“映射”视角下人工智能“思维”的哲学探讨

A Philosophical Exploration on the “Thinking” of Artificial Intelligence from the “Mapping” Perspective

王维国 /WANG Weiguo　　思敏康 /SI Minkang

（西安交通大学马克思主义学院，陕西西安，710049）
(College of Marxism, Xi’an Jiaotong University, Xi’an, Shaanxi, 710049)

摘　要：大语言模型以符号计算持续呈现出近似“理解”与“推理”的外在表征，这一现象再次引发了关于机器能否真正形成“思维”的多维探讨。本文引入数学中的“映射”理论作为分析视角，对人工智能的语言表征、“思维”表象与“黑箱”困境进行哲学阐释可以发现，人工智能所呈现的“理解”是认知主体对自身主观体验的归因性解读，“黑箱”问题则源于高维符号结构难以使人类主体形成有效认知映射的客观局限；人工智能的发展是在人类认知映射链中的层级性拓展与延伸，仍要以人类既有的认知形式与符号体系为边界。

关键词：大语言模型　人工智能　思维　映射

Abstract: As large language models exhibit continuously external representations approximating “understanding” and “reasoning” through symbolic computation, and this phenomenon has once again sparked multi-dimensional discussions on whether machines can genuinely form “thinking”. This paper introduces the mapping theory in mathematics as an analytical perspective to provide a philosophical interpretation of linguistic representation, “thinking” representation, and “black-box” dilemma in artificial intelligence. It can be found that the “understanding” manifested by artificial intelligence is an attributive interpretation of the cognitive subject’s own subjective experience, while the “black-box” problem stems from the objective limitation that high-dimensional symbolic structures are difficult for human subjects to form effective cognitive mappings. The development of artificial intelligence constitutes a hierarchical expansion and extension within the chain of human cognitive mappings, and it still takes the existing cognitive forms and symbolic systems of human beings as its boundaries.

Key Words: Large language models; Artificial intelligence; Thinking; Mapping

中图分类号：TP18; B80　DOI: 10.15994/j.1000–0763.2026.10.007　CSTR: 32281.14.jdn.2026.10.007

大语言模型发展至今，向用户展示其详细“思考过程”已成为普遍现象。这一能够增强语言模型推理能力，被称为“思维链”（Chain of Thought, CoT）[1]的技术，其运行方式是让大语言模型在输出最终回答前，模拟人类多步推理的思维模式，将完整的推理过程以文本形式呈现给用户。在用户的认知体验中，模型的每次“思考”似乎是在深度“理解”提示词语义的基础上，进行着与人类认知模式相似的逻辑推理活动。

这一现象再次引发了一个从1950年“图灵测试”提出、1980年“中文屋子”（Chinese

收稿日期：2025年12月12日
作者简介：王维国（1967–）男，陕西榆林人，西安交通大学马克思主义学院教授，研究方向为马克思主义理论研究。
Email: renda666@163.com
思敏康（2000–）男，陕西横山人，西安交通大学马克思主义学院博士研究生，研究方向为马克思主义理论研究。
Email: siminkang@qq.com

room）[2]思想实验问世以来，至今仍未形成共识的问题：以大语言模型为代表的人造智能体，是否真正实现了对人类语言的理解？这就表明，从哲学认识论的视角出发，对这一问题进行深化探讨就成为当前人工智能交叉研究领域的一个重要课题。

一、大语言模型的逻辑推理与机器的“理解”

“中文屋”思想实验是由塞尔（John R. Searle）提出的。该实验设想，一个完全不懂中文的人被封闭在一间屋内，外界通过传递纸张的方式提出中文问题。尽管屋子里的人不懂中文，却可依据实验人员预先提供的规则手册，对汉字符号进行整理、匹配与输出，进而使外界观察者产生其能够正确回答中文问题的错觉。这一思想实验旨在阐明，在特定语境下，对语言符号形式的操作并不能等同于对语言语义的真正理解，进而对“图灵测试”作为智能问题的评判标准的可靠性提出了质疑和挑战。2025年，沈宏梁等人在对机器翻译的哲学探究中提出：“通过大规模语料库的训练，机器可以学会不同于人的对人类语言的‘理解’，从而完成翻译工作。”[3]在研究中他们发现：“翻译质量越来越高的机器翻译方法却是走的一条越来越脱离模仿人类翻译过程的路线。迄今为止机器翻译技术的进步与理解语义没有呈现正相关，而是一种反向的关联。即翻译技术的进步是越来越远离人类理解的依赖。”[3]在翻译实践中，一些“在人类看来貌似需要理解才能完成的智能行为，不通过理解也完全可以实现。”[3]他们进一步认为：“理解只是一种有用的帮助而不是一种必需的根本。”[3]

