深度神经网络模型的认识论反思
——基于解释与经验支持度的区分

向 盾
(南开大学 哲学院,天津 300350)

摘 要:文章借鉴马尔乔尼对理论模型无法提供解释的批判性分析,区分"解释"与"经验支持度"概念,强调只有前者才能评估模型是否具有解释性,并规范"解释性理解"概念的适用场景。基于此,文章反驳沙利文关于深度神经网络(DNN)模型能够提供目标现象解释性理解的观点,指出DNN模型与理论模型在解释功能上不具可比性,即使消除链接不确定(Link Uncertainty)也无法弥补DNN模型的机制黑箱问题。进一步结合延展认知视角,"建模者—理论模型"认知对能通过推理承担解释任务,而"建模者—DNN模型"认知对则拥有隐秘的私人学习过程,缺乏解释所需的机制。

关键词:机器学习;科学建模;链接不确定;延展认知
中图分类号:N02 **文献标识码**:A **文章编号**:1674-7062(2025)06-0024-07

一、前沿

从黑色素瘤分类器到最近在蛋白质结构预测方面取得突破的AlphaFold3,机器学习(Machine Learning)在科学领域的应用越来越广泛。但由于它包含机制黑箱,导致DNN模型的科学认知作用近年来被哲学家们[1]110-111,[2]1-3,[3]323-325,[4]1823-1824广泛讨论。艾米丽·沙利文(Emily Sullivan)对机器学习中的黑箱问题进行了层级分类,主张认知不透明并不一定会导致解释性理解障碍[1]128-130。迈克尔·塔米尔(Michael Tamir)和埃莱·谢赫(Elay Shech)对DNN模型成分进行分类,区分功能近似无关性(functionally approximate irrelevance)与实施无关性(implementation irrelevance)[3]329-330,完善和发展了沙利文的观点。斯特凡·布伊斯曼(Stefan Buijsman)考察了个别DNN模型在发现因果图(learning causal graph)[2]3以及估算因果强度(the strength of causal relation)[2]8中发挥的解释作用,但对于多数DNN模型的因果解释能力仍持否定态度。相反,蒂姆·拉兹(Tim Räz)和克劳斯·贝斯巴特(Claus Beisbart)对此完全持反对态度[4]1823-1824。本文的核心问题是,利用机制不透明的模型进行科学研究时,模型是否有能力为建模者提供解释或解释性理解(explanatory understanding)。沙利文通过对比理论模型和DNN模型,认为DNN模型和目标之间的链接不确定——即缺乏支持模型与目标现象之间联系的科学经验证据——阻碍了人们的理解[1]30。我们可以通过显著性图(saliency map)、训练数据集以及相关经验性研究来降低链接不确定,让DNN模型中的神经元与目标现象具有充分的联系,从而提供关于目标的解释。本文通过卡特琳娜·马尔乔尼(Caterina Marchionni)对理论模型无法承担解释任务的批判性反思,区分模型的"解释"概念和"经验支持度"概念。前者涉及概念适当性问题,具有两极性(涉及"是""否");后者则是一个连续统,涉及程度差异(具有"高""低"之分)。本文认为,DNN模型的解释难题关乎前者而非后者。正是因为混淆了这两个概念,沙利文错误地将谢林模型具有的解

【收稿日期】 2025-02-18
【基金项目】 天津市哲学社会科学规划项目"模型表征的视角主义研究"(TJZXQN23-001)
【作者简介】 向 盾(1992-),男,湖北恩施人,南开大学哲学院博士研究生,研究方向为科学表征、机器学习。

释能力归于其经验支持度的提高,即通过独立的科学经验研究降低链接不确定。由此,DNN 模型也被误认为能够用该方法承担解释任务,导致模型的机制黑箱问题被忽视。除此之外,本文将从延展认知角度出发,指出"建模者——理论模型"认知对可以通过推理过程承担对目标的解释、论证等任务。相比之下,"建模者——DNN 模型"认知对则拥有隐秘的私人学习过程,无法提供解释所需的相关机制。

