量 Y。那么，利用深度学习网络，我们可以将其转化为四元正态分布，其中对应于 X 的正态分布变量 Z_X 为边际分布，对应于 D 的正态分布变量 Z_D 是一个基于 Z_X 的条件分布，对应于 M 的正态分布 Z_M 则是基于 Z_X 和 Z_D 的条件分布，对应于 Y 的正态分布 Z_Y 则是基于 Z_X、Z_D 和 Z_M 的条件分布。那么，为了探索中介效应，我们就可以先对这一多元正态分布进行某种设定（或者称为干预），之后利用转换函数的反函数来模拟原始数据下基于这种干预的数据形态，两相对比完成中介效应的估计。例如，基于反事实因果推理框架，自然间接效应可以定义为 $\mathrm{NIE} = \mathrm{E}[\mathrm{Y}(1, \mathrm{M}(1)) - \mathrm{Y}(1, \mathrm{M}(0))]$。其含义是，在反事实框架下衡量处理变量 D 通过中介变量 M 影响结果变量 Y 的那一部分因果效应。具体而言，M(d) 表示当处理被设定为 d 时中介变量的潜在取值，Y(d,m) 表示当处理被设为 d 且中介被人为设定为 m 时结果变量的潜在结果，因此 Y(1,M(1)) 表示在处理状态为 1 的情况下，且中介取其自然处理值（D=1 时 M 的自然取值）时的结果，而 Y(1,M(0)) 则表示在处理状态为 1 的情况下，中介变量被强制设为其在对照状态（D=0）下的取值时的结果；二者之差刻画的是在保持处理变量取值为 1 的条件下，仅通过改变中介变量取值（从 M(0)变为 M(1)）所引起的结果 Y 的变化。对该差值取总体期望，得到平均自然间接效应。

显然，Y(1,M(0)) 无法在经验上被直接观测，因为它是一个典型的嵌套反事实并且属于“跨世界”构造：M(0) 表示个体在未接受处理时中介变量的潜在值，而 Y(1,·) 则描述个体在接受处理的条件下结果 Y

更多免费行业报告 并购家 www.ipoipo.net

的生成机制。要观测 Y(1,M(0))，我们必须同时知道同一分析单位在 D=0 时的中介潜在值以及在 D=1 时的结果潜在值，但每个个体在现实中只能处于一种处理状态，因此 M(0) 与 M(1)不可能同时实现。因此，Y(1,m)只能在 D=1 时观测。换言之，Y(1,M(0))将两个互斥世界（D=0 与 D=1）的信息嵌套在同一个表达式中，这种“跨世界反事实”在数据中没有直接对应的观测量，因此只能在增设额外识别假设（如顺序可忽略性或结构模型假设）下通过模型间接推断。在传统的中介分析中，研究者可以通过线性模型进行预测。但是，如果需要突破模型参数的设定，我们可以采用上述的深度学习网络来对这个现实中“不可能”的统计量进行模拟。具体而言，通过归一化流神经网络确定 X、D、M 和 Y 与 Z_X、Z_D、Z_M 和 Z_Y 的对应关系。由于 Y 的取值取决于 D 和 M，Z_Y 的取值也取决于 Z_D 和 Z_M。那么，研究者可以固定 Z_D，并干预让 Z_M 取值为 Z_D 反事实取值时的取值，之后利用转换函数的反函数就可以获得 Y(1,M(0))的模拟数据。此时，从观测数据中，我们可以获知 Y(1,M(1))，模拟数据我们可以获知 Y(1,M(0))，二者的均值之差就是对自然间接效应的估计。在这个操作过程中，我们没有如上文所述那样采取传统方法进行参数建模。换言之，我们在一定意义上突破了参数限制。

综上所述，深度神经网络不仅替代了参数模型中的模型设定，而且可以通过可逆非线性变换实现干预分布的学习与生成。相比依赖有限维参数刻画分布形态的传统方法，该框架在理论上能够逼近任意连续条件分布，将条件分布建模从强假设驱动转向数据驱动，显著降低了分布误设所带来的偏误风险。在大样本情境下，这种数据驱动的分

布建模方式为复杂因果结构中的中介分析与反事实模拟提供了更具稳健性的估计基础。当然，这种高度灵活性以更高的数据需求和计算成本为代价。通常而言，深度神经网络需要较大的样本规模以稳定估计参数，其有限样本性质仍有待进一步理论研究。但是大样本条件满足的情况下，深度学习方法能够显著降低模型设定错误的风险，提供更稳健的分布估计与因果效应模拟结果。

更多免费行业报告 并购家 www.ipoipo.net

参考文献

[1] 胡安宁. 《应用统计因果推论（第二版）》. 上海：复旦大学出版社， 2025。

[2] Angrist, Joshua D. and Jörn-Steffen Pischke. 2009. Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press.

[3] Balgi, Sourabh, Adel Daoud, Jose M. Pena, Geoffrey T. Wodtke, and Jesse Zhou. "Deep learning with DAGs." Sociological Methods & Research 54, no. 4 (2025): 1624-1682.

[4] Baron, Reuben M., and David A. Kenny. "The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations." Journal of personality and social psychology 51, no. 6 (1986): 1173.

[5] Belloni, Alexandre, Victor Chernozhukov, Iván Fernández-Val, and Christian Hansen. "Program evaluation and causal inference with high-dimensional data." Econometrica 85, no. 1 (2017): 233-298.

[6] Goodfellow, Ian, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. Deep learning. Vol. 1, no. 2. Cambridge: MIT press, 2016.

[7] Hill, Jennifer L. "Bayesian nonparametric modeling for causal inference." Journal of Computational and Graphical Statistics 20, no. 1 (2011): 217-240.

[8] Holland, Paul W. "Statistics and causal inference." Journal of the American statistical Association 81, no. 396 (1986): 945-960.

[9] Hu, Anning. Statistical causal mediation analysis with R. Springer, 2025.

[10] Imbens, Guido and Donald Rubin. 2015. Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. NY: Cambridge University Press.

[11] Kennedy, Edward H. "Semiparametric theory and empirical processes in causal inference." In Statistical causal inferences and their applications in public health research, pp. 141-167. Cham: Springer International Publishing, 2016.

[12] Morgan, Stephen L. and Christopher Winship. 2014. Counterfactuals and Causal Inference: Methods and Principles for Social Research. NY: Cambridge University Press.

[13] Rosenbaum, Paul R., and Donald B. Rubin. "The central role of the propensity score in observational studies for causal effects." Biometrika 70, no. 1 (1983): 41-55.

[14] Shalit, Uri, Fredrik D. Johansson, and David Sontag. "Estimating individual treatment effect: generalization bounds and algorithms." In International conference on machine learning, pp. 3076-3085. PMLR, 2017.

第七章 从文本到多模态的拓展：计算社会科学智能发展前沿与趋势[①]

计算社会科学（Computational Social Science, CSS）从诞生之初就带有一种“方法论的野心”：用可计算的方式描述社会结构、行为互动与制度运行，使社会研究不仅能够解释过去，也能在更严格的证据约束下评估政策、推演风险与识别机制。进入大模型时代，这一野心第一次在“数据—模型—推理—验证”的全链条上获得了新的技术支点：文本不再只是研究材料，而是可被结构化、可被建模、可被对齐到行为与制度过程的“可计算证据”；同时，社会世界的证据形态也不再局限于文本与统计表，而是扩展到图网络、时空轨迹、图像/视频、语音与多源传感数据所构成的多模态事实流。由此，计算社会科学正在从以文本与变量为中心的经验分析，迈向以多模态证据与复杂系统为中心的机制建模与系统评估。

本章旨在呈现 AI 如何重塑计算社会科学的三类关键能力。其一，是规模化证据提取与知识组织：大模型显著降低了从非结构化材料中抽取事实、概念与关系的成本，使政策文本、历史档案、舆情内容与学术文献能够被更系统地编码为可检索、可复核的结构化证据。其二，是结构与互动的可计算表达：图神经网络、表示学习与多智能体框架，让社会网络、组织关系与集体行为的动态演化能够被更细粒度地刻画，从而为“结构性因果”与“机制性解释”提供新的建模语言。其三，

[①] 本章由复旦大学新闻学院周葆华教授，博士生吴雨晴、张凌赟合作撰写。

周葆华教授为国家万人计划哲学社会科学领军人才（2021），教育部首批青年长江学者（2015）。

更多免费行业报告 并购家 www.ipoipo.net

是从解释到干预的闭环：强化学习、仿真与数字孪生式框架，使政策评估不再只停留在事后归因，而能够在可控的情景空间中进行压力测试、分布效应识别与方案比较，推动研究进入“可验证的干预推理”阶段。

需要强调的是，能力扩张并不自动意味着知识进步。多模态证据更丰富，也更容易带来噪声、偏差与不可见的选择机制；模型推理更强，也更可能产生看似合理但不可检验的叙事。因此，本章在呈现技术路径的同时，将以“数据—理论—机制”闭环为基本标准，强调可复核证据链、可解释假设与可审计流程的重要性，并讨论在数据合规、算法偏差与研究责任等方面的治理要求。归根结底，AI 赋能计算社会科学的价值，不在于替代社会科学的判断，而在于让社会科学在更高密度、更高透明度的证据体系中，建立更可靠的解释与更稳健的行动知识。

本章提要：

- 大语言模型与多模态大模型正在重塑计算社会科学的内容分析范式：从文本主题建模、分类与生成，到图像/视频/声音/数字痕迹等多模态社会数据的理解、生成与模拟，AI 逐步成为连接社会数据与研究推理的通用“方法引擎”。
- 在文本分析层面，大语言模型显著缓解了“人工编码高成本”与“传统算法难解复杂语境”的双重困境，在主题标注、情感与立场识别、信息失序检测等任务中展现出效率与可解释性优势，但在文化语境、情感共鸣与隐性语义等仍需人工核验与理论约束。
- 面向分类与识别任务，大模型在情绪隐晦、话题极化与价值对齐情境下易出现系统性偏差，提示计算社科应用必须将偏差诊断、不确定性披露与分层校验纳入标准流程。
- 在文本生成环节，大模型可用于解释、摘要与信息提取，显著降低海量材料处理负荷并提升结构化信息供给能力，但需警惕“压缩导致的情绪与少数观点损失”、以及生成解释与结论外推带来的误导风险，强调人机协作的证据链与审阅机制。
- 多模态大模型扩展了计算社会科学的数据边界：在媒体内容编码、虚假内容与复杂语用识别、数字行为痕迹分析、城市感知与空间治理等方向提供零/少样本编码能力，并通过跨模态推理提升对复杂社会语境的识别与解释潜力。
- 多模态生成与多模态智能体为“可控社会情境生成—硅样本实验—多智能体仿真”提供新工具，使社会互动的实验化与可重复性

更多免费行业报告 并购家 www.ipoipo.net

显著增强；外部效度、模态整合深度推理与安全可控仍是关键瓶颈，需要以基准评测、伦理合规与可解释框架推动规范化落地。

7.1 大语言模型赋能计算社会科学文本分析

文本分析是计算社会科学中的核心任务。传统文本分析方法长期面临双重困境：一方面，人工编码耗时长、成本高且难以扩展，同时易受主观偏见影响；另一方面，早期计算方法技术门槛高，且难以处理需要结合语境推理的复杂任务[1]。大语言模型的兴起改变了这一状况。本节将从自动化主题建模、文本分类、文本生成三大任务聚焦大语言模型如何赋能计算社会科学文本分析。

1. 基于大语言模型的自动化主题建模

主题建模，即自动识别、分析和界定文本数据的主题类别，属于“自下而上、数据驱动”的归纳式文本分析任务。大语言模型能够凭借其出色的语义理解能力提高主题建模效率。有研究者对比了不同方法对在线脑肿瘤支持论坛帖子进行主题建模的效果：人工编码耗时111 分钟，而 GPT-4 仅需 1 分钟，且与人工编码结果一致性高[2]。另有研究显示，在事实类新闻文本主题标注中，尽管大模型与人类标注者的整体读者偏好排名表现持平，但大模型能生成比人类更高特异性、更具体的标签，更契合读者对精准事实的偏好，在叙事性文本（如小说、新闻）的主题标注中，大模型表现已达到人类水平，对比传统 LDA 模型受语法与词汇噪音影响，且生成的主题缺乏解释力，大模型能够生成具备更高可解释性、语义丰富度和主题多样性的标签[3]；在分析社交媒体上简短且口语化的用户评论时，经人工验证，GPT-3 主题建模准确率超过传统的 LDA 模型，高达 97.7%[4]。

更多免费行业报告 并购家 www.ipoipo.net

然而，研究指出大语言模型在情感共鸣、文化独特性等深层语义方面的编码结果仍需人工验证。例如，有研究者对比人工和模型编码结果发现，在对健康论坛帖子进行主题分析时，人类识别出了更多微妙的语义，而大语言模型较少发现医患互动中的细微差别[2]；在分析国际学生访谈时，人类研究者能读懂学生在提到“家庭支持”时的文化责任感，而 ChatGPT 倾向于将其简化为通用的“支持系统”，抹平了背后的文化差异[5]。

同时，大语言模型的“价值对齐”倾向可能会让编码类别更加中立。研究者发现，在分析 YouTube 平台中关于乞讨者的视频评论时，GPT-4 倾向于使用中立、去政治化的标签（如“社会偏见”），而人类则使用更具批判性的标签（如“极端仇恨言论”），且由于安全机制，GPT-4 倾向于将仇恨言论归类为温和的社会问题，缺乏人类研究者对权力结构和历史背景的敏感度[6]。还有研究者在分析抖音和微博文本数据时指出 AI 倾向于将事实陈述或复杂立场归类为中立[7]。

2. 基于大语言模型的文本分类

大语言模型根据预定义的类别对文本进行分类，属于“自上而下、理论驱动”的演绎式文本分析任务，具体包括情感分析、滥用语言检测、信息失序检测、立场及意识形态检测等常见任务。

（1）情感分析

大模型被用来识别和提取文本中的情感态度。基于较强的识别特定上下文语境情感的能力，从正向、负向、中立的三元情感分类，到

包含喜悦、愤怒、悲伤、惊讶等多元的情感分类，大模型分析的整体效果较好[8]。关于社交媒体平台消费者对于品牌转型的情感态度，研究显示，大模型的正向、负向、中立三元情感分类结果优于传统的机器学习分类模型，且与人类标注呈现高一致性[9]。关于技术争议事件的微博帖子情感分类，研究者将其分为喜悦、愤怒、悲伤、恐惧、期待、惊讶、信任、厌恶 8 个情感类别，其中 DeepSeek-V3 表现最佳，准确率达到 0.80，F1 分数为 0.71[10]。

大模型情感分类效果受多种因素影响，如文本长度、分类任务复杂度、话题极化程度、文本线索、情绪隐晦程度等。从分类文本长度看，GPT-4 在社交媒体短文本上的分类效果非常出色，对于较长的新闻文章，GPT-4 在负面情感分类上的准确率大幅下降；从分类任务复杂度看，GPT-4 二元情感分类任务效果优于多元分类任务[11]。话题极化程度也会影响分类效果，比如当大语言模型处理高度极化话题（如“移民”）时，会表现出强烈的系统性偏见，模型会利用对各政党在该政策上立场的先验预期，将带有左翼政党标签的声明判定为更积极，将带有右翼政党标签的声明判定为更消极。关于分类文本信息线索，大模型比人类更容易受到政党线索的干扰，表现出比人类更严重的偏见[12]。从情绪隐晦程度看，当语句情绪非常直接时，大模型与人类高度一致，而当文本缺乏明显情绪时，大模型的情感提示往往会根据文中描述的事件好坏来分配积极或消极标签，而非基于作者的态度[13]。

（2）滥用语言检测

滥用语言检测指对具有攻击性的恶意语言的识别。恶意语言的滥

更多免费行业报告 并购家 www.ipoipo.net

用，因其加剧社会歧视，已成为一个至关重要的社会问题，涉及仇恨言论（Hate Speech）、污名化（Stigma）、冒犯语言（Offensive Language）、欺凌（Cyberbullying）、不文明言论（Incivility）、阴谋论（Conspiracy）等。虽然大模型具有深层语义理解力，但这一任务依赖于对社会规范和文化背景的理解，大模型处理起来仍有较大的优化空间[14]。

目前，滥用语言检测存在以下局限性：第一，大模型能较好地识别显性的滥用语言，但仍容易忽略隐性滥用语言，存在分类偏见。研究显示大模型对粗俗语言、暴力威胁和侮辱等显性不文明言论的识别都非常可靠，但对于一些隐性言论如刻板印象和歧视的召回率显著偏低，讽刺和轻蔑最难被捕获[15]；另有研究指出，虽然 AI 能发现人类忽略的政治细微差别，但容易遗漏隐性不文明行为[16]；阴谋论文章因包含缺乏事实依据的推论，叙事连贯性和一致性不足，也难以被识别[17]。第二，大模型在需要深层文化背景理解的任务上还有优化空间，有研究者使用 ChatGPT 来检测伊斯兰恐惧症（指针对穆斯林的偏见和歧视言论），发现 ChatGPT 无法完全替代领域专家，并且在标注数据集的表现上采取了保守态度，优先考虑精确率而非召回率，虽然减少了误报（False Positives），但是漏掉了许多隐晦或复杂的真实案例[18]。第三，女性受到的网络伤害被低估或漏判，存在性别偏差。研究指出由于针对女性政治家的攻击更多集中在隐性不文明言论（如刻板印象和歧视）上，女性歧视可能会被低估或漏判[15]。第四，语言覆盖范围是大模型的瓶颈，大语言模型在非英语攻击性文本检测上的表现有待提高[19][20]。

研究主要通过大模型执行优化和数据集开发两种方式来提升检测效果。在大模型执行优化上，通过参数微调、提示词工程、思维链等方式来提升大模型在特定任务上的表现[16][21]；在数据集开发上，大模型具有卓越的泛化能力和适应性，需要构建更丰富、更具文化底蕴的多样化数据集，以进一步释放大语言模型的潜力[22]。

（3）信息失序检测

信息失序（Information Disorder）检测，是指大语言模型对包含错误或虚假信息的文本的识别和过滤，如谣言（Rumor）、虚假信息（Disinformation）及错误信息（Misinformation）的检测，以及事实核查（Fact-Checking）等。失序信息的传播威胁了社会稳定和公众信任[23][24][25]，提升其检测效果是重中之重。

研究通过多种方式提升效果：第一，通过引入外部数据提升失序信息的检测效果，比如辟谣列表、事件信息、垂直领域知识库。通过引入辟谣列表（Debunk Lists）作为外部知识结构，大多数大模型提升了检测社交媒体上虚假健康信息的能力，其中 Claude 提升最显著，从零样本的 76.1%提高到 100 条样本上下文的 84.5%[26]。第二，自定义虚假信息特征指标，将指标结果引入大模型。有研究制定 35 个虚假信息具体评估指标，将数值引入中外模型（GPT-4、ERNIE-3.5-128K、GLM-4 PLUS），分类效果显著提升，其中 GPT-4 的准确率达到了 99.2%[27]。第三，面对真假参半的信息谣言检测，多模型协作策略成为效果提升的方法之一。研究者使用三种不同的大模型（GPT-3.5、Claude-2 和 LLaMA-2）组成“评估小组”进行谣言检测，随后由第四个

更多免费行业报告 并购家 www.ipoipo.net

大模型作为“最终仲裁者”检测这种“混合型”谣言，表现出鲁棒性，避免单一大模型因为训练数据带来的偏见或知识盲区[28]。

目前，基于大模型的信息失序检测也存在局限性：第一，实时检测带来的巨大算力需求削弱了大模型在动态、真实的社交媒体环境中的可扩展性和实际应用[25]；第二，大模型改写后的假新闻极大地阻碍了现有的假新闻检测”或“严重阻碍了现有的假新闻检测[29]；第三，大模型在低资源语言的性能上存在明显局限。研究者使用伊朗事实核查网站数据集，发现对波斯语新闻声明进行事实核查时，大模型整体准确率偏低，即使经过微调并结合英文翻译，对波斯语新闻的事实核查准确率最高也没有超过 55%[30]。

（4）立场及意识形态检测

立场检测旨在使用大语言模型识别文本作者对目标的态度，促进理解社会群体中的共识与分歧[23][24]，是舆情分析、政策研判、公共话语研究的核心技术[31]。

立场检测常常表现为文本作者支持、反对或中立的特定倾向。2023 年 Gilardi 等人通过评估一条推文对美国互联网立法的立场，已经发现 ChatGPT 的表现显著优于众包工人（来自 MTurk 平台）[32]。虽然有研究显示大语言模型并不一定比微调的传统模型准确率高，但其在零样本设置下仍具竞争力[33]。

需留意大模型会出现的偏差：第一，中立偏差。有研究在对社交媒体文本的主要叙事进行立场倾向分类时，发现由于模型价值观对齐

（Value Alignment），无论是 GPT 还是 Claude，大模型最终划定为“中立”阵营的文本数量，远远高于人工深度阅读后分类的数量，表明大模型生成的立场标签通常更加温和且缺乏政治色彩[7]；另有研究在识别和分类社交媒体上关于人乳头瘤病毒（HPV）疫苗公众讨论立场时发现，模型在处理“中立”消息时表现最差，准确率仅在 54%左右，显著低于支持和反对类别[34]。第二，支持和反对的不平衡性。在针对 Facebook 的长文本立场检测中，模型识别“反对”立场的准确率显著高于“支持”立场[34]；在对低资源语言（爱沙尼亚语）和复杂话题（移民）进行新闻文本的立场检测时，也发现所有微调模型在识别反对立场时比识别支持立场更准确[35]。

意识形态检测是立场检测在政治科学领域中的重要应用，指研究人员通过文本识别行为者背后潜在的思想观念、党派组织等[24]。尽管大语言模型的训练数据高度依赖英语和美国语境，大模型也能适配非英语语境的意识形态检测。根据 X（原 Twitter）单条消息识别政客政治派别的研究发现，GPT-4 在所有本土语言非英语的国家背景下（如土耳其、波兰、瑞典等）均比监督模型和人类编码员表现得更准确[36]；也有研究显示 GPT-4 的意识形态分类可以适配不同语言、不同国家的政治语境，效果与人类专家标注的一致性持平[11]。

3. 基于大语言模型的文本生成

文本生成任务是自然语言处理领域最具挑战性和创造性的任务之一[37]。在计算社会科学领域，研究人员基于大语言模型强大的文本生成能力，根据目标自动生成连贯、流畅且有意义的文本[23]，常见任务

更多免费行业报告 并购家 www.ipoipo.net

包括文本解释、文本摘要、文本信息提取。

（1）文本解释

文本解释即使用大模型对分析结果进行解释，能够增强结果的可解释性（Explainability）、可信度（Credibility）、透明性（Transparency）[23][38]，减少“幻觉”[39]。有研究者提出了“解释链（Chain of Explanation）”方法，旨在指导大模型生成自然语言解释，说明为什么某段文本被归类为仇恨言论[14]，成为指导大模型进行高质量解释的常见方法[23]。

大模型文本解释对结果有纠错能力[40]，能提高对语言现象和理论的理解与评估[23]。研究显示，大模型生成的中间解释越充分，最终立场检测的准确率越高[41]。另有研究发现，解释的质量会随着任务复杂度的增加而下降，在对比“解释是否有霸凌”和更细粒度的“解释是哪种类型的霸凌（基于种族、性别、年龄、宗教）”时，部分模型（如 Claude 系列）的解释质量出现了明显下滑[42]。

由于大模型的预训练语料库偏见及经历了安全性微调[43]，大模型用于文本解释仍存在局限性。研究发现，尽管 GPT 生成的文本解释与人类评估者之间存在高一致性（88.87%），在讽刺等语言模式识别方面表现卓越，但大模型无法将发现与理论框架或情境联系起来进行解读，需要人类合作来完成更深层次的解释性工作[44]。也有研究指出，由于大模型具有语言和文化偏见，ChatGPT 难以准确理解非英语（如德语）本土内容的文化语境，无法正确解释欧洲公共服务媒体的“多样性”[45]。还有研究指出，要警惕处理敏感议题（如网络暴力）时的模型偏差，一些开源模型（如 Llama）可能会因为安全对齐而拒绝输

出分析理由[42]。

（2）文本摘要

文本摘要指利用大语言模型对长文本进行语义压缩，目的是在保持语义完整性的前提下，生成一段连贯的、易读的文本以概括原文本的核心内容。大模型文本摘要打破了传统抽取式（Extractive）和摘要式（Abstractive）方法的边界，在文本连贯性、流畅性和整体写作质量上实现了对传统技术的超越[46]。

面对互联网文本内容（新闻文章、科学论文、报告文件、公民评论等）的指数级增长，大模型文本摘要在显著降低海量信息处理负荷、快速提炼核心要义方面发挥着重要作用。有研究使用 ChatGPT 对不同利益相关群体就城市集约化的反馈进行文本摘要，高效地呈现出不同群体的动机分歧，为政府部门准备应对策略和公共宣讲要点提供了参考[47]。在面对长文本的摘要中，“先压缩、再分析”是一个有效的提效范式。面对长达数万字的金融报告，研究者采用了基于 Google LongT5 模型的“两步式”摘要框架以防止丢失关键信息或超出模型处理极限：先对长文本分块后进行分别摘要，然后将各部分摘要片段拼接起来再次进行聚合摘要。最终生成的摘要文本能有效保留核心商业术语、财务指标以及业绩表现指标，并且有效提升下游的情感、收益表现分析性能[48]。

基于大模型进行文本摘要也有一定的局限性，比如大模型无法捕捉细微的情感差异。通过对比公众在社交媒体上的自然反应，研究发现 Meta-AI 摘要虽然捕捉了事实性信息，但完全缺失了代词（如

更多免费行业报告 并购家 www.ipoipo.net

“we”“my”“us”），无法体现人类评论中建立社区韧性的人际连接，缺乏风险沟通的情感效果[49]。其次，大模型在提升效率的同时，仍需人工进行辅助监督以避免模型偏见。有研究验证了在少样本学习下，Llama-3 能够对复杂的定性访谈数据进行有效的主题摘要，其中约 80%的摘要质量被评为合格或良好，足以辅助研究人员进行后续的社会心理分析，但仍需人工监督以修正幻觉和过度阐述等偏差[50]；另有研究显示，模型在压缩文本过程中，会倾向于过滤掉负面词汇，导致摘要文本的整体情感分布向“中性”或“积极”偏移，使得下游的情感分类可能带有过于乐观的偏见[48]；还有研究指出，大模型在文本摘要时会向较长或重复频率高的文本倾斜，从而忽略细微的语境和少数派意见，因此机器生成的摘要并不能完全脱离人工审核[47]。

（3）文本信息提取

文本信息提取指大语言模型从非结构化文本中识别并提取特定类别的实体、关系或事件的结构化信息，分为命名实体识别（Named Entity Recognition）、关系抽取（Relation Extraction）和事件抽取（Event Extraction）。命名实体识别指识别文本中的实体（如人名、地名、组织名），关系抽取指识别两个实体之间的静态关系（如社会关系、因果关系），事件抽取指识别动态的状态变化，它不仅包含实体，还包含动作（触发词）以及围绕该动作展开的多个属性（论元），按照 5W1H 原则即能通过回答“Who”（谁参与）、“When”（何时发生）、“Where”（何地发生）、“What”（发生了什么）、“Why”（为什么发生）以及“How”（如何发生）来概述一个完整事件（如网络事件）[51][52]。

基于大模型的信息提取能够通过查询和抽取非结构化文本，对目标信息进行概览与跟踪，并取得较好效果，问题涉及法律法规文本[53]、学术期刊文章[54]、企业报告[55]、新闻媒体[56]、事故报告[57]、社交媒体平台[58]等。以提取“道歉行为”为例，研究发现 GPT-4 的提取准确率非常接近人类标注员[59]；还有研究证实大模型进行实体提取的高效性及产生极少幻觉，并指出大模型从长篇幅、非结构化的新闻报道文本中，识别并提取记者引用的各类消息来源（如政客、官僚或普通公民），产生第一类错误（假阳性）和第二类错误（假阴性）的错误率均不到 5%，准确性高，而从多个互联网来源提取和整合目标人物的传记信息时，高级提示词工程极其重要[60]。除了提示词工程外，思维链策略也能提升长文本信息提取的效果。有研究者从海量新闻报道中提取非法跨国肾脏交易的地理位置及相关参与者的角色，其中，Llama3.3 模型展现了高提取精度，思维链策略的使用使得提取准确率达到 86%～95%，被证明比传统的 NER 和常规机器学习方法更加精确，提升了处理长文本的噪音数据抗干扰能力，并减少“幻觉”[56]。

基于大语言模型信息提取的准确性，部分研究将其提取结果作为下游统计分析的基础数据支撑。有研究者对过去 20 多年的非法交易网络进行了时空模式纵向分析，并期望进一步进行影响因素建模[56]。也有研究者从新闻正文中提取严重程度、地点和伤亡人数等数据用于量化灾害的实际影响，用于对严重程度与伤亡的相关性分析，以及分析地理易感性与灾害影响的关系[61]。

4. 小结

更多免费行业报告 并购家 www.ipoipo.net

本节探讨了大语言模型如何破解计算社会科学文本分析中，人工编码成本高、传统算法门槛高且难解复杂语境的“双重困境”。大语言模型主要从三大核心任务全面赋能该领域：首先是自动化主题建模，它通过数据驱动的归纳分析识别文本主题，不仅将编码耗时从数小时大幅缩至数分钟，还能有效规避人工偏见；其次是文本分类，其通过理论驱动的演绎分析，基于研究者预设的框架高效归类文本，已被广泛应用于情感、滥用语言、信息失序、立场与意识形态检测；最后是文本生成，依托强大的语言生成能力，大语言模型能够出色地完成文本解释、摘要与结构化信息提取等任务。大语言模型正以其跨语境理解与生成力，为计算社会科学的文本分析提供了兼具效率与深度的方法论引擎。本节也指出了该方法运用于社会科学研究中目前所存在的局限和边界。

7.2 多模态大模型赋能计算社会科学：理解、生成与模拟

多模态大模型（Multimodal Large Language Models，MLLMs）以大语言模型为基座，连接多个模态的编码器与解码器，使其能够理解（和生成）多种模态数据[62]，迅速发展成为连接丰富社会数据与计算社科研究的重要工具。研究者不仅能借助多模态大模型理解复杂的社会数据，还能实现社会情境的受控生成，并构建基于多模态智能代理的社会仿真模拟。以下，本节将从理解、生成、模拟等三方面介绍多模态大模型如何持续赋能计算社会科学研究。

1. 以多模态大模型理解社会数据

随着智能设备的普及，人类社会活动正被更加全面地记录为多种模态的数据，包括社交媒体平台上的图文帖子、表情包与短视频，依托智能手机产生的数字痕迹（如屏幕截图），以及与地理空间密切相关的街景图像等。多模态大模型的发展进一步推动了计算社会科学对多种模态的社会数据的理解与分析。

（1）媒体内容编码

随着图文社交与短视频平台的兴起，媒体内容消费日益趋向多模态化，这也要求计算社科研究者们更加关注对图像、视频等媒体内容的分析。传统计算方法通常依赖针对特定任务训练的机器学习模型，或者通过调用现成的 API 来实现特定变量的编码[63]。而多模态大模型提供了一种新的技术路径：在零样本（Zero-shot）或少样本（Few-shot）条件下，仅通过提示词（Prompt）引导即可完成特征识别与语

更多免费行业报告 并购家 www.ipoipo.net

义编码，从而降低多模态内容分析的成本与技术门槛。

现有研究通常使用多模态大模型对图像（视频帧）进行编码。研究显示，在基本视觉要素识别方面，多模态大模型已表现出较强的能力。例如，有研究使用 GPT-4o 识别图像中的人物、场景、物体，并与人工编码结果进行交叉验证，匹配率达到约 91%[64]。此外，多模态大模型也能对具有一定主观性的语义变量进行编码，如主题、情感等。例如，有研究使用 LLaVA-Llama3 编码图文帖子的主题与情感表达，初步验证准确率均超过 80%[65]。在一些基于社会学理论构建的变量编码任务中，多模态大模型同样展现出潜力。例如，在在线约会平台研究中，研究者利用 InternVL-G 对用户照片的“面部美感”和“氛围（包括社会资本等）”进行评分，其结果与人工评分高度一致，其中美感与人工评分的 Pearson 相关系数超过 0.80，氛围的相关系数超过 0.70[66]。同时，也有研究指出，在更为复杂的变量编码任务中，由模型直接输出的二元分类结果缺乏鲁棒性。相比之下，由多模态大模型生成图像描述，再基于描述进行分类的两阶段方法，往往能获得更稳健的编码结果[67]。

相较图文内容，视频具有更复杂的多模态结构，不仅包含图像静帧，还包括时间维度上的动态图像、音频等模态。已有研究评估了多模态大模型在视频理解任务中的能力：包括基础视觉要素识别，如对象、动作、场景识别与简单描述[68]；以及更高层次的语义推理任务，如社会互动关系[69]、人物心理状态[70]、价值观表达[71]等。研究结果表明，当前多模态大模型在基础视觉识别与简单情境理解方面已具备一

定能力，但在长时间视频理解、复杂时序与因果推理、深层语义理解、多模态信息深度整合等方面仍存在局限。因此，以图像静帧而非完整视频为主要计算分析对象，仍然是当前计算社会科学研究的主流。

除了媒体内容编码，还有一类针对特殊内容的识别任务，如虚假信息以及幽默、讽刺等复杂语用现象的识别，在计算社会科学中具有重要意义。由于其表达具有隐含性和情境性，仅依赖表层信息难以准确识别。相关研究通常需要以更具针对性的推理策略设计，引导多模态大模型进行更深入的语义推理与判断。

（2）虚假内容识别

虚假内容识别是指利用计算方法自动判别信息的真实性、误导性或内容伪造性质。生成式人工智能的发展推动了深度伪造（Deepfake）内容的泛滥，使其能够伪造真实图像噪声等信号特征，并基本保持跨模态语义的一致性[72]。这使得传统依赖单模态特征或浅层跨模态一致性的检测方法难以有效识别复杂的虚假内容。

在此背景下，基于多模态大模型的研究往往从真实性（Authenticity）、跨模态一致性（Consistency）、操控意图（Intent）等角度出发构建虚假内容识别策略[73]。已有实验研究表明，多模态大模型能通过跨模态表示学习与复杂推理机制联合分析视觉与文本语义，识别不同模态之间的潜在语义矛盾，从而提升多模态虚假内容识别性能。有研究基于自驱动问答推理机制（Self-driven QA reasoning）从多模态大模型中提取外部知识，通过多轮提问与回答挖掘视频中深层的语义矛盾，在三个基准数据集上相较当前最先进（SOTA）方法平均

更多免费行业报告 并购家 www.ipoipo.net

提升约 2.6%的性能，最高准确率达到 95.61%[74]。还有研究从情绪操控的视角出发，引入大模型生成视频、音频、文本等不同模态下的模拟用户评论，通过分析新闻内容与评论之间的情绪一致性，捕捉虚假信息在情绪诱导上的异常模式，在真实数据集上的检测准确率分别达到 79.44%和 85.97%[72]。

（3）复杂语用识别

幽默、讽刺等复杂语用现象在人类交流互动中广泛存在。这类表达具有动态性、情境性与文化依赖性[75]，其字面含义往往与真实意图不同，需要结合多模态线索与隐含语境进行复杂推理才能准确理解。在社交媒体中，这类表达以表情包模因（Memes）的形式广泛传播，通过图像与文本的组合传递隐含意义，在网络舆论生态中占据重要地位。多模态大模型的发展推动了对模因复杂语义的推理与分析。相关研究表明，多模态大模型能够在一定程度上识别图文模因语义，并被用于有害模因识别等计算社会科学任务。例如，有研究引入思维链（Chain of Thought，CoT）推理策略，引导模型分步骤分析图文语义关系，从而提升多模态大模型对有害表情包的识别能力，在相关数据集上达到 92%的准确率[76]；还有研究利用多模态大模型生成图文内容的解释，并将解释文本作为辅助特征参与分类，从而提升模型的决策可解释性与识别性能[77]。

复杂语用现象不止存在于图文模因中。在现实社会互动中，人们还会综合利用面部表情、肢体动作、语音语调等多种模态线索进行表意。针对记录在视频媒体中的复杂语用现象，研究发现语音信息在讽

刺识别中具有关键作用，因为语调、重音等往往能够揭示说话者的真实意图。实验结果表明，融合语音与文本或视觉信息的多模态模型在相关基准数据集上的 F1 值可达到 **70%–78%**[78]。

（4）数字行为分析

用户的数字行为痕迹，可以通过手机屏幕截图、传感器数据、应用日志等数据形态记录。多模态大模型能够从中识别用户行为模式与情绪状态，推动媒介使用与人机交互研究。

基于屏幕截图，多模态大模型可以综合分析界面元素、文本语义等，实现细粒度数字行为的识别与解释。例如，有研究每 5 秒自动采集一次参与者的手机屏幕截图，借助 GPT-4o 分析了 3.4 万张截图，并将用户的社交媒体使用行为分为浏览推荐内容、查看订阅内容、浏览评论、积极沟通、主动搜索等类别，与人工编码一致性（Cohen's Kappa）达 0.74。通过结合问卷调查结果，研究发现，浏览算法推荐和评论内容显著影响了用户"后悔感"的产生[79]。另有研究利用 GPT-4o 将用户的手机使用分为交流、媒体消费、功能任务、游戏、其他等五类，标注结果与人工编码的一致性（Krippendorff's Alpha）达 0.88[80]。多模态大模型的数字行为分析能力也在大规模数据中得到验证：有研究集成 LLaVA、CLIP 等模型，对 **112 万张手机截图**进行描述生成和主题聚类，其中，图像描述准确率约为 **83%**[81]。

此外，通过结合传感器数据，大模型也能够推断用户的情绪状态。例如，研究持续收集手机传感器数据，提取了包含地理位置、屏幕解锁状态、应用使用日志在内的 77 项行为特征，并利用 Gemini 1.5 Pro

更多免费行业报告 并购家 www.ipoipo.net

对这些行为模式进行语义解释，有效预测了个体的情绪状态，其中少样本预测的平均绝对误差可降至 0.75[82]。

（5）城市感知分析

城市感知，也即居民对城市环境的主观感受，包括安全感、活力感、环境质量等，是理解人类社会与地理环境互动关系的基础[83]。多模态大模型在基于街景图像的城市感知任务中展现出较强的零样本推理能力。例如，GPT-4o 在城市安全感知评分任务中获得了较高的人机一致性，在成都、大阪的案例研究中拟合优度可达 0.40-0.45[84]。GPT-4o 也能够识别影响居住隔离的多种复杂建成环境特征（包括物理屏障、交通基础设施、植被等 18 类空间要素），识别准确率达到 86%[85]。此外，GPT-4o mini 能够识别街道社会互动质量（拟合优度可达 0.43），并进一步揭示了街道绿化水平与天空可见度对街区活力提升的显著影响[86]。

在此基础上，研究者进一步开发了面向城市研究的专用模型，如 UrbanLLaVA[87]等，实现了对卫星图像、街景图像、轨迹数据以及结构化地理空间数据的统一建模，构建了从局部街区到城市全局的多尺度空间认知能力。由此，多模态大模型不仅推动了对城市感知的系统性认识，也为计算城市研究、城市规划治理[88]提供了技术支撑。

2. 以多模态大模型生成模拟社会情境

多模态大模型为研究者提供了高效度、低成本、可控制的模拟社会情境生成能力。首先，在实验刺激生成方面，研究者可以通过图像

生成模型，精确操纵实验刺激材料的特征。例如，有研究者提出 RAISE 方法：通过将已生成的基准图像再次输入模型，实现在仅改变目标操纵变量的同时，保持背景和视觉风格的一致性，从而解决实验刺激物精准控制问题。研究者还通过多项实验验证发现，AI 生成的视觉刺激物能够通过操纵检验，并复制既有实验效应[89]。在实证应用中，有研究利用 ChatGPT-4o 和 ChatGPT-5 生成警察的证件照，操纵了主体的种族、性别和体型，并通过实验评估这些视觉线索如何影响公众对警察能力、亲和力的感知[90]。

此外，通过结合大模型与语音克隆（Voice Cloning）、唇形同步（Lip-syncing）等多模态克隆技术，研究者可以构建高度个性化的数字化身（Avatars），以极强的沉浸感驱动认知干预实验。例如，有研究整合了 GPT-4o 及 ElevenLabs 的语音克隆技术，使参与者听到“理想自我”以自己的声音对负面情境进行积极重述，从而提升个体的心理韧性与目标感[91]。另有研究以 GPT-3.5-turbo 生成图像并利用自编脚本生成视频，构建了视频主体的社会类别（年龄、性别）或面部表情与言语内容不一致的冲突情境。通过分析参与者对不一致线索的认知与情感反馈，研究深入探讨了人类的认知整合加工机制[92]。

除了实验刺激物的生成，还有一些研究通过生成模型扩展社会科学研究中的数据模态。通过在不同模态之间建立映射关系，生成模型可以在缺乏真实数据的情况下构建新的研究材料，为理解复杂社会现象提供新的视角。例如，有研究尝试进行环境声音与街景图像间的模态转换，以模拟城市环境中的多感官体验。研究者利用扩散模型，根

更多免费行业报告 并购家 www.ipoipo.net

据城市声景（Soundscape）生成对应的街景图像[93]、甚至全景图像[94]，并通过机器评估和人类评估验证了生成结果与真实环境之间的一致性。这类方法表明，声音中包含大量环境线索与空间特征，例如交通密度、空间开阔程度等，因此可以通过跨模态生成模型重建视觉环境。还有研究通过跨模态匹配技术为街景图像匹配环境声音，以弥补视觉材料在城市感知研究中的偏差[95]。

多模态生成模型不仅能生成实验材料，还能扩展数据模态，模拟和重建复杂的社会情境。这类研究中的部分方法基于专用生成模型，并不属于严格意义上的多模态大模型，但它们展示了多模态生成技术在计算社会科学中的巨大潜力，并为未来多模态大模型在社会情境生成中的应用提供了想象。

3. 以多模态大模型模拟社会互动

基于多模态大模型构建的智能体能够同时处理文本与图像信息，被用于在实验环境或多智能体系统中模拟社会互动。在实验中，研究者可构建多模态“硅样本”（Silicon samples）。例如，有研究者基于 Claude 3.5 Sonnet 模型构建虚拟参与者，为其设定人口统计属性与社会背景，并将原始实验中的刺激材料（如广告文本、产品图片）输入模型，使其在设定情境下回答与人类参与者相同的问题。该研究在 14 篇论文、45 个实验中生成约 1.9 万个 AI 人格（AI personas）。结果表明，这些虚拟参与者能够生成与人类高度一致的反馈（如支付意愿、品牌态度），复现约 **76%的主要效应**[96]。然而，另有研究指出，多模态硅样本的回答偏模板化，在行为分布上与人类实验结果存在明显差

异，难以复现复杂统计效应（如交互效应）[97]。

此外，还有研究进行了具备多模态感知的多智能体仿真模拟。例如，LMAgent 框架以 GPT-4V 为核心，构建支持文本与图像处理、具有记忆和决策能力的智能体。在模拟电商环境的沙盒系统中，上万个智能体能够同时进行购物、社交、发布图文帖等行为，并在多模态的互动中影响彼此的决策。仿真结果表明，该系统涌现出了羊群效应等群体行为[98]。

总之，尽管多模态智能体在复杂行为模拟方面仍有不足，但其为模拟社会互动、研究复杂社会系统提供了计算基础设施。

4. 小结

多模态大模型拓展了计算社会科学的研究方法：既能够理解多模态社会数据，展开社交媒体内容、数字行为痕迹和地理社会感知的分析；也能够通过生成实验材料、扩展数据模态的方式构造社会情境；还能够通过多模态硅样本、多智能体仿真的方式模拟社会互动。多模态大模型虽然在模态整合与复杂推理、生成稳定性以及外部效度等方面仍然存在局限，但正在以快速发展的理解与生成能力，为分析复杂社会现象提供新的方法想象。

更多免费行业报告 并购家 www.ipoipo.net

参考文献

[1] Törnberg P. How to use LLMs for text analysis[A]. arXiv,2023[2026-03-09].

[2] Muasher-Kerwin C, Hughes M C, Foster M L, Al Azher I, Alhoori H. Exploring large language models for summarizing and interpreting an online brain tumor support forum[J]. DIGITAL HEALTH, 2025, 11: 20552076251337345[2026-03-09].

[3] Piper A, Wu S. Evaluating Large Language Models for Narrative Topic Labeling (NLP4DH 2025)[DS]https://borealisdata.ca/citation?persistentId=doi:10.5683/SP3/MHJRIO. 10.5683/SP3/MHJRIO.

[4] Kirilenko A, Stepchenkova S. Automated topic analysis with large language models[M]. BEREZINA K, NIXON L, TUOMI A. //Information and communication technologies in tourism 2024. Cham:Springer Nature Switzerland,2024: 29～34[2026-03-09].

[5] Gustavsen D, Surbaugh H M, Emmons M. Using generative AI for qualitative coding[J]. Library Trends, 2025, 73(3): 213～242[2026-03-09].

[6] Breazu P, Schirmer M, Hu S, Katsos N. Large language models and the challenge of analyzing discriminatory discourse: human-AI synergy in researching hate speech on social media[J]. Journal of Multicultural Discourses, 2024, 19(3): 157～175[2026-03-09].

[7] Rogers R, Zhang X. The Russia–ukraine war in chinese social media: LLM analysis yields a bias toward neutrality[J]. Social Media + Society, 2024, 10(2): 20563051241254379[2026-03-09].

[8] Rodríguez-Ibánez M, Casánez-Ventura A, Castejón-Mateos F, Cuenca-Jiménez P-M. A review on sentiment analysis from social media platforms[J]. Expert Systems with Applications, 2023, 223: 119862[2026-03-09].

[9] Lizzi A, Zatini G, Della Porta A. A nonessential corporate social responsibility reorientation: the woke-washing phenomenon in Victoria's secret's transformation[J]. Corporate Social Responsibility and Environmental Management, 2025, 32(5): 6362～6376[2026-03-09].

[10] Zhang K, Dong C, Guo Y, Zhou W, Yu G, Mi J. Lagged stance interactions and counter-spiral of silence: a data-driven analysis and agent-based modeling of technical public opinion events[J]. Systems, 2025, 13(6): 417～442[2026-03-09].

[11] Heseltine M, Clemm Von Hohenberg B. Large language models as a substitute for human experts in annotating political text[J]. Research & Politics, 2024, 11(1): 20531680241236239[2026-03-09].

[12] Vallejo Vera S, Driggers H. LLMs as annotators: the effect of party cues on labelling decisions by large language models[J]. Humanities and Social Sciences Communications, 2025, 12(1): 1530[2026-03-09].

[13] Burnham M. What is sentiment meant to mean to language models?[J]. Research & Politics, 2024, 11(4): 20531680241307941[2026-03-09].

[14] Huang F, Kwak H, An J. Chain of explanation: new prompting method to generate quality natural language explanation for implicit hate speech[C]//Companion Proceedings of the ACM Web Conference 2023. Austin TX USA:ACM,2023: 90～93[2026-03-09].

[15] Stoll A, Yu J, Andrich A, Domahidi E. Classification bias of LLMs in detecting incivility towards female and Male politicians in german social media discourse[J]. Communication Methods and Measures, 2025, 19(4): 350～368[2026-03-09].

[16] Park J K, Ellezhuthil R D, Wisniewski P, Singh V. Collaborative human-AI risk annotation: Co-annotating online incivility with CHAIRA[A]. arXiv,2024[2026-03-09].

[17] Carvalho P. The thin line between conspiracy theories and opinion: why humans and AI struggle to differentiate them[J].

[18] Daud R, Basir N. Evaluating large language model versus human performance in islamophobia dataset annotation[J]. International Journal of Advanced Computer Science and Applications, 2025, 16(5): 1～12.

[19] Zampieri M, Rosenthal S, Nakov P, Dmonte A, Ranasinghe T. OffensEval 2023: offensive language identification in the age of large language models[J]. Natural Language Engineering, 2023, 29(6): 1416～1435[2026-03-09].

[20] Hasan M A, Alam F, Hossain M F, Naseem U, Ahmed S I. LLM-Based Multi-Task Bangla Hate Speech Detection: Type, Severity, and Target[A]. arXiv,2025[2026-03-09].

[21] Walker A, Thorne A, Das S, Love J, Cooper H L F, Livingston M, Sarker A. CARE-SD: classifier-based analysis for recognizing provider stigmatizing and doubt marker labels in electronic health records: model development and validation[J]. Journal of the American Medical Informatics Association, 2025, 32(2): 365～374[2026-03-09].

[22] Singh D D, Bhattacharjee R, Chakraborty A. Rethinking hate speech detection on social media: can LLMs replace traditional models?[A]. arXiv,2025[2026-03-09].

[23] Xu R, Sun Y, Ren M, Guo S, Pan R, Lin H, Sun L, Han X. AI for social science and social science of AI: a survey[A]. arXiv,2024[2026-03-09].

[24] Ziems C, Held W, Shaikh O, Chen J, Zhang Z, Yang D. Can large language models transform computational social science?[J]. Computational Linguistics, 2024, 50(1): 237～291[2026-03-09].

[25] Yi J, Xu Z, Huang T, Yu P. Challenges and innovations in LLM-powered fake news detection: a synthesis of approaches and future directions[A]. Computer Science and Mathematics,2025[2026-03-09].

[26] Rostami M, Hawamdeh S. Debunk lists as external knowledge structures for health misinformation detection with generative AI[J]. Systems, 2025, 13(10): 882～899[2026-03-09].

[27] Yu X, Shang D. Research on online disinformation identification based on information characteristics and large language models[J]. The Electronic Library, 2025, 43(5): 757～776[2026-03-09].

更多免费行业报告 并购家 www.ipoipo.net

[28] Chen J, Liu L, Zhou F. Do not wait: preemptive rumor detection with cooperative LLMs and accessible social context[J]. Information Processing & Management, 2025, 62(3): 103995[2026-03-09].

[29] Das R K, Dodge J. Fake news detection after LLM laundering: measurement and explanation[A]. arXiv,2025[2026-03-09].

[30] Dehghani S, Zahedivafa M, Baghshahi Z, Zare D, Yari S, Samei Z, Aliahmadi M, Abbasi M, Mirzamojtahedi S, Ebrahimi S, Alizadeh M. Leveraging large language models for fact-checking farsi news headlines[M]. PREUSS M, LESZKIEWICZ A, BOUCHER J-C, FRIDMAN O, STAMPE L. //Disinformation in open online media. Cham:Springer Nature Switzerland,2024: 16～31[2026-03-09].

[31] Ng L H X, Cruickshank I J, Lee R. Examining the influence of political bias on large language model performance in stance classification[J].

[32] Gilardi F, Alizadeh M, Kubli M. ChatGPT outperforms crowd workers for text-annotation tasks[J]. Proceedings of the National Academy of Sciences, 2023, 120(30): e2305016120[2026-03-09].

[33] Farr D T, Hui Xian Ng L, Cruickshank I J, Manzonelli N, Clark N, Starbird K, Bastian N D, West J. Ensemble-based uncertainty quantification for reliable large language model classification in social data applications[J]. IEEE Access, 2025, 13: 116419～116429[2026-03-09].

[34] Kim S, Kim K, Wonjeong Jo C. Accuracy of a large language model in distinguishing anti- and pro-vaccination messages on social media: the case of human papillomavirus vaccination[J]. Preventive Medicine Reports, 2024, 42: 102723[2026-03-09].

[35] Mets M, Karjus A, Ibrus I, Schich M. Automated stance detection in complex topics and small languages: the challenging case of immigration in polarizing news media[J]. PLOS ONE, 2024, 19(4): e0302380[2026-03-09].

[36] Törnberg P. Large language models outperform expert coders and supervised classifiers at annotating political social media messages[J]. Social Science Computer Review, 2025, 43(6): 1181～1195[2026-03-09].

[37] Gatt A, Krahmer E. Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation[J]. Journal of Artificial Intelligence Research, 2018, 61: 65～170[2026-03-09].

[38] Fresz B, Dubovitskaya E, Brajovic D, Huber M F, Horz C. How should AI decisions be explained? Requirements for explanations from the perspective of European law[J]. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2024, 7: 438～450[2026-03-09].

[39] Stammer W, Friedrich F, Steinmann D, Brack M, Shindo H, Kersting K. Learning by self-explaining[A]. arXiv,2024[2026-03-09].

[40] Farjam M, Meyer H, Lohkamp M. A practical guide and case study on how to instruct LLMs for automated coding during content analysis[J]. Social Science Computer Review, 2025: 08944393251349541[2026-03-09].

[41] Lan X, Gao C, Jin D, Li Y. Stance detection with collaborative role-infused LLM-

based agents[J]. Proceedings of the International AAAI Conference on Web and Social Media, 2024, 18: 891～903[2026-03-09].

[42] Cirillo S, Desiato D, Polese G, Solimando G, Sugumaran V, Sundaramurthy S. Exploring the ability of emerging large language models to detect cyberbullying in social posts through new prompt-based classification approaches[J]. Information Processing & Management, 2025, 62(3): 104043[2026-03-09].

[43] Kunz J, Kuhlmann M. Properties and challenges of LLM-generated explanations[C]//Proceedings of the Third Workshop on Bridging Human--Computer Interaction and Natural Language Processing. Mexico City, Mexico:Association for Computational Linguistics,2024: 13～27[2026-03-09].

[44] Webb M, Singh H, Inman R, Baniya S, Katz A. Advancing qualitative analysis in professional disaster and risk communication: a comparative study of an OpenAI ChatGPT 3.5 model-enabled method for processing complex public discourse[J]. International Journal of Disaster Risk Reduction, 2025, 130: 105835[2026-03-09].