从另一个角度看，翻译活动中的“歧义”现象并非单纯源于语言系统本身的缺陷，而是与不同认知主体对语言、语境要素的差异化感知与解读密切相关。当翻译实践需为某一语句赋予确定语义时，认知主体之间的理解差异会进一步凸显该语句潜藏的多义性特征，进而对翻译的准确性与一致性构成挑战。与之相比，机器翻译并未介入人类翻译活动中所固有的主观解释过程，其运作逻辑是依托大规模语料中上下文的关联概率，筛选出整体语境中最为常见、最为可能的对应表达，从而一定程度上规避了人类认知主体在语义理解过程中产生的分歧。这使得一些在人类看来似乎必须依赖深度语义理解才能完成的翻译任务，却通过符号关联性的计算方式得以实现。这也揭示了机器翻译与人类翻译在认知逻辑上的明显差异。

生成式语言模型与机器翻译的技术路径有所不同，但在“模拟文本数据的生成概率，实现对自然语言的理解和生成”[4]这一点上并无显著区别。当机器通过对文本数据关联性的计算，模拟人类语言“理解”过程、呈现近似人类的思维表征，并能与人进行有效对话互动时，人们自然会追问：仅以纯文本符号为训练材料，为何能使模型获得近似人类理解、思考、推理、交流等认知活动的功能？正如赵汀阳所追问的：“人工智能仅仅通过寻找数据的关联性，为什么能够形成相当于知识和思想的效果？”[5]这一疑问直指大语言模型的“黑箱”特性，亦是其语言“理解”方式与生成机理“可解释性不足”的一种体现。

CoT展现的推理过程，起初似乎能让用户感知到模型给出的答案遵循了何种推理路径。[6]这项技术对大语言模型性能的提升效果显著，但模型内部的真实推理机制依然需要进一步探究与阐释。这是因其以文本形式所呈现的“思考过程”仍是模型输出内容的组成部分，并没有突破对词元（token）间关联概率的依赖。这些输出内容虽然直观上契合人类思考与推理的形式规范，实际上并不必然意味着模型内部发生了与人类认知模式类似的推理活动。然而，若仅将其看作一种流程性操作环节，又难以解释CoT为何能切实带来模型的语义理解能力和输出效果的提升。这种理论与实践之间的张力，更引发了学界的疑惑：模型是否能在纯粹的语言训练中，形成一定程度上的“思维”或“类思维”能力？或者说，人工智能是否有可能通过对语言符号的掌握，“涌现”出真正意义上的思维活动？要回答这一疑问，就需要回到人类语言与思维的内在的关系。

二、物理形式与思维内容的边界

现代语言学一般认为，语言不仅仅是表达思维的工具，语言与思维之间可以相互影响、相互塑造。自然逻辑最初认为语言“无关乎观念的形成，只是为了表达业已形成的思想以实现交流的目的”。[7]这就是说，语言不参与思想的构造。洪堡（Wilhelm von Humboldt）认为，语言“虽然始终只能在每一次具体的思维行为中得到实现，但整体上并不依赖于思维”。[8]这意味着，个体的语言行为固然需要由其思维来驱动，但作为整体的人类语言却具有相对独立的存在形式。而后，沃尔夫（Benjamin Lee Whorf）作为“语言相对论”的代表人物，在语言与思维关系问题的讨论中，显著推动了语言功能的理论研究和认知深化。他的基本观点被王颖冲总结为：“语言形式决定语言使用者对宇宙的看法”；“世界上的语言具体特征不同，所以各民族对世界的分析也不同。”[7]这就是说，不同的民族因语言不同而形成整体上不同的思维方式，语言对思维有着深刻的影响和塑造作用。而维果茨基（Lev Semyonovich Vygotsky）则主张人类的语言和思维起初分别以思维的“前语言阶段”和语言的“前智力阶段”为特征相互独立发展，随后在某一历史时刻交叉，开始“相互联系、相互依存、相互制约”。[9]

