

人工智能的可解释性问题探析

——基于因果关系视角

谢佛荣,景海龙¹,邓菲菲¹

(南华大学 马克思主义学院,湖南 衡阳 421001)

[摘要] 随着人工智能技术的发展与应用,人工智能的可解释性问题受到了越来越多人的关注。如何解释输入与输出之间的联系,成为许多学科的研究重点。对此,相关关系的主张者与因果关系的主张者提出了各自的观点和理论。相比较而言,从因果关系视角阐释人工智能的可解释性问题是更为合理的一种路径。但要实现这一路径,还需要借助人工智能学者朱迪亚·珀尔对因果关系的三个层次区分,以消解传统因果关系理论彼此间不相容所造成的解释上的局限性,从而实现可解释的人工智能。

[关键词] 人工智能; 可解释性; 相关关系; 因果关系

[中图分类号] B82-057; TP18 **[文献标识码]** A **[文章编号]** 1673-0755(2022)04-0040-06

DOI:10.13967/j.cnki.nhxb.2022.0050

2021年世界人工智能大会法制论坛上,法布罗西尼(Fabrocini)做了题为“向可信赖人工智能迈进”的演讲^[1],指出人工智能发展的新趋势,即构建“可信赖”人工智能。虽然人工智能已经深入人们生活的各个领域,但是我们仍需进一步审视人工智能与我们人类之间的关系,思考当前的人工智能是否是“可信赖”的,以及什么样的人工智能才是“可信赖”的。当前学术界认为,实现可信赖人工智能的关键在于真正理解人工智能,明白人工智能为什么做出那些推理,并将这些内容概括为人工智能的可解释性问题。为此,一些学者试图从相关关系的视角把握人工智能的可解释性问题。但是这种做法难以澄清关系间的内在机制,从而无法科学解释推理背后的原因。据此,笔者试图借助朱迪亚·珀尔的因果关系思想,探析人工智能可解释性问题的因果路径。

一 人工智能的可解释性

巴尔的摩的一家非盈利组织近年来一直强烈支持在保释过程中引入人工智能。但是现在,该组织却改变了这一观点。组织负责人声称,不应该在审判前的司法活动中将人工智能作为风险评估工具,

因为这样做将会导致种族歧视的问题^[1]。这一事件反映出当前人工智能在各个领域的广泛应用引起了担忧,对此许多学者警告说,不应该盲目地相信人工智能。据报道,某信息安全研究所对北京一地的两个公租房小区进行了人脸识别系统测试,结果发现残障人士的顺利通过率低于一般正常人;国外某互联网公司发现,旗下一视频网站的推荐算法正以某种方式推动自己的用户观看儿童色情视频;一国际研究中心发现,其发明的物品识别系统在识别来自低收入国家的物品时表现较识别发达国家的物品时要差。在这些案例中,人工智能所产生的种种问题,一部分可以归结为缺乏相应的社会管理机制,另一部分则源于当前人工智能所面临的不可解释性问题。

由于人工智能的可解释性问题的存在,使得人工智能在实际应用中出现的问题难以得到解决,同时也阻碍了人们对某些领域成熟有效的人工智能的推广和应用^[2]。这些问题近年来在多个领域引起了广泛的讨论,这些讨论既是出于人工智能技术的实用性需求,同样也是出于理解人工智能本质的哲学需求。

[收稿日期] 2022-05-27

[基金项目] 教育部高校示范马克思主义学院和优秀教学科研团队建设项目“新时代弘扬大学生伟大民族精神研究”资助(编号:19JDSZK095)

[作者简介] 谢佛荣(1983—),男,江西赣州人,南华大学马克思主义学院副教授,哲学博士,博士后,
¹ 南华大学马克思主义学院硕士研究生。

(一)人工智能的可解释性是什么

目前对于人工智能的可解释性,学界还没有一个明确的定义,许多学者从不同角度提出了各自的观点。总的来说,人工智能主要由三部分组成:输入、模型、输出,可解释性指的是对人工智能模型的解释。

