·人工智能与未来社会（二十二）·

从大模型到世界模型：智能革命的认知跃迁与视界融通

肖　峰

【内容摘要】 世界模型是继大模型之后兴起的人工智能新技术。它克服了大模型作为一种“文本语言智能”的局限性，使人工智能具有面对物理世界进行因果推理、决策的能力，从而将人工智能推进到一个新水平，并启示我们对智能的本质形成更深刻的哲学认知。大模型与世界模型在模拟人类智能方面各有所长，两者的结合所形成的智能融合，可以在性能上实现优势互补，导向一场新的智能革命，体现出从多方面拓展智能观的哲学意义，由此走向多重意义上的哲学视界融通。

【关 键 词】 世界模型　大模型　智能观　智能融合

【作　　者】 肖峰，上海大学智能时代的马克思主义研究中心、上海大学智能哲学与文化研究院教授。（上海 200444）

【基金项目】 国家社科基金重大项目“数字智能技术与哲学发展及知识生产范式变革研究”（24&ZD320）

大模型的出现是人工智能发展历程中具有里程碑意义的事件。它将大数据、大算力与强算法结合，推动了人工智能的大规模应用和广泛参与，使智能技术日益广泛且深入地应用于社会各个领域，将人类带入智能时代。[①]然而，以 ChatGPT 为代表的大模型技术虽已推动社会生产力跃升，但其物理常识缺失、因果推理薄弱等局限，正成为制约产业深度应用的瓶颈。从工业机器人因环境理解不足导致的安全事故，到自动驾驶系统在复杂路况中的决策失误，再到医疗诊断 AI 因缺乏因果逻辑产生的“幻觉”风险——这些现实问题迫切要求人工智能实现从“文本智能”向“物理智能”的认知跃迁。于是，在大模型蓬勃发展、方兴未艾的当下，“世界模型”又跃入我们的眼帘，成为 21 世纪智能革命并行的技术浪潮。如果说大模型凭借对海量数据的学习和训练具有了强大的自然语言处理能力，尤其是文本生成与推理能力，那么世界模型则试图构建对物理环境的动态理解与行动能力，力图使人工智能水平实现跃迁。从大模型到世界模型，进而到两者的融合，正

① 参见肖峰：《大模型与智能社会：基于历史唯物主义的探察》，《中国社会科学》2024 年第 7 期。

在给人类带来一场新的智能革命，其丰富的哲学启示和意义也随之开显。

世界模型的含义及其与大模型的区别

世界模型（World Model, WM）正在成为人工智能领域的一种核心技术。其概念最早由大卫（David Ha）等人于2018年引入人工智能领域，旨在使智能体构建关于环境如何运作的内部模型，实现对物理世界及其运行机制的内部表征和动态模拟，从而使人工智能（AI）系统能够理解对象运动的规律和行动的结果，实时生成并更新对现实世界的预测，为人工智能提供规划和自主决策的能力，由此适应和应对复杂的环境。[①]世界模型包含物理规律、因果关系、空间关系等知识，通过构建一个虚拟的、高度逼真的环境来模拟现实世界的运行机制，从而具备对物理世界的理解和模拟能力，解决现有人工智能在理解现实方面的不足。世界模型并不是一个简单的仿真工具，而是融合了机器学习、深度神经网络、物理引擎等先进技术的综合体。

世界模型强调对物理规律和因果关系的理解，[②]其中各种实体、事件和关系都被精确地映射和呈现，并基于此来规划行动。而之前的人工智能技术，包括生成式大模型，主要依赖统计关联和数据驱动，缺乏对真实世界的动态建模能力。世界模型通过多模态数据（视觉、触觉、语言等）学习，建立动态的预测框架，能够回答“如果发生X，结果会如何”的反事实推理（counterfactual reasoning）问题。比如，在复杂的城市交通环境中，自动驾驶汽车必须实时处理其他车辆的移动、行人的行为、交通信号的变化等大量信息。世界模型就是要帮助自动驾驶系统预测这些实体未来的动作，包括行人轨迹和车辆碰撞的可能性，即模拟出不同条件下的路况变化趋势，进而给出合理的路径规划。

世界模型中的“世界”并非仅指物理意义上的地球或宇宙，而是指智能体所需要认知与交互的全部外部系统。其中的“模型”则是“内部表征”或“表示”，所以世界模型是对真实物理世界的建模，是智能体通过多模态感知与交互经验构建的对物理环境动态规律的可计算表征系统。它赋予人工智能系统类似人类的环境理解、预测未来及自主决策的认知能力，从而让机器像人类一样认知世界并与环境交互。从这个意义上讲，世界模型可被视为“环境模拟器”与“决策引擎”。我们亦可以从两大核心功能对世界模型加以归结，即构建对外部世界的内部表示以支持决策及预测外部世界的状态变化从而指导行动。[③]

