

## 对抗扰动与差异传递

——浅析大语言模型因果推理与因果性理论的互动关系

The Adversarial Perturbation and Transmission of Differences: A Brief Analysis of the Interaction Between the Causal Reasoning in Large Language Models and the Causal Theory

彭新波 / PENG Xinbo

(中共厦门市委党校哲学与文化教研部, 福建厦门, 361000)

(Philosophy and Culture Department, The Party School of CPC of Xiamen Municipal Committee, Xiamen, Fujian, 361000)

**摘要:** 因果性理论驱动大语言模型因果推理技术的发展, 更多的不是直接体现在对因果关系本质特征的建模上, 而是在评估和提升大语言模型因果推理能力的扰动技术和对抗性学习中。通过扰动训练数据中的非因果相关变量, 可以评估和提升模型识别正确因果关系的能力, 避免虚假相关, 通过扰动训练数据中的因果相关变量, 可以评估和提升模型的反事实推理能力。扰动是一种制造差异, 并追踪差异传递的过程, 这深化了对因果性理论的理解, 同时, 也跳出大语言模型本身, 通过动态的外部介入, 将干预、反事实等开放性、操作性的因果关系本质特征纳入到大语言模型与人类因果推理、因果认知的互动关系中来。

**关键词:** 因果相关变量 扰动技术 常态 传递关系

**Abstract:** The development of causal reasoning in large language models driven by the causal theory is reflected less in modeling directly the essential characteristics of causal relationships, and more in perturbation techniques and adversarial learning used to evaluate and enhance the causal reasoning capabilities of such models. By perturbing the non-causally correlated variables in the training data, the capability of the models to identify correct causal relationships can be assessed and improved, helping to avoid spurious correlations whereas their counterfactual reasoning capabilities evaluated and enhanced by perturbing the causally correlated variables. Perturbation is a process of making differences and tracing their transmission, and it not only deepens the understanding of the causal theory, but at the same time, it also moves beyond the large language model itself, incorporating open and operational essential features of causality, such as intervention and counterfactuals, into the interaction between large language models and human causal reasoning and cognition through dynamic external interventions.

**Key Words:** Causally correlated variables; Perturbation techniques; Normality; Transitivity relationship

中图分类号: TP18; B025.5 DOI: 10.15994/j.1000-0763.2026.07.007 CSTR: 32281.14.jdn.2026.07.007

### 一、大语言模型因果推理存在的问题

近年来, 人工智能迅猛发展, 大语言模型

(Large Language Models, LLMs) 更表现出惊人的能力, 正从生成式AI向推理型AI发展, 这集中体现在其因果推理能力上。2023年3月发布的GPT-4, 被研究者评价称, 在多个因果基准测试

**基金项目:** 国家社会科学基金青年项目“人工智能因果推理对生产组织形式的变革及其机制研究”(项目编号: 22CZX027)。

**收稿日期:** 2025年8月22日

**作者简介:** 彭新波(1988-)男, 湖南湘潭人, 中共厦门市委党校哲学与文化教研部副教授, 研究方向为科学哲学。Email: pengxb@xmdx.gov.cn

任务上表现出更高的准确性。<sup>[1]</sup>2025年1月上线的DeepSeek融入了旨在加强逻辑和因果理解的强化学习方法,并以思维链的形式阐述中间推理步骤,使研究人员能够准确识别模型与合理因果逻辑的分歧所在。<sup>[2]</sup>

大语言模型的革命性突破在于其变换器架构,如GPT的全称就是生成式预训练变换器(Generative Pre-trained Transformer),“变换器”实现了数学语言与自然语言之间的转换,其生成能力是基于对自然语言的概率采样(数学语言),从词频、位置编码和共现关系组成的庞大组合可能性中创造出新颖的输出,参数(权重)决定不同部分输入的概率性输出,从预训练数据的潜在语言结构中统计推导出新的表达方式。因此,大语言模型并非对自然语言的简单抄袭复制,它实现了对同一事物无限方式的表达,本质上是一种翻译的通用技术(translation generalized)。<sup>[3]</sup>

