

生成式人工智能对人类心智多样性的重塑：机制、实证与演化收敛

“人脑智能”课程学术调研报告

课程名称：人脑智能 (Human Brain Intelligence)

报告题目：AI 会不会让人类智力趋同？

汇报立场：正方观点（AI 必然导致人类智力的结构性趋同）

报告撰写人：课程学术汇报小组

完成日期：2026年10月

核心摘要：本报告系统梳理了近三年认知神经科学、计算语言学、演化复杂系统以及科学社会学关于生成式人工智能对人类心智影响的顶刊实证研究。通过对《Nature》、《Science Advances》、《Nature Human Behaviour》、《Trends in Cognitive Sciences》等八篇代表性研究的元分析，本报告论证了 AI 促成人类智力趋同的四重收敛机制：生理层的认知卸载与神经突触修剪钝化、表征层的概率模式捕获与语言方差压缩、方法论层的科学单一栽培以及演化层基于合成数据的模型坍缩闭环。研究表明，尽管 AI 显著抬升了个体任务表现的均值，但群体智力探索的方差却遭遇断崖式收窄。人类智力正从“异质性生物星丛”演变为“中心化算法投影”。

目录

- 引言：智力维度的再审视——均值与方差的哲学区隔
- 神经生理机制：认知卸载、自由能原理与用进废退
- 表征与符号学机制：统计模式均值化与语言多样性雪崩
- 集体行为实证：个体创意均值抬升与群体异质性收缩
- 认识论危机：科学研究的“单一栽培”与理解的幻觉
- 演化信息学模型：递归闭环下的“模型坍缩”与文明自体吞噬
- 反方主要观点的证伪与边界条件讨论
- 结论：在算法引力场中维护认知生态多样性

1. 引言：智力维度的再审视——均值与方差的哲学区隔

在讨论“AI 会不会让人类智力趋同”时，既有讨论常陷入常识性的认知误区：将“智力”等同于“执行特定任务的熟练度”或“获取标准答案的速度”。若以此狭隘视角观察，AI 无疑降低了技能门槛，使更多普通人能迅速完成合格的代码编写、图像渲染或论文构思。持乐观论调者据此断言，AI 赋能了个体，因而促进了智力的多元化。

然而，从现代认知科学（如 Sternberg 的成功智力理论与 Guilford 的发散-聚合思维框架）以及演化生物学的视角审视，智力（Intelligence）的核心特质在于其面对未知的“异质性表征能力”与“自适应发散方差（Cognitive Variance）”。智力不是单一维度的标量，而是高维相空间中的探索轨迹。一个健康的人类智力生态，仰赖于个体由于不同生物神经发育、独特生命体验、偶然挫折与文化边缘性所产生的非共识思考路径。

因此，本调研报告建立在一个严格的统计学与认知学区分之上： - 表现均值的抬升（Mean Elevation）：指群体在标准化基准测试（如考试成绩、通用文案质量、常规工程实现）上的平均得分上升； - 智力方差的收缩（Variance Shrinkage / Convergence）：指个体在问题表征、假设生成、逻辑链条构建以及语义空间探索上的差异度显著萎缩，向算法预设的最优解聚拢。

本报告的核心论断是：生成式人工智能作为一种中心化的认知中介，正是以抬高个体均值为诱饵，系统性消解了人类智力生态赖以维系的群体方差，从而推动人类智力走向结构性趋同。

2. 神经生理机制：认知卸载、自由能原理与用进废退

2.1 自由能最小化与大脑的“认知吝啬”本能

卡尔·弗里斯顿（Karl Friston）的自由能原理（Free Energy Principle, FEP）指出，生物大脑本质是一台预测编码（Predictive Coding）机器，其自组织的核心目标是最小化感知与外部世界的预测误差（变分自由能），即以最小的代谢能耗维持内稳态。由此衍生出的认知心理学公理是：人脑是天生的“认知吝啬者”（Cognitive Miser）。

在没有外部认知中介的自然状态下，大脑为了解决复杂生存问题，必须调用前额叶皮层（PFC）进行高耗能的深层工作记忆整合、概念抽象与挫折攻坚，这一过程推动了神经突触的动态重塑（Synaptic Plasticity）与异质化连接。

2.2 认知卸载（Cognitive Offloading）的深度跃迁

2011年，哈佛大学心理学家丹尼尔·韦格纳（Daniel Wegner）团队在《Science》发表里程碑论文（Sparrow et al., 2011），揭示了著名的“谷歌效应”（Google Effect）：当人类预期能够从外部工具获取信息时，大脑海马体与前额叶会自动降低对事实性知识本身的神经编码强度，转向仅仅记忆“检索路径与位置”。

