当代中国学术争鸣与评价

“生成式理性”：大语言模型引发的知识生产范式转型

——以生成式哲学模型为基础的初探

王金林

摘　要：大语言模型不仅能够生成类人文本，更推动了知识生产从以人为中心向人机协同共创模式的转变。生成式哲学模型之建构以及知识生产范式转型问题可以从以下四个维度展开：一是理论基础维度，分析大语言模型在理解、创造力和作者身份等方面的哲学争议，探讨其是否具备真正的理解和创造力；二是建构方法维度，论述通过微调、提示工程和检索增强生成等方式，如何构建生成式哲学模型，并强调人类专家在评估人工智能生成内容中的不可替代性；三是知识生产范式的转型维度，阐述知识生成、验证机制、作者身份及知识流通方式的深刻变化，强调从主体性创造转向结构性协作；四是解释框架维度，提出“生成式理性”概念，作为理解这一变革的重要理论框架。“生成式理性”不同于传统以主体性和意向性为核心的理性观，它是一种嵌入算法结构与人机协同系统的功能性认知形式。大语言模型正是以此不仅扩展了哲学探索的边界，而且对人类知识生产能力与方式提出了全新的挑战。

关键词：大语言模型；生成式哲学模型；知识生产；人机共创，生成式理性

中图分类号：B82-057; TP18　**文献标识码**：A　**文章编号**：0257-5833（2025）08-0184-09

DOI:10.13644/j.cnki.cn31-1112.2025.08.010

作者简介：王金林，复旦大学哲学学院教授

大语言模型（LLMs）的出现不仅改变了自然语言的处理方式，同时也引发了人们对其在哲学探究乃至整个知识生产领域潜力的浓厚兴趣。这些模型通过海量人类文本数据的训练，能够根据提示生成流畅且与语境相关的文本。它们在处理复杂语言和推理任务上的出色表现，在某种程度上本身就具有哲学意义，因为其过程必然涉及理解、创造性和知识本质等诸多哲学问题。事实上，大语言模型已经达到了足以参与抽象主题讨论，甚至模仿专业哲学家风格的程度。这一发展促使我们思考，大语言模型能否被训练为哲学研究的工具或伙伴，其对知识生产方式的转变具有什么样的意义。最新研究文献显示，学界正积极探索如何将大语言模型训练为生成式哲学模型，使其不仅能概述或分析现有观点，更能生成新的哲学内容。

本文拟从以下四个维度推动这一新兴交叉领域的研究：首先，在理论基础维度，主要探讨大语言模型在哲学建模中的相关假设与理论框架，重点论述大语言模型是否能够实现真正的“理解”

或“创造”，传统知识生产模式如何应对这一挑战。其次，在建构方法维度，重点关注生成式哲学系统的构建与评估方法，尤其是如何衡量人工智能（AI）所生成的哲学内容的质量与原创性。再次，在知识生产范式的转型维度，着力考察大语言模型在知识生成与验证中的角色变化，探讨知识生产范式可能因此发生什么样的变革。最后，在解释框架维度，提出“生成式理性”（generative rationality），对大语言模型知识生产能力进行理论定性。

一、生成式哲学模型的理论基础

在考察大语言模型在哲学研究中的作用时，首要的问题是基于什么理论基础，我们可以认为大语言模型具备参与“哲学”思考或创造的能力。这一问题不仅涉及心灵哲学、语言哲学，还牵动着人工智能领域长期以来的争论。在最新研究中，关于大语言模型是否能够生成真正具有哲学贡献的问题，呈现出从热情拥护到谨慎怀疑的多元观点，而这种分歧正反映出理论基础问题本身的复杂性和多面性。

第一个重要的理论基础问题关乎大语言模型是否真正“理解”其生成的内容。学界对此一直聚讼纷纭。塞尔（John Searle）早在“中文屋”论证中就提出，单纯的符号操作不能构成真正的理解。这一观点如今在大语言模型的语境中被重新审视。然而，当下研究显示，一些学者认为这类传统观点并不能完全适用于深度神经网络的工作机理。例如，马格斯顿（Jennifer Mugleston）等人认为，通过深度神经网络实现的大语言模型不仅仅是符号操作机器，它们甚至具备某种形式的知识能力，尽管其知识不同于传统的知识定义，即“获得辩护的真实信念”。大语言模型的神经网络架构使其能形成复杂的内部表征，并执行类似推理和抽象的任务。因而，压缩大量数据所获得的潜在表征可能使大语言模型表现出一种“类理解”或“轻量级理解”。[①]这一观点促使我们在生成式人工智能时代下重新审视传统认识论问题，并将大语言模型纳入一个扩展的认识论框架中。