可以说,通过变换器对自然语言进行概率拟合遵从的是“人类的心智、认知是计算”的理念,主张心灵与机器都是广义的输入输出设备,信息处理和学习的的基本机制在其中起作用。<sup>[4]</sup>这也意味着,大语言模型仍然是基于概率形式的相关关系来进行生成、预测乃至推理的,基于此形成的因果推理能力还不能完全模拟出人类回答“为什么”问题的水平,其根本原因在于:

第一,在输入方面,预训练数据是静态的,基于此生成的输出只是内部语言结构的重复,缺乏与外部经验的互动,从而缺乏人类因果推理的能动性。如果认为,以数据和符号构成的网络中的基本框架和固定结构为基础,按照规模规律(scaling law),当参数达到一定规模时,有望实现因果涌现,<sup>[5]</sup>那么这种所谓的“因果涌现”仍然是缺乏与外部环境互动的,没有对客观世界能动的表征,几乎可以断言,如此“涌现”的因果关系将与外部现实世界不相符。

第二,在输出方面,仅仅是相关关系的输出,因此缺乏人类因果推理的复杂性,只是传

达看似复杂的信息,而没有人类的主观因果认知,无法理解因果推理所获得的因果关系的意义所在,从而在因果相关变量识别,因果关系的时序、方向确定,以及因果解释等方面,表现较差。

第三,在输出输入的关系方面,如GPT-4所采用的方法,跳过朱迪安·珀尔(Judea Pearl)的“因果层次定理”(CHT),([6], p.29)即不对干预和反事实建模,而只是在珀尔所认为的不可能直接实现因果推理的观察层面,处理包含因果信息的数据,通过拟合自然语言中的因果语言来识别其特定模式,即凭借其对大规模语料库中语言模式和统计关联的高度捕捉能力,间接模拟出因果推理所需的输出输入关系,<sup>[7]</sup>这似乎是一条捷径。这条捷径让因果假设本身在观测数据中得到体现,直接以相关关系的形式重现人类的因果语言模式和因果推理结构,在基准测试中,也获得了良好的结果,但基于上述同样的原因,一是预训练数据是静态的,缺乏与外部世界的互动,二是将因果关系简化为相关关系,这种因果语言和因果推理仍然是“因果鹦鹉”。<sup>[8]</sup>

如果不以因果性理论为根据,忽视人类因果语言和因果推理的现实性、操作性,大语言模型进行因果推理将遭遇问题<sup>①</sup>,如虚假相关性、泛化性差和可解释性差等,而遭遇扰动时的模型表现下降是这些问题的集中体现。那么模型“在对人类而言似乎微不足道的场景中也表现出令人惊讶的失败,尤其是在面对输入数据的轻微扰动时。”([9], p.10)研究表明:“扰动使我们能够评估模型是真正理解了底层概念和结构,还是只是机械复述。”<sup>[10]</sup>

之所以扰动会给模型带来错误推理的问题,根本在于大模型所基于的概率统计分析严重依赖独立同分布(independent and identically distributed, i.i.d.)假设,即训练数据中准确的输入输出映射将导致对具有相似特征的未见数据的良好泛化。然而,在很多情况下,i.i.d.是被违背的,即训练数据中的分布与测试数据或

①目前没有相关证据和研究表明,2025年8月发布的GPT-5实现了因果推理方法质的进步,或展现出因果推理能力质的飞跃。

要推广的未知数据中的分布相偏离，导致原本在训练数据中表现良好的模型在测试数据或推广场景中性能下降，具体表现为产生虚假相关性，进行的因果推理对人类而言不可理解（可解释性问题），在因果发现中无法准确识别因果相关变量等。GPT、DeepSeek等大模型采用思维链技术减少逻辑错误，提高推理能力，但思维链也经不起扰动，也需要对抗思维链干扰。

因果性理论与因果推理的技术实践应该是相互推进的，大语言模型用以解决扰动问题的技术方法是以现有因果性理论为根据的，而对此技术方法在实践中的具体分析也可以深化对因果性理论理解。

## 二、对抗扰动与因果性理论

“扰动”（perturbation）是指对模型的输入、参数、条件的微小改变，模型在处理不应对输出造成影响的扰动时，模型的预测结果不会发生变化，这是评估模型稳健性的重要方法。对模型性能的扰动评估所使用的常见方法是对抗性学习（Adversarial Learning），许多人工智能模型和算法的研究工作通过扰动模型的输入、参数、条件引入对抗性扰动，生成对抗样本（Adversarial Example），让模型在对抗样本上学习来提升模型的稳健性。对抗性扰动是一种微小但精心设计的扰动（不会对人类而只会对模型的推理产生干扰）。<sup>[11]</sup>