为避免歧义,本文所讨论的模型主要指未被经验充分确证的理论模型①。它们(如谢林的棋盘模型)通常具有高度抽象化与理想化特征,例如在经济学模型中,建模者常常忽略行为主体的收入与政策因素,假定完全理性的行为人,或者使用棋盘式城市布局等。这些特征使模型既简洁又便于操作,旨在帮助建模者在复杂的交互现象(如商业周期、社会秩序、居住隔离现象等)中分离出潜在的因果机制,从而在理论层面上更好地把握研究对象。形式上,本文将模型视为"假设——推理"形式组合:假设包括实质性假设和辅助性假设(用来隔离干扰因素以及保障模型的数学可操作性),推理机制则由建模者指定。实质性假设和推理机制往往被视为模型解释的核心成分。本文第二节回顾马尔乔尼对理论模型的研究,区分了"解释"与"经验支持度"两个概念;第三节强调,决定模型能否承担解释任务的关键在于其"解释性"而非"经验支持度",并指出沙利文在此处的混淆;第四节从延展认知的视角考察 DNN 模型与理论模型之间的认识论差异,反驳 DNN 模型能够提供科学解释或解释性理解的观点;第五节对全文进行总结。

二、解释与解释的程度

马尔乔尼对模型解释的讨论与理论模型的解释悖论相关。朱利安·赖斯(Julian Reiss)[5] 指出,经济学模型存在解释悖论:

(1)经济学模型的假设包含错误(由理想化、抽象化等导致)②。

(2)然而,经济学模型提供解释。

(3)唯有真实的说明才能提供解释。

所以,经济学模型由于自身包含的大量不准确性导致其解释能力受到怀疑。科学模型的解释悖论受到多位哲学家的关注,安娜·亚历山德罗娃(Anna Alexandrova)和罗伯特·诺斯考特(Robert Northcott)、柯林·赖斯(Collin Rice)以及亚历山大·罗伊特林格(Alexander Reutlinger)等人都对该观点表示赞同[6-8]。相反,乌斯卡里·麦基(Uskali Mäki)、佩特里·伊利科斯基(Petri Ylikoski)、埃姆拉·艾迪诺纳特(Emrah Aydinonat)等人从模型族(model family)和论证有效性方面反驳该观点[9]268,[10]。其中,雅克·库里科斯基(Jaakko Kuorikoski)和伊利科斯基[11]3830-3831认为,理论模型的目的并不在于重现目标所有丰富的细节,它只试图捕捉部分被假定为核心的解释依赖关系。在此基础上,马尔乔尼[12]604-606认为,从模型真假角度来考察问题混淆了两个独立问题之间的区别,它们都与模型解释力(explanatory power)的评估相关:一个是概念问题(conceptual issue),涉及模型满足何种要求才能够解释;另一个是认识论问题(epistemological issue),涉及相信模型有多大可能满足这些要求。解释具有二极性,模型要么提供解释,要么没有提供解释。把概念问题和认识论问题都用二分法(dichotomy)来考察,混淆了经济学建模涉及的不同问题:一个问题是理想化模型就其本质而言不能提供解释;另一个问题是经济学建模在多大程度上免受经验证据的影响。前者涉及概念适当性;后者则是一个连续统。可见,消解解释悖论的关键在于强调模型包含不同组成部分。核心解释部分往往要求一定的经验真实性,而辅助假设部分往往出于研究实用性目的与现实相差很大。其中,辅助假设往往被分为伽利略假设和可处理性假设。前者用来隔离感兴趣的目标,比如忽视摩擦力对速度的影响;后者将核心解释部分用可处理的数学结构来表征,比如在谢林模型中用特定数学函数表达居民移动规则。后者的虚假性在很多时候无法通过去理想化来根除,同时模型解释又紧密依赖于它。在此基础上,马尔乔尼认为模型辅助假设是否被经验支持并不影响模型的解释性,只会影响模型在"潜在解释"(potential explanation)[12]609和"实际解释"(actual explanation)[12]609所构成的连续统中所处的位置。经验研究所提供的支持(经验支持度)无论是哪个极端,模型都至少被视为一种解释[12]624。从这个角度看,模型所包含的理想化特征并不影响其提供解释。