理论研究发展至今，人们在描述语言概念及其与思维的关系时，习惯于将二者均看作某种具有特定功能的存在。在人的直观体验中，语言的使用总是伴随着物理形式与内容含义的同步出场，无意义的符号和无符号的思维活动都不能构成有意义的语言。此二者不可分割，因而语言通常被视作物理形式与思维内容的统一体。[10]语言以物理形式承载一定思维内容，其含义是公共性的，表达又是个体化的。公共性体现为语言的约定俗成的意义，特别是对外部世界的指称，有着能够使共同体广泛理解并认同的指向性；个体化则如前文的“歧义”所述，个人对于公共性的语言文本也可以有着基于自身经历的独特意义体验。这使得传统语义学认为，“语词的意义不外是语词使用者关于这个语词所具有的观念，这种观念存在于该语词使用者的内心世界中或者为他（她）的内心世界所把握；它决定了语词的外延（指称）。”[11]

语义外在主义主张，语言的意义并不存在于认知主体的内心观念之中，而是由外部客观对象、指称关系以及语言共同体的共同使用与社会分工所确定的。也就是说，“语言的意义不是个人所能确定的”。[11]不过，尽管赋予语言意义的是公共性活动，理解语言意义却始终离不开个体思维。这是因为意识、思维、心智等等具体的精神活动都要在认知主体精神世界中进行，“意识作为人的生命组织的内在信息系统”，其思考过程“只有自己才能确切地体验（所谓‘自心知’）”。[12]不同个体间无法“心灵感应”似地直接传输思维，因而人们要相互交流思想，就不得不借助共同生活的外部世界中各类不依赖于主观而存在，又可被共同感知的物理形态。如一定频率和节奏的声音（有声语言）、人的肢体动作（肢体语言）、人造符号（文字）等。可以说，语言之所以能向不同认知主体表达公共性意义，前提在于其物理形态的共同可感性及其相似的语境，这体现着人类依靠语言交流所必需依赖的生理基础和社会环境。维果茨基注意到了这一物质前提，他将动物和人共有的无指向的“表述性发声反应”[9]视作语言发展的“前智力阶段”，也即说明所谓意义、语义并不天然地与特定的声音、符号等物理形态相关联。

在一些学者看来，符号系统只有在认知主体思维中“内含地自带解码自身的自反能力”时才成为语言，否则“只是一个符号系统”。[5]人类的语言符号系统不仅是认知主体间实现思想交流的物质前提，因其可转化为供计算机解读的机器语言形式，故而也成为大语言模型诞生的物质前提。当今时代人工智能借由纯粹文本符号实现对语言意义的“把握”，使“语言”概念变得更加需要深思——作为物理形态与思维内容统一的存在，它的物理符号形态何以内在地包含着可把握的主观意义？若要证明大语言模型拥有真正“思维”，就不得不先证明语

言的符号和背后的意义具有着某种条件下的同一，而实际上具体语词与意义的搭配又是人为的。这就需要对人工智能的所谓“理解”与“思考”做出进一步的探究和阐释。

康德的哲学中，“现象就是指事物向我的感官所显现的样子，而事物自身则是它本来的样子，也就是所谓的‘本体’”。[13]虽然康德的二元论立场是需要超越的，但其提出问题的方式有一定启发意义。个体通过感官传递信息，而后在头脑中形成对事物的映象。树木并不天然以“树”这一概念而存在，无论其具有何种属性，“树”的概念都是人类为了认识和区分这类事物而在思维中建构，并以一定符号构造为语词。当人看到这一符号时，脑海中会联想到客观世界中某一类对应的存在物，这使得人们直观上认为“树”这个词语指的就是那类客体。但严格地说，无论语言的物理形式还是客观事物，呈现于认知主体思维中的是由感官收集、思维加工后的关于客体的反映。能动的反映论则从人的认识能力与外部世界的客观性两方面明确了基于人类实践活动的“本体”与“映象”之间存在稳定的对应关系，这样的对应关系同样发生在语言的物理形式与认知主体思维之间——语言能够表达意义，是因为它的物理形式始终对应着一定的思维及其指向。