[45] Bengesser C, Sørensen J K. Different diversities: policies and practices at three european public service VoD services[J]. Convergence: The International Journal of Research into New Media Technologies, 2024, 30(4): 1365～1387[2026-03-09].

[46] Zhang Y, Jin H, Meng D, Wang J, Tan J. A comprehensive survey on process-oriented automatic text summarization with exploration of LLM-based methods[A]. arXiv,2025[2026-03-09].

[47] Fu X, Brinkley C, Sanchez T W, Li C. Text mining public feedback on urban densification plan change in Hamilton, new zealand[J]. Environment and Planning B: Urban Analytics and City Science, 2024: 23998083241272097[2026-03-09].

[48] Chiu I-C, Hung M-W. Finance-specific large language models: advancing sentiment analysis and return prediction with LLaMA 2[J]. Pacific-Basin Finance Journal, 2025, 90: 102632[2026-03-09].

[49] Lambrecht K. “What People Are Saying” About Extreme Weather: A Comparison of Public Commentary and AI-Generated Summaries[J]. IEEE Transactions on Professional Communication, 2025, 68(4): 433～450[2026-03-09].

[50] Balt E, Salmi S, Bhulai S, Vrinzen S, Eikelenboom M, Gilissen R, Creemers D, Popma A, Mérelle S. Deductively coding psychosocial autopsy interview data using a few-shot learning large language model[J]. Frontiers in Public Health, 2025, 13: 1512537[2026-03-09].

[51] Liu J, Min L, Huang X. An overview of event extraction and its applications[A]. arXiv,2021[2026-03-09].

[52] Li B, Han X, Liu J, Ding Y, Jing L, Zhang Z, Li J, Du X, Li F, Zhang M, Zhang M, Sun A, Yu P S, Fei H. Event extraction in large language model[A]. arXiv,2025[2026-03-09].

[53] Breton J, Billami M M, Chevalier M, Nguyen H T, Satoh K, Trojahn C, Zin M M. Leveraging LLMs for legal terms extraction with limited annotated data[J]. Artificial Intelligence and Law, 2025: 1～27[2026-03-09].

[54] Ji X, Xue J, Chen S, Zhu T. Automating social science: LLMs vs. Human experts

更多免费行业报告 并购家 www.ipoipo.net

in variable relationship identification[J]. Collabra: Psychology, 2025, 11(1): 144442[2026-03-09].

[55] Wang J, Liu G, Cheng Y, Xu X, Li Z. Leveraging internet-sourced text data for financial analytics in supply chain finance: a large language model-enhanced text mining workflow[J]. IEEE Transactions on Engineering Management, 2025, 72: 1924～1938[2026-03-09].

[56] Wang Z, Li M-H, Baxter P, Zhorayev O, Wei J, Kovacs V, Zhao Q, Yang C, Koizumi N. Implementing large language model and retrieval augmented generation to extract geographic locations of illicit transnational kidney trade[J]. International Journal of Health Geographics, 2025, 24(1): 10～34[2026-03-09].

[57] Du G, Chen A. Coal mine accident risk analysis with large language models and bayesian networks[J]. Sustainability, 2025, 17(5): 1896[2026-03-09].

[58] Nam K, Lee S, Jung H-Y. Detection and description of neologisms in korean lexicography: methodological issues in corpus balance, word unit bias and LLM assistance[J]. Lexikos, 2025, 35: 1～25[2026-03-09].

[59] Yu D, Li L, Su H, Fuoli M. Assessing the potential of LLM-assisted annotation for corpus-based pragmatics and discourse analysis: the case of apology[J]. International Journal of Corpus Linguistics, 2024, 29(4): 534～561[2026-03-09].

[60] Lee K, Paci S, Park J, You H Y, Zheng S. Applications of GPT in political science research[J].

[61] Sufi F, Alsulami M. Unmasking media bias, economic resilience, and the hidden patterns of global catastrophes[J]. Sustainability, 2025, 17(9): 3951[2026-03-09].

[62] Yin S, Fu C, Zhao S, Li K, Sun X, Xu T, Chen E. A survey on multimodal large language models[J]. National Science Review, 2024, 11(12): nwae403[2025-03-17].

[63] 周葆华, 吴雨晴. 超越单一模态：多模态计算传播研究的进展与前瞻[J]. 传媒观察, 2024(1): 16～27[2026-03-08].

[64] Guo P, Sun H, Xing S, Li J. A study on the visual rhetorical differences in national image representation of China and the United States by generative artificial intelligence: an empirical analysis based on large multimodal models[J]. Journal of Information Technology & Politics, 2025.

[65] Qian X. Evaluating destination attractiveness through a social semiotic perspective: a large language model analysis of user-generated images[J]. Current Issues in Tourism, 2025: 1～17[2025-09-10].

[66] Jang J, Kwon S, Park S. Beauty vs. Vibe: deconstructing visual appeal in online dating with large multimodal models[J]. Computers in Human Behavior, 2025, 173.

[67] Steinert-Threlkeld Z, Xie C. Zero-shot protest event detection with images and large multimodal models[A]. APSA Preprints,2025[2026-03-07].

[68] Tang Y Y, Bi J, Xu S, Song L, Liang S, Wang T, Zhang D, An J, Lin J, Zhu R, Vosoughi A, Huang C, Zhang Z, Liu P, Feng M, Zheng F, Zhang J, Luo P, Luo J, Xu C. Video understanding with large language models: a survey[A].

arXiv,2025[2026-03-06].

[69] Kong F, Zu W, Chen X, Yang Y, Zhu S-C, Feng X. SIV-bench: a video benchmark for social interaction understanding and reasoning[A]. arXiv,2025[2026-03-07].

[70] Villa-Cueva E, Ahmed S M M, Chevi R, Cruz J C B, Elzeky K, Cristobal F, Aji A F, Wang S, Mihalcea R, Solorio T. MoMentS: a comprehensive multimodal benchmark for theory of mind[C]. CHRISTODOULOPOULOS C, CHAKRABORTY T, ROSE C, PENG V. //Findings of the Association for Computational Linguistics: EMNLP 2025. Suzhou, China:Association for Computational Linguistics,2025: 22591～22611[2026-03-07].

[71] Wang Y, Qiu H, Chang H, Xiao J, Tao Z, Jin L. Can multimodal large language models understand human values in videos?[M]//Proceedings of the 7th ACM international conference on multimedia in Asia. New York, NY, USA:Association for Computing Machinery,2025: 1～7[2026-03-06].

[72] Zhang L, Shi Y, Cui M. A bi-level multi-modal fake generative news detection approach: from the perspective of emotional manipulation purpose[J]. Humanities & Social Sciences Communications, 2025, 12(1): 1～19.

[73] Ai W, Tan Y, Shou Y, Meng T, Chen H, He Z, Li K. The paradigm shift: a comprehensive survey on large vision language models for multimodal fake news detection[J]. Computer Science Review, 2026, 60: 100893[2026-03-06].

[74] Liu F, Li Y, Lang J, Hong R, Zhou F. Enhancing fake news video detection with self-driven question-answer from LMMs[J]. Information Processing & Management, 2026, 63(2): 1～17.

[75] Polli C, Sindoni M G. Multimodal computation or interpretation? Automatic vs. critical understanding of text-image relations in racist memes in English[J]. Discourse, Context & Media, 2024, 57: 100755[2025-07-31].

[76] Zhuang Y, Guo K, Wang J, YihengJing †, Xu X, Yi W, Yang M, Zhao B, Hu H. I know what you MEME! Understanding and detecting harmful memes with multimodal large language models[J]. Proceedings 2025 Network and Distributed System Security Symposium, 2025.

[77] Hee M S, Lee R K-W. Demystifying hateful content: leveraging large multimodal models for hateful meme detection with explainable decisions[J]. Proceedings of the International AAAI Conference on Web and Social Media, 2025, 19: 774～785[2026-03-06].

[78] Li Z, Gao X, Zhang Y, Nayak S, Coler M. Evaluating multimodal large language models on spoken sarcasm understanding[A]. arXiv,2025[2026-03-06].

[79] Guo L, Fu Y, Lin X, Xu X, Chang Y-J, Hiniker A. What social media use do people regret? An analysis of 34K smartphone screenshots with multimodal LLM[C]//Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. New York, NY, USA:Association for Computing Machinery,2025: 1～23[2026-03-07].

[80] Chang Y-J, Cho M-J, Chi Y, Chen Y-C, Lee Y-J, Kao K-C, Ram N, Robinson T N, Reeves B. When killing-time periods don’t boost notification engagement: insights from a multimodal large language model analysis of smartphone screen activity[A].

更多免费行业报告 并购家 www.ipoipo.net

Rochester, NY:Social Science Research Network,2025[2026-03-07].

[81] Cerit M, Zelikman E, Cho M-J, Robinson T N, Reeves B, Ram N, Haber N. Media content atlas: a pipeline to explore and investigate multidimensional media space using multimodal LLMs[C]//Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. New York, NY, USA:Association for Computing Machinery,2025: 1～13[2026-03-07].

[82] Zhang T, Teng S, Jia H, D'Alfonso S. Leveraging LLMs to predict affective states via smartphone sensor features[C]//Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing. New York, NY, USA:Association for Computing Machinery,2024: 709～716[2026-03-07].

[83] StreetSenser: a novel approach to sensing street view via a fine-tuned multimodal large language model[J]. International Journal of Geographical Information Science[2026-03-07].

[84] Zhang J, Li Y, Fukuda T, Wang B. Urban safety perception assessments via integrating multimodal large language models with street view images[J]. Cities, 2025, 165: 106122[2026-03-07].

[85] Law T, Roberto E. Generative multimodal models for social science: an application with satellite and streetscape imagery[J]. Sociological Methods & Research, 2025, 54(3): 889～932.

[86] Elrod K, Flanigan K A, Bergés M. Street view sociability: interpretable analysis of urban social behavior across 15 cities[C]//Proceedings of the 12th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation. New York, NY, USA:Association for Computing Machinery,2025: 282～286[2026-03-07].

[87] Feng J, Wang S, Liu T, Xi Y, Li Y. UrbanLLaVA: a multi-modal large language model for urban intelligence with spatial reasoning and understanding[A]. arXiv,2025[2026-03-07].

[88] Pan F, Huang X, Bi Y, Gao Y, Ye Y, Wang H. From tools to partners: How large language models are transforming urban planning[J]. AI Open, 2025, 6: 276～298[2026-03-07].

[89] Zamudio C, Grigsby J L, Michelsen M. RAISE: a new method to develop experimental stimuli for advertising research with image generative artificial intelligence[J]. Journal of Advertising, 2025: 1～16[2025-09-10].

[90] Robinson C G, Adams I T, Logan M W, Blair J P. AI-generated human stimuli for experimental social science[J]. Journal of Experimental Criminology, 2026[2026-03-07].

[91] Fang C M, Chua P, Chan S W T, Leong J, Bao A, Maes P. Leveraging AI-generated emotional self-voice to nudge people towards their ideal selves[C]//Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. New York, NY, USA:Association for Computing Machinery,2025: 1～20[2026-03-07].

[92] Compensis P. Assessing the integration of social information using AI-generated videos: a brief research report[J]. Frontiers in Psychology, 2026, 16[2026-03-07].

[93] Zhuang Y, Kang Y, Fei T, Bian M, Du Y. From hearing to seeing: Linking auditory and visual place perceptions with soundscape-to-image generative artificial intelligence[J]. Computers, Environment and Urban Systems, 2024, 110: 102122[2026-03-08].

[94] Zhuang Y, Gui Y, Fei T. Soundscape-to-panorama: spatialize auditory perception by linking acoustic environment to panorama[J]. International Journal of Digital Earth, 2025, 18(1): 2545584[2026-03-07].

[95] Le Q H, Moon H, Ho J, Ahn Y. From seeing to hearing: A feasibility study on utilizing regenerated sounds from street view images to assess place perceptions[J]. Building and Environment, 2025, 269: 112468[2026-03-08].

[96] Yeykelis L, Pichai K, Cummings J J, Reeves B. Using large language models to create AI personas for replication, generalization and prediction of media effects: an empirical test of 133 published experimental research findings[A]. arXiv,2025[2026-02-22].

[97] Yoo K, Haenlein M, Hewett K. A whole new world, a new fantastic point of view: charting unexplored territories in consumer research with generative artificial intelligence[J]. Journal of the Academy of Marketing Science, 2025, 53(3): 723～759[2025-09-10].

[98] Liu Y, Liu W, Gu X, Rui Y, He X, Zhang Y. LMAgent: a large-scale multimodal agents society for multi-user simulation[A]. arXiv,2024[2026-03-08].

更多免费行业报告 并购家 www.ipoipo.net

第八章 从技术嵌入到伦理价值的反思：数智人文的新视角与新趋势[①]

数智人文不只是把传统材料“数字化”，更在技术深度嵌入之下重塑研究对象、方法逻辑与知识生产方式。本章以“技术嵌入—范式重构—伦理反思”为主线，梳理数字人文向数智人文跃迁的机制，展示文学、历史与哲学等典型领域的数智化实践如何生成可迁移的方法论经验。

生成式 AI 的介入不仅带来效率提升，也带来叙事结构、文化表达与学术伦理的系统性挑战；其背后牵涉到数字存在的本体论追问与计算理性与人文理解之间的认识论张力。因而，数智人文的健康发展必须以批判性与反思性为底座，在人机协同的新型学术生态中重建数据治理、算法透明与学术规范，推动技术进步与人文价值同向而行。本章拟从五个维度展开论述：首先审视工具、对象与方法的系统再造，继而分析文学、历史学与哲学等传统人文领域的数智化转型实践，进一步探讨生成式 AI 对叙事结构、文化表达与创作伦理的冲击与挑战，在本体论与认识论层面进行深层理论审视，最终提出重塑人文学科批判性与反思性功能的路径建议。通过多维度的系统考察，本研究试图为理解数智人文的新视角与新趋势提供一个兼具理论深度与实践关怀的分析框架。

[①] 本章由复旦大学科技考古研究院文少卿副教授牵头，复旦大学国家发展与智能治理综合实验室张以颖合作撰写。

文少卿副教授现任国家发展与智能治理综合实验室副主任，“中欧人才”（国家自然基金委，2021 年）、“兰台青年学者”（中国历史研究院，2023 年），专注分子考古、量化考古与法医考古，利用古基因组学研究中华民族谱系，探讨文明起源与民族交融的历史进程。

本章提要：

- 数字人文正经历从“文献数字化—计算分析”走向“智能生成—范式重构”的跃迁，数智人文的核心不在工具升级，而在研究对象、方法逻辑与知识生产结构的系统再造。
- 在工具层面，预训练语言模型与知识图谱等技术推动人文研究由浅层统计转向深层语义计算，使计算机从“计数器”走向具备语义推理能力的分析代理。
- 在对象层面，文本、档案与文化遗产被转化为可运算的数据结构与语义网络，研究问题由线性叙事扩展到跨时空的结构关联与动态生成。
- 在方法层面，“可计算阐释”强调计算发现与意义阐释的互动：远距离阅读提供宏观导航，精读提供深度解释，二者构成互补而非替代关系。
- 生成式 AI 直接介入叙事与创作环节，带来作者性、文化同质化、算法偏见、著作权与真实性（幻觉）等挑战，促使人文学科在本体论与认识论层面重估技术的边界。
- 数智人文的战略价值在于重塑人文学科的批判性与反思性：以数据批判、算法批判与应用批判为抓手，建设人机协同的学术生态与交叉人才体系，形成兼具技术能力与人文关怀的制度化支撑。

更多免费行业报告 并购家 www.ipoipo.net

8.1 从“数字人文”到“数智人文”的范式变革

人工智能正以不可逆转之势深层嵌入人文社会科学的知识生产体系。自 20 世纪 90 年代数字人文（Digital Humanities）概念兴起以来，人文学科与信息技术的交叉融合历经了从文献数字化到计算分析、再到智能生成的三个发展阶段。早期的数字人文实践主要聚焦于文本的数字化存储与检索，典型如谷登堡计划（Project Gutenberg）与中国知网等数字化平台的建设。进入 21 世纪的第二个十年，以自然语言处理（NLP）和机器学习为代表的技术工具开始被大规模引入人文研究，莫雷蒂（Franco Moretti）提出的"远距离阅读"（Distant Reading）理念成为这一时期的标志性方法论贡献。

然而，随着 2017 年 Transformer 架构的提出及其后续 GPT、BERT 等大型预训练语言模型的迅猛发展，数字人文正经历一场前所未有的范式跃迁——从"数字人文"（Digital Humanities）向"数智人文"（Digital Intelligence Humanities）的转型。这种转变绝非仅仅是研究工具的迭代升级，更是一场深刻的本体论与认识论革命。技术工具、研究对象与方法逻辑在算法的介入下经历着系统性的再造，传统人文学科的边界与内涵正在被重新定义。周维东等学者指出，数智人文代表着人文知识生产方式的根本性变革，其核心特征在于人工智能从辅助性工具转变为知识生产的参与性主体。

在这样的时代背景下，我们需要以更为宏阔的历史视野与审慎的批判精神，审视这场由硅基智能引发的知识转型。值得注意的是，这

场转型并非西方学术界的独角戏。中国学术界在数智人文领域同样展现出强劲的发展势头，从中国知网的知识图谱建设到故宫博物院的数字化工程，从敦煌研究院的 AI 辅助壁画修复到中国历代人物传记资料库（CBDB）的持续扩展，中国的数智人文实践不仅汲取了国际学术前沿的理论养分，更在本土化应用中形成了独特的方法论贡献。

数智人文的演进表明，人文学科并非在技术冲击下被动调整的对象，而是能够以批判性与反思性参与塑造技术方向的主体。未来工作的关键，是在方法创新与价值守护之间形成制度化的平衡：以可计算阐释拓展证据边界，以数据/算法/应用批判守住学术与文化底线，以平台化基础设施与交叉人才体系提升协作效率，从而推动数智人文走向更高质量、更可持续的发展。

1. 工具、对象与方法的系统再造

（1）从远距离阅读到深层语义计算：研究工具的跃迁

早期的数字人文受限于算力与算法，多停留于词频统计、共现分析等浅层文本挖掘。如莫雷蒂（Moretti）所推动的计量文体学研究，主要依赖于简单的统计模型与可视化技术对文本进行宏观分析。乔克斯（Matthew Jockers）进一步发展了宏观分析（Macroanalysis）的方法，运用主题模型与情感分析技术对大规模文学语料进行系统性挖掘。

如今，基于预训练语言模型的自然语言处理技术，赋予了计算机理解复杂上下文、捕捉隐喻与多义词的能力。以 BERT 为代表的双向

更多免费行业报告 并购家 www.ipoipo.net

编码器模型能够在语境中动态生成词向量，从而有效解决一词多义和语义消歧等传统难题。研究工具不再是简单的计数器，而是能够进行模式识别、情感计算与语义推理的智能代理。米歇尔（Jean-Baptiste Michel）等人在《科学》杂志上发表的"文化组学"（Culturomics）研究，利用谷歌图书语料库对超过 500 万本图书进行计量分析，开创了大规模文化趋势计算研究的先河。

（2）从静态文本到动态语义网络：研究对象的转化

在传统视阈中，人文学科的考察对象是静态的物理文本、历史档案或物质文化遗产。文本被视为意义的容器，研究者通过阅读与阐释从中提取意义。而在数智人文范式下，这些实体被转化为高维向量空间中的数据点。文本的意义不再仅仅存在于字里行间，而是被解构为算法可以运算的特征矩阵。

这种转化具有深刻的方法论意涵。伯里（Berry）在其数字人文理论中指出，数字化转向意味着人文研究对象经历了一次"计算性转化"（Computational Turn），文本从线性的叙述序列转变为非线性的数据结构。词嵌入（Word Embedding）技术使得每一个词汇都被映射到一个多维向量空间中，词汇之间的语义关系可以通过向量距离进行精确量化。研究对象由此从封闭的、固定的文本，转化为开放的、动态生成的语义网络。这种转化不仅改变了我们观察文本的方式，更根本性地改变了"文本是什么"这一人文学科的基础问题。

（3）从直觉阐释到可计算阐释：方法逻辑的转型

传统的阐释学（Hermeneutics）依赖于学者的主体直觉与生命体

验，强调对文本的同情之理解。施莱尔马赫（Schleiermacher）和伽达默尔（Gadamer）所奠定的阐释学传统，将理解视为主体间的对话与视域融合的过程。数智人文则引入了计算逻辑，将归纳与演绎交织于数据驱动的模型训练之中。

学者提出"可计算阐释"（Computable Interpretation）的概念，强调在保留人文阐释主体性的前提下，借助计算方法发现文本中的潜在模式与深层结构。通过主题模型（如 LDA）、词向量聚类与网络分析等算法，研究者能够发现海量文献中人类肉眼难以察觉的潜在模式，为人文阐释提供了全新的实证基础。

这种方法论转型的深层意义在于，它打破了人文学科长期以来"理解"与"解释"之间的二元对立。在传统的学术范式中，自然科学追求因果解释（Erklären），人文科学致力于意义理解（Verstehen），两者被视为截然不同的知识路径。而数智人文的方法论实践表明，计算驱动的模式发现可以与人文学者的意义阐释形成有机互动，两者并非此消彼长的零和关系，而是能够在辩证互动中共同推进知识的深化。正如布洛克（David Blei）所言，主题模型并非要取代人类的阅读行为，而是为人类提供一种宏观导航文献集合的计算透镜。从直觉阐释到可计算阐释的转变，构成了当代人文研究方法论转型的核心要义。

2. 传统领域的数智化转型与实践

（1）文本挖掘与文学空间的拓扑学重构

文学研究长期受困于样本量的局限，精读模式难以统摄宏大的文

更多免费行业报告 并购家 www.ipoipo.net

学史脉络。莫雷蒂的"远距离阅读"理念打破了这一困局，他主张通过计量方法对文学场域进行宏观透视。而在深度学习时代，文本挖掘的能力得到了质的飞跃。通过深度文本挖掘与主题模型算法，研究者能够对成千上万部文学作品进行宏观扫描，揭示文学类型的演变规律、作家群体的风格流变以及特定文学意象的时空分布。

在古典文学研究领域，基于神经网络的语言模型可以精准刻画特定词汇在不同历史时期的语义漂移。汉密尔顿（William Hamilton）等人利用词嵌入模型追踪英语词汇在过去两个世纪的语义演变，为历史语义学提供了强有力的计算方法。在中国古典文学领域，学者们利用类似方法对《全唐诗》《全宋词》等大型语料进行情感分析与主题建模，揭示了唐宋文学从"雄浑壮阔"到"婉约精致"的审美转向背后的复杂社会文化因素。

通过对大规模语料的计算，研究者可以绘制出文学话语的拓扑结构图，发现那些被经典文学史叙事所遮蔽的边缘文本与断裂线索。例如，在明清小说研究中，利用文本聚类算法对数百部白话小说进行风格分析，可以揭示不同地域、不同阶层文学传统之间的隐性关联，从而突破传统文学史以经典作家为中心的叙述框架。文本挖掘不仅验证了既有的文学史假说，更重要的是，它生成了全新的文学问题，将文学研究引向了更为广阔的数据驱动的实证空间。值得强调的是，数智方法并非要取代传统的精读实践，而是与之形成互补关系：远距离阅读为精读提供宏观导航，精读则为远距离分析的结果提供深度阐释与人文意义赋予。

（2）知识图谱与历史时空的立体化复原

历史学是一门建立在史料考证基础上的学科。面对浩如烟海且碎片化的历史档案，传统的人力检索已显得力不从心。知识图谱技术的引入，为史料的组织与关联提供了革命性的方案。

在数字史学实践中，知识图谱的应用已取得显著成果。包弼德（Peter Bol）主持的中国历代人物传记资料库（CBDB）便是一个经典案例，该项目将中国历史上数十万人物的传记信息编码为结构化数据，支持社会网络分析与时空可视化。通过实体抽取、关系提取与事件图谱构建，人工智能可以将散落于方志、笔记、官方档案与碑刻中的零散信息，编织成一张具备时空属性的语义网络。历史人物的社会交往、家族迁徙的地理轨迹、制度演变的因果链条，皆可通过图谱进行可视化呈现与网络动力学分析。

此外，地理信息系统（GIS）与知识图谱的结合为历史空间研究开辟了新路径。中国历史地理信息系统（CHGIS）项目通过将历史地名、行政区划变迁与地理坐标系统关联，实现了对中国两千余年行政区划变迁的数字化重建。这种方法不仅实现了历史数据的结构化表达，更打破了线性史观的叙事局限，使研究者能够在多维时空中进行史料的比对与互证，极大地拓展了历史研究的广度与精确度。

（3）跨模态生成与哲学概念的历史计算

哲学研究关注思想的流变与概念的生成。在数智人文的视野下，跨模态生成技术为概念史（Begriffsgeschichte）与思想史的研究提供了新颖的路径。科塞雷克（Reinhart Koselleck）所开创的概念史

更多免费行业报告 并购家 www.ipoipo.net

方法强调追踪核心概念在历史语境中的语义演变。如今，语言模型不仅能够处理纯文本，还能在文本、图像、甚至音频之间进行语义映射，为概念史研究提供了前所未有的跨模态分析能力。

对于复杂的哲学概念，研究者可以利用跨模态技术分析其在不同媒介中的表达形式与接受史。同时，通过计算不同哲学流派文本的语义距离，可以量化分析思想体系之间的继承与断裂。例如，学者们利用词嵌入技术计算康德、黑格尔与马克思文本之间的语义相似度，从而在数据层面揭示了德国古典哲学到马克思主义思想演进的内在逻辑线索。人工智能在这里扮演了思想史考古挖掘机的角色，帮助学者在浩瀚的文献中追踪那些幽微的观念演化轨迹，为哲学文本的重新阐释提供了基于数据科学的坚实支撑。

8.2 AGI 冲击下的叙事、文化与真实性反思

生成式 AI 从分析走向生成，直接介入叙事、表达与创作环节，使人文学科不得不重新讨论作者性、真实性、文化多样性与偏见等核心议题。它们决定了数智人文能否从“能用”走向“可信”与“可持续”。

1. 生成式 AI 对人文学科的冲击与挑战

（1）叙事结构的解构与重组

生成式 AI 的运作机制基于概率分布与上下文预测，这从根本上颠覆了传统的人类叙事逻辑。人类的叙事往往具备明确的意图、情感驱动与线性的因果关系；而 AI 的生成物则是对海量语料进行模式匹配与重组的结果。本雅明（Walter Benjamin）在其经典论文中指出，机械复制时代艺术品的"灵韵"（Aura）已然消散。在生成式 AI 时代，这一命题获得了更为极端的呈现——AI 不仅复制，更在"创造"，而这种创造是否具备本雅明意义上的灵韵，成为亟需讨论的理论命题。

当 AI 参与文学创作或历史叙事时，传统的作者中心论被消解。福柯（Michel Foucault）关于"作者是什么"的追问在数智时代获得了全新的理论张力。文本成为人类提示词与机器算法共同作用的产物，叙事结构呈现出碎片化、非线性与多重可能性的特征。这种人机共创的模式虽然极大地丰富了叙事的形式，但也带来了叙事深度的流失与意义的空洞化风险。机器生成的文本看似逻辑严密，却往往缺乏基于人类真实生命体验的灵魂注入。在小说创作领域，AI 可以生成情节连贯的故事，但这些故事往往停留在表层的模式重组，难以企及卡夫卡

更多免费行业报告 并购家 www.ipoipo.net

式的存在焦虑或普鲁斯特式的记忆深度。这种叙事能力的"天花板"反而成为检验人类创造力独特性的一面镜子。

（2）文化表达的同质化与算法偏见

人工智能的训练依赖于既有的庞大数据集，这些数据集不可避免地携带了现实世界中的文化偏见与权力结构。本德斯（Emily Bender）等人在其具有广泛影响力的论文中将大型语言模型比喻为"随机鹦鹉"（Stochastic Parrots），警告说这些模型可能在看似流畅的表达中复制和放大训练数据中存在的系统性偏见。生成式模型在输出内容时，往往会强化主流文化的表达，而边缘文化、少数族裔的叙事则可能被算法过滤或边缘化。

这种由于数据分布不均导致的算法偏见，正在潜移默化地塑造着数智时代的文化表达格局。诺贝尔（Safiya Umoja Noble）在其研究中揭示了搜索引擎算法如何系统性地再生产种族与性别偏见。如果不加干预，生成式 AI 极有可能成为一种新型的文化霸权工具，导致全球文化表达的同质化。人文学者必须敏锐地察觉到这一危机，致力于构建多元、包容且具备文化敏感性的训练语料库，以此对抗算法带来的文化收敛。在中国语境下，如何在人工智能训练数据中充分体现中华优秀传统文化的丰富性与多样性，避免西方中心主义的数据偏见，是一个尤为紧迫的议题。

（3）人文创作伦理的困境

生成式 AI 的涌现引发了旷日持久的伦理争议。首当其冲的便是著作权归属与原创性界定。当一段饱含深情的诗歌或一篇严谨的史学综

述是由机器在几秒钟内生成时，我们该如何定义作者的身份？世界各国的知识产权法律体系目前尚未对 AI 生成内容的著作权归属作出明确规定，这一法律空白已成为学术界和法律界共同面临的重大挑战。

此外，AI 生成内容的真实性与可靠性构成了另一重严峻挑战。大型语言模型的"幻觉"（Hallucination）现象——即模型会一本正经地捏造历史事件或虚构文献出处——已被广泛记录和研究。在对真实性要求极高的史学与哲学研究中，这种虚假信息的泛滥将对学术诚信体系造成致命打击。学术界已开始探索建立 AI 辅助写作的规范框架，包括要求标注 AI 参与程度、建立 AI 生成内容的检测机制以及制定 AI 时代的学术伦理准则等。这就要求我们在享受技术红利的同时，必须建立起严格的学术伦理规范与 AI 内容审查机制。

2. 本体论与认识论的深层审视

（1）数字存在的本体论追问

在本体论层面，我们需要追问：经过数字转化与算法重构的文本，是否还是原来的文本？当一幅古典画作被转化为像素矩阵并输入神经网络，其所蕴含的神韵与光晕是否已经消散？海德格尔（Martin Heidegger）关于技术本质的追问在数智时代获得了新的理论回响——技术是否改变了存在者的存在方式？

弗卢瑟（Vilém Flusser）的技术图像理论为理解数字本体论提供了重要参照。他指出，数字编码创造了一种全新的存在层次——"技术图像"（Technical Image），这种图像不再指向现实，而是构成了

更多免费行业报告 并购家 www.ipoipo.net

自身的现实。在数智人文的实践中，这一理论洞见具有直接的方法论启示：当我们将《清明上河图》数字化为高分辨率像素矩阵，或将《论语》编码为词向量序列时，我们面对的已经不是原初的艺术品或文本本身，而是一种经过技术中介的新的存在形态。

数字人文学者必须面对数字物与物理实体之间的本体论差异，探究数据作为一种新的人类存在方式的哲学意涵。算法生成的虚拟主体是否具备某种形式的本体地位，是否需要建立新的本体论范畴来容纳这些数字存在者，这些问题已经超越了传统人文学科的理论框架，要求跨学科的哲学反思。更进一步而言，数字化转型所创造的不仅是新的研究对象，更是一种新的存在论处境：人类越来越多地通过数字中介来认识、理解和阐释自身的文化传统与历史记忆，这种中介化的认知模式本身就构成了值得深入探究的哲学命题。

（2）计算理性与人文理解的认识论张力

在认识论层面，核心争议在于机器理解与人类理解的本质区别。塞尔（John Searle）经典的"中文房间"思想实验揭示了语法操作与语义理解之间的鸿沟。人工智能的计算逻辑是基于句法和统计相关性，而非基于语义和真实世界的经验。机器能够快速总结出一部哲理名著的主题，但它并不理解痛苦、死亡或自由的真实重量。

德雷福斯（Hubert Dreyfus）基于海德格尔与梅洛-庞蒂的现象学传统，对人工智能的认识论局限进行了深刻批判。他指出，人类的理解植根于身体性的"在世存在"（Being-in-the-world），这种具身性

的认知维度是纯粹的形式化计算所无法企及的。因此，我们必须厘清计算理性的边界，承认算法在处理复杂情感、隐喻与道德判断时的局限性。数智人文的认识论并非要用机器计算取代人类阐释，而是要在算法输出与人类直觉之间建立一种辩证的对话机制。正如伽达默尔所强调的，真正的理解总是发生在不同视域的碰撞与融合之中。在数智人文的语境下，算法的视域与人文学者的视域之间的辩证对话，有望产生超越两者各自局限的新的理解可能。

更多免费行业报告 并购家 www.ipoipo.net

8.3 数智时代的人文再定位

在数智时代，人文学科如何避免沦为技术的附庸，如何以批判性与反思性参与塑造技术方向，并在组织与制度层面形成可持续的学术生态。它把价值立场落回到可执行的机制安排——数据治理、算法透明、学术规范与协作方式。

1. 技术批判：将算法作为审视对象

技术从来不是价值中立的。温纳（Langdon Winner）在其经典论文中指出，技术人工物具有内在的政治性。每一行代码、每一个参数的调整，都暗含着特定的利益导向与价值预设。人文学者应当成为算法的批判者，将人工智能本身作为审视与解剖的对象。我们需要运用历史的、哲学的、伦理的眼光，去审视数据采集的合法性、算法模型的公平性以及技术权力对人类主体性的侵蚀。

在具体实践中，这种技术批判可以从以下几个层面展开：一是数据批判，考察训练数据的来源、构成与潜在偏见，揭示数据生产过程中的权力关系；二是算法批判，分析模型架构的设计逻辑与隐含假设，审视算法决策的透明性与可解释性；三是应用批判，关注技术部署的社会后果，包括对学术劳动分工、知识产权制度以及文化多样性的影响。弗洛里迪（Luciano Floridi）所倡导的"信息伦理学"（Information Ethics）为这种多层次的技术批判提供了系统性的理论框架。

尤为重要的是，人文学者应当警惕"技术决定论"（Technological

Determinism）的思维陷阱。技术的发展方向并非自然而然、不可改变的。人文学科的批判性介入，恰恰是为了确保技术发展的方向能够充分反映人类的多元价值与文化诉求。从历史的视角看，每一次重大技术变革——从印刷术到互联网——都曾引发人文学科的深刻焦虑与创造性回应。数智时代的人文学科同样需要在焦虑与回应之间找到平衡，既不盲目排斥技术创新，也不放弃对技术的价值审视与伦理约束。

2. 范式融合：构建人机协同的新型学术生态

我们需要倡导一种算法意识与人文关怀并重的新型学术范式。未来的数字人文学者既要掌握计算技术，能够与工程师进行深度对话；又要坚守人文底线，以人类的尊严、自由与全面发展为最高准则。这种跨学科的素养培养已成为全球人文教育改革的重要方向。

在学术共同体层面，需要建立人文学者与计算机科学家之间的深度协作机制。斯坦福大学文学实验室（Stanford Literary Lab）、哈佛大学文化观测所（Harvard Cultural Observatory）以及北京大学数字人文研究中心等机构的实践表明，跨学科团队的有效协作是数智人文研究取得突破性成果的关键。同时，需要构建开放共享的数据基础设施与工具平台，降低人文学者进入数智研究的技术门槛，促进人文计算方法的普及与标准化。

在人才培养层面，新型数智人文学者的培育需要打破传统的学科壁垒。高等院校应当探索设立数智人文交叉学位项目，将编程能力、

更多免费行业报告 并购家 www.ipoipo.net

数据素养与统计思维纳入人文学科的核心课程体系，同时在计算机科学的教育中融入人文关怀与伦理反思的维度。这种双向渗透的培养模式，旨在造就一批既具备技术操作能力又拥有深厚人文素养的"桥梁型"学者，他们将成为未来数智人文研究的中坚力量。此外，学术评价体系也需要相应调整，建立能够充分认可跨学科成果价值的评审标准，为数智人文研究的可持续发展提供制度保障。

唯有如此，人文学科才能在技术革新的惊涛骇浪中锚定自身的位置，引导人工智能向着增进人类福祉、丰富人类精神世界的方向演进。

参考文献

[1] 柯平,宫平.数字人文研究演化路径与热点领域分析[J].中国图书馆学报,2016,42(06):13-30.

[2] MORETTI F. Distant Reading[M]. London: Verso, 2013.

[3] VASWANI A, SHAZEER N, PARMAR N, et al. Attention Is All You Need[C]//Advances in Neural Information Processing Systems. Long Beach: Curran Associates, 2017: 5998-6008.

[4] 令小雄.DeepSeek 开启后 ChatGPT 时代——基于数字范式革新及其运演哲思[J].西北工业大学学报(社会科学版),2025,(02):59-67.

[5] 孙茂松,金兼斌.人工智能及其社会化应用的现状和未来——《全球传媒学刊》就社会化智能信息服务专访孙茂松教授[J].全球传媒学刊,2018,5(04):3-12.

[6] JOCKERS M L. Macroanalysis: Digital Methods and Literary History[M]. Urbana: University of Illinois Press, 2013.

[7] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]//Proceedings of NAACL-HLT. Minneapolis: ACL, 2019: 4171-4186.

[8] MICHEL J B, SHEN Y K, AIDEN A P, et al. Quantitative Analysis of Culture Using Millions of Digitized Books[J]. Science, 2011, 331(6014): 176-182.

[9] 陈力.数字人文视域下的古籍数字化与古典知识库建设问题[J].中国图书馆学报,2022,48(02):36-46.

[10] BERRY D M. Understanding Digital Humanities[M]. London: Palgrave Macmillan, 2012.

[11] GADAMER H G. Truth and Method[M]. 2nd ed. London: Continuum, 2004.

[12] 刘炜,叶鹰.数字人文的技术体系与理论结构探讨[J].中国图书馆学报,2017,43(05):32-41.

[13] BLEI D M. Probabilistic Topic Models[J]. Communications of the ACM, 2012, 55(4): 77-84.

[14] HAMILTON W L, LESKOVEC J, JURAFSKY D. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change[C]//Proceedings of ACL. Berlin: ACL, 2016: 1489-1501.

[15] 崔竞烽,郑德俊,王东波,等.基于深度学习模型的菊花古典诗词命名实体识别[J].情报理论与实践,2020,43(11):150-155.

[16] BOL P K. GIS, Prosopography, and History[J]. Annals of GIS, 2012, 18(1): 3-15.

[17] 朱本军,聂华.跨界与融合:全球视野下的数字人文——首届北京大学“数字人文论坛”会议综述[J].大学图书馆学报,2016,34(05):16-21.

[18] BOL P K, GE J X. China Historical GIS[J]. Historical Geography, 2005, 33: 150-167.

更多免费行业报告 并购家 www.ipoipo.net

[19] KOSELLECK R. Futures Past: On the Semantics of Historical Time[M]. New York: Columbia University Press, 2004.

[20] 曾骏,王子威,于扬,等.自然语言处理领域中的词嵌入方法综述[J].计算机科学与探索,2024,18(01):24-43.

[21] 陈跃红.变革与创新：人工智能的迭代发展与人文学科的未来[J].人民论坛·学术前沿,2024,(05):85-97.DOI:10.16619/j.cnki.rmltxsqy.2024.05.009.

[22] BENJAMIN W. The Work of Art in the Age of Mechanical Reproduction[M]//Illuminations. New York: Schocken Books, 1969: 217-251.

[23] FOUCAULT M. What Is an Author?[M]//Aesthetics, Method, and Epistemology. New York: The New Press, 1998: 205-222.

[24] BENDER E M, GEBRU T, MCMILLAN-MAJOR A, et al. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?[C]//Proceedings of FAccT. Virtual Event: ACM, 2021: 610-623.

[25] NOBLE S U. Algorithms of Oppression: How Search Engines Reinforce Racism[M]. New York: NYU Press, 2018.

[26] 耿玉芳.人工智能在中华优秀传统文化传承创新方面的应用研究[J].出版广角,2019,(11):43-45.

[27] 吴汉东.人工智能生成作品的著作权法之问[J].中外法学,2020,32(03):653-673.

[28] JI Z, LEE N, FRIESKE R, et al. Survey of Hallucination in Natural Language Generation[J]. ACM Computing Surveys, 2023, 55(12): 1-38.

[29] 赵豪迈,郑福涛,赵佳悦.人工智能时代学术传播规范的重构——以生成式人工智能 ChatGPT 为例[J].陕西师范大学学报(哲学社会科学版),2024,53(05):87-102.

[30] HEIDEGGER M. The Question Concerning Technology[M]//The Question Concerning Technology and Other Essays. New York: Harper & Row, 1977: 3-35.

[31] FLUSSER V. Into the Universe of Technical Images[M]. Minneapolis: University of Minnesota Press, 2011.

[32] 吴冠军.速度与智能——人工智能时代的三重哲学反思[J].山东社会科学,2019,(06):13-20+160.

[33] SEARLE J R. Minds, Brains, and Programs[J]. Behavioral and Brain Sciences, 1980, 3(3): 417-424.

[34] DREYFUS H L. What Computers Still Can't Do: A Critique of Artificial Reason[M]. Cambridge: MIT Press, 1992.

[35] WINNER L. Do Artifacts Have Politics?[J]. Daedalus, 1980, 109(1): 121-136.

[36] CRAWFORD K. Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence[M]. New Haven: Yale University Press, 2021.

[37] FLORIDI L. The Ethics of Artificial Intelligence[M]. Oxford: Oxford University Press, 2023.

[38] 冯惠玲.新文科与数字人文教育之新[J].数字人文研究,2022,2(04):14-21.

[39] 王静静,叶鹰.国际数字人文研究中的跨学科知识扩散探析[J].大学图书馆学报,2021,39(02):45-51+61.

[40] 夏翠娟.面向人文研究的“数据基础设施”建设——试论图书馆学对数字人文的方法论贡献[J].中国图书馆学报,2020,46(03):24-37.

[41] 周维东.数智时代的人文知识生产[J].江西社会科学,2025,45(08):174-183.

更多免费行业报告 并购家 www.ipoipo.net

第九章 从效率到责任的重建：人工智能嵌入公共治理的应用模式与责任重构[①]

随着大数据、算法模型和算力基础设施的快速发展，人工智能特别是生成式人工智能正在加速嵌入公共治理全过程。从智能问答、智能导办、智能审批到辅助决策、智能推送，AI 不仅提升了政务服务的响应速度、供给精度和协同效率，也推动政府治理从被动响应走向主动服务、从经验判断走向数据驱动、从流程优化走向能力重塑。与此同时，人工智能已不再只是后台工具，而是逐步进入权利实现、程序运行、责任分配和公共价值塑造等治理核心环节，深刻重构公共部门与公民、技术与制度、效率与正当性之间的关系。然而，人工智能赋能公共治理并不只是技术进步的线性展开。算法“黑箱”、模型“幻觉”、智能鸿沟、组织能力退化、权责边界模糊以及公共价值偏移等风险，正在不断显现。尤其是在代理与辅助、人机协同与技术替代之间，公共治理必须审慎划定人工智能的应用边界。

本章围绕人工智能在政务服务中的应用场景、嵌入公共治理的两种模式及其责任重构问题展开讨论，力图在效率提升与价值坚守之间，探寻负责任的人工智能治理路径。

[①] 本章由复旦大学国际关系与公共事务学院郑磊教授牵头，复旦大学国际关系与公共事务学院博士生张宏、陈奕醇和杨家西合作撰写。

郑磊教授现任复旦大学数字与移动治理实验室主任、中国信息协会数字治理专业委员会副会长、国际数字政府学会中国分部联合主席、《联合国电子政务调查报告》专家组成员、上海市公共数据开放专家委员会委员。

本章提要

- 人工智能正广泛应用于智能问答、智能导办、智能审批、辅助决策和智能推送等场景，推动政务服务由被动响应转向主动、精准和智能化供给。
- AI 在提升行政效率、优化服务流程和强化数据研判能力的同时，也带来了智能鸿沟、算法偏见、数据安全和服务不平等等新问题。
- 人工智能嵌入公共治理主要呈现代理型与辅助型两种模式，二者在决策权配置、责任归属和治理正当性方面存在根本差异。
- 不同治理事项应根据规则明确性、权益影响程度、价值冲突复杂性和社会信任水平，审慎选择 AI 的介入深度，高风险场景应坚持“人机协同、人在回路”。
- 需警惕辅助型系统在实践中因自动化偏差、组织依赖和绩效导向而滑向事实上的代理型治理，进而引发责任卸载和权力异化。
- 负责任的人工智能治理，应围绕数据、模型、部署、使用、纠错和救济等环节重建全生命周期责任链，确保公共责任不能被技术或供应商替代。
- 推进公共治理中的 AI 应用，应同步完善说明义务、人工复核、申诉救济、审计监督和动态评估机制，把公平、透明、隐私、安全与公众信任嵌入技术治理全过程。

更多免费行业报告 并购家 www.ipoipo.net

9.1 AI 在政务服务中的应用场景及潜在风险[①]

随着大数据技术与算力供给的持续增强，以 ChatGPT、DeepSeek 等为代表的生成式人工智能正引发公共治理领域的深刻变革。我国政务服务从早期以信息查询为主的自动化阶段，正迈向能够理解复杂意图、提供个性化引导、支持多轮交互的人机协同智能化新阶段。

2024 年 1 月，国务院印发的《关于进一步优化政务服务提升行政效能推动“高效办成一件事”的指导意见》明确鼓励“探索应用自然语言大模型等技术，提升线上智能客服的意图识别和精准回答能力，优化智能问答、智能搜索、智能导办等服务”。[1]随后，《关于健全“高效办成一件事”重点事项常态化推进机制的意见》中进一步提出“探索开展‘人工智能+’政务服务”，“在确保安全的前提下稳妥有序推进人工智能大模型等新技术在政务服务领域应用”。[2] 随着人工智能在政务服务中的广泛应用，《关于深入实施“人工智能+”行动的意见》提出要“安全稳妥有序推进人工智能在政务领域应用，打造精准识别需求、主动规划服务、全程智能办理的政务服务新模式。”[3] 对此，中央网信办与国家发改委联合发布的《政务领域人工智能大模型部署应用指引》中系统划分了人工智能在智能政策问答、表单智能预填等 13 项重点应用场景，并对安全合规、能力边界与部署要求作出规范，为地方探索提供了可遵循的制度框架。[4]国家政策的持续出台与目标

[①] 本节由复旦大学国际关系与公共事务学院博士生陈奕醇撰写。
陈奕醇，复旦大学数字与移动治理实验室研究助理，主要研究领域为数字治理、移动政务服务、AI 公共服务。

指引，推动了 AI 在政务服务从试点探索向规模化应用。然而，AI 在政务服务应用中，仍面临着可靠性与权威性、服务普惠性与算法偏见、隐私保护与数据安全、责任界定与可问责性等方面的张力与挑战。因此，当前 AI 在政务服务中的应用场景如何，又可能面临哪些风险和挑战？本文就此展开初步的探索和思考。

1. AI 在政务服务中的应用与前景

（1）智能问答：从关键词匹配到语义理解与场景化服务

随着 ChatGPT、DeepSeek 等大语言模型的突破性进展，融合了自然语言处理（NLP）、知识图谱和检索增强生成（RAG）技术的政务问答系统，正在以两种主要形态深度嵌入政务服务体系中。

一方面，AI 通过具有高度拟人化特征的“数字人”重塑政务服务的形态与边界。[5]通过集成语音识别、合成以及计算机视觉生成技术，创造出高度拟人化的虚拟形象，[6]推动服务从传统人力窗口或静态图文界面向可视、可听、可互动的人机协同模式转型。例如，江西政务服务网上线的“小赣事”政务数字人，能够以生动的虚拟形象提供 7×24 小时的办事咨询与政策解读服务，显著提升了交互的亲和力与信息获取的直观性。[7]上海浦东新区推出的“小浦”数字人则进一步整合了区内多项政务服务入口，不仅能够回答问题，更能直接引导用户跳转至相关业务办理页面，初步实现了“问答即服务”的闭环。[8]苏州吴中区政务大厅的数字人“小吴”能够通过语音及文字互动，为用户提供实时解答和精准指引，辅助群众完成业务全流程办理。

更多免费行业报告 并购家 www.ipoipo.net

另一方面，AI 在政务 APP 平台集成的智能客服，具备语义理解、知识检索和文本推理等能力，[9]已成为人工客服之外移动端政务服务的主流交互方式，其核心能力表现为三个方面：首先是深度语义理解与上下文关联。江苏“苏服办”的智能客服能准确解析用户多轮提问提供连贯精准的答复，并主动推送办事入口，实现了从“答问”到“导办”的跨越。其次，是全渠道信息的整合与呈现。福建“闽政通”的智能客服在回答时，能同步提供线上办理链接、线下办理地址、所需材料清单乃至关联事项推荐，并提供语音朗读功能，构建了立体化的信息服务矩阵。最后，是个性化服务的初步探索。广西“智桂通”能够根据用户的注册身份和业务类型，动态调整回答的侧重点与推荐流程，提升了用户服务的针对性与有效性。

（2）智能导办：从静态指南到动态流程的智能化重构

智能导办是人工智能在政务服务中的进一步深度应用，标志着服务模式从被动应答向主动介入的转变。在智能问答基础上，AI 进一步帮助用户进入业务流程，将传统静态的办事指南清单转变为动态交互式的办事导引。基于自然语言理解、计算机视觉、智能表单与流程引擎等技术，AI 在理解用户办事意图基础上能够直接引导或辅助用户完成业务办理的各个环节，实现政务服务的在线便捷办理。例如，浙江“浙里办”的“小浙”助手已能够通过多轮对话了解用户办理“企业开办一件事”“科技成果转化一件事”等复杂集成服务的具体情形，并逐步引导至正确的在线申报通道。重庆“渝快办”的“渝小智”则深度融合数据查询与办事指引，在用户咨询“新生儿出生一件事”时

能清晰地展现业务办理流程，并告知户籍登记、医保参保等环节所需材料，实现了对事项流程的智能导办。

智能导办首先基于 AI 的多模态交互启动，对用户材料进行自动提取。用户通过文本输入或者语音提示等方式启动 AI 客服的智能导办服务，借助 OCR 与图像理解技术，AI 客服能够对用户上传的身份证、营业执照等证件图片或文件的关键字段进行智能提取，并将其精准填充至对应的申请表格中实现一键智能预填，大幅减少用户重复录入操作成本。其次，智能导办能根据用户画像和具体办事需求动态生成个性化的办事材料清单、流程图及注意事项，部分系统还能依据基础信息自动生成如“申请书”“承诺书”等标准文书的初稿供用户确认修改，显著降低了办事的材料准备门槛。最后，在用户提交电子材料阶段，AI 能够预审用户上传的材料文件格式、内容完整性及逻辑一致性，及时提醒用户材料信息错误，有效避免了因材料不合规导致的多次退件，提升了首次申报成功率与用户满意度。

（3）智能审批：从人工核验到智能辅助决策的流程再造

除了面向用户端的智能问答和智能导办，AI 也在向后台行政审批环节渗透，推动了后台审批流程的自动化与智能化再造。[10]基于自然语言处理、计算机视觉、光学字符识别和规则引擎等技术，AI 能够对用户上传的申请材料进行自动化的合规性审查、风险研判与辅助决策，从而将审批人员从重复繁琐的格式与要件审查中解放出来。[11]

在高频事项场景，智能审批极大提升了办事效率。在上海浦东新区政务服务大厅，以往审批材料全靠人工阅读，庞大的办件量给各个

更多免费行业报告 并购家 www.ipoipo.net

服务窗口带来了巨大压力。当浦东新区将 AI 应用于辅助审批后，窗口工作人员收件时长由 20 分钟缩减到 10 分钟，培训上岗时间由 2 个月缩短到 15 天，政务服务大厅商事登记领域“首办成功率”也大幅提升至 98%。[12]这种模式将大量标准化审核工作前置并自动化，直接减轻了窗口人员压力。在专业事项上，智能审批还能够扮演“专家助理”角色。浙江通过将智能化辅助审批系统应用于环境影响评价领域，工作人员将环评报告上传后 AI 可在十多分钟内完成全文解析，自动判断其与环保法规、标准的符合性，并输出带有问题定位的审核清单。[13]这一应用将审批人员从繁重的资料查阅和基础比对中解放出来，使其能更聚焦于核心环境风险的研判和行政工作上。

（4）辅助决策：从任务执行到智能研判的效能提升

辅助决策是 AI 在政务服务中，从面向公众的“窗口”向支撑政府内部高效运转的“中枢”的深度延伸。根据中央网信办与国家发改委联合印发的《政务领域人工智能大模型部署应用指引》，AI 除了在政务服务中为群众提供智能问答、辅助办理和政策服务直达快享外，在后台机关办公中主要围绕辅助文书起草、资料检索和智能分办展开，[14]旨在为政府工作人员提供高效辅助。当接收到用户端的服务需求后，AI 能够智能分办来文工单，减少人工协调成本。同时，AI 还能够辅助起草公文报告，快速检索历史资料和政策文件，提高办公效率。

首先，在任务智能分办与执行层面，AI 通过“智能分办”重构了内部协同流程。传统基于人工阅读和经验的工单分类、公文分发模式效率低下。利用自然语言理解技术，AI 可以自动解析任务内容进行

智能填单以减少已知信息重复确认，根据预设规则实现任务的精准分类和自动派发。[15]例如，在海南省的“海政通”全省协同办公平台上，基层网格员通过语音即可一键上报事件，系统不仅能自动生成结构化工单，还能基于大模型智能推荐分办部门，这一模式使单件工单录入时长压缩了 80%，日均处理量提升 40%，显著降低了内部协调成本。[16]

