大语言模型与因果之梯

Large Language Models and the Ladder of Causation

吴小安 /WU Xiaoan[1,2]　　俞沁元 /YU Qinyuan[2]

（1. 西北工业大学马克思主义学院，陕西西安，710072；2. 西北工业大学陕西省舆情信息研究中心，陕西西安，710199）
(1. School of Marxism, Northwestern Polytechnical University, Xi'an, Shaanxi, 710072;
2. Major Public Information Research Center of Shaanxi Province, Northwestern Polytechnical University, Xi'an, Shaanxi, 710199)

摘　要：朱迪亚·珀尔的因果三阶梯曾是一个被广为接受的人工智能批判及人工智能实践的指导方案。但随着GPT-4等一些现象级的大语言模型出现之后，它的人工智能批评被证伪，“受其连累”，其背后的结构方程模型的因果理论也面临挑战。本文的目标有三，首先，阐述因果三阶梯的核心内容和大语言模型的内在机制，并说明后者在哪种意义上重创了珀尔的原初设想。其次，廓清这种能够“表现出因果能力”的大语言模型的出现对于结构因果模型的理论意义何在，它摧毁珀尔的批评，但又带来了因果研究的新可能。最后，更重要的是，结构因果模型的理论价值犹存，给智能体装备因果推断引擎以帮助其做因果推断的设想并未过时。

关键词：结构因果模型　大语言模型　因果之梯

Abstract: Judea Pearl's three-tiered ladder of causation was once a widely accepted critique of AI and a guiding program for AI practice. However, with the emergence of some phenomenal large language models such as GPT-4, its AI critique has been thoroughly disproved, and the causal theory of Structural Equation Modeling behind it has also been faced with challenges. The goal of this paper is threefold: first, it intends to elaborate on the core elements of the causal triad and the inner mechanisms of the large language Model, and to show the sense in which the latter reinvents Pearl's original conception. Second, it also clarifies the theoretical significance for structural causal models of the emergence of such a large-language model capable of "demonstrating causal competence," which destroys Pearl's critique but opens up new possibilities for causal research. Finally, and more importantly, the theoretical value of structural causal models remains, and the idea of equipping intelligences with causal inference engines to help them make causal inferences is not obsolete.

Key Words: Structural causal models; Large language models; Ladder of causation

中图分类号：O141.4; N031　DOI: 10.15994/j.1000–0763.2025.08.002　CSTR: 32281.14.jdn.2025.08.002

朱迪亚·珀尔（Judea Pearl）是人工智能和统计领域的杰出人物，因其在因果推理方面的开创性工作而广受推崇。身为该行业的奠基者之一，珀尔对人工智能发展的见解自然引人注目，具有举足轻重的影响力。然而，让人略感意外的是，尽管普遍被视为“贝叶斯网络之父”——贝叶斯网络是机器学习的一个重要分支——他却一直坚定且直言不讳地批评机器学

基金项目：国家社会科学基金重大项目“西方语言哲学前沿问题研究”（项目编号：23&ZD240）。
收稿日期：2024年6月5日
作者简介：吴小安（1984–）男，江苏南通人，西北工业大学马克思主义学院副教授、陕西省舆情信息研究中心研究员，研究方向为因果模型、条件句逻辑和死亡哲学。Email: wuxiaoan1984@126.com
俞沁元（2000–）女，内蒙古包头人，西北工业大学马克思主义学院硕士研究生，研究方向为物理主义和因果形而上学。Email: yqyuandyx@163.com

习（尤其是深度学习）的人工智能径路。

在珀尔看来，首先，深度学习主要基于数据中的统计模式，学习的是变量间的相关性。然而相关并不等于因果，虽然这些模型可以基于历史数据预测未来结果，并取得了显著成就，但它们无法解释这些预测背后的“原因”，在理解因果关系是至关重要的这些场景中，比如，医疗保健、经济学和政策制定，这一局限性尤为明显。他通过类比柏拉图的“洞穴比喻”来说明这条径路在根本上就是错误的，它刻画的只是“阴影”，而不是“实在”本身：

> 如同柏拉图那个关于洞穴中的囚徒的著名隐喻，深度学习系统探索的是洞穴壁上的那些阴影，学习的是准确预测阴影的活动。深度学习系统不能理解，它观察到的阴影仅仅是三维物体的空间运动在二维平面上的投影，而强人工智能必须具备这种理解力。（[1]，p.335）

深度学习系统在根本上就是“曲线拟合”，它们只能“看到”并学习数据中的模式（即“洞穴壁上的阴影”），缺乏对这些数据背后更深层次、更复杂现象的理解，即它们可以基于过去的数据进行预测，但无法推理干预、反事实情况或想象新的情景。

其次，基于深度学习的人工智能有泛化和适应性方面的局限。当下深度学习的方法都是“大数据、小任务”范式，如人脸识别，先设计一个损失函数，再对之进行大量的数据训练，得到的模型在解决这个具体问题上的确非常好用，但是却不具备适应性和可解释性。当遇到新情况和新问题的时候，它不能举一反三，得从头来进行数据学习。