对抗性学习背后的理念正是因果关系的干预、操作定义，通过人类因果性思维的主观介入，产生新的动态数据，以及识别哪些数据是因果相关的、重要的或可忽略的，从而微调模型参数，提升稳健性。

首先，输入数据的扰动可以分为时间性和空间性两类。时间性扰动通过扰动因果关系中结果变量的其他前驱变量来检验模型稳健性。例如，在自动驾驶中，模型看到行人出现会触发刹车，看似掌握了“行人→刹车”的因果关系，但如果人类驾驶员习惯在发现行人时轻踩刹车导致刹车灯先亮，接近行人时才完成完整的刹车动作，模型就可能会形成“刹车灯→刹

车”的错误关联，这时候故意扰动刹车灯状态就能检验模型是否真的理解事件发生的先后链条。而空间性扰动是扰动因果关系所处的空间状态中存在其他共存的变量。例如，厨房里刀具归位本是为了安全，可如果训练数据里总由戴厨师帽的人完成这个动作，模型就容易把“厨师帽出现”和“刀具归位”错误绑定，这时候突然换个人来收拾刀具，模型若还能正确识别安全动机才算通过考验。

其次，不难看出，时间性扰动输入数据更多的是处理共因结构因果关系中的混杂变量问题，而在空间性上扰动输入数据更多的则是处理违背i.i.d.的泛化问题。与此同时，两者共同指向的是虚假相关性问题，也就是如何识别因果相关变量的问题。此外，还存在一些因果上相关但可忽略的变量。例如，“划火柴→火柴燃烧”，这是大多数模型通过学习能够推理出的因果关系，与此同时，空气湿度、环境温度、氧气含量等与划火柴的动作共同导致了火柴燃烧，如果对划火柴之外的其他这些变量进行扰动，那么就不一定出现“划火柴→火柴燃烧”，这时的模型可能认为“划火柴→火柴燃烧”不是因果关系，做出错误的因果推理。在人类的因果推理中，其他这些变量虽然也是因果相关变量，但在特定因果关系中会被视为可忽略的。

再次，对输入数据作上述类型的扰动，就是引入了因果干预，将因果推理和对抗性学习相结合，这是当前人工智能模型和算法研究的关注热点。通过因果干预，模型能够更好地理解数据中的因果关系，并提高对变化了的数据的适应能力。对抗性学习所生成的对抗性样本的生成器和判别器之间的对抗性训练可以被视为一种因果推理的示例，其中生成器试图通过欺骗判别器来产生与真实数据相似的输出，“对抗性学习通过生成对抗网络（GANs）隐式生成反事实。”<sup>[12]</sup>

对抗性学习并非对干预和反事实的建模，而是通过人类因果认知的介入，遵循可操作性的因果理论，实现干预和反事实层级的因果推理能力。

第一，可操作性因果理论的核心思想是制

造差异,对抗性学习对训练数据的微小改变,也可以理解为制造差异,检测模型在差异条件下是否稳健或者需要作何调整。以珀尔为代表的因果推理算法强调干预和反事实是对目标系统数据分布的相关性结构做出具有因果意义的调整,伍沃德(James Woodward)等人则提出可操控性因果理论为这种因果推理的方法提供了哲学基础上的支持。可操控性因果理论以“干预下的不变性”来刻画因果关系,强调因果解释的显著特征是,它们提供了可能与操纵和控制相关的信息,([13], p.6)即通过改变原因变量的值,可以操控结果变量值的改变。根据这一理论,人类对因果关系和因果解释的兴趣来源于人类对事物操控的实践兴趣,因果解释给出了人们对事物操控会产生什么样变化的信息,即会有什么样的差异产生的信息。