大模型与世界模型是人工智能领域两个不同却又紧密关联的概念，所以对世界模型的了解必须在与大模型的比较中进行。有“人工智能教父”称号的杨立昆（Yann LeCun）认为，大语言模型（LLM）存在无法对真实世界建模的缺陷，其理解和逻辑能力有限。仅仅依靠在更大的数据集上训练更大的语言模型，人工智能永远无法达到人类的智力水平，所以生成式人工智能的技术路线注定失败。而世界模型能像人类婴儿一样，通过观察和体验来学习和认知世界。它通过理解物理世界的运作规律、因果关系和常识推理，更接近人类智能的本质，从而克服大模型技术的局限，满足大众对人工智能的深层期待。[④]

可以说，包括大模型在内的先前的人工智能所面对的是人为设计的理想环境，它们专注于符号世界与封闭条件下的任务，如游戏或文本生成。这类人工智能仅能在规则明确、状态有限的环境（如围棋棋盘、文本生成任务的语义空间）中，依赖人类预先定义的目标函数（如胜率最大化、文本流畅度）运行。其底层逻辑是基于数据分布的统计相关性，侧重于语言和知识的统计关联，

① David Ha and Jürgen Schmidhuber, “World Models,” 2018-05-27, https://arxiv.org/abs/1803.10122.

② Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, et al., “Toward Causal Representation Learning,” 2021-02-22, https://arxiv.org/abs/2102.11107.

③ Jingtao Ding, Yunke Zhang, Yu Shang, et al., “Understanding World or Predicting Future? A Comprehensive Survey of World Models,” 2024-12-01,https://arxiv.org/pdf/2411.14499.

④ 苇草智酷：《当前大语言模型最终都会被淘汰：杨立昆万字演讲实录》，2025-03-30，https://k.sina.com.cn/article_6792525966_194ddb88e019017d94.html；李冲：《杨立昆：构建“世界模型”》，《华东科技》2024年第7期。

缺乏对底层物理定律的理解。而世界模型技术由物理引擎支持，能理解现象背后的物理机制，实现对真实世界的物理仿真，完成对物理规律、时空关系及因果推理的建模。它基于因果链推理（如“打雷→下雨→路面湿滑”的逻辑推导）和反事实想象（模拟未发生事件的后果，如“如果移除支撑物，建筑是否会倒塌？”）来理解事物之间的联系，由此突破对数据表面关联的依赖，克服大模型因物理常识缺失而不能“理解世界”的根本缺陷，从而可以处理真实世界中动态、复杂的任务。

我们还可以从更多具体区别来理解世界模型的特征。例如，从驱动方式来看，大模型以数据为驱动，依赖对大规模文本 / 多模态数据（如图像、视频数据）的训练，需大量标注数据；而世界模型以环境驱动或物理规律驱动，依赖与物理环境交互（如通过传感器）分析物体运动模式，自主推导出诸如摩擦力、流体动力学等规律，无须人工标注物理参数。就“感知”能力而言，大模型没有感知能力，形象地说，它没有“眼睛”，“看”不到外面的世界，只能埋头于文本进行文字接龙，即预测下一个“Token”，或依赖已有的文本信息来生成内容。而世界模型有“眼睛”，它通过摄像头、激光雷达、力传感器等设备实时感知环境，动态构建对物理世界的内部表征。在多模态感知的条件下，世界模型能将视觉、触觉、听觉等输入转化为统一的潜在空间表示，并据此生成控制指令。从对象角度看，大模型的“世界”是离线、静态的文本数据集，它通过人类提供的文本指令或历史数据来间接“理解”世界，类似于通过阅读百科全书来学习知识，而非亲身探索，其输出的是符合统计规律的响应，其中贯穿的是概率导向；世界模型的直接对象是现实环境，它能够理解世界的多维特性，能够在三维空间中进行推理和互动，能够根据环境反馈调整行动，形成“感知—预测—行动”闭环，例如机器人通过视觉识别抓取物体位置，并在触碰后修正力度，其输出的是能够预测环境变化的安全行动策略，其中贯穿的是目标导向。从理解力上看，大模型的典型应用场景是对话、创作、信息检索等；世界模型的典型应用场景是机器人控制、自动驾驶、游戏 AI 等。从人工智能流派或范式上看，大模型技术基于深度学习，属于联结主义；世界模型则涉及环境交互和强化学习，属于行为主义框架下的技术。综合上述区别可以看到，大模型主要是“知识的容器”，其智能受限于训练数据分布与架构设计者的先验认知；而世界模型是“规律的探索者”，通过与环境的持续对话构建动态认知框架，其知识体系具有开放性与进化性。

世界模型迄今并未成熟，还处于探索阶段，或者说当前的人工智能尚未达到世界模型技术的水平，但这一概念或人工智能进路已经引起极大的关注。随着这一技术的不断发展，在不久的将来，我们或将见证真正的世界模型诞生。当然，即使在概念阶段和初期探索中，世界模型也预示着我们对智能本质的哲学认知必定会经历新的跃迁。