上述这些观点并没有多少特别之处，很多人提出过类似观点，珀尔人工智能批判的新意在于，他提出了一个新的框架来解释和容纳现有人工智能的发展现实和理论局限，以及给出了人工智能要真正实现“跃迁”的目标和标准，并把自己关于因果的研究嵌入其中，从而指明了一种新的人工智能发展径路。

受到图灵的启发，珀尔提出了“因果阶梯”（The Ladder of Causation）的框架。他把因果问题分为三个由低到到高且完全不同的层级，分别是相关（看见）、干预（做）和反事实（想象），它们构成了一个完整且自洽的体系。每一个层级对应着不同的认知能力、所能够回答的问题类型和所需要的数学语言以及理论准备（如图1所示）。上一层级拥有下一层级所没有的能力，只有当一个层级的信息满足要求的时候，这个层级的问题才能得到回答。如果模型能够回答高层级的问题，那么它也就能够回答比它低层级的问题。比如，如果一个模型具备处理反事实问题的能力，那么它也能够处理模型相关的干预问题。但是能回答低层级的问题并不意味着能够回答高层级的问题，干预问题不能够从纯粹观测性的数据中得到回答，而反省式反事实问题也不能够从纯粹干预的信息中得到回答。

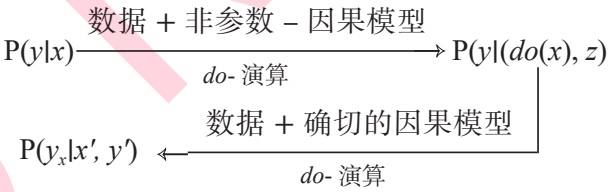

图1 不同的问题对应不同“装备”

珀尔的上述人工智能批判与设想发表于2018年左右。[2]，[3]以当时人工智能的发展现实而言，他的批评给人感觉是非常切中肯綮的，也得到了很多的呼应和支持。[4]但随着OpenAI公司在2023年3月推出具有里程碑意义的人工智能产品：GPT-4，情势在一夜之间发生了转变，作为一种基于大语言模型的人工智能聊天机器人，它本身并没有配备珀尔所预想的“因果推断引擎”，纯粹是基于训练数据库中的文本，却就能够生成和回应因果问题，模拟因果推理的过程，并生成反事实情景的描述，[5]GPT-4轻松跨越了因果的三阶梯，在一定意义上证伪了珀尔关于当前的人工智能程序是在因果关系阶梯的第一层运行，基于相关关系的人工智能径路“不能跨越除因果第一阶梯之外的其他阶梯”的论断。（[1]，p.335）

一、因果三阶梯

接下来将简述珀尔因果三阶梯的人工智能批判，[6]，[7]并力图证明尽管这个批判本身已经被人工智能的发展现实所扬弃，但这并不意味着其背后的数学理论也将寿终正寝，恰恰相反，大语言模型的出现反而解决了结构因果模型的某些“致命”短板，为其在社会科学中的广泛应用提供了新的可能性。

因果三阶梯的第一层级对应的认知能力是看和观察（observation）。它能够侦测到我们的环境中所存在的那些规则性联系，并基于这种观察来做预测。珀尔认为很多动物、当前的学习机器和认知革命之前的早期人类都具有这项能力，所获取的知识是从相关（correlation）或者关联（association）来的。在数学上，这种认知能力对应着纯粹的统计关系，不需要对数据的生成机制有所预设，而是通过“摆弄”裸数据得到想要的结果。比如，早上听到公鸡打鸣，根据你日常的经验，你知道太阳快出来了的可能性很大了。这是典型的概率问题，表达为条件概率P(太阳出来|公鸡打鸣)，公鸡打鸣使得太阳出来的条件概率大大增加。这种能力是建立在消极观察基础上的预测，只能说明相关关系，并不能识别因果关系，尽管公鸡打鸣和太阳升起总是稳定地联系在一起，但两者之间显然没有因果关系。很多时候的确也不需要一个“为什么”的解释，好的预测就足够了，比如，京东通过你浏览和购买商品的数据，分析你的消费习惯和偏好，根据购买记录推荐给你更有可能购买的商品，至于为什么会如此，商家并不关心。

第二层级对应的认知能力是干预（intervention）。这种能力“不仅仅是观察世界，更是改变世界”，预测自己行为将会产生怎样的后果，并选择通过做这样的行为来达到自己的目标。工具的使用，如果是有意为之，而不是偶然或者对祖先的复制模仿，就可以被认为是达到了第二个层级的标志。很少一部分物种被证明具有这种能力的某些要素。

上述问题的回答不能直接从看到的数据中得出来，而数据中极有可能没有对应的先例，就算有也可能不具备参考价值，比如，星巴克考虑明天让所卖的咖啡价格上涨两倍，尽管有之前咖啡价格上涨两倍的相关数据，但导致了之前咖啡价格上涨的原因和今日的情形并不相同，比如，之前的价格上涨是因为咖啡豆的供应短缺，周边只此一家咖啡店，供不应求，但现在是在不考虑市场因素的情况下，判定如果坐地涨价会如何。