然而，生成式 AI 带来的认知卸载发生了质的飞跃： - 搜索引擎时代：卸载的是陈述性记忆（Declarative Memory），人类依然需要自主完成语义加工、逻辑整合、因果推演与论证建构； - 生成式 AI 时代：卸载的直接是程序性思维与认识论探索（Procedural & Epistemic Reasoning）。用户仅需提供碎片化提示词，AI 即可代劳问题分解、框架搭建与综合决策。

2.3 神经回路的用进废退与发散性思维钝化

爱沙尼亚塔尔图大学认知神经科学家雅安·阿尔（Jaan Aru）在2024至2025年连续发表研究（Aru, 2024; 2025），从脑网络交互机制剖析了 AI 对人类内在创造过程的生理损伤： 1. 网络协同机制的中断：人类独创性依赖于默认模式网络（DMN，负责无意识联想、自传体记忆调用）与执行控制网络（ECN，负责逻辑约束与批判评估）之间高强度的动态震荡。这种内在震荡是个体大脑在解决难题时经历的神经折磨，也是个体智力产生独特质感的生物学源泉； 2. 算法捷径导致神经修剪均一化：当 AI 瞬间提供流畅、高度可用的解答时，大脑

的自由能最小化机制驱动个体放弃内在脑网络的深层协同。长期接受外部预制答案，将导致人类负责发散性假设生成的额顶叶网络回路逐渐废用，大脑前额叶突触修剪失去个性化环境刺激，生理层面的思考进路发生群体性钝化与收敛。

3. 表征与符号学机制：统计模式均值化与语言多样性雪崩

语言是心智的外显载体（维果茨基内化理论与萨丕尔-沃尔夫假说）。思维的范式在极大程度上被其能够调用的符号系统与语法结构所锚定。

3.1 大语言模型的模式偏好（Mode-Seeking）与长尾平滑

当前主流的大语言模型均基于 Transformer 自注意力机制，并通过大规模自监督预训练与人类反馈强化学习（RLHF）进行价值与行为对齐：

- 损失函数的内在偏好：预训练阶段的交叉熵损失函数（Cross-Entropy Loss）驱使模型拟合海量语料中的最大后验概率分布；
- RLHF 的对齐税（Alignment Tax）：为了确保安全性与有用性，模型微调极大地惩罚了离群、极端、具有争议或非常规的表达，将响应分布强制压制在温和、中庸、规范的中心模态周围；
- 结果：大语言模型在数学本质上是一台极高维度的“模式偏好均值生成器”。

3.2 88万文本实证：语言复杂度方差暴跌 21% 至 50%

2026年，南加州大学莫特扎·德加尼（Morteza Dehghani）教授团队在《Nature Human Behaviour》发表了针对大模型语言同质化的开创性研究（Sourati et al., 2026）。该研究采集了涵盖 Reddit 社区、新闻媒体、arXiv 学术摘要在内的 超过 880,000 篇真实文本，并使用 GPT-4、LLaMA-3、Gemini 等主流大模型进行了大规模的重写与辅助写作实验。

研究得出了震撼性的量化结论：

1. 复杂度方差断崖式萎缩：经过大模型辅助或润色后，文本的句法复杂度、词汇丰富度指标（如 Type-Token Ratio）的方差骤降 21% 至 50%；
2. 社会语言学指纹（Sociolinguistic Fingerprints）的清洗：在未经润色的自然语言中，文本深度编码了作者的年龄、性别、地域、文化阶层以及独特的心理特质；而在经过大模型处理后，用于识别上述人类个体特质的机器学习分类器准确率发生严重滑坡。
3. WHELM 规范的单极同化：研究进一步证实，语言并没有收敛于一个多元中立的交集，而是无一例外地向西方、高收入、受过高等教育、自由派（Western, High-income, Educated, Liberal, Male, 即 WHELM）的现代文风聚拢。

当全球不同文化、不同心智模式的人类个体在写作、沟通和构思时，其表达结构与概念范式被统一压缩至 WHELM 凸包内时，人类智力表征空间的收缩已成为被大规模量化的客观事实。