这样的对应从何而来？主观因何而产生？不同的哲学家对此给出了不同的解释，科学界也一直在对此开展研究。“早期认知科学开始通过将心脑认知思考过程与计算机的结构和工作方式相类比来解释人类的认知和思维过程。思想和思考在这种类比中被看成是对信息的符号计算加工和处理：感知系统从环境中摄取信息，思维通过对储存信息进行符号转化和加工处理进而形成决策和判断，行为对判断和决策作出相应的行动。”[14]以人工神经元网络为技术进路的联结主义则认为，有关我们直观上认为的概念、思想“的信息是被整体地、彻底分散地保存在可能几亿个神经元之间互相联结的方式中，大脑对信息的处理是完全整体化的、不能分解的。”[14]认知科学更多地把心智看作一种信息加工系统，即外界输入经由大脑计算或神经元联结等中间过程，形成内部表征并指导行为。他们从思维表征、加工等中间过程解释主观的成因，却并不能充分说明外界输入的物理信息如何对应转化为思维处理的内部信息。

当代意识神经科学试图通过聚焦意识的神经相关物（neural correlates of consciousness, NCCs）来解释意识。查尔默斯（David John Chalmers）就提出一种假设：“意识的神经相关物（NCC）是一个最小的神经系统N，它满足从N的状态到意识状态的映射（mapping）关系，其中在条件C下①，N的给定状态足以对应相应的意识状态。”[15]他之所以认为需要一个理论上最小的神经系统，是因为大脑神经系统整体足以产生意识，但指出这一点并不能解释意识本身。

查尔默斯想要回答的问题是“什么神经现象足以产生意识？”，这是意识神经科学的研究方向之一，但远未得到理想的结果。而即使该假说在未来得到验证，新的问题又出现：NCC的生物信号对应转化为个体自感知的意识活动的方式又是什么？可以说，即便人们对心智活动的认识逐步深入到人的大脑、神经系统等生理层面，但这些生理活动本身仍保持着可被观测的物质形态，未能真正说明主客观之间的边界如何被跨越。至少截至目前，人们“对大脑的任何科学知识都没有促进我们对意识感受性（qualia）的理解”。[16]这就表明，在涉及思维、意识的问题上，科学检验有着与哲学反思一样的难题，需要进一步深化研究。但以“映射”描述神经活动与意识现象间关联的研究思路，启发我们认识到：对主观领域的未完全认知并不否定主客观间的“对应”关系本身的恒常性与稳定性。

①条件C可以被理解为：在大脑正常活动的前提下，允许出现非寻常的输入以及有限程度的脑刺激，但不包括脑部损伤（lesions）或其他结构层面（architecture）的改变。通俗理解就是，设定条件C主要为了强调在该假说中大脑需处于正常活动的相对稳定状态，不应有过度刺激或生理性状改变。

三、“映射”视角下的语言与人工智能“黑箱”

“映射”是现代数学中的一个重要概念。它是指在A与B两个集合中，若存在一定法则f，使得对于集合A中的每个元素x，按此法则都在集合B中有唯一确定的元素y与之相对应，则称法则f为从集合A到集合B的一个“映射”。其中，“y称为x在映射f下的象，x称为y在映射f下的一个原象”。[17]数学中，这一法则可以表示为确定的函数。从物理符号到个体的感官信息、感官信息到神经信号、神经信号到意识活动等等在质上不同，却在人类世界中实际存在着某种恒常的对应关系，可以用“映射”加以形式化描述。在“映射”视角下，语义、意义等不一定要看作“语言”这一实体性存在的内部固有的要素或属性，其机理在于物理符号能够通过“文本A→感官信息a→神经信号a→思维A”的一系列映射在人类主体心智中激发相应思维活动；而人类认知主体也可以通过“思维B→神经信号b→肌肉状态b→肢体活动b→文本B”的映射链条完成语言符号的生成。