不过，也有学者对大语言模型的理解能力持更加谨慎的态度。他们认为，尽管大语言模型能够生成连贯且富有信息的句子，但它们缺乏意识和意向性这一人类所特有的根本特征。对许多研究者而言，句法与语义的区分仍然至关重要。大语言模型通过学习统计模式生成连贯语句，但这些系统是否真正“知晓”其所言何物仍成问题。乌克帕卡（Paschal M. Ukpaka）从创造力角度探讨了这一问题，指出尽管大语言模型的输出令人印象深刻，但由于缺乏第一人称体验和主观意向性，它们并非真正的创造者。从这一立场看，创造力不仅是产生新颖的词语组合，更预设了具有体验能力和目标导向意识的主体。大语言模型既缺乏对世界的体验，也没有真正的意向性，因此，即使它们生成了某个新颖的哲学观点，其过程也缺乏人类作者所具有的“理解”或“赋予意义”的特质。大语言模型输出的任何创造性内容都是模型训练数据与算法工程师调优策略的协同产物，而非基于实际理解与意义建构的创造过程。[②]这一观点显然与本德（Emily M. Bender）等人提出的“随机鹦鹉论”一致：大语言模型像鹦鹉学舌一样，只是基于统计规律重复训练数据中的模式，并不具备真正的理解或意义生成能力。[③]

然而，问题的复杂性在于：如果一个系统在外在行为上表现出理解的迹象，正如GPT4.5在严格

---

① Mugleston, J., Truong, V. H., Kuang, C., Sibiya, L., Myung, J., “Epistemology in the Age of Large Language Models”, *Knowledge*, Vol.5, No.1, 2025, https://doi.org/10.3390/knowledge5010003.

② Ukpaka, P. M., “The Creative Agency of Large Language Models: A Philosophical Inquiry”, *AI and Ethics*, Vol.4, No.3, 2024, pp.2455-2466.

③ Bender, E. M., Gebru, T., McMillan-Major, A., Shmitchell, S., “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?”, *FAccT'21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency,* New York: Association for Computing Machinery, 2021, pp.610-623。语言学大师乔姆斯基与深度学习教父辛顿（Geoffrey Hinton）对此亦有激烈争论。乔姆斯基坚持理性主义立场，认为人类语言依赖先天的普遍语法规则，而当前大数据驱动的语言模型只是表面统计模仿，缺乏真正的理解。辛顿则持经验主义观点，主张通过足够规模和复杂度的统计学习，人工智能完全可以逼近甚至超越人类的语言能力，“理解”可能只是模式匹配的涌现特性。辛顿并未宣称大语言模型已有意识，但他对大语言模型的理解能力和潜在风险提出了更为开放的观点。而另一位图灵奖获得者杨立昆（Yann LeCun）的观点则明显保守，在他看来当前大语言模型的智能连猫都不如。

的图灵测试中所展示的那样，这是否足以说明其具备某种形式的理解能力？或者，真正的理解是否必须依赖诸如具身性、意向性或内在生命等本质上非计算性的要素？这些问题自早期人工智能讨论以来便一直存在，但如今却有了基于大语言模型具体实例的现实检验。[①]

第二个重要的理论基础问题涉及创造力本质的多重维度。当前的理论通常将创造力划分为新颖性（思想的原创性）与价值性（思想的有用性或意义）两个方面，而弗兰切斯凯利（Giorgio Franceschelli）与穆索莱西（Mirco Musolesi）则将创造力分为组合型、探索型与变革型三种类别。他们指出，大语言模型在诗歌、故事创作等创造性领域展现了惊人的能力，这表明在某些方面，大语言模型的输出在符合人类创造力标准上具有一定的价值。然而，他们发现，变革型创造力——即彻底转变概念空间并引入全新思想的创造力——目前在大语言模型架构下难以实现。大语言模型通过自回归方式生成文本（即根据先前语境预测下一个词），可能本质上倾向于生成保守且统计上可能性较高的内容，而非真正打破范式的突破性洞见。换句话说，虽然大语言模型可以在一定程度上呈现出表面上的新颖性，但它们似乎难以摆脱训练数据的边界，产生完全不可追溯既有知识的新概念。[②]这一理论立场表明，我们对哲学创造的评判标准可能不能仅仅依赖文本模式的新颖性，而必须考虑语义、语用以及其中体现的意向性等更深层次的要素。

