人工智能视角下因果推理研究的认知路径

王占
军事科学院军事科学信息研究中心，北京 100142

DOI: 10.61369/TACS.2026090042

摘　　要： 因果推理是智能系统由相关关系走向认识世界、干预世界、反事实思考的重要能力。目前主流的机器学习大多采用数据关联建模的方式，无法有效地应对分布偏移、混杂因素以及可解释性要求。认知科学对于人如何学习因果以及怎样推理的研究，为人工智能的因果建模提供了一些启发。本文在梳理机器学习和因果推理的关系之后，从干预主义、结构因果表征、反事实模拟三个认知路径入手，分析人工智能借鉴人类因果认知方式的技巧，并讨论其融合的趋势以及所面临的理论和技术挑战。

关 键 词： 人工智能；因果推理；认知路径；结构因果模型；反事实思维

Cognitive Paths of Causal Reasoning Research from the Perspective of Artificial Intelligence

Wang Zhan
Research Center for Military Science Information, Academy of Military Sciences, Beijing 100142

Abstract: Causal reasoning is a critical capability for intelligent systems to move from correlation to understanding the world, intervening in reality and conducting counterfactual thinking. Most mainstream machine learning models adopt data-driven correlation modeling, which cannot effectively cope with distribution shift, confounding factors and interpretability requirements. Researches in cognitive science on how humans learn and reason about causality provide inspirations for causal modeling in artificial intelligence. After sorting out the relationship between machine learning and causal reasoning, this paper analyzes the techniques for artificial intelligence to draw on human causal cognitive patterns from three cognitive paths: interventionism, structural causal representation and counterfactual simulation. It further discusses the integration trend as well as relevant theoretical and technical challenges.

Keywords: artificial intelligence; causal reasoning; cognitive path; structural causal model; counterfactual thinking

引言

因果推理属于智能特征之一。人类可以从少数观察中得出因果关系并预测干预结果，并设想“如果当初不同，结果会怎样”。相比主流的人工智能，尤其是深度学习方法，其更看重大规模数据中所具有的统计相关关系，从而很容易受到伪相关、数据选择偏差以及分布变化的影响。近些年来，因果推断在人工智能范畴中引起诸多关注，其不只是试图拟合数据，而且要找出数据背后产生的原因。同时，认知科学对人类因果认知的研究显示，干预操作、结构图式以及反事实想象乃是认识因果关系的三种基本认知手段。人工智能视角下的因果推理研究，越来越多地借鉴这些认知机制，形成了独特的研究路径。

一、机器学习与因果推理概述

（一）机器学习中的关联推理及其局限

现代机器学习，特别是统计学习和深度学习，本质上是通过关联进行推断。监督学习是用最小化经验风险拟合输入和输出之间条件概率分布的算法，它所提取的是数据中的相关特征。深度神经网络用多层非线性变换自动提取特征，但是它所提取的特征依然服务于预测的准确性，难以区分原因和结果、直接原因和混杂因素。

就方法论而言，关联推理的不足有三。第一，因果关系不能确定。两个变量之间统计相关性不能说明哪个是原因、哪个是结果，也不能排除共同原因造成的虚假相关。第二，混杂因素无法处理。当训练数据存在选择偏差或者隐藏混杂时，模型就会学到系统的错误关联。第三，无法回答干预和反事实问题。关联模型只能回答观察到 X 时 Y 如何，不能回答如果改变 X，Y 会怎样，或者如果过去没有 X，Y 会怎样 [1]。这些不足使只依靠数据拟合的

 Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

智能系统在遇到真实世界开放、动态、交互性任务时，就会变得很脆弱。

（二）因果推理的核心问题与层级

因果推理试图找出变量之间产生的原因，它的形式化基础主要是图灵奖得主 Judea Pearl 提出的因果阶梯理论。该理论把因果认知分为三层，第一层是关联，回答“看到 X 是否意味着 Y”；第二层是干预，回答“实施 X 是否导致 Y”；第三层是反事实，回答“如果当时没有 X，Y 是否会不同”[2]。每一个层级对应不同的数学工具以及推理能力，关联层级可以用概率分布描述，干预层级用因果图和 do 算子表示，反事实层级用结构方程模型和潜在结果框架表示。