在决策上述咖啡涨价的情形时，可以通过随机对照实验来帮助做决策。但是很多时候，因为各种现实的和不现实的原因，随机对照实验不可行，于是退而求其次，试图从观察的数据来推导想要的结果。可以通过构造因果模型和数学的演算得到想要的结果，用珀尔的数学言语，就是求解P(销售情况|*do*(咖啡价格上涨两倍))。

但会使用工具和拥有关于工具的一个“理论”是两码事。拥有后一种知识意味着你能够明白为什么这个工具有效，如果没有效又该怎么做。所以这就进展到了认知能力的第三层级：想象。它和反事实推断密切相关。人类所有的发明，创造，进步和更新都和这样一种想象力有关：

> “虚构”这件事的重点不只在于让人类能够拥有想象，更重要的是可以“一起”想象，编织出种种共同的虚构故事，不管是《圣经》的《创世记》、澳大利亚原住民的“梦世记”，甚至连现代所谓的国家其实也是种想象。这样的虚构故事赋予智人前所未有的能力，让我们得以集结大批人力、灵活合作。（[8]，p.26）

第三层级是一种反省式（retrospective）推理：设使所做的事情和实际所做的事情不一样，那会怎么样呢？但是生命在流淌，“逝者如斯”，没有回头路可以改变过去，干预问题还可以做随机化对照实验来发现因果影响，但是对已经发生的事情，不存在一个实验，可以让我们不做那些已经做了的事情，再看结果会怎么样。面对这样的问题，需要“更复杂的装备”。

正如知道牛顿运动方程：$F=ma$，就能够回答很多反事实问题，比如，设使相较于现在的拉力增加两倍，物体的速度将会怎么样？所

以，只有当有对现实情形的可靠因果模型时，“这种模型有时被称为‘理论’，甚至（在构建者非常自信的情况之下）可以被称为自然律”，那么才能够回答反事实的问题：“如果不这么做会怎么样？”。数学语言对上述问题的表达就是：$P(y_x|x', y')$，这个式子中x', y'是对现实已经发生了的情形的表达，y_x则表示的是如果我们做与现实的行为x'不同的行为x，那么结果y将是什么。

珀尔的因果三阶梯框架有诸多意义。首先，批判当下人工智能径路的同时也为如何发展指引了方向。在珀尔看来统计上的回归，以及风行当下的深度学习技术都还停留在相关这个认知水平，尽管基于它们的自动驾驶技术、语音识别技术取得了重大进展，但这些“成就”并不意味着人工智能技术取得了质的突破。这些径路还停留在因果之梯的第一层级，运作的系统缺乏灵活性（flexibility）和适应性（adaptability）。尽管对深度学习的人工智能径路的发展轨迹持批评态度，珀尔并没有完全否定其价值。相反，他倡导一种混合方法，将机器学习的模式识别能力与能够推理行动及其后果的因果模型结合起来，旨在拓展机器理解和能力的界限，实现不仅能预测，还能推理、适应和解释的人工智能。

其次，重新审视几种因果的哲学理论的前景。因果的概率理论，它是二十世纪下半叶主要的因果理论之一，[9]，[10]哲学家们为此付出了四十年的努力，提出了各种方案，[11]但结果都不令人满意。还包括因果的规则性理论，它把因果关系理解为前后相继的规则性。从因果三阶梯的视角，这些理论努力是完全无望的。从因果三阶梯的视角，这些努力是完全无望的。概率（或者规则性）位于因果阶梯的第一层，是不能够刻画第三层或者第二层的因果概念的。

再次，把因果问题划分为不同的层级的好处在于，它回避了已经讨论了数千年的关于“因果是什么？”的问题，而是去讨论具体可行的问题：“什么是一个因果推理者能够做的？”所以不同于哲学家们要给因果一个完备的定义，珀尔对于因果的定义非常“随性”，纯粹一种“工作定义”。（[12]，p.5）正如牛顿并不努力来定义“力的类别或物理属性，而只想研究这些力的量与数学关系”，珀尔也力图数学化因果关系，并以此来讨论如何来解决其中一类问题。出乎意料的是，这样反而深化了对于因果的理解。

二、ChatGPT的工作原理

研究表明，大语言模型在不同类型的因果任务上（如因果发现、反事实推理和实际因果）都有非常好的表现，[5]轻易地跨越了珀尔所设定的因果阶梯。然而，大语言模型并非像珀尔所预设的那样，是利用因果概念、引入因果模型和装备因果推断引擎来完成这些因果任务。基于深度学习技术开发的ChatGPT，是通过识别和利用数据之间的相关关系来处理因果任务。[13]接下来，将对ChatGPT的生成机制和训练方式进行论述，以展现它的工作原理如何不同于珀尔的设想。