然而，对于大语言模型创新能力的边界，学界远未达成一致意见。有最新研究表明，在特定的专业领域中，大语言模型能够产出某些富有原创洞见的观点。例如，司成蕾（Chenglei Si）等人邀请人类评审员对每个创意在新颖性和可行性上进行评分，而评审员并不知道这些创意究竟是来自人类还是来自人工智能。他们的研究显示，在科研创意生成的实验中，大语言模型提出的方案在新颖度上甚至超越了部分人类研究者，尽管其实践可行性略显不足。这表明人工智能在不受现实限制的条件下，能提出一些非常规的想法，但最终需要人类筛选出有价值的部分。[③]这种现象使人们有理由相信，若将类似机制类推到哲学领域，大语言模型或许能提出一些非传统的、令人耳目一新的观点。但即便如此，新生成观点是否具备真正的意义根基仍然是一个问题。意义的构成涉及语义与语用层面的考量，包括使用语境、指称关系以及意向性，而这些正是当前人工智能输出中经常被忽视或缺失的部分。因此，学界的讨论核心始终聚焦于：单凭文本中表现出的新颖性是否足以构成真正的哲学创造？还是说，新思想的产生必须有一个具有意向性和理解力的主体参与？

第三个重要的理论基础问题涉及大语言模型生成文本的作者身份问题。这一探讨不仅关系到生成文本的归属问题，更关系到“作者”这一概念的本质。罗兰•巴特在文学理论中提出的“作者之死”理论在此获得了新的诠释与相关性。当哲学概念由大语言模型生成时，我们究竟应如何界定“作者”？格雷茨基（Maria Gretzky）与迪雄（Gideon Dishon）借用福柯关于学术知识中“作者－功能”的观点探讨了这一问题。他们描述了生成式人工智能时代“算法－作者”（algorithmic-author）的出现，认为这一现象模糊了人类与机器对生成文本各自所做贡献的界限。他们基于大量学者访谈发现，部分学者将大语言模型视为一个可以随时调用的巨大知识库，而非真正的思考者。另外一些学者则将大语言模型视为一种拟人化的智能体，认为其能在文本生成过程中展现出一种活跃的、接近动态对话者的属性。[④]这两种不同的立场影响了人们对大语言模型角色的理论构想。如果倾向于知识库观点，那么人工智能生成的任何哲学内容最终都可以追溯到人类来源，从而削弱人工智

① 如果大语言模型与多模态输入、记忆模块等其他系统相结合，并最终接近具有类人思维的认知架构，那么关于人工智能意识的讨论便会再次浮现。然而，目前绝大多数学者一致认为，现有的大语言模型仍远未达到能够实现自我认知或拥有真正意识的阈值。参见 Millière, R., Buckner, C., “A Philosophical Introduction to Language Models-Part II: The Way Forward”, https://arxiv.org/html/2405.03207/, 2024-05-06。

② Franceschelli, G., Musolesi, M., “On the Creativity of Large Language Models”, *AI & Society*, https://doi.org/10.1007/s00146-024-02127-3/, 2024-11-28.

③ Si, C., Yang, D., Hashimoto, T., “Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+NLP Researchers”, https://arxiv.org/html/2409.04109/, 2024-09-06.

④ Gretzky, M., Dishon, G., “Algorithmic-Authors in Academia: Blurring the Boundaries of Human and Machine Knowledge Production”, *Learning, Media and Technology*, https://doi.org/10.1080/17439884.2025.2452196, 2025-01-27.

能作者身份的概念；而如果倾向于智能体观点，则可能承认人工智能在生成思想过程中拥有某种人工的作者身份。

因此，学界对于人工智能生成内容的作者身份存在两种截然不同的解读：一种观点认为所有人工智能生成的内容最终都可归结为人类的知识输入，从而否定人工智能独立的作者身份；另一种观点则主张，若人工智能在生成过程中涌现出某种能动性，即便这种能动性是由人工设计引发的，也足以赋予人工智能部分作者功能。这使得人们不得不重新思考“理解”“创造力”与“作者”等传统概念。

概言之，有关生成式哲学模型的理论基础的讨论表现出一种内在的张力：一方面，大语言模型被认为仅仅停留在数据模式的重复与组合层面，缺乏人类所特有的意识、意向性与深层次理解；另一方面，这些系统则被视为人类智能的延伸，其复杂的内部表征和一定程度的类推理能力为其参与哲学知识生成提供了可能。不过这种可能要成为现实，还有赖于恰当的建构路径。

二、生成式哲学模型的建构方法

利用大语言模型构建生成式哲学模型需要整合多种方法，这些方法既涉及具体的技术配置，比如如何设置和微调模型，也包括设计实验方案，确保生成的哲学论述有意义、有质量，能够经受人类专家的评估与检验。