在谢林的种族隔离模型中,模型从一个类似棋

① 为行文方便,下文中凡提及"模型",均特指"理论模型"。

② 笔者添加。

盘的城市中随机分布的行为人开始。行为人有黑白两种肤色,可以原地不动,也可以无偿迁移。他们愿意与同肤色居民结群相居,即用于解释居住隔离现象的核心因素。当该偏好得不到满足时,行为人将移动所在位置直至满足偏好。谢林模型的突破性见解在于,它证明了种族隔离和其他形式的空间隔离都不一定来自强烈的歧视性偏好。但该模型包含了大量的抽象化和理想化。建模者忽略制度、政策、人均收入等情况对迁移的影响,从而隔离对目标现象有影响但不感兴趣的因素;建模者将搬家行为进行离散化,以回合推进形式呈现;建模者对个人偏好的差异性、多样性进行理想化处理,呈现出同质性、单一性特征,帮助实现数学可操作性。与本文相关的方面在于,这些抽象化、理想化等手段让模型与目标之间具有极大差异,建模者需要经验支持来增加模型作为实际解释的信心,这正是独立经验研究发挥作用的地方。然而,建模者常常无法对模型进行直接的经验验证,这常常使建模者对模型处于怀疑、可能、很有可能的连续统中[9]275。但无论处于连续统(潜在解释——实际解释)中哪个位置,它都已经是一种解释。所以,经验支持不会剥夺模型的解释性。

这里存在潜在的反对意见:在建模活动中,如果完全失去经验支持,模型似乎不能建立与目标的相关性,进而无法解释目标,甚至无法表征目标。这个问题的本质是:经验支持是建立模型与目标之间相关性的必要条件吗?毫无疑问,在模型的核心解释部分,经验元素是必要的。否则,建模仅仅只是一个独立于经验的抽象活动。但是,在本文的讨论语境中,经验性支持主要针对辅助假设(特别是可处理性假设)而言。也就是说,可处理性假设缺乏经验支持不会剥夺模型提供解释性理解的资格。比如,谢林模型在提出之后的数十载,其核心解释部分都可以被总结为:种族隔离和其他形式的空间隔离都不一定来自强烈的歧视性偏好。但针对可处理性假设采用的去理想化,稳健性分析等手段,则增加了对该解释部分的信心。正是在这个意义上,沙利文将理论模型和 DNN 模型在不同范畴上进行了错误的对比:理论模型中用来增加建模者信心的证明活动被误解为能够赋予 DNN 模型解释能力。在此基础上,她认为降低"链接不确定"能够让 DNN 模型提供解释,即增加支持模型与目标现象之间联系的科学和经验性证据[1]3。根据沙利文对深度患者模型(deep patient model)的考察可知,相关的科学和经验证据主要包括建立"更多统计模型以提高结果的稳健性,进行临床试验,或者开展各种纵向研究等方式[1]31。虽然不同学科对相关证据的数量、质量的要求不一样,但我认为可以大致总结为通过经验、观察、实验或统计分析等科学方法获得的,可以用来支持或证伪模型与目标现象之间解释关系的证据。简而言之,它是关于目标的科学经验证据。请注意,笔者并不否认经验支持的重要性。事实上,具有经验支持的实际解释意味着科学家和决策者的高信任度。相对于潜在解释,它更加直接影响科学共识和社会决策,对未来的科研进程具有强大的影响力。只是,这不能作为模型能否提供解释的依据。

可见,理想化、抽象化并不妨碍模型提供解释。若模型本身具有解释力,那么经验证据有助于增强模型解释的经验支持度;若模型提供预测、猜想或启发,则提供相应的经验支持度。马尔乔尼主要从理论模型角度考察解释问题,本文试图将相关的概念区分应用到 DNN 模型讨论中,强调理论模型与 DNN 模型之间的对比研究存在上文提到的模糊之处,解释力和解释信心属于不同范畴。相关的独立经验证据虽能降低沙利文所说的链接不确定,但仅提升预测或启发的经验支持度,不足以赋予其解释力或解释性理解。最后,关于科学解释的哲学讨论纷繁复杂,笔者无意提供一种新的或坚持某种已有的科学模型解释观。为了论证的方便,本文沿用当下广为接受的机制观和因果观[13],即当模型成功地表征了目标现象的部分原因或机制时,模型就解释了目标现象[9]272。