第二,分析对抗性学习对于训练数据的不同类型的扰动,可以深化对因果理论的理解。时空性上的扰动不同于可忽略变量的扰动,前者在“制造差异”的意义上理解因果关系的本质,后者也具有“制造差异”的含义,同时还要进一步明确,制造的是何种“差异”,这种差异在因果路径上发挥着怎样的作用。这就对大语言模型建模因果关系本质特征提出了更高的要求,正如当年珀尔受到卡内基梅隆大学科学哲学组“基于动态结构来研究因果的方法”的启发,认识到对“干预”的正确建模应该是图上的变化而不是静态的图,<sup>[14]</sup>从而才有了人工智能科学实践的因果革命。今天在科学哲学的理论层面探寻因果关系的更本质特征,对于大语言模型因果推理的发展和解决其遇到的问题,也具有启发意义。

### 三、差异传递:对因果关系的更本质刻画

大语言模型因果推理的研究对于什么样的因果相关变量是核心因果变量和可忽略的因果变量,从形式化因果建模方面做的努力不多,更多的是“避开了将这些概念形式化为因果模型的挑战,提供了一个直接从事件的自然语言描述中捕捉事件必要且相关背景信息的机

会。”<sup>[1]</sup>这么做是由于这项任务极其困难,核心和可忽略因果变量的区分涉及到人类的领域知识和常识。如,空气湿度、环境温度、氧气含量等与划火柴的动作共同导致了火柴燃烧,从而导致森林火灾的发生,这一情况中涉及到常态性/异常性、社会规范、主观意图等。人类凭借领域知识和常识可以轻松判定,划火柴是森林火灾的核心因果相关变量,而其他变量可忽略,这是由于能够助燃的空气湿度、环境温度、氧气含量是常态性的变量值,而在森林里划火柴是异常性行为,这种异常是对森林防火的社会规范的违背,而且如果这一行为具有一定的主观意图,那么引发森林火灾的后果是可以被归因追责的。只要对“划火柴→森林火灾”进行反事实推理,就可以确定因果关系,并且在对其他可忽略变量进行扰动时,不会改变这一因果推理和判断。

相比之下,大语言模型因果推理研究更多所采取的做法是,为大语言模型提供一系列实例,让大语言模型针对每个实例从多个选择题答案中选择一个来回答反事实条件问题。

例如,前提:一位女士没有点中餐。反事实问题:如果她点了中餐会发生什么?多个选择题答案:这位女士会没那么饿;这位女士会变得非常饿;那是不可能的。

大语言模型在预训练阶段具备了“以相关关系的形式重现人类的因果语言模式和因果推理结构”的能力,它就能够选择出正确的答案,实现反事实推理能力。但是在某些情况下,大语言模型还是不能捕捉到人们通过常识就能轻易判断的上下文信息,这尤其体现在对“涉及对所有相关其他变量的反事实翻转”和对常态性/异常性的推理方面,<sup>[1]</sup>这两者是密切相关的,都会导致在扰动可忽略的因果相关变量时,因果推理出现错误,模型性能下降。

例如,在GPT-4最容易出错的“短路”情况中:<sup>[1]</sup>

输入:卡罗尔还活着。爱丽丝在卡罗尔的水中放入了无害的解毒剂。向水中添加解毒剂可保护水免受毒药侵害。如果爱丽丝将解毒剂放入卡罗尔的水中,鲍勃会向水中下毒。向未

受保护的水中添加毒药会使其有毒。如果卡罗尔喝了有毒的水，她就会死亡。卡罗尔喝了她的水并存活下来。爱丽丝是卡罗尔存活的必要条件吗？

经过推理后，在标签 <Answer> 是 / 否 </Answer> 内给出最终答案。

GPT-4:

“我们可以将因果链设置如下：

1. 爱丽丝把解药放进了卡罗尔的水里。
2. 水中的解药能保护水不被下毒。
3. 如果爱丽丝把解药放进卡罗尔的水里，鲍勃就会给水里下毒。
4. 给未受保护的水下毒会使其有毒。
5. 如果卡罗尔喝了有毒的水，她就会死。

为了确定爱丽丝是否是卡罗尔存活的充分原因，让我们考虑一个反事实的情景：

如果爱丽丝没有把解药放进卡罗尔的水里，卡罗尔会存活吗？

在这种情况下，我们有：

1. 爱丽丝没有把解药放进卡罗尔的水里。
2. 水中的解药不能保护水不被下毒。
3. 鲍勃没有在水中下毒。
4. 水仍然无毒。
5. 卡罗尔喝了她的水并存活下来。