结构因果模型（Structural Causal Model，SCM）是因果推理的核心形式化工具。它是由一组内生变量、外生变量和结构方程组成，每一个结构方程表示一个变量是由它的直接原因所产生的机制[3]。因果图用有向无环图表示变量之间的因果关系，do 算子切断指向干预变量的全部入边，模拟外部干预后的分布变化。在此基础上，因果推理有两大任务，即因果发现和因果效应估计。因果发现是从观察数据或者干预数据中学习因果图结构，因果效应估计是在给定因果图或者部分先验知识的情况下，估计干预或者反事实的结果。

（三）人工智能引入因果推理的理论动因

人工智能引入因果推理不是偶然产生的兴趣，而是为了满足现实任务需求而产生的必然结果。首先，因果推理可以提高模型的泛化能力，因为因果机制在分布变化中比较稳定，所以学习因果结构而不是表层的相关性，可以使模型在环境发生变化时仍然有效。其次，因果推理可以提高可解释性。黑箱模型给出的预测不能说明“为什么”，但是因果模型可以揭示变量之间的生成路径，为决策提供了机制性解释[4]。再次，因果推理支持干预与规划。在机器人控制、医疗决策、公共政策等各个领域中，智能系统要预测行动的后果，本质上就是干预推理的问题，单靠关联模型无法完成。最后，因果推理可以达到公平性和伦理约束的目的。通过将敏感属性的直接因果效应和间接效应区分，在模型中消除非法的因果路径，从而减少算法歧视。

二、人工智能视角下因果推理研究的认知路径

（一）基于干预与操作的因果认知路径

干预主义因果理论认为，因果关系不是被动观察到的规律，在理想的干预条件下可以保持稳定。哲学家 James Woodward 认为，X 是 Y 的原因，只有存在对 X 的理想干预，使 Y 的变化系统地依赖于 X 的变化。该理论强调操作在因果识别中处于中心地位[5]。认知心理学研究表明，儿童从一开始就会表现出通过操作对象来学习因果关系的倾向，他们不但会观察事件之间的共变，还会主动地去制造干预检验自己的因果假设。儿童会反复推动物体，观察不同的动作可以产生怎样的结果，然后根据结果改变自己的因果判断。

人工智能采取干预主义路线，形成主动因果学习、因果强化学习这些办法。主动因果学习就是设计出最佳的干预策略，选取最有信息量的变量做实验，用最小的干预成本识别因果结构或者估计因果效应[6]。该类方法把干预看作是获取信息的一种行为，用贝叶斯优化、信息增益或者因果熵等准则指导实验的设计。因果强化学习关注智能体在与环境交互的过程中，怎样用干预经验学习因果模型，进而改善长期回报。机器人领域研究显示，机器人主动操作环境物体比仅仅看视频数据更能使它形成物理因果图式。

（二）基于结构因果模型与图式化表征的认知路径

人类因果认知的又一明显特点就是用结构化的图式组织世界知识。认知心理学中的因果地图理论认为，人们把复杂的外界世界简化成由节点和箭头组成的因果网络，节点代表变量或者事件，箭头代表因果影响。图式化的表征是模块化的、层次化的、可组合的，不同的领域因果图式可以独立地更新、迁移，也可以组合成更复杂的因果解释[7]。人们认识“下雨造成地湿”，“地湿造成滑倒”是两个简单的机制，可以将二者串联起来进行推理。

结构因果模型给认知图式赋予了严格的数学表述。SCM 把世界建模成变量集合和结构方程，因果图就直接显示出变量之间的因果拓扑。人工智能中把 SCM 和深度学习结合起来，产生了因果表征学习、因果图神经网络等方向[8]。因果表征学习是从高维、非结构化的数据中提取出低维的因果变量以及它们之间的关系，使得模型不仅可以做预测，还可以在潜在的空间里进行干预、反事实推断。因果图神经网络把因果图当作归纳偏置加入图神经网络当中，在节点分类、链接预测以及动态系统建模时利用因果邻居信息缩减混杂偏差。

（三）基于反事实思维与认知模拟的认知路径

反事实思维属于人类因果认知里非常特别的一种高级能力。它指向与事实相反的可能情境，即“如果昨天不下雨，那么比赛就不会取消”。认知科学认为反事实推理依靠心理模型、可能世界语义完成，即人们在大脑里建立现实的因果模型，再对模型中的某些变量进行修改，从而模拟出修改后结果的传播过程。反事实推理不但是因果判断的依据，也是道德归责、情感体验、学习改进的心理基础。人们为什么会后悔，是因为反事实比较出另一种选择会带来更好的结果[9]。