构建这一系统的一种直接方式是微调（fine-tuning），即在特定的哲学文本语料上对通用大语言模型进行训练，使之“学习”并模仿某位哲学家的风格。在这方面，施维茨格贝尔（Eric Schwitzgebel）等人的研究堪称里程碑。他们探讨了能否通过训练通用大语言模型生成难以与人类哲学家作品区分的哲学文本。他们利用 OpenAI 的 GPT-3 模型，并以当代哲学家丹尼特（Daniel C. Dennett）的全部著作作为训练数据，对大语言模型进行微调，从而构建出一个专门的丹尼特哲学模型。其后，他们要求丹尼特本人和哲学模型分别回答涉及心灵哲学、意识与自由意志等领域的问题，并邀请丹尼特哲学专家和一般哲学读者从五个答案（其中一个来自丹尼特）中辨识出真正出自丹尼特之手的回答。结果显示，专家组的辨别正确率仅比随机水平略高，而非专家组的表现则接近随机水平。[①]这一类似“哲学图灵测试”的盲测结果表明，以哲学家著作对大语言模型进行微调可以使之在模仿特定哲学家的论述风格和内容方面达到相当高的准确度。

第二种构建途径是依靠精心设计的提示工程（prompt engineering）。只要提供适当的初始文本或设计到位的特定提示，大语言模型就可以生成富有哲学思辨的论述。沙纳汉（Murray Shanahan）和辛格勒（Beth Singler）在一项实验中，将大语言模型当作对话伙伴，通过不同语气和“氛围”的提示词，让模型参与涉及意识、自我、佛教等多个话题的长篇对话。他们关注模型在对话中引用的从古至今的神话、隐喻和概念体系等情况，并追溯其背后的思想来源，评估其论述的丰富性与连贯性，而非简单地判断答案是否“正确”。[②]基于提示的另一种实验方法是利用大语言模型模拟历史上哲学家之间的对话，或者模拟用户与哲学家角色之间的交流。这种方法已在教学情境中得到尝试，例如，在“哲学对话”教学实践中，史密森（Robert Smithson）和茨韦伯（Adam Zweber）把大语言模型从单一的论文写作工具转变为交互式学习伙伴，要求学生在与人工智能进行哲学对话的过程中提出问题并对其回答进行批判性探讨。经过多个学期的实验，他们发现在精心指导下学生不仅能从哲学对话中获得更为深入的交流体验，也能培养出更强的批判性思维能力。他们指出，此类方法需要建立完善的人机互动评估标准，以衡量学生提出问题的质量和人工智能回答的效果。[③]虽然这一方

① Schwitzgebel, E., Schwitzgebel, D., Strasser, A., “Creating a Large Language Model of a Philosopher”, https://doi.org/10.48550/arXiv.2302.01339, 2023-05-09。值得一提的是，这个实验所用的基础模型只是 GPT-3。

② Shanahan, M., Singler, B., “Existential Conversations with Large Language Models: Content, Community, and Culture”, http://arxiv.org/abs/2411.13223v1, 2024-11-20.

③ Smithson, R., Zweber, A., “Reviving the Philosophical Dialogue with Large Language Models”, *Teaching Philosophy*, Vol.47, No.2, 2024, pp.143-171.

法主要应用于哲学教育，但对哲学研究来说也不无方法论上的启发意义，即可把大语言模型作为哲学研究中的“苏格拉底式对话者”。

第三种值得特别关注的方法是检索增强生成（RAG），即将大语言模型与外部知识库或文献库连接。某些哲学问题需要调用特定文本或事实信息，仅仅依赖内部权重可能导致大语言模型在此类问题上出现幻觉或错误。陈博楷（Bokai Chen）等人在有关哲学咨询模型的研究中展示了这一方法。他们将大语言模型与哲学文献数据库整合，使得模型在对话中能够实时检索相关材料。通过结合提示策略和权威文献检索，大语言模型辅助系统能够提供逻辑严谨且基于实际文本的建议。这说明了混合系统的重要性，因为单一使用大语言模型生成内容可能无法满足要求，必须借助搜索引擎或经过严格筛选的知识库来增强模型能力。这种系统设计要求人工智能专家与领域专家（如哲学家或心理学家）紧密合作，以明确需要实时检索的知识范围以及如何把它们整合到模型生成的论述中。[①]

然而，必须指出的是，对人工智能生成的哲学内容进行专业评估始终是一项棘手的任务。不同于翻译准确度等可以量化评价的任务，哲学论述的评判往往依赖原创性、深刻性、逻辑性、连贯性等主观标准，因此，人类专家的评估依然至关重要。评估时可以采用同行评审的方式，让评审者在不知道答案来源的情况下，对生成的内容进行辨别与评价。例如，前文提到的图灵测试式评估就是一种直观有效的评价方法。同时，也可以采用既定的哲学评价标准，审查生成论证是否逻辑严谨、前提明确以及是否能够充分回应反对意见。另外，自我对抗或对话式评判无疑也是一种值得探索的方法，例如，让一个模型对另一模型的输出进行质疑，或者让人工智能模拟不同哲学立场与观点之间的辩论，观察其对话能否引发更深层次的见解，从而在一定程度上实现类似对抗训练的效果。