从大模型到世界模型：智能跃迁与智能观拓展

人工智能从大模型走向世界模型的技术进阶，将使人工智能变得更加智能，更接近人的智能。这一进程不仅标志着人工智能技术范式的新跃迁，更折射出人类对智能本质之哲学探索的新拓展，彰显了智能革命的多重哲学意义。

（一）从文本语言智能到物理实景智能

如前所述，包括大模型在内的先前的人工智能，主要用于解决人为设计的理想环境中的问题。例如，根据提示词生成文章时，其任务的边界被严格限定在“给定提示词→生成响应”的封闭循环中，这样的大模型并不理解真实世界，它所做的只是统计某个词语出现的概率，只是对文

本而非对世界进行建模，像鹦鹉“说话”一样随机生成看起来合理的字句，所以被本德（Emily M. Bender）形象地比喻为“随机鹦鹉”（stochastic parrots）。[①]

可以说，大模型所进行的这种“文字接龙”过程，没有面对真实的物理世界，不是与实际存在或运作的客观对象打交道，只是与人创造的代码、符号或抽象世界打交道。由此所表征的智能，被直接简化为规则推导与符号操作，这样的人工智能系统也成为脱离物理世界的抽象符号处理器。当一种智能只能与虚拟的文本世界打交道时，这种智能只能显现出“似乎具有智能”或“看起来像有智能”，但不是真正具有智能，而是“虚假”的或“人工”的智能，“artificial”就是一个兼具这两种含义的单词，所以“artificial intelligence”除了有“人工智能”的意思，还有“虚假智能”的意味。生成式人工智能的“幻觉”或编造、杜撰虚假事实问题，本质上源于大模型对世界运行机制的符号化模拟和统计建模本质，这使其模型倾向于输出训练数据中的高频模式，而非符合物理规律的动态推理和真实答案，由此使其不时显现出“虚假智能”的特性。

大模型所模拟的智能不以真实世界为对象和基础，不能处理真实世界中的问题，只能面对文本世界、处理语言符号问题，所以只能称其为“文本智能”或“语言智能”。真实的智能应是能够面对物理对象，与真实世界打交道并处理实际情境中的问题。智能是在真实世界中产生的，只有回归真实世界，才是真实的智能。世界模型以真实的物理世界为对象，与实际环境互动并处理开放世界中的现实问题，这样的智能可称为“物理智能”或“物理实景智能”，它为人工智能达到真实智能的水平奠定了本体论基础。

能够理解真实世界的人工智能才能通向真理。大模型的“知识”都是从“书本”中习得的，也是基于书本知识去生成“新知识”。从拟人化的角度看，这样的知识是抽象的、无生活经验作为依托的知识，是从符号的推演中生成的内容，或是对多种观点看法的统计学再现。大模型的知识生成依赖数据集的概率分布，是统计规律的产物，而非客观真理的集合。它只以内容的生成质量（如文本连贯性）作为核心评价指标，而不是以内容是否符合客观实际为标准。尽管能生成流畅的解释，但缺失物理常识，缺乏对因果机制的把握，其生成内容可能违背现实规律（如能生成关于“浮空椅子”的流畅描述，却无法解释其违背重力定律的原因），这表明其生成的“知识”本质上是符号间的概率关联，而非对世界的真实理解。[②]这也类似柏拉图的“洞穴寓言”：大模型只能“看”到墙上的影子（文本符号），而非真实世界。

世界模型将大模型的书本习得知识方式，转变为从观察中建立认知，在实操中增长能力。通过分析环境交互数据（如机器人抓取视频、自动驾驶路测视频），世界模型可以自主发现或提炼物理规律（如摩擦力、刚体运动规律），形成可泛化的“常识”。这种范式更贴近生物智能的学习机制，如人类婴儿就是通过观察与试错而非背诵百科全书来建立认知。由此，世界模型不再将知识编码为静态的符号网络，而是通过诸如“预测误差最小化”等机制，不断动态地调整对世界的理解，形成可随对象变化而更新的知识。它可以作为一个强大的模拟器，创建环境的全面元素并建模它们之间的现实关系。[③]因此，在杨立昆看来，世界模型的本质是对人工智能发展路线的认知纠偏。他认为，大语言模型缺乏对物理世界的真实理解，世界模型才代表了对智能本质的更深刻理解：智能不是数据的统计拟合，而是主体与环境持续交互中涌现的适应性能力。世界模型通过这种交互才可能具备复杂推理和规划的能力，所以人工智能 需要像婴儿一样通过观察和互动来学习知识、增长智能。而“仅仅依靠语言和文字训练出来的人工智能系统，永远无法逼近人类的理解力”，[④]因为语言文本中蕴含的信息，远远不如物理世界所蕴含的信息那样丰富、复杂，大语言

① David J. Chalmers, “Could a Large Language Model Be Conscious?” 2023-08-09, https://arxiv.org/abs/2303.07103.