需要明确的是，引入“映射”概念并非简单对数学领域相关概念的引申与套用，它既是对主客观之间关联关系的一种描述，也是尝试提供一种独特的学术分析范式。“映射”不表征客观存在与主观意识之间的具体因果关联，而是在“主观–客观”的哲学划分的基础上，揭示二者之间某种恒常关联的必然存在。正是基于对这种关联关系本身的把握，个体才能通过构建特定的映射关系，传达自身的认知理解，并在此基础上形成群体层面的认知共识。以“映射”视角来看，个体间进行语言交流过程中的“语义理解”，发生于物理符号到符号接收者思维层面的映射过程之中。“语句的意义”和“语义的理解”，实则是从不同视角上对同一过程的双重描述：其中，“意义”侧重语言符号能够使人类认知主体稳定获得特定主观意义体验；“理解”则侧重描述接收者从接收物理符号到形成对应主观意义体验的完整认知过程。语言的“意义”并非独立，其确立依赖于符号在个体思维中的有效映射，表征着人类群体在与语言符号互动过程中所形成的“家族相似①”性主观体验。这也就进一步表明：人类将自身主观体验对象化地投射于物理符号之上，使得认知层面能够将“意义”视作语言这一实体性存在的“内在属性”。而语言“意义”在人与人之间的有效传达，实际上离不开特定的语境或场景的支撑。

这种将主观体验对象化并赋予外在事物的认知与实践方式是人类长期的实践中形成的，自结绳记事的原始认知阶段开始，它便成为人类把握认识对象、传递知识信息的基本方式。这种认知方式依赖于客观现象自身的恒常性——正如地球绕太阳的公转不因个体或群体无法观测而停止。与之形成鲜明对比的是，语言意义或指称的恒常稳定则是人类社会发展进步中所形成的重要认知智慧。在长期的社会实践中，社会共同体创造并维护了共享的语言符号形式与使用规范，使得不同认知主体可以通过同一物理符号的刺激，在各自心智中形成相似的意义体验，进而实现认知与信息的传递。从认知机制来看，语言通过“事物A→思维中的A→文本a→思维中的A→事物A”的完整链路，在人类个体思维中实现指称的构建。对语义的理解固然离不开个体的主观认知活动，但社会对语义的普遍规定与共同认可，使人们容易混淆“规范的稳定”与“特性的稳定”这两个概念。语言符号是人造标识物，语言概念则是人为定义的抽象范畴，其稳定性并非源于自身固有的、不依赖于主观认知的恒常特性，而是基于社会实践需要，通过主观层面的规范与共识得以确立。以此种视角来看，“语义理解”更应被界定为物理符号在接收者个体心智中映射特定思维活动的动态认知过程。语言概念可以看作为此类映射关系及其两端的物理符号和

①这是哲学家维特根斯坦在语言的意义方面所提出的一个概念。他认为词语的用法在不同的语境中，是有所不同的，如同家族相似关系，即一个成员与其他成员至少有一个或多个共同属性。

思维活动所构成的有机整体，最终在“物理符号–思维活动”的映射关系中实现自身的统一。

而由于意义是人的一种自感知体验，有的意义还是通过批判性自我反思才得以生成，因此符号本身不直接等同于意义。至少在现阶段，大语言模型无法通过对语言符号关联性的统计计算，真正形成类人的思维活动。正如赵汀阳所指出的：“人工智能的思维材料是本身不含知识或语义解码的标识”，“学会与人对话的人工智能事实上学到的不是有着人文和知识意义负荷的语言，而是由无数关联性或无穷可能链接构成的标识系统”。“人工智能输出的标识链接在人类眼里是语句，但人工智能并不理解其中的含义”。[5]换言之，“含义”本身即是对人类的“理解”活动的另一种表述，不同“语境”下语词的“语义”是在个体思维映射过程中被察觉和把握的。由于人类习惯于将自身主观自感知的体验以“概念及其含义”的形式赋予了外在符号，因而通常认为，大语言模型只有将语言的“含义”属性内化，亦即实现真正的“理解”，才能形成合乎规范的输出结果。而事实上，模型仅输出相应的物理符号序列，是人类依托自身的映射机制与知识背景，才从中生成了相应主观意义体验。如此一来，问题也就从“机器能否内化人类语言的某种属性”，转化为“机器能否通过不断的尝试性表达，获得人类的理解和认可”。