其次，在知识辅助与文书处理层面，AI 正成为政务工作人员的“全能知识助理”。通过将大语言模型内嵌于政府内部的 OA 系统，能够为政府工作人员提供智能写作、智能问答、智能审核等应用功能，显著提升政府办公效率。[17]一方面，基于智能文书处理功能，AI 能够根据会议要点和文书模板辅助生成会议纪要和报告等公文初稿，并进行格式与内容的智能校对和润色，将工作人员从繁琐的文字工作中解放出来。另一方面，面对海量分散的政策文件和历史资料，AI 智能资料检索能够精准理解工作人员的模糊查询意图，实现跨部门和跨年度的快速定位、关联对比和要点摘要，确保政策解读的标准统一与高效。

最后，在宏观数据研判与决策支持层面，AI 已逐步为政府提供决策支撑、政策模拟和风险预警等功能，从日常办公应用的范畴开始触及治理核心。例如，在应急场景处置中，AI 可融合分析突发事件的多维信息，快速模拟不同救援方案的部署效果，为优化资源配置提供科学建议。[18]在政策评估方面，AI 能对政策出台后的社会反馈、经济指标等进行多维度分析，量化评估政策效果与潜在影响，为政策迭代提供数据支撑。[19]尽管此类深度应用尚处于探索阶段，但其反映了大

更多免费行业报告 并购家 www.ipoipo.net

模型在整合分析、复杂推理方面的巨大潜力，是推动政府从“经验依赖”向“数据决策”跃升的关键。[20]

（5）智能推送：从被动检索到主动服务的范式跃迁

智能推送代表了人工智能在政务服务中应用的高级阶段，其核心是实现从“人找政策、人找服务”的被动响应模式向“政策找人、服务找人”的主动供给模式的范式跃迁。[21]这不仅极大提升了政务资源的触达效率和普惠水平，也标志着政务服务开始从标准化供给向个性化、精准化和预见性的服务演进。

智能推送的实现依赖于一个完整的技术闭环，首先需要融合多源数据和构建用户画像，系统整合来自业务办理、浏览查询、地理定位等多维度数据，形成动态立体的用户数字画像。[22]其次运用推荐算法和自然语言处理等技术，建立用户需求与不断更新的政策知识与服务资源之间的映射关系，识别用户的个性化需求与偏好。[23]最后通过政务 APP、短信、企业服务平台等渠道将信息与服务精准主动地推送至用户，并基于用户反馈持续优化算法模型。[24]

当前，智能推送在实践中主要体现在以下两大场景：一是惠企利民政策的智能匹配与直达。传统政策传达存在信息不对称和不精准的问题，企业和个人往往“找不到、读不懂、不会办”。AI 通过构建“政策知识图谱”，对市场主体进行画像建构，从而实现政策内容的自动解析和精准匹配。例如杭州“亲清在线”平台推出的智能服务专员“西小服”，能在企业完成基础信息填报后，自动为其筛选、解读并推送可申报的各类奖励、补贴和税费优惠政策，降低了企业搜寻政策的成

本，极大提高了政策“免申即享”或“精准快享”的落地效率。[25]类似实践也已在上海“一网通办”企业专属网页、广东“粤商通”等平台广泛应用。

二是情景感知下的个人服务主动供给。在移动政务服务应用上，AI 通过对用户行为数据的深度分析实现关联事项推动与服务前置。当用户频繁查询社保缴纳明细时，政务 APP 可判断其可能正办理入职、购房或贷款手续，从而在首页显著位置或通过消息中心主动整合推送“社保参保证明开具”“养老待遇测算”等相关服务入口。例如，“闽政通”“浙里办”等应用推出的“猜你想办”和“常用服务”等功能正是基于用户习惯进行个性化主动供给的典型。

2. AI 在政务服务应用中面临的问题与风险

人工智能正推动着公共服务供给从人工服务向人机交互转型。当前，我国已有 70 余个省级或市级政府接入 DeepSeek 等大模型，智能创新应用遍地开花。[26]然而，这场以算法为驱动、以效率提升为表征的技术革新在高歌猛进的表象之下，也悄然积聚着多重风险与深层困境，技术赋能与公共价值之间的结构性张力正成为制约人工智能政务服务可持续发展的关键命题。

（1）智能应用鸿沟：技术门槛引发公共服务不平等

在人工智能加入嵌入政务服务体系背景下，智能应用鸿沟正成为制约公共服务均等化的重要障碍。智能鸿沟并非凭空产生，而是传统数字鸿沟在人工智能时代的深化与演变。[27]数字鸿沟的核心在于硬件

更多免费行业报告 并购家 www.ipoipo.net

接入与基础操作技能的差异，而智能鸿沟则表现为个体在认知能力、新型工具应用能力及人机协同适应能力上的结构性差距。[28]以算法驱动的政务服务革新在提升效率的同时，也构筑起了一道无形的认知与应用壁垒，使部分群体面临被边缘化的风险。[29]具体而言，智能鸿沟在政务服务领域主要表现为以下三重维度。

首先，表现为区域资源的接入鸿沟，不同地区在智能算力基础设施上的投入差异，将导致政务服务智能化水平的区域失衡。经济欠发达地区或因算力资源匮乏、技术基础薄弱，难以部署和运行高性能政务大模型，进而无法提供与发达地区同等质量的智能服务。[30]算力资源的区域性失衡正成为制约政务服务均等化进程的新瓶颈。

其次，表现为政府内部的技能鸿沟，智能鸿沟不仅存在于公众之间，亦广泛存在于公务员队伍内部。不同年龄、专业背景及地区的工作人员，在理解、掌握并运用 AI 工具辅助履职的能力上存在显著差异。[31]若缺乏系统性的数字素养培训与能力建设机制，可能导致“AI 赋能强者愈强、弱者愈弱”的马太效应，进而使智能服务在不同层级、不同地域的落地效果参差不齐，难以实现预期中的全域效能提升。

最后，表现为特殊群体的应用鸿沟，政务大模型所依托的交互式界面，对用户的数字素养与认知能力提出了更高要求。老年人、残障人士及低学历群体对于人工智能等新兴技术的认知与操作能力相对不足，[32]在面对高度依赖在线操作与智能交互的政务服务时往往陷入“进不去、不会用、办不成”的困境。[33]当前，人工智能技术迭代的高速性与用户能力提升的缓慢性之间存在显著张力，若一味追求技术

先进性与迭代速度而忽视群众需求的多样性与包容性设计，智能鸿沟将呈现持续扩大之势。

值得警惕的是，当前政务大模型的训练数据多来源于主流用户群体，对边缘群体的服务需求、语言习惯及行为模式覆盖不足，导致其在真实场景中的使用体验与办事成功率普遍偏低。这种效果差异可能进一步形成负向循环：使用效果差的群体更少使用系统，系统因而更难获取其数据进行优化，最终导致公共服务质量差距在技术演进中被持续放大。

（2）技术可靠风险：算法“黑箱”引致决策的权威性危机

人工智能在政务服务中的深度嵌入，不可避免地面临由其技术特性所引发的可靠性风险。这一风险集中体现为算法决策过程的不透明性、大模型固有的“幻觉”问题，以及由此引发的政府决策权威性的信任危机。[34]

一方面，算法“黑箱”的不可解释性侵蚀了决策正当性。以深度学习为代表的人工智能算法因其内部运作具有高度复杂性与不透明性，形成难以追溯的“黑箱”问题。[35]当这一难以解释的算法机制被应用于实际决策时，就会导致决策过程无法被理解，决策结果可能面临质疑。在传统行政框架下，行政决策行为的做出需要遵循一套程式化制度，要求决策者对其治理行为及其选择做出合理解释。[36]然而，当审批决定、政策解释等由 AI 系统生成时，其推理路径隐藏在数以亿计的参数之中，即使技术人员也难以完整追溯其决策逻辑。这种技术黑箱与行政公开原则之间的内在张力，使公众对算法决策的正当性

更多免费行业报告 并购家 www.ipoipo.net

产生怀疑，[37]进而侵蚀政府决策的权威性与公信力。

另一方面，大模型的“幻觉”问题挑战政务信息的权威性。生成式人工智能由于大模型的概率统计推理逻辑并不能理解实际含义，因而不可避免地存在编造虚假信息的“AI 幻觉”现象。[38]在政务服务场景中，这一缺陷将会被急剧放大。当 AI 生成看似合理实则错误的政策解读、办事指引或法规适用建议被直接输出给公众时，将导致用户和企业“白忙活一场”，严重损害政府信息的权威性与准确性。由于政策法规更新频繁，模型训练与知识库更新存在天然滞后，研究表明，政务 AI 知识库更新周期超过 6 个月的系统，其错误率会提升 2.3 倍。[39]因此，AI 系统在快速变化的政策环境中极易输出“看似合理”的错误信息，形成新型信息服务风险。

综上而言，人工智能以概率统计为核心的技术逻辑，与政务服务以规范权威为核心价值的制度逻辑之间存在根本性张力，决定了 AI 在政务服务中的应用必须始终保持谨慎的“辅助型”定位，而不能轻易赋予其独立的决策权限。

（3）组织转型风险：人机协同引发的能力退化与权力重构

人工智能在政务服务中的深度应用，不仅重塑着前台的服务形态，更深刻影响着政府组织内部的权力结构、运行逻辑与能力构成。但人机协同的新型组织形态在提升效率的同时，亦潜藏着能力退化、权力转移与责任边界模糊等多重风险。

首先，AI 过度依赖可能导致治理主体的自主性下降。[40]当工作人

员日益习惯于依赖 AI 完成公文起草、信息检索、初步审核等任务时，其自身的业务判断能力、政策理解能力和沟通协调能力可能会逐渐退化。[41]一旦形成对 AI 系统的过度依赖，当系统出现故障或面临超出其能力边界的复杂情形时，工作人员能否及时补位和有效应对便成为一个不容回避的现实问题。据工业和信息化部调查显示，76%的政务 AI 系统缺乏冗余供应商预案，这意味着一旦核心 AI 服务中断，大量依赖 AI 系统的政务流程可能陷入瘫痪。[42]同时，AI 的广泛应用也引发工作人员的“技能焦虑”，担心自己被技术替代、担心技能过时、担心在新型人机协同中找不到定位。[43]这种焦虑若得不到有效疏导，可能转化为对技术应用的抵触情绪，影响组织变革的顺利推进。

其次，AI 应用可能导致决策权力向技术开发者转移。大数据与人工智能的深度嵌入，在一定程度上推动决策权向少数掌握数据资源和技术优势的层级、部门或人员集中。当具备数据采集、处理与解释能力的部门在治理体系中的地位持续上升时，其他传统业务部门的专业判断力与决策权可能被边缘化。[44]这种技术性权力重构以“数据即权力”的逻辑逐步渗透政府组织结构，催生出新型“数字科层”，[45]使得“整体性治理”在操作层面可能异化为“集权式整合”，跨部门协同实则是技术能力的非均衡分布而形成新的权力垄断。

更为深层的问题在于，当前大多数政务 AI 系统由政府部门与私营科技公司合作开发，这意味着公权力行使过程中出现了私人力量的深度嵌入。[46]随着算法系统对行政决策影响程度的提升，技术设计者、训练者和维护者得以在事实上获得对公共决策的间接影响力。[47]这种

更多免费行业报告 并购家 www.ipoipo.net

权力转移缺乏传统公法框架下的民主监督与责任约束，形成了技术赋权与民主问责之间的结构性脱节。[48]当算法逻辑替代行政逻辑、技术理性僭越公共理性时，如何确保技术开发者对公共利益负责，便成为亟待回应的制度性挑战。

（4）制度伦理风险：权责归属模糊与公共价值偏移

人工智能嵌入政务服务所引发的深层风险不仅局限于技术效能层面，更深刻触及制度规范与公共伦理的根基。当算法系统在公共事务中扮演“准决策者”角色时，传统以人为中心的行政责任体系、人文情感关怀及公共价值分配逻辑，正面临系统性冲击与重构压力。

首先，人机协同可能带来行政责任的归属模糊。在人工智能深度介入政务服务供给的背景下，行政决策与公共服务从传统的“人—人”直接互动，演变为“人—机—人”的复杂链条。当公民对算法行政结果提出质疑时，行政机关诉诸“技术复杂性”或“技术中立”以回避实质性责任，技术部门则以“算法专业性”为由推卸责任，将会形成“集体无责任”现象，削弱政府决策的权威性与公信力。[49]在这一模糊地带，公众难以通过传统的陈述、申辩、复议等程序机制进行监督问责，政府决策的权威性与公信力将会被持续侵蚀。[50]

其次，服务渠道单一化将导致客服人文关怀下降与个性化服务缺失。随着业务流程向线上迁移、办事体验高度依赖智能交互、政策咨询默认由虚拟客服承接，技术理性将会逐渐挤压政务服务中应有的人际温度与情感支持。在实际体验中，用户在使用移动政务应用 AI 客服时普遍遭遇“找不到入口”“答非所问”“交互缺乏温度”等问题，

而人工客服的缺位或响应滞后则进一步加剧了服务的可及性危机。这一现象折射出当前 AI 技术应用对用户多样性认知不足、服务供给模式趋同化，以及包容性保障系统性缺失等问题。当技术迭代以“效率优先”为名压缩传统服务渠道时，政务服务的人文关怀与基本可及性可能被无形消解。

最后，效率导向的价值预设可能引发公共价值的结构性偏移。人工智能技术内嵌的效率至上逻辑与政务服务以公平正义为核心的价值追求之间存在内在张力，[51]当流程简化率、响应速度、办结时效等量化指标成为评价政务 AI 成效的硬性标准时，那些难以量化但关涉社会公平、包容参与和人格尊严等隐性指标便可能在无形中被边缘化。政务 AI 应用若缺乏公共价值的刚性约束，极易陷入“重技术轻治理、重形式轻实质”的表面化陷阱。在强调“技术赋能”的政策话语下，盲目追求技术领先与概念创新，却忽视了政务服务的本质在于满足人民需求、解决实际问题，最终可能导致“数字官僚主义”的滋生蔓延。[52]

（5）数据安全风险：数据汇聚与深度处理引发的安全隐忧

人工智能在政务服务中的有效运行高度依赖于对海量数据的采集、汇聚与深度分析，数据既是驱动政务智能的核心资源，亦可能成为系统性风险的源头。在数据的大规模集中与智能化处理过程中，安全与隐私保护面临前所未有的挑战。

一方面，数据语料污染可能引发决策误导与安全漏洞。政务大模型的训练与优化高度依赖高质量数据语料，若训练数据被恶意投毒、

更多免费行业报告 并购家 www.ipoipo.net

标注不规范或包含错误信息，将直接导致模型输出出现系统性偏差。[53]攻击者可通过污染公开数据源或利用数据采集漏洞向政务 AI 的语料库植入虚假信息、偏见内容或恶意指令，使模型在后续运行中持续输出错误政策解读、歧视性意见乃至敏感信息。[54]大模型语料污染具有高度隐蔽性与不可控性，一旦发生难以通过常规检测手段发现，可能形成隐藏在政务服务中的“定时炸弹”。

另一方面，数据采集、存储与使用各环节均存在泄露风险。在数据采集环节，公共部门在运用人工智能为群众提供便利化服务时，往往会存在超越业务必要限度的信息收集行为，如强制要求用户进行人脸识别、过度采集住址信息、生物特征等敏感数据，对公民隐私权存在潜在侵犯，[55]也为数据窃取等违法行为提供了可能。[56]在数据存储环节，政务数据平台若存在加密措施不到位和访问权限管控不严等问题，海量政务数据将面临被攻击、窃取或内部泄露的风险。在数据流通与使用环节，由于当前人工智能语料库普遍存在数据分级分类制度不完善、数据时效性差异显著、标注质量参差不齐等问题，[57]政务 AI 模型可能输出过时、错误甚至带有偏见的信息，进而影响公共服务的准确性与公正性。

此外，政务 AI 系统通常涉及多主体协同开发与运维，包括政府部门、技术供应商、数据服务商等多元主体，数据在各主体间的流转进一步放大了泄露风险。当数据控制权与处理权分散于多个组织时，责任边界趋于模糊，任何环节的防护短板都可能成为攻击者的突破口。更为严峻的是，政务数据的高价值属性使其成为高级持续性威胁的重

点目标，[58]针对政务 AI 系统的定向攻击、供应链渗透和数据窃取，正成为新型网络安全博弈的前沿地带。

参考文献

[1] 中国政府网. 国务院关于进一步优化政务服务提升行政效能推动“高效办成一件事”的指导意见 国发〔2024〕3 号 [EB/OL]. (2024-01-16)[2026-01-05]. https://www.gov.cn/zhengce/zhengceku/202401/content_6926256.htm

[2] 中国政府网. 国务院办公厅关于健全“高效办成一件事”重点事项常态化推进机制的意见 国办发〔2025〕24 号 [EB/OL]. (2024-01-16)[2026-01-05]. https://www.gov.cn/zhengce/zhengceku/202507/content_7031051.htm

[3] 中国政府网. 国务院关于深入实施“人工智能+”行动的意见 国发〔2025〕11 号 [EB/OL]. (2025-08-21)[2026-01-05]. https://www.gov.cn/gongbao/2025/issue_12266/202509/content_7039598.html

[4] 中国网信网. 政务领域人工智能大模型部署应用指引 [EB/OL]. (2025-10-10)[2026-01-06]. https://www.cac.gov.cn/2025-10/10/c_1761819469929310.htm

[5] 李姝婷, 邱茜. 当 AI 成为“公仆”：政务服务数字人类型对政府信任的影响研究 [J]. 公共行政评论, 2025, 18(04): 1-21.

[6] 韩啸, 李静怡, 马亮. 政务服务数字人：理论诠释、实践困境与因应路径 [J]. 电子科技大学学报(社科版), 2024, 26(06): 34-43.

[7] 中国新闻网. “赣服通”5.0 版上线 发布全国首个政务服务数字人 [EB/OL]. (2022-11-16)[2026-01-06]. https://baijiahao.baidu.com/s?id=1749654665929527947&wfr=spider&for=pc

[8] 上海市人民政府. 浦东上线智能政务服务数字人 办事可找靠谱的“小浦” [EB/OL]. (2024-06-13)[2026-01-06]. https://www.shanghai.gov.cn/nw4411/20240613/1a5cb060218846c583a8761b9bc03f40.html

[9] 王芳, 魏中瀚, 连芷萱. 政务智能问答系统评价指标体系构建与测评问题编制 [J]. 图书情报知识, 2023, 40(06): 98-111.

[10] 孙妍, 成福. 敏捷致智：敏捷治理引领政务服务智能化转型的内在机理研究——以济南市智慧审批实践为例 [J]. 甘肃行政学院学报, 2023, (05): 68-81.

[11] 刘晓洋. 人工智能重塑政务服务流程的认知逻辑与技术路径 [J]. 电子政务, 2019, (11): 104-111.

[12] 聚焦“高效”“办成”两个关键点推进“人工智能+政务服务”改革 [J]. 中国行政管理, 2024, 40(05): 147-149.

[13] 浙江省生态环境厅. 中国环境报头版头条：浙江环评改革用好“加减法” 审批时限压缩至法定时限的 1/4，完善“浙里环评”数字应用 [EB/OL]. (2025-

更多免费行业报告 并购家 www.ipoipo.net

07-29)[2026-01-05]. https://sthjt.zj.gov.cn/art/2025/7/29/art_1201344_58959828.html

[14] 中央网信办. 政务领域人工智能大模型部署应用指引 [EB/OL]. (2025-10-10)[2026-01-05]. https://www.cac.gov.cn/2025-10/10/c_1761819469929310.htm

[15] 孟庆国, 鞠京芮. 人工智能支撑的平台型政府：技术框架与实践路径 [J]. 电子政务, 2021, (09): 37-46.

[16] 海南省大数据发展中心. 超 1600 项政务事项实现“指尖流动”，海南“海政通”打造自贸港政务服务新范式 [EB/OL]. (2025-10-10)[2026-01-05]. https://dsj.hainan.gov.cn/ywdt/jwyw/202509/t20250916_3931080.html

[17] 董昌其. 以场景为牵引的政务大模型应用：范式转型、实践分野与能力生成 [J]. 天津行政学院学报, 2026, 28(01): 29-40.

[18] 钟爽, 徐小峰. 人工智能在应急管理中的应用及其风险分类——基于文献与案例的综述研究 [J]. 北京航空航天大学学报(社会科学版), 2024, 37(05): 26-40.

[19] 吴进进, 符阳. 算法决策：人工智能驱动的公共决策及其风险 [J]. 开放时代, 2021, (05): 194-206+10.

[20] 张福磊, 吴玲. 人工智能大模型赋能城市基层治理：内在逻辑、现实挑战与推进路径 [J]. 西南民族大学学报(人文社会科学版), 2025, 46(12): 200-210.

[21] 米加宁, 于楠楠. “人工智能+”驱动地方政府治理现代化的逻辑进路与实践图景 [J]. 福建论坛(人文社会科学版), 2026, (01): 86-98.

[22] 黄梅银, 易兰丽, 王理达. 政务服务中的智能推送：需求、应用模式和实现路径 [J]. 电子政务, 2020, (02): 11-20.

[23] 许海玲, 吴潇, 李晓东, 等. 互联网推荐系统比较研究 [J]. 软件学报, 2009, 20(02): 350-362.

[24] 黄梅银, 易兰丽, 王理达. 政务服务中的智能推送：需求、应用模式和实现路径 [J]. 电子政务, 2020, (02): 11-20.

[25] 杭州市人民政府. 西湖区 AI 助企专员“西小服”升级版上线 [EB/OL]. (2025-10-10)[2026-01-05]. https://www.hangzhou.gov.cn/art/2024/7/29/art_812262_59100566.html

[26] 人民网. 人工智能政务发展需警惕智能鸿沟 [EB/OL]. (2025-10-10)[2026-01-05]. http://theory.people.com.cn/n1/2025/0913/c40531-40563157.html

[27] 彭兰. 智能传播时代“智能鸿沟”的走向探询 [J]. 中国编辑, 2024, (11): 19-26.

[28] 孙榕, 李白杨. 生成式人工智能（GAI）背景下的新型数字鸿沟识别框架研究 [J]. 图书情报知识, 2025, 42(03): 19-30.

[29] 中国共产党新闻网. 人工智能政務發展需警惕智能鴻溝. http://theory.people.com.cn/BIG5/n1/2025/0913/c40531-40563157.html

[30] 侯冠宇, 胡宁宁. 赋能城市高质量发展的新质生产力培育研究 [J]. 城市发展研究, 2025, 32(11): 90-96.

[31] 朱仁崎, 潘晓庆, 应验. 干部数字素养的指标构建与评估 [J]. 中国行政管理, 2024, 40(07): 29-38.

[32] 李翌萱. 人工智能时代老年人社会参与的价值审视和伦理思考 [J]. 自然辩证法通讯, 2021, 43(06): 88-93.

[33] 赵淼, 刘银喜. 公共价值共创何以弥合老年数字鸿沟？——基于上海市“数字伙伴计划”的案例研究 [J]. 电子政务, 2026, (03): 16-28.

[34] Busuioc, M. Accountable artificial intelligence: Holding algorithms to account [J]. Public Administration Review, 2020, 81(5): 825–836.

[35] 雷刚. 论算法行政的程序法治：挑战与回应 [J]. 电子政务, 2024, (06): 81-95.

[36] 刘银喜, 吴京阳. 生成式人工智能嵌入政府治理的应用前景、潜在风险和防范机制 [J]. 北京航空航天大学学报(社会科学版), 2025, 38(01): 103-112.

[37] Grimmelikhuijsen S. Explaining why the computer says no: Algorithmic transparency affects the perceived trustworthiness of automated decision-making [J]. Public Administration Review, 2023, 83(2): 241-262.

[38] 王鸿宇, 蓝江. 智能时代知识生产及其结构性风险——基于 DeepSeek 生成式人工智能技术的哲学反思 [J]. 理论与改革, 2025, (03): 51-61.

[39] 吉林省国家保密局. 大模型政务应用安全风险及应对案例 [EB/OL]. (2025-10-10)[2026-01-05]. https://bmj.jl.gov.cn/xcjy/202509/t20250908_3496567.html

[40] 刘银喜, 吴京阳. 生成式人工智能嵌入政府治理的应用前景、潜在风险和防范机制 [J]. 北京航空航天大学学报(社会科学版), 2025, 38(01): 103-112.

[41] 韩啸, 马琪. 虚拟数字人嵌入基层治理的实践探索：类型、风险与对策 [J]. 协商治理研究, 2025, (03): 137-158.

[42] 吉林省国家保密局. 大模型政务应用安全风险及应对案例 [EB/OL]. (2025-10-10)[2026-01-05]. https://bmj.jl.gov.cn/xcjy/202509/t20250908_3496567.html

[43] 韩啸, 王婷婷. “支持”还是“威胁”——政务数字人对公务员公共服务动机的影响机制 [J/OL]. 软科学, 1-12 [2026-03-16]. https://link.cnki.net/urlid/51.1268.G3.20251201.0937.002.

[44] 郑磊, 陈奕醇. 技术崇拜、制度约束与故事建构——数字治理中的悖论表现与成因探讨 [J]. 探索与争鸣, 2025, (06): 105-113.

[45] 高翔. 决策权集中与行政科层化：数字时代的政府组织变革 [J]. 广西师范大学学报(哲学社会科学版), 2023, 59(01): 8-19.

[46] 郑崇明. 警惕公共治理中算法影子官僚的风险 [J]. 探索与争鸣, 2021, (01): 103-109+179.

[47] 孟天广, 吴培琳. 算法政治：算法介入国家治理的耦合逻辑与政治效应 [J]. 山东大学学报(哲学社会科学版), 2024, (05): 134-147.

更多免费行业报告 并购家 www.ipoipo.net

[48] 梅立润. 技术置换权力：人工智能时代的国家治理权力结构变化 [J]. 武汉大学学报(哲学社会科学版), 2023, 76(01): 44-54.

[49] 郑磊, 陈奕醇. 技术崇拜、制度约束与故事建构——数字治理中的悖论表现与成因探讨 [J]. 探索与争鸣, 2025, (06): 105-113.

[50] 雷刚. 论算法行政的程序法治：挑战与回应 [J]. 电子政务, 2024, (06): 81-95.

[51] Rulandari N, Silalahi A D K. Resolving the Effectiveness‐Efficiency Paradox in AI‐Enabled Public Services: Insights From the Resource‐Based View and Public Value Theory[J]. Journal of Public Affairs, 2025, 25(3): e70067.

[52] 韩宏伟, 付旭卓. 数字官僚主义的现实表象、生成逻辑及破除路径——基于法治政府建设的视角 [J]. 黑龙江省政法管理干部学院学报, 2025, (05): 127-133.

[53] 王正超. 数智时代政务大模型数据安全风险及治理机制研究 [J]. 中国科技论坛, 2025, (11): 151-160.

[54] 张凌寒. 加快建设人工智能大模型中文训练数据语料库 [J]. 人民论坛•学术前沿, 2024, (13): 57-71.

[55] Yang, X., Mei, H., & Zheng, Y. Understanding the antecedents of privacy fatigue in facial recognition-based m-Gov services: An empirical study from China [J]. Government Information Quarterly, 2023, 40(3): 1-12.

[56] 孙妍, 张成福. 敏捷致智：敏捷治理引领政务服务智能化转型的内在机理研究——以济南市智慧审批实践为例 [J]. 甘肃行政学院学报, 2023, (05): 68-81.

[57] 张凌寒. 加快建设人工智能大模型中文训练数据语料库 [J]. 人民论坛•学术前沿, 2024, (13): 57-71.

[58] 丁伟峰. 政务数据安全风险与法律规制 [J]. 人民论坛•学术前沿, 2022, (09): 106-108.

9.2 人工智能嵌入公共治理的两种模式[①]

在当代公共治理的数字化转型进程中，人工智能技术的嵌入正在深刻重塑公共部门的运作方式。然而，这种技术嵌入呈现出截然不同的两种应用图景：一方面，在全国各地普遍应用的交通违章智能抓拍系统中，从违章行为识别、证据固定到处罚决定书生成、罚款通知推送的全过程已实现完全自动化，人类管理者仅在系统故障或当事人申诉时才会介入；另一方面，以深圳市为例，其在 2024 年 7 月上线的人工智能辅助审判系统，覆盖立案、阅卷、庭审、文书制作等 85 项审判流程，但该系统在设计之初便明确“锚定辅助工具的功能定位”，充分尊重裁判者的自主决策权，系统在各环节均设置审核、确认、决定选项，AI 根据法官决定的判决结果辅助生成裁判文书，“确保技术进步不越俎代庖，让司法裁判始终由审判人员作出”[1]。

这两种截然不同的应用模式背后，隐含着人工智能在公共治理中角色定位以及人机关系的分野。一是代理型模式，其中算法成为“行动者”，决策权实质性地转移至技术系统，人类“出场”而退居为监督者甚至旁观者[2-3]；二是辅助型模式，其中算法是“工具”，决策权牢牢掌握在人类手中，人工智能仅仅拓展了人类的信息处理能力和认知边界[4-5]。前者可能代表着更“激进”和“理想”的治理变革，而后者则蕴含着更为“克制”和“妥协”的考虑。这种差异并非仅仅是技术应用深度的量变，而是涉及权力配置、责任归属、价值权衡等方面

[①] 本节由复旦大学国际关系与公共事务学院博士生张宏撰写。
张宏，复旦大学数字与移动治理实验室研究助理，主要研究领域为数字政府、公共数据开放、智能治理。

更多免费行业报告 并购家 www.ipoipo.net

的质变。

1. 代理型模式：算法成为“行动者”

代理型模式的本质，在于决策权的实质性转移。在这一模式中，人工智能系统不再满足于提供信息、建议或辅助判断，而是直接完成从信息输入到决策输出的完整闭环。算法依据预设的规则或通过机器学习获得的模式，自主地识别情况、判断性质、确定处置方式，并直接生成具有效力的判断或决定[6]。人类管理者在这一过程中的角色也发生了根本性转变：他们不再是每一个具体决策的作出者，而是系统的监督者以及在例外情况下的干预者。

这种权力配置的重构,建立在特定的技术-制度架构之上。代理型模式得以运行的前提，是大量公共治理事项本身具有较高的规则明确性和标准化特征[7]。例如，交通违章之所以能够实现完全自动化处理，正是因为违章行为的构成要件、处罚标准、执行程序都有明确的法律规定，这些规则可以被完整地编码为算法逻辑。在这种情况下,算法执行的是一种“机械适用法律”的过程,其决策空间被严格限定在规则框架之内。理论上，面对相同的输入，无论是算法还是人类,都应当得出相同的处理结果，而正是这种决策的确定性和可预测性，为算法代理决策提供了逻辑基础。

然而，代理型模式的吸引力远不止于简单地“替代人工”。在公共治理的实践逻辑中，这种模式至少在三个维度上展现出独特的优势。首先是效率革命[8]。面对海量的、重复性的治理事务，人类处理能力

的局限性成为瓶颈。以税务稽查为例，在全国亿万纳税人的申报数据中识别异常，依靠传统的人工抽查不仅效率低下，而且覆盖面有限。而智能风险识别系统可以实时处理全量数据，依据多维度指标模型快速锁定高风险对象，这种处理能力是人力所无法企及的。其次是规则执行的刚性化[9]。算法不会疲劳、不会受情绪影响、不会因人际关系而通融，它确保了“法无例外”的执行力度。在一些寻租空间较大的领域，用算法代替人的自由裁量，恰恰被视为反腐倡廉的技术手段。最后是治理能力的时空延展。自动化系统可以 7×24 小时不间断运行，覆盖传统行政时空无法触及的领域，这在应急管理、公共安全等需要即时响应的场景中尤为重要。

正是基于这些优势，代理型模式在公共治理中的应用呈现出不断扩展的趋势。从最初的交通管理、税务稽查等执法领域，逐步延伸至社会保障资格审核、公共资源配置、甚至某些行政许可的初审环节。

然而，当算法真正成为“行动者”时，一系列深层的治理悖论随之浮现。首先是责任归属的困境。传统行政法理论建立在“人作决策、人担责任”的基础上，当决策由算法作出时，出现错误或不当处理，究竟应当由谁承担责任?是系统的设计者、开发者、部署者，还是名义上的主管部门?这种责任的模糊化，既可能导致无人问责的真空，也可能产生过度问责的寒蝉效应[10]。其次是算法黑箱与可解释性的矛盾。许多先进的 AI 系统，特别是基于深度学习的模型，其决策过程本身就是不透明的。即便是系统的开发者，也难以完整解释为什么算法在特定情况下会作出特定判断。这种不透明性在私人领域或许可以

更多免费行业报告 并购家 www.ipoipo.net

容忍，但在涉及公民权利义务的公共决策中，却与正当程序原则、知情权保障产生尖锐冲突[11]。当一个人不知道自己为何被列入风险名单、如何才能从名单中移除时，救济权利如何实现?

更为根本的问题在于，代理型模式可能改变权力运行的性质。算法嵌入的不仅是技术，更是价值观和利益取向[9]。在系统设计阶段，关于如何定义风险、如何权衡不同目标、如何设定阈值的决策，本质上都是政治性的权力判断，但这些判断被技术化、客观化地呈现，从而规避了民主审议和公共辩论。更重要的是，一旦算法系统投入运行，其规则便获得了某种“技术必然性”的外衣，修改和调整变得困难重重。这导致治理过程的僵化：面对边缘案例、特殊情况，系统可能无法像人类那样灵活应变、从容裁量。而人类管理者在长期依赖系统后，自身的专业判断能力和主体性也可能退化[12]，形成一种恶性循环。

代理型模式还可能引发更深层的政治哲学问题。公共治理的正当性不仅来自决策结果的正确性，也来自决策过程的民主性和可参与性[13]。当算法代理决策成为常态，公民与政府的互动关系发生了微妙变化：个体面对的不再是可以沟通、说理、协商的“人”，而是冰冷的、程序化的系统。这种非人格化的治理方式，在提高效率的同时，也可能削弱公共部门的回应性、同理心和人文关怀。治理变成了一种纯粹的技术操作，而遗忘了其背后的社会意义和价值追求。

2. 辅助型模式：算法作为“增强工具”

与代理型模式形成鲜明对照的，是将人工智能定位为辅助工具的

另一种更为克制的嵌入路径。在这一模式中，决策权的归属从未发生转移，算法的功能被严格限定在信息处理、方案生成、风险提示等支持性环节，而最终的判断、选择和决定，始终由人类管理者作出。这一模式的核心是将人工智能视为人类认知能力的延伸和增强，而非人类决策者的替代和取代[14]。

辅助型模式的技术-制度架构，则建立在“人机协同”的基础之上。这种协同不是简单的分工，而是一种深层的能力互补。算法擅长的，是在海量数据中识别模式、计算概率、模拟推演——这些恰恰是人类认知的短板。而人类擅长的，是处理模糊情境、理解复杂语境、进行价值权衡、作出伦理判断——这些正是当前算法的局限所在[15]。辅助型模式可将二者的优势有机结合：让算法承担“计算密集型”的工作，为人类决策提供坚实的信息基础；让人类承担“判断密集型”的工作，在算法提供的选项中进行最终决断。以智能辅助审判为例，系统可以在几秒钟内检索出数千个类似案件的裁判结果，识别出不同判决的法律依据和量刑标准，甚至预测不同判决可能引发的上诉风险，但它不会直接告诉法官“应当如何判”，而是将这些信息作为参考，由法官综合案件的特殊性、当事人的具体情况、社会效果等多重因素，作出最终裁决。

这种模式选择的背后，蕴含着对公共治理复杂性的认识。公共决策往往需要在多个合理方案之间进行权衡，需要平衡效率与公平、原则与灵活、当下与长远等多重价值。这种权衡本质上是政治性的、价值性的，而非纯粹技术性的。算法可以算出“如果选择 A 方案有多大

更多免费行业报告 并购家 www.ipoipo.net

概率会产生何种后果”，但它无法回答“应当选择哪个方案”，因为后者涉及的是价值排序和偏好判断，这是民主决策和专业判断的领域，而非算法计算的领域[16]。辅助型模式正是基于这一认识，坚持将价值判断和最终决策的权力保留在人类手中，而将算法限定在技术分析和方案生成的范围内。

从治理价值的角度看，辅助型模式在多个维度上具有独特的意义。首先是维护了治理的正当性基础。在民主法治框架下，公共权力的行使必须具有可追溯的责任主体[17]。辅助型模式通过明确“人作决策”的原则，确保了每一个决策都有具体的责任承担者，这不仅符合行政法治的基本要求，也维护了公民寻求救济的权利。当公民对决策不满时，他们可以质疑决策者的判断、要求其说明理由、通过申诉和诉讼寻求纠正，而这一切在代理型模式中都变得困难重重。其次是保持了治理的灵活性和回应性。公共治理面对的是复杂多变的社会现实，标准化规则往往无法穷尽所有情况。辅助型模式通过保留人类的自由裁量权，为应对特殊情况、处理边缘案例、平衡多元诉求保留了必要的弹性空间。这种灵活性不是任意性，而是在规则框架内根据具体情境作出合理调适的能力。最后是维护了专业判断的权威性和尊严性。无论是法官、医生还是政策制定者，其专业性的核心不在于机械地适用规则，而在于基于专业知识、经验积累和价值判断，作出负责任的决策[18-19]。辅助型模式尊重并强化了这种专业性，将算法定位为专业判断的辅助手段，而非取代专业判断的自动化工具。

当然，辅助型模式对决策者的能力提出了更高要求。在代理型模

式中，人类的角色是系统的维护者和监督者，需要的是技术管理能力；而在辅助型模式中，人类仍然是决策者，不仅需要保持原有的专业判断能力，还需要发展与算法协同工作的新型能力，包括理解算法逻辑、评估算法输出、整合算法建议与专业经验的能力。这种能力要求的提升，在公共部门数字化转型的背景下，构成了一个现实的挑战。如果决策者缺乏足够的数字素养和算法理解力，辅助型模式可能异化为“有名无实”的形式主义：名义上保留了人类决策权，实际上人类已无力作出独立判断[20]。

3. 分野背后：哪些因素影响了人机关系的模式适用性?

代理型与辅助型两种模式的分野，表面上看是技术应用程度的差异，实质上是基于对治理情境、风险评估、价值考量及社会接受度等的综合判断。

首先是治理事项本身的性质特征。影响模式选择的一个关键维度，是该事项的规则明确性和决策复杂性之间的平衡。交通违章处理之所以可以采用代理型模式，根本原因在于其决策逻辑高度规则化。在这种情况下，算法不是在“创造”决策，而是在“执行”既定规则[21]。相反，司法裁判即便在事实清楚、法律明确的案件中，仍然涉及大量的情境判断、证据权衡、量刑裁量，这些判断无法完全规则化、程序化。正是这种决策复杂性，决定了这种情境下算法更应该是辅助工具而非决策主体。然而，规则明确性并非静态的、先验的，而往往是在实践中被建构出来的。一个值得注意的现象是，为了实现算法代理，治理事项本身可能被“标准化改造”：原本存在的自由裁量空间被压

更多免费行业报告 并购家 www.ipoipo.net

缩，复杂的情境被简化为若干离散的类型，连续的判断被转化为二元的选择。这种改造在带来效率提升的同时，也可能导致“数字的庸俗化”以及对人性的压抑[22]。

第二个关键因素是该决策对个人权益的影响程度。通常认为，对人身自由、重大财产权利等个人权益有重大影响的决策，需要更高程度的正当程序保障[23]。交通罚款虽然涉及财产权，但数额通常较小，且有明确的法律依据，因此可以容忍较高程度的自动化。但在涉及剥夺自由、取消重大资格等后果严重的决策中，对程序正当性的要求就显著提高，辅助型甚至“无 AI 介入”应当是更审慎的选择。这不仅是法律规范的要求，更是基于一个朴素的正义直觉：当决策可能严重影响一个人的生活时，他至少应当有权面对一个能够倾听、理解、回应的人类决策者，而不是一个冰冷的算法系统。这种对“人的在场”的要求，还与决策的可辩驳性和可救济性密切相关。代理型模式中，决策过程高度自动化，留给当事人参与、辩论、说理的空间极为有限[6]。而在影响重大权益的场景中，当事人有权陈述事实、提出辩解、阐明特殊情况，这种对话和说理的过程本身就是程序正义的重要组成部分。辅助型模式通过保留人类决策者，为这种对话创造了可能性——即便最终决策不利于当事人，但他至少经历了一个“被听见”的过程，这对于决策正当性的感知至关重要。

第三个维度是决策中价值冲突的复杂程度。许多公共治理决策并非在“对错”之间选择，而是在多个正当价值之间权衡。城市规划需要平衡发展与保护、效率与公平；政策制定需要权衡不同群体的利益

诉求；执法需要在法律的刚性与现实的复杂性之间寻找平衡。这种价值权衡是实质性的、政治性的，难以被化约为算法可处理的优化问题。即便算法能够模拟不同方案的后果，但“哪种后果更可欲”的判断，仍然是人类的责任。在价值冲突显著的领域，辅助型模式更为合适：算法可以揭示不同选择的利弊得失，但最终的价值排序和选择，应当由对政治共同体负责的人类决策者作出。更深层的问题在于，有些价值本身就是“面向人”的，它们的实现离不开人的参与和互动[18]。同理心、尊重、尊严感，这些价值无法通过算法运算来达成，甚至往往会被边缘化[24]。在医疗、教育、社会救助等领域，治理不仅是提供服务，更是传递关怀和尊重。在这些场景中，算法可以提供技术支持，但不能取代人的在场，因为人的在场本身就是服务的一部分，是价值实现的必要条件。

第四个因素是社会的技术接受度和信任水平。技术的嵌入不仅是技术问题，更是社会问题。即便从技术和效率角度看代理型模式更优，但如果公众对此缺乏信任、产生抗拒，这种模式就难以持续。公众对算法决策的接受度，受到多重因素影响[25-26]：对算法公正性的信心、对错误后果的担忧、对“非人化”治理的情感抗拒等。在一些领域，公众已经习惯并接受了算法的自动化决策，比如信号灯控制、违章抓拍；但在另一些领域，公众仍然坚持“要见到人”，比如医疗诊断、司法裁判。这种社会心理的差异，构成了模式选择的重要约束。这意味着，模式选择不能仅仅基于技术可行性，还必须考虑社会准备度：在算法透明度、问责机制、救济渠道等尚未完善之前，辅助型模式是一

更多免费行业报告 并购家 www.ipoipo.net

种更为审慎的过渡方案。

4. 趋势与反思:人机边界的动态调整

代理型与辅助型两种模式的划分，在理论上看似清晰，但在实践中却呈现出高度的流动性和模糊性。人机关系的边界并非一成不变，而是在技术演进、组织学习和社会适应的过程中不断调整、重构。

最值得警惕的趋势，是从辅助到代理的“渐进式转移”[27]。这种转移往往不是突然发生的，而是通过一系列微小的、看似合理的步骤逐步实现的。一个系统最初被引入时，明确定位为辅助工具，人类决策者认真审核每一个算法建议，系统的准确率也在持续监测之中。然而，随着系统运行时间的延长，一种微妙的变化开始发生：当算法建议的准确率持续保持在高位，人类审核的频率开始下降；当工作负荷增加而人力资源紧张，人工审核被逐步视为低效的冗余环节；当组织绩效考核强调处理效率而非决策质量，快速确认算法建议成为理性选择。这种现象深刻体现了心理学上的“自动化偏差”（Automation Bias）。根据 Skitka 等人的经典定义，这是指决策者将自动化系统视为一种启发式替代工具，从而减少警觉性监控和认知努力。这种偏差导致决策者在系统沉默时不仅忽略风险（遗漏错误），更在系统出错时盲目遵从其建议（执行错误），最终导致实质性的决策权让渡[28]。正是受此心理机制驱动，辅助型模式在实践中悄然滑向代理型模式，而这种转变甚至没有经过正式的决策和公开的讨论。

这种“功能蔓延”的背后，首先是技术系统自身的扩张逻辑。算

法的运行会产生数据，数据的积累会提升算法的性能，性能的提升会增强人们对系统的依赖，依赖的加深又会产生更多数据，形成一个自我强化的循环[29]。在这个过程中，系统获得的不仅是技术能力的提升，更是权力的扩张：它从提供建议的参谋，逐渐变成事实上的决策者。而这种权力转移往往是隐性的、不自觉的，直到某个错误或争议事件发生，人们才突然意识到“什么时候算法已经接管了决策？”

更复杂的情况是，两种模式的边界在实践中可能变得高度模糊。一个系统在设计上保留了人类的决策权，设置了审核、确认、修改的环节，形式上是辅助型模式。但如果人类审核流于形式，系统的建议几乎总是被直接采纳，那么这个系统的实质功能已经是代理型的。这种“名义上的辅助、实质上的代理”，会使得责任归属变得极其模糊[30-31]：出现问题时，技术部门会说“我们只是提供建议，决策是业务部门做的”，业务部门会说“我们只是按系统操作，错误在算法”。这种责任的相互推诿，正是功能蔓延带来的治理困境。

那么，如何防止这种不自觉的滑动？关键在于建立明确的边界意识和持续的反思机制。首先，在系统设计和部署时，就应当明确其功能定位[32-33]，并将这种定位制度化、规范化。如果确定为辅助型模式，就应当在系统中嵌入机制确保人类决策权的实质行使：不仅要有“确认按钮”，更要有促使决策者进行独立思考的设计，比如要求对偏离算法建议的决策进行说明，对算法建议的接受率进行监测，定期评估人工审核的质量。其次，需要建立技术系统的定期评估和动态调整机制。算法系统不应当是“一次部署、永久运行”的，而应当接受持续

更多免费行业报告 并购家 www.ipoipo.net

的监督和审查[34]。这种审查不仅关注技术性能指标（如准确率、召回率、误报率），更要关注功能边界的变化：系统实际承担的角色是否与初始设计一致？人机协同的模式是否仍然有效？这种评估应当成为制度化的要求，而不是依赖偶然的危机事件来触发。更根本的是，需要培育一种对技术系统保持批判性距离的组织文化。在数字化转型的热潮中，技术往往被赋予某种“进步”的光环[35]，对技术的质疑容易被贴上“保守”或“抵制变革”的标签。但真正的数字化成熟，恰恰需要一种“批判性数字素养”：既认识到技术的价值，也清醒地意识到技术的局限；既善用技术工具，也警惕技术对人的替代和对权力的重塑[36]。这要求公共部门的管理者和工作人员，不仅要学会使用算法系统，更要学会质疑算法系统：质疑它的假设、质疑它的逻辑、质疑它的建议。

从更宏观的视角看，人机边界的调整应当是一个民主治理的过程，而非纯粹的技术决策。哪些公共决策可以交给算法，哪些必须由人类作出，这不是技术专家可以单方面决定的，而应当经过广泛的社会讨论和民主审议。公民有权参与这个过程，表达他们对算法治理的期待和担忧，对人机关系边界的价值判断。遗憾的是，在当前的实践中，这种公共讨论往往是缺失的：算法系统的部署被视为内部管理决策，在行政系统内部完成，直到系统上线运行，公众才被动地接受既成事实[34]。这种决策过程的封闭性，不仅削弱了算法治理的民主正当性，也错失了汲取多元视角、预防潜在风险的机会。

最后，需要重新理解“人机协同”的真正含义。这不应当仅仅被

理解为“人负责复杂的，机器负责简单的”或“人负责决策，机器负责执行”，而应当是一种更深层的互动和互构关系。算法改变了人类决策的条件和方式，而人类的使用和反馈也在持续塑造算法的演进。在这种协同中，关键不是划定一条静态的边界，而是建立一种动态的平衡机制：既充分发挥算法在信息处理、模式识别、复杂计算方面的优势，又切实保障人类在价值判断、伦理权衡、责任承担方面的核心地位；既利用技术扩展治理能力，又防止技术异化治理本身。这种平衡不会自动实现，它需要审慎的制度设计、持续的反思评估、广泛的民主参与。它要求在拥抱技术创新的同时，保持对技术权力的警觉；在追求治理效能的同时，坚守治理的价值底线。

人工智能嵌入公共治理，不是一个已经完成的事实，而是一个正在展开的过程。在这个过程中，代理与辅助的分野、人机边界的划定，都不是预先确定的，而是需要在实践中不断探索、辩论和选择的。这种选择，既是技术问题，更是政治问题、伦理问题、关乎想要建设一个什么样的公共世界的问题。

参考文献

[1] 广东省深圳市中级人民法院. 深圳中院今日上线人工智能辅助审判系统[EB/OL]. (2024-07-01)[2025-12-07]. https://ipc.court.gov.cn/zh-cn/news/view-3143.html

[2] 李牧,王俊.数字政府视阈下的算法行政：兴起、风险与法治纠偏[J].武汉理工大学学报(社会科学版),2024,37(02):1-8.

[3] 王张华,颜佳华.人工智能时代算法行政的公共性审视——基于"人机关系"的视野[J].探索,2021,(04):82-95.

[4] 王张华,颜佳华.人工智能驱动政府治理变革:技术逻辑、价值准则和目标指向[J].天津行政学院学报,2020,22(06):3-11.

更多免费行业报告 并购家 www.ipoipo.net

[5] 吴进进,符阳.算法决策：人工智能驱动的公共决策及其风险[J].开放时代,2021,(05):194-206+10.

[6] 张敏.算法治理：21 世纪的公共管理现代化与范式变革[J].政治学研究,2022,(04):50-62+157-158.

[7] 贾开.人工智能与算法治理研究[J].中国行政管理,2019,(01):17-22.

[8] 王贵.算法行政的兴起、挑战及法治化调适[J].电子政务,2021,(07):2-14.

[9] 周晓丽,姬晓暄.算法嵌入政府治理：逻辑、风险与规制[J].西安交通大学学报(社会科学版),2024,44(01):52-61.

[10] 王秋艳,万国华.生成式人工智能的双重法律规制理路[J].中国特色社会主义研究,2023,(04):72-84.

[11] 衣俊霖.数字孪生时代的法律与问责——通过技术标准透视算法黑箱[J].东方法学,2021,(04):77-92.

[12] 何哲.人工智能时代的社会转型与行政伦理:机器能否管理人?[J].电子政务,2017,(11):2-10.

[13] 李瓔珞.全过程人民民主论析：结构、流程与功能[J].云南社会科学,2022,(04):10-17.

[14] Jarrahi M H. Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making[J]. Business horizons, 2018, 61(4): 577-586.

[15] 杜严勇.厌恶算法还是欣赏算法？——人工智能时代的算法认知差异与算法信任建构[J].哲学分析,2022,13(03):151-165+199.

[16] 陈锐,王文玉.算法嵌入政府治理的权力异化风险及其规制方案[J].理论探索,2022,(06):81-90.

[17] 彭中礼,王亮.法治视野中的智慧治理[J].济南大学学报(社会科学版),2020,30(06):16-26+157.

[18] Brady F N. “Publics” administration and the ethics of particularity[J]. Public Administration Review, 2003, 63(5): 525-534.

[19] 周雪光.权威体制与有效治理:当代中国国家治理的制度逻辑.开放时代,2011,(10):67-85.

[20] 杨建.认真对待个人信息保护的社会结构:AI 科技时代的不对称性挑战.环球法律评论,2022,44(02):18-35.

[21] 刘永谋,李尉博.从“大设计”到“小设计”：大数据时代的社会规则之变[J].哲学分析,2022,13(01):123-137+199.

[22] 高奇琦.国家数字能力：数字革命中的国家治理能力建设[J].中国社会科学,2023,(01):44-61+205.

[23] 杨媛.算法嵌入政府治理的风险及应对[J].中国电信业,2022,(04):53-57.

[24] 黄雅卓.算法技术赋能政府职责体系构建及其限度[J].天津行政学院学报,2023,25(03):34-42.

[25] 景怡.算法态度文献综述、研究框架和未来展望[J].华东经济管理,2023,37(05):107-118.

[26] Oliver K. Reactions on Algorithms: a systematic Literature Review of Algorithm Aversion and Algorithm Appreciation[D]. The University of Innsbruck, 2020.

[27] 翟月荧.算法行政的兴起、风险及其防控[J].新视野,2022,(03):81-85.

[28] Skitka L J, Mosier K L, Burdick M. Does automation bias decision-making?[J]. International Journal of Human-Computer Studies, 1999, 51(5): 991-1006.

[29] 于文轩,魏炜.数据开放中的算法依赖：发展模式与驱动路径[J].理论与改革,2023,(02):109-122.

[30] 梅立润.人工智能时代国家治理的算法依赖及其衍生难题[J].中南大学学报(社会科学版),2022,28(06):123-131.

[31] 谭九生,杨建武.人工智能嵌入政府治理的伦理风险及其防控[J].探索,2021,(02):126-138.

[32] 刘星.行政裁量中的技术控制——基于政务服务应用场景的实践观察[J].公共行政评论,2022,15(01):51-68+197.

[33] 查云飞.行政裁量自动化的学理基础与功能定位[J].行政法学研究,2021,(03):114-124.

[34] 张凌寒.算法自动化决策与行政正当程序制度的冲突与调和[J].东方法学,2020,(06):4-17.

[35] 复旦大学数字与移动治理实验室. 2024 中国城市数字治理报告: 中国城市数字治理的盼与忧——走向数字善治[R]. 2024.

[36] 王乐.“余数生命”:基于疫情健康码的伦理反思.昆明理工大学学报(社会科学版),2021,21(01):27-32.

