doi:10.19306/j.cnki.2095-8110.2022.06.003

人工智能可解释性评估研究综述

李　瑶[1,3]，左兴权[2,3]，王春露[1,3]，黄　海[2]，张修建[4,5]

（1. 北京邮电大学网络空间安全学院，北京 100876；
2. 北京邮电大学计算机学院，北京 100876；
3. 可信分布式计算与服务教育部重点实验室，北京 100876；
4. 北京航天计量测试技术研究所，北京 100076；
5. 国家市场监管重点实验室（人工智能计量测试与标准），北京 100076）

摘　要：近年来，可解释人工智能（XAI）发展迅速，成为当前人工智能领域的研究热点，已出现多种人工智能解释方法。如何量化评估 XAI 的可解释性以及解释方法的效果，对研究 XAI 具有重要意义。XAI 的可解释性评估涉及主、客观因素，是一个复杂且有挑战性的工作。综述了 XAI 的可解释性评估方法，首先，介绍了 XAI 的可解释性及其评估的概念和分类；其次，总结和梳理了一些可解释性的特性；在此基础上，从可解释性评估方法和可解释性评估框架两方面，综述和分析了当前可解释性评估工作；最后，总结了当前人工智能可解释性评估研究的不足，并展望了其未来发展方向。

关键词：可解释性评估；人工智能可解释性；主观评估；客观评估；评估方法；神经网络；深度学习
中图分类号：TP18　**文献标志码**：A　**文章编号**：2095-8110(2022)05-0013-12

Research Progress of Artificial Intelligence Interpretability Evaluation

LI Yao[1,3], ZUO Xing-quan[2,3], WANG Chun-lu[1,3], HUANG Hai[2], ZHANG Xiu-jian[4,5]

(1. School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing 100876, China;
2. School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China;
3. Key Laboratory of Trustworthy Distributed Computing and Service, Ministry of Education, Beijing 100876, China;
4. Beijing Aerospace Institute for Metrology and Measurement Technology, Beijing 100076, China;
5. Key Laboratory of Artificial Intelligence Measurement and Standards for State Market Regulation, Beijing 100076, China)

Abstract: In recent years, explainable artificial intelligence(XAI) has developed rapidly and is becoming a research hotspot in the field of artificial intelligence. Many artificial intelligence explanation methods have emerged. How to quantitatively evaluate the interpretability of XAI and the effect of explanation methods are of great significance to the study of XAI. The evaluation of XAI's interpretability involves subjective and objective factors, which is a complex and challenging task. This paper reviews the interpretability evaluation methods of XAI. Firstly, the concepts and classifications of XAI's interpretability and its evaluation are introduced. Secondly, we summarize and sort out some characteristics of XAI's interpretability. On this basis, current studies on interpretability evaluation are reviewed and analyzed from both aspects of interpretability evaluation method and interpretable assessment framework. Finally, shortcomings of current studies on XAI's interpretability evaluation are summarized, and the future development directions

收稿日期：2022-08-20；**修订日期**：2022-10-22
作者简介：李瑶（1997-），女，硕士研究生，主要从事可解释人工智能、人工智能可解释性评估方面的研究。
通信作者：左兴权（1971-），男，教授，博导，主要从事智能优化与决策、人工智能、数据挖掘及智能交通方面的研究。

are prospected.

Key words: Interpretability evaluation; Artificial intelligence interpretability; Subjective evaluation; Objective evaluation; Evaluation method; Neural network; Deep learning

0 引言

近年来,从智能推荐系统、智能电子邮件过滤到自动驾驶,人工智能的应用越来越广泛,其面对的问题越来越复杂,机器学习模型的复杂程度越来越高。为了获得更好的性能,机器学习模型的参数数量可达上亿量级。性能提高的代价是模型透明性的缺失,人们无法理解模型的决策逻辑,因而无法信任其做出的决策。为此,学术界在 2004 年提出了可解释人工智能(Explainable Artificial Intelligence,XAI)[1]的概念。为使用户理解、信任和管理新一代人工智能系统,2016 年 10 月,美国国防部高级研究计划局(Defense Advanced Research Projects Agency,DARPA)启动了"可解释的人工智能"[2]项目。