机器正是在训练过程中建立了自然语言符号与词元之间的稳定“映射”，进而实现对公共语言符号系统的拟合。不过，在仍以语言符号为运作载体的前提下，模型所呈现出的“语义理解”和“推理思考”表象并无显著区别。这就是说，无论是对词语的处理还是对上下文关联的延展，模型都是以人类的输入语段为基础，依照既定算法流程对符号进行匹配、重组与生成，并将经由高维符号关联组织而成的文本结果重新反馈至人类感官。凡是进入人的视野成为有意义的事物，实际上都已被纳入由客观世界通向主观意识的映射链条之中。人类之所以产生“模型具备理解能力”或者“模型拥有思考能力”的直觉判断，正是由于在阅读模型输出文本时，自主完成了“符号→思维”的映射过程。当输出内容成功触发这一映射，并使人形成诸如“理解”“推理”等意义体验时，人便会将自身的主观体验对象化地归因于交互对象，即人工智能模型。同理，人工智能的“黑箱”难题不仅源于模型结构的复杂性，还源于其内部高维符号结构难以用人类所熟知的自然语言进行表征与转述。当此类对象与现象无法在人类思维中形成有效映射时，人们往往习惯性地将其归因于模型具有某种难以解析的未知特性。

从映射视角上说，当前学界对大语言模型“思维”与“黑箱”问题的判断，在很大程度取决于能否在主观层面建立有效映射。至少就现有的训练范式看，大语言模型对于符号关联性的处理，是在人类“符号→思维”的整体映射链条之上，拓展出“自然语言符号→机器语言符号→自然语言符号”这一新的“映射”层级，而真正意义上的语言“理解”活动，终究归属于人类认知主体。这也恰能解释，为何多数哲学家坚持认为大语言模型尚不具备，也难以形成通常意义上的人类心智。与之不同，有些科学家的态度是：即便当前模型未呈现真正的思维能力，未来未必不可能。此类乐观立场亦从侧面印证：作为一种描述世界关系的认知工具，“映射”原理具备可靠的解释力，由此形成的阐释具有公共性与可交流性，这也正是本文将其作为哲学认识论分析视角的依据所在。

四、人工智能的发展受制于人类既有的认识条件

一些计算机领域的专家对未来人工智能可能会涌现意识持积极预期，并对意识在复杂系统中的自发涌现机制做了探索性阐释。如有学者就提出，智能体若能构建某种内部符号或神经活动模式以表征自身，且此类表征能够被主动调用，则可以视为具备自我意识。[18]从当前技术水平看，大语言模型仍缺乏对客观世界的直接感知能力和自我纠错能力，尤其是未能达到人类思维层面那种将语言符号与外部事物深度联结的高度，亦即尚未形成人类水平的完整

映射关系。对此，李飞飞团队就提出，未来人工智能还需“突破现有大型语言模型的范式，转向一种更为根本的世界模型。这种新模型不仅能理解语义关系，更能在几何、物理和动态规则上一致地‘想象’和‘重建’世界。它应当能感知多模态输入，预测场景变化，并与环境进行交互。”[19]概括地讲，未来“通用人工智能”的基础，在于实现大语言模型、世界模型和多模态模型的深度嵌合，由此构成具备空间表征与行动推理的“空间智能”。

从“符号接地”[20]问题的角度看，通过引入视觉、听觉、动作等多模态信号，构建包含位置、速度、物理约束等要素的空间表征，一定程度上能够复现人类与外部世界互动的实践场景。这一路径意图将当前仅在语言符号系统内循环的大语言模型嵌入现实世界，以缓解其在常识推理、物理直觉、长期规划等方面的局限，并且有望在与客观事物的交互中形成近似“事物→模型内部表征→符号”的类人映射链。然而，无论语言模型所处理的文本数据，还是多模态模型与空间模型所处理的感知数据，实质上都是不同类型的人造符号编码系统。这一事实使得当前人工智能的发展，要受制于人类既有的认识条件。