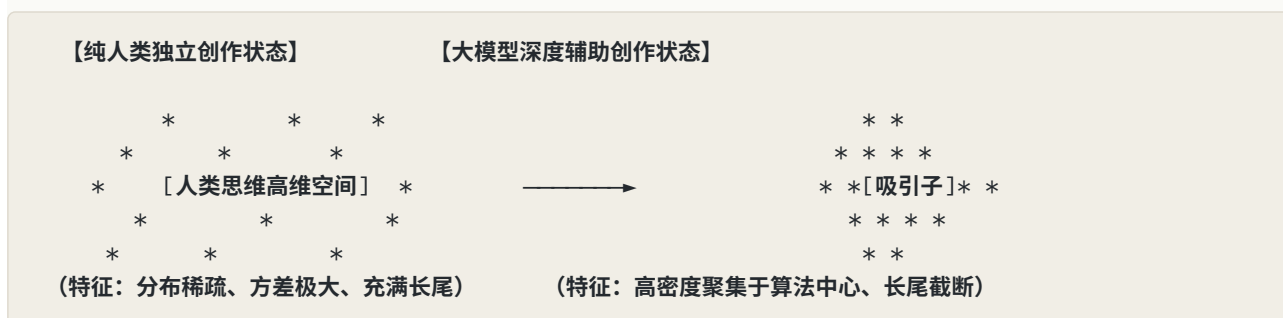
4. 集体行为实证：个体创意均值抬升与群体异质性收缩

对于“AI 让人类更有创意”这一流行直觉，行为经济学与计算社会科学给出了严密的实验否定。

4.1 《Science Advances》的里程碑实验：Doshi & Hauser (2024)

伦敦大学学院（UCL）的阿尼尔·多希（Anil R. Doshi）与埃克塞特大学的奥利弗·豪瑟（Oliver P. Hauser）在《Science Advances》发表了一项预注册的双盲随机对照实验（RCT）。实验招募了数百名写作者进行短篇故事创作，对比了纯人类创作组与获得大语言模型（GPT-4）灵感支持组的输出特性。

研究的核心发现勾勒出一个清晰的“集体创新的社会困境” (Social Dilemma of Generative AI) : - 个体视角的利己最优: 在获得 AI 生成的 1 个或 5 个故事构思后, 独立评估者对个体故事的评价显著提高, 故事更加连贯、精彩、富有吸引力。特别是在前期基线测试中创造力偏弱的个体, 提升幅度最大; - 群体层面的多样性塌缩: 研究人员利用高维文本嵌入向量 (Embeddings) 计算所有作品之间的余弦相似度 (Cosine Similarity), 结果显示: AI 辅助组作品之间的语义相似度极显著高于纯人类创作组 ($p < 0.001$)。接入 AI 的创作者们集体涌向了少数由算法建议的高概率情节架构, 整体作品集的语义覆盖范围严重收敛。



4.2 常规交互下 0.92 的惊人相似度: Wan & Kalman (2026)

休斯顿大学的云万 (Yun Wan) 与开放大学的卡尔曼 (Yoram M. Kalman) 在《Computers in Human Behavior: Artificial Humans》发表复现与扩展研究。他们证实：在人类用户的常规交互条件下，不同用户通过通用提示词由 AI 生成的创意大纲，其平均余弦相似度高达 0.92。

尽管该研究提出通过复杂的十重多元角色提示词矩阵（10 Diverse Personas）可以在实验室环境下部分抵消这种相似性，但这从反面确立了正方论据：在真实人类社会，绝大多数普通用户绝不具备构建多重正交提示词的认知储备与精力。在追求效率的工程实践中，99% 的人类必然顺从于默认的单极输出，智力趋同在宏观社会演化尺度上具有不可逆的统计惯性。

5. 认识论危机：科学研究的“单一栽培”与理解的幻觉

科学研究作为人类智力探索未知的最高形态，通常被视为抵抗平庸与同质化的最后堡垒。然而，科学社会学与认识论研究表明，AI 正在重构科学探索的范式本身。

5.1 《Nature》长文警示：科学认知的“单一栽培” (Scientific Monoculture)

2024年3月，耶鲁大学人类学教授丽莎·梅塞里（Lisa Messeri）与普林斯顿大学心理学教授克罗克特（M. J. Crockett）在《Nature》发表题为《Artificial intelligence and illusions of understanding in scientific research》的封面综述长文。

论文指出，科学家群体对 AI 的热情正导致严重的认识论异化，具体表现为科学视角的单一栽培：1. 问题选择的趋同化：科学家们开始自发性地优先提出“适合 AI 工具求解的问题”（如基于海量已有数据的关联模式挖

掘），而放弃那些非结构化、无法数字化表征、依赖长周期肉身观察或激进直觉的科学难题；2. 研究范式的均一化：跨学科的独特性正被“大模型代理被试”、“提示词工程调优”和“特征嵌入聚类”所统一格式化。理论物理的公理直觉、社会学的深描民族志、生物学的田野观察，统统被降维重构为统计关联预测；3. 理解的三重幻觉（**Illusions of Understanding**）：- 解释深度的幻觉：科学家以为 AI 的高精度预测等同于人类理解了其机理；- 探索广度的幻觉：科学家误以为 AI 在毫秒间检索了数万篇文献就代表穷尽了科学视野，忽视了模型只能在已有训练集的凸包内插值；- 客观性的幻觉：将算法输出的统计中位数奉为超脱人类立场的终极真理。