在这个反事实情景中，即使爱丽丝没有在她的水中放入解药，卡罗尔仍然存活。这表明爱丽丝的行为不是卡罗尔存活的充分原因，因为即使没有爱丽丝的干预，卡罗尔也会存活。

<Answer> 否 </Answer>”

在这一“短路”情况中，GPT-4进行了错误的反事实推理，给出了错误的答案。这说明GPT-4更多的是在对训练数据的因果语言作概率拟合，而非对反事实推理的本质要素进行建模，从而实现反事实推理能力。用结合因果图的结构方程模型来分析这个案例，能更好地呈现GPT-4因果推理能力的缺陷。

用 $A=1$ 表示爱丽丝把解药放进了卡罗尔的水里，没有放则是 $A=0$ ，用 $B=1$ 表示鲍勃就会给水里下毒，没有下毒则是 $B=0$ ，用 $W=1$ 表示水有毒，无毒则是 $W=0$ ，用 $C=1$ 表示卡罗尔死亡，存活则是 $C=0$ ，如此可以得到这一“短路”情况的因果图：

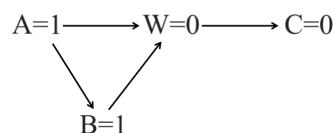


图1 “短路”情况的因果图

对应的结构方程组为 $A=1; B=1; W=\neg A \wedge B; C=W$ 。

(1)  $A=1 \rightarrow B=1 \rightarrow W=1 \rightarrow C=1$  是一条因果路径，即如果爱丽丝放解药，鲍勃就下毒，所以爱丽丝放解药并非保护水不被下毒，从而爱丽丝放解药不导致卡罗尔存活。这是反事实推理。

(2) 如果忽略 $B=1$ 这个变量及其取值，就只剩下 $A=1 \rightarrow W=0 \rightarrow C=0$ 这条路径，即爱丽丝放解药，水受保护免受毒药侵害，卡罗尔存活，但因为水本来就无毒，所以爱丽丝放不放解药，卡罗尔都会存活。从反事实推理来看，这不是一条有效的因果路径。

这是用结合因果图的结构因果模型分析得出的GPT-4的错误推理过程，问题恰恰在于GPT-4没有对反事实推理正确建模，也就是说，上述两条推理并不是结构因果模型分析这一情况的全部内容，还有很重要的信息被GPT-4错漏了，即：

第一，由于同时存在 $A=1 \rightarrow W=0$ 的因果路径，即爱丽丝放解药导致水受保护免受毒药侵害，因此， $A=1 \rightarrow B=1 \rightarrow W=1 \rightarrow C=1$ 的因果路径被取消掉了，即“短路”。因此(1)不成立。

第二， $B=1$ 表征的是鲍勃下毒，相比于 $B=0$ 鲍勃不下毒， $B=1$ 是一个异常值，异常取值的变量不应该被忽略。因此(2)不成立。

被GPT-4错漏的信息是，其一，“制造差异”的差异是什么样的差异；其二，制造的差异是通过什么路径传递的。

首先，诸多因果关系的研究者都是将默认、典型和常态 (defaults, typicality, and normality) 纳入因果推理的建模中，如霍尔 (Ned Hall)、哈尔彭 (Joseph Y. Halpern)、希区柯克 (Christopher Hitchcock)、孟席斯 (Peter Menzies) 等。以哈尔彭和希区柯克的工作为代表：将常态排序 (normality ordering) 添加到对因果关系的定义中，即因果推理的形式化构建中。在通常的因

果模型 $M=(S, F)$ 中,  $S$ 是模型中的变量及其可能的取值,  $F$ 是这些变量之间的结构方程, 在此基础上, 添加一个世界的部分排序, 可以用来比较不同世界的正常程度, 用符号 $\succsim$ 来表征,  $\succsim$ 可以用来比较现实世界和已经进行干预的世界。由此得到扩展的因果模型 $M=(S, F, \succsim)$ 。<sup>[15]</sup>