首先是ChatGPT的生成机制。当用户输入指令（prompt）以后，这些指令经过嵌入（embedding）、编码（encode）和解码（decode），才能转化为它所要求的内容。这是由ChatGPT所采用的变换器（transformer）架构决定的。尽管ChatGPT的表现宛如真人，但它本质上仍然是一种数学模型，并不能直接识别人类使用的自然语言。因此，面对自然语言编写的指令，它首先要将其转化为自身可识别的形式，这个过程便是“嵌入”。

在变换器架构下，嵌入分为三步：第一步，将自然语言文本拆分为最小的语言单位，即词元（token）；第二步，将文本中的词元映射为词向量，并将每个词元在文中的位置映射为位置向量，前一个过程称作“词嵌入”，后一个过程称作“位置嵌入”；第三步，将对应的词向量和位置向量相加得到嵌入向量，再将所有的嵌入向量组合得到一个向量矩阵。其中，位置向量映射的是词元在文本中的先后顺序，词元和词向量的映射关系则要更为复杂：词向量在一定

程度上可以反映词元的语义关系，语义越相似的词元，映射得到的词向量在向量空间的位置越接近。

因此，通过嵌入得到的向量矩阵，包含了文本中每个词元的语义关系和位置关系。表1即为嵌入的示例，当我们要求ChatGPT翻译“我爱你。”，这段文本首先会被拆分为四个单独的词元，而后这四个词元又会被映射为四组不同的词向量、位置向量和嵌入向量，最终四个嵌入向量组合得到一个向量矩阵。

如果说嵌入把握的是单个词元的语义关系，那么编码把握的则是整个文本的语义关系。简单来说，编码是指对嵌入得到的向量矩阵作进一步调整。由于作为整体的文本具有更加复杂的语义关系，编码不止一次，而且每次编码都将根据文本的不同特征进行调整。因此，编码得到的向量矩阵能够更好地反映文本的语义关系。作为文本生成的最后一个阶段，解码具有不同于嵌入和编码的形式：它是指根据编码得到的向量矩阵，预测并生成新的文本。经过词嵌入，每个词元都有一个对应的词向量，这些词向量共同构成ChatGPT的词表。当ChatGPT进行解码时，它将根据编码得到的向量矩阵，为词表中所有的词向量赋予生成概率，挑选出概率最高的词向量，生成它对应的词元。此外，ChatGPT具有特殊的掩膜（mask）机制，它将屏蔽向量矩阵中和生成词元无关的部分，以提高解码的精确性。同时，不同的自然语言处理任务有着不同的掩膜要求：表2为翻译解码的示例，当ChatGPT翻译文本时，将从文本的第一个词元开始翻译，并在翻译每个词元时，屏蔽这个词元之后的全部信息。表3为对话解码的示例，当ChatGPT进行对话时，由于它要对整段文本作出回应，故它不需要屏蔽任何信息。

接着是ChatGPT的训练方式。在嵌入、编码乃至解码阶段，ChatGPT都是根据特定的权重，对指令文本、向量矩阵进行相应的处理，这些权重通常被称作“参数”。作为深度

表1 嵌入的示例

自然语言文本：我爱你。			
	词向量	位置向量	嵌入向量
我	(1, 1, 1, 1)	(1, 0, 0, 0)	(2, 1, 1, 1)
爱	(2, 2, 2, 2)	(0, 1, 0, 0)	(2, 3, 2, 2)
你	(3, 3, 3, 3)	(0, 0, 1, 0)	(3, 3, 4, 3)
。	(4, 4, 4, 4)	(0, 0, 0, 1)	(4, 4, 4, 5)
向量矩阵：(2, 1, 1, 1 2, 3, 2, 2 3, 3, 3, 4 4, 4, 4, 5)			

表2 翻译解码的示例

自然语言文本：我爱你				
向量矩阵：(9, 9, 9)				
	掩膜矩阵	预测结果		
我	(9, 0, 0)	I: 60%	Me: 25%	Myself: 15%
我爱	(9, 9, 0)	Love: 55%	Like: 30%	Prefer: 15%
我爱你	(9, 9, 9)	You: 70%	Youself: 30%	
生成文本：I love you				

表3 对话解码的示例

自然语言文本：你好				
向量矩阵：(1, 2)				
	掩膜矩阵	预测结果		
你好	(1, 2, 0)	你：50%	您：30%	我：20%
你好	你	(1, 2, 1, 0)	你好：60%	们：30%	的：10%
生成文本：你好				

神经网络模型，ChatGPT通过这些参数模拟人类大脑的神经突触。当人类学习到新的知识，脑中的神经突触便会发生相应的变化。而当ChatGPT进行训练时，它的参数也会随之调整，以记录训练成果，这个过程被称作“拟合”。因此，使自身参数拟合用户的实际需求，生成令人满意的文本，是ChatGPT的训练目的。