人工智能模型为什么会失去可解释性?传统计算机软件的设计原理都是严格遵循刚性逻辑而建立的,解决任何问题都需要首先满足三个先决条件:(1)问题中输入和输出间的数量关系;(2)数量关系间的算法解;(3)用计算机可识别的语言进行编程^[3]。因此,传统计算机中的模型是一个具有明确数量关系的算法解,并不存在可解释性的问题。但是随着人们实践需求的提高,这一做法被发现具有很大的局限性,难以实现更为复杂的推理。于是一些学者将目光转向相关关系,通过对数据间相关特征的拟合来完成复杂推理的输出。并且通过对人脑活动的模拟,使拟合不断深入,从而获得更为准确的输出。随着拟合程度不断深入,模型越来越复杂,这种相关关系已经难以被认知,人工智能也就失去了可解释性。

由于模型高度的非线性和调参技术,人们无法探知模型究竟从数据中学到了什么以及如何通过这些数据来进行决策,这种“端到端”的模式导致当前人工智能的解释性极弱^[4]。这导致了人工智能在实际应用中常出现两类问题:(1)当一种模型在实践中表现出良好的效果时,我们不知道是什么使其有效,从而无法在其他模型中进行推广。(2)随着人工智能应用范围的扩大,错误的输出也时常发生。当这些错误出现时,我们无法知道是什么原因导致了这些错误,从而难以在模型中作出相应的修正。这同时也造成了一个伦理上的问题,即难以分辨出前文所述的那些引发歧视的错误输出究竟是由于模型造成的,还是由于在设计之初就已存在。

(二)人工智能是否具有可解释性

有一种观点认为,当前人工智能的不可解释所造成的可解释性问题不是人工智能的一种缺陷,而是它的一种特征^[5]。笔者认为,这一观点值得商榷。因为,这一观点错误地将人工智能等同于基于相关关系的模型,认为当前模型的不可解释是人工智能的一种特征。暂且不论基于相关关系的模型是否可以被解释,人工智能的目的并非创造一个不可被理解的“人”。而是协助人们实现一些能行过程,促进人们对自身认知、行为机制的进一步理解。因

此,当前人工智能的不可解释不应该被视为人工智能的一种特性,而应该视为同传统计算机不够“智能”一样的一个待解决的问题。并且,人工智能也应该具有可解释性。在某些领域,如人脸识别,人工智能的不可解释所造成的问题只是识别的失败,不会带来重大影响。但在诸如医疗等领域,人工智能的不可解释所可能带来的严重后果是不可接受的,并且与传统医学理念背道而驰^[6]。2018年5月25日实施的欧盟《通用数据保护条例》中规定,人工智能必须符合一定要求的可解释性。2021年9月25日,我国国家新一代人工智能治理专业委员会发布的《新一代人工智能伦理规范》中规定,要在多个层面提升人工智能的可解释性。多个国家与地区已经开始立法对人工智能的可解释性进行要求,可见无论是从理论还是实践的角度,人工智能都应该具有可解释性。

由上可知,人工智能的可解释性问题主要是由于当前人工智能模型的不可解释造成的。这种模型基于相关关系对数据进行拟合,造成了输入与输出之间的认知困难。

二 基于相关关系的可解释性困境

对于人工智能可解释性问题的探讨,当前学界提出了两种路径:一种是基于相关关系;一种是基于因果关系。但是基于相关关系的可解释性在理论和实践中都面临着一些问题,下面将阐明和分析基于相关关系的可解释性困境。

如今,消费者通过互联网进行商品购买;企业通过大数据进行客户发掘;工厂通过智能机器进行产品生产;汽车通过自动驾驶系统进行无人驾驶,等等。这些事例都表明,21世纪是一个数字化的时代,数据正代替现实事物,成为人们实践的对象,大数据正代替现实世界,成为人们实践的领域。一些学者认为,这个数据—大数据“世界”的本质是相关关系。所谓“数据”,不是指传统意义上的数字,而是指一切将对象电子化的记录。“大数据”则不仅仅是数据在量上的增加,而是将世界上的一切事物都看作是由数据构成的,皆可量化,并用数据加以表示^[7]。从“数据”到“大数据”的转变所造成的影响在于,数据只是对于事物的一种反映,就其本质来说仍是客观现实世界的一部分。而大数据本身已经构成了一个完整的“世界”,其内部具有一定的规律和法则。大数据哲学家罗伯·基钦(Rob Kitchin)就曾指出,数据就其本身而言,或许只是一个中立的现实