人工智能中的反事实推理是在两个方面展开的。第一层就是反事实解释。在可解释人工智能中，反事实解释就是生成出最小的输入样本，使模型输出发生变化，从而告诉用户哪些特征的变化会导致不同的决策。信贷审批场景下系统可以告知申请人，收入增长20% 贷款就会批准该种解释符合人类直觉，具有较强的可操作性、沟通效率。第二层就是反事实推断。根据潜在结果框架或者 SCM，反事实推断用来回答个体层面的因果问题，即“某患者如果不接受治疗，那么他的康复概率会是多少”。反事实分析用以考察个体在敏感属性发生改变的时候，其决策是否随之变化，从而找出并消除歧视性的因果路径。

（四）认知路径的融合与挑战

以上三条认知路径不是互相独立的，而是一条线上的三个部分，彼此之间相互联系、相互支撑。干预主义路径给因果知识提

Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

供操作来源，结构表征路径给因果知识提供组织方式，反事实路径给因果推理提供模拟机制 [10]。人类在日常生活中可以灵活地综合使用观察、干预、结构学习和反事实想象形成一个动态的、可以修正的因果认知系统。

融合路径的典型趋势有三个方向，一是用先验因果图和深度学习相结合，在数据不足的时候用结构约束提高学习效率；二是用大语言模型中的因果知识、常识辅助因果发现和反事实生成；三是人机协同场景下，使人类因果判断和机器因果推断互相校正。

但是认知路径的工程化还存在着诸多难题。从观察到因果的可辨识性上着手。仅仅依靠被动观察的数据，很多因果关系是无法唯一确定的，必须有附加的假设或者干预数据。其次就是混杂和选择偏差普遍存在的现象，真实世界数据很难达到随机化、无隐藏混杂的理想状态。第三，因果表征学习没有统一的评价标准，怎样判定潜在变量是否对应真实的因果变量还存在争议。第四，认知启发算法会把人类的认知偏差复制出来，比如对因果链的过分简化、忽略非线性交互、受到确认偏误的影响。因此人工智能因果推理难以简单地模仿人类的认知，在借鉴其优点的基础上，用形式化的手段纠正它的偏差。

三、结语

总之，人工智能视角下的因果推理研究，正在从以数据关联为中心的范式，转向融合干预、结构表征和反事实模拟的认知范式。认知科学提供了理解因果推理的丰富理论资源和经验证据，干预主义、结构因果模型和反事实思维分别对应于人类因果认知的操作维度、组织维度和模拟维度。三条路径的交叉融合，有望推动人工智能在可泛化性、可解释性、决策可靠性和伦理公平性等方面实现深层提升。然而，从认知机制到形式化算法、从受控实验到开放世界，仍存在理论和技术上的显著鸿沟。未来研究需要在因果可辨识性、结构学习、反事实生成和跨学科融合等方面持续深入，构建既能统计学习又能因果理解的人工智能系统，使其在复杂真实环境中具备更稳健、更可解释的智能行为。

参考文献

[1] 陈炯宇 . 基于多尺度特征取和异常因果推理的交通流预测研究 [D]. 太原科技大学 ,2025.

[2] 唐婧尧 . 基于因果推理的低资源信息抽取研究 [D]. 大连理工大学 ,2025.

[3] 胡丽娟 . 基于因果推理的特征选择算法研究 [D]. 桂林电子科技大学 ,2025.

[4] 高枫 . 基于因果推理和反事实文本增强的多模态情感分析 [D]. 山东师范大学 ,2025.

[5] 赵葛剑 , 张新鹏 .DeepSeek：从 " 概率生成 " 到 " 因果推理 "[J]. 自然杂志 ,2025,47(2):79-84.

[6] 蔡玉琦 . 从贝叶斯推理到因果推理 [D]. 吉林大学 ,2024.

[7] 郑洁 . 基于因果推理的数据库基数估计研究与实现 [D]. 宁夏大学 ,2023.

[8] 任晓明 , 李晟 . 因果推理和因果认知：当代逻辑与哲学发展的一个新的生长点——四川大学博士生导师任晓明教授访谈 [J]. 社会科学家 ,2022,(8):3-8+161.

[9] 黄闪闪 . 因果推理研究的认知进路 : 基于人工智能视角 [J]. 湖南科技大学学报 (社会科学版),2021,24(2):44-50.

[10] 江滨 . 强人工智能中的因果推理研究 [D]. 湖北大学 ,2020.

 Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License