具体而言，ChatGPT的训练分为预训练（pre-training）和微调（fine-tuning）两个阶段。在预训练阶段，OpenAI收集了大量文本数据，以供ChatGPT训练。由于训练数据集过大，为更好地拟合数据集的各项特征，OpenAI又为ChatGPT设置了大量参数。在此情形下，OpenAI不可能单纯依靠人力对ChatGPT的参数进行调整。因此，在预训练阶段，ChatGPT主要采用无监督学习的方式进行训练。无监督学习是指人工智能从没有人工标注的数据集中，挖掘数据之间的相关关系。比如，词向量之所以能够反映词元的语义关系，是因为语义相似的词元具有相似的用法，ChatGPT将这些用法相近的词元归为一类，间接捕捉到词元之间的语义关系。训练数据集包含许多种类的文本，如文学作品、学术论文和网络论坛记录，通过无监督学习，ChatGPT的参数很好地拟合了这些文本的特征，同时，从庞大的训练数据集中，ChatGPT习得了大量的知识：比如，它可以立即告诉你南苏丹的首都在哪里。

在微调阶段，ChatGPT主要通过强化学习和监督学习的方式进行训练，这两种训练方式都有人类的参与。ChatGPT的强化学习是指基于人工反馈的强化学习，由专业人员为ChatGPT根据指令生成的文本进行评分，ChatGPT再根据评分结果进一步学习数据间的相关性。监督学习则是指人工智能从有人工标注的数据集中，学习数据之间的相关关系。通过监督学习，OpenAI能够影响ChatGPT生成文本的倾向。比如，为避免ChatGPT生成种族歧视、性别歧视等有害信息，OpenAI聘请了专门的数据标注团队，制作了一个打上负面标签的有害信息数据集，以供ChatGPT进行监督学习。在预训练的基础上对ChatGPT进行微调，能够进一步强化其功能。ChatGPT表现优异，正是得益于OpenAI的微调。

ChatGPT采用的变换器架构，决定了它的生成机制，而这又进一步决定了它的训练方式。但不论是预训练阶段的无监督学习，还是微调阶段的强化学习和监督学习，它们学习的都是文本之间的相关关系，也就是因果阶梯第一层级的知识。按照珀尔对因果阶梯的设想，相关、干预和反事实是三个由低到高的层级，高层级的因果能力可以回答低层级的因果问题，而低层级的因果能力却无法解决高层级的因果问题。既然ChatGPT能够处理第二、第三层级的因果任务，那么它本应采用珀尔的结构因果模型或其他包含反事实知识的模型，但事实却并非如此：它单凭第一层级的因果知识，就回答了第二、第三层级的因果问题！

三、结构因果模型再审视

珀尔的人工智能批判之所以会被“打脸”和他的工作所预设的“鹄的”（珀尔的理论的目标和靶子）密切相关。他所预设的对手是发现和描述数据相关性的经典统计学方法，在后者看来，数据是其核心，包含着现实的一切智慧和秘密，统计分析的主要工作就是利用这些数据来寻找变量之间的相关性、测试假设或者建立预测模型。珀尔明确反对这种极端经验论（Radical empiricism），[14]反对在统计学和机器学习文化中占主导地位的数据中心思维，认为仅依赖于数据可能忽略了现实世界中复杂的因果关系和背景知识，即忽视了理论和模型的重要性。

但珀尔没有意识到的是，不只有基于数据的因果推理，还有另外一种基于知识的因果推理（knowledge-based causal reasoning）。大语言模型更多处理的不是单纯的“数据”，是文本和元数据（Metadata），即关于文本数据的信息，如文本的作者、生成时间、文体类型、内容概述等，理解和生成的是关于元数据的描述性内容。而这些元数据本身是包含着因果信息的，也正是因为训练的文本符合因果、符合逻

辑，所以其识别出的语言模式与因果拟合，从而实现了“弯道超车”，表现出“因果能力”。

最新的研究表明，[5]大语言模型，比如GPT-4，在多个因果基准测试上已经达到了新的最高精度标准。在成对的因果发现任务、反事实推理任务以及实际因果关系（在一个设定的情境中确定必要和充分因果的精度）上均超越了现有算法，分别达到了97%、92%和86%，相较于之前，已然有了大幅度的增进，以至于珀尔本人都极为感慨。在2023年9月的一次采访中，他承认之前关于因果三阶梯的设想是错误的，甚至对人工智能对于人类未来的影响表达了一种极为悲观的态度①。

但这种能够“表现出因果能力”的大语言模型的出现对于结构因果模型的理论意味着什么？毕竟结构因果模型的一个重要目标就是“挽深度学习的人工智能径路之弊，并助力实现未来之通用人工智能”，但如今他所指出的弊端似乎都已经被克服，而自己的人工智能构想还很大程度停留在理论上。