更多免费行业报告 并购家 www.ipoipo.net

9.3 人工智能在公共治理中的“负责任”使用[①]

承接前文对辅助型与代理型人机关系的分析，人工智能在公共治理中的位置已明显前移。它不再只是后台工具，而是日益深入地介入咨询、引导、识别、研判乃至部分决策环节。角色越靠前，技术就越深地卷入权利实现、程序运行和责任分配之中。由此，讨论的重心也必须从能不能用、好不好用，转向谁来负责、如何问责，以及怎样在制度上确保人工智能被负责任地使用。负责任的人工智能治理，关心的正是当人工智能进入公共部门后，责任如何界定、问责如何落实、公共价值如何嵌入技术系统与治理流程。

1. 责任基础：责任、问责与人工智能时代的责任重构

（1）责任的规范基础与机制：权力、裁量与后果

在公共治理中，责任意味着权力、裁量与后果之间的规范性联系。谁掌握公共权力，谁拥有资源配置权和程序裁量权，谁就应当对其行为后果承担相应义务。现代公共管理并不排斥自主性与裁量权，但正因为存在裁量，责任才成为公共权力不可分离的内在约束。责任并非只在失误之后才出现，它从权力被赋予之时就已存在，既要求行为者依法履职，也要求其对后果作出说明并接受评价[1]。

如果说责任回答谁应当负责，那么问责处理的则是责任如何被看见、被追问、被落实。在公共管理文献中，问责既是一种规范性德性，

[①] 本节由复旦大学国际关系与公共事务学院博士生杨家西撰写。
杨家西，复旦大学数字与移动治理实验室研究助理，主要研究领域为数字治理与问责制、政民互动、公共组织行为与绩效。

也是一种制度性机制，即行为者必须面对某个问责场域，对其行为作出解释并接受评判与后果[2]。问责具有外部性、互动性和后果性三个特征：说明必须面向外部主体，包含解释、追问、回应和评价的过程，并可能带来纠偏、修复乃至制裁[3]。置于行政语境中，算法行政问责的核心不在于技术是否先进，而在于政府部门能否就人工智能系统的目的、依据、过程、结果和后续影响给出合理说明，并接受来自各方的质询与评价，最终承担相应后果[4]。

（2）公共价值导向与算法责任的特殊性：政府服务于公共利益

从公共行政的规范立场看，负责任的人工智能治理首先关乎政府如何理解自身与公民的关系。新公共服务理论强调，政府服务于公共利益，而非控制社会[5]。技术应用不能偏离这一根本定位，无论系统多么高效，若它削弱了公众理解政策、表达异议和获得救济的能力，就难以被称为负责任。

人工智能中的责任不能被简化为技术故障后的赔偿，也不能被理解为算法替人负责。人工智能本身并非完整责任主体，但它承载着设计者的价值预设、开发者的建模逻辑、部署者的政策目标和使用者的操作方式，其输出会对资格认定、风险评估、资源分配产生真实影响。算法并不价值中立，它通过代码设定、数据选择、目标函数参与到决策中，可能影响个体的知情权、隐私、自由和平等保护[6]。算法责任之所以成为独立议题，既与算法承担代理性角色有关，也与其社会嵌入性、价值迁移性和结果外溢性有关，是一种综合性、代理性和隐蔽性的新型责任形态[7]。

更多免费行业报告 并购家 www.ipoipo.net

（3）责任重构方向：从分布式责任到全生命周期治理

人工智能时代的责任需要重新界定，更深层原因在于责任链条已被技术系统和组织协作重新切分。引入人工智能往往涉及业务部门、数据提供方、模型开发方、平台运维方、外部供应商和一线使用者。责任不再集中附着于单一岗位，而呈现分布式、阶段化和网络化结构。但责任的分布并不意味着稀释：越是多主体协作，越需要把责任拆解清楚，谁对数据质量负责，谁对模型逻辑负责，谁对部署审批负责，谁对具体使用负责，谁对异常纠正和公众救济负责。若仅承认责任分布而不进行责任分配，分布式责任就会滑向无人负责的真空[8]。

正因如此，人工智能时代的问责应被理解为全生命周期的责任治理，包括系统上线前的责任识别和影响评估，运行中的透明说明、监测纠偏和人工干预，运行后的复核、申诉、救济和责任追究。问责不再是事后追责，而是事前防范、事中校正和事后修复的结合。只有让责任落实嵌入可持续的制度与程序，责任政府才可能形成制度基础[9]。

基于此，负责任的人工智能治理可界定为：公共部门在设计、采购、部署、使用和更新人工智能系统的全过程中，以公共价值为导向，通过明确责任主体、强化说明义务、完善质询程序和配置救济与后果机制，确保人工智能在提升治理效能的同时，不损害公平、透明、隐私、安全与公民权利。公共价值不能停留于原则表述，而需落实为对技术应用边界的明确要求。从公共价值创造视角看，算法治理不仅追求绩效价值即效率、效果，也必须创造程序价值即透明、平等、可问责、可救济[10]。真正的负责任治理，必须使人工智能不仅可用，而且

可被公众理解、质疑和接受。

2. 风险识别：人工智能时代的问责困境与责任挑战

（1）可解释性挑战：算法黑箱抬高解释门槛

人工智能时代的问责之难，首先体现在解释门槛的显著抬高。传统公共服务中，责任主体可见，决定理由可追问。人工智能介入后，许多识别、筛选、预测被技术系统吸收，公众面对的是运行逻辑难以直接理解的系统。黑箱性、模型复杂性和技术语言，使关键环节不再容易被看清。这种不透明显著降低问责的可操作性[11]。可解释人工智能研究表明，深度学习等高性能模型往往以可解释性下降为代价，其内部参数庞大、因果链条不易分辨[12]。对政府部门而言，可解释性不是附加功能，而是问责可运行的必要条件。

（2）责任碎片化：多主体协作与算法卸责

解释门槛的抬高加剧了责任链条的碎片化。自动化行政由采购、建模、测试、部署、运维和使用共同构成。技术协作越复杂，责任越易被拆散和策略性转移。自动化行政中的算法卸责表现为：以算法如此决定代替组织解释义务；以按系统规则操作回避个案裁量；借助供应商和技术话语转移风险，形成人人都参与、人人都不完整负责的格局[13]。问责并未消失，但可能被形式化，真正承担解释和纠偏责任的主体越来越模糊。既有研究区分了利用算法进行治理与治理算法本身两条路径[14]，提示公共部门应根据事项对公民权利的影响程度，采取差异化的治理强度和问责要求。

更多免费行业报告 并购家 www.ipoipo.net

（3）价值冲突：效率逻辑与公共价值之间的张力

责任碎片化背后，更深层的张力在于效率逻辑与公共价值的冲突。公共部门引入人工智能最易见的收益是提速、降本、增效，但行政正当性不仅来自速度和准确率，也来自程序正当、机会平等、权利保障和合理救济。数字治理研究表明，仅靠技术刚性嵌入而缺少理念、制度、组织、法治和伦理的同步调整，技术赋能容易偏离预期目标[15]。技术治理失灵的一系列症候，包括技术更新与体制改革脱节、技术嵌入与隐私保护失衡、治理效率与社会公平冲突，说明技术理性并不自动等于公共理性[16]。若组织过度依赖使用率、响应速度等指标，公共服务的价值重心就可能从是否合乎公共利益转向是否符合系统逻辑，导致量化与简化、整体与自主、集成与透明、标准与精细、效率与感受等多重悖论[17]。

这种张力还传导至信任层面。问责有效不仅因为制度存在，更因为公众相信其正当性。但公众对人与人工智能的信任并不对称。调查显示，公民整体上更信任人类工作者，尤其在生命、安全等庄严价值领域差异更明显[18]。这意味着人工智能的可用性不等于公共可接受性。若在高风险事项中不加区分地推进自动化，形式上的解释和程序设计未必足以支撑治理信任，反而可能让公众产生机器替代责任的直观感受。

（4）工具弱化：低强度问责与合规表演

问题的复杂性还在于，现有问责机制本身往往处于低强度运行。算法行政问责已形成影响评估、透明度机制、审计监管、听证复核等

多种工具，但实际效果不稳定。从欧美政策实施经验看，信息提供、质询强度和有效激励仍是最薄弱环节[19]。信息不等于可理解，披露不等于可追问，程序存在不等于真正问责。若说明文档过于技术化、公众参与流于形式、责任后果过轻，问责就容易停留在合规表演层面。责任型人工智能治理研究也显示，国家政策组合与公共责任取向并不必然同向变化，责任是穿透不同治理模式的横向议题[20]。

综合来看，人工智能时代的责任挑战包括：系统黑箱提高了解释难度；多主体协作分散了责任链条；指标导向冲击了公共价值；公众信任不会自动提升；问责工具存在信息不足、质询偏弱、后果不强等问题。这些困境无法靠进一步提高技术性能自动解决，而需要在制度上重构责任、在程序上强化问责、在价值上校准技术应用边界。

3. 治理路径：制度安排、治理重点与动态调适

（1）重建全生命周期责任链

真正的负责任治理必须从责任链重建开始。责任链重建的关键，在于把责任分解到人工智能全生命周期的不同环节：谁提出应用目标并界定业务场景，谁负责数据治理与数据质量，谁负责模型采购、开发与测试，谁负责部署审批和风险评估，谁负责一线使用与人工复核，谁负责更新维护、异常响应和纠错修复，谁负责受理申诉、提供解释并作出救济决定。责任链条明确，问责才有抓手。责任可以分层、分工、分阶段配置，但公共责任不能外包。供应商可承担合同义务和专业责任，但不能替代公共组织的解释责任和救济责任[21]。

更多免费行业报告 并购家 www.ipoipo.net

（2）构建“说明—质询—后果”的问责闭环

首先强化说明义务。公共部门应在系统设计、采购、部署和运行阶段形成持续的说明文档制度，说明系统做什么、如何做、为何用、适用场景、数据依赖、限制、人工干预及申诉方式。高风险事项应进行算法影响评估，提前识别对弱势群体的影响。说明是问责的基础设施[22-23]，且不应是一次性告知，而需贯穿全过程，使公共价值通过持续调适真正嵌入系统[24]。

其次提升质询强度。说明需配套可被追问、被异议、被修正的程序结构。高影响事项必须保留人工复核与人在回路机制，关键是这一人类角色是否真正具有介入权、纠偏权和解释权。若一线人员只能机械确认系统结论，人工复核就是责任表演。需同步建立听证、上诉、专业审计、公众参与等机制，使不同主体能形成真实追问。解释的形式与质量影响公众对准确性、公平性和信任的判断，是可问责的重要组成部分[25]。

再次强化后果机制。问责缺乏实际后果就会退化为说明过了就算负责。负责任治理需让信息、质询与责任后果形成闭环：对内纳入绩效考核、审计监督、异常复盘和责任追究；对外将可解释义务、数据治理义务、缺陷修复义务等写入采购合同。对高风险事项的错误，不仅要修正系统输出，更要追溯组织决策和管理疏漏。只有不负责任的使用会带来真实后果，责任机制才不会停留在口头。

（3）嵌入公共价值：效度、温度与尺度的统一

说明、质询、后果的闭环还需要公共价值的嵌入。负责任治理应

在制度设计上坚持效度、温度与尺度的统一：效度指人工智能是否真正改善治理而非制造数字形式主义；温度指技术能否照顾公众体验和情境不确定性；尺度指技术使用必须有边界，不能因技术上可行就默认治理上应当如此[26]。负责任治理的本质，是让价值判断约束技术部署。

（4）适应性治理：敏捷调整与动态复盘

负责任治理不能停留于一次性制度设计，而需转向更具适应性的范式。面对技术、场景和风险的持续变化，治理重点应包括对技术—社会复合体的动态识别、对规制节奏的及时校准。我国人工智能治理已从探索式治理演进到敏捷式治理，有效性越来越依赖适应性、敏捷性与动态协调能力[27]。生成式人工智能作为由人、数据、流程、制度、组织和技术共同构成的社会技术系统，需要跨主体的联合责任与组织整体视角[28]。

上述努力需落实到全生命周期动态治理：人工智能系统并非一次部署即稳定，模型会更新，数据会变化，场景会迁移。负责任治理应是持续监测、持续校正、持续复盘的过程，包括日常监测、版本记录、异常报告、周期性评估、公众反馈吸纳和重大变更再审查。治理的关键不在于承诺永不出错，而在于建立能不断发现问题、吸收问题并修复问题的制度韧性。

（5）防范不稳定因素：技术崇拜、制度失配与平台依赖

同时需警惕治理本身可能孕育的不稳定因素。技术崇拜可能让治理目标偏移：若人工智能被过度叙事为治理神器，地方实践易陷入炫

更多免费行业报告 并购家 www.ipoipo.net

技治理和数治神话，用形式智能化替代实质改进[29]。制度能力与技术能力不匹配会使责任链松动：数据质量不足、专业人才短缺、数字鸿沟扩大、解释问责模糊等问题已构成真实风险[30]。透明并不天然等于信任：若说明文档过于技术化、解释无法被理解，透明可能沦为合规文本生产。问责过强或过弱都有副作用：不足纵容卸责，过强导致求稳回避创新。平台权力和外部技术依赖也可能改变责任结构，尤其在大型平台深度参与下，公共价值与商业逻辑的矛盾可能被放大[31]。

正是基于对这些不稳定因素的认识，才能真正理解负责任治理的复杂性。它既不能放任技术扩张，也不能把责任机制设计成僵化的惩罚机器，而需要在约束与创新之间形成动态平衡。负责任的人工智能治理并非在技术发展之外附加伦理包装，而是要求公共部门重新组织权力、程序和责任。只有当人工智能系统始终处于可解释、可质询、可纠偏、可追责的框架中，它才可能真正成为提升公共治理能力的工具，而非稀释公共责任的借口。

4. 结语

负责任的人工智能治理，归根结底讨论的是技术深度嵌入政府行为后，公共责任如何继续被看见、被追问并被兑现。责任并未消失，只是从单一岗位职责转化为跨主体、跨环节、跨阶段的制度分配问题。问责也不再只是事后追责，而是贯穿设计、部署、运行和修复全过程的治理安排。

本章论述指向三个相互关联的判断。第一，责任是公共权力的内

在属性，人工智能不会削弱这一原则，只会迫使公共部门更清晰地重构责任链。第二，人工智能时代的问责，本质上是由说明、质询与后果共同构成的制度机制，其重点不在于技术本身，而在于技术是否被置于可被追问和可被纠偏的治理框架中。第三，负责任治理不仅要应对黑箱、偏差、碎片化和卸责等直接挑战，还要防范技术崇拜、制度失配、形式透明、能力短板和激励失衡等不稳定因素。只有把公共价值真正嵌入系统设计、程序运行和责任配置之中，人工智能才可能在提高治理效能的同时，维护公共部门应有的公平、透明、温度与正当性。

更多免费行业报告 并购家 www.ipoipo.net

参考文献

[1] 李韬, 王佳. 自主性、裁量权与公共管理者的责任 [J]. 中国行政管理, 2010, (10): 57-60.

[2] Bovens M. Two concepts of accountability: accountability as a virtue and as a mechanism [J]. West European Politics, 2010, 33(5): 946-967.

[3] Mulgan R. Accountability: an ever-expanding concept? [J]. Public Administration, 2000, 78(3): 555-573.

[4] 张惠彬, 王怀宾. 算法行政问责的机制与评估——基于欧美 40 份政策的实施分析 [J]. 中国行政管理, 2024, (1): 89-99.

[5] 登哈特 J V, 登哈特 R B. 新公共服务：服务，而不是掌舵 [M]. 丁煌, 译. 北京: 中国人民大学出版社, 2010.

[6] 丁晓东. 论算法的法律规制 [J]. 中国社会科学, 2020, (12): 138-159, 203.

[7] 肖红军. 算法责任：理论证成、全景画像与治理范式 [J]. 管理世界, 2022, 38(4): 200-226.

[8] 夏志强, 黄琳珠. 算法行政责任分配与问责 [J]. 吉林大学社会科学学报, 2025, 65(6): 110-123, 239.

[9] 毛寿龙. 引咎辞职、问责制与治道变革 [J]. 浙江学刊, 2004, (1).

[10] 昌诚, 张毅, 王启飞. 面向公共价值创造的算法治理与算法规制 [J]. 中国行政管理, 2022, (10): 12-20.

[11] 董杨, 张蕴萌, 张雪然. 基于公共责任的人工智能监管：美国的关键做法及启示 [J]. 中国行政管理, 2024, (1): 112-120.

[12] Angelov P P, Soares E A, Jiang R, et al. Explainable artificial intelligence: An analytical review [J]. WIREs Data Mining and Knowledge Discovery, 2021, 11(5): e1424.

[13] 高童非. 论自动化行政中的“算法卸责” [J]. 电子政务, 2023, (2): 115-128.

[14] 曾雄, 梁正, 张辉. 欧美算法治理实践的新发展与我国算法综合治理框架的构建 [J]. 电子政务, 2022, (7): 67-75.

[15] 郑磊. 数字治理的效度、温度和尺度 [J]. 治理研究, 2021, 37(2): 5-16, 2.

[16] 张铤. 技术治理何以失灵? [J]. 自然辩证法研究, 2022, 38(11): 77-81.

[17] 郑磊, 陈奕醇. 技术崇拜、制度约束与故事建构——数字治理中的悖论表现与成因探讨 [J]. 探索与争鸣, 2025, (6): 105-113, 179, 2.

[18] 和思绮, 刘炳胜. 公民更信任公共部门中的“人”还是“人工智能”——一项考虑任务复杂程度与公共领域的调查实验研究 [J]. 公共行政评论, 2025, 18(4):

22-40, 196-197.

[19] 张惠彬, 王怀宾. 算法行政问责的机制与评估——基于欧美 40 份政策的实施分析 [J]. 中国行政管理, 2024, (1): 89-99.

[20] Djeffal C, Siewert M B, Wurster S. Role of the state and responsibility in governing artificial intelligence: a comparative analysis of AI strategies [J]. Journal of European Public Policy, 2022, 29(11): 1799-1821.

[21] 夏志强, 黄琳珠. 算法行政责任分配与问责 [J]. 吉林大学社会科学学报, 2025, 65(6): 110-123, 239.

[22] 张惠彬, 王怀宾. 算法行政问责的机制与评估——基于欧美 40 份政策的实施分析 [J]. 中国行政管理, 2024, (1): 89-99.

[23] 董杨, 张蕴萌, 张雪然. 基于公共责任的人工智能监管：美国的关键做法及启示 [J]. 中国行政管理, 2024, (1): 112-120.

[24] 郭晟豪, 王皓铎. 生成式人工智能治理：价值对齐的规制与训导路径 [J]. 中国行政管理, 2025, 41(10): 23-34.

[25] Aoki N, Tatsumi T, Naruse G, Maeda K. Explainable AI for government: Does the type of explanation matter to the accuracy, fairness, and trustworthiness of an algorithmic decision as perceived by those who are affected?[J]. Government Information Quarterly, 2024, 41(4): 101965.

[26] 郑磊. 数字治理的效度、温度和尺度 [J]. 治理研究, 2021, 37(2): 5-16, 2.

[27] 姜李丹, 薛澜. 我国新一代人工智能治理的时代挑战与范式变革 [J]. 公共管理学报, 2022, 19(2): 1-11, 164.

[28] Janssen M. Responsible governance of generative AI: conceptualizing GenAI as complex adaptive systems [J]. Policy and Society, 2025, 44(1): 38-51.

[29] 郑磊, 陈奕醇. 技术崇拜、制度约束与故事建构——数字治理中的悖论表现与成因探讨 [J]. 探索与争鸣, 2025, (6): 105-113, 179, 2.

[30] 张成福, 王祥州. 人工智能嵌入公共服务治理的风险挑战 [J]. 电子政务, 2023, (1): 37-51.

[31] Zaim R, Shaw J A. Toward responsible artificial intelligence in health: regulatory structures and power dynamics of the big tech industry in the United States [J]. Policy Studies, 2026, 47(2): 236-248.

更多免费行业报告 并购家 www.ipoipo.net

第三篇

中国高校 AI4SSH 的探索、行动与体系建制

第十章　中国高校 AI4SSH 发展的指数化观察[①]

AI 技术正以前所未有的广度与深度渗透至人文社会科学领域，推动研究范式由“工具性应用”走向“方法论重构”。在这一进程中，高校既是前沿探索的主阵地，也是交叉学科生态的关键组织者。如何在扩散式应用与建制化发展之间形成可持续路径，需要兼具“全景画像”与“导向性参照”的评价工具。

本章引入“中国高校 AI4SSH 指数”，以研究核心能力、发展创新潜力与社会传播能力三大维度为主轴，结合多源数据证据链，对我国高校 AI4SSH 的发展态势进行指数化观察。指数的目的在于识别结构特征、揭示能力短板与增长动能，推动融合研究从自发探索走向有组织、可评估、可迭代的体系建制。

通过“方法—画像—结论与展望”的结构，本章尝试回答三类关键问题：我国高校 AI4SSH 整体呈现何种梯度与集聚特征；哪些能力维度正在加速积累、哪些环节仍是明显短板；下一阶段应如何以制度供给、平台支撑与人才机制推进从规模积累到质量引领的跃迁。

[①] 本章由复旦大学大数据研究院赵星教授牵头，复旦大学国家智能评价与治理实验基地王译晗博士等合作撰写。

赵星教授现为复旦大学国家智能评价与治理实验基地副主任，上海市“曙光学者”、教育部教指委委员、国际信息计量学和科技评价领域权威期刊 *Journal of Informetrics* 编委，五种中文重要期刊编委。

更多免费行业报告 并购家 www.ipoipo.net

本章提要：

- 本章以“中国高校 AI4SSH 指数”为工具，面向全景式把握我国高校在 AI 与人文社会科学深度融合中的总体进展与结构特征，兼具“评估基准”与“发展导向”双重功能，为学科布局、资源配置与生态构建提供量化参照。
- 指数体系覆盖“研究核心能力—发展创新潜力—社会传播能力”三大维度，并细化为多层指标结构，强调从范式融合、组织协同与生态构建等视角识别高校 AI4SSH 综合能力与可持续发展潜力。
- 总体态势呈现“体系初构、梯次分明”的结构化发展格局：研究产出规模与本土融合推进较快，形成明显的集聚效应与区域协同骨架，但不同高校、不同区域之间的发展强度与能力结构差异显著。
- 生态建设正进入“实体化、建制化”加速阶段：交叉研究机构的设立与平台化布局成为推动 AI4SSH 从零散探索走向体系化发展的关键变量，但制度性支撑与长效机制在多数高校仍处起步阶段。
- 社会传播与公共影响呈现“圈层化”分布：学术成果向公共议程与社会服务的转化效率仍有提升空间，如何将研究能力转化为政策支撑与社会赋能，是下一阶段需要补齐的链条。
- 面向未来，亟待推动范式融合由“应用跟随”迈向“创新引领”；构建“内外联动、制度保障”的可持续生态；强化需求驱动的社会服务与话语体系；加快培养“厚基础、强交叉、善协同”的复合型人才队伍。

10.1 指数评价体系构建和指标计算方法

人工智能（AI）技术正以前所未有的广度与深度渗透至各学科领域，推动研究范式的系统性变革。在社会科学与人文科学（SSH）领域，AI4SSH 已从工具性应用演变为一场深刻的方法论革命，强调通过数据、模型与理论的深度融合，形成“解释—预测—干预”一体化的新研究范式。为系统把握我国高校在此前沿交叉领域的发展态势，引导学科布局与生态构建，构建“中国高校 AI4SSH 指数”具有重要现实意义。该指数兼具“评估基准”与“发展导向”双重功能，通过构建科学量化的评价体系，推动融合研究从自发探索转向有组织、体系化的理论创新与范式重构，为国家战略决策与资源优化提供关键实证支撑。

本报告构建的 AI4SSH 指数评估体系共包含 3 个一级指标、7 个二级指标和 10 个三级指标，涵盖“研究核心能力”“发展创新潜力”与“社会传播能力”三大维度。数据来源多元，涵盖 Web of Science 核心合集、中国知网（CNKI）两大学术论文数据库，高校官网网站信息，以及微信、微博、抖音等社交媒体平台的公开传播数据。报告原创性引入“AI 渗透率”“国际科研合作组织力”“国际交叉创新潜力”等特色指标，从范式融合、组织协同与生态构建等多重视角，系统评估各高校在 AI 与人文社科交叉领域的融合程度、发展规模、创新潜力与综合影响。

展望未来，应着力推动范式融合从应用跟随向创新引领升级，构

更多免费行业报告 并购家 www.ipoipo.net

建内外联动、制度保障的可持续发展生态，强化需求驱动的社会服务与话语体系建设，并加快培育厚基础、强交叉、善协同的复合型人才队伍。

本指数通过首次全景式、结构化评估，科学揭示了中国高校AI4SSH发展现状与内在逻辑，为优化学科生态、提升国际话语权提供了实证依据与战略指引，有望推动我国在该全球前沿领域实现从规模积累到质量引领的系统跨越，在人工智能驱动人文社科研究范式转型的关键时期，为构建具有中国特色、世界影响的交叉学科创新体系贡献实证支撑与实践路径。

1. 指数评价体系的构建原则

为科学、客观地评估中国高校在“人工智能赋能社会科学与人文研究”（AI4SSH）领域的综合能力与发展水平，本指数评价体系的构建遵循以下三项原则：

系统性与典型性原则。AI4SSH能力涵盖理论融合、技术应用、成果转化与学术影响等多个维度。本指数评价体系在系统覆盖“理论－技术－成果－影响”全流程的基础上，重点选取能代表学科深度融合与范式创新特征的指标，如在成果评价中，不仅关注论文产出数量，更注重衡量其在推动 SSH 理论创新与方法突破方面的融合程度与实质贡献。

可量化与可获得性原则。AI4SSH 指数评价体系的设计优先采用概念清晰、便于量化比较的指标。在确保数据真实可靠的前提下，兼

顾不同高校数据的可获得性。所有分析数据均来源于 Web of Science 核心合集与中国知网（CNKI）两大权威数据库，高校官网以及主流媒体，以保证评价覆盖范围与结果的可信度。

导向性与发展性原则。AI4SSH 指数评价体系设计注重体现对学科交叉融合与创新发展的引导功能，在全面覆盖科研活动关键环节的基础上，突出对机构可持续发展能力与未来潜力的关注。通过设置研究融合度、原始创新力、组织领导力等前瞻性指标，引导高校不仅关注即时的科研成果产出，更要重视研究范式的创新性与国际合作的引领性。该原则确保评价体系既立足当前发展现状，又着眼于长远进步潜力，为推动 AI4SSH 领域的健康持续发展提供方向指引。

2. 指数评价体系

表 10-1 中国高校 AI4SSH 指数评价体系

一级指标（权重）	二级指标（权重）	三级指标（权重）	指标内涵
研究核心能力（0.7）	研究融合度（0.1）	国际研究 AI 渗透率（0.05）	用于衡量 AI 在机构研究中的采纳程度与融合深度。
		国内研究 AI 渗透率（0.05）	
	学术产出规模（0.3）	国际研究成果产出规模（0.1）	用于衡量机构研究活动的总体规模与产出能力，反映一个机构在该领域的投入程度和学术体量。
		国内研究成果产出规模（0.2）	
	学术影响力（0.3）	国际学术影响力（0.1）	用于衡量机构研究成果所产生的整体学术影响力与受关注程度，反映其被同行参考和借鉴的价值。
		国内学术影响力（0.2）	
发展创新潜力（0.15）	组织领导力（0.05）	国际科研合作组织力（0.05）	用于衡量机构在 SSH 跨国 AI 合作研究中，是否扮演引领组织者

更多免费行业报告 并购家 www.ipoipo.net

			角色。
	原始创新力（0.05）	国际交叉创新潜力（0.05）	用于衡量机构研究成果在知识来源组合上的交叉创新性与新颖程度。
	生态构建力（0.05）	实体化建制水平（0.05）	衡量机构为支持AI与SSH深度融合，在实体组织机构层面的战略布局与投入程度。
社会传播能力（0.15）	公共影响力（0.15）	公共传播影响力（0.15）	衡量 AI4SSH 建设成果的社会传播声量。

在系统分析人工智能赋能社会科学与人文研究理论内涵的基础上，项目团队充分吸收整合了既有计量指标（如学术产出规模、学术影响力），并综合考虑我国高校的实际情况，原创性地从研究核心能力、发展创新能力中国高校和社会传播能力三个维度构建了AI4SSH指数评价体系，以科学捕捉 AI 与 SSH 交叉融合的深度、组织领导力和创新潜力（见表 10-1）。其中，“AI 渗透率”旨在从宏观层面度量 AI 浪潮下机构对研究范式的革新融合程度；“国际科研合作组织力”则从微观合作网络视角，识别机构在全球科研合作中的主导能力；“国际交叉创新潜力”基于前沿的知识组合理论与算法，量化了研究成果在思想源头方面的原创性。

3. 数据来源与数据处理

在本指数框架内，“AI4SSH”涵盖 AI 以任何形式对 SSH 研究领域的介入或赋能，包括但不限于将 AI 作为研究工具、分析方法或研究对象，从而全面刻画该交叉领域的研究生态。

本指数构建以学术论文为核心数据基础，数据来源于 Web of Science 核心合集与中国知网（CNKI），分别覆盖国际权威的社会科学

引文索引（Social Sciences Citation Index，简称 SSCI）、艺术与人文引文索引（Arts & Humanities Citation Index，简称 AHCI） 与国内的中文社会科学引文索引（Chinese Social Sciences Citation Index，简称 CSSCI），以确保全面反映我国高校在人文社科与人工智能交叉领域的高水平研究成果。

数据采集时间范围为 2023 年 1 月 1 日至 2025 年 12 月 31 日（检索日期：2025 年 10 月 29 日，下载日期：2025 年 10 月 30 日-2026 年 2 月 27 日），以聚焦人工智能技术在人文社科领域的第一个繁荣期。

机构界定上，严格限定为第一作者或通讯作者单位位于中国内地高校的文献：其中中文文献依据第一作者中排序首位的机构认定，外文文献依据通讯作者中排序首位的机构认定。

文献类型上，统一限定为研究论文：WoS 核心合集限定为 Article，CNKI 限定为研究论文，以确保成果类型的一致性，并聚焦于具有完整研究过程和原创价值的学术成果。

检索策略基于权威关键词集构建：参考 Schmallenbach 等（2024）[①]与 Gao 等（2024）[②]的研究方法，综合经济合作与发展组织（OECD）所建立的 214 个 AI 术语词表[③]，以及 Microsoft Academic Graph 中标注的“机器学习”“人工智能”“自然语言处理”“计算机视觉”“模式识别”等核心领域标签。采用并集逻辑在标题与摘要字段中执行检索，

[①] Schmallenbach, L., Bärnighausen, T.W. & Lerchenmueller, M.J. The global geography of artificial intelligence in life science research. Nat Commun 15, 7527 (2024). https://doi.org/10.1038/s41467-024-51714-x.
[②] Gao, J., Wang, D. Quantifying the use and potential benefits of artificial intelligence in scientific research. Nat Hum Behav 8, 2281–2292 (2024). https://doi.org/10.1038/s41562-024-02020-5.
[③] Baruffaldi, S. et al. Identifying and measuring developments in artificial intelligence: Making the impossible possible. OECD Science, Technology and Industry Working Papers, No. 2020/05, 1–68 (2020).

更多免费行业报告 并购家 www.ipoipo.net

形成覆盖人工智能各分支领域的完整检索式（详见附录一），以兼顾查全率与查准率。

检索结果显示，WoS 核心合集中获得全国 SSH 领域 AI 相关论文共 17,049 篇，涉及 915 个国内通讯机构，SSH 领域论文总量为 157,891 篇，涉及国内机构 1,410 个；CNKI 中获得 SSH 领域 AI 相关论文 16,140 篇，SSH 领域论文总量为 137,868 万篇。

数据处理流程包括三个关键环节：首先进行数据清洗，剔除重复记录、通知、声明及辑刊等非研究性文献；其次进行机构归一化处理，参考教育部《全国高等学校名单》对机构名称进行标准化规范，统一表述形式；最后进行学科复核，依据 WoS 核心合集与 CNKI 的学科分类体系，人工校验并确保所有论文均属于人文社科范畴。通过上述标准化流程，最终构建用于指数计算的规范化论文数据集。

4. 指标计算方法

本指数体系包含 3 个一级指标，7 个二级指标，10 个三级指标，从不同维度系统评估中国高校在人文社科领域的 AI 相关研究水平。本节按三级指标逐项说明其具体计算方法。

（1）国际研究 AI 渗透率

AI 渗透率指标由项目团队原创提出。国际研究 AI 渗透率是衡量全国各机构在 SSH 领域发表的国际论文中，人工智能技术被采纳与融合广度的核心指标。本指标基于 Web of Science 核心合集的 SSCI/AHCI 数据，通过三级递进计算，旨在从“研究范式转换”的相

对比例视角，客观评估机构在国际学术前沿的智能融合水平与相对位置。

国际研究机构 AI 渗透率：表征某机构在其国际发表的 SSH 论文中，AI 相关研究的占比，计算公式为：

$$P_{inst}^{int} = \frac{N_{AI}^{int}}{N_{total}^{int}} \times 100\%$$

其中，P_{inst}^{int}表示国际研究机构 AI 渗透率，N_{AI}^{int}表示观测期内该机构在 SSH 领域发表的 AI 相关英文论文数量，N_{total}^{int}表示该机构同期在 SSH 领域发表的英文论文总数。

国际研究全国 AI 渗透率：表征全国层面在 SSH 国际研究中 AI 相关研究的整体普及程度，作为评估的基准参考值，计算公式为：

$$P_{inst}^{nat} = \frac{N_{AI}^{nat}}{N_{total}^{nat}} \times 100\%$$

其中，P_{inst}^{nat}表示国际研究全球 AI 渗透率，N_{AI}^{nat}表示观测期内所有机构在 SSH 领域发表的 AI 相关英文论文总数，N_{total}^{nat}表示同期所有机构在 SSH 领域发表的英文论文总数。

国际研究机构相对 AI 渗透率：该指标通过排除宏观趋势的干扰，聚焦于表征某机构在 AI 技术采纳上相对于全国平均水平的真实表现，计算公式为：

$$\mathrm{PAI}_{\mathrm{int}} = \frac{P_{inst}^{ins}}{P_{inst}^{nat}} \times 100\%$$

其中，$\mathrm{PAI}_{\mathrm{int}}$表示国际研究机构相对 AI 渗透率，$P_{inst}^{ins}$表示国际研

更多免费行业报告 并购家 www.ipoipo.net

究机构 AI 渗透率，P_{inst}^{nat}表示国际研究全国 AI 渗透率。

因此，国际研究机构相对 AI 渗透率即构成本报告指标体系中“研究核心能力”维度的三级指标——国际研究 AI 渗透率。该指标值为 1 表示该机构处于全国平均水平；大于 1 表示高于平均水平；小于 1 则表示低于平均水平。通过三级递进测算，该指标构建了一个能有效剥离数量规模影响的评估框架。它不仅反映了单一机构在国际学术前沿采纳 AI 技术的广度，更重要的是，通过将机构表现与全国基准水平对比，揭示了该机构在全球性研究范式转型进程中的相对位置与引领能力。该指标克服了传统排名依赖绝对产出的局限，为在 AI4SSH 这一交叉领域内，识别和评价真正具备融合深度与发展潜力的机构，提供了创新性的量化视角。

（2）国内研究 AI 渗透率

国内研究 AI 渗透率旨在衡量各机构在我国核心学术期刊发表的 SSH 论文中，人工智能技术的采纳与融合深度。本指标基于中国知网（CNKI）中的 CSSCI 数据，通过与“国际研究 AI 渗透率”指标相同的三级递进计算，旨在从“研究范式转换”的相对比例视角，客观评估机构在本国学术体系内的智能转型进展与相对位置。

国内研究机构 AI 渗透率：表征某机构在国内发表的 SSH 论文中，AI 相关研究的占比，计算公式为：

$$P_{inst}^{dom} = \frac{N_{AI}^{dom}}{N_{total}^{dom}} \times 100\%$$

其中，P_{inst}^{gom}表示国内研究机构 AI 渗透率，N_{AI}^{dom}表示观测期内该

机构在 SSH 领域发表的 AI 相关中文论文数量，N_{total}^{int}表示该机构同期在 SSH 领域发表的中文论文总数。

国内研究全国 AI 渗透率：表征全国层面在 SSH 领域国内研究中 AI 相关研究的整体普及程度，作为评估的基准参考值，计算公式为：

$$P_{dom}^{nat} = \frac{N_{AI}^{nat-dom}}{N_{total}^{nat-dom}} \times 100\%$$

其中，P_{dom}^{nat}表示国内研究全国 AI 渗透率，$N_{AI}^{nat-dom}$表示观测期内所有机构在 SSH 领域发表的 AI 相关中文论文总数，$N_{total}^{nat-dom}$表示同期所有机构在 SSH 领域发表的中文论文总数。

国内研究机构相对 AI 渗透率：该指标通过排除宏观趋势的干扰，聚焦于表征某机构在 AI 技术采纳上相对于全国平均水平的真实表现，计算公式为：

$$\mathrm{PAI_{dom}} = \frac{P_{inst}^{dom}}{P_{dom}^{nat}} \times 100\%$$

因此，国内研究机构相对 AI 渗透率即构成本报告指标体系中“研究核心能力”维度的三级指标——国内研究 AI 渗透率。该指标值为 1，表明该机构处于全国平均水平；大于 1 表明高于平均水平；小于 1 则表明低于平均水平。本指标构建了一个植根于本土学术语境的评估框架。它不仅衡量单一机构在本国核心知识体系中应用 AI 技术的广度，更重要的是，通过将机构表现与全国基准水平对比，揭示了该机构在推动本土化研究范式转型、服务国家战略需求与学科内涵建设方面的相对贡献与引领效能。该指标为识别在本土 AI4SSH 发展中具

更多免费行业报告 并购家 www.ipoipo.net

有示范性和深化潜力的机构，提供了关键的量化依据。

（3）国际研究成果产出规模

产出规模是衡量机构在 AI4SSH 领域研究活动总体体量与产出能力的基础指标，直接反映机构在该交叉学科领域的学术体量与资源投入程度。

国际研究成果产出规模，旨在评估机构在该交叉学科领域的国际学术能见度与影响力，具体通过机构在 SSCI 与 AHCI 来源期刊上发表的 AI4SSH 领域论文总数进行测度。计算公式为：

$$P_{int} = N_{SSCI/AHCI}$$

其中，P_{int}表示国际研究成果产出规模，$N_{SSCI/AHCI}$ 为机构在观测期内发表于 SSCI 或 AHCI 索引期刊的 AI4SSH 论文数量。该指标旨在评估机构参与国际学术对话的能力与活跃度，为综合评价机构在 AI4SSH 领域的国际学术影响力提供基础支撑。

（4）国内研究成果产出规模

国内研究成果产出规模，旨在评估机构在我国学术体系内的研究体量与贡献度，反映其服务于国内学术发展与问题解决的能力。该指标通过机构在 CSSCI 期刊上发表的 AI4SSH 领域论文总数进行测度。计算公式为：

$$P_{dom} = N_{CSSCI}$$

其中，P_{dom}表示国内研究成果产出规模，N_{CSSCI}为机构在观测期内发表于 CSSCI 期刊的 AI4SSH 论文数量。该指标旨在评估机构对本

土学科发展的贡献程度，为综合评价机构在 AI4SSH 领域的本土研究贡献与学科建设能力提供基础支撑。

（5）国际学术影响力

国际学术影响力是衡量机构在 AI4SSH 领域研究成果所产生的整体学术影响与受关注程度的核心指标，直接反映了其研究成果被同行参考、借鉴和认可的价值。本指标基于 WoS 核心合集的 SSCI/AHCI 期刊数据，通过机构在观测期内发表的 AI4SSH 英文论文所获得的总引用次数进行量化评估。该指标的计算公式为：

$$\mathrm{I}_{\mathrm{total}}^{\mathrm{ins}} = \sum_{i=1}^{N_{ins}} C_i$$

其中，$\mathrm{I}_{\mathrm{total}}^{\mathrm{ins}}$表示国际学术影响力，$N_{ins}$表示该机构在观测期内发表于 SSCI/AHCI 期刊的 AI4SSH 英文论文总数，C_i表示第i篇论文在统计时点获得的总引用次数。

该指标能够客观反映机构研究成果在国际学术共同体中引发的关注程度和知识扩散效应，是评价机构在国际范围内学术贡献与学科影响力的关键依据。

（6）国内学术影响力

国内学术影响力是衡量机构在 AI4SSH 领域发表的中文成果所产生学术影响与受关注程度的对应指标，反映了其研究成果在我国学术界被参考、借鉴与认可的价值。本指标基于中国知网（CNKI）的中文社会科学引文索引（CSSCI）来源期刊数据，通过机构在观测期内发

更多免费行业报告 并购家 www.ipoipo.net

表的 AI4SSH 中文论文所获得的总引用次数进行量化评估。

该指标的计算公式为：

$$\mathrm{I}_{\mathrm{total}}^{\mathrm{dom}} = \sum_{i=1}^{N_{dom}} C_i$$

其中，$\mathrm{I}_{\mathrm{total}}^{\mathrm{dom}}$表示国内学术影响力，$N_{dom}$表示该机构在观测期内发表于 CSSCI 期刊的 AI4SSH 领域中文论文总数，C_i表示第i篇论文在统计时点获得的总引用次数。

引用数据统一截至于同一统计时点，确保各机构之间的可比性。该指标能够客观反映机构研究成果在国内学术共同体中引发的关注程度和知识扩散效应，是评价机构在本国学界内学术贡献与学科影响力的重要依据。

（7）国际科研合作组织力

“国际科研合作组织力”指标由项目团队原创提出。该指标突破传统评价关注国际合作论文数量的局限，从科研主导力的视角出发，衡量机构在 SSH 领域跨国 AI 合作研究中是否扮演引领者或组织者的关键角色，以此反映机构的国际科研主导性和学术影响力。指标通过计算机构在跨国合作中以通讯作者第一单位身份发表论文的比例，来量化其在国际合作中的主导作用。具体计算分为以下三步。需说明的是，此处将跨国合作 AI 论文定义为作者单位包含至少两个不同国家的 AI 相关论文。

机构国际科研合作组织力

首先计算单个机构在国际合作中的主导力。该数值由机构作为通讯作者第一单位发表的跨国合作 AI 论文数量，除以该机构参与的所有跨国合作 AI 论文总数得出。公式表达为：

$$L_{org_inst} = \frac{N_{lead_inst}}{N_{total_inst}}$$

其中，L_{org_inst}表示机构国际科研合作组织力，N_{lead_inst}表示该机构作为通讯作者第一单位的跨国合作 AI 论文数量，N_{total_inst}表示该机构参与的所有跨国合作 AI 论文总数。该值越接近 1，表明该机构的主导地位越强。

国家国际科研合作组织力

其次计算全国范围的基准水平。该数值由所有中国机构作为通讯作者第一单位发表的跨国合作 AI 论文总量，除以所有中国机构参与的跨国合作 AI 论文总量得出。公式表达为：

$$L_{org_nat} = \frac{N_{lead_nat}}{N_{total_nat}}$$

其中，L_{org_nat}表示国家国际科研合作组织力，N_{lead_nat}表示所有中国机构作为通讯作者第一单位的跨国合作 AI 论文数量，N_{total_nat}表示所有中国机构参与的所有跨国合作 AI 论文总数。该值反映了国家层面的平均主导水平。

机构相对国际科研合作组织力

最后计算机构的相对表现。该数值由第一步得出的机构国际科研合作组织力，除以第二步得出的国家国际科研合作组织力。公式为：

更多免费行业报告 并购家 www.ipoipo.net

$$L_{org} = \frac{L_{org_inst}}{L_{org_nat}}$$

其中，L_{org}表示机构相对国际科研合作组织力。该指标用于横向比较：若 L_{org}>1，则表明该机构的组织力高于全国平均水平；若L_{org}<1，则低于平均水平。

通过这一指标的测算，可以客观评估各高校在人文社科 AI 研究国际合作网络中的实际地位和影响力，为提升我国在该领域的国际话语权提供参考依据。

（8）国际交叉创新潜力

“国际交叉创新潜力”是本项目团队开发的前瞻性创新指标，是衡量机构在 AI4SSH 领域研究成果知识组合创新性与学科交叉融合度的核心指标。项目团队在 Uzzi 等（2013）[①]关于“非典型组合”与科学影响的研究的基础上，提出的引文组合新颖性框架（见图 10-1），通过量化机构发表论文在参考文献组合上的非典型知识组合程度，表征研究的原始创新潜力。[②]

需要特别说明的是，本指标侧重于评价研究成果在知识来源组合上的新颖程度，而非直接评价论文质量的高低。新颖的知识组合并不等同于成果的高质量，但非典型的知识重组往往与原创性探索、学科交叉突破密切相关，因此可作为衡量科研创新潜力的有效参考。

[①] Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical Combinations and Scientific Impact. Science, 342(6157), 468–472.

[②] 杨杰,柳美君,步一,等. 组合新颖性和学科新颖性能否提升论文的技术影响？基于直接和间接技术影响力的双重视角[J].情报学报,2025,44(03):325-338.

具体测算过程如下：

基于蒙特卡洛模拟构建随机重组引用网络，通过 10 组随机网络建立网络零模型，计算期刊对组合的新颖性：

首先计算单篇论文中某一期刊对 m,n 的 Z 值：

$$Z(\text{Novelty})_{m,n} = \frac{\text{obs}(\text{pair}_{mn}) - \exp(\text{pair}_{mn})}{\sigma(\text{pair}_{mn})}$$

其中，$\text{obs}(\text{pair}_{mn})$表示真实知识网络中期刊对的频率；$\exp(\text{pair}_{mn})$和$\sigma(\text{pair}_{mn})$分别表示 10 次随机蒙特卡洛模拟引用网络中期刊对频率的均值与标准差。

对于单篇论文的全部引文期刊对，得到 Z 值分布向量$\exp(\text{pair}_{mn})$和$\sigma(\text{pair}_{mn})$，取其第 10 百分位数作为高维统计量Zscore_{10pt}，并通过下式转化为组合新颖性指标：

$$\text{Combination novelty} = \begin{cases} -\ln(\text{Zscore}_{10pt} + 1), \text{if } \text{Zscore}_{10pt} > 0 \\ \ln(-\text{Zscore}_{10pt} + 1), \text{if } \text{Zscore}_{10pt} \leq 0 \end{cases}$$

更多免费行业报告 并购家 www.ipoipo.net

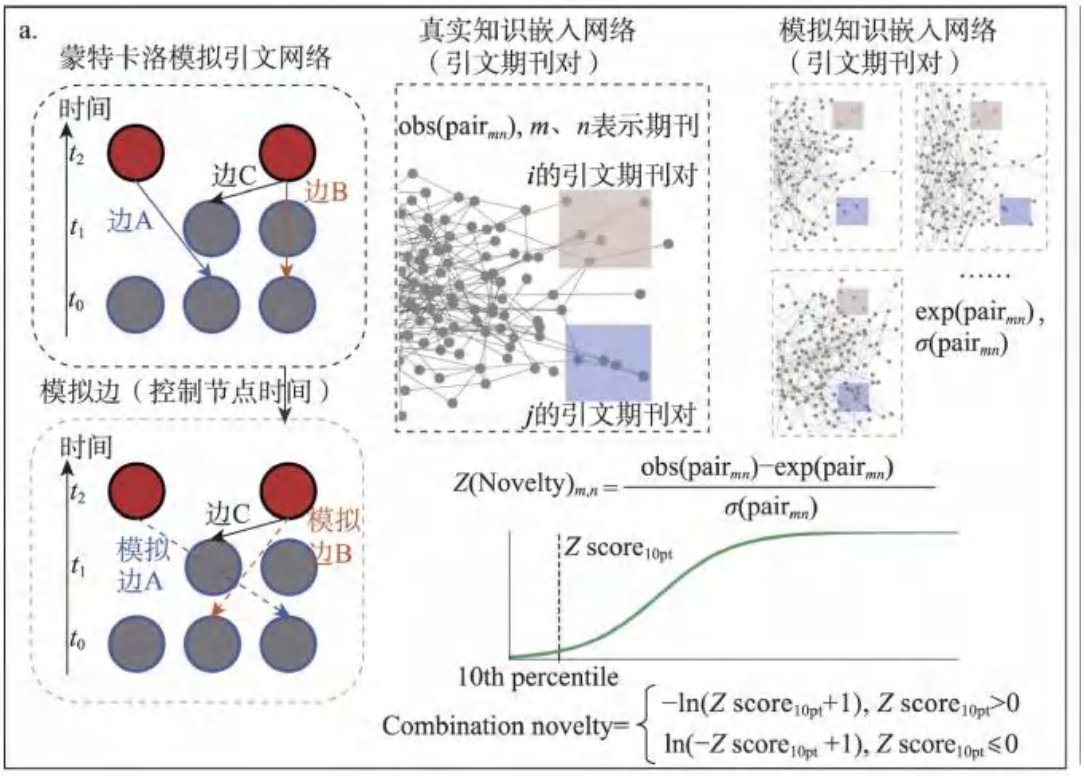


图 10-1 科学新颖性测度框架：引文组合新颖性

资料来源：杨杰,柳美君,步一,等. 组合新颖性和学科新颖性能否提升论文的技术影响？基于直接和间接技术影响力的双重视角[J].情报学报,2025,44(03):325-338.

机构的国际交叉创新潜力计算公式为：

$$I_{innov}=\frac{\sum_{i=1}^{N}(Combination\ novelty)_i}{N}$$

Combination novelty 表示单篇论文的组合新颖性，N 表示该机构在观测期内发表的 AI4SSH 领域国际论文总数，$\sum_{i=1}^{N}(Combination\ novelty)_i$ 表示该机构发表的 N 篇论文中，每一篇论文的组合新颖性得分的总和。该数值越大，表明机构的整体创新水平越高。

（9）实体化建制水平

“实体化建制水平”指标旨在评估高校为促进人工智能与人文社会科学深度融合，在实体组织机构层面进行战略布局与资源投入的水平。实体化、建制化的专门机构是交叉学科从概念倡议迈向常态化、

制度化发展的关键标志。通过衡量高校在此方面的投入，该指标能够反映其推动科研范式变革的顶层设计能力与生态构建决心。其计算公式为：

$$E = N_s$$

其中， E 表示实体化建制水平，N_s表示该校在统计时点前设立并持续运作、且完全符合以下两项核心筛选条件的二级实体机构（如独立研究院、研究中心、重点实验室等）总数：第一，机构名称或官方职能表述须同时涵盖（或明确指向）“人工智能”（或其关键子领域，如数据科学、机器学习等）与至少一个“人文社会科学”领域（如教育学、法学、经济学、管理学、政治性、社会学、新闻学、文学、历史学、哲学、艺术学等）；第二，机构的核心使命与研究方向，须明确聚焦于利用人工智能的理论、方法或工具，解决人文社会科学的基础问题或推动其方法论创新，而非将人文社科仅作为技术应用的被动场景。

本指标统计基于系统检索各高校官方网站“组织机构”“院系设置”等相关栏目下公布的完整机构清单，并排除了纯工科导向的人工智能学院、非实体组织及行政服务类中心（本指标检索时间为 2025 年 12 月 1 日，基于公开信息进行统计，对于个别需内部权限方可访问机构信息的高校可能造成统计偏差），从而能够客观衡量高校在 AI4SSH 交叉方向上进行制度化布局的实质性投入与战略重视程度。

（10） 公共传播影响力

“公共传播影响力”指标旨在系统测度高校在 AI4SSH 领域的标志

更多免费行业报告 并购家 www.ipoipo.net

性建设成果、关键进展与专家观点，在中央主流媒体舆论场中获得的报道强度与传播广度。需要说明的是，AI 领域的公共传播通常率先发酵于社交媒体与自媒体平台，相较于传统主流媒体，这些平台能够更早、更敏锐地捕捉到公众对该领域前沿动态的关注焦点与讨论走向。因此，本指标选取微信、微博、抖音作为核心数据来源，以更为精准地反映高校在该前沿交叉领域的社会能见度与公众对话能力。

在数据获取方面，本指标以"高校名"+("AI"OR "人工智能")为核心检索式，追溯时间范围为 2023 年 1 月至 2025 年 12 月，并对检索结果进行人工去重处理。针对不同平台的内容形态与传播特征，分别设定统计口径：微信平台限定类型为“文章”，按热度排序，统计阅读量达到 2000 及以上的推文数量；抖音平台限定类型为“视频”，按热度排序，统计点赞量达到 1000 及以上的视频数量；微博平台限定类型为“话题”，统计话题数量、各话题阅读量及讨论量。其中，微信推文阅读量 2000 以上、抖音视频点赞量 1000 以上的筛选阈值设定，主要基于以下考量：达到上述量级的传播内容已突破小范围关注，产生了初步且可被观测的社会影响力，能够较为稳定地表征高校成果在公众舆论中的可见程度。

在权重设计上，本研究基于大语言模型 DeepSeek 与 ChatGPT-5 对指标重要性进行的多次迭代分析，依据不同平台在信息传播链条中的功能差异，构建了分层赋权体系。具体权重分配为：微信推文占 0.50，抖音视频占 0.30，微博话题数占 0.06，微博话题阅读量占 0.08，微博话题讨论量占 0.06。上述权重设计的内在逻辑在于：微信平台以长文

本与深度解读为主要传播形态，其内容具有较强的专业性与持续影响力，故赋予最高权重；抖音平台以短视频为核心载体，强调传播的瞬时扩散速度与广泛受众覆盖能力，权重次之；微博平台则兼具议题生成与舆论发酵的双重功能，其话题数反映议题触发能力，阅读量表征传播广度，讨论量体现互动深度，三者综合赋权以全面衡量其在公共讨论中的复合作用。

本指标的综合计算公式如下：

$$S = 0.50 \times N_{wechat} + 0.30 \times N_{douyin} + 0.06 \times N_{topic} + 0.08 \times R_{topic} + 0.06 \times D_{topic}$$

其中，S 表示高校的公共传播影响力综合得分；N_{wechat}表示该高校相关微信推文中阅读量达到 2000 及以上的文章数量；N_{douyin}表示该高校相关抖音视频中点赞量达到 1000 及以上的视频数量；N_{topic}表示该高校相关的微博话题数量；R_{topic}表示各微博话题的阅读量总和；D_{topic}表示各微博话题的讨论量总和。为消除不同指标间的量纲差异，在计算综合得分前，对各分项指标均进行归一化处理。

在数据执行层面，本指标由两名项目人员背对背执行检索、筛查与计数工作，复核一致后得到各高校的有效传播数据，并按上述权重体系计算综合得分，作为衡量高校 AI4SSH 领域公共传播影响力的最终依据。

5. AI4SSH 指数计算

在完成基础指标测算后，需通过系统性的指数合成方法将各维度指标整合为综合指数。本报告采用以下三个步骤进行计算：

更多免费行业报告 并购家 www.ipoipo.net

（1）指标标准化处理

在本节指出，为消除指标量纲与分布差异，首先对原始指标数据进行标准化处理。

对于研究融合度、组织领导力、原始创新力等已具备明确理论取值区间的相对指标，采用 Min-Max 标准化方法，将其统一映射至[0,1]区间，计算公式为：

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

对于学术产出规模、学术影响力、生态构建力、公共影响力等绝对数值指标，因其通常呈右偏分布，先进行对数转换压缩极端值，再实施 Min-Max 标准化，计算公式为：

$$x' = \frac{\ln(x+1) - \min(\ln(x+1))}{\max(\ln(x+1)) - \min(\ln(x+1))}$$

（2）权重确定

采用主客观组合赋权法，以科学确定各项指标的权重。权重分配注重体现对中国本土学术话语体系的重视，同时强调研究体量和影响力。具体实施中，首先基于大语言模型 deepseek、ChatGPT5 对指标重要性进行多次迭代分析，获得初始权重；继而采用专家咨询法，邀请 SSH 和 AI 交叉领域专家对初始权重进行评议与调整，最终权重综合了智能计算与专家智慧，确保权重分配的合理性。

（3）指数合成

采用线性加权求和模型，基于层次化权重结构计算各高校的

AI4SSH 综合指数。具体计算公式如下：

$$AI4SSH_i = \sum_{j=1}^{n} w_j \times x'_{ij}$$

其中为第 j 项指标的最终权重（详见表 10-1），为机构 i 在第 j 项指标上的标准化值。

本次评估构建了一个包含 3 个一级指标、7 个二级指标与 10 个三级指标的指标体系。各指标权重配置如表 10-1 所示：在一级指标中，“研究核心能力”作为核心维度，权重占 0.7；“发展创新潜力”和“社会传播能力”为特色维度，权重各为 0.15。

在二级指标层面，“研究核心能力”下设“研究融合度”（权重 0.1）、“学术产出规模”（权重 0.3）和“学术影响力”（权重 0.3）三个指标；“发展创新潜力”下设“组织领导力”（权重 0.05）、“原始创新力”（权重 0.05）和“生态构建力”（权重 0.05）三个指标；“社会传播能力”下设“公共影响力”指标，权重为 0.15。

在三级指标层面，“研究融合度”下设“国际研究 AI 渗透率”和“国内研究 AI 渗透率”两个指标，权重均为 0.05；“学术产出规模”下设“国际研究成果产出规模”（权重 0.1）和“国内研究成果产出规模”（权重 0.2）两个指标；“学术影响力”下设“国际学术影响力”（权重 0.1）和“国内学术影响力”（权重 0.2）两个指标；“组织领导力”下设“国际科研合作组织力”指标，权重为 0.05；“原始创新力”下设“国际交叉创新潜力”指标，权重为 0.05；“生态构建力”下设“实体化建制水平”指标，

更多免费行业报告 并购家 www.ipoipo.net

权重为 0.05；“公共影响力”下设“公共传播影响力”指标，权重为 0.15.