三、沙利文论证

沙利文的论证如下:谢林模型的解释能力来源于模型与目标之间的链接(也被她称为高层次黑箱,阻碍对现象的解释),具体的运算细节则是低层次黑箱(不必然阻碍对现象的解释)。同样,DNN 模型可以通过显著性图以及模型训练数据去除链接不确定,使模型能够提供解释或解释性理解(explanatory understanding)。沙利文所关注的解释性理解,也被称为"理解——为什么(understanding——why)"[14]。"解释'为什么'有助于我们理解'为什么'"(explaining why helps us to understand why)。DNN 模型提供解释性理解的最大障碍往往在于,建模者无法从 DNN 模型中获得关于目标系统的因果关系或机制,模型内部的推理规则是黑箱。为解决该问题,沙利文区分"理解模型"(understanding models)和"用模型去理解"(understanding with mod-

els),“理解和解释模型的工作原理与使用该模型理解感兴趣的现象是有区别的”[1]112。DNN 模型黑箱只会影响我们理解模型,但不必然影响我们对目标现象的理解。她认为,在理论模型中也存在与模型工作原理相关的算法实施黑箱(implementation black box),它并不会影响我们“用模型去理解”目标现象,只影响我们“理解模型”,因为对现象的理解并不需要模型在实施(算法)方面完全透明。比如,谢林模型需要计算阶乘函数(factorial function),而这可以通过几种不同的计算方式实现。她认为用模型提供目标系统的解释性理解不需要理解谢林模型中阶乘函数的具体实施方式,而只需要相信该阶乘函数实施正确即可。在这种情况下,实施黑箱是无害的。只要模型不存在最高层次的黑箱,就不会妨碍理解。因为最高层次黑箱意味着只有输入和输出已知,使得解释和理解的可能性非常有限[1]116。所以,实施黑箱似乎并不会对理解造成影响。当然,如果所解释的问题涉及实施过程,或者实施过程对高层次结果有影响,那么实施黑箱当然会造成解释障碍。但至少,模型的实施黑箱并不必然影响建模者对目标系统的解释性理解。相反,沙利文认为链接不确定才是阻碍建模者解释目标现象的主要因素。本文赞同沙利文认为实施黑箱并不与解释相冲突,但同时认为链接不确定也不是阻碍解释的原因。相反,模型的机制黑箱才是 DNN 模型面临的解释挑战。根据对沙利文思想的介绍,本节试图从模型类比、链接不确定与“用模型去理解”现象之间的关系去批判沙利文的观点。

(一)黑箱类比

在理论模型中,建模者为模型提出假设(包含理想化、抽象化等手段)并指定相关推理规则,最后通过具体的运算操作得到相关假设的结果。从这个角度看,它可以被视为一个论证过程,与思想实验没有本质区别。进行建模时,建模者首先会通过对目标系统进行抽象化、理想化等方式建立起模型假设,同时指定模型被允许的推理方式(归纳推理、演绎推理等方式),建立起必要的相关性。我们可以将其称为“启发阶段”。在启发阶段完成后,模型成为一个封闭的系统,具有自己的生命[15],随后的实际操作是使结果清晰化的过程(类比于科学中利用草稿进行计算的过程)。在该过程中,建模者在假设上进行具体操作,得出关于目标系统的具体信息,简称“操作阶段”(算法实施阶段)。涉及模型的运算规则、模型与目标的相关性等假设在启发阶段被建模者利用背景知识明确下来。沙利文所说的模型实施黑箱源自第二阶段的操作过程,因为操作的复杂性导致建模者往往无法对整个过程清晰明了。但该过程严格意义上并不存在黑箱,因为只要建模者愿意,他能够检查运算的每个步骤。从这个角度看,操作阶段是对启发阶段所提出的假说的确证。用更通俗的话说,操作阶段通过科学界公认(可信的)的数学算法将启发阶段提出的推理规则加以实施。所以,在模型中,算法是推理规则得以实现的工具。这使具体的实施方式虽然对于建模者而言并不清晰明了,但只要它们可靠,模型就能够对目标现象进行推理。