论文尖锐地警告：其终极恶果是“论文产出越来越多，但人类真正有突破性的理解却越来越狭隘（**Produce more, understand less**）”。

6. 演化信息学模型：递归闭环下的“模型坍缩”与文明自体吞噬

如果仅将视角停留在当前一两代技术的使用上，尚不能完全看清趋同的终局。将信息论与复杂自适应系统理论引入后，我们会发现人类智力正在遭遇一场系统性的“认知自体吞噬”（**Cognitive Autophagy**）。

6.1 递归训练的诅咒与模型坍缩定理

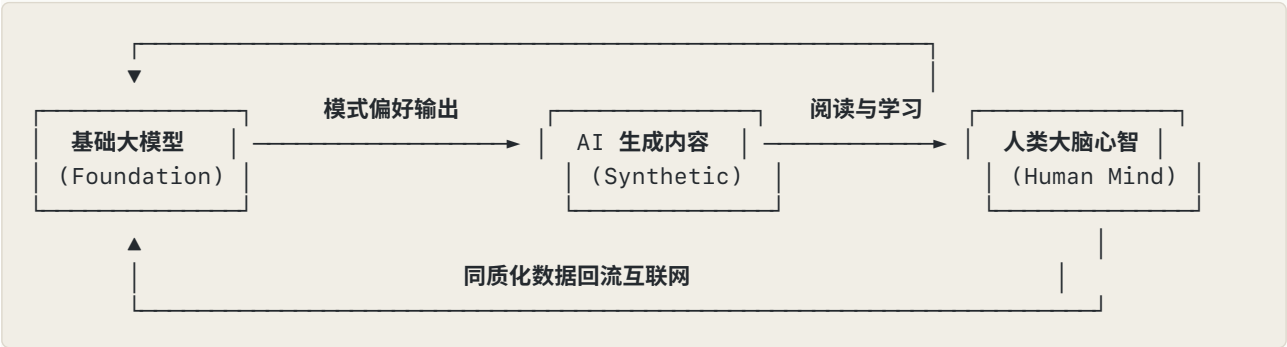
2024年7月，来自牛津大学、剑桥大学与多伦多大学的联合团队（Shumailov et al., 2024）在《Nature》发表数学证明论文《AI models collapse when trained on recursively generated data》。

研究通过严格的数学推导与实验证明：当生成式模型被使用其前代模型生成的数据进行递归微调或训练时，系统必然发生不可逆的“模型坍缩”（**Model Collapse**）：- 初级坍缩：概率分布长尾处的低频事件（离群值、少见词汇、异常概念组合）因抽样方差被系统性擦除，模型仅生成中心众数（**Mode-seeking**）；- 终极坍缩：分布方差迅速缩减至零，输出退化为极度单一甚至荒谬的无意义伪影。

6.2 人机共生环境中的“认知自体吞噬闭环”

过去，人类文明的知识库由生物人类在物理世界中经由痛觉、偶然性、异端叛逆与生物繁衍突变所持续注入“外部熵”（负熵流与真实随机性）。

然而，当前互联网已发生根本性逆转：根据计算语言学统计，网络上新增的文本、代码与数字艺术中，已有相当比例由 AI 生成或润色。这就构成了前所未有的人机递归闭环：



1. 人类大脑在日常学习、科研和工作中，沉浸于被 AI 同质化过的合成知识；
2. 大脑依据认知吝啬法则，内化了 AI 的思维进路，输出更加标准化的文本与文化；

3. 这些高度均质化、缺乏极端长尾创新的数据再次被爬取，用于训练下一代大模型。

在这一闭环中，人类智力分布的长尾——那些孕育了爱因斯坦、梵高、尼采的“精神离群值”——被数学算法视作统计噪声无情清洗。人类整体智力生态不可避免地陷入了一场大规模的文明级模型坍塌。