② 蔡基刚、林芸：《学术论文写作的挑战与变革：借助ChatGPT直接生成一篇学位论文的实验》,《北京第二外国语学院学报》2024年第4期。

③ Jingtao Ding, Yunke Zhang, Yu Shang, et al., “Understanding World or Predicting Future? A Comprehensive Survey of World Models,” 2024-12-01, https://arxiv.org/pdf/2411.14499.

④ 杨立昆：《AGI即将到来是无稽之谈，真正智能要建立在世界模型之上》, 2025-03-22, https://finance.sina.cn/2025-03-22/detail-ineqpzky2533883.d.html。

模型作为世界的文本投射，远不如世界本身厚重，所以唯有“世界建模”（world modeling）才是机器达到人类智能的正确途径。①

可以说，从“解决人为问题”迈向“理解真实世界”，就是从“语言智能”向“物理智能”抑或从“文本智能”向“实景智能”的跃迁。在这一跃迁中，人工智能从符号处理器进化到世界模拟器，所面对的对象从文本世界转向现实世界，智能哲学的主题也就从人工智能与人的智能的哲学关系，进阶到人工智能与实在世界的关系，且通过解决这一关系，来更好地解决人工智能与人的智能的关系问题。

人工智能从大语言模型到大世界模型的进阶，使智能的本质问题进一步凸显。如果说大语言模型主张“智能的本质是语言”，那么形成的是与真实世界脱节的“语言智能观”。世界模型主张智能的本质是对物理世界的动态建模，力求形成与真实世界打交道的智能，由此才可能形成真正的类人智能。可以说，人工智能必须从文本世界通过世界模型走向物理世界，实现符号系统与物理约束的深度耦合，才能突破“文本游戏”智能观的局限，实现智能研究从“语句接龙”向“物理实践”的范式转移，使智能迈向对真实世界的理解，由此获得智能的“真谛”。

（二）从符号操作智能到具身交互智能

大模型基于语言文本来理解智能，相当于秉持符号操作的智能观；而世界模型基于物理实景来理解智能，即主张具身交互的智能观。从大模型到世界模型，意味着对智能的关注焦点从认知智能转向行动智能，抑或从离身智能转向具身智能。

传统人工智能秉持的是笛卡尔身心二元论的认知观，它诉诸从抽象符号或数据中学习，所进行的认知是脱离物理世界和身体活动的抽象符号处理过程。尽管大语言模型将其推进到“语词接龙”的新层次，但本质上仍是与身体无关的符号操作。在其视野中，认知或智能活动都是离身进行的，是去身体化的现象，而非与世界动态交互的“具身认知”过程。所以，传统人工智能的知识既脱离物理世界，也脱离身体活动。由于不具身，尽管大模型可以生成关于“疼痛”的医学论文，却无法体验疼痛的生理意义，这也暴露了“符号操作智能”的本质缺陷。由于不具身，大模型不能与物理世界形成互动，既无法感知真实对象，也无法对真实对象施加作用，形不成改变现实世界的行动。因此，这样的智能至多是认知智能，而不可能是行动智能，从而不是对人的智能的完整模仿。

如果将具身智能视为一种专门的人工智能技术，那么世界模型与具身智能（embodied intelligence）具有紧密的关系。世界模型被视为发展具身智能的关键步骤，或者说世界模型的核心使命就是实现具身智能，就是将智能过程从“被动数据拟合”转向“主动环境交互”，即通过虚拟或真实环境中的持续交互去完成复杂任务。这样的人工智能才能成为与真实世界打交道的人工智能，才是真正能改变世界的具身智能。世界模型与具身智能的密切关系还可以形象地理解为：世界模型是具身智能的“大脑”，提供环境理解与规划能力；具身智能则是世界模型的“身体”，通过物理交互验证并优化世界模型。由此，世界模型是具身智能的认知基础，具身智能是世界模型的物理延伸，世界模型与具身智能有机结合，可以实现从“语言理解”到“物理行动”的闭环；结合具身智能的世界模型，通过传感器与执行器构建“身体—环境”反馈环路，形成具身交互闭环，实现“感知→建模→决策→反馈”的完整行动链（如自动驾驶的实时路径规划）。由此导向的智能观契合梅洛–庞蒂的现象学具身认知理论：智能必须扎根于与世界的互动，认知必须产生于身体与环境的持续对话，脱离物理交互的纯符号系统仅是“无身的幽灵”，不是真正的智能。如果说

① 明朝：《“世界模型”时代——人工智能影像的世界性实践》，《电影新作》2024 年第 4 期。

大模型的智能观受限于语言数据，那么世界模型通过具身认知更新了我们对智能的看法，将智能的边界扩展至语言无法覆盖的具身经验领域，在通过机器人、自动驾驶系统等载体将“智能身体化”的基础上，构建出能有效与真实物理世界互动的系统，由此推动人工智能从数字空间向物理世界渗透，从认知智能向行动智能延伸，进而推动通用人工智能（AGI）的发展。