XAI 可解释性对于机器学习研究和用户至关重要,一方面,可解释性是衡量模型决策合理性和算法安全性的基础;另一方面,缺乏可解释性会阻碍人工智能在一些关键领域的落地应用,例如,信用评分、医疗保健、自动驾驶以及军事领域。2018 年 5 月,欧盟出台了《通用数据保护条例》,首次引入了关于自动决策的条款,要求为用户提供获得解释权的权利。

人工智能可解释性研究作为人工智能领域中一个新兴的分支,目前已出现了多种解释技术,包括 LIME[3]、SHAP[4]、显著图(Saliency Map)[5]及注意力机制(Attention Mechanism)[6-7]等,但如何评估 XAI 的可解释性还没有形成共识[8],XAI 可解释性评估研究还处于早期发展阶段。可解释性评估工作需要针对具体的应用场景、解释模型和用户,需要考虑主客观因素,因而难以形成普遍认可的科学评估体系[9]。对于可靠性要求高的领域[10],只有科学评估 XAI 的可解释性,才能促进 XAI 在这些领域中应用。

XAI 的可解释性评估涉及人机交互(Human Computer Interaction,HCI)、人类科学(Human Science)、可视化(Visualization)、机器学习(Machine Learning)和数据科学(Data Science)等多领域的交叉[11],具有较大挑战性。本文综述了当前 XAI 可解释性评估方法。首先,介绍了 XAI 的可解释性及其评估的基本概念和分类;然后,总结和梳理了 XAI 可解释性的一些特性;进而,从可解释评估方法和可解释性评估框架两方面,分析和讨论了可解释性评估工作;最后,总结了当前人工智能可解释性评估研究的不足,并展望了其未来发展方向。

1 XAI 的可解释性及其评估

人工智能可解释性研究可追溯到 1991 年,G. D. Garson 等[12]提出了从敏感性分析的角度,分析和解释机器学习模型的预测结果。2004 年,首次提出了 XAI 这一术语。目前还没有关于人工智能可解释性标准的且普遍接受的定义,对于如何评估可解释性也没有标准的普遍认可的体系[9]。

从可解释性的角度,机器学习模型可分为透明盒(Transparent Box)模型与黑盒(Black Box)模型[13]。透明盒模型的决策机理是透明的,本身具有可解释性,例如,决策树模型、规则模型及线性模型等。黑盒模型的决策机理不透明,用户无法得知决策的依据,例如,神经网络、支持向量机等。目前的解释技术主要用于解释黑盒模型。

1.1 可解释性人工智能(XAI)

可解释性涉及多领域交叉,其定义需融合不同领域对解释的理解和需求,因而难以形成统一定义[14]。目前,可解释性定义主要从技术和用户两方面考虑,如 DARPA 定义 XAI 为"XAI 向用户提供解释,用以使用户理解系统的整体优势和不足,理解系统在未来或不同情况下的行为,并可能允许用户纠正系统的错误"[2]。文献[14]指出,还需考虑可解释性功能性需求,如公平性、因果性等,因此,从解释受众和解释功能的角度出发,定义 XAI 为:"XAI 能够提供其功能的细节和原因,使其功能对用户而言是清晰的或容易理解的",说明 XAI 在不同应用场景下的功能性目标需考虑具体的用户。

1.2 XAI 的可解释性评估

许多文献强调了可解释性评估的必要性和评估指标的缺乏[11,14-15]。文献[11]调研了 381 篇 XAI 相关文献,其中只有 5%的研究尝试评估 XAI 的可解释性。类似地,文献[16]发现,78%的关于决策支持系统解释的研究缺乏结构化的评估工作。只有系统科学地评估 XAI 的可解释性,才能提高 XAI

的可靠性和实用性，推动 XAI 的研究和应用。具体来说，可解释性评估的目标包括：1）为解释方法之间的比较提供科学、有效的评价标准；2）评价 XAI 是否实现了预期的可解释性目标[17]。

文献[18]将可解释性评估分为三类：1）基于应用（Application-ground）：在实际应用场景下，由用户（尤其是专业人员）评估可解释性；2）基于人（Human-ground）：设计简化的任务，利用基于用户实验获得的评价指标来评估解释性；3）基于功能（Function-ground）：无需用户参与，通过可解释代理模型或量化指标来评估可解释性，例如，决策树的深度、模型预测的不确定性等。基于应用的评估是最理想的，因为它评估了 XAI 在实际应用中用户对解释的反馈，然而用户的参与导致评估成本较高，且评估结果依赖于所选的专业人员的领域。基于功能的评估无需人的参与，但其评估结果的有效性难以保证，因为量化指标可能并不能很好地反映可解释性。基于人的评估是一个折中方法，比基于应用的评估成本低，但比基于功能的评估更有效。