要素。但是在大数据时代,对于数据的认识将不得不考虑地方、主体性、政治经济学、制度规制和思想体系等因素。或许数据已经成为了一个认识论单元,以至数据如何被本体论的定义不再是一个中立的技术过程,而是一个规范的、政治的和道德的问题^[8]。因此,数据已经不只是一个科学概念,同时也是一个哲学概念。在此基础上,客观世界中事物之间的联系法则已经在数据一大数据“世界”中失去了效力,其所产生的作用被相关关系所取代。

相关关系是指两个变量中,一个变量发生的变化与另外一个变量发生的变化是关联的^[9]。例如,一名未成年人的年龄与他的身高之间就存在相关关系,因为随着年龄的增加,他的身高也会增加,见图1。对于一名未成年人的年龄与身高之间的这种相关关系,通常寻找一个函数来进行表达,由此,当我们知道这名未成年人的年龄时,可以大致知道他的身高。此外,现实中一些变量之间并不如同上述例子一样具有明显的相关关系,例如某地咖啡的销量与博物馆参观人数之间的关系,见图2。那么就需要收集一定的数据,然后再对这些收集到的数据进行拟合,根据拟合得出的函数来获得其他未知信息。这种拟合构成了当前人工智能模型的基础,通过在拟合的层次与方法上的深入与创新,模型越来越复杂,也越来越智能。

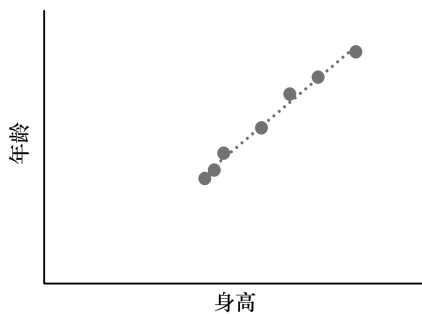


图1 未成年人身高、年龄相关性

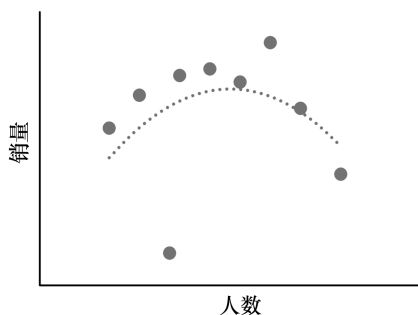


图2 咖啡销量、博物馆参观人数相关性

但是这种做法面临着三个问题:(1)有些相关关系比较复杂、难以解释。对于不同的相关关系,一个直观的感受是它们具有不同强度的相关关系。未成年人年龄与身高之间的相关关系和某地咖啡销量与博物馆参观人数之间的相关关系中,前者明显强于后者。统计学上通常的做法是通过线性相关系数来度量相关关系的强度,称之为相关性强度。系数的绝对值越趋近于1表示相关性强度越强,越趋近于0表示相关性强度越弱。但是,相关性强并不一定意味着线性相关系数高。因为线性相关系数只能度量简单的相关关系,但是有一些相关关系是复杂的、非线性的。这时就需要引入更多新变量,来度量高维的相关性。随着更多新的变量引入,数据整体的其他指标又会下降,因此,对于复杂数据相关性强度的度量需要综合考量。而这个问题就比较复杂、难以解释。(2)相关关系的寻找依赖经验判断。这主要表现在对数据的选择上。首先,在实际情况中我们选用的数据不一定就具有代表性,可能无法反映出数据的基本分布情况。例如对某种疾病致死率的研究,如果只选用医院的急诊科数据,那么所得到的致死率将会比社会整体人群的实际致死率高得多。反之,对某种疾病痊愈率的研究,选用急诊科的数据将会导致比率较社会整体人群低得多。上述只是一个简单的例举,很多实际研究将会复杂得多。其次,除了一开始的数据选择,对于复杂相关关系中其他变量的选择也依赖于经验。例如某位商人研究天气情况与冰激凌销量之间的相关关系,而所得的数据并未反映充分的相关性强度。这时他就需要选择其他变量以期获得显著的相关关系来指导之后的决策。选择时间、职业或是其他变量很大程度上决定了他是否会错失一些有意义的关联。(3)相关关系难以描述所有数据关联。相关关系的前提是相互依存的变量变化,没有变化就没有相关关系。但现实中某些明显关联的数据之间可能体现出令人疑惑的变化,或者没有变化。以辛普森悖论为例,该悖论指出:存在这样的数据,总体上的统计结果与其每一个子部分的统计结果相反^[10]。例如,在一种新药的测试中,根据总体统计,服用该药的人群中男性患者比不服用该药的男性患者痊愈率高,女性患者也比不服用该药的女性患者痊愈率高。但是总体上,数据显示服用该药比不服用该药的患者痊愈率低。也就是说,根据统计数据,当知道患者性别时,服用该药有益于治疗;当不知道患者性别时,服用该药无益