智能观的这种跃迁也体现在学习的效率和效果上。大模型脱离了具身的行动，因此由其决定的机器学习方式也效率极低，成本极高，要具备某种特定的性能水平，所需的训练样本数量或试错次数极为庞大。而人类或动物通过身体与对象的互动去理解世界的运作方式，学习新任务的速度就非常快，可以通过较少的样本甚至零样本学习某些知识或技能，这就是世界模型所要模拟的学习方式。人工智能的智能观由此实现从“封闭抽象”到“开放具身”的范式转换，智能被视为在与环境的具身互动中生成，而不是在静止封闭的符号世界中出现。

（三）从统计相关智能到因果推理智能

从大模型到世界模型，对智能本质的理解还显示了从统计相关到因果推理智能观的跃迁。

智能系统的核心使命在于揭示事物间复杂关系的本质规律。当前主流的人工智能范式以深度神经网络为基础，依托海量数据建立输入与输出的统计关联性。这种基于概率建模的范式虽然取得了显著成就，但其本质仍停留于数据拟合的层面，未能实现对物理世界因果机制的深层理解。大语言模型作为概率建模的典型代表，其运作机制建立在文本序列的连贯性预测之上。通过自监督学习捕捉词汇共现模式，模型构建起庞大的条件概率分布矩阵，这种基于 Transformer 架构的注意力机制擅长发现数据中的表面相关性，却无法建立实体间的因果链。例如在图像生成领域，扩散模型可能产生“漂浮的椅子”这类违背物理规律的图像，其根本原因在于模型仅学习到椅子的视觉特征分布，却未掌握支撑结构与重力作用的因果机制。这表明基于相关性而非因果性的模型无法触及实在对象的本质，从而无法更深入、更准确地理解和解释世界。

世界模型的构建标志着智能系统认知能力的重大突破。该模型借鉴人类的反事实推理机制，采用结构因果模型（structural causal model, SCM）构建环境动态方程。具体而言，通过引入物理先验约束（如能量守恒、动量定理）保证系统的因果一致性。在训练范式上，采用分层强化学习框架：底层通过物理模拟器建立基本因果图，中层构建反事实推理模块，顶层实现目标导向的规划能力。这种结构使系统不仅能回答“是什么”，更能解释“为什么”和“如何改变”，能够推演“如果改变某个变量，结果将如何变化”这类反事实问题。由此，将人类的物理直觉与因果推理能力赋予机器，使人工智能获得理解物理规律的能力，能够理解自身行为所产生的后果，从而拥有关于世界的因果模型。

由于不能理解动态的因果关系，大模型缺乏对动态世界的理解和预测能力，尽管它在模式识别与概率推断方面展现出强大的能力，但其内在机制仍存在根本性局限：无法建立动态因果链条的完整认知框架，通常只能从统计相关性中形成静态知识的表示，当面对真实世界中连续变化的交互场景时，往往陷入“相关非因果”的认知困境。以自动驾驶为例，现有模型虽能识别道路标志与车辆轮廓，却难以推演突发情况下各要素的联动效应，譬如湿滑路面导致刹车距离增加与后方车辆追尾风险的连锁反应。世界模型在把握因果关系后形成的新型认知架构，不仅能够解析当前环境状态的拓扑结构，更能通过反事实推理模拟不同决策路径的可能结果，即预测一个行动或事件可能引起的后果，展示对世界的动态建模能力。这种能力使得人工智能从“理解当前”迈向“预测未来”，为自动驾驶、机器人和社会仿真等领域提供了全新的解决方案，也为构建自主决策

系统奠定了认知科学基础。

从大模型到世界模型所贯穿的这种人工智能能力提升或跃迁，标志着哲学智能观的新转变，意味着我们不能将智能简化为模式识别，而需要将智能视为能够自主构建因果图并进行反事实推演的系统，从而看到智能的本质是理解与预测的统一，抑或说真正的认知能力应当包含构建动态因果图谱的能力，正如人类面对新场景时会自发建立“如果—那么”的假设验证机制。传统人工智能强调对规则的遵循，而世界模型则通过动态建模实现了对复杂系统的理解和预测。这种转变重新定义了智能的边界，使其从静态的规则遵循迈向动态的适应与创造，使得智能体能够突破监督学习的限制，在强化学习的试错过程中形成类人的因果推理架构。当人工智能系统开始理解物理规律背后的因果关系链，并能通过想象补全未观测变量时，其展现出的创造性解决问题的能力就具备了类似人类的认知特征，从而显示出“真正智能”的特征。

总之，从大模型到世界模型，是人工智能发展的必然追求，如果说大语言模型铺垫了基础，下一个突破应当是构建大世界模型。[①]否则，人工智能就只能停留在虚拟空间中解决人为设定的封闭条件的理想化问题，如下棋、根据提示词写作文等，而不可能面对开放、复杂多变的条件进行感知、推理和决策，从而解决物理世界中的真实问题。基于此，世界模型技术成为人工智能进一步发展的关键，也是走向通用人工智能的具有里程碑意义的技术。它赋予人工智能系统类似人类的认知能力，如环境理解、预测未来状态及自主决策等，所以也被视为“后大模型时代”的标志性技术，意味着人工智能从以语言处理为中心，向更全面、更复杂的智能系统转变，并通过技术路径重新诠释了“何为智能的本质”。