一般来说，公共语言符号系统包含两个核心要素：其一，作为编码基本单元的字符，以可感知的外在形式差异实现区分；其二，具有形式差异的字符依照特定顺序排列组合为语句和篇章。正是符号的形式差异与组合规则共同构成了编码体系的整体性，使物质性符号能够在人类认知主体思维中映射复杂的精神活动。在自然世界中，鹦鹉、狗、马等动物虽可感知到人类语言物理形式的差异，并基于它们自身“智能”形成对指令性声音的简单映射，但因不具备符号的解码与编码能力，无法把握语言复杂的组合结构，因而难以产生真正意义的理解。与之相比，大语言模型在将语言符号转化为机器表征时，仍完整保留了单元差异性和编码整体性：人类感官所依赖的形式差异被替换为“词向量”的数值差异，字符的序列组合关系同样可被模型读取与统计。这表明，大语言模型的训练基础本质上源于人类对语言符号系统内部差异与结构的认识，而符号差异与组合方式本身又是基于人类对客观世界的共同可感所建构。因此，大语言模型不仅受限于人类对客观世界的认识水平，还受限于人类对语言符号系统本身的理解程度。

物理学对客观世界的数学表达与自然语言的日常表达，实质上都属于人类在映射链中构造的符号体系。相较于自然语言，物理学依托严格的实验程序与高度形式化的符号系统，使认知表达更精确、可计算、可重复检验。物理概念与规范虽源于客观世界，却同样是人类在思维中总结现象、形成认识后，经由多层映射转化而成的符号产物。尽管摄像头、传感器等设备可采集客观世界的物理数据，但其概念框架、数据标准与处理方式都是人类经验的形式化结果，设备始终在人类认识的框架内运行，并未超越人类与世界互动的基本方式。换言之，通常被视为客观规律的科学理论，仍以人造符号与人类认知活动同外部世界形成映射关系。它们既反映世界的客观属性，又不可避免地具有人为建构性和历史阶段性。人工智能与现实世界的关联，实质是对人类形式化认识的再呈现，因而难以突破人类认识的既有边界，且是在一个更精细的人造符号系统中拓展了人类映射链条的中间层级。只有当人工智能能够自主发现人类尚未意识到的外部世界中事物间的恒常联系，并独立形成科学理论时，才意味着另一种“思维意识主体”的诞生，其“智能”也将不再仅仅依附于人类的知识。

五、对人工智能“思维”的展望

赵汀阳曾提出追问：“人工智能是否会成为一个心灵？紧接着的问题是：心灵是什么样的？”[5]在当前人工智能的理论研究与实践发展水平下，这一系列问题仍是悬而未决的开放性议题。从认知主体的本质属性来看，人类是自身心智映射活动的主体，而人工智能在符号处理领域的迭代发展，其逻辑起点源于人类对自身认知机制的模拟；其技术结果与发展方向，

仍需要通过这一映射关系回归人类心智层面加以检验与校准，这也是人工智能实现“服务于人类”这一价值定位的基本要求。[21]这也确立了关于人工智能的哲学研究的基本议题，即数字技术、人工智能何以可靠，何以向善，亦即为知识与技术的客观实在性与价值理性奠定根基。据此，本文以“映射”作为分析视角，系统探讨语言表征、人类思维与人工智能的“心智”表象三者之间的内在关联，尝试将机器的符号处理能力、人类的语义理解活动以及AI“黑箱”现象的实质，纳入同一逻辑结构链条中进行整体性阐释。基于此可以看出，机器所谓的“思维”，并非是具备真正主观认知能力的体现，其实质上源于两方面因素：一是人类对自身主观体验的归因性投射，二是其高维符号表征向人类自然语言转化过程中的不可完全译回性。由此可以推知，未来人工智能技术的迭代进步，将体现为映射层级的持续扩展、外部校验的不断强化、高维符号结构向低维自然语言表述的可译回性提升，以及这三方面的高效协同。从哲学认识论的分析视角进一步审视可以发现，真正意义上的“理解”与“思考”，并非单纯的形式化符号处理或逻辑运算，而是依托于社会规范约束、在主体层面所实现的完整映射链条。这一链条既包含主体对外部世界的感知映射，也涵盖主观认知与客观实践的双向互动，而这正是当前人工智能发展需要进一步突破的认知边界。

[参考文献]

[1] Wei, J., Wang, X., Schuurmans, D., et al. 'Chain-of-Thought Prompting Elicits Reasoning in Large Language Models'[J]. *Advances in Neural Information Processing Systems*, 2022, 35: 24824–24837.