文献[19]根据用户是否参与评估，将可解释性评估分为主观、客观评估两类。主观评估利用用户或专家反馈来评估 XAI 可解释性；客观评估利用客观评估指标来量化评估可解释性。以上基于应用的评估和基于人的评估属于主观评估，而基于功能的评估属于客观评估。

2　可解释性的特性

可解释性的特性是指可解释性应具备的特性，用于评估和比较 XAI 的可解释性。文献[20]从解释方法（Explanation Methods）和个体解释（Individual Explanations）两方面总结了可解释性的特性。

解释方法的特性包括 4 个方面：1）表达能力（Expressive Power）：是指解释方法生成的解释的形式，如 if-then 规则、模糊逻辑、直方图、决策树、线性模型、有限制的自然语言等；2）半透明性（Translucency）：是指解释方法对机器学习模型内部工作原理的依赖性，例如，模型无关的解释方法与模型内部工作原理无关，其半透明性为零；3）可移植性（Portability）：是指解释方法可应用的范围，高半透明性的解释方法的可移植性低；4）算法复杂性（Algorithmic Complexity）：是指解释方法的计算复杂性。此外，解释方法的稳定性[21]、鲁棒性[22]、敏感性[23]等也是评估可解释性的重要指标。

个体解释是指解释方法生成的解释内容，其特性包括 9 方面：1）准确性（Accuracy）：是指解释对未知实例预测的准确性，例如：规则形式的解释的预测准确性；2）保真度（Fidelity）：是指解释是否反映模型真实预测行为，对于局部解释，保真度是指解释是否很好地反映模型在某一实例附近的预测行为；3）一致性（Consistency）：是指对同一任务（如数据集）训练得到的两个模型的解释的相似程度，如果这两个模型对相似实例的解释越相似，则一致性越高；4）稳定性（Stability）：是指对相似实例生成的解释的相似程度，与一致性不同，稳定性是指同一模型对相似实例解释的相似性；5）可理解性（Comprehensibility）：是指用户对解释的理解程度，是偏主观的特性；6）确定性（Certainty）：是指解释能否反映模型预测的确定性，许多模型只提供预测结果，而不提供模型预测正确性的置信度；7）重要性（Importance）：是指解释能否反映其所包含的信息（如特征）间的重要性程度，例如，规则集形式的解释中各条规则的重要程度；8）新颖性（Novelty）：是指解释能否反映来自新区域（远离训练数据分布的区域）的解释实例；9）代表性（Representativeness）：是指解释覆盖实例程度，解释可能覆盖整个模型行为，或只能解释部分实例。

由于解释是面向用户的，因此解释需要以用户能理解的形式呈现。文献[24]从用户角度出发，分析了人容易理解的解释的特性，主要包括 7 方面：1）对比性（Contrastiveness）：又称反事实忠实性（Counterfactual Faithfulness），人们倾向于反事实思考，通常会问为什么不是其他预测结果。好的解释应能突出事实和相关事件之间的差异性。2）选择性（Selectivity）：人们往往并不期望解释能涵盖模型预测的完整原因，而更倾向于从多个可能的原因中选择主要原因作为解释。因此，解释方法应能明确模型预测结果的主要原因。3）社会性（Sociality）：解释需要解释者和被解释者之间的互动，因此需考虑社会环境和目标用户，以适用于不同领域和环境。4）异常关注（Focus on the Abnormal）：人们更关注异常事件（实例）发生的原因，分析异常事件的原因可提供更好的解释。5）真实性（Truth）：解释应反映真实的决策逻辑。6）先验知识一致性（Consistent with Prior Knowledge）：人们更倾向于忽略与其先验知识不一致的信息。7）普遍性（Generality）：好的解释应能应用于大多数实例。

一些文献从其他角度分析了解释的特性，例如：

文献[14]从解释目标的角度分析了解释的可信性(Trustworthiness)、因果性(Causality)、可转移性(Transferability)、信息性(Informativeness)、置信能力(Confidence)、公平性(Fairness)、可访问性(Accessibility)、互动性(Interactivity)及隐私意识(Privacy Awareness)。文献[19]从解释概念的角度分析了因果性、完整性(Completeness)等 36 个相关解释特性。