① 李飞飞：《从语言模型迈向世界模型，AI智能的新前沿》，2024-11-28，https://www.sohu.com/a/831060356_121798711。

大模型与世界模型的互补：智能融合与视界融通

从大模型到世界模型，体现了智能模拟水平的提升，也蕴含了智能观的跃迁，表征了我们对智能本质的哲学理解不断深化，尤其是将大模型对智能模拟中所忽视或未及的方面进行了补充。但这并不意味着大模型所体现的认知框架或智能实现方式对于我们的哲学智能观就不再有价值。恰恰相反，其价值和意义仍旧重大，尤其是它具有世界模型所不具有的优势。只有将两者的优势结合起来，实现大模型与世界模型的交汇融合，才能迈向更高水平的人工智能，也才能获得对智能本质的更全面把握。

如果说大模型的问世给人类带来了一场智能革命，而世界模型的构想则预示着又一场新的智能革命，那么更确切地说，新的智能革命不仅发端于从大模型走向世界模型，更成熟于两者的交汇融合。大模型和世界模型不应被视为相互对抗的人工智能技术，而应是相互支撑和促进的技术。例如，大模型与世界模型之间可以进行任务的分解与协作，其中大模型解析意图，像在机器人任务中，大模型可解析人类指令（如“把杯子放在桌子左侧”“将书本放入书架第二层”等），将其转化为世界模型所需的空间目标；世界模型则执行规划，根据物理约束生成可行路径（如避开障碍物、计算抓取力度）。又如，将大模型的生成能力与世界模型的预测能力相结合，既能保持大模型的数据驱动优势，又能注入世界模型的逻辑推演内核，使智能体在复杂环境中实现更高层次的自主决策。还有研究将大模型和世界模型视为人工智能发展中不断走向交叉融合的两个赛道，语言作为一种通用的表示形式，可以成为将大模型和世界模型联结起来的纽带，基于语言的世界模型可以适应更广泛的任务。还可以从世界模型的角度关注大语言模型中的世界知识，这种知识

可分为三部分：一是全球物理世界的知识；二是局部物理世界的知识；三是人类社会的知识。也就是说，有的大语言模型确实获得了关于世界的时空知识，而不仅仅是收集表面统计数据，尽管这种世界知识的质量还有待提高。[①]这一技术的进化轨迹显示：真正的智能革命不在于单一模型的颠覆式创新，而在于不同范式之间的创造性融合。

如果说大模型代表语言智能，世界模型代表具身智能，那么无疑这两种智能都是必要的，偏于一隅的智能是不健全、不全面的。如大语言模型虽然具有非凡的文本生成能力，即语言智能，但在需要物理常识推理的任务中表现显著下降，暴露出脱离具身体验的知识缺陷。反观世界模型研究，虽然可以在特定的场景中显示出较强的行动能力，但其语言理解能力仍局限于有限场景。这警示我们：单一技术路径难以突破智能发展的天花板。近期，斯坦福大学提出的“具身语言模型”（embodied language model）概念，通过将语言模型与物理仿真环境耦合，把现实世界的连续传感器模态直接整合到语言模型中，从而建立单词和感知之间的联系，为实现两种智能模型的结合提供了思路。[②]

在人的身上也可以看到，通过感知—行动具身地与物理世界交互来增长知识或智能，与通过语言交流和学习来增长知识与智能，都是必不可少的。认知科学研究表明，人类智能的演进遵循双模态协同发展的规律。认知科学中的具身认知理论（embodied cognition theory）指出，人类对世界的理解建立于身体与环境的互动基础之上，这种具身性体验构成了概念形成的物质基础；同时，语言符号系统的发展使人类能够突破时空限制，通过文化传承实现知识的指数级积累。神经影像学研究证实，人类大脑在处理语言信息时，会同步激活感觉运动皮层，[③]这种神经机制揭示出语言理解与具身经验的深层耦合。

从以上分析可以看到，更接近人的智能机器，应该是语言和具身两种智能模拟路径的结合，所以要实现达到人类水平的人工智能，仅有大模型或仅有世界模型都是不够的。在这个意义上，全盘否定语言智能和大模型路径也有失偏颇，将生成式人工智能与世界模型完全对立起来是值得商榷的。合理的态度或许是，我们需要了解大语言模型和世界模型各自能做什么和不能做什么，在此基础上，对两种模型取长补短，使二者协同发挥作用。