于治疗,这显然是荒谬的。问题的原因在于数据间的内部关系是复杂的,简单的数据分解可能导致不同子组间具有不平衡的表示形式。

因此,因果关系的主张者认为相关关系事实上只是在对数据进行总结,想要获得数据之间的确定联系还需要对数据进行解释,而要获得数据间的确定联系需要在因果关系这一层次才能得以完成。数据不是万能的,它只能够告诉你服药的患者比不服药的患者更容易康复,但不会告诉你原因何在。也许,那些服药的患者之所以更容易康复只是因为他们的病情较轻,即使不服用这种药物,他们依然能够很快康复。

三 基于因果关系的可解释性路径

不同于人工智能中通过拟合实现数据(事实)与数据(事实)之间的联系,哲学中,一些学者认为因果关系是事实(数据)与事实(数据)之间的一种联系法则,并试图借助不同方法对这种法则进行解释,以期澄清这种联系。对因果关系的分析最早可以追溯到亚里士多德,他提出“四因说”,认为一个因果关系至少应该从动力、目的、形式、质料四个角度来进行说明。现代意义上对因果关系的探讨起源于休谟,他在《人性论》中提出了两种不同的因果理论,规则性理论与反事实理论。在休谟之后,许多哲学家都对因果关系提出了各自的看法,主要分为五种:规则性理论(Regularity Theory)、反事实理论(Counterfactual Theory)、概率理论(Probability Theory)、过程理论(Process Theory)和干预理论(Interventionist Theory)。在此基础上,朱迪亚·珀尔(Judea Pearl)提出了基于因果关系的人工智能模型基础:关联、干预、反事实。

(一)关联

关联层回答的问题是:“如果观察到……结果会怎样?”这一问题涉及的是人的观察能力,即发现事物间相关性的能力^{[11]7}。珀尔认为,这一能力是所有动物都具有的,只是因果关系中最低级的一层。“一只猫头鹰观察到一只老鼠在活动,便开始推测老鼠下一刻可能出现的位置,这只猫头鹰所做的就是通过观察寻找相关性。”但是,问题的关键是什么老鼠下一刻会出现在那里?相关关系的支持者们发明了很多复杂的方法来总结老鼠下一刻最可能出现的位置在哪里,但是他们无法解释老鼠为什么总是出现在该处。因此,基于数据拟合的相关性事实

上解释的是一种动物的“智能”,而非人的“智能”。要想实现真正的“人”工智能,还需要进入到更深的层级。

(二)干预

干预层回答的第一个问题是:“如果做了……结果会怎样?”这个问题涉及的是人的行动能力,即主动干预事件的能力^{[11]10}。珀尔认为,干预比关联更高级。无论数据集多大,神经网络多深,只要是被动的收集数据来寻找相关性,都是无法回答诸如“在此处种一棵树,下一次老鼠还会在这出现吗”这样的问题。正如观察到一场火灾和制造一场火灾,二者意味着完全不同的事实。这一层级回答的第二个问题是:“怎么做?”^{[11]11}因为在珀尔看来,如果你希望干预一个变量,那么你还需要思考的一个问题是如何对这个变量进行干预,因为有些行为会直接改变变量,而有些行为则只是对变量产生影响,两者的意义是不一样的。尽管干预使我们进入到一个更高的层次,但是要想实现因果关系,珀尔认为还需要进入到下一个层次。