可解释性特性可用于评估和比较可解释性水平,但有些特性的量化方法尚不明确,这是可解释性评估工作的重要挑战之一;另一个挑战是:"好"解释应满足什么特性方面还未形成共识[25],目前研究主要从直觉出发,分析"好"解释应满足的特性[24]。如何结合具体应用场景、评估目标、用户类型,合理地选择、组合、量化上述特性,对可解释性评估至关重要。

3 可解释性评估方法

根据是否需要用户参与可解释性评估,可将评估方法分为主观评估方法和客观评估方法。

3.1 主观评估方法

若解释有助于用户建立 XAI 的决策逻辑的心理模型,则该解释是有效的[19]。大多数可解释性的评估工作以用户为中心进行评估,基于用户的反馈评估可解释性。评估过程一般涉及两类用户[26]:普通用户和专家用户,普通用户是指没有 AI 专业知识或技能的用户,专家用户是指具有一定专业水平的数据专家和 AI 专家等。文献[27]分析了 653 篇 XAI 文献,将主观评估研究分为定性研究、定量研究、定性和定量结合研究。

3.1.1 定性评估

定性评估基于开放式问题,通过采访、问卷调查、量表分析等方式评估解释的有用性、用户满意度和信任等[17]。DARPA 的 XAI 项目中,R. R. Hoffman 等[15]的工作是 XAI 可解释性定性评估的代表,其通过建立 XAI 解释过程的概念模型,从解释的优良、用户满意度、用户心理模型、用户信任与依赖以及好奇心的影响等方面评估可解释性,并对用户实验设计给出具体建议和示例,示例包括:1)设计一组询问用户对解释效果的感受的问题,评估解释对用户好奇心的影响,如"我想知道我是否正确理解这个人工智能系统";2)设计 5 分利克特量表(Likert Scale)评估用户满意度和用户信任与依赖,量表问题如:"我喜欢用该 XAI 系统来决策"。心理模型是指用户对 XAI 系统的理解,该评估工作列出 11 种提取用户心理模型的方法,并分析了各方法的优缺点,其中典型的方法包括:1)预测任务(Prediction Task):用户对给定的样本进行预测并解释预测的原因;2)自解释任务(Self-explanation Task):用户在完成指定任务后,描述自己的理解;3)有声思考问题解决任务(Think-aloud Problem Solving Task):用户在完成任务的过程中,说出自己的想法、感受、意见等。

文献[28]在众包平台上召集了 120 名用户,每个用户完成 4 分利克特量表和 5 分利克特量表,从有效性、效率、说服力、满意度、可审查性、透明性和信任 7 个方面来评估一个用于推荐领域的 XAI 的可解释性。

解释方法有效性验证方面,一般通过小规模用户实验来验证,例如,文献[29]设计了一些描述题、选择题、判断题,请 47 位学过机器学习课程的学生回答,通过统计用户预测的准确度来验证解释方法的有效性。文献[30]通过 70 位学习机器学习课程的本科学生在线用户实验,以验证解释方法的有效性。

3.1.2 定性与定量评估结合

主观评估的定量研究以封闭式问题为基础,计算任务完成的效果[19],例如:计算人机任务性能测试的准确性、反应时间等指标[17]。

一些评估工作结合定性和定量分析来评估和比较可解释性[26],用户完成预测任务后,除定量分析预测准确性、所用时长等指标外,还需用户完成填空、量表等定性调查,以进一步分析用户满意度、理解性等。文献[31]为评估医学领域中 LIME 解释方法生成的 XAI 的解释性,计算 XAI 预测结果中医生赞同的比例、XAI 的解释与医生的解释的相似性,以评估解释的准确性和充分性,同时请医生完成 2 份 5 分利克特量表来评估医生对解释的满意与信任程度。文献[32]为了研究复杂性对 XAI 的解释性的影响,在众包平台上召集 900 名用户,每个用户完成 3 个任务和 1 份 5 分利克特量表,通过计算任务完成时间、准确度、解释的使用难度来评估具有不同复杂性的 XAI 的可解释性。一些研究利用辅助专业设备来评估可解释性,例如,文献[33]和文献[34]在用户实验中利用眼球追踪(Eye Tracker)设备来判断用户的注意力,同时利用量表来评估解释的可信性和可依赖性。