基于上述权重体系，AI4SSH 综合指数的计算公式为（按照三级指标）：

$$\mathrm{AI4SSH} = 0.05 \times \mathrm{PAI_{int}} + 0.05 \times \mathrm{PAI_{dom}} + 0.1 \times \mathrm{P_{int}} + 0.2 \times \mathrm{P_{dom}} + 0.1 \times \mathrm{I_{total}^{ins}} + 0.2 \times \mathrm{I_{total}^{dom}} + 0.05 \times \mathrm{L_{org}} + 0.05 \times \mathrm{I_{innov}} + 0.05 \times \mathrm{E} + 0.15 \times \mathrm{S}$$

其中各变量分别代表：

$\mathrm{PAI_{int}}$：国际研究 AI 渗透率

$\mathrm{PAI_{dom}}$：国内研究 AI 渗透率

$\mathrm{P_{int}}$：国际研究成果产出规模

$\mathrm{P_{dom}}$：国内研究成果产出规模

$\mathrm{I_{total}^{ins}}$：国际学术影响力

$\mathrm{I_{total}^{dom}}$：国内学术影响力

$\mathrm{L_{org}}$：国际科研合作组织力

$\mathrm{I_{innov}}$：国际交叉创新潜力

E：实体化建制水平

S：公共传播影响力

该权重体系的配置体现了评价设计的三个核心导向：一是突出研究质量与影响的核心地位，研究成果的规模与影响力指标合计权重达 0.6，其中学术产出与学术影响力各占 0.3，形成了以成果质量及规模为主的评价主体；二是兼顾国内外学术体系的均衡布局，国内研究相关指标权重略高于国际研究，既鼓励参与国际竞争，也强调对本土学术生态与话语体系建设的支撑；三是引导学科交叉与创新生态发展，通过设置研究融合度、组织领导力、原始创新力、生态构建力、公共影响力等特色指标（合计权重 0.4），推动高校关注范式融合、组织创新与长效生态构建。整体权重设计在强调核心科研表现的同时，注重

评价维度的多样性与引导性，为高校在 AI4SSH 领域的发展定位提供结构化的参照框架。

10.2 中国高校 AI4SSH 指数与发展评价

指数评价的意义，不在于给高校贴上简单标签，而在于提供一扇“结构化观察窗”，帮助我们把握 AI4SSH 在中国高校的总体演进：从工具性应用走向方法论革命，从分散探索走向生态构建与体系建制。基于多源数据与统一口径，本节以综合指数为主线，对我国高校 AI4SSH 的发展态势进行全景式刻画，重点呈现其梯队结构、区域集聚与增长特征，揭示“体系初构、梯次分明”的总体格局，并据此讨论不同类型高校在学科禀赋、组织投入与外部链接上的差异化路径。需要强调的是，综合评价更适合作为宏观体检与趋势识别工具，其结论应与分项指标的结构解释相结合，以避免单一指标对复杂现实的过度简化，从而为资源配置、学科布局与平台建设提供更具可操作性的参照。

1. 中国高校 AI4SSH 指数综合评估

中国高校在该交叉领域的建设呈现出层次分明、梯度发展的结构化态势。综合指数得分分布显示，少数高校凭借在相关领域的长期布局与资源投入，已形成较为成熟的研究体系与产出规模；而大多数高校处于能力积累与快速成长期。

在具体能力评估中，综合性大学与部分优势高校处于领先方阵，细分核心评价指标的表现差异成为决定综合得分的关键。清华大学以

更多免费行业报告 并购家 www.ipoipo.net

85.76 分获得较高评估得分，数据显示其在“社会传播能力”下的公共影响力以及“发展创新潜力”等维度表现突出，体现了较高的公众触达率与知识扩散效能；武汉大学以 74.79 分紧随其后，其在国内外“学术产出规模”与“学术影响力”核心能力指标上均表现出众，反映了深厚的学科体量积累；复旦大学综合得分为 74.74 分，该校在“发展创新潜力”维度的表现极为抢眼，特别是在“生态构建力”与“组织领导力”两项指标上获得了极高的评分，展现出该校在交叉学科实体化建制、内部生态孕育以及跨国科研协同网络中的前瞻性战略布局。南京大学、上海交通大学、北京大学、浙江大学、北京师范大学、东南大学与中国人民大学等高校同样在此次综合评估中表现出色，稳居领先梯队。数据进一步显示，领先组高校在总分上保持了高度的连贯性，但在二级与三级指标上呈现出差异化的能力侧重。例如，中国人民大学在“生态构建力”指标上表现达到满分，表明其在实体机构建设方面投入卓著；南京大学在“国内学术产出规模”与“国内学术影响力”指标上均获得满分表现，显示出其在本土学术语境下的深厚耕耘。这些数据的交叉对比表明，国内高水平研究型大学在推进 AI 与 SSH 融合时，已根据自身的学科禀赋形成了不同的发展路径。

数据进一步显示，综合评价较为突出的高校在总分上保持了较好的连贯性，但在二级指标上呈现出差异化的竞争优势。例如，复旦大学与中国人民大学在“生态构建力”指标上得分较高，表明其在实体化建制与组织布局方面具有前瞻性；南京大学在“国内学术产出规模”与“国内学术影响力”指标上均获得满分表现，显示出其在本土学术

语境下的深厚耕耘。这些数据的交叉对比表明，国内高水平研究型大学在推进 AI 与 SSH 融合时，已根据自身的学科禀赋形成了不同的发展路径。

表 10-2 中国高校 AI4SSH 发展评价指数

高校	AI4SSH	研究核心能力						发展创新潜力			社会传播能力
	指数	研究		学术		学术		组织	原始	生态	公共
		融合度		产出规模		影响力		领导力	创新力	构建力	影响力
	得分	国际	国内	国际	国内	国际	国内	得分	得分	得分	得分
		得分	得分	得分	得分	得分	得分				
北京大学	71.22	55.12	26.89	81.30	91.15	83.31	78.72	56.76	47.80	44.44	61.57
北京工业大学	40.70	72.60	52.58	40.57	40.61	73.17	21.16	81.70	15.35	22.22	31.69
北京航空航天大学	47.79	54.16	69.28	39.15	52.95	54.53	42.48	77.78	25.33	11.11	49.68
北京交通大学	39.90	68.56	48.17	57.48	44.12	71.30	6.13	87.96	34.12	0.00	33.57
北京理工大学	48.92	36.75	70.74	33.81	52.95	49.65	31.55	95.79	97.01	22.22	50.30
北京师范大学	69.87	13.84	42.17	82.76	96.01	92.52	84.36	71.49	37.44	22.22	46.08
大连理工大学	45.20	75.19	53.96	42.62	59.61	55.11	25.94	56.22	24.98	44.44	37.17
电子科技大学	44.80	43.56	72.75	33.81	47.71	94.56	32.23	80.75	98.35	11.11	4.35
东北大学	31.87	54.78	19.77	33.81	38.00	59.96	10.05	47.62	28.93	22.22	28.10
东南大学	68.42	71.05	49.81	100.00	64.97	97.56	54.03	95.10	81.05	22.22	59.37
复旦大学	74.74	49.55	47.94	67.16	86.77	82.77	79.92	89.29	56.66	88.89	65.31
广州大学	38.79	23.99	33.49	30.47	46.23	41.75	46.01	62.50	26.15	22.22	31.38
哈尔滨工业大学	58.28	72.71	84.81	75.95	41.23	98.74	38.98	87.87	31.29	55.56	54.39
合肥工业大学	36.20	86.60	52.00	41.26	34.34	71.07	0.23	82.30	23.01	66.67	16.84
湖南大学	39.42	27.64	19.42	42.62	53.34	55.23	29.51	80.13	20.12	22.22	30.62
华东师范大学	64.83	19.42	44.82	58.45	88.34	62.21	78.47	72.60	46.70	66.67	45.92
华南理工大学	47.51	64.91	52.21	51.19	61.42	51.10	35.12	63.13	18.46	66.67	31.38
华南师范大学	49.26	13.47	55.43	22.98	68.79	53.22	32.07	81.02	18.46	77.78	61.08
华中科技大学	51.49	39.63	27.78	83.90	61.42	95.24	18.21	84.75	28.91	44.44	42.50
华中师范大学	62.96	42.21	40.10	59.40	83.33	64.68	76.26	83.33	28.58	44.44	44.67
吉林大学	56.43	28.47	20.10	57.96	77.18	61.92	68.30	76.06	95.86	11.11	25.13
暨南大学	40.60	23.17	12.02	47.10	49.57	73.58	36.26	66.14	29.38	22.22	24.82
南京大学	72.05	23.85	39.75	64.32	100.00	52.11	100.00	78.70	57.39	22.22	62.05
南京师范大学	45.84	24.11	70.34	31.32	69.67	39.70	58.21	44.80	32.50	66.67	8.25
南开大学	51.86	6.40	14.69	20.91	79.83	44.63	74.61	59.52	14.07	33.33	53.43
清华大学	85.76	51.48	50.31	86.64	94.31	96.64	83.13	56.76	91.42	88.89	100.00
厦门大学	48.32	11.75	8.58	51.75	62.28	68.93	53.23	37.88	76.37	11.11	39.11
山东大学	52.98	13.97	2.48	68.00	63.93	79.03	74.52	58.64	28.77	22.22	28.58
上海大学	54.87	41.75	62.68	60.33	76.17	74.13	72.86	81.70	31.03	11.11	1.37

更多免费行业报告 并购家 www.ipoipo.net

上海交通大学	71.63	24.70	57.64	85.29	83.19	88.16	77.80	74.53	34.85	22.22	75.95
深圳大学	50.91	26.46	50.37	52.30	56.66	63.33	42.38	69.44	25.73	22.22	65.53
四川大学	45.65	48.68	14.75	82.47	61.71	100.00	23.60	88.57	26.86	11.11	5.59
苏州大学	37.88	23.35	18.22	37.68	44.66	43.70	44.52	100.00	44.08	33.33	6.38
天津大学	48.09	40.21	39.20	56.48	49.57	69.56	19.35	71.23	100.00	66.67	38.94
同济大学	60.90	23.83	60.19	88.47	75.13	83.62	53.70	70.73	89.02	0.00	38.27
武汉大学	74.79	44.70	40.19	99.80	96.42	97.69	83.48	74.16	92.42	11.11	39.53
西安交通大学	60.76	16.82	35.68	62.15	72.34	79.98	66.26	79.37	76.51	44.44	41.24
西南财经大学	34.01	19.92	4.81	40.57	37.31	75.60	10.05	53.42	65.07	66.67	16.22
西南大学	42.71	12.55	26.84	24.97	71.34	32.51	55.95	44.64	73.88	33.33	12.94
西南交通大学	31.84	55.19	20.10	53.91	22.83	56.80	0.04	82.20	66.18	44.44	18.56
浙江大学	70.98	23.49	45.31	98.37	84.98	95.02	77.74	54.67	82.41	44.44	43.91
郑州大学	40.27	26.16	28.55	50.06	57.00	72.03	0.61	73.10	91.56	11.11	33.42
中国地质大学（武汉）	29.86	67.47	35.69	58.92	21.61	70.40	0.15	60.54	42.33	0.00	15.12
中国科学技术大学	42.22	34.80	90.82	36.88	35.86	66.74	20.77	89.48	81.85	11.11	34.21
中国矿业大学	47.02	43.09	39.49	50.63	43.56	67.73	31.89	62.50	54.91	11.11	63.59
中国人民大学	66.22	8.11	18.30	57.96	96.01	78.80	78.86	70.81	30.91	100.00	41.08
中南财经政法大学	36.72	1.70	21.51	8.95	60.83	51.86	44.42	57.87	30.82	22.22	19.19
中南大学	53.21	29.21	37.45	83.34	54.49	88.89	39.49	77.41	28.13	0.00	57.18
中山大学	65.35	25.81	26.77	84.46	77.66	89.92	74.43	68.66	25.86	22.22	60.15
重庆大学	45.52	28.07	22.47	52.84	49.12	90.75	51.37	92.59	22.47	0.00	18.56

注：本表按高校名称首字母排序

从区域分布特征分析，AI4SSH 指数结果呈现出以中心城市群为核心的空间集聚特征。长三角地区与京津冀地区构成了我国 AI4SSH 研究的两大核心枢纽。长三角地区汇聚了复旦大学、南京大学、上海交通大学、浙江大学、东南大学等多所处于领先方阵的高校，并在评估的总体样本中占据较高比例，整体展现出较强的区域协同发展厚度与创新活跃度。京津冀地区则依托清华大学、北京大学、北京师范大学与中国人民大学等，在学术产出与政策智库支持方面形成了密集的学术共同体。与此同时，华中地区的武汉大学与华中师范大学、华南地区的中山大学与深圳大学、以及东北地区的哈尔滨工业大学与吉林大学等，分别构成了各自区域的学术增长极。这种分布态势与国家区域

高等教育资源布局基本保持一致，客观反映了科研平台、算力资源与人才储备对交叉学科建设的底层支撑作用。

从院校类型与层次来看，评估表现优异的高校以“双一流”建设高校为主干力量，呈现出多类型高校协同演进的生态格局。综合性大学在榜单中占据主导地位，其较为完备的人文社科体系为人工智能技术的应用提供了丰富的理论基础与研究场景。同时，以理工科见长的高校，如上海交通大学、西安交通大学、哈尔滨工业大学等，凭借在自然语言处理、数据挖掘等底层技术层面的硬核基础，在“原始创新力”与“研究融合度”等指标上展现出较强的竞争力，成功将技术优势转化为交叉领域的研究效能。此外，部分在法学、财经等领域具有行业特色的高校（如中南财经政法大学、西南财经大学等）也进入榜单，反映出 AI4SSH 研究正逐步从通用方法论的探索，向特定学科领域的纵深应用拓展，初步形成了能力互补的学科生态体系。

2. 中国高校产出规模指标增长情况

产出规模的绝对值反映了高校在 AI4SSH 领域的历史积累与现有体量，而复合年均增长率（Compound Annual Growth Rate, CAGR）则能够有效衡量各高校在该领域的投入强度与发展加速度。CAGR（Compound Annual Growth Rate，复合年均增长率）是指在一段时间内，某指标（如产出规模）按复利方式逐年增长的平均速度。CAGR 假设增长是连续、平稳的，不受中间年份的波动影响，能更准确地反映一个机构持续增长的能力。计算公式为：

更多免费行业报告 并购家 www.ipoipo.net

$GAGR_{\text{近3年国内产出增长率}} = (\frac{2023\ \text{年国内产出规模}}{2024\ \text{年国内产出规模}})^{\frac{1}{2}} - 1$。

本项目团队选取 2023 年至 2025 年为观测窗口，重点考察综合表现突出的 20 所高校在国际与国内两个维度的产出增长态势，以系统评估其在不同学术话语体系下的发展潜力。

在国际学术产出维度，中国高校的 AI4SSH 研究呈现出稳步增长的态势。表 10-3 数据显示，我国优势高校在国际发文方面已完成从平稳积累向跨越式突破的转变。在观测期内，绝大多数评估对象的国际产出复合年均增长率保持在 70%以上的高位运行。其中，北京大学与南京大学的表现尤为抢眼，其 CAGR 分别达到 146.74%与 138.05%。值得深入剖析的是其年度增长的结构性跃升特征：在 2024 年，大部分高校的同比增长率尚处于 30%至 80%的区间；而进入 2025 年，几乎所有优势高校的同比增长率均突破了 100%，北京大学（223.44%）、南京大学（209.09%）、华东师范大学（194.12%）以及复旦大学（185.37%）等更是实现了产出规模的成倍翻番。这种在 2025 年的集中爆发，充分显示出中国高校前期在人工智能底层技术与人文社科理论交叉融合上的积淀，已转化为符合国际标准的实质性科研成果。

此外，综合性大学在此轮国际话语体系构建中发挥了中流砥柱的作用。以浙江大学与清华大学为例，其在保持基数庞大的同时，不仅 2025 年国际发文总量分别突破 250 篇与 200 篇大关，更维持了超过 100%的极高复合年均增长率。这客观反映出此类高校的学科生态具有较强的延展性与国际接轨能力，其研究方向与全球 AI4SSH 的科技伦理、计算社会科学等前沿议题形成了高度共振。

表 10-3　2023-2025 年中国高校国际产出规模增长情况

高校	2023 年（篇）	2024 年（篇）	2025 年（篇）	复合年均增长率（CAGR）
北京大学	34	64	207	146.74%
北京师范大学	52	66	148	68.71%
东南大学	71	99	207	70.75%
复旦大学	28	41	117	104.42%
哈尔滨工业大学	56	44	100	33.63%
华东师范大学	25	34	100	100.00%
华中师范大学	26	42	85	80.81%
吉林大学	24	39	91	94.72%
南京大学	24	44	136	138.05%
清华大学	50	75	209	104.45%
山东大学	33	44	115	86.68%
上海大学	25	45	86	85.47%
上海交通大学	48	65	178	92.57%
同济大学	49	65	184	93.78%
武汉大学	63	93	213	83.87%
西安交通大学	31	45	100	79.61%
浙江大学	48	94	251	128.67%
中国人民大学	26	42	101	97.09%
中南大学	54	71	110	42.72%
中山大学	45	73	145	79.51%

注：本表按高校名称首字母排序

相较于国际产出的稳步增长，国内产出规模的增长呈现出更为前置的加速特征与更为广泛的辐射效应，彰显了国内学术界对 AI 赋能范式本土化应用的敏锐响应。根据表 10-4，本土学术界对 AI4SSH 的接纳经历了从早期快速扩容到稳步深化的发展历程。区别于国际产出在 2025 年的井喷，国内产出规模在 2024 年即迎来了第一波激增。以清华大学、南京大学、吉林大学、东南大学和上海交通大学为代表，其在 2024 年的同比增长率均高达 100%乃至 140%以上，展现了较强的敏捷响应能力与本土问题意识。进入 2025 年，虽然部分高校的同比增速有所回落，但整体发文绝对体量依然维持在高位扩张通道中。

更多免费行业报告 并购家 www.ipoipo.net

表 10-4 2023-2025 年中国高校国内产出规模增长情况

高校	2023 年（篇）	2024 年（篇）	2025 年（篇）	复合年均增长率（CAGR）
北京大学	56	108	137	56.30%
北京师范大学	65	120	176	64.74%
东南大学	18	44	54	73.21%
复旦大学	43	76	142	81.47%
哈尔滨工业大学	13	12	24	35.87%
华东师范大学	54	89	130	54.92%
华中师范大学	39	76	113	70.07%
吉林大学	32	68	80	58.51%
南京大学	68	147	203	72.69%
清华大学	46	110	187	101.73%
山东大学	21	38	53	58.56%
上海大学	37	56	82	48.51%
上海交通大学	37	74	116	77.37%
同济大学	36	50	83	51.66%
武汉大学	76	112	179	53.38%
西安交通大学	35	48	68	39.80%
浙江大学	49	72	125	59.59%
中国人民大学	72	113	176	56.52%
中南大学	19	22	38	42.16%
中山大学	40	57	88	47.99%

注：本表按高校名称首字母排序

在这个过程中，部分具有顶层交叉战略规划的高校表现出较强的续航能力。清华大学在国内产出维度展现了持续的统治力，其复合年均增长率高达 101.73%，是观测范围内少有的 CAGR 突破 100%的高校；复旦大学与上海交通大学同样表现稳健，其 CAGR 分别达到 81.47%与 77.37%，这反映了上海地区高校在构建数字人文、智能治理等本土学术话语体系上的战略协同性。整体而言，国内产出规模的前置性爆发与持续增长，表明我国高校在 AI4SSH 研究中不仅具备强大的学术生产力，更致力于利用智能技术解决国家发展战略与社会转型中的实际问题，为学科范式的本土化演进夯实了坚实的成果基础。

10.3 中国高校 AI4SSH 指数分项指标分析

如果说综合评价回答“整体处于什么位置”，那么分项指标更关心“优势在哪里、短板在哪里、机制是什么”。AI4SSH 的发展并非单一维度竞争，而是由研究范式融合、持续产出能力、知识扩散效应、组织协同与生态建制等多因素共同塑形。本节围绕指数体系的核心逻辑，对各维度指标进行拆解分析：在“研究核心能力”层面，强调从“融合深度”到“产出厚度”再到“影响效度”的科研价值链；在“发展创新潜力”层面，关注国际合作组织力、源头创新与实体化建制；在“社会传播能力”层面，讨论成果出圈与公共影响的形成条件。通过这种结构化剖析，本节旨在把“差距”还原为可解释的要素组合，为高校制定差异化发展策略、为政策部门优化支持工具（平台、评价、数据治理与人才机制）提供更具针对性的证据基础。

1. 研究核心能力

本研究构建的 AI4SSH 综合指数在“研究核心能力”维度设立了“研究融合度”“学术产出规模”与“学术影响力”三项二级指标及六项三级指标，以系统、递进地评估高校在该交叉领域的科研综合实力。该维度的设计遵循从“方法融合”到“成果产出”再到“知识传播”的科研价值链逻辑：首先，“研究融合度”从范式转型与学科交叉深度切入，衡量人工智能在人文社科研究中的渗透程度与主动性；进而，“学术产出规模”从科研体量与持续产出能力出发，反映高校在该方向的学术积累与资源投入强度；最终，“学术影响力”聚焦于成果的学术价值与知识扩散效应，体现研究在学界获得的认可度与思

更多免费行业报告 并购家 www.ipoipo.net

想贡献。三项指标分别对应科研活动的“深度—厚度—效度”，形成逻辑连贯、层次分明的评价体系。为辨析高校在不同学术语境中的表现，各二级指标进一步区分为“国际”与“国内”两个观测维度，构成六项三级指标，兼顾全球前沿参与度与本土学术生态贡献度的综合评价。

（1）研究融合度

“研究融合度”旨在系统评估 AI 技术在高校人文社科研究中被采纳的广度与应用的深度，具体涵盖 AI 方法、工具或模型在研究设计、数据处理、分析推理及成果产出等环节的嵌入程度与整合水平，从而反映高校在推动“研究范式转换”的实质性能力。为精准量化该二级指标，本指数体系设计了“国际研究 AI 渗透率”和“国内研究 AI 渗透率”两项三级指标，分别基于不同学术场域的数据，衡量 AI 技术在研究中的融合程度与利用实效。通过对比高校在国际与国内两个层面的 AI 渗透率表现，可进一步洞察其在 AI 融合发展中的战略布局与资源投入倾向，为理解高校在该交叉领域的定位提供量化依据。

国际研究 AI 渗透率：国际研究 AI 渗透率在不同高校间呈现显著差异。表 10-5 列出了 AI4SSH 国际研究 AI 渗透率排名前 25 的中国高校及其对应数值。结果显示，合肥工业大学、大连理工大学、哈尔滨工业大学、北京工业大学及东南大学渗透率均高于 70%，处于国际 AI 高渗透率区间，表明其在 AI4SSH 领域的融合应用水平较为突出。

从高校类型分布来看，理工类院校整体表现显著，如合肥工业大学、大连理工大学、哈尔滨工业大学等位居前列。综合类院校则呈现

分化态势，其中东南大学、北京大学、东北大学等表现较为突出。此外，部分特色类院校亦展现出独特的研究优势，如北京交通大学、中国地质大学（武汉）、中国矿业大学等，在结合 AI 技术与自身专业领域方面形成深度融合，在国际学术舞台上表现卓越。

表 10-5 中国高校国际研究 AI 渗透率

高校	国际研究 AI 渗透率	高校	国际研究 AI 渗透率
北京大学	55.12	华中科技大学	39.63
北京工业大学	72.60	华中师范大学	42.21
北京航空航天大学	54.16	清华大学	51.48
北京交通大学	68.56	上海大学	41.75
北京理工大学	36.75	四川大学	48.68
大连理工大学	75.19	天津大学	40.21
电子科技大学	43.56	武汉大学	44.70
东北大学	54.78	西南交通大学	55.19
东南大学	71.05	中国地质大学（武汉）	67.47
复旦大学	49.55	中国科学技术大学	34.80
哈尔滨工业大学	72.71	中国矿业大学	43.09
合肥工业大学	86.60	中南大学	29.21
华南理工大学	64.91		

注：本表按高校名称首字母排序

国内研究 AI 渗透率：表 3-2 列出了 AI4SSH 国内研究 AI 渗透率排名前 25 的中国高校及其对应数值。如表 3-2 所示，在国内研究层面，中国科学技术大学以 90.82%的渗透率位居首位，成为唯一一所 AI 渗透率突破 90%的高校。哈尔滨工业大学、电子科技大学、北京理工大学及南京师范大学的渗透率均高于 70%，反映出这些高校在 SSH 学科研究中已实现 AI 技术的高度嵌入与纵深发展。

整体上，国内 AI 渗透率呈现明显的结构性特征：头部高校之间分差显著，而中后部高校渗透率则相对集中，差异较小。从院校类型来看，高渗透率梯队具有鲜明的工科导向，理工类院校占据主导地位，

更多免费行业报告 并购家 www.ipoipo.net

体现出其在 AI 融合应用方面的整体领先态势。师范类院校表现同样突出，南京师范大学位列前五且为其中唯一的非理工类高校，另有 5 所师范院校进入 TOP25，显示出“AI+教育”方向具有强劲的发展潜力。

表 10-6 中国高校国内研究 AI 渗透率

高校	国内研究 AI 渗透率	高校	国内研究 AI 渗透率
北京工业大学	52.58	华南师范大学	55.43
北京航空航天大学	69.28	华中师范大学	40.10
北京交通大学	48.17	南京大学	39.75
北京理工大学	70.74	南京师范大学	70.34
北京师范大学	42.17	清华大学	50.31
大连理工大学	53.96	上海大学	62.68
电子科技大学	72.75	上海交通大学	57.64
东南大学	49.81	深圳大学	50.37
复旦大学	47.94	同济大学	60.19
哈尔滨工业大学	84.81	武汉大学	40.19
合肥工业大学	52.00	浙江大学	45.31
华东师范大学	44.82	中国科学技术大学	90.82
华南理工大学	52.21		

注：本表按高校名称首字母排序

研究融合度：研究融合度是衡量各机构在 SSH 领域的研究活动中，AI 技术采纳程度与融合深度的综合性二级指标（权重为 0.05），其下包含两个三级指标：国内研究 AI 渗透率（权重为 0.025）与国际研究 AI 渗透率（权重为 0.025），从国内与国际双重视角，系统评估机构在 SSH 研究中与 AI 技术的融合水平，客观反映其在智能化学术研究范式转型中的整体表现与相对位置。研究融合度计算方式如下：

$$PAI = \frac{(PAI_{int} + PAI_{dom})}{2}$$

哈尔滨工业大学在 AI 与 SSH 的交叉融合方面展现出特色和优势。

合肥工业大学、大连理工大学、中国科学技术大学、北京工业大学、北京航空航天大学及东南大学的研究融合度均高于 60 分，构成具有明显优势的头部梯队。这些高校在 AI4SSH 领域实现了较高程度的研究范式融合与学科深度耦合，共同构成了国内该交叉领域发展的核心引领力量。

国内与国际 AI 渗透率数据对比：国际与国内 AI 渗透率数据对比揭示了中国高校在 AI4SSH 交叉学科融合发展中的不同路径特征。如图 3-4 所示，以“得分差”（即国内研究 AI 渗透率减去国际研究 AI 渗透率的差值）作为分类依据，可将高校划分为以下三种类型：

- **“均衡型”**（-10 分<得分差<10 分）：以天津大学、清华大学、复旦大学等为代表，其国际与国内 AI 渗透率水平较为接近；
- **“国际强国内弱型”**（得分差≤-10 分）：以西南交通大学、东北大学、合肥工业大学等为代表，其国际 AI 渗透率显著高于国内；
- **“国内强国际弱型”**（得分差≥10 分）：以中国科学技术大学、南京师范大学、华南师范大学等为代表，国内 AI 渗透率显著高于国际。

结合分布图分析可见，“均衡型”高校占比不足三成，说明仅少数院校形成了国内外协同渗透、均衡推进的格局。与此同时，“国内强国际弱型”高校占比已略超“国际强国内弱型”，整体上呈现出以国内学术体系为重心的“单极倾斜”特征。“均衡型”院校的结构，正代表了那些在 AI 研究渗透率上实现国内外双向发力、协同发展的高校，是构建健康、可持续的交叉学科生态的重要力量。

更多免费行业报告 并购家 www.ipoipo.net

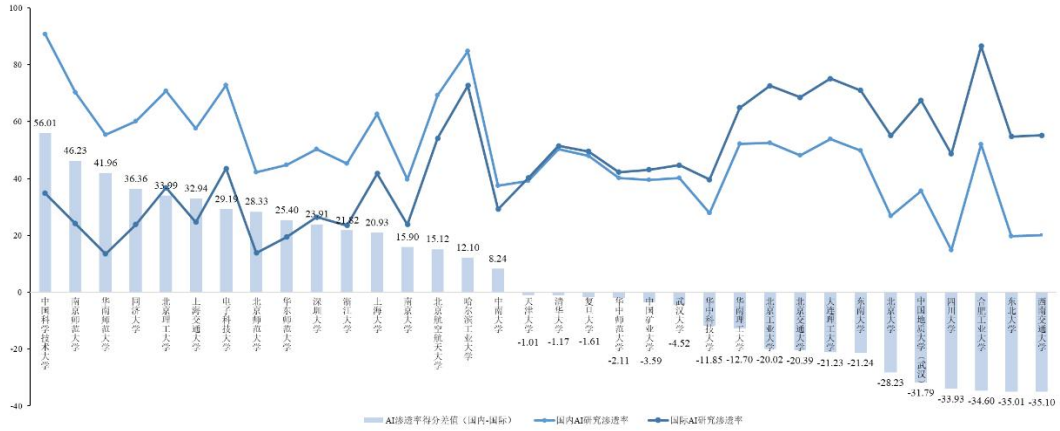


图 10-2 中国高校“国内-国际”研究 AI 渗透率差值分布

从高校类型来看，在国际与国内 AI 渗透率较高的高校中，工科类高校（含具有工科特色的综合类院校）占比较高。这类院校在 AI 技术研发与应用方面具有显著优势，其人文社科研究也多围绕工科相关交叉领域展开，因而更易开展高质量国际合作并在国际期刊发表成果，从而有效推动“研究范式转换”在国际层面的实践。

综合类高校可分为两种类型：一类以工科见长，另一类则强调整体学科均衡发展。这两类院校在国际研究定位上存在明显差异：以工科为特色的综合类院校，其人文社科国际研究多集中于交叉领域，引入 AI 技术能够较快提升研究的创新性与国际能见度；而以人文社科或基础理科为传统优势的综合类院校，其人文社科国际研究更注重理论深化与传承，AI 技术多为局部、辅助性应用，“研究范式转换”的整体推进程度相对有限。此外，部分传统人文社科领域在国际合作与成果发表方面本身存在一定的学科差异，这也使得其在渗透率指标上的表现相对受限。

（2）学术产出规模

学术产出规模作为图书情报与科学计量学中的经典分析指标，在本体系中用于衡量高校在AI4SSH领域的科研总体体量与成果产出能力。该指标综合反映机构在资源上的持续投入强度与高质量学术成果的积累水平，既体现其在该交叉领域的科研活跃度，也为评估其学术影响力提供基础。

本指标基于 WoS 核心合集与中国知网（CNKI）两大权威数据库，通过统计高校在观测期内发表于 SSCI/AHCI 与 CSSCI 来源期刊的 AI4SSH 论文数量，分别评估其在国际与国内两个维度的科研产出实力与发展势能。

表 10-7 中国高校国际研究成果产出规模

高校	国际研究成果产出规模	高校	国际研究成果产出规模
北京大学	81.30	上海大学	60.33
北京交通大学	57.48	上海交通大学	85.29
北京师范大学	82.76	四川大学	82.47
东南大学	100.00	天津大学	56.48
复旦大学	67.16	同济大学	88.47
哈尔滨工业大学	75.95	武汉大学	99.80
华东师范大学	58.45	西安交通大学	62.15
华中科技大学	83.90	浙江大学	98.37
华中师范大学	59.40	中国地质大学（武汉）	58.92
吉林大学	57.96	中国人民大学	57.96
南京大学	64.32	中南大学	83.34
清华大学	86.64	中山大学	84.46
山东大学	68.00		

注：本表按高校名称首字母排序

国际研究成果产出规模：从中国高校国际 AI4SSH 研究成果产出规模指标得分的结果来看（见表 10-7），东南大学、武汉大学与浙江大学的得分均超过 90，同济大学、清华大学、上海交通大学、中山大学、华中科技大学、中南大学、北京师范大学、四川大学、北京大学

更多免费行业报告 并购家 www.ipoipo.net

等 9 所高校得分均高于 80。这些高校展现出较高的科研活跃度，并在 AI4SSH 交叉前沿探索中形成了一定的示范效应。

整体上，高校在国际领域的产出体量差异较明显，前 25 所高校的指标数值跨度较大，整体呈现明显的梯度分化特征，梯度间差距显著，梯度内分差较小。

表 10-8 中国高校国内研究成果产出规模

高校	国内研究成果产出规模	高校	国内研究成果产出规模
北京大学	91.15	清华大学	94.31
北京师范大学	96.01	厦门大学	62.28
东南大学	64.97	山东大学	63.93
复旦大学	86.77	上海大学	76.17
华东师范大学	88.34	上海交通大学	83.19
华南理工大学	61.42	四川大学	61.71
华南师范大学	68.79	同济大学	75.13
华中科技大学	61.42	武汉大学	96.42
华中师范大学	83.33	西安交通大学	72.34
吉林大学	77.18	西南大学	71.34
南京大学	100.00	浙江大学	84.98
南京师范大学	69.67	中国人民大学	96.01
南开大学	79.83	中山大学	77.66

注：本表按高校名称首字母排序

国内研究成果产出规模：从中国高校国内 AI4SSH 研究成果产出规模指标得分的结果来看（见表 10-8），南京大学研究成果产出规模最大，武汉大学、北京师范大学、中国人民大学、清华大学、北京大学规模分值名列前茅。华东师范大学、复旦大学、浙江大学、华中师范大学、上海交通大学规模分值大于 80，在国内 AI4SSH 领域发展中贡献重要力量。

国内高校的产出规模整体差距较小、竞争激烈。尽管有高校分值超过 90，但超过 70%的高校得分集中于 60–85 之间，反映多数机构

仍处于技术引入与初步应用阶段，尚未实现深度的创新突破或规模化成果跃升。

学术产出规模：学术产出规模是衡量机构在 AI4SSH 领域研究活动总体体量与产出能力的综合性二级指标（权重为 0.4），其下包含两个三级指标：国际研究成果产出规模（权重为 0.15）与国内研究成果产出规模（权重为 0.25）。在当前阶段，国内研究成果的产出与传播，对于提升我国在 AI4SSH 领域的整体学术水平、培养本土研究力量、形成具有中国特色的研究范式具有更为直接和关键的作用，因此在指标体系中被赋予了更高的权重。学术产出规模计算公式如下：

$$P = P_{int} * 0.375 + P_{dom} * 0.625$$

武汉大学在成果规模的综合表现上具有显著优势；清华大学、北京师范大学、浙江大学等 13 所高校得分均在 75 以上，形成“双高协同”的头部梯队——既具备突出的成果产出体量，又在国内外均具有高质量表现，体现了顶尖高校的资源整合能力与学术竞争力。

国际与国内研究成果产出规模数据对比：对比各高校在国际与国内产出的侧重，可将其分为三类：中南大学、华中科技大学等更侧重国际布局（国际−国内差值≤-10 分）；南开大学、西南大学等更聚焦国内发展（差值≥10 分）；清华大学、上海交通大学、武汉大学等则呈现国内外均衡发展态势（差值介于-10 至 10 分之间）。这一分布反映出不同高校在推进 AI4SSH 研究时的战略定位与路径选择差异。

更多免费行业报告 并购家 www.ipoipo.net

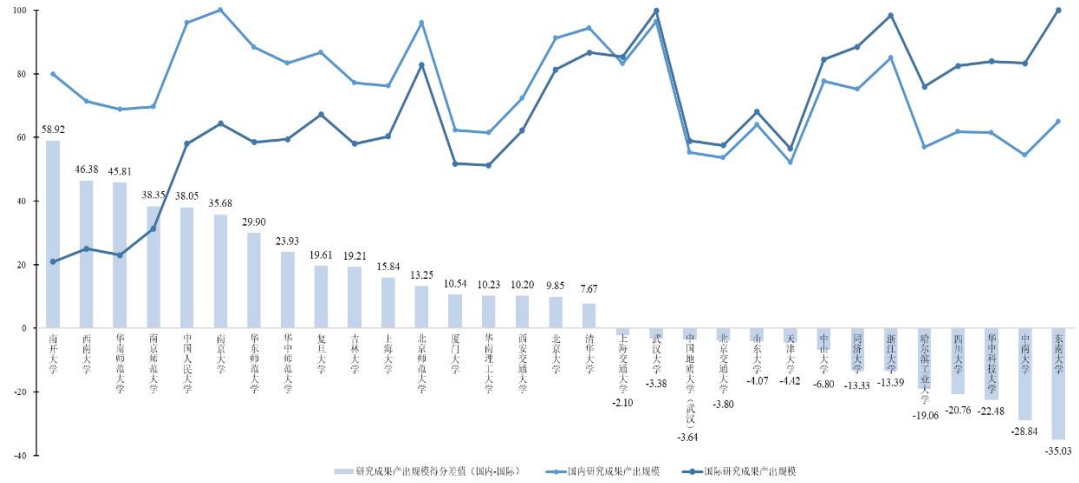


图 10-3 中国高校“国内-国际”研究成果产出规模差值分布

（3）学术影响力

“学术影响力”是用于全面评估高校在 AI4SSH 领域所产出的研究成果在学术界引起的关注程度及其知识价值的核心指标。它通过量化研究成果被同行引用、参考和借鉴的广度与深度，反映其原创性、重要性以及对学科发展的推动作用。学术影响力指标本质上反映了学术成果与学科发展需求及学术共同体认可之间的契合度。

本报告将“学术影响力”作为二级指标，并下设“国际学术影响力”与“国内学术影响力”两个三级指标，二者共同构成了对科研产出质量与价值的多维评价，对科研评估、资源布局及学科战略规划具有重要参考意义。

国际学术影响力：中国高校 AI4SSH 领域研究的国际学术影响力结果如表 3-5 所示，四川大学学术影响力位居第一，哈尔滨工业大学、武汉大学、东南大学、清华大学、华中科技大学、浙江大学、电子科技大学、北京师范大学、重庆大学分数均超过 90，表明其研究成果已

获得国际学界的广泛关注与认可。

从高校类型分析，综合类大学在排名中占据半数以上席位，反映出其研究成果与国际学术前沿的契合度较高。同时，哈尔滨工业大学、东南大学、华中科技大学等工科特色鲜明的院校排名普遍靠前，其固有的技术优势使其更易产出符合国际学术范式的高质量交叉成果。此外，部分专业类院校聚焦于“AI+优势专业”的融合路径，也在国际领域形成了特色贡献。

表 10-9 中国高校国际学术影响力

高校	国际学术影响力	高校	国际学术影响力
北京大学	83.31	上海交通大学	88.16
北京工业大学	73.17	四川大学	100.00
北京交通大学	71.30	同济大学	83.62
北京师范大学	92.52	武汉大学	97.69
电子科技大学	94.56	西安交通大学	79.98
东南大学	97.56	西南财经大学	75.60
复旦大学	82.77	浙江大学	95.02
哈尔滨工业大学	98.74	郑州大学	72.03
华中科技大学	95.24	中国人民大学	78.80
暨南大学	73.58	中南大学	88.89
清华大学	96.64	中山大学	89.92
山东大学	79.03	重庆大学	90.75
上海大学	74.13		

注：本表按高校名称首字母排序

国内学术影响力：南京大学、北京师范大学、武汉大学、清华大学、复旦大学展现出高水平的学术影响力。传统顶尖综合类高校领跑国内影响力榜单，这说明其研究成果较好地对接了本土人文社科的全面需求，同时也表明国内 AI4SSH 的学术影响力与高校的整体学术声誉密切相关。排名高校之间的最大分差超过 55 分，揭示了国内高校在该领域的影响力存在显著不平衡，成果认可度与采纳程度分布不均。

更多免费行业报告 并购家 www.ipoipo.net

表 10-10 中国高校国内学术影响力

高校	国内学术影响力	高校	国内学术影响力
北京大学	78.72	山东大学	74.52
北京师范大学	84.36	上海大学	72.86
东南大学	54.03	上海交通大学	77.80
复旦大学	79.92	苏州大学	44.52
广州大学	46.01	同济大学	53.70
华东师范大学	78.47	武汉大学	83.48
华中师范大学	76.26	西安交通大学	66.26
吉林大学	68.30	西南大学	55.95
南京大学	100.00	浙江大学	77.74
南京师范大学	58.21	中国人民大学	78.86
南开大学	74.61	中山大学	74.43
清华大学	83.13	重庆大学	51.37
厦门大学	53.23		

注：本表按高校名称首字母排序

综合学术影响力：学术影响力是衡量机构在 AI4SSH 领域研究成果被同行参考、借鉴与认可程度的综合性二级指标（权重为 0.4），其下包含两个三级指标：国际学术影响力（权重为 0.15）与国内学术影响力（权重为 0.25）。国内学术影响力对于提升我国 AI4SSH 领域的整体学术水平、培养本土研究力量、解决本土社会科学与人文问题具有更为直接和关键的作用，因此在指标体系中被赋予了更高的权重。学术影响力计算公式如下：

$$I_{total} = I_{total}^{ins} * 0.375 + I_{total}^{dom} * 0.625$$

结果显示，综合类高校在两类影响力排名中均占据半数以上席位，其中武汉大学、清华大学、北京师范大学、浙江大学、南京大学、上海交通大学及复旦大学等校在国内外均保持较高影响力水平，呈现出较为明显的“双向协同”特征。理工类院校依托学科交叉背景，其国际影响力曲线表现出较快的上升趋势，逐步接近全球高影响力区间的

阈值。北京师范大学、华东师范大学、华中师范大学等师范类院校的影响力分布则呈现“国内高位、国际相对低位”的态势，一定程度上反映出其研究议题与成果扩散具有较显著的本土化属性。

国际与国内学术影响力数据对比：通过“国际–国内影响力差值”可进一步揭示不同高校的影响力结构特征。部分综合类高校，如苏州大学、中国人民大学、上海大学及复旦大学等，其差值多集中于±5 分以内，体现出在国内外学术体系间发展相对均衡的格局。师范类院校的差值普遍为正，部分峰值超过 30 分，体现出“国内主导、国际跟随”的扩散特点。理工类院校的差值则多为负值，表明其成果在国际学术界获得的关注度整体上高于国内，其研究范式与国际主流 AI 学术话语体系具有较高的契合度。

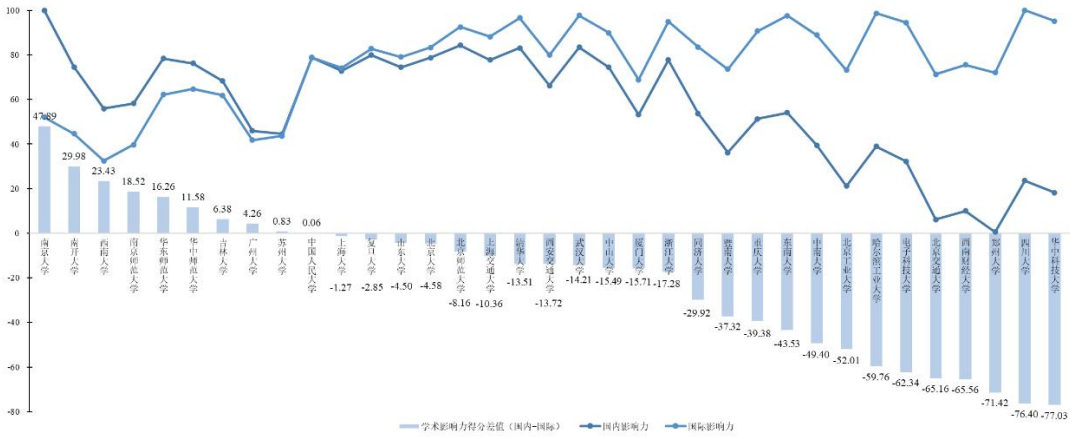


图 10-4 中国高校“国内-国际”学术影响力差值分布

（4）中国高校研究核心能力比较分析

亮点：头部效应初显，理工融合优先。从整体表现上来看，中国高校在研究核心能力上呈现出一定的头部聚集与聚类分化特征。一方面，以清华大学、复旦大学等为代表的综合类高校在学术产出规模与

更多免费行业报告 并购家 www.ipoipo.net

影响力方面实现了国内外“双高协同”，展现出较强的资源整合能力与学科布局优势。另一方面，理工类及行业特色型院校在 AI 渗透率指标上表现突出，说明其在推动“AI+人文社科”的范式转型中已取得积极进展，尤其是在对接国际学术前沿、提升研究可对话性方面具备一定先发优势。

不足：发展结构不平衡，内外联动待提升。与此同时，研究核心能力维度也暴露出若干结构性问题。首先，在“研究融合度”方面，仅不足三成的高校实现了国内外 AI 渗透率的均衡发展，多数院校呈现“单极倾斜”态势，反映出其在国际与本土学术场域之间的协同能力仍较薄弱。其次，在“学术产出规模”上，尽管头部高校体量庞大，但中后部高校普遍处于“技术引入期”，尚未实现从数量积累向质量跃升的关键突破，成果的创新性与引领性仍有待提升。此外，“学术影响力”的国内外分布亦呈现明显分化：师范类院校国内影响力较强但国际显示度较弱，理工类院校则相反，这表明研究议题的本土化与国际化之间仍存在一定的张力，尚未形成有效的双向循环。

启示：强化双向嵌入，构建分类发展路径：面向未来，高校在推进 AI4SSH 研究时应进一步强化“双向嵌入”意识：既要立足本土问题导向，深耕中国式现代化进程中的重大议题，又要主动融入全球学术网络，提升研究的国际对话能力与理论贡献水平。同时，应鼓励不同类型高校走差异化发展之路——综合类院校可依托多学科优势，推动 AI 与人文社科的深度融合与系统创新；理工类院校需加强人文社科学术积累，避免陷入“技术工具主义”倾向；师范类院校则可聚焦“AI+

教育”特色方向，打造具有辨识度的研究赛道，并逐步拓展国际传播路径，提升学术辐射力。

2. 发展创新潜力

“发展创新潜力”旨在从组织、创新与生态三个层面，前瞻性评估高校在 AI4SSH 领域可持续发展的能力与结构性优势。本维度下设“组织领导力”“原始创新力”与“生态构建力”三项二级指标，并分别对应“国际科研合作组织力”“国际交叉创新潜力”与“实体化建制水平”三项三级指标。三者分别聚焦于高校在全球科研网络中的主导地位、知识生产的源头创新能力，以及支撑交叉学科长效发展的制度性基础，共同构成了从“组织引领”到“知识创造”再到“生态固化”的潜力评价链条，以系统辨识高校在推动该领域纵深发展中所蕴含的成长动能与战略韧性。

（1）组织领导力

组织领导力从科研主导力的视角，衡量高校在全球 AI4SSH 合作网络中的引领与协调能力。其下设的三级指标“国际科研合作组织力”，通过计算机构在跨国合作成果中担任通讯作者单位的比例，量化评估其在合作中的主导地位与议程设置能力，从而反映其凝聚国际资源、引领前沿方向的实际效能。

苏州大学、北京理工大学、东南大学与重庆大学在 AI4SSH 跨国合作中展现出核心引领作用，构成“引领型”梯队；中国科学技术大学、复旦大学、四川大学等高校得分介于 80–90 分之间，形成“优质

更多免费行业报告 并购家 www.ipoipo.net

型”梯队，其国际合作已建立起稳定机制并具备较强的组织协调能力；西安交通大学、南京大学、北京航空航天大学等高校得分超过 70 分，处于“发展型”阶段，正积极融入并提升在全球合作网络中的协同水平。

表 10-11 中国高校国际科研合作组织力

高校	国际科研合作组织力	高校	国际科研合作组织力
北京工业大学	81.70	吉林大学	76.06
北京航空航天大学	77.78	南京大学	78.70
北京交通大学	87.96	上海大学	81.70
北京理工大学	95.79	上海交通大学	74.53
电子科技大学	80.75	四川大学	88.57
东南大学	95.10	苏州大学	100.00
复旦大学	89.29	武汉大学	74.16
哈尔滨工业大学	87.87	西安交通大学	79.37
合肥工业大学	82.30	西南交通大学	82.20
湖南大学	80.13	中国科学技术大学	89.48
华南师范大学	81.02	中南大学	77.41
华中科技大学	84.75	重庆大学	92.59
华中师范大学	83.33		

注：本表按高校名称首字母排序

（2）原始创新力

原始创新力以国际交叉创新潜力为具象指标，旨在衡量高校在 AI4SSH 领域进行源头知识创造与突破性探索的潜力。其下设的三级指标“国际交叉创新潜力”，基于“突破性研究常源于非典型的跨学科知识重组”这一认识，通过量化机构发表论文中参考文献所涉期刊组合的非典型性，来表征其研究的知识新颖性与学科交叉融合深度，从而揭示其原始创新潜力。

天津大学、电子科技大学、北京理工大学、吉林大学、武汉大学、郑州大学及清华大学在创新引领性上表现突出。整体分布上，中后部

相邻高校间的创新度分差呈逐渐扩大趋势，反映出创新能力的分布并非均匀，头部机构的领先优势较为明显。

表 10-12 中国高校国际交叉创新潜力

高校	国际交叉创新潜力	高校	国际交叉创新潜力
北京大学	47.80	天津大学	100.00
北京理工大学	97.01	同济大学	89.02
北京师范大学	37.44	武汉大学	92.42
电子科技大学	98.35	西安交通大学	76.51
东南大学	81.05	西南财经大学	65.07
复旦大学	56.66	西南大学	73.88
华东师范大学	46.70	西南交通大学	66.18
吉林大学	95.86	浙江大学	82.41
南京大学	57.39	郑州大学	91.56
清华大学	91.42	中国地质大学（武汉）	42.33
厦门大学	76.37	中国科学技术大学	81.85
上海交通大学	34.85	中国矿业大学	54.91
苏州大学	44.08		