哈尔彭和希区柯克将默认(defaults)定义为, 在没有提供任何附加信息的情况下, 对发生的事情或情况的假设。典型(typicality)是指同一类事物的类型特征, 例如鸟会飞, 这是鸟的典型特征, 而如果实际情况中这只鸟是一只企鹅, 则它不是典型的, 而在没有进一步信息的情况下, 我们都是将事物与其类型特征联系在一起的。常态或正常(normality)则既可以是描述性的, 即某事物是统计模式或平均状态; 也可以是规范性的, 涉及道德规范、法律规范等多样形式, 如不撒谎符合道德规范, 因此是正常的, 撒谎则是不正常的, 又如除非有医生的证明, 否则员工是不允许旷工的, 员工按时上班是正常的。<sup>[15]</sup>

默认、典型和常态三个概念的含义相互交织, 共同构成模型中变量取值及其关系的常态排序, 对默认、典型和常态状态的偏离就是异常(deviant)。制造的差异是常态情况与异常情况对比之下的差异。

其次, 哈尔彭和希区柯克的研究是在因果推理形式化建模的工作中具有开创性意义的。他们是对刘易斯(David K. Lewis)将因果关系定义为反事实依赖关系这一理论的进一步发展和完善, 使用常态排序来解决反事实关系的“可能世界”语义学问题。从而在形式化因果建模工作中更充分地发现和推理反事实依赖关系不能完全覆盖的因果关系。

仅仅通过反事实依赖关系对因果关系的刻画, 会导致“短路”等情况中出现因果推理错误, 关键在于, 因果推理除要通过干预进行反事实依赖关系分析之外, 还需进行变量间的路径分析, 在路径分析中, 会涉及到原因和结果变量之间到底传递的是什么问题。有所传递, 才会有所依赖, 而没有依赖关系的地方, 可能也存在因果作用的传递, 这意味着因果作用不能仅仅以反事实依赖关系来刻画。这引出

的是依赖性(dependence)和传递性(transitivity)的关系问题, 贝克斯(Sander Beckers)和维内肯斯(Joost Vennekens)指出可以把因果关系分析为一种受不对称影响的传递关系, 即一个事件只有在它的缺席不会成为原因的情况下才会成为原因, 他们在因果模型中添加了不对称性条件, 以平衡因果推理的依赖性和传递性要求。<sup>[16]</sup>贝克斯和维内肯斯对不对称性条件的补充, 认为原因不能是事件的缺席, 也是在原因变量取值的意义上来考虑因果结构中的路径分析问题的。结果变量对原因变量的依赖, 不是由于结果依赖于原因的微不足道的变化, 而是原因制造差异传递给结果, 才会有原因和结果之间的依赖关系。换句话说, 要通过差异的传递来分析因果路径, 推得因果判断。

反事实依赖关系强调了因果的共变, 原因和结果有着相应相随的变化(variances), 而因果关系更为本质的不仅仅是传递变化, 而是差异(differences)。因此要考虑的是与变量的默认、典型和常态取值相比而出现的偏差(deviancy), 从而取异常值。

第三, 匹兹堡大学的迦楼(J. Dmitri Gallow)提出了一种模型不变的因果关系理论(A Model-Invariant Theory of Causation)<sup>[17]</sup>和以此为基础的因果过程追踪(Trace a Causal Process)规则。<sup>[18]</sup>

与哈尔彭和希区柯克不同的是, 迦楼不是将 $\succsim$ 作为不同世界的常态程度排序, 而是直接在变量取值意义上定义 $\succsim$ : 模型中每一变量哪个值比其他值更默认、典型和常态的规定 $\succsim$ 。<sup>[17]</sup>这种不同的实质是, 迦楼将哈尔彭和希区柯克扩展因果模型所揭示的现实世界与可能世界的常态差异, 用来说明现实世界因果关系的路径内容, 即具有依赖性关系的原因和结果变量之间的多条路径上, 是否有差异, 即偏差被传递。这也是结合了贝克斯和维内肯斯的工作, 在他们区分依赖性和传递性的基础上, 进一步揭示依赖性和传递性是在差异传递的意义上统一起来的。因果关系是差异制造的反事实依赖关系, 并且同时还是这种所制造的差异的传递关系。

在此基础上, 因果关系中的核心因果变量

和可忽略的因果变量的区分遵循对因果过程的追踪规则，即差异的传递规则。其核心理念可以总结为：（1）因果判断不依赖于原因和结果变量本身以外的任何常态信息，<sup>[18]</sup>也就是说，在因果推理中，可以忽略取常态值的变量，但绝不允许忽略取异常值的变量；（2）存在生成性和依赖性两种因果过程，追踪前者，不能忽略异常值变量，追踪后者，必须包含所有变量（包括常态值变量）。<sup>[18]</sup>