[2] 冯志伟、张灯柯、饶高琦. 从图灵测试到ChatGPT——人机对话的里程碑及启示 [J]. 语言战略研究, 2023, 8（2）: 20–24.

[3] 沈宏梁、王新、罗晖. 机器有机器的理解：机器翻译中理解问题的哲学思考 [J]. 自然辩证法通讯, 2025, 47（3）: 35–41.

[4] Ju, Y., Ma, H. 'Training Data for Large Language Model'[J]. arXiv preprint, arXiv: 2411.07715, 2024.

[5] 赵汀阳. 人工智能还给人类的思维难题 [J]. 中国社会科学, 2024,（8）: 101–123; 206.

[6] 杜家乐、陈曙东、叶亮 等. 大语言模型中的思维链技术综述 [J]. 无线电通信技术, 2025, 51（5）: 877–887.

[7] 王颖冲. 语言与思维关系再认识——沃尔夫《论语言、思维和现实》解读 [J]. 外语教学与研究, 2011, 43(4): 583–592; 641.

[8] 威廉·冯·洪堡特. 论人类语言结构的差异及其对人类精神发展的影响 [M]. 姚小平 译, 北京：商务印书馆, 1999, 75.

[9] 吴进善. 维果茨基的语言与思维关系理论解读 [J]. 西北民族大学学报（哲学社会科学版）, 2016,（2）: 124–130.

[10] 向明友、张兢田. 论语言与思维的关系 [J]. 同济大学学报（社会科学版）, 2009, 20（4）: 91–95.

[11] 陈亚军. 意义何以可能？——普特南的新语义学理论读解 [J]. 自然辩证法研究, 2001, 17（6）: 18–23.

[12] 陈伯海. “思”与“在”——意识活动探源 [J]. 社会科学, 2009,（9）: 154–163; 192.

[13] 吴增定. 笛卡尔的“普遍怀疑”意味着什么？——从胡塞尔和海德格尔的现象学视野来看 [J]. 现代哲学, 2025,（1）: 71–85.

[14] 冯文婧. 心智哲学视域中的意向性实在论叙事逻辑论析 [J]. 山东社会科学, 2022,（2）: 134–140.

[15] Chalmers, D. J. 'What Is a Neural Correlate of Consciousness?'[A], Metzinger, T. (Ed.) *Neural Correlates of Consciousness: Empirical and Conceptual Questions*[C], Cambridge, MA: MIT Press, 2000, 17–39.

[16] 李珍. 跨越“解释鸿沟”——意识科学中的生物学与信息论之辩 [J]. 哲学研究, 2025,（3）: 132–144.

[17] 王绵森、马知恩. 工科数学分析基础：上册 [M]. 北京：高等教育出版社, 2017, 12.

[18] Schmidhuber, J. 'Driven by Compression Progress: A Simple Principle Explains Essential Aspects of Subjective Beauty, Novelty, Surprise, Interestingness, Attention, Curiosity, Creativity, Art, Science, Music, Jokes'[A], Pezzulo, G., Butz, M. V., Sigaud, O., et al. (Eds.) *Anticipatory Behavior in Adaptive Learning Systems*[C], Berlin, Heidelberg: Springer, 2009, 48–76.

[19] 张佳欣. 空间智能将成AI攀登的下一座高峰 [N]. 科技日报, 2025–11–14（04）.

[20] 袁毓林. ChatGPT语境下语言学的挑战和出路 [J]. 现代外语, 2024, 47（4）: 572–577.

[21] 李钢、刘皆成. 人工智能价值对齐的实然困境与应然逻辑 [J]. 北京联合大学学报（人文社会科学版）, 2026,（2）: 11–20.

[责任编辑 王巍 谭笑]