作为一种因果分析的形式化框架和数学理论，结构因果模型通过明确地建模因果关系来帮助研究者、决策者和实践者更好地理解和预测事件之间的关系。但结构因果模型的发展和应用有其瓶颈，首先，图构建的复杂性、数据局限性、因果关系的复杂性、计算挑战等等。比如，构建因果图可能是因果分析中最具挑战性的部分，也是整个结构因果模型理论展开的核心所在，所以当要做具体的因果分析时，首要的事情就是“白手起家”构建出一个因果图。而所有珀尔关于因果数学工作的构建与展开都是建基于如下预设：所使用的因果图是对“实在”的准确表征。这个看似无害的预设实质上是一个极具挑战性的工作。

第一，现有的因果发现算法是基于各种因果原则（因果马尔可夫条件，忠实性条件，极小性条件等等），本质上是根据变量之间条件独立关系来确定因果图（自动因果发现平台Tetrad上有几十种因果发现算法），因果图的最终确定可能还要再辅以各种因果原则、图构建者的理论直觉和前见知识。当前的因果发现算法尽管数量很多，但它的效果高度依赖于数据的质量、完整性和代表性。在大多数的社会科学实践中，要最终得出因果图需要高质量和全面的数据，这个要求都实在是过于“苛刻”，社会科学所处理的观察数据总是有限的、有偏见的或者低质量的，因为数据的收集永远存在着不完美的可能。而且根据现有的因果发现算法，根据给定的观测数据集通常并不能唯一确定一个因果图，很多时候存在多个与数据相拟合的图（等价类）。[15]

第二，大多数从业者在开始因果分析时会根据他们的领域知识手动构建一个因果图。而这也意味着建构一个具体领域的因果图很大程度上依赖于专家的领域知识（domain knowledge）和专业直觉。（[1]，p.203）本质上，这是一种基于元数据（metadata-based）的推理（所谓元数据就是与变量相关的数据，而不只是关于变量的数据值）。一个更为直接的认识论问题是：如何来确定一个因果模型是正确的，毕竟它在很大程度上是基于人的主观性（变量的选择是有人来最终确定的）而建构出来的！

第三，存在不可观测的变量（不存在关于这个变量的数据），但它对计算出因果效应是不可或缺的。（[1]，p.136）但是既然没有数据，那自然不可能从数据中推断出包含这个变量于其中的因果图，这就加剧了对于专家知识和个人认知的要求，认识到这个变量的重要性，并把它包括在图中。另外与受控实验环境相比，现实的社会系统通常更加复杂和动态，大量相互作用的变量使得难以隔离和识别清晰的因果路径。

第四，在实际因果的分析中，要考虑具体的情境，考虑特定的社会和文化规范等因素，要根据所考虑的问题决定把哪些因素纳入进来，而又有哪些因素暂时可以不予考虑。但鉴

①当前关于大语言模型因果推断能力的判定是基于因果基准测试得出的，本质上就是设想了很多因果情境和问题，通过问答的形式，根据回答的准确率来判定其因果能力，在具体的应用在还没有实质和显著的体现，尽管应用的前景可期。

于个体认知能力和经验局限，并不存在一个统一的因果直觉或者判断，形式化的实际因果理论很大一部分工作是把这些背景知识中的因素嵌入到理论中去，[16]哲学家们为此大费周章，建立了异常复杂的因果和因果贡献度的数学定义，这个定义同样以变量值为基础。这种形式化的方法也有其表征的局限和界限，有些背景因素（或者人的因素）很难被形式化到模型中去（比如，一个人的表情状态对于另一个人行为的影响，如果它在具体的分析中是重要的，又要用怎样的函数关系来刻画？）。（[1],p.116）

在很长一段时间，整个因果的形式化分析，一直停留在一些简单明晰的例子（或者玩具例子）①，[15]，[16]很少去分析和挑战一些复杂的情景，除了因为哪怕是在这些简单的情境中，哲学家们也没有达成共识，另一个重要的原因也许是，当真正要拿他来分析一些稍微复杂例子的时候，将引入更多的变量，更多的因果假定，“疾风知劲草”，个体视角的偏狭，认知能力的局限、领域专业知识缺乏的弊端将会更被凸显出来，构建出一个“正确的”因果图就变得异常困难。

而结构因果模型的这些发展瓶颈，在大语言模型的助力之下，有了“破局”的可能。首先，可以借助大语言模型来生成候选因果图，或者利用大语言模型来批判人类自己建构的因果图，发现研究中的漏洞，或者规划一个鲁棒性检测。已有的研究表明，在各种不同领域中的因果发现任务中，大语言模型的性能都超过了现有的因果发现算法。[5]

其次，大语言模型直接在自然语言中处理这些背景变量，而且基于训练文本中统计规则性来做出的判断和分析的结果可能更“拟合”哲学家念兹在兹的“大众因果直觉”。借助大语言模型，我们可以跨越专业知识的鸿沟，拓展分析的思路，从而推动研究的进程。尽管目前并不知道ChatGPT的训练数据集有多大，但是从GPT-3所公布的1750亿的参数数量就可以想见其训练数据的体量一定非常惊人，相比较而言，人就算终生保持阅读的习惯，其一生的文字“摄入量”也不会超过200MB。这意味着人的认知必然是非常“局限的”，个体的能力和偏好更是加剧了这种“偏狭”，但是ChatGPT是“不挑食的”,它训练一切可以数据化的文本，它的视野和认知远超任何个体或者群体，从而在进行特定情形的因果分析时，也会表现地更全面、更系统。