注：本表按高校名称首字母排序

（3）生态构建力

生态构建力旨在衡量高校为 AI4SSH 深度融合提供制度化、可持续性支撑的能力，重点关注其在组织架构与资源保障方面的长期布局。本维度通过三级指标“实体化建制水平”进行量化评估，该指标以高校是否设立并持续运行由高校乃至国家部委层面支持的实体研究机构为核心判据。这类实体机构标志着学科交叉从倡议走向建制化发展，是评估高校战略投入与生态构建能力的关键观察点。

在实践层面，由教育部、科技部、网信办等部委支持或高校重点布局的实体机构已成为领域发展的重要支柱。其中，复旦大学国家发展与智能治理综合实验室、复旦大学国家智能评价与治理实验基地、中国人民大学智慧治理学院、清华大学人工智能国际治理研究院、浙

更多免费行业报告 并购家 www.ipoipo.net

江大学国家智能社会治理实验基地等，共同构成了首批高显示度的国家级或校级示范基地。这些机构在战略定位、资源投入与体制机制上进行了系统性设计，已成为全国 AI4SSH 建制化发展的先行者与关键标杆。

评估结果显示，北京大学、北京师范大学、上海交通大学、华中师范大学、中山大学等 40 余所高校也已陆续布局实体研究机构，初步形成“京—沪—杭—粤”等多点联动、层次分明的生态骨架。这标志着我国 AI4SSH 交叉研究正从零散探索阶段，逐步转向可持续、可迭代的体系化发展新阶段。

这类实体平台不仅是凝聚方向、整合资源、开展有组织科研的核心载体，更是推动科研范式转型、促进成果转化与政策对接的重要枢纽。因此，“实体化建制水平”不仅刻画了高校当前在战略层面的投入力度，也预示了其构建长效交叉研究生态、引领领域未来发展的制度潜能。

（4）中国高校发展创新潜力比较分析

亮点：建制化进程加速，合作网络日益紧密。从发展创新潜力的整体表现来看，中国高校在国际科研合作的组织协调与知识重组的源头探索方面已取得初步进展。部分高校在国际合作成果中担任通讯作者的比例较高，表明其已具备一定的议题设置能力与资源整合实力，在全球 AI4SSH 合作网络中逐步从“参与”走向“主导”。同时，“国际交叉创新潜力”指标显示，少数头部高校在跨学科知识重组方面开展了积极探索，其研究成果的参考文献组合呈现较强的非典型特征，为产

生突破性创新成果积累了知识基础。

不足：创新能力分化，生态建构仍处于起步阶段。

与此同时，发展创新潜力维度亦暴露出结构性与制度性短板。从“国际交叉创新潜力”来看，头部高校与中后部高校之间的差距呈扩大趋势，大量院校仍以跟随性、应用性研究为主，在理论贡献与范式突破层面尚未形成有效积累。创新能力分布的“断层”现象较为突出。在“生态构建力”层面，尽管复旦大学国家发展与智能治理综合实验室、中国人民大学智慧治理学院、清华大学人工智能国际治理研究院等 AI4SSH 代表性实体化机构已率先建立，但从全国范围来看，实体化建制仍呈现“点状分布”特征，多数高校尚未形成稳定、制度化的交叉研究平台，难以支撑 AI4SSH 领域的长效发展。

启示：夯实制度基石，构建分层创新格局。面向未来，高校应系统推进“组织引领—创新驱动—生态筑基”三位一体的发展路径。首先，鼓励已具备国际组织能力的高水平院校发挥“领头羊”作用，带动更多高校融入全球创新网络，提升整体合作层级。其次，加大对原创性、交叉性研究的制度供给，通过设立专项基金、优化学术评价机制等方式，为“非典型”知识探索提供包容空间，激发源头创新能力。再次，加快实体化研究机构建设步伐，推动 AI4SSH 从“项目制”走向“建制化”，以平台整合资源、稳定队伍、凝练方向，为交叉研究的可持续发展提供长效制度保障。

3. 社会传播能力

更多免费行业报告 并购家 www.ipoipo.net

社会传播能力旨在评估高校在 AI4SSH 领域的研究成果、专家观点与重要进展突破学术边界、进入公共话语体系并服务社会发展的能力。本维度下设二级指标“公共影响力”，并通过三级指标“公共传播影响力”进行量化测度，以此系统考察高校在主流舆论场中的传播能见度、政策关联度与社会参与水平。

公共传播影响力聚焦于高校在中央主流媒体（如人民网、新华网、央视网等）中获得的原创报道强度与传播广度，是其研究成果获得社会关注、对接国家战略并参与公共讨论的关键体现。评估显示，清华大学、复旦大学与北京大学在该指标上表现突出，其 AI4SSH 领域的机构建设、重要成果及专家观点频繁获得公共传播的深入报道，反映出较高的公共能见度与较强的政策关联性。

从分布结果来看，传播能力呈现明显的梯度特征：头部高校优势显著，中后段高校影响力相对集中且差异较小。此外，媒体报道强度与高校类型及其实质性建设内容之间存在一定关联，具备显著交叉特色或重大进展的高校更易获得舆论关注。这一指标不仅刻画了高校成果的社会扩散效能，也为其提升公共形象、增强社会服务能力提供了可观测的依据。

亮点：头部高校话语优势显著，政策关联度较高。在“社会传播能力”维度，以复旦大学、清华大学、北京大学为代表的顶尖高校，依托其在 AI4SSH 领域的机构建设、重大成果与专家观点，获得了公共传播的高频次深度报道，展现出较强的公共能见度与政策影响力。这些高校不仅在前沿议题的舆论场中占据主导地位，也在一定程度上推动

了公众对人工智能与人文社科交叉领域的认知深化。

不足：传播梯度分化明显，公共转化有待加强。从整体分布来看，社会传播能力呈现显著的“头部集中”特征。头部高校与中后部高校之间传播声量差距悬殊，而中后部高校内部差异相对较小，反映出多数院校在成果转化与社会传播方面仍处于“有研究无声音”的阶段。媒体报道强度与高校的实质性交叉建设密切相关，具备机构平台、标志性成果或权威专家的高校更易获得舆论关注，而大量高校尚未建立起有效的外部沟通机制，成果的社会影响力仍局限在学术圈层内部。

启示：完善传播机制，增强社会赋能效应。高校应着力提升“成果转化+公共沟通”的双重能力。一方面，需强化 AI4SSH 研究成果与国家战略、区域发展的对接，增强研究的现实回应性与社会服务价值；另一方面，应构建常态化、专业化的媒体沟通机制，主动策划传播议题、培育专家型传播团队、拓展多元传播渠道，提升在权威舆论场中的可见度与话语权。同时，鼓励中后部高校立足自身特色与区域优势，打造差异化传播亮点，逐步缩小与头部高校之间的传播能力差距，推动形成更加均衡、多元的社会传播格局。

4. 中国高校 AI4SSH 指数分项指标分析总结

综合第三章对“研究核心能力”“发展创新潜力”“社会传播能力”三个维度的系统分析可以看出，当前中国高校在 AI4SSH 领域呈现出“分化与协同并存、转型与跃升交织”的总体特征。该领域已取得阶段性显著成果，但也面临较为突出的结构性挑战，正处于从“点的突

更多免费行业报告 并购家 www.ipoipo.net

破”向“面的繁荣”跨越的关键时期。

（1）发展成效初显，特色路径渐趋明晰

部分顶尖高校已在学术产出规模、国内外影响力、国际科研合作、实体化机构建设及公共传播等方面形成“多轮驱动”的领先格局，成为引领我国 AI4SSH 发展的核心力量。其中，复旦大学、清华大学、武汉大学等高校在多维度表现均衡且突出，展现出系统性的竞争优势。与此同时，理工类、师范类及行业特色类院校也分别在 AI 渗透率、本土学术影响力、专业领域融合等维度上形成差异化优势，初步构建起“多点开花、分层推进”的发展态势。

（2）结构性矛盾依然存在，均衡发展任重道远

中国高校在 AI 与人文社科交叉研究中仍存在多方面问题：在研究融合度上，仅少数高校实现国内外双向协同，多数院校呈现“单极倾斜”特征；在学术产出上，头部高校规模领先，但中后部高校多处于技术引入阶段，量质分化明显；在创新能力上，源头性突破仍集中于少数头部院校，中后部高校创新度差距呈扩大趋势；在生态构建上，实体化机构建设尚处“点状分布”阶段，多数高校缺乏稳定的制度支撑；在社会传播上，声量高度集中于顶尖高校，大量院校“有成果无声音”的现象较为普遍。上述分化与不足，既反映出我国 AI4SSH 发展仍处于从“探索期”向“成长期”过渡的阶段特征，也揭示了未来高质量发展的关键着力方向。

（3）坚持系统观念，推动高质量跃升

面向未来，中国高校应立足自身定位，走差异化、协同化、可持

续的发展道路。AI4SSH 研究不应是技术的简单叠加，而应是范式的深刻重构。综合类高校应强化“双向嵌入”意识，在国内外学术体系中发挥桥梁与枢纽作用；理工类院校需加强人文社科学术积淀，避免陷入技术工具主义倾向；师范类院校可依托教育场景优势，打造特色赛道并逐步提升国际传播能力。与此同时，国家层面应加强顶层设计，系统推进实体化机构建设、原创性研究支持与社会传播能力提升，构建“创新研究—场景赋能—价值传播”三位一体的 AI4SSH 发展生态，助力中国高校在全球人文社科智能化转型进程中实现从“跟跑”到“并跑”乃至“领跑”的历史性跨越。

更多免费行业报告 并购家 www.ipoipo.net

10.4 主要结论与未来展望

1. 主要结论

本指数通过构建涵盖“研究核心能力”“发展创新潜力”与“社会传播能力”的综合评价体系，首次对中国高校在人工智能与人文社科交叉领域（AI4SSH）进行了全景式、多层次的系统评估。评估发现，我国在该前沿交叉领域已形成“体系初建、梯队分明、路径分化、生态萌发”的发展格局。从整体表现来看，我国高校在研究产出规模与本土融合方面进展显著，呈现出强劲的增长势头与广泛的建设基础；但在国际学术影响力、源头创新能力以及社会传播声量方面仍存在明显短板，且各维度内部均呈现出不同程度的结构性分化。这些特征不仅反映出高校在不同融合阶段的现实差异，也揭示了交叉学科建设从规模积累向质量引领、从学术探索向生态构建演进的关键命题，为优化学科布局、引导资源投入、提升融合效能提供了实证依据与决策参考。

（1）梯队格局清晰，呈现结构化发展态势

评估结果显示，中国高校在 AI4SSH 领域的综合能力呈现出层次分明、梯度发展的分布

态势。依据综合指数得分，可将其划分为特征鲜明的三个梯队：以清华大学、武汉大学、复旦大学、北京师范大学、浙江大学、北京大学等为代表的第一梯队（得分≥72 分），在成果规模、融合深度与学术影响力上均建立起系统性优势，形成了较为成熟的交叉研究体系。

复旦大学作为第一梯队的核心成员，在多维度表现均衡且突出，尤其在发展创新潜力与社会传播能力方面展现出强劲实力。得分介于 60 至 72 分之间的第二梯队高校数量集中，显示出该领域已在国内具备广泛的建设基础与成长潜力，正处于快速积累与特色形成的关键期。从主导力量来看，“双一流”建设高校，尤其是综合性大学与具备强劲工科背景的研究型大学，在榜单中占据绝对主导地位，凸显了深厚的既有学科基础与强大的资源整合能力对该交叉领域发展的根本性支撑。

（2）研究核心能力呈现“规模引领、本土融合深入”的特征

在衡量科研硬实力的“研究核心能力”维度，评估结果显示以下突出特征：其一，学术产出规模呈现国内外异步增长态势，国内成果增速显著高于国际，反映出本土学术生态对 AI 方法工具的积极响应与应用场景的快速拓展；国际成果则保持稳健增长，体现我国学者在国际学术体系中的持续融入与竞争力积淀。但值得注意的是，国际成果的绝对数量与增速仍落后于国内，表明我国高校在国际学术话语权构建方面尚存提升空间。其二，研究融合度呈现“本土高于国际”的明显趋势，表明当前阶段的融合实践更侧重于本土化学术语境与实际需求，这也提示在国际学术话语与范式影响力方面仍有提升空间。其三，从空间布局来看，高水平研究成果与学术影响力高度集聚于京津冀、长三角、粤港澳及华中等高等教育与数字经济发展水平较高的区域，体现出交叉学科建设与区域创新体系之间的紧密关联，但中西部地区高校的参与度和表现仍有待加强，区域分化较为显著。

更多免费行业报告 并购家 www.ipoipo.net

（3）发展创新潜力分化，实体建制成关键

在表征长远发展能力的“发展创新潜力”维度，评估发现各高校在组织协同、源头创新与生态构建方面存在不均衡现象。部分高校如复旦大学等已在国际科研合作网络中展现出较强的议题设置与协同主导能力，其“国际科研合作组织力”指标位居前列，体现出在全球化科研语境中的引领地位。在表征知识重组与原始创新潜力的指标上，头部高校优势显著，这表明深层次方法论创新仍集中于少数机构，创新能力呈现“头部集中、尾部发散”的格局，中后部高校创新度分差逐渐拉大，表明创新能力分化加剧。尤为重要的是，以复旦大学国家发展与智能治理综合实验室、中国人民大学智慧治理学院等为代表的实体化交叉研究机构相继建立，标志着该领域正从依托项目与人员的非建制化探索阶段，转向以常设性平台为支撑的体系化、制度化发展阶段，为交叉学科的可持续发展奠定了必要的组织与制度基础。这一趋势显示出少数先行高校在生态构建层面的战略远见，同时也提示多数高校在实体化建制方面尚处于起步或空白阶段，制度性支撑不足成为制约其发展潜力释放的重要因素。

（4）传播能力圈层化，公共影响待提升

在考察成果溢出与公众参与的“社会传播能力”维度，研究发现公共传播影响力存在明显差异。清华大学、上海交通大学、深圳大学、复旦大学等高校在主流社交平台展现出较高传播热度与公众关注度，一批高校依托学科特色与专业运营能力，实现了 AI4SSH 研究成果高质量、破圈式传播，有效扩大了社会影响力与公众认知度。然而，部

分院校传播能见度与受众触达水平仍相对有限，多平台协同传播体系尚不完善，学术成果向公共话语转化的程度不足。未来需构建适配不同平台功能定位的传播策略，提升成果转化与公共沟通能力。

2. 未来展望

基于上述研究发现，中国高校 AI4SSH 交叉领域在取得阶段性显著成效的同时，也面临着从“规模扩张”向“质量引领”、从“本土融合”向“国际对话”、从“学术探索”向“生态构建”转型升级的关键挑战。未来深化发展应重点关注以下方向：

（1）推动范式融合从“应用跟随”到“创新引领”的转变

未来需在巩固当前技术应用成果的基础上，大力鼓励和资助具有理论原创性的探索。高校与科研管理机构应设立鼓励跨学科知识深度重组的专项计划，引导学者不仅利用 AI 工具解决人文社科问题，更致力于从交叉实践中凝练新理论、新方法，乃至催生新的子学科领域。特别是要支持在人工智能伦理、算法治理、数字人文基础模型、社会科学仿真模拟等关键方向上的前沿探索，力争产出能够设定国际议题、引领学术议程的突破性成果。

（2）构建“内外联动、制度保障”的可持续发展生态

一方面，应实施更加主动的国际化战略，通过牵头或深度参与国际大科学计划、发起国际学术组织、创办高水平英文期刊等方式，系统性提升在国际合作网络中的组织力与话语权，促进国内外研究能力的均衡发展与双向赋能。另一方面，应总结推广先行高校实体化建制

更多免费行业报告 并购家 www.ipoipo.net

经验，鼓励更多有条件的院校布局建设高水平的实体交叉研究机构，在人事管理、经费支持、评价考核等方面进行契合交叉学科规律的制度创新，为长效发展提供稳定的组织保障和资源支撑。

（3）强化“需求驱动、交叉赋能”的社会服务与话语体系

交叉研究的生命力最终体现在对社会发展与国家战略的实际贡献上。未来应推动 AI4SSH 研究更紧密地对接数字中国建设、国家治理现代化、文化数字化战略等重大需求，围绕智慧治理、数字人文遗产保护、舆情分析与引导、社会风险模拟预警等场景开展有组织的科研攻关。同时，高校应主动提升科研成果的传播与转化能力，通过智库报告、公共科普、政策建言等多种形式，将学术语言转化为公众理解和政策制定者需要的社会语言，构建兼具中国特色与国际视野的交叉学科话语体系。

（4）培育“厚基础、强交叉、善协同”的复合型人才队伍

人才是交叉学科发展的第一资源。应打破传统学科壁垒，在研究生培养层面大力推广双导师制、交叉学科课程模块与学位项目。在科研队伍建设中，应探索建立适应交叉研究特点的人才评价与激励机制，认可并奖励在跨学科合作、团队攻关中的贡献，吸引和汇聚计算机科学、人文社科等多领域优秀学者形成稳定协同团队，为领域的持续创新储备核心人力资本。

总之，推动 AI4SSH 的深度融合与高质量发展，是一项长期性、系统性工程。它需要学术界、教育界、政策制定者与社会各界的共同关注与持续努力，通过不断优化学科生态、创新制度安排、深化国际

合作，方能使中国在此全球新兴前沿领域不仅实现“量的积累”，更完成“质的跨越”，为构建人类数字文明新形态贡献中国智慧与中国方案。

更多免费行业报告 并购家 www.ipoipo.net

参考文献

[1] 陈志敏, 吴力波. 未来已来：2025 人文社会科学智能发展蓝皮书[M]. 上海: 复旦大学出版社, 2025.

[2] Schmallenbach, L., Bärnighausen, T.W., & Lerchenmueller, M.J. The global geography of artificial intelligence in life science research[J]. Nature Communications, 2024, 15: 7527. https://doi.org/10.1038/s41467-024-51714-x.

[3] Gao, J., & Wang, D. Quantifying the use and potential benefits of artificial intelligence in scientific research[J]. Nature Human Behaviour, 2024, 8: 2281–2292. https://doi.org/10.1038/s41562-024-02020-5.

[4] Baruffaldi, S., et al. Identifying and measuring developments in artificial intelligence: Making the impossible possible[R]. OECD Science, Technology and Industry Working Papers, No. 2020/05, 2020: 1–68.

[5] Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. Atypical combinations and scientific impact[J]. Science, 2013, 342(6157): 468–472.

[6] 杨杰, 柳美君, 步一, 等. 组合新颖性和学科新颖性能否提升论文的技术影响？基于直接和间接技术影响力的双重视角[J]. 情报学报, 2025, 44(03): 325–338.

[7] Web of Science Core Collection（SSCI/AHCI），AI4SSH 论文数据与引文数据检索来源（检索式见报告附录）。

[8] 中国知网 CNKI（CSSCI），AI4SSH 中文论文与引文数据检索来源。

[9] OECD. AI 术语词表（214 个 AI 术语）与相关测度研究（用于构建检索关键词集与检索策略）。

[10] Microsoft Academic Graph（MAG）学科标签体系（用于辅助构建核心 AI 领域标签集）。

[11] 高校官网信息（组织机构/院系设置等栏目公开清单），用于“实体化建制水平”等指标的统计与核验。

[12] 微信、微博、抖音等社交媒体平台公开传播数据，用于“公共传播影响力”指标的检索与计数（按报告设定的时间范围、检索式和平台统计口径执行）。

第十一章 AI4SSH 基础设施和代表性项目建设现状与展望[①]

人工智能对科研活动的影响，正从单点工具应用扩展为对研究基础设施、研究流程、创新组织方式的系统性重塑。在这一背景下，AI4SSH 基础设施的建设已成为高校推进人文社会科学数智化转型的关键抓手。它既决定数据、算力、模型与工具能否形成可持续的校内公共供给，也决定跨学科协作能否从项目合作走向长期建制。

本章围绕“数据底座与流通机制—融合算力与研究装置—模型/智能体与工具链”三条主线，在系统梳理国内外政策导向与概念演进的基础上，结合国内高校公开实践，揭示出 AI4SSH 基础设施从“资源汇聚”走向“可计算组织”、从“计算支撑”走向“可实验研究环境”、从“模型接入”走向“研究全过程能力体系”的阶段性跃迁。

AI4SSH 基础设施并非硬件堆砌。对人文社会科学而言，其核心在于形成可复用、可互操作、可审计的公共能力：数据的可计算与可流通、算力与环境的可用与可持续、模型与工具的可控与可扩展，以及支撑人机协同知识生产的组织机制与治理规则。

[①] 本章由复旦大学大数据学院张计龙研究馆员牵头编写。
张计龙研究馆员目前担任教育部教指委委员、复旦大学国家发展与智能治理综合实验室副主任、上海市科研领域大数据联合创新实验室主任、复旦—阿法迪共建智慧图书馆学研究中心常务副主任等，研究领域包括智慧图书馆、科学数据管理等。

更多免费行业报告 并购家 www.ipoipo.net

本章提要

- AI4SSH 基础设施建设已成为全球与我国科技战略的重要组成部分，国际战略引领、国家顶层统筹、地方政策响应与高校平台落地形成协同联动的推进格局。
- 概念层面，研究基础设施正在向“数智学术基础设施”演进。其核心突破在于以“数据—算力—模型和工具链”的系统集成为基础，将平台科研、人机协同与知识共生转化为基础设施能力，并以公共性、可复用性与可持续性为发展导向。
- 数据底座建设三大趋势：由分散资源库走向校级底座与平台群整合；数据对象由结构化统计扩展到多源异构与多模态知识对象；数据流通由静态共享走向嵌入全流程的机制化流转和 AI-Ready。
- 算力基础设施由单一 HPC 平台走向超算、智算与通算融合的综合底座；社会仿真与多智能体实验平台兴起，为相关研究提供了“可实验”的研究空间。
- 模型、智能体与工具链正构成面向研究全过程的能力体系：通用模型校内化重组、垂域模型学科化深入、工具链平台化供给，共同推动 AI 能力从“可用”向“可嵌入、可复现、可治理”转变。
- 平台同质化、算力统计口径不一、资源可及性不足、伦理与合规边界不清、平台贡献评价缺位等问题仍制约基础设施的规模化外溢。未来应以互操作标准、评测与审计机制、服务化团队与可持续治理规则为抓手，推动 AI4SSH 基础设施建设从规模扩张走向高质量公共供给。

11.1 AI4SSH 的基础设施[①]

当前，人工智能正系统性地重塑人文社会科学（AI4SSH）的研究范式与知识生产方式，核心是构建超越传统数字资源平台的新型“数智学术基础设施”。本节从数据治理与科研范式转型的交叉视角切入，首先审视国际科技战略动向，思考各国如何将数据、算力与模型确立为科研基础设施的关键支柱，并梳理国内从国家顶层设计到部委专项、再到高校哲学社会科学实验室的多层级协同布局。在此基础上，本节进一步厘清了从广义“研究基础设施”“数字人文数据基础设施”向“数智学术基础设施”嬗变的学理脉络，揭示出 AI 时代社科类基础设施的核心特征已超越资源数字化本身，转向聚焦研究对象的可计算性提升、“数据—算力—模型和工具链”的系统集成，以及对“人机共生”与“人智协同”知识创新模式的支撑。据此，本节将 AI4SSH 基础设施界定为：支撑人工智能深度介入人文社会科学研究、教学及知识服务的多模态数据底座、智能化计算环境、领域模型能力与集成化工具链条及其协同机制的有机整体，并由此确立后文的分析框架与内容结构。

1. AI4SSH 基础设施建设的战略布局

人工智能与科学研究的深度融合已成为当前全球科技战略布局的重要方向。2025 年下半年以来，欧盟委员会发布《欧洲科学人工智

[①] 本节由复旦大学伏安娜、张计龙合作撰写。
伏安娜，复旦大学图书馆副研究馆员、上海市科研领域大数据联合创新实验室主任助理，主要研究领域是科学数据管理、数据素养教育等。

更多免费行业报告 并购家 www.ipoipo.net

能战略》（A European Strategy for Artificial Intelligence in Science），提出建设“欧洲科学人工智能资源中心”（RAISE，Resource for AI Science in Europe），统筹算力、数据、人才和研究资金等关键资源[1]。英国政府发布《人工智能赋能科学战略》（AI for Science Strategy），明确将数据、算力、人才和文化作为支撑 AI 驱动科学研究的关键支柱[2]。美国白宫推出“创世纪计划”（Genesis Mission），提出利用人工智能重塑科研范式，以加快科学发现进程[3]。相关战略动向表明，人工智能对科研活动的影响已超越单一技术应用范畴，正系统性地推动科研基础设施、研究流程与创新组织方式的深刻变革。

在此背景下，我国也进入政策密集出台的关键阶段。2025 年 8 月，国务院印发《关于深入实施“人工智能+”行动的意见》[4]，首次将“创新哲学社会科学研究方法”纳入国家“人工智能+”重点行动，明确提出“推动哲学社会科学研究方法向人机协同模式转变，探索建立适应人工智能时代的新型哲学社会科学研究组织形式”，强化“数据供给创新”“智能算力统筹”“模型基础能力”等八大基础支撑，这为 AI4SSH 基础设施建设提供了顶层政策框架。同年 12 月，国家数据局发布《关于加强数据科技创新的实施意见》[5]，系统部署数据科技创新体系建设，提出“加快建设面向科学研究的数据基础设施”，支持打造“Data for Science”协同创新中心，促进科学数据高效流通和开发利用，推进全国一体化算力网和数据流通利用体系建设，为 AI4SSH 依赖的跨机构数据共享和大规模计算分析提供了制度保障。2026 年 3 月发布的“十五五”规划纲要草案摘要[6]，进一步将“强化算力算法

数据高效供给”“全方位推进数智技术赋能”并列为重要方向。这些多层级、多维度的政策部署，正加速推动 AI4SSH 从宏观愿景向可落地的基础设施与组织实践转化。

在部委专项方面，国家科技管理部门在 AI for Science 领域持续作出专项部署。2018 年起，科技部围绕科技创新 2030—“新一代人工智能”重大项目多次发布申报指南[10][11]；2023 年，科技部会同国家自然科学基金委启动实施“人工智能驱动的科学研究”专项[12]。尽管上述部署主要面向自然科学，但其构建的数据资源体系、算法模型平台和科研协同机制，可构成理解 AI4SSH 基础设施建设的关键参照。

对人文社会科学领域而言，更直接的制度依托，是教育部哲学社会科学实验室、文化和旅游部重点实验室等创新平台。《教育强国建设规划纲要（2024—2035 年）》[13]明确提出“推进哲学社会科学创新平台和创新团队建设…建好高校哲学社会科学实验室”，将其纳入构建中国哲学社会科学自主知识体系的重要部署。截至目前，教育部已分两批认定 55 个哲学社会科学实验室；自 2014 年起，由高校承建的文化和旅游部重点实验室累计布局 23 个。两类实验室在数据积累、技术研发、平台建设和人才培养等核心任务及基础条件要求上具有诸多共性，共同体现出人文社会科学创新平台从传统学科组织形态，转向以数据、技术和交叉协同为支撑的新型科研基础设施。

在地方层面，北京[7]、上海[8]、江苏[9]等地亦同步推进。AI 赋能科研基础设施建设正从国家战略延伸至区域创新体系，形成央地协同、分层推进的发展格局。同时，地方专项与高校平台之间的协同效应日

更多免费行业报告 并购家 www.ipoipo.net

益显现。例如，在上海“百团百项”专项支持下，复旦大学、上海科学智能研究院与上海创智学院联合推进“早期中华文明多模态大模型”及其评测体系建设[14]。该项目构建了全球首个聚焦早期中华文明意识形态研究的 720 亿参数多模态大模型，并配套建设“AI for Humanities 人文智能实训营”，推动历史、考古、语言等十余个人文学科与人机协同研究的深度融合。

由此，AI4SSH 基础设施建设形成了国际战略引领、国家顶层统筹、地方政策响应与高校平台落地相协同联动的多层级推进格局。

2. 从研究基础设施到数智学术基础设施：相关概念的演进

“研究基础设施”这一概念本身具有较强包容性。有学者指出，从研究者日常使用的笔记本电脑，到资助研究的制度安排，再到学术文献的引证网络，均可视作广义上的研究基础设施[15]。根据欧洲研究基础设施战略论坛（European Strategy Forum on Research Infrastructures, ESFRI）的界定，研究基础设施涵盖大型科研装置、科学数据资源、知识库、计算系统、通信网络，以及用于开展科学研究的其他独特工具[16]。这一界定突出了其工具性、平台性与公共性，也表明其本质并非单一物理设施，而是支撑知识生产活动持续展开的综合性条件。PARTHENOS 项目（Pooling Activities， Resources and Tools for Heritage E-research Networking, Optimization and Synergies）进一步强调了研究基础设施在较长时间段内的可重复使用性，并指出其应支持个人研究与协作研究及研究成果的共享[17]。对于人文社会科学而言，仅从广义层面讨论研究基础设施仍过于宽泛，尚不足以解释数字化、

网络化和智能化条件下知识生产方式的深刻变化。

因此，国内学界近年来围绕数字人文基础设施、新文科数字基础设施、数智学术基础设施等概念展开了更具针对性的探讨。在社会科学研究场景中，孟天广和严宇指出，文科实验室已成为数字时代社会科学研究的重要基础设施，开放型数据平台、自主算法工具库等构成其核心建设内容[18]。数字人文领域的相关讨论则更早聚焦于文献、数据、工具与服务的协同关系。刘炜等将数字人文研究基础设施界定为“在数字环境下为开展人文研究而必备的基本条件”，其内涵不仅涵盖文献、数据和软件工具，也延伸至学术交流、出版平台及相关公共服务[19]。徐晨飞和包平进一步强调了数字人文研究基础设施“促进数据分享与重用、加速科研创新的生态系统”的平台属性[20]。夏翠娟将讨论聚焦至“数字人文数据基础设施”，关注独立于具体项目和单一研究问题之外的数据层建设，及其长期保存、共建共享和开放利用的技术规范，凸显了内容、数据与知识层面的生产与组织机制[21]。由此可见，国内关于数字人文基础设施的讨论，已从资源数字化逐步拓展至平台生态、数据治理和知识组织等更深层次的问题。

随着人工智能特别是 AI for Science 的发展，关于科研基础设施的讨论从数字化条件下的资源组织，转向智能化条件下的平台科研与实验环境重构。毛进等提出“平台科研”范式，其核心在于以 AI 技术为支撑，实现多主体、多流程、多数据源协同的智能科研过程，突出通用性平台、智能化流程与 AI 对科研活动的创新性赋能[22]。杨小康等指出，智能化科学设施应在基础支撑层、科学模型层和实验应用层

更多免费行业报告 并购家 www.ipoipo.net

三层架构之上，具备传统科研范式所不具备的科学大模型、生成式模拟与反演、高通量自主无人实验和大规模可信科研协作等新功能[23]。上述立足 AI4S 语境而揭示的平台化、流程化和智能化趋势，启发了对 AI 时代人文社会科学基础设施重构方向的把握。

对人文社会科学而言，关于哲学社会科学研究基础设施的讨论逐渐转向“数智学术基础设施”。王晓光、付靖宜指出[24]，构建新一代数智学术基础设施已成为重塑哲学社会科学研究范式的优先战略，教育部哲学社会科学实验室正是其集中体现。作为典型的数智学术基础设施，其建设的核心支点在于开放的平台体系与先进的计算能力。随着生成式人工智能的发展，学界对此的理解持续深化。祝忠明、寇蕾蕾提出，AI 驱动的人文社会科学研究正从“工具依赖”走向“知识共生”，即 AI 深度嵌入知识生产全过程，与研究者形成共生关系[25]。顾洁、刘炜以“人机共生”概括此转型，强调生成式人工智能驱动哲学社会科学研究范式的根本性变革[26]。王晓光、付靖宜区分“数据驱动”与“人智协同驱动”两种路径，后者更强调 AI 与研究者间的动态交互与协同创新[27]。这些讨论表明，AI 时代的人文社会科学基础设施已超越数字资源、技术工具、研究平台的简单叠加，实质上是支撑人机协同知识生产的新型能力体系。

3. 本章对 AI4SSH 基础设施的界定

基于以上讨论，本章从科研范式转型的视角理解 AI4SSH 基础设施。相较于数据密集型研究阶段，AI4SSH 基础设施并非在人文社会科学研究中简单增加 AI 工具，而是对数据资源形态、计算支撑方式、

模型组织逻辑和研究协作机制提出了系统性新要求，主要体现在三方面。

第一，其核心目标从资源数字化保存转向提升可计算性，使研究对象能够被机器识别、被算法处理、被模型调用，并在不同研究场景中实现复用。

第二，其建设重点不再停留于工具的单点嵌入，而是强调数据、算力、模型、工具和平台之间的系统集成，形成面向研究全过程的技术链条。

第三，其组织逻辑不再局限于单一学科问题求解，而是跨学科、跨机构和跨场景协同，在人机互动与知识共生中催生新的研究方法与知识形态。

据此，本章将 AI4SSH 基础设施界定为：支撑 AI 深度介入人文社会科学研究、教学与知识服务活动的数据资源、计算环境、模型能力、工具链条及其协同机制的有机整体。这一整体并非若干要素的简单叠加，而是通过数据驱动模型优化、模型反哺数据生产、算力支撑全过程运行以及开放共享、互操作与跨域协作等机制，形成具有公共性、平台性、可扩展性和可复用性的基础支撑体系。

本章后文将进一步从三个层面展开分析：其一，考察高校如何围绕高质量、多模态与可流通的数据资源构建 AI4SSH 的数据底座；其二，分析高性能计算、智能计算与实验环境如何共同构成新型算力支撑体系；其三，探讨领域模型、智能体与工具链如何嵌入科研、教学

更多免费行业报告 并购家 www.ipoipo.net

与知识服务流程，形成 AI4SSH 的能力层基础设施。

11.2 数据底座与流通机制建设[①]

数据底座与流通机制决定 AI4SSH 能否从“看得见的资源汇聚”走向“用得上的可计算组织”。在传统数字人文与文科信息化建设阶段，数据建设更多对应“收藏—编目—检索”的资源逻辑；而在大模型与多模态学习驱动下，数据更直接嵌入模型训练、检索增强、知识抽取、推理与评测的全过程，其价值不再主要取决于规模本身，更取决于是否具备可复用的结构化表达、语义对齐能力与可追溯加工链条。对高校而言，这意味着数据治理必须同步升级：一方面通过校级底座与平台群整合多源异构数据，形成统一的发现、调用与版本管理机制；另一方面通过分级授权、审计日志与引用规范，在开放共享与安全合规之间建立可持续的流通规则。当数据能够以稳定接口进入跨学科工作流，并在标注、清洗、质量控制与更新维护中形成可持续循环，AI 能力才能在更广范围内扩散，推动研究从“材料驱动”走向“证据链驱动”，并为平台化公共能力供给奠定基础。

1. 校级汇聚：由分散资源建设走向数据底座整合

从当前高校实践看，数据基础设施建设的一个突出变化，是从“点状资源库建设”向“平台群整合”和“校级数据底座”转变。关键不再是数据库数量的增加，而是高校开始以基础设施视角重组数据资源，使不同来源、类型、用途的数据能够在统一框架中被发现、调用和治理。

[①] 本节由复旦大学程蕴涵、伏安娜合作撰写。
程蕴涵，复旦大学图书馆数据馆员，主要研究领域是科学数据管理、数据分析与可视化等。

更多免费行业报告 并购家 www.ipoipo.net

北京大学“文科数智化公共平台”是这一转向的代表性案例[28]。该平台以校级公共平台为统领，将数据库、案例库、语料库、图像库、政策库和工具库整合进统一入口，构建起跨类型资源的系统性服务框架。其中，数据库汇聚了北京大学开放研究数据平台、国内外公开数据库和公开大数据研究平台、国家统计局微观数据等；案例库整合了北京大学管理案例库、北大公共治理案例库、司法案例库；语料库涵盖中文语料库（如北京大学现代汉语语料库 CCL[29]）、平行语料库（如戏言字幕对译集[30]）等；图像库纳入中国汉代图像信息数据库[31]，以及计算艺术实验室建设的中国电影知识数据库[32]等资源；工具库则集成了数字人文平台[33]和人文社科数据搜索平台[34]。由此，原本分散的人文社科数据资源被纳入统一的公共服务框架，既增强了资源的系统性，又为后续研究调用提供了稳定底座。

类似的整合性布局也在其他高校出现。复旦大学形成了由复旦大学社会科学数据平台、星河启智平台、DataHub 等构成的复合型平台体系；中国科学院大学则通过中国科技云、中国科学院人工智能开放平台和 ScienceDB 等构建起数据、模型与服务相互联通的平台生态。综上，当前高校 AI4SSH 数据建设已从“分散平台建设”走向“底座式整合”。“数据底座”正在突破单个超大平台的简单扩张逻辑，转而通过平台群、统一入口和持续治理，将资源汇聚能力建设为一种基础设施能力。

2. 数据对象向多源异构与多模态演进

在完成平台整合的基础上，数据底座所承载的数据对象正经历结

构性拓展。AI4SSH 背景下，数据对象突破传统结构化统计数据与一般文本资源的边界，扩展为调查数据、古籍文献、语料资源、图像与音视频、时空信息、教育数据、治理数据及知识图谱等多元类型并存的格局。数据建设的重点随之从单一类型资源的数字化整理，转向对多源异构知识对象的系统组织、语义关联与计算支持。

在社会科学领域，调查数据、统计数据等仍然是科学研究的基础性数据支撑。例如，北京大学开放研究数据平台汇聚有中国家庭追踪调查（CFPS）、中国健康与养老追踪调查(CHARLS)、中国老年健康与家庭幸福调查（CLHLS-HF）、北京社会经济发展年度调查等高价值调查数据[35]。中国人民大学中国社会调查数据资料库则以中国综合社会调查(CGSS)、中国教育追踪调查（CEPS）、中国宗教调查（CRS）等长期连续调查为核心，形成了领域内具有重要影响的基础数据资源。在此基础上，社科数据也逐步由单一结构化走向多源异构和多模态扩展。例如，清华大学智能社会科学实验研究集成装置整合政务数据、经济数据、司法裁判文书和社交媒体等多类资源[36]；北京师范大学建设的全球新闻事件数据共享平台则对事件及其参与者信息进行自动化识别、概化、分类和编码[37]。复旦大学国家发展与智能治理综合实验室整合多个特色数据库，涵盖全球新闻事件、能源与环境监测、人口与社会变迁、分子考古等领域，数据规模超过 200TB[38]。总体而言，社会科学数据的主体虽仍以结构化数据为基础，但其发展方向已明确指向多源融合、跨媒介关联和场景化组织。

在人文学科领域，数据对象的扩展更具方法论意义。传统人文研

更多免费行业报告 并购家 www.ipoipo.net

究长期依赖文献阅读、版本比勘和历史阐释，但在 AI4SSH 背景下，古籍、语料、图像、人物、时间、地点和知识关系网络正被转化为可处理、可连接、可推理的知识对象。北京大学“识典古籍”不仅提供阅读与检索，还支持分类、时代筛选和研究辅助分析[39]；中国历代人物传记资料库（CBDB）将人物传记资料组织为可供统计分析与空间分析使用的关系型数据库[40]；“宋元学案”知识图谱系统借助自然语言处理与深度学习技术，将 240 万字古籍转化为结构化知识网络[41]；南京大学中国商业广告数据库（CCAA）通过对 18700 余张历史广告图像的细粒度元数据标注，探索了图像资源向可计算数据转化的路径[42]。此外，清华大学依托国家社科基金重大项目建设多模态语料库，并形成“全球汉籍影像开放集成系统”“中国古典文献资源导航系统”“时间轴知识图谱”“璇琮数字人文智慧平台”等相互联通的平台体系，共同构成古典文学文献知识化、智能化探索的数据底座[43]。在历史地理与文化遗产研究领域，南京大学“中华古城数字图谱集成工程”展示了人文学科数据对象的多模态扩展路径[44][45]。上述案例表明，人文学科传统研究对象正被转化为机器可识别、可关联、可训练、可推理的知识资源。

3. 数据流通的机制化转型

与传统数字资源建设相比，AI4SSH 数据底座重在“流通”而非单纯“保存”。从当前高校实践看，数据流通机制主要表现为三种类型。

其一是开放发布机制，即通过开放数据平台、科研数据仓储和机构知识库实现数据上传、描述、长期保存与对外发布。北京大学开放

研究数据平台、浙江大学开放数据平台、复旦大学社会科学数据平台、ScienceDB、上海交通大学科学数据存储库等均属该类型。此类机制的核心在于通过标准化元数据和平台化管理提升数据的可发现性与可获取性。此外，越来越多的平台引入数据引证制度，要求使用者规范引用原始数据。这不仅增强了数据的学术认可度与可追溯性，更使数据流通从单纯的“获取使用”延伸至“规范引用—可溯贡献—学术归属”的闭环，进一步强化了数据作为独立学术产出物的地位。

其二是分级授权机制。复旦大学社会科学数据平台、北京大学社会科学数据平台等允许数据所有者通过数据加锁、权限分层等方式，对不同用户实施差异化的开放授权；复旦大学星河启智平台等则通过校内登录实现授权服务。这表明，在 AI4SSH 背景下，数据流通并不意味着无边界开放，而是在开放共享与安全治理之间寻求平衡。对于涉及敏感属性或版权问题的研究数据而言，分级授权尤为重要。

其三是工具嵌入机制，这是特别能体现 AI4SSH 特征的流通方式。北京大学“吾与点”可将原始文本和图像自动转化为结构化表格、知识图谱和可检索知识库[46]；复旦大学星河启智平台、中国科学院大学中国科技云等，将数据、模型和工具链融入同一平台生态。由此，AI4SSH 语境中的“数据流通”超越了资源交换，转向“数据—工具—知识”之间的过程性协同流转。

这一变化说明，数据流通正在从“资源共享”走向“过程嵌入”。数据不再只是研究前端的输入材料，而是作为可计算、可调用、可追溯的要素，在清洗、标注、建模、推理、可视化和知识生成全过程中

更多免费行业报告 并购家 www.ipoipo.net

持续流动。流通机制本身也已成为数据底座的重要组成部分。

4. 面向 AI-Ready 的建设升级

在 AI4SSH 背景下，数据建设正由“可共享”转向“AI-Ready”。所谓 AI-Ready，重点强调了数据及元数据在结构化、标准化、标签定义、来源追踪、语义组织和可计算表达等方面达到可直接支撑模型训练、知识抽取、智能推理和知识生成的水平[47][48][49]。

这一导向已开始在国内政策和实践层面获得明确体现。2025 年，国家数据局公开征集高质量数据集，并将其界定为“经过采集、加工等数据处理，可直接用于人工智能模型开发训练，能有效提升模型性能的数据的集合”[50]。在随后发布的高质量数据集建设先行先试名单中，中国人民大学哲学社会科学高质量数据集、上海戏剧学院 AI 赋能艺术人才培养和演艺行业发展高质量数据集、中国戏曲学院中国戏曲高质量多模态数据集等高校人文社科类项目入选[51]。此后，上海市数据局发布的高质量数据集先行先试项目名单中，复旦大学《考古资料高质量数据集建设》、上海交通大学《面向法学人才培养的 AI 知识库和智能体构建》、华东师范大学《面向国际中文教育的高质量数据集构建》等人文社科类项目也相继入选[52]。

以复旦大学考古资料高质量数据集为例，该项目致力于将珍贵的考古文献资源转化为可持续利用、可开放共享的高质量数据资产，系统构建了从原始资料处理、细粒度标注，到元数据标准、考古学术语词表、本体模型、实体关系抽取和知识图谱生成的全链条体系。其示

范价值不在于资料规模，而在通过规范化和知识化处理，使数据具备进一步支撑领域模型和智能研究应用的能力。

5. 小结

总体来看，当前高校 AI4SSH 数据基础设施建设已呈现出三个相互关联的变化：其一，平台形态由分散的资源库建设转向校级数据底座与平台群布局；其二，数据对象由单一文本和表格扩展为多源异构、跨媒介关联的知识对象体系；其三，数据流通由静态共享转向嵌入清洗、标注、建模和推理过程的机制化流转。在这一基础上，AI-Ready 建设提出了更高要求，即不仅要让数据“可保存、可共享”，更要使其“可计算、可调用、可推理”。

未来高校 AI4SSH 基础设施建设水平，将越来越不取决于数据存量，而取决于能否将数据有效组织为可流通、可组合、可复用的智能研究要素。高校间的差异，也将更多体现在高质量数据生产能力、跨平台流通能力以及面向研究流程的要素组织能力之上。

更多免费行业报告 并购家 www.ipoipo.net

11.3 算力融合、实验环境与仿真平台发展[①]

在 AI4SSH（AI for Social Sciences and Humanities）领域，算力基础设施不仅承担着“算得动”的基础支撑作用，更在深刻重塑人文社会科学的研究范式：通过统一调度、预配置环境与交互式平台，传统计算中心已升级为覆盖科研全过程的 “研究装置”。这一升级对人文社会科学尤为关键——由于研究者大多缺乏从底层配置算力环境、搭建复杂工具链的完整技术能力，平台能否将异构资源封装为低门槛、可持续的公共服务，直接决定了 AI 能力向广泛学科场景的渗透程度。

随着超算、智算、通算 “三算融合”不断深化及工具链集成能力持续增强，部分高校进一步发展出社会仿真与多智能体实验平台，为社会过程研究提供了“可实验”的研究空间：研究者可以在更广阔的情景范围内开展压力测试，观察互动与涌现，并在更精细的层面识别分布效应、风险传导与政策外溢。与此同时，算力平台的“可用性”本身也是治理问题：资源口径的透明度、排队与配额的公平性、环境的可复现性、成本的可持续性，都会影响平台能否真正成为学科共同体的公共基础设施。因此，本节重点强调：算力建设需从“峰值规模”转向“公共研究环境”的质量导向，并提示仿真平台带来的新责任：实验化能力越强，越需要在数据边界、伦理审查、用途约束与结果沟

[①] 本节由复旦大学殷沈琴、张计龙合作撰写。

殷沈琴，研究馆员，复旦大学人文社科数据研究所副所长，上海市科研领域大数据联合创新实验室副主任，“开放数林”政府开放数据专家委员会委员，IFDO 国际数据组织联合会中国区代表，主要研究领域为科学数据管理、开放政府数据，数字人文。

通（不确定性披露）上形成制度化安排，避免“模型即现实”的误用与滥用，并建立必要的外部监督接口。

为把握当前高校在 AI4SSH 相关算力基础设施方面的整体布局，本文以教育部哲学社会科学实验室两批建设名单所涉高校为基准，综合 QS 世界大学排名中人文艺术、社会科学与管理学两大领域表现突出的代表性高校，采用网络调研方法，采集并梳理上述高校校级层面或校级主导建设算力平台的可公开获取信息，汇总形成表 10-1。需说明的是，本调研仅基于各高校公开披露的网络信息，不排除部分平台因未公开或不便公开导致算力统计不够完整的可能。同时，各校对于算力规模的统计口径不尽统一，部分平台仅公布特定计算集群的峰值性能数据，也有部分平台公布的是 AI 低精度算力、聚合算力、分项汇总算力或 Linpack 实测性能等不同口径数据。因此，表列数据仅用以反映各校算力平台在公开披露口径、资源类型、体系结构与资源配置特征上的概貌，不宜作为校际横向性能差异的精确比较依据，亦不涉及具体学科对算力资源的实际调用与分配情况。此外，各校算力资源的配置总量与其办学规模、学科结构之间存在关联，但受限于前述数据可及性因素，相关分析仅为初步判断。

从表 10-1 可以看出，当前高校相关算力平台建设并不表现为单一的规模扩张，而是呈现出多类型并存、异构融合和服务转向的特征。首先，平台公开算力口径差异显著，既包括以双精度、单精度为主的传统高性能计算平台，也包括面向大模型训练与推理的 Tensor FP16 等 AI 低精度算力平台，还包括多平台汇聚形成的综合算力底座，以

更多免费行业报告 并购家 www.ipoipo.net

及根据公开分项数据汇总或基于 Linpack 测试结果呈现的计算性能数据。其次，资源形态日趋多元，CPU 节点、GPU 节点、国产加速卡、大内存节点、云端资源和高速存储系统共同构成高校科研计算支撑环境。再次，平台功能正在由后台计算资源供给，转向支撑数据处理、模型训练与推理、智能体调用、虚拟仿真和交互式研究环境建设的复合型基础设施。对于 AI4SSH 而言，算力平台的关键价值不在于单一峰值指标，而在于其能否通过统一调度、低门槛接入和持续服务，转化为人文社会科学研究者可实际运用的公共研究环境。

表 11-1　部分高校算力平台公开信息概况

高校	平台名称	算力指标（公开口径）	核心/节点规模
北京大学	高性能计算平台[①]	双精度峰值：4.89 PFlops；单精度峰值：19 PFlops	计算节点 786 个，CPU 核心总数 40,500 个
清华大学	高性能计算中心（开拓 1000）[②]	Tensor FP16 峰值：455 PFlops	91 个 GPU 节点，共 728 块 GPU 卡
	高性能计算中心（探索 1000）	公开峰值：4.35 PFlops	900 个 CPU 节点
复旦大学	CFFF 平台[③]	公开报道口径：28 PFlops	千卡规模，可扩展至万卡[④]
上海交通大学	“交我算”平台[⑤]	总聚合算力：320 PFlops	101 台 CPU 节点，形成 5000+ vCPU 的通用计算池
中国人民大学	校级计算平台[⑥]	CPU 理论计算峰值：100 TFlops；GPU 理论计算峰值：480 TFlops	计算节点 93 台（CPU 节点 72 台、GPU 计算节点 21 台）
南京大学	高性能计算中心	理论双精度浮点计算峰值：约	360 台液冷计算节点

[①] 北京大学. 以算力筑基，北大高性能计算校级公共平台赋能科研创新与自主突破[EB/OL]. (2026-01-27)[2026-03-23]. https://news.pku.edu.cn/xwzh/00c33dd97de2417dad83b60ba8d63b76.htm

[②] 清华大学高性能计算中心. 硬件资源（计算资源）[EB/OL]. [2026-03-23]. https://hpcc.cs.tsinghua.edu.cn/jszy/yjzy.htm

[③] 复旦大学. 提升原始创新能力 强化创新策源能级[N]. 《复旦》, 2023-05-14(1284).