大模型和世界模型的确各有优势和局限。大模型在语言处理上展现出惊人潜力，具有极强的生成能力，但其对物理世界的具身认知存在根本性缺陷，其知识边界被训练数据框定，无法涵盖人类未经验或未记录的现象（如特定物理场景的具身感知）。世界模型通过强化学习与环境交互构建认知体系，擅长实时感知、因果推理与行动规划，在动态决策层面具备独特优势，却在抽象符号处理方面存在明显短板，在文本处理和语义理解上存在不足，难以解析复杂指令。因为依赖领域特定数据，大模型所获得的知识难以迁移到新场景，泛化能力较弱。通俗地说，两者之间存在认知的层级差异：大模型停留在“知”的层面，是“语言学家”，擅长符号推理但困于文本洞穴；世界模型聚焦“行”的维度，是“行动家”，精于物理交互却短于抽象思考。由此可见，单一的大模型或单一的世界模型只能覆盖智能的局部维度，只有结合二者才可突破“符号—物理”的二元割裂，通过引入多模态认知架构新范式，构建“语言—物理”的联合表征空间，催生“具身–符号”融合的智能系统：其中大模型解决人机交流问题，使机器理解人的要求，然后世界模型按人的要求去解决实际问题，由此合成智能的全过程。这样的智能系统既能通过“眼睛”观察世界，又能通过“语言”反思世界，还能通过“行动”应对物理世界，最终在物理与语义的双重维度中逼近人类智能。

① Jingtao Ding, Yunke Zhang,Yu Shang, et al., “Understanding World or Predicting Future? A Comprehensive Survey of World Models,” 2024-12-01, https://arxiv.org/pdf/2411.14499.

② Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, et al., “An Embodied Multimodal Language Model,” 2023-05-06, https://arxiv.org/abs/2303.03378.

③ Olaf Hauk, Ingrid Johnsrude and Friedemann Pulvermüller, “Somatotopic Representation of Action Words in Human Motor and Premotor Cortex,” *Neuron*, vol.41, no.2, 2004, pp. 301-307.

换句话说，大模型代表的语言智能与世界模型代表的具身智能有机融合可以催生新型认知—行为智能架构，其本质是通过“感知—认知—决策—执行”的闭环链路，赋予机器类人的完整心智模型，由此构成“认知—行为智能系统”，使人工智能既能像人一样思考，也能像人一样行动，从而成功模拟人的全面智能——一种认知与行动融合的智能，即知行合一的智能，既能说又能做的智能。这种融合智能系统展现出三个贯通性特征：其一，在认知维度上实现了从离散符号到连续语义的跃迁，能够像人类一样建立跨模态知识表征；其二，在行为层面突破了传统预设指令的执行限制，可通过自主探索形成目标导向的动作序列；其三，在交互层面构建了动态认知—行动耦合机制，能够在复杂环境中实时调整策略。医学领域的达芬奇手术机器人已验证了这种融合的价值，这种机器人不仅依靠符号系统解析万亿级医学影像数据，更通过具身感知实现微米级器械操控，手术精度超越人类平均水平。[①]在自动驾驶领域，融合智能正在重塑车辆决策系统，该系统既可基于符号规则解析交通法规，又能通过具身学习积累不同气候条件下的驾驶经验，甚至在突发状况中模拟人类司机的直觉反应。这种认知与行为的深度融合，也使得智能系统的发展动力由单驱动跃迁为数据和环境的双驱动，与人的智能发展受亲身实践经验和书本知识的双重推动一样，由此从哲学上触及智能本质的视野整合：真正的智能必须既是符号世界的思考者，又是物理世界的作用者。为此，要摒弃非此即彼的路径之争，构建语言智能与具身智能的协同进化体系，这不仅是技术发展的必然选择，更是对人类智能本质的正确回归。当符号推理的抽象性与具身经验的具象性互渗融合时，人工智能不再局限于特定任务的工具属性，而是逐步展现出类人的整体性智能特征。这种突破也预示着未来人机协同将进入更深层的认知协作，最终实现真正意义上的智能跃迁或实质性突破。

大模型与世界模型所实现的智能融合，还可以通过知识的共享与增强来展现。在这个融合的系统中，大模型可作为世界模型的“知识库”，它从文本中学习的常识（如“冰受热会融化”）可为世界模型提供先验知识，弥补纯物理仿真的不足，辅助物理模型快速适应新环境。而世界模型可以向大模型输入物理反馈，修正大模型的错误假设；对于大模型生成的内容可能违反物理规律的幻觉问题，世界模型可通过物理反馈约束符号生成的任意性，过滤不合理输出，从而降低幻觉风险。

大模型常被誉为“通才”，它擅长语言与知识泛化，但缺乏对物理世界的深度理解。而世界模型主要充当“专才”角色，它聚焦环境建模与行动规划，但依赖领域特定数据。将两者结合，通过知识输入、多模态交互、联合训练等方式，让世界模型成为大模型的补充，提升具身智能的物理可信性。从长期看，若实现大模型与世界模型之间的技术融合，或将推动通用人工智能（AGI）的突破，构建兼具语言理解与物理推理的智能体。鉴于大模型和世界模型在功能和应用场景上存在互补性，它们的融合将会共同推动人工智能向更高的层次迈进。或者说，大模型与世界模型所代表的两种智能的有机融合不仅是技术演进的必然趋势，更是构建更全面、更强大、更可信赖的人工智能系统的关键路径，甚至成为通向通用人工智能的核心路径。