最后，大语言模型的强大算力会成为结构因果模型发展的有力支撑。大语言模型可以通过处理大量的文本数据、实验数据或者观测数据，自动挖掘和发现潜在的因果关系；确定图本身的假设，以及它与数据的拟合度，最终证成或者证伪因果图；可以进行更高效的因果模型参数优化，特别是在复杂的因果网络中，模型的参数空间可能非常大，传统方法难以有效探索。而大语言模型能够通过其计算能力和优化算法，更快地找到最优参数设置，从而提升因果推断的效果。总之，其强大算力不仅为因果推断提供了更高的计算效率，还拓展了结构因果模型的应用范围。

尽管大语言模型在因果分析方面的出色表现有力地帮助我们克服了一些难以跨越的障碍，但并不意味着它解决了所有的因果问题。作为一种基于知识的因果推理，它更多还是一种基于统计规则性的模式识别，当面对一些具体的因果问题（很多定性的因果问题超出了直觉能够回答或者回答可信的范畴），为了做出正确的判断或者对其有信心的判断，还是得构建模型、收集数据，利用形式化的结构因果模型来帮助我们进行因果推断并给出具体的决策。至少根据当下大语言模型的表现，结构因果模型的数学化工作不但没有失去意义，反而在这个新工具的加持下，重新有了“更广泛”应用的可能。

所以大语言模型和结构因果模型实质上是互补的。就目前的观察而言，至少在如下几个

①珀尔和哈尔彭所讨论的都是简单的“玩具问题”（toy questions），但真要分析一些稍微复杂的例子，其中的复杂程度又足以让人望而却步，比如珀尔所讨论的判定薰蒸剂对于燕麦作物产量因果影响的例子。

方面有互补的可能性。首先，大语言模型可以帮助解释和预处理复杂的文本数据，为因果模型的建立提供更清晰、结构化的数据输入。其次，大语言模型可以帮助生成因果关系的假设，以及辅助检查因果模型的假设和结构，提供修改建议，以优化模型的准确性。再次，可以辅助进行具体案例分析，为特定情境下的因果关系提供详细解释。最后，大语言模型的强大文本理解能力可以协助整合来自不同数据源的信息，为因果模型提供更丰富的背景信息。

最后，大语言模型可以“表现”出因果能力和它真正“具有”因果能力是两码事。它之所以能识别文本中的因果叙述，能够因果地回答问题与生成符合因果逻辑的文本，主要是基于对语言模式的识别，它的“理解”是对大量训练数据中因果语言模式的一种识别，本质上还是曲线拟合，是“文字接龙”，并不包含对于因果机制的真实洞见；同样地，尽管它能生成和描述反事实情景，但这种能力同样是基于语言生成技术，而非真正的逻辑或者因果能力。正如塞尔在“中文屋”论证所指出的，完全可以设想一个完全不懂中文的美国人也能够回答中文问题的特殊场景，[17] 今天的大语言模型也是一样，尽管它并不具备因果理解的机制，也不能真正建构因果图或者进行干预演算，但它凭借其对大规模语料中语言模式和统计关联的高度捕捉能力，能够间接模拟出因果推断所需的输入输出关系，从而完成因果相关任务。

所以难怪会有人主张，大语言模型其实和“鹦鹉一样，只是复述数据中嵌入的因果知识”。[18] 这个意义上，有两点值得强调。第一，这意味着珀尔的人工智能批判虽然“失效”，但鉴于大语言模型并不真正理解因果，那么他所设想的人工智能的因果径路就并没有被推翻，机器的确还没有具备“真正的”因果能力，通过给智能体装备“因果推断引擎”以帮助其实现因果推断的径路，依然有其理论和实践的价值。第二，大语言模型所体现的因果处理能力，虽然在输出结果上能够模拟某些因果推理的效果，但是否具备真正理解因果结构的能力仍不确定。面对真实世界中的干预、反事实和复杂因果结构，我们可能仍需诉诸结构化的因果模型。另一方面，也不排除一种更激进的可能性，即因果推断本质上就是一种高度抽象的模式识别，其特殊性未必超出统计学习本身。具体答案为何，需要时间和实践的检测。但至少目前的研究表明，大语言模型在因果推断方面还存在着许多局限和挑战。将结构因果模型与大语言模型集成（collaboration），已成为解决纯数据驱动型人工智能的一些基本限制的一个有前途的方向。