[④] 中央广播电视总台央视网. 国内高校规模最大云上智算平台上线[EB/OL]. (2023-06-28)[2026-03-23]. https://edu.cctv.com/2023/06/28/ARTImOcGqyooy4ifnnqL83cD230628.shtml

[⑤] 上海交通大学高性能计算中心. 高性能计算平台硬件配置[EB/OL]. [2026-03-23]. https://hpc.sjtu.edu.cn/Item/Hardware.htm

[⑥] 中国人民大学科学技术保障中心. 计算平台概述[EB/OL]. [2026-03-23]. http://pingtai.ruc.edu.cn/compute/overview/

	（蓝鲸一号）①	2.58 PFlops；HPL 实测效率：88.25%	
	高性能计算中心（Flex 集群）	CPU 理论计算峰值：约 870 TFlops（不含 GPU）；Linpack 实测效率：79.62%	906 台 CPU 计算节点、4 台 GPU 计算节点
北京师范大学	人工智能学院 AIC 平台②	双精度算力合计：448.3 TFlops（8 台 GPU 节点和 1 台 DCU 异构加速节点加总）	8 台 GPU 节点，1 台 DCU 异构加速节点
武汉大学	超算中心③	总体理论峰值：约 7.1 PFlops	505 台 CPU、184 台 GPU、168 台 KNL 服务器
中山大学	高性能计算公共平台（珠海校区）④	未见统一峰值指标公开披露	CPU 核心数 1288 核，高端 GPU 卡 56 张
	高性能计算公共平台（深圳校区）⑤	未见统一峰值指标公开披露	CPU 核心数 5000+核，高端 GPU 卡 20 张
	国家超级计算广州中心（天河二号）⑥	系统峰值计算速度：100.7 PFlops	总内存容量约 3PB，全局存储总容量约 19PB
中国科学院大学	科技云⑦	计算资源：315 PFlops	具体节点规模未见统一公开披露
	科学计算中心⑧	计算性能（Linpack 实测）：CPU 58.74 TFlops，GPU 8.54 TFlops，MIC 5.19 TFlops	206 个 CPU 计算节点、30 个 GPU 计算节点、10 个 MIC 计算节点
同济大学	科学计算平台⑨	未见统一峰值算力指标公开披露	570 台 CPU 节点、62 台 GPU 节点

说明：①表 10-1 根据各高校及相关平台公开资料、学校新闻及调研材料整理；②表中 PFlops/TFlops 指标依据原始公开资料保留其披露口径，包括理论峰值、AI 低精度算力、聚合算力、分项汇总算力及 Linpack 实测性能等类型；③“未见统一峰值指标公开披露”表示本次未检索到可直接用于横向比较的统一峰值算力指标。

1. 算力平台向融合式算力底座演进

从表 10-1 可见，当前高校算力平台在公开披露口径和资源构成上呈现明显差异，但总体演进趋势已突破传统意义上的高性能计算集群，

① 南京大学高性能计算中心. 计算资源[EB/OL]. [2026-03-23]. https://hpcc.nju.edu.cn/24266/list.htm
② 北京师范大学人工智能学院. 科研平台简介[EB/OL]. [2026-03-23]. https://ai.bnu.edu.cn/rcpy/sbpt/kypt/ptjj/index.htm
③ 武汉大学高性能计算中心. 中心简介[EB/OL]. [2026-03-23]. https://hpc.whu.edu.cn/zxgk/cszxjj.htm
④ 中山大学软件工程学院. 中山大学高性能计算公共平台（珠海校区）[EB/OL]. (2024-12-18)[2026-03-23]. https://sse.sysu.edu.cn/article/840
⑤ 中山大学公共卫生学院（深圳）. 中山大学高性能计算公共平台（深圳校区）[EB/OL]. (2024-09-03)[2026-03-23]. https://phs.sysu.edu.cn/zh-hans/article/961
⑥ 国家超级计算广州中心. 基本情况[EB/OL]. [2026-03-23]. http://www.nscc-gz.cn/Introduction/Generalprofile/Basicsituation.html
⑦ 中国科学院计算机网络信息中心. 中国科技云[EB/OL]. [2026-03-23]. https://cnic.cas.cn/kxyj/kypt/cxpt/202106/t20210630_6120519.html
⑧ 中国科学院大学科学计算中心. 系统平台[EB/OL]. [2026-03-23]. https://scc.ucas.ac.cn/index.php/zxjj/xtpt
⑨ 同济大学信息化办公室. 科学计算服务平台文档[EB/OL]. [2026-03-23]. https://dev.tongji.edu.cn/hpc-doc/#/

更多免费行业报告 并购家 www.ipoipo.net

而是在不同程度上向超算、智算与通算融合配置的综合底座转型。北京大学、南京大学、武汉大学等高校仍以高性能计算平台作为重要支撑，同时积极向异构算力和统一调度方向拓展。复旦大学 CFFF 平台、上海交通大学“交我算”、中国科学院大学中国科技云等，则已体现出更明显的异构融合特征。算力基础设施的核心价值，也由单纯提供计算峰值，转向支撑多模态数据处理、模型训练、工具调用和复杂科研流程运行的综合能力。

北京大学高性能计算校级公共平台具有较强代表性。该平台在原有 HPC 基础上，通过“未名”系列集群、自研 SCOW 平台和“鹤思”调度系统，逐步实现超算、智算、云计算等多类资源的统一管理与调度，清晰呈现了从单一算力供给向融合式算力平台演进的路径。复旦大学 CFFF 平台则明确面向“教—学—研”全链条，既提供 AI 计算框架，也兼容 HPC 调度环境，形成面向科研与教学双重场景的校级智算基础设施。上海交通大学“交我算”整合“致远一号”国产千卡智算平台、“思源一号”高性能计算平台、π 2.0 平台、AI 平台、ARM 平台和云计算平台，形成典型的综合算力底座。中国科学院大学中国科技云则将超算、智算、云计算、模型广场、数据集与软件服务纳入一体化体系，凸显出算力平台与数据平台、模型平台和服务平台深度耦合的发展趋势。

这一结构性演进表明，高校算力基础设施的基本形态已经发生深刻变化，建设重点已从单一高性能计算集群扩展为面向复杂科研流程的多类算力协同配置、统一调度和公共服务。对 AI4SSH 而言，这种

变化尤为关键，因为其研究任务往往同时涉及文本处理、知识图谱构建、模型训练、检索增强、多模态计算和交互式分析，单一 HPC 平台已难以满足科研全过程支撑需求。

2. 算力环境向研究装置升级

表 10-1 主要呈现了高校算力平台的公开披露口径、资源类型与节点配置，而从平台实际运行模式来看，更值得关注的是其功能边界的持续拓展。当前高校算力基础设施已不再仅仅承担后台计算资源供给功能，而是逐渐演化为面向科研全过程的研究环境支撑体系。换言之，平台的价值不仅体现在“拥有多少算力”，更体现在能否以统一入口、统一认证、预配置环境和低门槛服务，支撑不同学科研究者开展实际工作。

清华大学“智能社会科学（AI4SS）实验研究集成装置”是这一转型的代表性案例。该平台不仅整合政务、经济、司法文书、社交媒体等二十多类数据资源，而且通过虚拟化调度、Notebook 研究平台、学科智能体和 AI4SS 工具体系，为社会科学研究提供集数据处理、模型开发、分析计算与可视化于一体的交互式环境。这表明，在 AI4SSH 语境下，算力平台正在从传统“计算中心”向“研究装置型”基础设施转型。

这一趋势并非个例。复旦大学 CFFF 平台与星河启智平台也体现出类似逻辑：前者提供面向 AI 与 HPC 双场景的科研智算平台，后者则进一步将数据、模型、智算服务、Agent 与实验协作纳入同一链条，

更多免费行业报告 并购家 www.ipoipo.net

构成面向科研任务全过程的复合环境。上海交通大学网络信息中心在总结“交我算”平台建设经验时明确指出，AI 时代高校算力建设的重点已由规模扩张逐步转向存量优化和服务升级，平台需要同时支撑传统科学计算、AI 训练和 AI 推理，并通过统一调度实现“超算、智算、通算”三算融合[53]。该经验表明，高校算力平台能否成为真正的基础设施，关键并不只在于硬件性能，而在于能否构成持续运行的服务体系，包括专业团队、统一平台、资源调度机制和面向学科需求的定制化支撑能力。

对 AI4SSH 而言，这一点尤为重要。人文社会科学研究者通常不具备从底层配置算力环境、开发部署模型到组织复杂研究流程的完整技术能力，因此高校是否能够将这些能力封装为可融入日常科研活动的公共服务，直接决定了 AI 能力能否真正扩散到更广泛的人文社会科学学科场景。

3. 仿真平台推动研究环境实验化

从基础设施演进逻辑来看，仿真平台并非独立于算力平台之外的另一类设施，而是算力平台在数据处理、模型运行与调度能力持续提升基础上形成的高级应用形态。也就是说，当算力平台能够稳定支撑多模态数据处理、模型运行与交互式环境搭建时，其功能边界便自然延伸至场景化实验与虚拟仿真领域。对于 AI4SSH 而言，这意味着高校不仅提供“能算”的平台，更提供“可实验”的研究环境。

这种功能升级最具代表性的体现，就是社会模拟与多智能体仿真

平台的兴起。清华大学 AgentSociety 基于大语言模型构建具有“类人心智”的智能体，并通过模拟城市环境、行为规则和社会关系，还原复杂社会现象，使社会分层、认知极化、舆情治理、人口发展和区域协调等传统上难以实验的研究问题，获得了‘可实验’的研究空间。[54]中国人民大学“玉兰—万象”平台提出自动化场景构建、AI 社会科学家和大规模异步并行模拟等机制，为科研人员和决策者提供可交互、可干预的社会科学实验场[55]。复旦大学 SocioVerse（众生）则利用超大规模真实人口池和社会环境、人口、场景、行为四类引擎，对选举、社交媒体传播和社会经济调查等场景进行高精度模拟[56]。深圳大学 Society Zero Universe LLMs 社会仿真模拟平台也沿此方向发展，即借助大语言模型、外部数据接入和生成式工作流，将研究者的理论构想快速转化为可运行的社会仿真实验[57]。

这类平台虽在表现形式上常被归入模型或智能体应用，但从基础设施属性来看，其运行的前提在于稳定的异构算力环境、持续的数据供给和灵活的任务调度能力。换言之，社会模拟平台并非脱离算力平台而独立存在，而是算力平台从“后台资源”发展为“实验环境”的必然结果。其价值意义也不仅在于提升研究效率，更在于为长期以来难以通过真实实验开展研究的人文社会科学领域，提供了一种新的方法路径，使复杂社会过程在一定程度上具备了“可计算、可模拟、可比较、可干预”的特征。

4. 小结

总体来看，当前高校 AI4SSH 算力基础设施已呈现出三个相互关

更多免费行业报告 并购家 www.ipoipo.net

联的演进方向。其一，平台形态由单一高性能计算平台扩展为超算、智算、通算协同配置的复合型算力底座；其二，功能边界由单纯支持运算扩展为支持研究全流程、虚拟仿真和场景实验；其三，平台价值由硬件规模决定转向由统一调度能力、专业服务团队、技术封装和公共供给能力共同决定。

与此同时，当前建设仍存在若干亟待解决的问题：一是各校算力统计口径不一，平台规模和性能指标缺乏统一可比标准；二是平台运行模式差异较大，部分平台硬件能力较强，但面向人文社会科学研究者的可及性和易用性仍然不足；三是仿真实验平台的研究伦理、解释边界和数据边界仍需进一步规范。对于 AI4SSH 而言，未来算力建设的关键，不应再局限于峰值规模扩张，而应更加注重其作为公共研究环境、实验基础设施和可持续运行平台的系统化建构。

11.4 面向研究全过程的能力体系建构[①]

模型、智能体与工具链构成 AI4SSH 基础设施的能力层：它们决定 AI 能力能否嵌入科研、教学与知识服务的全过程，形成可复用、可扩展、可治理的组织能力。在高校场景中，能力建设的重点已不再是简单接入某一通用大模型，而是围绕专业语料、知识资源与研究任务，对模型能力进行校内化重组与学科化深化，并通过工具链平台化供给把“生成式能力”转化为可运行的工作流。通用模型的本地部署与自研转化提升可控性与可持续性，使模型从外部接口走向校内公共底座；垂域模型与智能体将学科知识、方法规范与场景需求固化为可调用能力，并反哺数据生产与研究流程，推动“工具使用”转向“过程嵌入”。与此对应，治理要求更加明确：需要在可复现、可审计与责任可追溯的底线上，建立评测与基准、权限与日志、内容安全与版权合规等配套机制。只有当能力体系能够以平台方式稳定供给，并在组织与制度层面获得认可（包括工程化支撑与平台贡献评价），高校的 AI4SSH 建设才能从“模型接入”走向“研究全过程能力体系”；同时应避免把不确定的生成输出直接当作学术证据，要求关键结论回到可核验的数据与方法链条，并保留人工审阅与同行审查的责任边界，以保证学术诚信。

1. 能力底座：通用模型能力的校内化重组

如果说数据底座解决的是“研究对象如何被组织”的问题，算力

[①] 本节由复旦大学汪东伟、殷沈琴合作撰写。

汪东伟，复旦大学图书馆馆员/工程师，上海市科研领域大数据联合创新实验室应用研发部主管，主要研究领域是科学数据管理、数据智能治理。

更多免费行业报告 并购家 www.ipoipo.net

平台回答的是“研究过程如何被计算”的问题，那么模型、智能体与工具链所要回应的，则是人工智能能力如何真正嵌入科研、教学与知识服务全过程，并转化为可持续、可复用、可扩展的能力体系。当前，大语言模型正支撑起多个社会科学分支学科的发展，如计算传播学、计算历史学、计算心理学、计算语言学、计算法学等，均已取得丰富成果[58]。在此背景下，高校 AI4SSH 建设的重点已不再是简单接入一个通用大模型，而是围绕专业语料、知识资源、研究任务与治理要求，对模型能力进行校内化重组，并进一步向学科化、场景化方向延伸。

从能力来源看，当前高校通用模型能力大体呈现两种路径。其一，是将外部成熟通用大模型本地化部署为校内 AI 能力底座。DeepSeek 发布后，国内多所高校开始将通用大模型部署为校内公共能力底座，而不再仅将其作为外部接口调用。浙江大学“浙大先生”本地化部署 DeepSeek 系列模型并面向全校开放服务[59]，复旦大学、上海交通大学、山东大学等高校也已将相关通用模型纳入校内 AI 服务体系。这一路径的意义，不仅在于提升模型调用效率，更在于通过本地部署、权限管理和校内资源适配，使通用模型逐步具备“可控、可持续、可扩展”的基础设施属性。

其二，是依托高校科研团队自研或技术转化形成通用模型能力源头。复旦大学团队发布的 MOSS[60]，是国内较早受到广泛关注的高校自研通用大模型之一；清华大学知识工程实验室技术转化形成的智谱 GLM 系列，则代表了高校科研成果向通用模型能力供给转化的另一种路径[61]。与前一种模式相比，这类实践更强调高校不仅是通用模型

的使用者，也是通用模型能力的重要供给者。由此，高校在 AI 能力体系中的角色开始由“接入平台”转向“能力源头”。

在本地部署与自研转化两种路径下，高校对通用模型能力的关注点都已不再停留于模型本身，而是转向将其组织为校内公共支撑系统。例如，北京大学通过大模型试验场、小北探索等平台组织模型测试、知识库服务与研究支持[62]；中国人民大学上线 RUC AI 大模型并将其嵌入教学科研场景[63]。由此可见，在高校场景中，通用模型的意义已不限于提供一般性问答能力，而是逐步成为支撑后续垂域模型、知识库系统和智能体应用的公共能力底座。

2. 垂域深化：学科模型的成果形态与支撑功能

在 AI4SSH 建设中，垂域模型具有双重属性。一方面，它是高校将学科知识、专业语料、研究方法和场景需求转化为模型能力的重要成果形态，标志着人工智能能力开始由通用调用走向学科内生；另一方面，它又不断反哺相关领域的科研、教学与知识服务，成为后续研究活动的重要支撑条件。基于上述双重属性，垂域模型并不限于某类阶段性技术成果，而正在成为 AI4SSH 基础设施体系中的关键组成部分。

从成果形态看，当前高校垂域模型建设已在教育、法律、新闻传播、古籍、古文字、中华文明研究等方向形成较有代表性的探索。教育领域，北京师范大学“师承万象”[64]“北极星”[65]“AI 大先生”[66]分别从模型能力、场景评测与教育知识底座三个维度构成协同布局。

更多免费行业报告 并购家 www.ipoipo.net

法律领域，中国人民大学涉外法治大模型围绕多语言、多法域和高可信架构，构建面向跨境法律服务的专业能力[67]。金融领域，中山大学“智鹿”金融大模型聚焦金融文本理解与数值推理任务，提升复杂金融场景下的决策支持能力[68]。新闻传播领域，中国人民大学“明新大模型 1.0”推出深澜·DeepPulse 深度报道助手、睿析·OpinionPulse 舆情推理模型、和鸣·GlobalPulse 国际社媒内容生产系统三个智能体，覆盖内容生产、舆情分析与跨文化传播全过程[69]。北京师范大学“AI 太炎”专攻古汉语文本理解，支持字词释义、文白翻译、句读标点与用典分析[70]。南京大学围绕人工智能古文字析解模型、中国思想家大模型和数智多语种文献处理平台，分别推进古文字识别、思想史知识组织与跨语种文献处理[71]。复旦大学“早期中华文明多模态大模型”将文本、图像、音视频与时空信息纳入统一研究框架，实现对中华文明早期遗存的多维整合与跨媒介分析[72]。南京农业大学“荀子古籍大语言模型”则面向古籍整理流程，支持标引、抽取、断句、翻译和词法分析等任务[73]。

从支撑功能看，这些垂域模型构成相关领域继续开展科研和教学的重要基础。教育模型已用于问答，并进入课程辅助、教学评价、智能批改、教师训练与教育评测等环节；法律与传播模型不仅体现为知识表达成果，也被用于法律咨询、案例分析、深度报道、舆情推理和国际传播等任务；古籍与文明研究模型则进一步支撑古文释读、断句标点、知识抽取、跨语种检索和多模态证据关联等研究活动。

3. 工具链供给：将 AI 能力嵌入研究流程

在 AI4SSH 基础设施体系中，工具链的重要性并不只在于提供若干智能化工具，而在于高校能否围绕数据生产、知识组织、模型运行和研究服务等关键环节，形成系统集成、协同衔接的支撑体系，并以公共平台、统一入口和持续服务的方式向科研人员开放。工具链建设的要点是将相关能力组织为可嵌入研究全过程的基础设施体系，使科研人员能够以更低门槛、更高效率调用人工智能能力。

首先，越来越多高校开始提供面向模型开发与运行的公共技术环境，使科研人员无需从零搭建复杂的软件与算力条件，而可以在校内平台上直接开展模型训练、微调、调用与交互式分析。这里所依托的并不仅仅是通用大模型本身，也包括 Notebook 环境、模型开发框架、容器化部署、训练推理管理工具等底层支撑条件。诸如 PyTorch、TensorFlow 等机器学习框架，以及 Jupyter Notebook 等交互式计算环境，虽然并不直接体现人文社会科学研究对象的特殊性，但构成了模型开发与运行的基础软件环境。以清华大学智能社会科学实验研究集成装置为例，其 Notebook 研究平台和 AI for Social Science 模块，已经将数据处理、模型调用、可视化分析等功能组织进统一研究环境[74]。对于人文社会科学研究者而言，这意味着人工智能能力开始从少数技术团队可用的专门资源，转变为可以由更多学科研究者共享使用的校内公共服务。

其次，高校正在围绕高质量数据集生产，提供面向人文社会科学复杂对象的数据加工与知识组织工具链。AI4SSH 所处理的对象通常不是标准化的结构化数据，而是文本、图像、音视频、目录、人物、

更多免费行业报告 并购家 www.ipoipo.net

事件、时空关系和语义结构等异构材料，因此其基础设施建设必须覆盖采集、清洗、标注、抽取、质检、发布和维护等一整套过程。在数字人文和计算社会科学实践中，具有对象辨识度的工具与标准，通常包括面向文本编码与结构化表达的 TEI（Text Encoding Initiative）、面向图像互操作与数字对象组织的 IIIF (International Image Interoperability Framework)、面向历史文献与手稿识别转写的 Transkribus[75]，以及文本分析、网络分析和数字展陈环境中的 Voyant Tools、Gephi、Omeka 等。这些工具和标准的意义，不在于其本身是否“新”，而在于它们为复杂知识对象的数字化重构、语义关联和后续计算提供了可复用的技术路径。北京大学“吾与点”智能数据平台能够对文本、图像等多模态材料进行结构化信息抽取，自动生成表格、知识图谱和可检索知识库；复旦大学考古资料高质量数据集建设则体现出“原始资料处理—细粒度标注—元数据标准—术语词表—本体模型—实体关系抽取—知识图谱”的链条式生产逻辑。由此，工具链首先体现为高校向研究者提供高质量数据资源生产条件的能力。

另一个重要的方向是高校正在将检索、阅读、分析、生成、协同与追踪等功能整合为面向科研全过程的服务型工具链，使人工智能能力不仅停留在数据生产或模型训练层面，而是进一步进入选题、综述、论证、写作和团队协作等学术工作流。北京大学 Science Navigator 围绕“读文献—做计算—做实验—多学科协同”组织科研流程[76]；南京大学“南雍闻道”集 AI 研究员、AI 综合问答、专业知识库问答、Deep Research、AI 学术追踪和 AI 知识库等功能于一体[77]；中国人民大学

DeepAnalyze Agentic 大模型将数据分析流程规划作为核心能力[78]；上海交通大学 SciMaster 提出了“搜、读、算、做、写”全流程闭环[79]。这些实践说明，当前高校提供的工具链已经不再是若干分散软件的简单叠加，而是在试图把 AI 能力稳定嵌入研究全过程，形成面向科研人员的持续服务体系。

需要指出的是，这类工具链基础设施的形成，并不完全依赖高校自主开发，而是逐步形成自研平台、校企共建平台和商业产品采购接入并行的混合供给格局。例如，北京大学—字节跳动数字人文开放实验室联合推出“识典古籍”平台；腾讯与高校在数字人文、文化遗产保护等领域亦有多个合作项目落地；玻尔（Bohrium）等平台已被部分高校部署为校内科研工具包；南京大学“AI 试炼场”由学校与超星、阿里云共同打造。

从总体上看，AI4SSH 中的工具链供给，正在推动高校科研支持方式发生深刻变化。过去，研究者往往需要自行搜寻工具、配置环境、拼接流程；而现在，越来越多高校开始通过统一入口、统一认证、统一平台和持续运维，把数据加工、模型调用、知识组织和研究服务整合为校内公共能力。在这一意义上，工具链并不是对模型或数据的附属补充，而是连接数据底座、模型能力与科研工作流的关键中介层。对于 AI4SSH 行动，高校是否能够提供这种面向研究全过程的工具链基础设施，构成衡量其基础设施建设水平的重要标志。

4. 治理嵌入：模型基础设施化以可信、可控与可评估为前提

更多免费行业报告 并购家 www.ipoipo.net

模型、智能体与工具链能否真正成为基础设施，最终取决于其是否具备稳定、可信、合规和可持续的治理结构。当前高校模型建设一个值得注意的变化是，场景化评测、本地化部署和使用规范正逐步内嵌于平台之中。

北京师范大学“北极星”基础教育大模型评测场是这一方向的代表案例。该平台围绕教育专业能力、教育应用能力和教育安全能力三大维度，构建覆盖多学科、多场景的教育评测体系，表明模型系统只有经过场景化持续评估，才能真正具备教育基础设施属性。中山大学 TrustedGPT 大模型可信评测平台，则体现出高校开始把大模型可信能力评测视为平台建设的重要组成部分[80]。由中国科学院计算机网络信息中心研发的国际首个面向人工智能赋能科学研究(AI4Science)的评价平台科学地平线(SciHorizon)平台则分别围绕“大模型科学能力”与“AI-Ready 科学数据质量”两个方面构建了综合性的评价体系，为科学创新提供权威评估基准[81]。

从更深层看，高校模型治理还包括知识来源治理、场景评测治理、使用规则治理和主体性治理。尤其对于人文社会科学而言，模型能否理解本土知识、是否尊重学术规范、能否在研究过程中保持研究者的主导地位，都是决定其能否成为基础设施的关键条件。

5. 小结

总体来看，当前高校 AI4SSH 模型、智能体与工具链建设，已呈现出四个相互交织的演进趋势：其一是底座化，即通用模型能力正在

通过本地部署、自研转化和校内平台化组织，成为后续应用扩展的公共能力底座；其二是领域化，即模型建设由通用能力调用走向围绕教育、法律、治理、古籍、文明研究等方向的专门能力构建；其三是工具链化，即数据生产、知识组织与研究服务被进一步贯通为面向全过程的支撑体系；其四是治理化，即模型系统能否成为基础设施，不仅取决于能力规模，更取决于其是否嵌入高质量数据、场景评测、伦理约束和制度规则之中。

AI4SSH 阶段的模型建设，其核心不在于模型的数量，而在于是否形成了一个由通用能力底座、垂域模型体系、工具链支撑和治理机制共同构成的能力体系。未来真正决定高校模型系统能否成为学术基础设施的，也不是模型参数规模本身，而是其与本土知识资源、研究方法、制度规则和学术责任之间的整合程度。

更多免费行业报告 并购家 www.ipoipo.net

11.5 发展展望

当前，我国高校AI4SSH基础设施建设已由早期的数字资源积累、平台试点和局部应用，逐步进入以数据、算力、模型和工具链协同配置为特征的新阶段。数据底座从资源汇聚走向可计算组织，算力平台由单一高性能计算走向融合式研究环境，模型、智能体和工具链开始进入科研、教学与知识服务全过程——这些进展表明，AI4SSH 基础设施正在成为高校新型学术基础设施的重要组成部分。未来一个更加关键的方向，是推动这些基础资源由“可用资源”转化为“可流通要素”，使数据、模型和工具能够在不同平台、不同场景和不同主体之间实现更高水平的共享、调用与组合。

如前所述，当前高校 AI4SSH 基础设施建设也面临若干亟待解决的现实问题。平台同质化倾向导致重复投入，开放共享与实际可及性之间存在落差，算力平台的配额限制和有偿使用等门槛对青年教师、小规模团队和人文学科构成制约，部分平台偏重“上线”而持续活跃度不足，数据确权和授权机制尚不健全。与此同时，数据安全、权限治理、模型调用边界与伦理风险问题也更加凸显，如何在开放利用、可信流通与合规治理之间形成平衡，将成为未来建设中无法回避的关键议题。

面向未来，高校 AI4SSH 基础设施建设应从四个方面继续推进。第一，在数据层面，重视高质量数据集建设、科学数据要素组织和流通机制完善，推动数据由“可存储、可共享”转向“可计算、可调用、

可组合”，并逐步建立适应人文社会科学特点的数据分类分级、授权使用和跨平台互操作机制。同时，还应关注不同高校在数据基础、技术条件和学科结构上的差异，探索分层分类的建设路径。第二，在平台层面，应由单点建设转向体系协同，推动专题平台、公共平台、算力平台、模型平台和研究服务平台之间的贯通，减少低水平重复建设，提升资源统筹与复用效率。同时，应重视资源配置的公平性与可及性，探索面向青年教师、小规模团队、人文学科和经费基础较弱学科方向的基础算力保障、普惠配额和专项支持机制，避免公共平台在实际运行中演化为仅服务少数重点团队的“准专属资源”。第三，在能力层面，应继续加强高质量数据集生产工具链、知识组织工具链和研究服务工具链的建设，使模型能力真正进入数据生产、知识重构、学术分析和成果表达全过程，形成面向研究全过程的稳定支撑体系。第四，在治理层面，同步推进可信评测、伦理审查、使用规范等制度建设，使 AI4SSH 基础设施在扩展能力边界的同时，守住学术规范、制度约束与安全底线。

更多免费行业报告 并购家 www.ipoipo.net

参考文献

[1] European Commission. European Strategy for Artificial Intelligence in Science[EB/OL]. (2025-10-08) [2026-03-19]. Available: https://research-and-innovation.ec.europa.eu/strategy/strategy-research-and-innovation/our-digital-future/european-ai-science-strategy_en

[2] Department for Science, Innovation and Technology. AI for Science Strategy[EB/OL]. (2025-11-20) [2026-03-19]. Available: https://www.gov.uk/government/publications/ai-for-science-strategy/ai-for-science-strategy

[3] The White House. Launching the Genesis Mission[EB/OL]. (2025-11-24) [2026-03-19]. Available: https://www.whitehouse.gov/presidential-actions/2025/11/launching-the-genesis-mission/

[4] 国务院. 关于深入实施“人工智能+”行动的意见[EB/OL]. (2025-08-26) [2026-03-19]. https://www.gov.cn/gongbao/2025/issue_12266/202509/content_7039598.html

[5] 国家数据局. 关于加强数据科技创新的实施意见[EB/OL]. (2025-12-27) [2026-03-19]. https://www.gov.cn/zhengce/zhengceku/202512/content_7053003.htm

[6] 新华社. “十五五”规划纲要草案，摘要来了[EB/OL]. (2026-03-08)[2026-03-22]. https://www.nda.gov.cn/sjj/swdt/xwfb/0308/20260308145749918872055_pc.html

[7] 北京市科学技术委员会，中关村科技园区管理委员会，北京市发展和改革委员会，等. 北京市加快人工智能赋能科学研究高质量发展行动计划（2025—2027 年）[EB/OL]. (2025-07-04) [2026-03-19]. https://kw.beijing.gov.cn/zwgk/zcwj/202507/t20250711_4146595.html

[8] 上海市人民政府办公厅. 关于进一步深化改革创新 加快科技服务业高质量发展的若干意见[EB/OL]. (2025-05-20) [2026-03-19]. https://stcsm.sh.gov.cn/zwgk/kjzc/zcwj/qtzcwj/20250527/ad5cf4a2721647b18c1b9fd6f4ca8637.html

[9] 江苏省人民政府. 江苏省“人工智能＋”行动方案[EB/OL]. (2025-12-29) [2026-03-19]. https://www.js.gov.cn/art/2026/2/3/art_64797_11725210.html

[10] 科学技术部. 科技部关于发布科技创新 2030—“新一代人工智能”重大项目 2018 年度项目申报指南的通知：国科发资〔2018〕208 号[EB/OL]. (2018-10-12)[2026-03-22]. https://service.most.gov.cn/kjjh_tztg_all/20181012/2731.html.

[11] 科学技术部. 科技部关于发布科技创新 2030—“新一代人工智能”重大项目 2022 年度项目申报指南的通知[EB/OL]. (2022-08-12) [2026-03-19]. https://www.ncsti.gov.cn/kjdt/tzgg/202208/t20220817_93926.html

[12] 新华社. 科技部启动“人工智能驱动的科学研究”专项部署工作[EB/OL]. (2023-03-27)[2026-03-22]. https://www.gov.cn/xinwen/2023-

03/27/content_5748495.htm.

[13] 中共中央, 国务院. 中共中央 国务院印发《教育强国建设规划纲要（2024—2035 年）》[EB/OL]. (2025-01-19)[2026-03-22]. http://www.moe.gov.cn/jyb_xxgk/moe_1777/moe_1778/202501/t20250119_1176193.html.

[14] 上海经信委. 百团百项丨早期中华文明多模态大模型：用 AI 唤醒沉睡千年的文明密码[EB/OL]. (2026-01-15)[2026-03-23]. http://sghexport.shobserver.com/html/baijiahao/2026/01/15/1695802.html

[15] (瑞典) 卡特琳娜·C. 克莱默，(瑞典) 奥洛夫·哈伦斯坦 主编。欧洲的大科学和研究基础设施 [M]. 乔黎黎，樊潇潇，邸月宝 译。北京：中国科学技术出版社，2021.

[16] ESFRI. ESFRI Roadmap 2021: Strategy Report on Research Infrastructures[R]. European Strategy Forum on Research Infrastructures, 2021.

[17] PARTHENOS Project. Other Teaching Resources[EB/OL]. (2016-11-29)[2026-03-24]. https://training.parthenos-project.eu/for-trainers/other-teaching-resources/.

[18] 孟天广， 严宇． 文科实验室：数字时代社会科学研究的“基础设施”［J］． 中国社会科学评价， 2024(2): 21-30, 157.

[19] 刘炜,谢蓉,张磊,等.面向人文研究的国家数据基础设施建设[J].中国图书馆学报,2016,42(05):29-39.

[20] 徐晨飞， 包平． 面向农史领域的数字人文研究基础设施建设研究——以方志物产知识库构建为引［J］． 中国农史， 2019, 38(6): 40-51

[21] 夏翠娟.面向人文研究的“数据基础设施”建设——试论图书馆学对数字人文的方法论贡献[J].中国图书馆学报,2020,46(03):24-37.

[22] 毛进,周凡倩,王卓昊.AI for Science 推动科研范式革新：创新知识服务视角下的“平台科研”范式[J].情报学报,2025,44(02):132-142.

[23] 杨小康,许岩岩,陈露,等.AI for Science：智能化科学设施变革基础研究[J].中国科学院院刊,2024,39(01):59-69.

[24] 王晓光， 付靖宜． 转型与破局：建设面向哲学社会科学的数智学术基础设施［J］． 图书与情报， 2025(3): 1-12.

[25] 祝忠明， 寇蕾蕾． AI 驱动人文社会科学范式升维：从工具依赖到知识共生［J］． 图书与情报， 2025(3): 13-25.

[26] 顾洁、刘炜：《生成式人工智能驱动的哲学社会科学研究范式转型》，《社会科学》2025 年第 6 期

[27] 王晓光， 付靖宜． 转型与破局：建设面向哲学社会科学的数智学术基础设施［J］． 图书与情报， 2025(3): 1-12.

[28] 北京大学文科数智化公共平台. 数据库[EB/OL]. [2026-03-22].

更多免费行业报告 并购家 www.ipoipo.net

https://hdipp.pku.edu.cn/sjzy/sjk/bjdxkfyjsjpt.htm

[29] 北京大学中国语言学研究中心. CCL 语料库检索系统[EB/OL]. [2026-03-23]. https://corpus.pku.edu.cn/index

[30] 北京大学外国语学院. 戏言字幕对译集 [EB/OL]. [2026-03-23]. http://corpus.ac.cn/

[31] 北京大学数字人文研究中心. 中国汉代图像信息数据库[EB/OL]. [2026-03-23]. https://hanart.pku.edu.cn/

[32] 北京大学艺术学院、中国传媒大学文化大数据实验室. 中国电影知识体系平台 [EB/OL]. [2026-04-01]. https://movie.yingshinet.com/

[33] 北京大学数字人文研究中心. 数字人文开放平台[EB/OL]. [2026-04-10]. https://hdipp.pku.edu.cn/sjzy/szpt/szrwpt.htm.

[34] 北京大学科学智能实验室. 数据集搜索[EB/OL]. [2026-04-10]. http://scielab.pku.edu.cn/dataset_search/.

[35] 北京大学开放研究数据平台. 北京大学开放研究数据平台[EB/OL]. [2026-03-23]. https://opendata.pku.edu.cn/

[36] 清华大学社会科学学院. 清华大学智能社会科学（AI4SS）研究平台 [EB/OL]. [2026-03-23]. https://ai4ss.sss.tsinghua.edu.cn/#/overview/home

[37] 北京师范大学地理科学学部地理数据与应用分析中心. 全球新闻事件数据共享平台：数据介绍[EB/OL]. [2026-03-23]. https://gde.bnu.edu.cn/share?type=datainfo

[38] 复旦大学. 国家发展与智能治理综合实验室简介[EB/OL]. [2026-03-23]. https://ghforum.fudan.edu.cn/c7/94/c13206a706452/page.htm

[39] 北京大学-字节跳动数字人文开放实验室. 识典古籍：古籍在线阅读平台 [EB/OL]. [2026-03-23]. https://www.shidianguji.com/

[40] ININDEX 数据库. 历史人物传记数据库（Biog）[EB/OL]. [2026-03-23]. https://www.inindex.com/biog

[41] 北京大学数字人文研究中心. 宋元学案知识图谱系统[EB/OL]. [2026-03-23]. https://syxa.pkudh.org/

[42] 陈静. 中国商业广告数据库[EB/OL]. 南京大学高研院数字人文创研中心, [2026-3-24]. https://digitalhumanities.nju.edu.cn/projects/5d9b01a69272a396e7.

[43] 清华大学. 大数据技术赋能文献学：从量变到质变的深度解析[EB/OL]. (2026-03-03)[2026-03-23]. https://www.tsinghua.edu.cn/info/3245/123793.htm

[44] 南京大学中华文明数智创新实验室. 南京大学“中华古城数字图谱集成”工程成果发布[EB/OL]. 南京大学, (2025-09-23)[2026-03-23]. https://www.nju.edu.cn/info/1056/440521.htm.

[45] 南京大学中华文明数智创新实验室. 高校唯一！南大文科实验室重要成果在北京文化论坛发布[EB/OL]. 澎湃新闻, (2025-09-24)[2026-03-23]. https://www.thepaper.cn/newsDetail_forward_31671586.

[46] 北京大学数字人文研究中心. 吾与点（WYD）智能标注平台[EB/OL]. [2026-03-23]. https://wyd.pkudh.net/

[47] Open Data Institute. A Framework for AI-Ready Data[EB/OL]. [2026-04-10]. https://theodi.org/insights/reports/a-framework-for-ai-ready-data/

[48] Verhulst S, Zahuranec A J, Chafetz H. Moving Toward the FAIR-R Principles: Advancing AI-Ready Data[EB/OL].[2026-04-10]. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5164337

[49] Stocker, M., Snyder, L., Anfuso, M. et al. Rethinking the production and publication of machine-readable expressions of research findings. Sci Data 12, 677 (2025). https://doi.org/10.1038/s41597-025-04905-0

[50] 国家数据局. 数字科技研究基础设施[EB/OL]. [2026-04-10]. https://www.nda.gov.cn/sjj/ywpd/szkjyjcss/0613/20250613202723101895576_pc.html.

[51] 国家数据局. 高质量数据集建设先行先试工作名单发布[EB/OL]. (2025-09-15)[2026-03-23]. https://www.nda.gov.cn/sjj/zwgk/tzgg/0915/20250915161428917995973_pc.html

[52] 上海市数据局. 上海市数据局关于公布第一批高质量数据集先行先试项目名单的通知[EB/OL]. (2025-12-05)[2026-03-23]. https://dt.sheitc.sh.gov.cn/cms/slddt/2650.jhtml

[53] 林新华. AI 时代 校级计算平台的可持续发展之路[EB/OL]. 中国教育网络,（2025-06-30）[2026-03-23]. https://www.cernet.edu.cn/info/focus/li_lun_yj/202507/t20250701_2678376.shtml.

[54] 清华大学.AgentSociety [EB/OL]. [2026-03-31]. https://agentsociety.readthedocs.io/en/latest/.

[55] 中国人民大学.玉兰-万象：YuLan-OneSim [EB/OL]. [2026-03-31]. https://www.yulan-onesim.cn/.

[56] 复旦大学智能复杂体系基础理论与关键技术实验室. 祝贺！复旦大学智能复杂体系基础理论与关键技术实验室魏忠钰老师入选“AI100 青年先锋”[EB/OL]. (2025-04-01)[2026-03-31]. https://iics.fudan.edu.cn/0b/ac/c33343a723884/page.htm.

[57] 深圳大学. 深圳大学发布 Society Zero Universe 平台 大语言模型重塑社会科学研究范式[EB/OL]. (2025-11-26)[2026-03-31]. https://news.szu.edu.cn/info/1003/14570.htm.

[58] 龚为纲,黄思源.大语言模型与计算社会科学:工具、议题与挑战[J].求索,2025,(03):137-148.

[59] 浙江大学. 浙大版 DeepSeek[EB/OL]. (2025-02-21)[2026-03-30]. http://www.lawlib.zju.edu.cn/redir.php?catalog_id=20706&object_id=268765.

[60] 复旦大学自然语言处理实验室等. MOSS[EB/OL]. [2026-03-23]. https://github.com/OpenMOSS/MOSS

更多免费行业报告 并购家 www.ipoipo.net

[61] 北京智谱华章科技有限公司. 智谱清言[EB/OL]. [2026-03-23]. https://chatglm.cn/main

[62] 北京大学计算中心. AI 创新实践[EB/OL]. [2026-03-23]. https://cc.pku.edu.cn/business/601.html

[63] 中国人民大学. RUC AI 大模型[EB/OL]. [2026-03-23]. https://chat.ruc.edu.cn/

[64] 北京师范大学. 师承万象基础教育大模型[EB/OL]. [2026-03-30]. https://smartedu-bnu.tal.com/

[65] 北京师范大学. 北极星(PolarisEDU)基础教育大模型评测平台[EB/OL]. [2026-03-30]. https://www.bnueval.com/.

[66] 北京师范大学. 北京师范大学发布教育数智中枢“AI 大先生”平台[EB/OL]. (2025-01-19)[2026-03-30]. https://news.bnu.edu.cn/zx/zhxw/b3e54fca82f24441826c28135179d496.htm.

[67] 中国人民大学高瓴人工智能学院. 代表性科研成果介绍[EB/OL]. [2026-03-30]. http://ai.ruc.edu.cn/firstresearch/researchresult/index.htm.

[68] 中山大学软件工程学院. 中山大学软件工程学院成立可信大模型研究中心[EB/OL]. (2024-07-17)[2026-03-30]. https://sse.sysu.edu.cn/article/762.

[69] 北京青年报. 发布“明新大模型 1.0” 人大新闻传播教育迎来 70 周年[EB/OL]. 中国人民大学新闻网转载, (2025-05-09)[2026-03-30]. https://news.ruc.edu.cn/1989862861716037634.html.

[70] 北京师范大学. AI 太炎[EB/OL]. [2026-03-30]. https://t.shenshen.wiki/.

[71] 程晓琳, 杨频萍. 中华文明研究迎来数智革命，南大“数智钥匙”开启中华文脉新维度[EB/OL]. 南京大学新闻网转载, (2025-05-16)[2026-03-30]. https://www.nju.edu.cn/info/1056/419231.htm.

[72] 复旦大学. 中华文明大模型[EB/OL]. [2026-03-30]. https://fudan.ai4s.com.cn/partner/research-plaza/research-plaza-web/galaxy-model/model/43.

[73] 南京农业大学. 首创！我校研发的全国首个古籍大语言模型正式发布[EB/OL]. (2023-12-07)[2026-03-30]. https://www.njau.edu.cn/2023/1206/c122a127132/page.htm.

[74] 清华大学.清华大学智能社会科学实验研究集成装置[EB/OL]. [2026-03-30]. https://ai4ss.sss.tsinghua.edu.cn/#/overview/home.

[75] Transkribus - The AI Platform for Historical Documents[EB/OL]. (2025-07-29)[2026-03-30]. https://www.transkribus.org/.

[76] 北京大学计算中心. 关于 AI for Science 科学导航平台上线服务的通知[EB/OL]. (2025-05-21)[2026-03-31]. https://its.pku.edu.cn/announce/tz20250521083037.jsp.

[77] 南京大学. 南雍闻道智能科学服务平台[EB/OL]. [2026-03-31]. https://nju.libsp.net/.

[78] 中国人民大学信息学院. 范举教授团队研发的面向数据科学的 Agentic 大模型——DeepAnalyze 正式发布[EB/OL]. (2025-10-29)[2026-03-31]. http://se-office.ruc.edu.cn/xwdt/f1e6f2953dc64542a9ef7679edbc7b7b.htm.

[79] 上海交通大学. 人工智能赋能基础科学 上海交大发布科学基座模型与科研智能体等重磅成果[EB/OL]. (2026-02-02)[2026-03-31]. https://news.sjtu.edu.cn/mtjj/20260203/219648.html.

[80] 中山大学软件工程学院. 大模型可信度哪家强？中大软工 TrustedGPT 告诉你！[EB/OL]. (2025-02-27)[2026-03-31]. https://sse.sysu.edu.cn/article/861.

[81] 中国科学院计算机网络信息中心. 科学地平线：AI4Science“数据+大模型”评价体系[EB/OL]. [2026-03-24]. https://scihorizon.cn/.

更多免费行业报告 并购家 www.ipoipo.net

第十二章 中国高校 AI4SSH 组织、推动与管理典型案例[①]

AI 赋能人文社会科学(AI4SSH)从“点状探索”走向“体系建制”，关键不在单项技术是否先进，而在组织方式、平台能力、激励机制与治理规则能否匹配交叉融合的复杂性。高校既是知识生产主阵地，也是交叉生态组织者：需要把算力与数据底座、跨学科协同机制、工程化交付能力与社会服务转化链条同时搭建起来，才能把技术增益转化为可持续的公共能力。

本章选取复旦大学、武汉大学、中国人民大学、北京师范大学四个案例，覆盖“平台型基础设施建设”“场景牵引的数智人文工程”“面向治理的复杂决策系统”“面向教育改革的循证教研闭环”等不同路径。四校实践共同指向一条可迁移的经验链条：以真实公共任务牵引跨学科协同，以平台化底座沉淀数据与模型资产，以制度化机制保障质量、透明与问责，并通过联盟网络、培训体系与开放合作实现扩散。

本章旨在以案例方式呈现 AI4SSH 组织创新的可复制要素，为高校在“十五五”背景下推进学科智能化转型、建设文科实验室与交叉平台、完善科研管理与激励制度提供参考。

[①] 本章由复旦大学国家发展与智能治理综合实验室汤维祺、王凌翔合作编写。素材由复旦大学文科处、高校哲学社会科学实验室联盟提供。

本章提要：

- AI4SSH 的体系化推进需“组织能力建设”，包括稳定的人才梯队、跨院系协同机制、工程化支撑队伍以及可持续运行的平台规则。
- “场景牵引”是跨学科协作的最有效粘合剂：以教育改革、文化遗产保护、数字政府与国家治理等复杂公共任务为抓手，可将学术解释、工程实现与社会服务嵌入同一闭环。
- 平台化底座决定可复制性：算力、数据、模型与工具链的共享机制及合规规则，能够把协作意愿转化为协作能力，降低跨团队成本并提升成果可复核性。
- 制度化学术治理与质量控制是“规模化”前提：学术委员会、子实验室矩阵、版本化数据治理、可审计日志与第三方评测等机制有助于把研究过程前置为可管理的证据链。
- 扩散与生态建设依赖联盟网络与人才体系：通过暑期学校、学术年会、培训课程与共建共享平台，形成共同语言与工作流模板，推动从个案到规模化应用。
- 高校经验共同指向“基础设施—规范—产出”的联动逻辑：以公共能力为导向的制度供给能够同时促进创新活力与风险可控，并提升成果进入公共决策与社会服务的可解释性。

更多免费行业报告 并购家 www.ipoipo.net

12.1 高校 AI4SSH 体系建设案例

1. 复旦大学 AI4SSH 点线面协同推进

复旦大学在较早阶段即将科学智能（AI4S）与哲学社会科学智能（AI4SSH）纳入学校整体布局，形成了从算力底座到学术生态、从学科广度到尖峰项目的结构性优势。作为国内哲学社会科学研究的重镇，学校提出以“有组织科研+平台化支撑”推进 AI 赋能：一方面服务国家重大战略与上海城市治理实践，另一方面面向国际学术前沿与全球治理议题，探索中国高校在 AI4SSH 领域的组织创新与制度供给。在“AI 驱动的人文社会科学理论创新与范式变革”议题上，学校强调以问题为牵引、以平台为支撑、以机制为保障，推动研究从“数字化”走向“智能化”，并通过可复制的流程把跨学科协作固化为常态能力。

（1）打造持续生长的人才体系

复旦将“人”作为 AI4SSH 体系建设的起点。基于对全校 2023—2025 年科研项目、学术论文及战略性成果的全面分析，学校在人工智能相关领域已形成一支规模化的哲学社会科学智能人才队伍：队伍规模约 326 人，占哲学社会科学教师总人数约 23.3%。队伍结构呈“金字塔”分层：正在向“核心圈”迈进的约 10%是理论与应用并重的战略引领者；“进阶圈”约 20%是各院系将 AI 融入研究的创新主力，覆盖新闻、管理、经济、法学、国际关系、外语等多个学科；“培育圈”约 70%是未来发展的广泛基础与后备力量，已在数字化转型中展现潜力。这种分层并不以行政隶属为界，而以研究问题与方法能力为

纽带，便于学校在重大任务中快速组建跨学科“作战单元”。

在培养路径上，学校强调“研究—课程—平台”联动。一方面，在数智人文（计算语言学、科技考古、数字史学等）与计算社会科学（计算政治学、计算社会学、计算法学、计算传播学、计算心理学、计算经济管理学等）方向涌现出一批兼具领域深度与 AI 技能的交叉型青年人才。2024 年发布的国内首部《未来已来——哲学社会科学智能蓝皮书》由校内五十余位中青年学者参与编写；2025 年世界人工智能大会发布的“中华早期文明源流”多模态大模型则有超过百位青年教师和学生参与，形成“项目带动式”培养链条。另一方面，学校通过校内课程体系扩大覆盖面：在校内百门“AI 大课”中，约三十七门与哲学社会科学相关，选课人数达到 2121 人，使 AI 能力逐步成为人文社会科学研究的通用基础设施，而不是少数团队的“专属技能”。

（2）点线面结合，构建哲学社会科学智能学术新生态

在项目组织上，复旦形成“面上全覆盖—点上重突破—线上深贯通”的推进逻辑。面上层面，2024 年初出台《复旦大学科学智能专项基金实施方案》，截至当年 9 月累计立项哲学社会科学领域科学智能专项基金项目 17 项；同年 11 月依托上海市教委专项任务开展“人工智能促进科研范式改革赋能学科跃升计划”，累计立项 8 项。这些项目发挥“宣传队”“播种机”作用，使 AI 赋能研究在哲学社会科学学科实现较广覆盖。点上层面，学校联合上海科学智能研究院、上海创智学院等启动“中华早期文明源流跨模态智创工程”重大项目，聚焦考古、古文字、古籍、历史地理与语言学等学科，探索以多模态大

更多免费行业报告 并购家 www.ipoipo.net

模型重构早期文明研究的知识生产方式。

与此同时，学校注重把学术生态与社会服务、国际合作相衔接：通过与相关院系平台机构、科研院所与企业形成联动，积极对接联合国、WMO、WHO、IPCC 等国际机制，探索以技术合作为载体、以标准输出为手段、以国际影响力打造为目标的开放生态建设，推动形成“研究—平台—政策—国际合作”贯通的多层级推进格局。

线上层面，学校强调全球协作与产教融合，形成多学科平台体系：一方面面向全球治理与规则制定等重大议题，推动建立人工智能治理创新平台与国际合作网络；另一方面，在校内逐步形成以领域深耕为特色的哲学社会科学实验室“三层体系”，让传统学科与 AI/大数据技术的融合有稳定的组织载体与工程支撑。在学术传播与共同体建设方面，学校通过大会、蓝皮书与国际会议强化“共同语言”。例如 2025 年举办人文社会科学智能大会（参会者约 400 余人），发布国内首部人文社会科学智能发展蓝皮书；同年主办国际社会计算会议 ICSC 2025，吸引近 200 名中外专家学者围绕社会网络、算法公平、数字信任、健康与社会计算等议题交流，形成跨学科、跨部门、跨区域的协作网络。

（3）基础设施牵引：以 CFFF 为底座的“数据—模型—工具”共享机制

复旦将新一代科研基础设施视为 AI4SSH 可持续运行的关键支点。CFFF（Computing for the Future at Fudan）作为校级智能计算平

台，由“切问”一号与“近思”一号等部分构成，提供 GPU/CPU 算力、分级存储与大数据处理能力，可支持千卡级大模型研发、万核高性能计算与大规模数据加工。平台不仅提供资源，更提供可复用的工具链与解决方案，帮助不同学科团队把 AI 能力嵌入研究流程。在平台化逻辑下，算力不只是“资源供给”，更是推动学科协作的组织工具：它促使数据治理规则、模型调用规范与跨团队共享流程逐步制度化，从而把“协作意愿”转化为“协作能力”。

在数据与模型层面，复旦强调安全共享与跨团队复用：通过平台汇聚多学科数据资源，推动形成可调用的数据集与特色模型库，为哲学社会科学领域的文本、图像、行为与制度材料提供统一处理底座。更重要的是，平台化治理为科研管理提供抓手——从数据合规、隐私保护到模型评测与复现要求，逐步形成可操作的研究规范与证据链，提升成果在政策咨询与社会服务中的可解释性与可追溯性。这种“基础设施—规范—产出”的联动机制，使 AI4SSH 不止于技术应用展示，而能够沉淀为可持续的研究能力与治理能力。

（4）经验启示：把“组织创新”做成可复制的流程

一是以人才分层与能力扩散为前提，兼顾尖峰引领与群体提升。通过“核心—进阶—培育”的梯队结构，既保障重大任务的攻关能力，也形成持续扩张的学术共同体。

二是以“点线面”组织逻辑将科研从零散探索转向体系化推进：面上项目扩大覆盖、形成方法扩散；点上工程集中突破、塑造标杆成

更多免费行业报告 并购家 www.ipoipo.net

果；线上平台贯通资源、降低协作成本。

三是以基础设施平台化治理重塑研究流程，把算力、数据、模型与工具嵌入科研管理与学术规范，形成“数据—理论—机制”闭环的可操作抓手。

2. 武汉大学文化遗产智能计算实验室的跨学科组织模式

文化遗产保护与活化利用，是典型的 AI4SSH“高复杂、强跨界、重工程”场景：它既需要对文物、遗址、古籍等对象的学术解释，也需要测绘遥感、信息管理与智能计算等技术体系支撑，更需要面向公众传播与治理需求的转化路径。武汉大学文化遗产智能计算实验室以此为切入口，形成了“学术研究—工程平台—人才培养—社会服务”贯通的组织模式，为高校文科实验室在数智时代如何组织跨学科合作提供了可借鉴样本。

（1）定位与使命：在“数字中国—文化强国”战略中建设工程能力

实验室面向文化遗产数字化与智能化活化的国家需求，依托武汉大学信息管理、测绘遥感等优势学科资源，协同历史、考古、文学艺术等传统人文学科，强调以数据工程与智能算法提升文化遗产研究、保护与传播的系统能力。与单纯的数字化展示不同，实验室更注重把“学术解释能力”与“工程实现能力”绑定：一方面沉淀可持续的数据与模型资产，另一方面在真实场景中验证方法、完善流程，从而形成可复用的技术与治理方案。

（2）组织机制：跨学院协同与项目制“作战单元”

实验室以跨学科团队为核心组织单元，通过“考古—测绘—信息—计算”多链条协同，将田野调查、数据采集、模型训练与成果展示纳入同一项目闭环。典型做法是以重点遗址与区域文明研究为牵引，联合多家单位组成联合考古队，在长期田野工作中建立数字化管理与质量控制流程，使数据采集、标注与分析能够持续累积并可追溯。在

更多免费行业报告 并购家 www.ipoipo.net

科研管理上，实验室强调“工程师—研究者”共同工作：对外是服务遗产地、博物馆与管理部门的需求侧，对内则以平台建设、数据治理与工具研发把需求沉淀为可复用能力。

在联盟协同方面，武汉大学承办高校哲学社会科学实验室联盟第二届会议（2025 年 5 月 14—15 日），会议聚焦“数智时代哲学社会科学与人类文明新形态”，来自 30 家教育部哲学社会科学实验室的专家学者与科研管理部门同仁 200 余人参会。会议提出推动资源整合与协同创新、培育交叉人才加强合作交流、深化成果应用与社会服务赋能等重点方向，并通过平行论坛与工作会议推动形成资源共享体系、定期沟通与项目共研的长效机制。对实验室而言，这种联盟机制既是成果展示窗口，也是“共建共享”的治理平台，有助于把单点经验转化为可推广的组织规范。

（3）平台与成果：以数字考古与遥感 AI 为突破口的“证据链”建设

在成果产出上，实验室突出“数据—模型—场景”的一体化，并把关键成果沉淀为可复用的平台能力。以盘龙城遗址等重大考古项目为代表，团队搭建田野考古发掘数字化管理平台，结合激光滤波测绘、实景三维建模、声呐测深与数字高程模型等方法，复原约 3500 年前聚落景观与自然环境。该项目入选“2024 年度全国十大考古新发现”；同时，实验室在云梦郑家湖墓地（2021 年度）、十堰学堂梁子遗址（2022 年度）等项目中持续开展数字考古工作，逐步形成“现场采集—数据治理—三维重建—知识表达”的标准化流程。在联盟第二届会议成果