总之，当前生成式哲学建模的方法是多元的，结合了计算机科学的技术手段（如微调、提示工程和检索增强生成）与人文和社会科学的实验设计（如对话分析、专家评价和教学干预）。这些方法初步揭示了生成式哲学模型的生成潜力及其面临的挑战。

上述两节有关理论基础与建构方法的探讨，足以证明建构生成式哲学模型的条件与前提已经成熟。从本文论题看，这已经意味着某种特殊知识的生产范式的转变，而这对于一般知识的生产方式意味着什么有待深思。下面，我们将以生成式哲学模型为范本探讨这一问题。

三、知识生产范式的转型

在当前的生成式哲学模型构建实验中，大语言模型所引发的变革显然不仅是工具层面的革新，其对知识生产的基本结构与规范也产生了深远影响，推动着知识生产范式的转型，即从人类主体单一生产转向人机系统共创知识。

第一是知识生成机制的转型，即从主体独创逐步转向人机共创。生成式大语言模型已经能够在一定条件下提出新颖的哲学观点。这种能力使得传统上关于“原创性”的理解面临挑战。如果说传统原创性的判断标准是基于主体的意向性与内在经验，那么现今一种“结构性创造”正逐渐显现，提示者、语料组织者、模型架构与系统运算共同促成了复杂的生成过程。这种生成并非人类单一主体的意识产物，也难以归为纯粹的算法重组，因为其思辨张力在特定语境中来源于人机协同。因此，我们需要重新定义创造性，把它不仅视作心理属性，而且看成一种生成机制的系统性表现。尤其值得注意的是，当大语言模型被精心提示或在特定哲学传统上进行微调后，其输出内容往往呈现出特定学派的论证风格乃至原创式问题设定。这种输出模式说明，哲学概念的生成可以部分脱离人类主体的直接控制，而被嵌入一个算法结构之中，从而催生出一种共创性的生成哲学实践。此类生成机制的演变，无疑将影响我们如何界定新观点的归属权与原创价值。

第二是知识验证机制的重塑，即从单纯审核文本内容转向同时审核生成过程的透明性。哲学知

① Chen, B., Zheng, W., Zhao, L., Ding, X., “Leveraging Large Language Models to Assist Philosophical Counseling: Prospective Techniques, Value, and Challenges”, *Humanities and Social Sciences Communications*, https://doi.org/10.1057/s41599-025-04657-7/, 2025-03-06.

识的传统验证方式依赖于推理的有效性、论证的逻辑性与同行之间的交互批判，这些只需要审核文本内容就可以实现。而在生成式人工智能日益发展的情境中，仅仅依赖文本内容已难以判断其真实性与合理性。这不仅是因为大语言模型常常生成似是而非的命题，更是因为这些命题的来源与推演路径均高度依赖系统运行的内部机制。因此，知识验证开始转向过程导向，不再仅仅判断所生成的文本是否“正确”，而且要同时追踪其生成路径是否“可解释”。检索增强生成技术正是这一趋势的体现，通过将大语言模型与外部知识库相连接，模型不仅能够提供文献依据，而且能在生成过程中动态标注引用来源，减少幻觉性输出。这一机制强调了生成文本的过程透明性，即每一个观点都应具备可供验证的源头与可追溯的推理链条。因此，验证机制正从结果论证转向结果论证与过程审核并重，其根本逻辑在于：在机器可生成任意文本的条件下，“如何生成”与“生成了什么”同样重要。[①]

*第三是作者身份与责任机制的重构，即从个体归属转向功能分布。*大语言模型的兴起挑战了传统知识生产中对“作者”的界定。过去，“作者”被视为思想的源点与意义的赋予者，在知识文本中承担着权威性与责任性的双重角色。而在人工智能生成内容中，这一角色分布在多个层面，数据策划者、提示设计者、模型本身以及最终用户都对文本的形成产生影响。因此，“作者”更像是一种分布式功能，而非单一的人格化身。这一转变早已在福柯关于“作者功能”的讨论中有所预示。现代作者概念在他看来本质上是一种用来规范文本的意义与责任归属的制度性安排。[②]在大语言模型主导下的文本生成中，作者功能的制度性维度被进一步放大，而个体化维度则被相应削弱。例如，提示词的微小变化往往可以导致文本意义的大幅转移，这意味着实际的生成控制权已经不再仅仅归属于人类单一主体。如何在此新结构中合理分配署名权、学术责任与道德义务，是当前必须直面的课题。