但在 DNN 模型中,模型的推理方式对于建模者而言是不透明的,导致“模型——目标”机制方面的相关性无法被建立。如果建模者通过相关独立研究挖掘相关机制和因果关系,提供 DNN 所需要的相关性,那么 DNN 模型的解释责任实际上已经被转移到相关的独立实证研究中。理论模型则相反,在启发阶段,建模者必须给出明确的机制和必要的假设,确保模型自身提供关于目标的推理。值得强调的是,在建模过程中追求操作阶段的心理上的明晰对于科学解释而言并不是必要的。从这个角度看,科学模型类似于数学草稿上的运算推理。它们帮助我们执行运算细节,并且有人会认为草稿上的演算推理构成科学研究的黑箱。所以,谢林模型在建立成功时,就已经能够提供解释,因为模型能够利用背景知识提供关于目标的反事实依赖关系或解释机制①(居民移动规则)。相关的算法只是实现该机制的手段,“模型——目标”之间的机制相关性能够被建立,并不依靠后来独立实证研究提供的“模型——目标”链接。相反,DNN 模型或许可以建立模型元素相关的链接,类似于谢林模型中“人——棋子”“棋子颜色——人的肤色”“棋子周围的颜色分布——居民邻居肤色分布”,但它无法提供清晰的模型内部推理机制,机制黑箱使得 DNN 模型无法像理论模型那样提供关于目标的解释。除了建模所必要的模型与目标之间的相关性之外,模型和目标之间的链接只是帮助谢林模型具相更高的实证支持,增加建模者的主观信心,而不赋予模型本身解释能力。

由于混淆了模型概念问题和认识论问题的差

① 这取决于采纳何种解释观。但无论采用哪种,谢林模型都能够满足。

异,沙利文的论述中包含着模糊和矛盾之处。沙利文认为,“建立现象与模型之间的必要联系并不因此取代对模型的需求或模型的认识价值。即使模型与现象之间的联系不再不确定,模型仍然可以解释现象。”[1]113那么,DNN 模型提供解释的能力是来源于模型本身,还是来源于后期相关的独立实践研究?如果来源于前者,那么后期用来降低链接不确定的实践研究就没有发挥解释作用,这与沙利文的主张相悖;如果来源于后者,那么 DNN 模型就只是一个预测机器,沙利文同样无法接受。除此之外,沙利文还面临一个难题,如果链接不确定是一个程度概念,那么到达何种程度时一个模型才能被算作解释呢?她可能会回答,这取决于实践需要。但这并不能彻底摆脱问题,谢林模型实际上已经提供了解释所必需的元素(居民肤色情况、相关肤色容忍度、肤色满意程度等)和机制(移动规则)的相关性,满足了实践所需,为何不能被称作解释呢?笔者认为,在该问题方面,沙利文无法提供独立的评价标准,因为她混淆了解释和经验支持度之间的差别,而链接不确定只关乎后者。沙利文在文中提到的显著性图手段,也只是通过突出输入数据(如图像或文本)中对模型输出影响最大的区域或特征,揭示模型在决策时重点关注的部分。可见,它并不能提供相关的推理机制、因果关系等解释元素。依靠独立经验研究增加模型与目标之间的关联程度也只涉及决策、权重的关联。从这个角度看,DNN 模型是强大的启发装置,提供相关性和预测,而非解释。

(二)链接不确定

在沙利文对谢林模型和 DNN 模型的分析中,链接不确定是指缺乏独立的科学经验证据支持“模型——目标”之间的认知基础(epistemic foundation)。这主要缘于建模者的认知局限性以及研究目标的复杂性,导致模型需要抽象化、理想化处理,从而剥削了经验支持。这种经验支持可以针对解释,也可以针对预测,其本身并不足以赋予原模型解释力。换句话说,笔者并不否认相关科学证据有助于建立一个具有较高经验支持度的模型,甚至提供一个较为实际的解释性理解,而是认为建立解释性模型并不必然依赖于上述科学证据。建模者完全可以建立一个初始实际意义较弱的模型,但仍然提供解释①。然而,在 DNN 模型中,独立实证研究或许能够后期提供模型与目标之间的相关性,但关键在于建模者既无法根据背景知识,也不能通过独立实证研究提供解释所必要的机制。如果相关独立研究在后期提供了 DNN 模型的机制,那么 DNN 模型的解释责任实际上转移到了相关实践研究上,DNN 模型本身仍然无法提供解释。这一点在理论模型中有所不同,因为理论模型本身所提供的机制能够指导研究。