(三)反事实

反事实层回答的问题是:“如果发生了……那么原因是什么?”这个问题涉及的是人的想象与反思能力,即对没有发生的事物进行思考的能力^{[11]12}。前两个层次中,即使是干预层,对象也只是基于数据的事实,但是事实无法告诉我们一个没有发生的情况会导致什么结果,对此则需要在反事实的世界中进行想象。珀尔认为,“在弹性限度内,假如加在这根弹簧上的砝码重量是原来的两倍,弹簧伸长的长度也会加倍”这样的论述就是一种反事实的论述。它首先需要观察到弹簧伸长长度与施加砝码的重量之间存在相关关系(关联),然后需要主动改变砝码重量来发现变量之间的相对关系(干预),最后这一相对关系被奉为定理,在不进行实际行动的基础上得出某一砝码重量下弹簧的伸长长度(反事实)。

珀尔认为,对相关关系的推理只是推理过程中的最低层次。要想达到更高层次,还需要借助干预与反事实,以实现因果关系的推理。并且通过三个层次的数理化,构建更智能的、可被解释的人工智能。根据珀尔的论述,他所指的因果关系并不局限于某一传统因果关系理论之中,而是对这些理论的一种综合。(1)在关联层,对相关关系的发现事实上是一种对事物间恒常连接的观察,这是对规则性理论的一种运用。珀尔同休谟一样,认为这种或然

性中蕴含着因果关系。但是休谟囿于经验的考察而止步于此,珀尔则进行了更深入的探讨。(2)在干预层,通过珀尔提出的两个问题,“如果做了……结果会怎样?”与“怎么做?”,是对于干预理论的一种运用,将因果关系视为可用于操纵和控制目的的一种潜在关系。(3)在反事实层,珀尔认为日常生活中习以为常的反事实条件句中显现着从事实到非事实的认知革命。因此,人工智能对人脑的模仿不仅在于神经网络层次的不断深入,还应该包含反事实中现实与想象间的跨越,并且后者才是思维的关键所在。(4)三个层次共同构成了人工智能完整的因果推理,一个完整的因果推理所回答的是“假如当时发生的事情与实际不一样,结果依然如此的概率是多少?”这是对概率理论的一种应用,是对这个世界中事物发生的可能性的一种度量。佩德罗·多明戈斯(Pedro Domingos)在《终极算法:机器学习与人工智能如何重塑世界》中提出了一种处理人工智能中相关关系的观点,构建综合各种机器学习算法学派优点的终极算法以实现理想的人工智能^[12]。他的观点并不是希望设计出一种通用算法来回应所有问题,而是讨论现有各类算法的功能与理论边界,探索这些算法间内在相容的可能性,以及如果存在这样一个解决所有问题的终极算法,那么它至少应该具有哪些特征,以期当前算法的设计提供一些启示。在这一视角下,珀尔所做的是一件相似的事情,即探索一种“终极因果”来解决人工智能的可解释性问题。

概率因果关系的主张者古德(I. J. Good)认为,因果关系不应该被分解到对各自有利的层面上进行讨论。在珀尔看来,以往的因果关系哲学讨论正陷入这一困境当中。规则性理论、反事实理论、概率理论、过程理论和干预理论都将因果关系的某一特征上升到了本体论地位,构建出彼此独立的因果关系理论。因此,珀尔从因果关系的认知基础出发,认为因果关系真正的认知功能在于指导人的推理过程,是对事物间特殊联系的一种发现。对于因果关系的认知基础的传统讨论分为两种观点。一种是由亚里士多德所提出的“解释论”,认为因果关系的功能在于解释事物的存在为何这样,而非其他。另外一种观点是由休谟所提出的“联系论”,认为因果关系的功能在于联系不同的事物。由于“联系论”下的因果关系通常包含大量非因果性的内容,其后的分析多以“解释论”作为因果关系的认知基础。但是