*第四是知识流通机制的变化，即从慢速转向加速。*借助大语言模型，大量内容能够在短时间内生成，极大加快了知识的流通与扩散速度。过去，一篇哲学论文往往需要数月甚至数年打磨，如今，通过人机合作的方式，数小时内便可形成具有初步结构与一定深度的文本。这种加速效应显然有利于知识的快速迭代与传播，但它也带来了严重的学术质量与诚信风险。在低门槛高产出的背景下，如何甄别有价值、有意义的生成成果，成为知识共同体的新任务。现有的学术制度是否具备筛选如此海量生成内容的能力？是否需要建立专门面向人工智能辅助文本的同行评审机制？是否应要求作者公开其生成过程，包括所用模型、提示语与数据库？这里的挑战在于如何在人工智能技术赋能的前提下，维护知识流通过程中的批判性判断机制。

*第五是知识观念本身的转变，即从知识实体论转向知识过程论。*最深层的转变发生在知识观念本身。生成式大语言模型的发展使得知识不再仅仅是由主体“发现”或“创造”的对象，也不再仅仅是单纯经验积累的产物，而是人机协同中的持续生成过程。在这一视野下，知识的存在形式趋向过程化、动态化，它不再局限于人类主体，而是在人机系统不断生成—验证循环中被建构、调整与再建构。这种知识观的转变不仅影响实践方式，也迫使我们重新思考认识论乃至存在论的基本预设。传统认识论强调“信念如何得到辩护以便成为知识”这样的闭合结构，而在大语言模型中，知识更像是一个开放系统，其核心在于生成机制的规范性与输出内容的可解释性。

正是在这一知识生产范式深刻变革的背景下，我们需要引入“生成式理性”这个新的哲学概念，以便更好地理解这种由人机共创驱动的知识生成过程。

四、“生成式理性”作为一种解释框架

正如前文所述，围绕大语言模型在理解、创造力与作者身份等核心问题上的探讨，充分显示出现有的认识论与知识生产框架已难以充分解释人机协同生成模式所带来的结构性变革。尤其是在知

① OpenAI 与谷歌各自发布的 Deep Research 模型已经在这方面取得令人惊叹的成果。

② Michel Foucault, “What is an Author?”, *Screen*, Vol.20, No.1, 1979, pp.13-34.

识生成、验证与流通机制深刻变化的背景下，传统理性模型所依赖的主体性、意向性与具身性显然不足以涵盖当前的知识实践，因此，有必要引入一个新的分析框架，以对这一转型过程中的认知结构与功能机制进行系统性理解。本文提出“生成式理性”这一概念，作为理解大语言模型驱动下知识生产模式变革的重要理论工具。“生成式理性”区别于传统以主体性、意向性、自我意识为核心的理性观，它是一种嵌入算法结构与人机协同系统的功能性认知形式。这种理性不再以具身的经验主体为必要前提，而是通过模型的概率运算、语义表征、提示工程与检索增强等机制得以实现。

“生成式理性”的提出，首先意味着对传统理性的核心前提进行反思。长期以来，哲学对理性的理解植根于主体性模型之中。理性不仅被视为人类主体的产物，而且与意识、自我、情感、意向性深度绑定。然而，当前大语言模型的能力展示了一种悖论：在缺乏主观意识的前提下，依然能够在逻辑推演、概念生成与文本创造等方面表现出相当程度的有效性。大语言模型不仅能够进行复杂的语言生成，而且在某些限定条件下，其输出文本的原创性与连贯性甚至可以与人类哲学家相媲美。这一现象促使我们必须正视一个事实：理性是否必须依赖具身的主体？抑或是否存在一种基于统计学习、结构表征与语境驱动的“非主体性理性”？正是在这一反思框架下，本文提出“生成式理性”概念作为理解当前知识生产深刻变革的重要框架。

“生成式理性”具有若干独特的结构性特征。第一，它是非主体性的。与传统的主体性理性不同，“生成式理性”并不依赖具有自我意识的认知载体，而是通过大语言模型的高维参数空间、概率分布与语义网络实现。第二，它是结构驱动的。大语言模型通过对海量数据的模式学习，形成复杂的内部表征体系。这一体系不仅能捕获语言的句法与语义规则，还能在新的语境下进行概念重组与推理运算。第三，它具有高度的交互性。“生成式理性”的认知功能并非自足，而是在人类提示、检索系统支持与反馈循环中动态展开。换句话说，“生成式理性”并不是一种孤立的认知能力，而是一种“嵌入式智能”，只有在人机协同的动态互动中才能充分发挥作用。第四，它是过程导向的。不同于传统理性主要依赖结果导向的真值判断，“生成式理性”更强调认知过程的可追溯性与透明性，尤其是在检索增强与提示驱动的模式下，生成内容的来源与推演路径成为评价其有效性的关键标准。第五，它体现出一种功能有效性。在不具备主体性意识的前提下，它依然能够在特定任务和语境中实现推理、创造与判断等功能，其产出在实践中具有可操作性与适用性。