究其根本,DNN 模型所能提供的是相关权重和统计概率方面的预测,而不是机制。相反,理论模型在机制方面并不存在黑箱。就目前关于科学模型解释的哲学讨论而言,统计概率并不必然提供科学解释,这极大程度削弱了沙利文认为 DNN 模型能够提供解释的论点。比如,当下被广泛考察的“鲁索 - 威廉姆森理论”(Russo - Williamson - Theory),它由费德丽卡·鲁索(Federica Russo)和乔恩·威廉姆森(Jon Williamson)提出,简称 RWT。该理论从因果认知理论(epistemic theory of causality)角度对因果关系的确立进行研究,“要确立因果关系,科学家需要机制和依赖关系的相互支持”[16]。可见,单纯的统计学相关性和机制的泛化稳定性都不足以提供完整的因果关系,它们需要共同作为证据支持健康医学研究中的因果关系。统计证据面临混杂因素(confounding factor)和非因果关联相关的难题,即确立了变量 A 和 B 之间的关联后,仍可能存在共同因素 C 导致二者。“鲁索 - 威廉姆森理论”强调机制在提供解释时的必要性,而这正是 DNN 模型所欠缺的。相反,理论模型则能够依靠建模者的背景知识提供模型运转机制。

四、延展推理视角

为了更清晰的呈现 DNN 模型和理论模型之间的认知区别,我在讨论中引入科学建模的延展推理说明(extended cognition inferentialism account),这种思想建立在安迪·克拉克(Andy Clark)和戴维·查莫斯(David Chalmers)提出的延展认知[17]基础上。在该视角的帮助下,我们能够更清晰地看出 DNN 模型和理论模型之间的差异。

模型延展推理说明由库里科斯基、阿基·莱蒂宁(Aki Lehtinen)和伊利科斯基提出。库里科斯基和莱蒂宁针对科学模型指出:“当借助外部媒介(如纸笔、计算机或棋盘)和一套形式化的保真推理规则完成从假设到解释的推理时,相关的认知系统就

① 后续的经验证据可能改变该模型的经验支持度。

不是建模者一个人,而是'模型——建模者'认知系统对(model – modeler cognitive system pair)。"[18]库里科斯基和伊利克斯基强调"表征性人工制品所能做出的解释性推论的范围和可靠性取决于用户和模型之间的交互"[11]3834。它强调建模者和模型研究目标时共同发挥的认知作用。在"认知对"的帮助下,建模者和科学模型共同构成的认知系统(简称Sm)能够提供对目标系统的解释。所以,"科学的目标不是为科学家提供令人满意的体验——这显然可以用更少的资源实现——而是为了提高我们集体对世界做出正确假设性推断的能力"[11]3823。由此可见,电脑模拟或者玩具模型在推理论证方面和科学家在草稿纸上进行运算推理是没有本质差异的,也正是因为这个原因,谢林模型才能够通过"要求明确的模型假设"[11]3824"提供推理可靠性"[11]3824以及"回答跟更多的反事实问题"[11]3825提供解释性理解。

然而,建模者和DNN模型所组成的认知系统(简称Sd)与Sm在认识论方面存在差异。在Sd中,模型并不是建模者推理能力的延伸,而是启发(heuristic)能力的增强,它提供目标现象中各元素之间的相关性。相比于理论模型的公开的推理过程,它的分类决策过程是隐秘的,习得的。以黑色素瘤分类器为例,与皮肤科医生相比,该分类器能更好地识别黑色素瘤。如果我们将"认知对"概念引入该模型的考察中,在成功训练出DNN模型后,"DNN模型——建模者"认知对就能直接从皮肤图像中识别出某人是否患有黑色素瘤。通过比较Sd和Sm,前者的认知过程是相关性的,隐秘的,启发的;后者的认知过程是推理性的,公开的,证明的。

五、总结

沙利文通过比较谢林模型和DNN模型,指出实施黑箱不会阻止谢林模型提供解释性理解,所以DNN模型的黑箱也不会。同时,降低链接不确定能够帮助DNN模型解释目标现象。本文首先通过马尔乔尼对解释的概念问题和认识论问题之间的区分,指出谢林模型的解释能力并不来源于独立的科学经验研究;然后,在此基础上强调解释所需要的条件,尤其是模型在机制方面的透明性。在谢林模型中,建模者能够利用背景知识明确相关关系或机制,算法作为其实现手段并不会影响模型的解释能力。然而,DNN模型中的机制是隐秘的,习得性的,这是其提供解释的最大障碍。倘若科学经验研究能够提供相关机制,那么解释责任便从DNN模型转移到了经验研究上。沙利文的论证并不清晰,并且存在矛盾之处。最后,本文试图引进科学模型中的延展推理说明,指出Sm作为一个整体能够通过推理实现对目标的解释、论证等任务。相比之下,Sd并不能展现其推理能力,它只拥有隐秘的私人认知过程,无法回答反事实问题或呈现相关机制。"DNN模型——建模者"认知对作为一个整体能够实现对目标的感知,讨论的范畴是意识、预测、表征等领域。

【参 考 文 献】

[1]SULLIVAN E. Understanding from machine learning models [J]. The British journal for the philosophy of science, 2022, 73 (1).