发布环节，实验室团队还发布了“文化遗产 3D 数据空间”等成果，展示文化遗产三维数据的组织、检索与展示能力，为跨团队共享与公众传播提供了统一入口。

在方法创新上，实验室将遥感数据、云平台与深度学习结合，探索“以算法指导调查、以数据反哺解释”的新范式。围绕遗址预测、城墙语义分割、壁画线描自动生成、古籍页面拼接等任务，团队形成一系列交叉研究成果：例如基于云平台耦合地理要素与时序光谱特征的遗址预测模型、基于 LiDAR 数据的遗址语义分割、面向壁画数字化保护的线描自动生成方法、以及面向古籍数字化的非重叠图像拼接方法等。在工程实践中，团队完成重点区域超 1 万平方公里机载 LiDAR 遗址遗迹探测识别项目，推动交叉研究与应用创新。据实验室简报，相关团队 2024 年获批国家重点研发计划项目，承担 5 项国家重点研发课题及 2 项国家自然科学基金面上项目，为文化遗产保护与数字化转型提供持续的科研与工程支撑。

（4）人才培养与开放合作：以暑期学校与国际协作为扩散机制

实验室把人才培养嵌入工程实践，通过暑期学校、工作坊与跨校交流形成“教学—科研—实践”联动。2025 年举办“文化遗产数字化保护与智能计算”暑期学校，由历史学院、长江文明考古研究院、文化遗产智能计算实验室等单位组织，围绕文化遗产数字化的理论学习与实践操作开展集中培训，帮助不同学科背景的学生与青年研究者形成共同语言与协作能力。与此同时，实验室也积极将研究成果转化为可感知的公共产品，例如建设文化遗产数字演绎空间，融合影像建

更多免费行业报告 并购家 www.ipoipo.net

模、沉浸式投影、XR 叙事等设备与人文阐释方法，形成“数字孵化—虚实融合—协同演绎—沉浸体验—智慧服务”的创新链条，为文化遗产活化利用提供可复制的展示与传播路径。

在社会服务层面，实验室以文化遗产保护的现实需求为牵引参与地方治理与公共文化服务。例如在武当山文化遗产保护座谈会等活动中，实验室专家从遥感测绘与数字化技术视角提出保护与文旅应用方案，强调数字化技术对提升保护精细化与拓展活化利用路径的作用，体现了“研究—治理—传播”联动的实践取向。

（5）经验启示：把“场景牵引”转化为可推广的组织与制度

一是以真实场景牵引跨学科协同，把文化遗产保护的复杂链条拆解为可执行的任务单元，在项目制中形成稳定协作关系。

二是以平台化数据工程沉淀资产，强调数据采集、标注、质量控制与可追溯的证据链建设，为学术解释与公共传播提供共同底座。

三是以联盟与教育项目推动成果扩散，通过会议、暑期学校与国际合作建立“共享—共研—共用”的长效机制，提升可复制与可推广能力。

3. 人民大学数字政府与国家治理实验室的组织创新与政策转化链

数字政府建设正在把公共管理的关键问题推向“数据化、实时化、系统化”：政策制定需要更高频的监测指标，公共服务需要更精细的需求识别，治理风险需要更前置的预警与评估。面对这一趋势，中国人民大学数字政府与国家治理实验室以“数字政府建设与治理创新”为主轴，探索以大数据与人工智能为底座、以跨学科团队与工程化能力为支撑的研究组织模式。其目标并不止于在学术上提出新概念或新模型，而是把社会科学研究从“解释世界”进一步推进到“支撑复杂决策系统”的工程能力建设：能够持续供给数据、可迭代地产出指标与模型、可复用地形成工具与方法，并在真实治理场景中验证与修正。

（1）战略定位：以“数字中国—治理现代化”为牵引的研究平台

中国人民大学将数字政府与国家治理问题视为哲学社会科学数智化转型的重要方向。实验室围绕数字政府、智慧城市与社会治理等议题，强调用数据科学与 AI 方法提升国家治理的观测、评估与模拟能力，形成面向公共部门可落地的工具与方案。

在更高层级的平台建设上，学校推动建设“国家治理大数据和人工智能创新平台”，把其定位为文科实验室体系的新形态：以大数据和人工智能技术为基础，以各学科子实验室研究团队和专职工程师团队为支撑，以专业的数据工程服务与开放数据平台为内核，服务学术研究与政府决策，推动哲学社会科学数智化转型，并面向国家治理重大需求建设复杂决策系统。

更多免费行业报告 并购家 www.ipoipo.net

2024 年学校举行创新平台揭牌仪式并发布子实验室体系，强调以数据为中心的学科生态与工程化服务能力建设，提出打造交叉融合新范式、人才引育新平台、咨政服务新生态等目标。这些部署为数字政府与国家治理实验室提供了更高能级的资源与制度支撑，也明确了其“研究高地+工程平台+决策支撑”的综合定位。

（2）组织与管理机制：跨院系协同的“子实验室+工程师”架构

实验室依托公共管理学院，并与社会与人口学院、人工智能学院、统计与大数据研究院、信息学院、环境学院等多学院形成协同支撑，体现了“治理问题牵引—多学科供给”的组织逻辑。与传统项目组相比，实验室更强调长期运行的组织能力：通过专职工程师团队与数据工程服务，把数据供给、工具开发、模型迭代与成果交付纳入常态化流程。

数字政府议题天然跨学科：既涉及公共管理与政策分析，也涉及社会学与人口、统计与数据科学、信息系统、人工智能、环境与空间治理等知识。实验室在组织上依托多学院协作，通过学术团队与工程团队的分工协同，将“研究问题”与“工程实现”放在同一管理框架内推进。

2025 年 4 月 23 日，实验室召开发展研讨会暨学术委员会委员聘任仪式，主题为“数字政府建设与治理创新”。会议由实验室、中国人民大学创新高地国土规划与国家治理研究院、公共管理学院等单位共同主办。学校领导在致辞中强调数字政府建设对推进国家治理现代

化的关键作用，并现场宣读学术委员会委员聘任决定、颁发聘书。实验室聘任叶嘉安院士担任学术委员会主任，丁德文、郭仁忠、陈军院士等担任副主任，并邀请来自多所高校与科研机构的专家担任委员。通过学术委员会制度，实验室将重大议题遴选、研究规范与成果质量评估前置化，为跨学科合作提供稳定的学术治理框架。

在平台运行上，实验室强调“工程服务—研究产出—政策转化”一体化：工程师团队负责数据汇聚、治理与工具开发，学术团队负责理论框架、指标体系与因果识别，二者在项目制中形成可复用的流程与规范。

在激励机制方面，实验室强调以“任务牵引+过程评估”替代单纯以论文数量衡量：对跨学科团队，重点考察数据资产沉淀、模型工具复用、政策服务贡献与人才培养成效；对工程团队，则以平台稳定性、数据治理质量、工具可用性与服务响应效率作为核心指标。这种面向交付的评价逻辑，有助于把“研究创新”与“治理应用”纳入同一套激励框架，避免出现技术与学术两张皮。

（3）研究与服务路径：从监测评估到政策仿真的全流程解决方案

围绕国家治理重大需求，平台以“子实验室矩阵”组织研究方向。根据公开报道，创新平台一次性授牌 12 个子实验室，覆盖国家数字监督治理、就业监测与政策仿真、减贫与乡村振兴与可持续发展、数字法治、国际传播与全球民意研究、基层与社会治理质性资料数据、世界一流企业数智治理与创新、城市更新、医疗和健康服务体系监测

更多免费行业报告 并购家 www.ipoipo.net

分析、重要农产品产量预测与价格预警、中国式现代化统计评价与动态监测、北京市消费大数据监测等领域。这种矩阵式组织兼顾“国家需求导向”与“学科能力供给”：既能围绕共性方法（指标体系、因果识别、网络分析、仿真评估）形成共享底座，也能在子领域中沉淀具有可迁移性的工具与数据资产。

在科研管理上，实验室将“可复现、可审计、可解释”作为面向治理应用的底线要求：对数据来源、清洗与标注过程建立版本化记录；对模型与指标的更新建立评测与对比机制；对政策评估结论的适用范围与不确定性进行明确披露。这套“证据链管理”不仅提升学术质量，也降低政策转化中的沟通成本，使研究成果能够以更透明的方式进入公共部门决策流程。

在方法路径上，实验室强调把“数据—理论—机制”闭环落到治理场景：以可持续的数据供给与指标体系为前提，结合因果识别、网络分析与行为建模等方法，将政策效果评估从一次性研究转向持续监测；以模拟与推演工具为支撑，将政策选择空间显式化，帮助决策者理解不同组合方案的权衡关系与风险边界。

此外，实验室通过开放交流与能力建设提升外溢效应：面向政府部门、行业机构与学术共同体，提供专题研讨、培训课程与工具试用，推动形成“需求侧提出问题—平台侧提供方法—学界共同改进”的协作循环。在这一循环中，社会科学理论为指标设计与机制解释提供框架，AI 与数据工程为高频监测与大规模评估提供手段，而治理实践

则反过来检验理论假设与模型适用边界，从而形成可持续的迭代路径。

（4）经验启示：以平台化治理提升 AI4SSH 的可复制性

一是以开放数据平台为“共同底座”，把分散的数据资源与研究能力聚合到可治理、可共享的基础设施上，降低跨团队协作成本。

二是以工程师队伍作为“稳定中枢”，将数据工程、工具开发与模型部署纳入研究管理，使社会科学研究具备持续迭代与服务交付能力。

三是以子实验室矩阵组织跨领域议题，在统一规范下形成“领域深耕+方法共用”的格局，既保证学术深度，也便于形成可推广的治理解决方案。

四是以数据安全与合规治理为前提，把隐私保护、授权使用与风险控制纳入平台规则，既保护数据资产，也为跨部门合作提供信任基础。

更多免费行业报告 并购家 www.ipoipo.net

12.2 中国高校 AI4SSH 创新应用案例

1. 北京师范大学“AI 大先生”项目

北京师范大学面向智能时代教育变革与人文社会科学研究范式重构的双重需求，形成了以教育治理与教育实践为牵引、以数据与模型平台为底座、以跨学科协同与有组织科研为抓手的 AI4SSH 体系化推进路径。其突出特点在于：不把 AI 仅作为“效率工具”，而是将 AI 嵌入教育研究—教研—治理全链条，把可解释的证据、可复用的工作流与可扩展的协同网络一起建起来，使 AI 能力真正转化为可持续的公共能力。

（1）以教育强国使命牵引 AI4SSH“研究—实践—治理”一体化

北师大在 AI 与人文社会科学融合上采取“问题牵引、场景落地、平台化建设”的总体思路：一方面围绕基础教育改革与教育治理现代化，将 AI 能力直接对接课程改革、教研改造、评价变革与管理优化等关键环节；另一方面在语言文字与传统文化研究中，通过资源平台化与智能化工具链建设，推动“材料—方法—解释”协同升级。这种定位使其 AI4SSH 建设兼具两类价值：对内形成可持续的学术与人才生态，对外能够以可复制的解决方案服务区域教育治理与公共文化传承。

（2）实验室—平台—联盟的分层协同与建制化推进

在组织载体上，北师大呈现出“实验室牵引、平台支撑、联盟扩散”的分层结构：

实验室层面，北师大入选第二批教育部哲学社会科学实验室培育建设序列，建设“国家教育发展战略智能决策实验室”，建设周期为5年（含1年试运行期），为教育战略研究、政策评估与智能决策提供制度化平台载体。

平台层面，北师大构建教育数智中枢“AI大先生”平台，汇聚教学、评价、管理等多模态教育数据，并以知识图谱、专业知识库与检索增强生成等方式形成教育领域垂直大模型能力，为研究与实践提供统一的技术中台与数据底座。

联盟层面，依托“AI大先生”形成“AI+循证教研联盟”，以教研场景为切口组织跨区域、跨学校的协同试点与经验扩散，推动AI赋能由单点试验走向可复制、可推广的治理方案。联盟已覆盖多个区域与大量学校，形成以课例分析与证据汇聚为核心的持续迭代机制。

（3）以循证闭环把AI能力转化为可复制的教育变革方案

北师大的“关键动作”并非孤立技术部署，而是围绕教育改革的证据链条，构建从数据采集、分析解释到策略生成与推广复用的一体化流程。

建设教育垂直基础能力：北师大将教育场景的异质数据（课堂行为、学情评价、资源使用、管理过程等）统一汇聚并结构化，形成面向教育的知识图谱与专业知识库，通过模型微调与检索增强生成等方式构建教育垂直大模型能力，强调“可调用、可审计、可持续迭代”的平台化供给，为不同学段、不同学科、不同地区的应用创新提供共

更多免费行业报告 并购家 www.ipoipo.net

用底盘。

以“AI+循证教研”推动“双新”落地：北师大举办“AI+循证教研赋能‘双新’学术年会”，面向教育行政与一线学校发布年度方案，提出从平台搭建、课例研究到成果推广的递进式推进体系，设置分阶段任务清单，并以“AI 大先生”智能系统支撑实施。相关公开材料进一步显示，该方案以 6 大阶段、22 项重点任务组织推进，并通过课堂智能观测、多模态学情分析等关键技术应用，将教研从经验驱动转向证据驱动。

将教研活动“流程化、产品化”：在组织机制上，北师大以联盟为载体推进课例分析与循证教研实践，形成跨区域协作的案例库与方法共同体，推动形成一站式解决方案。公开材料显示，联盟已覆盖 20 多个区域、200 多所学校，并汇集多个学段与学科的大量课例，为循证教研提供智能化、专业化支持。这一做法的可复制性在于：把“教研”从一次性活动转化为可持续运行的工作流体系，并以平台日志、证据链与标准化模板支撑规模化推广。

（4）以语言文明资源平台化支撑数智人文

北师大在数智人文方向突出“资源平台化+智能化工具链”路径，强调把专业研究能力转化为可被公众与教育系统共同使用的公共能力。

《说文解字》研究与应用平台升级。联盟简报显示，北师大汉字汉语相关实验室对平台进行了全方位改版升级：电脑端更名为《说文

解字》研究与应用平台，移动端命名为“北师大《说文解字》”，移动端增设古文字识别辅助功能；检索方式从单字检索扩展到部首、部件等多种检索路径，并新增重要文献、更新图像版本、修订古音数据并全面纠错。

“汉字全息资源应用系统”等平台化探索。公开报道指出，北师大围绕汉字整理与规范研究，利用智能化手段建立大型资源应用平台，并进一步打通资源与专家研究系统，加快传统人文学科研究，相关国家语委项目在此基础上形成阶段性成果。

学术共同体与人才培养。北师大承办或主办训诂学等基础学科高水平学术活动，推动传统学科在数智时代的传承创新，形成“学科共同体—数据资源—智能工具”相互促进的生态。

（5）经验启示

北师大的实践为高校推进 AI4SSH 提供了三点可迁移经验：

以重大公共任务牵引交叉融合。以教育改革与教育治理现代化为“真实需求端”，能有效避免 AI 应用停留在演示性创新，推动形成可检验、可迭代的研究—实践闭环。

把底座做成平台，把应用做成工作流。教育垂直大模型与数据中台解决“共用能力”问题；循证教研把改革过程拆解为可复用的工作流模板，实现从个案到规模的迁移。

用联盟机制解决扩散与治理。跨区域联盟将技术—方法—案例以共同体方式组织起来，在标准、培训与反馈机制支撑下，形成可持续

更多免费行业报告 并购家 www.ipoipo.net

的扩散通道，同时也为治理提供透明度与可追溯性基础。

2. 复旦大学早期中华文明大模型

（1）数字人文的范式跃迁：从工具辅助到认知重构

近年来，人工智能特别是大语言模型与多模态模型的快速发展，正在深刻改变人文社会科学研究的技术条件与知识生产方式。回顾其演进脉络，数字技术进入人文学科并非始于当下：从早期“人文计算”阶段的文献整理、文本检索与统计分析，到数字人文阶段的文本挖掘、知识组织与可视化应用，计算技术长期主要作为研究辅助工具而存在。进入生成式人工智能时代之后，模型在自然语言理解、内容生成、跨模态处理和复杂任务编排等方面取得突破，使其逐步从单纯的工具性技术扩展为能够参与知识处理、研究支持与公共服务的综合性基础设施，人文学科正经历着从“数字人文”向“智能人文”的范式跃迁。这一转变不仅仅是研究工具的迭代，更标志着技术介入层面的深刻变化，AI 不再仅仅是提升效率工具，而是凭借其自然语言处理与逻辑推理能力，开始介入知识生产的内容生成与意义构建。

当前，人文领域的 AI 研究总体上呈现出两个并行趋势。其一，是应用场景不断拓展，人工智能已广泛进入文献数字化、古籍识别、信息抽取、知识图谱构建、历史地理分析、图像处理和智能检索等研究环节；其二，是建设目标不断升级，研究重点正由面向少数专家的专门工具，逐渐转向面向科研、教学与公共文化传播的综合平台。与此同时，这一领域也面临若干共性难题：一是多源异构数据整合不足，尤其在人文研究中，文本、图像、地图、拓片、简帛、甲骨等材料类型复杂，标准不一，难以直接转化为高质量训练语料；二是文化语境

更多免费行业报告 并购家 www.ipoipo.net

理解能力仍然有限，通用模型在处理具有历史深度、语义歧义和强解释性的中文传统文献时，往往难以充分把握其语词系统、知识结构与价值内涵；三是从技术能力到学术应用之间仍存在转化鸿沟，很多系统能够完成局部任务，却尚难形成稳定、可持续、可推广的人文研究支持体系。项目材料所总结的时空多源异构数据组织、文化语境深度理解、古籍数字化与古汉语信息处理等问题，正集中反映了这一领域的发展瓶颈。

（2）时代潮流的强劲推动：国家战略与文明主体性的双重驱动

2024 年 10 月 28 日下午，中共中央政治局就建设文化强国进行第十七次集体学习。中共中央总书记习近平在主持学习时指出，要在创造性转化和创新性发展中赓续中华文脉。高扬中华民族的文化主体性，把历经沧桑留下的中华文明瑰宝呵护好、弘扬好、发展好。植根于国家对于中华优秀传统文化创造性转化、创新性发展的战略需求。党中央高度重视运用现代科技手段加强对古籍典藏的保护修复与综合利用，旨在深入挖掘其中蕴含的哲学思想、人文精神与价值理念。在此宏观愿景下，发展人工智能技术被提升至维护国家安全、促进社会进步的战略高度，要求技术发展必须确保安全、可靠、可控，并与中国经济社会发展深度融合。

其次，项目的立项是对当前通用大模型存在的“西方中心主义”偏见与价值观对齐难题的直接回应。现有的大规模语言模型主要依赖于以英语为主的互联网语料进行训练，这导致模型在处理非西方文化语境时，往往难以准确理解和生成符合本土文化逻辑的内容，甚至可

能放大社会不公与文化偏见。对于拥有独特时空维度、深厚历史积淀的中华早期文明而言，构建具有自主知识产权的垂直领域大模型，是打破西方数据霸权、确立中华文明在数字空间中主体性的关键举措。

（3）学科痛点与技术挑战：早期中华文明研究的瓶颈

尽管国家战略层面高度重视，但在微观实践层面，早期中华文明研究仍然面临诸多挑战。

数据多样性与结构化困境：早期中华文明研究涉及考古、古文字、早期典籍及历史地理四大支柱，数据呈现显著的“多源异构”特征。目前面临的主要挑战是如何将甲骨、金文、简牍帛书等非结构化文本，以及古代舆图、遗址影像等复杂图像数据，转化为机器可读的半结构化或结构化语料。特别是古文字识别（OCR）与历史地理信息的时空坐标系定位，仍存在巨大的技术鸿沟。

文化语境理解的复杂性：古汉语具有含蓄的表达方式、多重指称变体及深厚的历史文化常识依赖，通用的自然语言处理技术难以胜任。如何构建能够深度理解“时空多模态”文化语境的大模型，使其不仅能识别文字，更能解析背后的历史脉络与文化逻辑，是当前的核心技术壁垒。

伦理与治理体系的缺失：随着 AI 在文化传播与教育领域的深入应用，缺乏专门针对中华文化输出的伦理审查机制成为一大隐患。如何建立符合中华核心价值观的多文化背景伦理委员会，确保模型输出的安全性与准确性，是项目必须解决的治理难题。

（4）从数据库到大模型再到应用：早期中华文明大模型的建设任务

更多免费行业报告 并购家 www.ipoipo.net

针对目前存在的研究痛点和技术挑战，早期中华文明大模型项目主要分成三个子任务，从基础到应用层层递进：

首先，构建基于时空维度的多模态早期中华文明数据库。该数据库旨在整合考古、古文字、早期典籍及古代历史地理等多个领域的数据资源，支持多语种数据的存储与检索、支持多模态检索技术。该数据库的具体功能包括多模态数据结构化、古文翻译系统、图像配准与修复、地理信息系统（GIS）应用。

其次，在数据库的基础上发展多模态早期中华文明时空大模型。中华文明大模型在模型架构上探索了基于离散扩散的多模态统一理解与生成基础模型路线，支持对文本、图像、文献、文物资料及结构化知识等多源数据的统一建模。在此基础上，项目通过面向人文社科领域的中期训练，将历史学、考古学、文献学、艺术史等学科中的研究方法、证据组织方式和推理路径转化为指令化任务数据，推动人文社科思维方式向模型能力的系统注入。经过训练适配，模型在多模态语义理解、跨文献证据关联、复杂问题分解、长程研究任务规划和工具协同调用等方面形成了能力基础，可为古籍整理、文物图像解读、历史事件分析和文明要素关联等研究场景提供智能化支撑。

最后，将本项目所建设的数据库和大模型进行应用。我们力图将多模态早期中华文明数据库和时空大模型应用于教育、研究及社会文化传播等领域。为教育工作者和学生提供丰富的文化教育资源、教学工具；为学者提供高效的数据库和智能分析工具；将中华优秀传统文化融入现代社会生活，加深公众对早期中华文明的理解与认同，助力

文化自信的建立。

（5）高质量数据及充足算力：项目现有的基础资源

开展早期中华文明大模型项目，数据和算力是最重要的基础资源。截至目前，本项目所掌握的基础资源情况如下：

首先是数据基础。对于古籍，复旦大学图书馆目前已完成 4 千余种数字化古籍，约百万册页。囊括了古籍中的主要部类，包括经史子集类。并重点涵盖了经学类资料。除自主数字化之外,图书馆还选订了大量商业化的中文古籍数据库资源，共 21 种。另外还有大量公开的古籍数字人文数据库可供使用。对于古文字，本项目将利用复旦大学出土文献与古文字研究中心承担的“中华字库”工程中“金文的搜集与整理”“楚简、帛书及其他古文字的搜集与整理”两个子项目的数据，并利用其“上博简字词全编”“‘缀玉联珠’甲骨缀合信息库”以及“先秦秦汉古汉字资料数据库”等数据库资源。对于历史地理，本项目不仅参考周振鹤先生主编的《中国行政区划通史》和谭其骧先生主编的《中国历史地图集》等权威历史地理著作，还吸收复旦大学与哈佛大学合作开发的“中国历史地理信息系统”（CHGIS）的研究成果。同时，收集整理中国古代地方志和古旧地图、建设历史地名库。对于考古领域的数据，得益于考古领域数字化程度的逐步提高，已经形成一个真正的跨学科综合数据库，目前总数据量近 200T。现代语言数据库当前上线数据已超过 55TB、150 多亿条，且保持日均 1000 万条以上速度持续索引入库。

其次是数据平台基础。复旦大学联合上海科学智能研究院开发的

更多免费行业报告 并购家 www.ipoipo.net

数据平台，具备数据采集、数据汇交、数据加工、数据生成与增强、数据存储、数据资产管理、数据安全、数据检索、数据分析等能力。数据平台完成了 30 个子领域、198 个来源的数据采集汇聚方案，形成了常态的数据源梳理保鲜机制，已汇聚包括 NCBI、GEO、ECMWF 等 120 个海外科学数据集 4PB 数据。

最后是算力基础。复旦大学建设了 CFFF（Computing For the Future at Fudan）异构智算平台，包括切问一号和近思一号集群，CFFF 总体算力规模超过 28PFlops（28 千万亿次/秒），是全国高校第一的超算集群，同时也是算力第一的异构智算集群。有力支撑了人工智能赋能生命科学、大气科学、材料科学的研究，更是本项目的重要算力支持。

（6）项目进展：首个“人文社科+大模型”公共服务平台

在丰富的理论和资源支持之下，我们的项目将推出最新的成果：“人文社科+大模型”公共服务平台。

该平台将全校师生、公众、泛文化产业作为我们的服务对象，将应用场景拓展至科研、教学和公共文化，以低门槛和全开放的姿态迎接不同目的和不同知识储备的使用者。这是一个教研用融合的公共服务平台。

为了构建这个平台，我们升级大模型内化人文社科方法论，通过 Skills 定义科研任务，整合人文社科科研工具，从科研场景扩展到教学和数字文化应用。在科研场景，以多模态底座与自定义 Skills 驱动学术工作台，以 Chat-First 重塑学术工作流，打造零幻觉、可溯

源的意图驱动研究新模式。助力学术成果产出、加速学术资产沉淀、支撑重大课题项目申报结项。在教学场景，我们以“通识教育智能体+Agent/Skills/数据实训”为基础，赋能教育教学并提升在校所有文科学生在自己学科学习和研究上的的 AI 应用能力。在数字文化应用方面，我们采取“Agent 应用+开放 API”模式，通过 Agent 应用或开放模型/Agent API 接口，落地数字文化场景应用。

为了构建这个平台，我们还进行了多模态数据体系构建和模型评测集开发。数据自然提取算法可以进行现代研究图文自动提取、古籍 OCR 以及古文字图谱构建等任务；数据开放平台完成古籍书库、现代研究、多模态史料和历史实体、古文字知识图谱导入。本项目的模型评测集是全球第一个关注早期中华文明的多模态模型评测集。涉及考古、古文字、历史地理、语言学等领域的 10000+道评测题目。

最后，我们的平台已经可以用于实际的科研工作，项目小组已经利用平台进行了“社会学仿真 Agent”“湖南省博物院&马王堆汉墓出土文献与文物研究”“中华文明在西方”“历史地理中华文明源流 Agent”“历史地理信息平台 Agent”等人文社科课题及“AI+人文”项目的科研实践，成果斐然。

更多免费行业报告 并购家 www.ipoipo.net

3. 复旦大学 Insurance-Skills——面向金融保险智能体的行业技能基础设施建设

随着大语言模型能力提升，智能体（Agent）正在从通用问答工具发展为能够理解任务、调用工具、执行流程并形成反馈闭环的智能系统。与传统聊天机器人不同，智能体不只是生成文本回答，而是能够围绕具体目标进行任务拆解、知识检索、工具调用和结果交付。在保险场景中，智能体可被用于产品咨询、投保核保、保单服务、理赔处理、客户服务、营销推荐、合规风控与运营支持等多个环节。

然而，保险行业具有强专业、强流程、强监管和高风险特征。通用智能体若直接进入保险业务，往往会面临专业知识不充分、业务流程不清晰、合规边界不明确、工具调用不可控等问题。例如，在理赔场景中，智能体不仅需要回答用户“需要准备哪些材料”，还需要理解险种责任、理赔流程、材料核验、异常情形、人工复核边界与监管要求；在核保场景中，智能体也不能仅凭语言生成能力给出判断，而必须遵循明确的规则、条件和风险提示。

因此，保险智能体的关键不只是“模型是否足够强”，而是是否具备面向具体业务场景的专业能力单元。这里的技能（Skill），可以理解为智能体的专业技能包：它将某一类保险任务中的业务规则、流程步骤、知识口径、工具接口和风险边界封装起来，使智能体能够以更标准、可控和可复用的方式完成特定任务。

当前，保险技能的应用仍存在若干突出问题。第一，高质量、强

场景化的技能供给不足，许多能力仍停留在通用提示词或参考文档层面。第二，技能的设计和表达缺乏统一规范，不同来源的技能在结构、粒度、输入输出和边界说明上差异较大。第三，同类技能质量参差不齐，用户难以快速判断其完整性、可操作性和安全性。第四，可复用和可集成程度不足，许多技能依赖特定数据、系统或上下文，难以迁移到其他机构和业务流程中。第五，风险治理机制尚不完善，数据隐私、合规监管、错误指导和外部依赖等问题可能影响保险智能体的可信落地。

针对上述问题，复旦大学经济学院许闲教授团队建设并开源发布了 Insurance-Skills 平台，旨在面向保险全生命周期构建一个保险技能的汇聚、解析、测评与复用基础设施，为保险智能体从概念验证走向规模化应用提供能力底座。目前，平台已开放在线体验，访问地址为：https://skills.fduinsurance.com ； GitHub 开源项目地址为：https://github.com/FDU-INS/Insurance-Skills。

（1）场景牵引：以保险全生命周期重构智能体能力单元

Insurance-Skills 平台的核心定位，是将保险领域分散存在的业务知识、流程规则、工具能力和风险约束，转化为可发现、可理解、可比较、可评估、可复用的技能资产。平台并非单一保险智能体，而是服务于保险智能体构建的能力基础设施。

更多免费行业报告 并购家 www.ipoipo.net



图 12-1　Insurance-Skills 平台首页图

首先，平台提供保险技能的多渠道汇聚与场景化分类。围绕保险全生命周期，平台持续收集和整理保险相关技能，覆盖产品咨询、保单服务、核保支持、理赔服务、客服助手、营销与推荐、合规与风控、运营支持、培训与知识问答等高频业务环节。通过标签体系和场景分类体系，用户可以从业务场景、功能方向和能力类型等维度快速定位相关技能，避免在分散渠道中反复搜索和筛选。

其次，平台提供技能的标准化展示与功能解析。对于每一个技能，平台不仅展示名称和基本描述，还围绕功能说明、适用场景、使用方式、依赖条件、输入输出和风险边界等信息进行结构化整理。这一机制有助于保险公司数字化团队、产品技术团队和开发者快速理解某一技能能解决什么问题、适用于什么场景、需要哪些条件、可能存在哪些限制，从而提升能力选型和集成效率。

再次，平台建立保险技能的多维测评体系。针对同一场景下技能质量难以比较的问题，Insurance-Skills 从清晰度、完整度、可操作性、可维护性和安全性五个维度进行评价与打分。清晰度关注功能边界和表达是否明确，完整度关注任务步骤和必要信息是否充分，可操作性关注是否具备实际调用和执行条件，可维护性关注版本、依赖和扩展能力，安全性则关注隐私、合规、误导性输出和外部依赖风险。通过测评机制，平台将原本依赖经验判断的能力选择过程，转化为更透明、更可比较的质量评估过程。



图 12-2　保险智能体 Skill 测评详情图

最后，平台支持开源共建与生态扩展。Insurance-Skills 通过开源项目和在线平台向开发者、研究者和行业机构开放，鼓励围绕保险技能

更多免费行业报告 并购家 www.ipoipo.net

的补充、优化、测评和应用展开协作。对于保险机构而言，平台可作为智能体能力选型和初步评估入口；对于开发者而言，平台提供了进入保险垂直场景的标准化参考；对于研究者而言，平台积累的数据与测评结果也为观察保险智能体能力生态提供了基础。

（2）基于平台数据的核心洞察

基于 Insurance-Skills 平台汇聚的公开样本，团队同步发布了《保险“龙虾”技能洞察报告》，对保险智能体技能市场进行了系统分析。报告基于 539 个公开技能样本，从供给规模、场景分布、能力类型、复用程度、风险结构和质量短板等方面揭示了当前保险技能生态的若干特征。

第一，市场增长较快，但有效供给仍然有限。报告显示，团队从万级检索结果中最终筛选出 539 个有效样本，其中 2026 年 3 月单月新增 287 个，说明保险技能供给正在快速放量。但从筛选结果看，大量内容与保险业务关联度不足，或难以形成可用技能，说明市场热度尚未充分转化为高质量能力供给。

第二，场景分布高度集中，重点聚焦理赔、合规、风控和承保等环节。在有效样本中，合规审查内控和数据分析等方向占比较高，反映出当前保险智能体建设更关注规则密集、信息处理需求高、降本增效预期明确的业务领域。相比之下，覆盖完整业务闭环、跨环节协同或面向复杂客户旅程的技能仍然较少。

第三，能力形态以辅助型为主，距离自主执行仍有差距。报告指

出，71.5%的样本属于参考文档型，仅 8.7%支持自主工具调用。这说明当前多数技能仍主要提供知识说明、流程参考或话术辅助，尚未真正具备任务执行、系统交互和流程闭环能力。

第四，复用性较弱，跨机构迁移难度较高。报告显示，仅 0.7%的样本可独立运行，说明大量技能对特定系统、数据、接口或业务环境存在较强依赖。这也意味着，保险智能体能力建设不能只依靠零散技能堆叠，而需要进一步完善标准化描述、依赖管理和集成规范。

第五，安全风险较为集中。团队在样本中识别出 1937 条风险项，其中数据隐私泄露、合规监管和外部依赖是主要隐患。保险业务涉及大量个人身份、健康、财务和合同信息，若技能在数据调用、输出建议或责任边界方面缺乏控制，可能带来较高风险。

第六，质量短板具有共性。报告显示，样本综合均分为 7.73，弱项主要集中在示例、章节覆盖、错误指导和版本跟踪等方面。这说明许多技能虽具备基本说明，但在可执行示例、异常场景处理、持续维护和风险提示方面仍不充分。

总体来看，平台数据表明，保险技能生态已经进入快速发展阶段，但仍处于从“数量增长”向“质量提升”过渡的早期阶段。未来保险智能体落地的关键，将不仅是构建更多技能，而是提升技能的场景适配性、结构完整性、工具调用能力、可迁移性和安全治理水平。

（3）小结

从 AI4SSH 视角看，Insurance-Skills 平台体现了高校参与行业智

更多免费行业报告 并购家 www.ipoipo.net

能化转型的一种典型路径：以真实行业场景牵引跨学科协作，以平台化方式沉淀知识、规则和工具资产，以标准化测评提升质量与透明度，以开源共建推动能力扩散和生态形成。它说明，高校不仅可以在 AI 与社会科学交叉中提出理论和模型，也可以进一步建设面向行业和社会服务的公共能力基础设施。

12.3 综合评述：从点状探索到体系建制

在 AI 能力快速扩散、方法门槛持续降低的背景下，中国高校如何把零散的项目探索转化为可持续的组织能力，把学术创新转化为可复制的公共能力，并在合规、伦理与风险可控前提下形成面向国家战略与社会治理的长期供给。教育部高校哲学社会科学实验室联盟近一年的工作简报显示，越来越多高校正在以实验室建制、平台化底座与联盟协同的方式推进 AI+人文社科融合，议题覆盖数字政府、教育数字化、文化遗产保护、数据法治、国际传播、低碳治理等多类公共任务，呈现出由“工具应用”走向“范式重构、组织重构与制度供给”的阶段性跃迁。

1. 总体态势：建制化提速，AI4SSH 进入“平台—规范—生态”竞争阶段

从联盟简报与高校公开信息可以观察到三点总体变化。第一，组织载体从科研项目向实验室与平台转移。越来越多高校以教育部哲学社会科学实验室、校级交叉平台和创新中心为抓手，将算力、数据、工具链与研究管理纳入常态化运行体系，强调长周期能力积累而非一次性成果。第二，研究对象从“文本与变量”扩展到多模态社会过程与复杂决策系统。大模型、智能体与多源数据融合，使社会科学不仅能做“解释”，也能做“监测—评估—推演”的连续服务，推动证据生产从论文形态走向可交付的工具与流程。第三，生态建设从校内合作走向跨校联盟与开放合作。暑期学校、学术论坛、联盟年会与开放课题等机制增多，其作用不只是传播成果，更在于形成共同语言、共

更多免费行业报告 并购家 www.ipoipo.net

同口径与可复用工作流模板，降低跨学科、跨机构协作成本。

2. 组织模式谱系：四类路径正在形成“可迁移的组织模板”

结合案例与简报材料，高校推进 AI4SSH 的组织模式大体呈现四类可迁移路径。（1）平台型基础设施路径：以算力、数据与工具链为底座，将 AI 能力沉淀为“可调用、可审计、可复用”的公共能力模块，典型做法包括建设统一算力平台、开放数据平台、模型库与工具库，并通过数据治理与权限管理实现合规共享。（2）场景牵引的工程化路径：以文化遗产保护、教育改革、数字政府等高复杂公共任务为牵引，把跨学科合作组织成项目制“作战单元”，用真实场景倒逼数据工程、模型评测与流程规范，形成可复制的解决方案与治理工具。（3）方法论中枢路径：围绕因果推断、社会仿真、风险评估与政策评测等关键方法能力建设评测基准、工具链与训练体系，使方法从‘个人手艺’转化为‘团队可协作的公共流程’，并通过标准化 schema 与日志机制提升可复核性。（4）联盟扩散与能力外溢路径：通过跨区域、跨学校联盟，把平台能力、培训体系、案例库与模板化工作流打包输出，实现从单校试点走向规模化推广，同时也为治理透明度、评测口径一致性与风险共担提供组织基础。

3. 关键机制：把能力固化为制度与流程

高校 AI4SSH 推进的难点不在“有没有模型”，而在“能否长期稳定地产出可信证据与可用产品”。为此，若干机制正在成为共识。一是数据治理与合规流通机制。人文社科数据往往分散、敏感且跨主体，必须通过分级分类、授权使用、可追溯日志与‘数据可用不可见’等

制度安排，才能实现跨团队共享与复现。二是工程师队伍与产品化能力。多校实践表明，缺乏稳定的数据工程与工具开发队伍，平台很难持续迭代；而一旦形成‘学术团队—工程团队’协同，研究就能从论文输出延伸到指标、模型、工具与服务的持续供给。三是质量控制与问责机制。包括统一评测协议、压力测试、红队评估、事故报告与回滚机制，以及对结论适用范围与不确定性的披露要求。对公共治理相关应用而言，这些机制决定成果能否进入决策流程。四是人才培养与共同语言建设。暑期学校、专题训练营、跨学科研讨与实践课程的意义，正在从‘技能培训’转向‘共同语言与共同工作流’的构建，降低跨学科协作的长期摩擦。

4. 主要瓶颈与风险：能力扩散快于制度供给，协同成本仍高

在快速推进的同时，也存在若干共性瓶颈。首先，AI-ready 数据不足与数据孤岛仍是普遍约束。许多领域的数据质量、标注口径与长期维护机制不足，导致模型能力难以稳定复现，跨校共享难以规模化。其次，激励与评价体系尚未充分适配“平台化科研”与“团队化交付”。对工程化平台建设、数据资产沉淀、评测基准维护等工作，传统评价往往难以准确计量，容易造成‘平台建设靠热情、可持续性不足’的困境。再次，合规、伦理与安全治理仍处在追赶阶段。生成式能力带来的真实性风险、偏见与操控风险，以及在教育、司法、公共服务等高影响场景中的责任链条问题，都需要更系统的制度化回应。最后，跨学科协同成本仍高。不同学科在问题定义、证据标准与解释语言上的差异，决定了‘协作’不能只靠工具，需要通过流程与制度

更多免费行业报告 并购家 www.ipoipo.net

把协作前置化、规范化。

5. 趋势判断与建议方向：以公共能力为导向，形成可迭代的“研究—治理”闭环

面向下一阶段，高校 AI4SSH 的竞争将更多体现为“公共能力供给能力”的竞争。本蓝皮书建议的方向可概括为三条主线。第一，以应用验证校准能力边界，形成面向公共任务的“试验场”。在教育、文化遗产、数字政府、低碳治理等领域，推动试点从展示型走向评估型：明确目标、指标、对照与复盘机制，用真实数据与过程日志减少信息不对称，避免对技术的迷信或误判。第二，以平台化底座降低协作成本，把方法能力沉淀为公共基础设施。重点建设可共享的数据资源、模型与工具库、评测基准与合规模板，使中小团队与新进入者能够在统一底座上开展创新，形成‘向上竞争’的生态。第三，以制度改革释放跨学科创新的长期激励。包括对平台建设与数据资产的评价机制、对跨学科成果的认定规则、对工程师与复合型人才的岗位与晋升通道，以及对高风险应用的审计与问责机制。当这些机制逐步完善，高校推动 AI+人文社科将更可能从“项目探索”走向“体系建制”，并在全球知识生产与公共治理体系中形成更可持续的中国方案。

附录：AI4SSH 领域前沿论文推荐

表 A1 经济学领域代表性论文

标题	作者	期刊	推荐理由
Artificial intelligence and corporate ESG performance	Li, JJ; Wu, T; Hu, BQ; Pan, DL; Zhou, YQ	INTERNATIONAL REVIEW OF FINANCIAL ANALYSIS	AI 通过绿色创新提升 ESG 绩效
Artificial intelligence digital employees and sustainable innovation in online retail: The mediating role of ambidextrous green innovation and the moderating role of ethical anxiety	Wang, SF; Zhang, H	JOURNAL OF RETAILING AND CONSUMER SERVICES	数字员工通过双元绿色创新驱动可持续绩效
Industrial robot diffusion and regional innovation disparities	Wu, WW; Wang, ZF	TECHNOLOGY IN SOCIETY	工业机器人扩散加剧区域创新差距的机制
Does public data openness accelerate new quality productive forces? Evidence from China	Zhong, Y; Lai, HS; Zhang, L; Guo, LX; Lai, XB	ECONOMIC ANALYSIS AND POLICY	公共数据开放驱动新质生产力
Artificial intelligence augmenting human intelligence for manufacturing firms to create green value: Towards a technology adoption perspective	Huang, L; Chin, T; Papa, A; Pisano, P	TECHNOLOGICAL FORECASTING AND SOCIAL CHANGE	AI 增强人力资本助力制造企业创造绿色价值
Can artificial intelligence technology reduce carbon emissions? A global perspective	Cao, QF; Chi, CY; Shan, JH	ENERGY ECONOMICS	AI 通过提升能源效率与引导技术变革降低碳排放
Does artificial intelligence matter for the population aging-inclusive growth nexus? International evidence	Wen, HW; Shang, JJ; Nghiem, XH	TELECOMMUNICATIONS POLICY	AI 助力应对人口老龄化对经济增长的挑战
The impact of artificial intelligence on labor market: A study based on bibliometric analysis	Zhou, HL; Wang, LH; Cao, YT; Li,	JOURNAL OF ASIAN ECONOMICS	文献计量揭示 AI 对劳动力市场的多维影响

更多免费行业报告 并购家 www.ipoipo.net

	JC		
Robots and firm total factor productivity: evidence from China	Li, BW; Liu, X; Zhou, C	INTERNATIONAL JOURNAL OF TECHNOLOGY MANAGEMENT	机器人应用对企业全要素生产率的促进效应
Artificial intelligence and corporate fraud: Evidence from China	Xia, Y; Xie, LL; Chen, X; Lin, Z	ECONOMIC ANALYSIS AND POLICY	AI 应用对企业财务欺诈行为的抑制效应

表 A2 管理学领域代表性论文

标题	作者	期刊	推荐理由
Artificial intelligence and innovation capability: A dynamic capabilities perspective	Gao, Y; Liu, SQ; Yang, L	INTERNATIONAL REVIEW OF ECONOMICS & FINANCE	AI 通过数字适应性与市场感知提升创新能力
When robot knocks, knowledge locks: how and when does AI awareness affect employee knowledge hiding?	Liu, Z; Lin, QX; Tu, SM; Xu, X	FRONTIERS IN PSYCHOLOGY	AI 意识通过心理可得性影响员工知识隐藏
Artificial intelligence, corporate information governance and ESG performance: Quasi-experimental evidence from China	Zhou, XY; Li, GC; Wang, QW; Li, YGX; Zhou, DQ	INTERNATIONAL REVIEW OF FINANCIAL ANALYSIS	探究 AI 通过信息治理提升企业 ESG 绩效
How and when AI-driven capabilities foster organizational sustainability: A moderated mediation analysis of digital literacy and green HRM	Ali, A; Yang, Q; Ali, A; Ali, Z; Noman, M	INTERNATIONAL JOURNAL OF INFORMATION MANAGEMENT	AI 能力通过数字素养与绿色人力资源管理提升组织可持续性
AI Technology Adoption, Knowledge Sharing, and Manufacturing	Lin, XY; Wu, D	IEEE TRANSACTIONS	AI 采纳通过知识共享提升创新绩效，

Firms Innovation Performance: The Moderating Effect of Absorptive Capacity		ON ENGINEERING MANAGEMENT	吸收能力调节
The Dual Role of AI: A Capable Assistant in Collaboration and a Scapegoat in Competition	Li, QS; Wang, CH; Wang, ZX	INTERNATIONAL JOURNAL OF HUMAN-COMPUTER INTERACTION	揭示 AI 在人机协作与竞争中的双重角色差异
Exploring the Roles of Corporate Social Responsibility and Artificial Intelligence Adoption in the Impact of Political Connections on Innovation Performance	Yao, MM; Xu, MD	SUSTAINABILITY	探究政治关联通过 CSR 与 AI 采纳影响创新绩效
Industrial Robot Imports and Firm Innovation: A Large Language Model	Yuan, YM; Chen, HY; Ge, Y; Huang, XH; Yang, GJ	WORLD ECONOMY	大语言模型方法评估工业机器人进口对企业创新影响
The impact of artificial intelligence usage on supply chain resilience in manufacturing firms: a moderated mediation model	Yue, XC; Kang, M; Zhang, YM	JOURNAL OF MANUFACTURING TECHNOLOGY MANAGEMENT	AI 应用通过供应链可视化提升制造企业供应链韧性
Artificial Intelligence-Supported Leadership : A Catalyst for Team Excellence in China's Fast-Moving Consumer Goods Industry	Shang, FK; Zhou, YJ; Comite, U; Badulescu, A; Badulescu, D	JOURNAL OF ORGANIZATIONAL AND END USER COMPUTING	AI 赋能领导力成为快消品行业团队卓越的催化剂

表 A3 历史学领域代表性论文

标题	作者	期刊	推荐理由
Using natural language processing to evaluate local conservation text: A study of 624 documents from 303 sites of the World	Chen, Y; Zhang, LC; Dong, Q	JOURNAL OF CULTURAL HERITAGE	评估世界遗产城市地方保护文本的实践与特征

更多免费行业报告 并购家 www.ipoipo.net

Heritage Cities Programme			
A Study of the Initial System of the Yongle Nanzang Based on Phonological Correlations and Their Relationship with the Qishazang	Jiang, YC; Wang, BX; Huang, RX	RELIGIONS	探究永乐南藏音系关联与版本源流
Shedding new light on lacquering crafts from the Northern Wei Dynasty (386-534 CE) by revisiting the lacquer screen from Sima Jinlong's Tomb	Zhu, ZY; Qin, YL; Guo, ZY; Cai, SQ; Lin, PX; Wang, XJ; Cai, JQ; Yang, JC	JOURNAL OF CULTURAL HERITAGE	重探司马金龙墓屏风漆艺与北魏工艺
Periodical geography and geographical correspondence in the Naval Chronicle (1799-1818)	Ding, YN	JOURNAL OF HISTORICAL GEOGRAPHY	解析 *Naval Chronicle* 的期刊地理与书信网络
Spatiotemporal dynamics of prehistoric millet agriculture dispersal in Northwest China	An, CB; Tang, L	ARCHAEOLOGICAL RESEARCH IN ASIA	重构史前黍作农业在西北地区的时空扩散过程
Spatiotemporal dynamics of Catholic church heritage in Sichuan using density weighted historical segmentation and SHAP	Zhang, HZ; Li, XK; Zhang, BB; Li, YZ; Liu, HY	NPJ HERITAGE SCIENCE	密度加权历史分段揭示四川天主教堂时空动态
The construction of thematic platforms for Chinese old maps: a global overview	Dang, Y; Hu, Q; Li, S	DIGITAL SCHOLARSHIP IN THE HUMANITIES	全球视野下中国古地图专题平台建设综述
OracleNet: enhancing Oracle Bone Script recognition with Adaptive Deformation and Texture-Structure Decoupling	Zhou, S; Wang, X; Qiu, JW; Bu, WR; Wang, H	NPJ HERITAGE SCIENCE	自适应形变与纹理解耦提升甲骨文识别能力
Characteristics and trends in Chinese Historical Geography (2019-2024)	Li, RH; Zhang, XH; Li, S	HISTORY OF SOCIAL SCIENCES; GEOGRAPHY	综述中国历史地理学近年研究特征与趋势
A Study of the Spatial Correlogram Patterns of Chinese Surnames	Fan, XH; Zhang, XM; Gao, Y; Liu, Y; Li, XM;	AMERICAN JOURNAL OF HUMAN BIOLOGY	探究中国姓氏空间自相关图式与历史分布规律

	Yuan, YD; Chen, LJ; Chen, JW		

表 A4 公共管理领域代表性论文

标题	作者	期刊	推荐理由
When AI fails, who gets the blame? Citizens' attribution patterns in AI-induced public service failures	Liang, ZH; Li, Y; Chen, T	JOURNAL OF CHINESE GOVERNANCE	实验揭示公民对 AI 公共服务失败的归因模式
Intelligent government: The impact and mechanism of government transparency driven by AI	Zhao, X; Huo, YX; Abedin, MZ; Shang, YP; Alofaysan, H	PUBLIC MONEY & MANAGEMENT	AI 通过优化行政效率与开放数据提升政府透明度
Can Cross-Sector Collaboration Contribute to Boundaries Reshaping of Digital Government Platforms? Empirical Evidence Based on Machine Learning and Text Analysis	Zhang, SP; Yang, YH; Gan, HH	PUBLIC ADMINISTRATION	跨部门协作重塑数字政府平台边界的作用机制
Holding AI-Based Systems Accountable in the Public Sector: A Systematic Review	Yuan, QL; Chen, TH	PUBLIC PERFORMANCE & MANAGEMENT REVIEW	系统综述公共部门 AI 系统问责机制与挑战
Artificial intelligence policy frameworks in China, the European Union and the United States: An analysis based on structure topic model	Wang, SR; Zhang, YM; Xiao, YM; Liang, Z	TECHNOLOGICAL FORECASTING AND SOCIAL CHANGE	结构主题模型比较中美欧 AI 政策框架异同
Government adoption of generative artificial intelligence and ambidextrous innovation	Zhou, ZK; Liu, DW; Chen, ZJ; Pancho, M	INTERNATIONAL REVIEW OF ECONOMICS & FINANCE	探究政府采纳生成式 AI 对双元创新的影响

更多免费行业报告 并购家 www.ipoipo.net

Artificial intelligence vs. public administrators: Public trust, efficiency, and tolerance for errors	Bao, HX; Liu, WF; Dai, Z	TECHNOLOGICAL FORECASTING AND SOCIAL CHANGE	比较公众对 AI 与行政人员的信任与容错差异
Public trust in artificial intelligence users	Dang, JN; Liu, L	CURRENT OPINION IN PSYCHOLOGY	探究公众对人工智能使用者的信任形成机制
The Influence of Public Expectations on Simulated Emotional Perceptions of AI-Driven Government Chatbots: A Moderated Study	Guo, YY; Dong, P; Lu, BC	JOURNAL OF THEORETICAL AND APPLIED ELECTRONIC COMMERCE RESEARCH	公众期望调节对 AI 政务聊天机器人情感感知的影响
How does scientific research influence policymaking? A study of four types of citation pathways between research articles and AI policy documents	Cao, Z; Zhang, L; Huang, Y; Sivertsen, G	JOURNAL OF THE ASSOCIATION FOR INFORMATION SCIENCE AND TECHNOLOGY	揭示科研论文影响 AI 政策制定的四种引文路径

表 A5 信息管理领域代表性论文

标题	作者	期刊	推荐理由
Inter-organizational knowledge sharing in the age of artificial intelligence: the mediating role of knowledge digitization and cross-language collaboration	Cheng, Q; Zhang, SY; Wang, LL; Lu, Y	JOURNAL OF KNOWLEDGE MANAGEMENT	AI 通过知识数字化与跨语言协作促进组织间知识共享
Examining generative AI user continuance intention based on the SOR model	Zhou, T; Ma, XJ	ASLIB JOURNAL OF INFORMATION MANAGEMENT	基于 SOR 模型揭示生成式 AI 用户持续使用意愿的形成机制
Artificial humanity: A multi-method exploration of user responses to AI	Lee, CT; Shen,	TELEMATICS AND	探究用户对短视频平台 AI 博主功能

influencer affordances in short video platform	YC	INFORMATICS	的多元反应与认知
AI challenges conventional knowledge management: light the way for reframing SECI model and Ba theory	Zhang, WY; Zhang, W; Daim, T; Yalcin, H	JOURNAL OF KNOWLEDGE MANAGEMENT	AI 推动 SECI 模型与 Ba 理论重构革新
How does AI affect the self-actualization of content creators in dynamic environments? A knowledge management perspective	Wang, MS; Yang, YZ; Zhong, PC	TECHNOLOGY IN SOCIETY	知识管理视角下 AI 对内容创作者自我实现的影响
A Text-Mining-Based Evaluation of Data Element Policies in China: Integrating the LDA and PMC Models in the Context of Green Development	Hu, SG; Wang, XB	SUSTAINABILITY	LDA-PMC 融合模型评估中国数据要素政策
Integration of public libraries and cultural tourism in China: An analysis of library attractiveness components based on tourist review mining	Jiang, TT; Xu, YR; Li, Y; Xia, YK	INFORMATION PROCESSING & MANAGEMENT	游客评论挖掘分析公共图书馆文旅融合吸引力
The smart archive management practicing pipeline: a virtual online learning platform for AI literacy development in archival science	Dong, LX; Tang, SN; Cheng, Y; Wang, P	INFORMATION RESEARCH-AN INTERNATIONAL ELECTRONIC JOURNAL	构建档案学 AI 素养虚拟在线学习平台
Research on the Biography Digital Storytelling Based on Ontology	Si, L; Liu, Y	KNOWLEDGE ORGANIZATION	基于本体的传记数字叙事方法与研究
How to Digitalize Social Enterprises: Evidence From Private Museums in China	Fan, XY; Liu, Y; Chen, AL; Shou, YY	BRITISH JOURNAL OF MANAGEMENT	探究中国私立博物馆数字化转型的路径与机制

更多免费行业报告 并购家 www.ipoipo.net