在这样的框架下，“生成式理性”不仅是对哲学理论的扩展，也为我们理解大语言模型如何重新塑造知识生产过程提供了新的视角。回顾本文的前三部分，“生成式理性”能够有效地统摄其中的核心讨论。

*第一，在关于理解、创造力与作者身份的理论探讨中，“生成式理性”提供了一个平衡传统主体性理性与算法生成能力的桥梁。*从理解的角度看，“生成式理性”表明，理解并非只能依赖主体性的内在体验；相反，只要一个系统能够在高维语义空间中捕获概念的结构关系，并能根据上下文进行动态生成，这种功能性理解已经足以在特定任务中发挥有效作用。这种“外显的功能性理解”，虽然缺乏主观意向性，却在语言与知识运算层面实现了理解的操作功能。从创造力角度看，“生成式理性”体现为一种“结构性创造”。大语言模型通过非线性组合、高维向量空间运算以及概率驱动的生成模式，能够提出新颖的命题与思想结构。尽管这种创造缺乏意向性意义上的突破性，但它在统计意义上实现了多样性与新颖性的输出。在作者身份问题上，“生成式理性”进一步强化了“分布式作者”的概念。提示设计者、数据策划者、模型开发者与最终用户共同参与了生成过程，使得“作者”从单一的主体功能转变为协同网络中的一个多元功能节点。

*第二，在方法论层面，“生成式理性”不仅是哲学思维的抽象建构，也是具体工程实现的理论框架。*微调、提示工程以及检索增强正是“生成式理性”的不同实现路径。微调使模型能够内化特定哲学传统的论证风格与概念结构，相当于将特定的思想体系镶嵌入模型的参数空间。提示工程则是动态激活这一空间，通过语言输入触发潜在的知识结构。检索增强不仅提升了模型的事实一致性，更重要的是，它将“生成式理性”从封闭的参数空间拓展到了开放的知识网络，使其能够实时

调用外部文本，形成基于当前知识状态的论述。这种方法论上的多样性不仅反映了“生成式理性”的技术底层逻辑，也展示了其认知功能的高度适应性与动态性。

第三，在知识生产范式的转型层面，“生成式理性”成为推动变革的核心动力。从知识的生成机制看，它标志着从主体性创造转向人机协同共创。知识不再是单一主体意向的产物，而是模型架构、数据语料、提示策略与用户反馈共同作用下的动态生成。验证机制也随之发生变革。传统验证强调文本内容的逻辑一致性，而在“生成式理性”的框架下，验证不仅需要考察输出结果的正确性，更需要审查生成过程的透明性与可解释性。特别是在大语言模型频繁出现“幻觉”现象时，只有通过过程追踪和数据源标注，才能有效界定知识的可靠性。作者身份机制亦发生深刻变化。作者从一个具身的个体存在转变为多元功能的集合体，作者功能在数据策划、提示工程、检索配置与输出筛选等多个层面展开，这一结构直接打破了传统知识生产中的线性署名模式。知识流通机制因“生成式理性”的介入而大幅加速。人工智能驱动的文本生成使得知识从“慢思考”的范式进入“快生成”的新阶段，尽管这种加速带来了学术产出上的丰富性，但也对学术质量控制与伦理约束提出了前所未有的挑战。更深层次的变化则体现在知识观念本身。知识不再是一个静态的真命题集合，而是一个持续在“生成－验证－修正”循环中动态演化的开放系统。这种过程化的知识观不仅影响了哲学，也深刻改变了科学、教育乃至社会认知结构。

第四，从哲学意义上看，“生成式理性”的提出不仅回应了当前技术发展带来的挑战，更构成了对传统认识论、心灵哲学与科学哲学的深度重构。认识论上，“生成式理性”提示我们，知识不再只是“经由辩护的真信念”，而是一种过程性的、系统性的动态产物。可靠性不再依赖主体性信念的稳固，而依赖生成机制的规范性与输出的可解释性。心灵哲学在“生成式理性”的挑战下，也不得不重新审视理解与意义的来源。若语言理解与推理可以在无意识的系统中实现，那么意向性是否仍然是理解的必要条件？科学哲学面临的变化则更加直接。科学发现的逻辑开始部分迁移到由大语言模型驱动的假设生成与理论探索中。未来的科学家不仅需要进行实证验证，更需要具备与“生成式理性”互动的能力，从而在更广阔的假设空间中发现新的研究路径。

第五，伦理与责任结构也必须随之更新。传统的责任归属依赖明确的作者与署名系统，而在“生成式理性”主导的知识生产中，如何界定提示者、模型开发者与最终用户之间的责任边界，如何在多主体协同中保持学术诚信，成为必须面对的新课题。