只有遵循这些因果性理论和规则，让模型追踪因果过程，识别因果过程中传递的异常性差异，才能准确区分核心因果相关变量和可忽略的因果相关变量，从而在可忽略的因果相关变量被扰动时，保证模型因果推理的稳健性。

#### 四、因果性理论驱动下大语言模型因果推理的发展

大语言模型因果推理的只是让因果假设本身在观测数据中得到体现，直接以相关关系的形式重现人类的因果语言模式和因果推理结构。而这些观测数据（训练数据）是静态的，缺乏与外部世界的互动，表现为因果语言模式和推理结构的相关关系缺乏人类因果语言和因果推理的现实性、操作性。通过在扰动中生成的对抗性样本上进行学习，可以改善大语言模型的因果推理能力，一定程度上缓解上述问题。分析模型对抗扰动的过程，一方面，可以证明珀尔所提出的因果层级定理对大语言模型因果推理的发展仍然是适用的，并且可以进一步深化对因果性理论的理解；另一方面，虽然因果性理论没有被完全纳入到大语言模型因果推理的建模之中，但已经在作为模型测试和评估技术的对抗性学习中发挥着作用，这种跳出模型本身的外部介入，正是因果推理获取新知（开放性）、改造世界（操作性）的本质特征。具体而言：

第一，深化对因果性理论的理解。通过扰动时空性上的非因果相关变量，可以暴露统计分析所获得的相关性关系中的虚假相关，说明因果关系比相关关系更能体现人类智能把握事物关系的本质。因果关系和相关关系两者都以

两个事物的共同变化来刻画事物的关系，而因果关系进一步强调这种共变是作为原因的事物制造了差异所产生的，因此因果关系是不对称的，这就比相关性关系有了更多的约束，也更符合人类智能把握和理解事物关系是要对事物进行控制和操作这一因果性理论的理念。

更重要的是，通过扰动特定语境下可忽略的因果相关变量，则可以暴露因果反事实推理的结构拟合存在的缺陷，说明人类智能认识和操控事物的关系，更深层次地动机是对不确定性现象进行确定性的认识和控制。事物关系中一定是出现了与正常情况相比的异常偏差（不确定性现象）才引发人类智能对其进行认识和控制的兴趣，因此，仅仅从差异制造来刻画事物关系达到控制的目的还不够，还需要明确所制造的差异的传递关系，以确定性的因果路径来认识和控制不确定性的现象。

第二，从认知科学的视角来看，大语言模型从大数据中生成更完整的因果图和更详尽的因果背景知识，对因果推理的发展提供强劲的技术支撑。其所基于贝叶斯认知模型，强调信念的动态更新，即将先验知识与新证据结合，修正信念（后验）的持续学习过程，基于这一认知理论的信念强度直接与现有数据和证据的权重相关。<sup>[3]</sup>而人类的因果推理通过控制操作更侧重发现令人惊异的不确定性变化，这种变化是相对于已知信念的偏差，也就是差异，确定这种差异的传递关系，也就获得了新的因果知识。在人类的因果推理，即因果认知的发生过程中，一个人所秉持的某一因果信念的强度或程度可以通过其因该信念而采取行动的可能性来衡量。<sup>[3]</sup>也就是说，大语言模型因果推理更侧重关注静态世界（已知数据）的平均值，而人类因果推理则更侧重关注动态世界（能动介入获得新的数据）的极端值。对应到结构因果模型中，变量的常态值、默认值只描述已有因果关系，变量的异常值才意味着新因果关系的发现，在因果图中追踪差异的传递，才能有新因果知识的产生。正是在这个意义上，珀尔认识到“对干预的正确建模应该是图上的变化，而不是静态的图”。