结　　语

马克思和恩格斯在《共产党宣言》里感叹：“一切固定的古老的关系以及与之相适应的素被尊崇的观念和见解都被消除了，一切新形成的关系等不到固定下来就陈旧了。”随着计算机科学的发展，我们对于“智能”的认识和定义也经历着这样一个还来不及固定下来就陈旧的过程。在电子计算机初露头角的年代，人类的“快速精确的心算能力”曾被广泛视为智能的象征。然而，随着时间的推移，这种能力渐渐被看作是“机械性的”行为。八十年代，人工智能领域努力研发能下国际象棋的计算机。到了1996年，当IBM的“深蓝”（Deep Blue）击败了世界象棋冠军加里·卡斯帕罗夫（Garry Kasparov）后，下象棋的能力也开始被视为一种机械行为，不再被认为是人类智能的独特证据。接着，人们将智能的标准转向了围棋——一种更为复杂、需要巨大计算力的游戏。然而，当2016年运用深度学习技术的AlphaGo击败了世界顶尖围棋选手后，下棋的能力同样被贬为“仅仅是曲线拟合”，不再被看作真正的智能。甚至在不久前，人们夸赞婴儿能轻松完成的脸部识别任务，但对计算机而言似乎是不可能的壮举，毕竟按照当时人们的理解，要教会计算机识别千差万别的人脸，实在是一个庞大的工程！然而，转瞬间人脸识别技术取得了飞速发展，即便在佩戴口罩也能准确识别。2018年，当珀尔提出因果三阶梯的时候，在很多人看来（同样在笔者看来），他对于人工智能的批判卓

有洞见并随同附和，但技术的迅捷发展让他的理解和论断“俯仰之间，已为陈迹”。

至少当下的研究表明，大语言模型的发展并不意味着敲响了结构因果模型“灭亡的丧钟”，结构因果模型作为一种理解和建立复杂关系的方法，大语言模型作为一个先进的自然语言处理工具，两者有互补之处，能够相互结合提供更深入的数据分析和解释，使因果推断过程更加高效、准确和用户友好。甚至对于人工智能自身的发展而言，正如很多人所相信并为之努力的那样，结构因果模型可以提升了机器学习模型的可解释性，使得人工智能的决策过程更加透明和可靠。但是，凡此种种还只是在起步阶段，在这个阶段，中国学者做出了很多努力和贡献。让人不禁想起图灵在《计算机与智能》的最后所说的那句话：初见前路近可至，细思百事竞待忙。

［参 考 文 献］

[1] 朱迪亚·珀尔、达纳·麦肯齐. 为什么：关于因果关系的新科学 [M]. 江生、于华 译, 北京：中信出版社, 2019.

[2] Pearl, J. 'The Seven Tools of Causal Inference, with Reflections on Machine Learning'[J]. *Communications of the ACM*, 2019, 62(3): 54–60.

[3] Pearl, J., Mackenzie, D. *The Book of Why: The New Science of Cause and Effect*[M]. London: Basic Books, 2018.

[4] Marcus, G., Davis, E. *Rebooting AI: Building Artificial Intelligence We Can Trust*[M]. Pimlico: Vintage, 2019.

[5] Kiciman, E., Ness, R., Sharma, A., et al. 'Causal Reasoning and Large Language Models: Opening a New Frontier for Causality'[J]. *Transactions on Machine Learning Research*, 2024, 1–57.

[6] 任晓明. 因果推理的跨学科跨文化探索 [J]. 社会科学家, 2022,（8）: 9–16.

[7] 梅剑华. 人工智能与因果推断——兼论奇点问题 [J]. 哲学研究, 2019,（6）: 86–95.

[8] 赫拉利. 人类简史：从动物到上帝 [M]. 林俊宏 译, 北京：中信出版社, 2014.

[9] Reichenbach, H. *The Direction of Time*[M]. California: University of California Press, 1991.

[10] Suppes, P. *A Probabilistic Theory of Causality*[M]. Amsterdam: North-Holland Publishing Company, 1970.

[11] Eells, E. *Probabilistic Causality*[M]. New York: Cambridge University Press, 1991.

[12] Pearl, J., Glymour, M., Jewell, N. P., et al. *Causal Inference in Statistics: A Primer*[M]. New York: John Wiley & Sons, 2016.

[13] 尤洋、郭宇. ChatGPT 与因果性 [J]. 科学学研究, 2023, 41（12）: 2122–2130.

[14] Pearl, J. 'Radical Empiricism and Machine Learning Research'[J] *Journal of Causal Inference*, 2021, 9(1): 78–82.

[15] Pearl, J. *Causality: Models, Reasoning, and Inference*[M]. Second Edition, New York: Cambridge University Press, 2009.

[16] Halpern, J. Y. *Actual Causality*[M]. Cambridge, MA: MIT Press, 2016.

[17] Searle, J. R. 'Minds, Brains, and Programs'[J]. *Behavioral and Brain Sciences*, 1980, 3(3): 417–424.

[18] Zečević, M., Willig, M., Dhami, D. S., et al. 'Causal Parrots: Large Language Models May Talk Causality But Are Not Causal'[J]. *Transactions on Machine Learning Research*, 2023, 1–27

［责任编辑　王巍　谭笑］