从哲学意义上看，大模型与世界模型的融合还可导向“互补认识论”。人工智能与认识论密切相关，人工智能的不同流派就秉持不同的认识论立场，[②]而大模型与世界模型所贯穿的智能融合，则映射出认识对象、认识来源、认识方式和认识形态的多重互补。其一，认识对象不再是大模型所面对的单一的文本世界，也不是世界模型所指向的单一的物理对象，而是两者的结合，这就极大地扩展了人工智能所面对的世界范围。其二，知识不再仅来源于大模型对数据的

① Su Saqib and Adeel Ahmad Bajwa, “The Role of Da Vinci Xi Robotic Simulation Curriculum in Enhancing Skills in Robotic Colorectal Surgery,” *Annals of Medicine and Surgery*, vol.85, no.12, 2003, pp.6001-6007; Giovanni Gerardo Muscolo and Paolo Fiorini, “A New Cable-Driven Model for Under-Actuated Force–Torque Sensitive Mechanisms,” *Machines*, vol.11, no.6, 2023, p.617.

② 肖峰：《人工智能与认识论的哲学互释：从认知分型到演进逻辑》，《中国社会科学》2000年第6期。

加工，而是整合了世界模型直接与物理对象交互的行动过程，具身的和非具身的活动都能为人工智能提供知识，知识来源的范围因此而最大化。其三，在认知方式上，大模型与世界模型的结合，使人工智能的认知过程可以将基于表征的符号加工与直接感知对象统摄为一体，由此形成更强大和更全面的理解能力。其四，在认知形态上，大模型与世界模型的结合体现了经验与先验的整合。大模型（如 GPT）依赖海量数据的统计规律，通过“观察—归纳”模式学习知识，此时的知识源于数据输入而非先验逻辑，类似于经验主义的知识获取路径。世界模型则倾向“理性主义”。因其中的知识源于物理定律的形式化表达（如能量守恒方程的内嵌），亦即对物理规律、因果关系的建模，试图通过内部推理（如物理引擎模拟）预测未来，知识源于先验逻辑与推理，更接近理性主义的知识形成路径。二者的结合类似康德提出的“先验综合判断”，即大模型提供经验性知识（后验），世界模型嵌入物理规则（先验），共同构成对世界的完整认知。这种融合的认知形态表明了知识既非纯粹先天也非纯粹后天，而是人类与技术共同进化中涌现的交互建构。

综上可见，大模型与世界模型的融合所实现的智能边界扩展，可带来新的智能革命，它引领我们走向智能观的辩证大综合，实现多重意义上的哲学视野融通。

结语

从大模型到世界模型以及两者融合的人工智能发展趋向，除了在哲学智能观上促进认知的跃迁和视界的融通外，也引发许多未尽的哲学问题，如语言与行动谁更代表智能本质？当维特根斯坦说“语言的界限就是世界的界限”时，意味着人所理解的世界就是他所运用的语言，此时如何看待两者之间的复杂关系？或许对人来说使用语言比应对物理世界更复杂，而对 AI 来说理解语言则比理解物理世界更容易，这是否也印证了莫拉维克悖论？又如，大模型与世界模型融合后形成的智能体或 AI 系统还是人的被动的工具吗？它是人的“新型伙伴”“认知生态的参与者”，甚至成为“准主体”进而“新主体”或“双主体”吗？于是，融合的智能体引发了一个根本性的哲学问题：当机器既能“言说”又能“行动”时，人类对智能的垄断是否终结？当机器能够自主构建对世界的解释模型时，人类是否仍是“意义”的唯一立法者？这或将引发继哥白尼革命、达尔文革命之后的又一次认知革命——人工认知革命。如果 AI 与人结为双主体，以后是否还会发展为比人更强大的主体（从强 AI 到超级 AI），从而发生人机之间的主客易位？其后果是否意味着硅基主体强于碳基主体将成为不可避免的趋势？人的技术史定位是否就是充当从碳基智能到硅基智能的“垫脚石”？人类如何规制人机之间这种主客易位的情况？这就引向了人工智能与人类价值对齐的问题。

无论如何，世界模型是值得期待的人工智能发展趋向。从大模型到世界模型，智能革命的这两次浪潮相辅相成、层层递进。在两者的融合中，大模型为世界模型的构建奠定了坚实的技术基础，而世界模型则将大模型的能力推向了一个全新的高度，它们可以共同拓展人工智能的应用边界。可以预见，随着智能革命的不断深入，世界模型将与大模型结合，在更多领域发挥重要作用，当然，其前提是保持与人类的价值对齐。而如何实现这种对齐，则是需要我们持续关注和深入探讨的另一个重要问题。

编辑 张 蕾