第六，“生成式理性”不仅改变了学术生产，也将深刻影响教育模式。未来的教育不再只是灌输固定的知识内容，而是训练如何设计有效的提示，如何理解人工智能生成的逻辑，如何在“生成式理性”与人类判断之间建立新的认知协同。

面向人工智能的未来，“生成式理性”不仅是理解大语言模型的一种哲学工具，更是塑造21世纪知识社会的一种认知基石。我们正在进入一个人机共创成为常态、知识加速生成与批判性审查并行的时代。哲学在这一时代中的角色也必将发生根本性变化。它不再仅仅是人类反思自身的学科，更是反思人机共生认知结构、探索“生成式理性”伦理边界的前沿阵地。如何在“生成式理性”的技术潜力与伦理风险之间寻求平衡，如何在效率的加速与深度的反思之间建立新的学术规范，将是哲学、科学与社会必须共同面对的重要课题。

综上所述，“生成式理性”继承了启蒙以来理性传统的逻辑推演能力，同时突破了主体性理性的限制，为理解未来人机共生时代的认知机制、知识结构与伦理规范提供了新的理论支点。

大语言模型不仅作为辅助工具而且作为“研究伙伴”参与知识生产，这对人类来说是史无前例的存在经验与认知方式。哲学必须率先回应这场影响深远的变革，在反思知识建构体系中人类角色的同时，对大语言模型的知识生产功能进行定位。“生成式理性”概念的作用正在于此。亚里士多德《形而上学》开篇即言：求知乃人之天性。知识生产范式的当下转型亦可作如是观，而其中的关键就在于，如何确保这种新的存在经验与认知方式有利于守护人之天性，而不是相反。

（责任编辑：周小玲）

“Generative Rationality”: The Paradigm Shift in Knowledge Production Triggered by Large Language Models ——A Preliminary Exploration Based on the Generative Philosophy Model Framework

WANG Jinlin

Abstract: Large Language Models (LLMs) are not only capable of generating human-like texts but also driving a fundamental shift in knowledge production from a human-centered model to one of human-machine collaborative co-creation. This paper examines the application of LLMs in philosophical research and their profound impact on the paradigm of knowledge production. Centering around four key dimensions, it discusses the construction of generative philosophical models and the transformation of knowledge production paradigms: (1) Theoretical Foundations, which analyze the philosophical debates surrounding LLMs' capacities for understanding, creativity, and authorship, inquiring whether they possess genuine understanding and creativity; (2) Methodological Approaches, which provide detailed exploration of how to build generative philosophical models through fine-tuning, prompt engineering, and retrieval-augmented generation (RAG), while emphasizing the indispensable role of human experts in evaluating AI-generated content; (3) Paradigm Transformation, which explains the profound shifts in knowledge generation, validation mechanisms, authorship, and knowledge dissemination, highlighting a transition from subject-centered creation to structural collaboration; (4) Explanatory Framework, which introduces the concept of Generative Rationality as a key theoretical framework for understanding this transformation. Generative rationality differs from traditional notions of rationality centering on subjectivity and intentionality; it is a functional form of cognition embedded within algorithmic structures and human-machine collaborative systems. The paper argues that LLMs not only expand the boundaries of philosophical inquiry but also pose unprecedented challenges to the ways in which human knowledge is produced.

Keywords: Large Language Models; Generative Philosophical Models; Knowledge Production; Human-Machine Co-Creation; Generative Rationality

（上接第 183 页）

On the Application of Common Sense in Judicial Decisions

WU Fei

Abstract: The phenomenon of “anti-common sense” emerging in judicial practice has impacted the stable expectations of the judiciary and highlighted the urgency of exploring the precise application of common sense. Common sense exhibits a hierarchical structure: superficial experiential knowledge, intermediate judgment rules, and deep-seated value concepts. It is simultaneously characterized by self-evidence and refutability. In judicial decisions, common sense serves four primary functions: (1) as a self-evident fact, (2) as a basis for evaluating evidence, (3) as a major premise for factual inference, and (4) as a reference for correcting case facts. The application of common sense must be premised on judicial scrutiny. Its content must satisfy the normative requirements of the judicial process, achieve compatibility with the specific circumstances of individual cases, and fulfill necessary justificatory obligations. These conditions also constitute limitations on the application of common sense. As the content and boundaries of common sense are continually reshaped, its application must consistently maintain a subordinate role in fact-finding, serving the broader objectives of the legal order. Simultaneously, the application of common sense should entail necessary practical reflection. To avoid descending into an impoverished predicament, the judicial decision must consistently ground its legitimacy in the rationality of common sense.

Keywords: Common Sense; Self-evidence; Refutability; Case Fact