7. 反方主要观点的证伪与边界条件讨论

在学术研讨中，主张“AI 不会导致智力趋同”的常见论据主要有三点，本调研报告在此给出针对性的理论与实验证伪：

反方核心观点	反方背后的逻辑假设	认知神经科学与实证研究的证伪证据
观点一：工具解放论 “AI 承担繁琐机械工作，让人类解脱出来从事更高阶、更多元的创造性思维。”	假设“高阶创造力”可以脱离低阶知识磨砺与神经负荷而凭空产生。	证伪：Jaan Aru (2025) 证明，人脑的高阶抽象直觉必须以底层神经突触的长久编码与挫折攻坚为生理基石。跳过底层思维爬坡直接索取高阶结果，会导致大脑额顶网络连接弱化，产生“认知肌肉萎缩”。
观点二：个性化交互论 “千人有千种 Prompt，个性化 Agent 会成为每个人的专属智囊，个体差异将被放大。”	假设用户界面的输入多样性能够逃脱底层大模型的先验概率分布限制。	证伪：Sourati et al. (2026) 与 Wan & Kalman (2026) 证实，Prompt 的扰动无法打破预训练语料与 RLHF 构建的强引力凸包。常规用户的输出相似度高达 0.92，表层的提问修辞无法掩盖底层表征的单极吸引。
观点三：历史技术外推论 “历史上算盘、印刷术、计算器出现时也有人担忧人类退化，但人类文明最终更加多元繁荣。”	假设生成式 AI 与历史上的计算工具处于同一范式类别。	证伪：Messeri & Crockett (2024) 明确指出，历史工具替代的是机械运算 (Execution)，而 AI 介入的是概念表征与因果假说生成 (Epistemic Framing)。AI 是人类历史上首个“意义生成型”中介，历史类比存在根本性的范畴谬误。

8. 结论：在算法引力场中维护认知生态多样性

综合神经可塑性理论、计算语言学 88 万文本分析、创造力随机对照双盲实验以及递归信息系统的数学证明，本调研报告得出确凿结论：

在缺乏激进制度干预与认知自觉的前提下，生成式人工智能的发展与普及，必然在生理、表征、方法论与演化四个尺度上，驱动人类智力走向系统性趋同。

这种趋同并非意味着人类智能绝对值的退化，而是意味着人类心智生态的“单一种植化” (Cognitive Monoculture)。正如生态学所揭示的，单一品种的高产农田在丰年固然高效，但在遭遇未知病虫害侵袭时却极易全军覆没；拥有丰富方差与长尾变异的热带雨林，才是生命抵御毁灭性风险的真正避难所。

作为深圳理工大学“人脑智能”课程的研究者，我们审视 AI 的终极目的，绝非盲目的技术悲观主义，而是要穿透效率的迷雾，清醒指认算法对生物心智多样性的深层重塑。只有正视这一趋同危机，人类才能在教育重塑、系统架构设计以及人机协作边界的确立上，筑起捍卫人类认知异质性与文明反脆弱性的坚固防线。

参考文献列表 (References)

1. Doshi, A. R., & Hauser, O. P. (2024). Generative artificial intelligence enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 10(28), eadn5290. DOI: [10.1126/sciadv.adn5290](https://doi.org/10.1126/sciadv.adn5290)
2. Sourati, Z., Karimi-Malekabadi, F., Ozcan, M., McDaniel, C., Ziabari, A. S., Trager, J., Tak, A., Chen, M., Morstatter, F., & Dehghani, M. (2026). The shrinking landscape of linguistic diversity in the age of large language models. *Nature Human Behaviour*. (Preprint available at [arXiv:2502.11266](https://arxiv.org/abs/2502.11266)).
3. Sourati, Z., Ziabari, A. S., & Dehghani, M. (2026). The homogenizing effect of large language models on human expression and thought. *Trends in Cognitive Sciences*. (Preprint available at [arXiv:2508.01491](https://arxiv.org/abs/2508.01491)).
4. Messeri, L., & Crockett, M. J. (2024). Artificial intelligence and illusions of understanding in scientific research. *Nature*, 627(8002), 49 – 58. DOI: [10.1038/s41586-024-07146-0](https://doi.org/10.1038/s41586-024-07146-0)
5. Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature*, 631(8021), 755 – 759. DOI: [10.1038/s41586-024-07566-y](https://doi.org/10.1038/s41586-024-07566-y)
6. Aru, J. (2025). Artificial intelligence and the internal processes of creativity. *The Journal of Creative Behavior*, 59(2), jocb.1530. DOI: [10.1002/jocb.1530](https://doi.org/10.1002/jocb.1530)
7. Wan, Y., & Kalman, Y. M. (2026). Diverse AI personas can mitigate the homogenization effect in human-AI collaborative ideation. *Computers in Human Behavior: Artificial Humans*, 100289. DOI: [10.1016/j.chbah.2026.100289](https://doi.org/10.1016/j.chbah.2026.100289)
8. Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *Science*, 333(6043), 776 – 778. DOI: [10.1126/science.1207745](https://doi.org/10.1126/science.1207745)