“解释论”的问题正在对相关关系的排除。因为因果关系不同于相关关系这一认知同时也包含了另外一个认知,即因果关系同相关关系的区别并不等同于和无关关系的区别,有些时候因果关系与相关关系的界限甚至是模糊的。后一认知促使我们去诉求一种对相关关系更具包容性的因果关系理论,而这正是珀尔所做的。除了上述理论意义之外,珀尔因果关系思想的另外一个优点在于对现有人工智能相关关系探索的肯定与继承。使得因果关系不是对于相关关系否定后的另辟蹊径,而是站在相关关系之上的一种方法论指导。并将这一方法论指导视为人工智能可解释性问题的解决之道。即在相关关系的基础上,通过“干预”与“反事实”来解决当前基于相关关系拟合的人工智能模型所面临的问题,实现具有可解释性的人工智能。

四 结 语

综上所述,对于人工智能的可解释性的探索不单单是一个模型设计的问题,还关系到如何理解人工智能以及人自身的推理能力。珀尔对因果关系的理解为走出当前困境提供了一条可借鉴的路径。在珀尔的因果关系观点下,因果关系与相关关系并非两个对立的观念,而是推理的不同阶段。通过对相关关系之下更深层次的“干预”和“反事实”的应用,能够解决当前人工智能在“关联”层次所面临的困境,迈向实现推理的、可解释的人工智能。

[参考文献]

- [1] 菲利波·法布罗西尼. 向可信赖人工智能迈进[J]. 上海质量, 2021(8):16-20.
- [2] 吴丹, 孙国焯. 迈向可解释的交互式人工智能: 动因、途径及研究趋势[J]. 武汉大学学报(哲学社会科学版), 2021(5):16-28.
- [3] 何华灿. 重新找回人工智能的可解释性[J]. 智能系统学报, 2019(3):393-412.
- [4] 化盈盈, 张岱璿, 葛仕明. 深度学习模型可解释性的研究进展[J]. 信息安全学报, 2020(3):1-12.
- [5] 刘艳红. 人工智能的可解释性与 AI 的法律责任问题研究[J]. 法制与社会发展, 2022(1):78-91.
- [6] 吴新松, 马珊珊, 赵春昊. 人工智能可解释性与评估方法[J]. 信息技术与标准化, 2021(7):21-26;32.
- [7] 贺天平, 宋文婷. “数—数据—大数据”的历史沿革[J]. 自然辩证法研究, 2016(6):34-40.
- [8] 赵伟, 李迪. 大数据科学的因果解释性理论难题刍议[J]. 科学技术哲学研究, 2019(5):106-111.

- [9] [美] 萨曼莎·克莱因伯格. 别拿相关当因果! [M]. 郑亚亚, 译. 北京: 人民邮电出版社, 2018: 48.
- [10] NICHOLAS P, JEWELL. 统计因果推理入门 [M]. 杨矫云, 安宁, 李廉, 译. 北京: 高等教育出版社, 2020: 2.
- [11] [美] 朱迪亚·珀尔, 达纳·麦肯齐. 为什么: 关于因果关系的新科学 [M]. 北京: 中信出版社, 2019.
- [12] [美] 佩德罗·多明戈斯. 终极算法: 机器学习与人工智能如何重塑世界 [M]. 北京: 中信出版社, 2016: 33-34.

Analysis of the Interpretability of Artificial Intelligence

——Based on a causal relationship perspective

XIE Forong, JING Hailong, DENG Feifei

(University of South China, Hengyang 421001, China)

Abstract: With the development and application of artificial intelligence technology, the interpretability of artificial intelligence has attracted more and more attention. How to explain the connection between input and output has become the focus of research in many disciplines. In this regard, proponents of correlation and causation put forward their own viewpoints and theories. Comparatively speaking, it is a more reasonable path to explain the interpretability of artificial intelligence from the perspective of causality. But to achieve this path, artificial intelligence scholar Judea Pearl's three-level distinction of causality is needed in order to eliminate the interpretation limitations caused by the incompatibility of traditional causality theories, thus explainable artificial intelligence can be realized.

Key words: artificial intelligence; interpretability; correlation; causality

(本文编辑: 刘衍永)