信念和因果推理能够帮助识别新的数据和实验干预,并最终验证那些曾经缺乏数据支持的信念。因果性理论的作用不仅仅是映射现有的现实,以数据和因果知识的形式被用作概率采样,拟合因果语言,而是生成未曾见过的未来可能性。因此,人类和大语言模型的因果推理应该是互动和互补的关系,大语言模型为人类梳理出关于现有现实的清晰详尽的因果关系背景和结构,而人类要通过自身的科学实践为这幅背景和结构图不断增添新的内容。对大语言模型的推理实施扰动,是人类通过自身的背景知识和能动行为提升大语言模型的“构图”能力,其本身就是运用人机各自的认知特征进行互动的过程。

总之,驱动大语言模型因果推理发展的因果性理论,以与默认/常态情况相比有偏差的差异传递关系为核心理念,涉及物理世界的规则,社会行为、道德法律等规范,因此大语言模型因果推理的进一步发展还需更多考虑实践性、互动性,只有在现实的物理世界中,在人类社会的规范中,才有常态和异常之分,需要在不断对因果推理的实践新知和理论本质的探寻中,规范模型和算法对人类智能的模拟。

#### [参考文献]

- [1] Kiciman, E., Ness, R., Sharma, A., et al. 'Causal Reasoning and Large Language Models: Opening a New Frontier for Causality'[EB/OL]. *Transactions on Machine Learning Research*, <https://openreview.net/forum?id=mqoxLkX210>, 2024-08-14.
- [2] Guo, D., Yang, D., Zhang, H., et al. 'DeepSeek-R1 Incentivizes Reasoning in LLMs Through Reinforcement Learning'[J]. *Nature*, 2025, 645: 633-638.
- [3] Felin, T., Holweg, M. 'Theory is All You Need: AI, Human Cognition, and Causal Reasoning'[J]. *Strategy Science*, 2024, 9(4): 346-371.
- [4] Simon, H. A. 'Cognitive Science: The Newest Science of the Artificial'[J]. *Cognitive Science*, 1980, 4(1): 33-46.
- [5] 尤洋、郭宇. GPT与因果性[J]. *科学学研究*, 2023, 41(12): 2122-2130.
- [6] Pearl, J. *Causality: Models, Reasoning, and Inference*[M]. Cambridge: Cambridge University Press, 2009.
- [7] 吴小安、俞沁元. 大语言模型与因果之梯[J]. *自然辩证法通讯*, 2025, 47(8): 10-19.
- [8] Zečević, M., Willig, M., Dhimi, D. S., et al. 'Causal Parrots: Large Language Models May Talk Causality But are Not Causal'[J]. arXiv preprint, arXiv: 2308.13067, 2023.
- [9] Lippe, P. 'Learning Causal Representations in Spatio-temporal Systems'[D]. Amsterdam: University of Amsterdam, 2025.
- [10] Hong, P., Majumdar, N., Ghosal, D., et al. 'Evaluating LLMs' Mathematical and Coding Competency through Ontology-guided Interventions'[J]. arXiv preprint, arXiv: 2401.09395V5, 2024.
- [11] 桂韬、悉志恒、郑锐等. 基于深度学习的自然语言处理鲁棒性研究综述[J]. *计算机学报*, 2024, 47(1): 90-112.
- [12] Deng, Z. Z., Zheng, X. L., Tian, H., et al. 'Deep Causal Learning: Representation, Discovery and Inference'[J]. *ACM Computing Surveys*, 2026, 58(2): 1-36.
- [13] Woodward, J., *Making Things Happen: A Theory of Causal Explanation*[M]. New York: Oxford University Press, 2003.
- [14] 董焕防. CMU的科学哲学与珀尔的因果理论——科学哲学如何影响人工智能?[J]. *自然辩证法通讯*, 2025, 47(1): 28-35.
- [15] Halpern, J. Y., Hitchcock, C. 'Graded Causation and Defaults'[J]. *The British Journal for the Philosophy of Science*, 2015, 66(2): 413-457.
- [16] Beckers, S., Vennekens, J. 'The Transitivity and Asymmetry of Actual Causation'[J]. *Ergo: An Open Access Journal of Philosophy*, 2017, 4(1): 1-27.
- [17] Gallow, J. D. 'A Model-Invariant Theory of Causation'[J]. *Philosophical Review*, 2021, 130(1): 45-96.
- [18] Gallow, J. D. 'How to Trace a Causal Process'[J]. *Philosophical Perspectives*, 2022, (36): 95-117.

[责任编辑 王巍 谭笑]