[2]BUIJSMAN S. Causal scientific explanations from machine learning[J]. Synthese, 2023,202(6).

[3]TAMIR M, SHECH E. Understanding from deep learning models in context [C]// LAWLER I, KHALIFA K, SHECH E. Scientific understanding and representation modeling in the physical sciences. New York: Routledge Press, 2022.

[4]RAZ T, BEISBART C. The importance of understanding deep learning[J]. Erkenntnis, 2024, 89.

[5]REISS J. The explanation paradox[J]. Journal of economic methodology, 2012, 19: 49.

[6]ALEXANDROVA A, NORTHCOTT R. It's just a feeling: why economic models do not explain[J]. Journal of economic methodology, 2013, 20(3): 262.

[7]RICE C. Moving beyond causes: optimality models and scientific explanation[J]. Noûs, 2015, 49(3): 589 – 615: 589.

[8]REUTLINGER A, HANGLEITER D, HARTMANN S. Understanding with(toy) models[J]. British journal for the philosophy of science, 2017, 69(4): 1095.

[9]MAKI U. On a paradox of truth, or how not to obscure the issue of whether explanatory models can be true[J]. Journal of economic methodology, 2013, 20(3).

[10]YLIKOSKI P, AYDINONAT E. Understanding with theoretical models [J]. Journal of economic methodology, 2014, 21(1): 23 – 24.

[11]KUORIKOSKI J, YLIKOSKI P. External representations and scientific understanding[J]. Synthese, 2015, 192.

[12]MARCHIONNI C. What is the problem with model – based explanation in economics? [J]. Sciendo, 2017, 9(47).

[13]HAUSMAN D. Paradox postponed[J]. Journal of economic methodology, 2013, 20(3): 250.

[14]CHRISTOPH B, CLAUS B, GERG B. What is understanding? An overview of recent debates in epistemology and philosophy of science[C]// CHRISTOPH B, SABINE

A. Explaining understanding: new perspectives from epistemology and philosophy of science. New York: Routledge, 2017: 13.
[15] HUGHES R I G. Models and representation[J]. Philosophy of science, 1997, 64(4): S331.
[16] RUSSO F, WILLIANSON J. Interpreting causality in the health sciences[J]. International studies in the philosophy of science, 2007, 21(2):159.
[17] CLARK A, CHALMERS D. The extended mind[J]. Analysis, 1998, 58(1): 7 - 19.
[18] KUORIKOSKI J, LEHTINEN A. Incredible worlds, credible results[J]. Erkenntnis, 2009, 70:122.

Epistemological Reflections on Deep Neural Network Models

——Based on the Distinction Between Explanation and Degree of Empirical Support

XIANG Dun

(*College of Philosophy*, *Nankai University*, *Tianjin* 300350, *China*)

Abstract: This paper builds upon Marchionni's critical analysis that theoretical models fail to provide explanations, distinguishing between the notion of "explanation" and that of "empirical support", and highlighting that only the former serves as a genuine criterion for evaluating a model's explanatory power. It also clarifies the appropriate contexts for employing the concept of "explanatory understanding". Based on this framework, the paper challenges Sullivan's recent claim that deep neural network (DNN) models can provide explanatory understanding of target phenomena. The paper argues that traditional scientific models and DNN models are not comparable in terms of their explanatory function, and that reducing link uncertainty does not suffice to enable a model to provide explanations. Rather, the black - box nature of DNN mechanisms poses a significant challenge to their explanatory capacity. Furthermore, adopting an extended cognition perspective, the paper contends that the "modeler - traditional scientific model" cognitive system can engage in explicit reasoning to fulfill explanatory and argumentative tasks, whereas the "modeler - DNN model" cognitive system remains embedded within private learning processes, lacking the necessary mechanistic visibility for explanation.

Key words: machine learning; scientific modelling; link uncertainty; extended cognition

(责任编辑　赵　雷)