3.1.3 总结和分析

如前所述,当前有很多以用户为中心的主观评估

方法，但还没有用户实验设计的统一标准[16,21]。一些研究工作提出用户实验设计的建议[35]，例如：在DARPA 的 XAI 项目中，S. T. Muller 等[36]围绕解释的类型、实验设计、用户模型的可靠性、用户信任等 9 个方面，调研和总结了从 1987 年至 2018 年间 XAI 可解释性主观评估工作，提出了一组以用户为中心 XAI 设计原则[37]，为可解释性主观评估方法提供指导。

由于解释是面向用户的[38]，因此，用户实验是一种高效且直接的可解释性评估方式。采用这种方式时，解释的有效性依赖于用户认知能力和解释的应用环境。合理的用户实验可以评估解释方法的实际应用效果[39]。然而，用户实验具有随机性和主观性，不同用户可能倾向于不同类型和程度的解释，用户认知的局限性也可能导致用户对解释的合理性做出错误判断。此外，用户实验是基于“好的解释能提升用户表现”的假设，然而此假设成立的条件有待进一步探究。文献[39]中，一项涉及 3800 名参与者的研究表明，清晰、详细的解释反而会损害用户表现。文献[40]指出，基于用户反馈的用户实验可能会导致研究人员过于追求设计一个更有说服力的解释方法，而不是设计一个与解释对象一致的解释方法。

3.2　客观评估方法

客观评估无需用户参与，利用客观指标来评估 XAI 的可解释性。可解释性虽然涉及人的主观感受，但也可以通过量化评估指标实现客观评估[23]。客观评估方法能快速地、自动地评估 XAI 的可解释性[41]。相比主观评估，客观评估工作相对较少。文献[19]调研了 70 篇可解释性评估文献，其中客观评估工作的占比为 38.02%。

客观评估研究可解释特性的量化方法。本文总结了一些常用的量化特性，见表 1。

表 1　XAI 可解释性的客观评估工作

Tab. 1　Objective evaluation works for XAI's interpretability

特性	含义	评估工作
稳定性	XAI 对相似/邻近样本生成解释的相似性[19]	[43][44][45][46] 47]
敏感性	XAI 对输入样本变化的敏感程度[19]	[23][42][50][51][52][53]
保真度	XAI 解释描述模型行为的准确程度，即解释与黑盒模型的一致程度[47]	[5] [23][42][46][49][54][55][56][57]
复杂性/稀疏性	XAI 解释的复杂程度	[45][46][49][57][58][61][62][63]
因果性	XAI 解释与黑盒模型预测的因果关系水平[64]	[64][65][66][67][68][69][70]
有效性	XAI 解释能准确地反映出黑盒的决策逻辑[44]	[31][41][44][45]

3.2.1　稳定性

稳定性是指 XAI 对相似/邻近样本生成解释的相似性[19]。对于同一样本或相似的样本，XAI 应产生相似的解释，若生成具有较大差异的解释，则会影响用户对 XAI 的信任。例如，自动驾驶领域中，若在行驶情况没有发生明显变化时，XAI 向用户提供几种不同的解释，则用户会对自动驾驶系统失去信任[42]。

一些评估工作用某种解释方法对相同样本进行多次重复解释，然后用解释的相似性来评估该解释方法的稳定性。例如，文献[43]、[44]及[45]中，利用解释中包含的特征集之间的相似性来评估解释的相似性；文献[43]和[46]中，利用同一样本生成的解释中特征集之间的 Jaccard 系数来评估解释的稳定性。文献[47]计算给定样本 x_i 的邻近样本中相对离散利普希茨常数(Relative Discrete Lipschitz Constant)最大的样本 $\hat{L}(x_i)$，用其相对离散利普希茨常数来表示解释的稳定性

$$\hat{L}(x_i)=\arg\max_{x_j\in B_\varepsilon(x_i)}\frac{\|f_{\text{expl}}(x_i)-f_{expl}(x_j)\|_2}{\|h(x_i)-h(x_j)\|_2} \quad (1)$$

其中，f_{expl} 为解释方法；B_ε 为邻近样本集合；h 为聚合函数。一般来说，解释内容的基本单元是样本中的变量(特征或像素)，当该变量为高度、面积等用户可理解的信息时，$h(x_i)=x_i$；当该变量为像素等用户难以理解的信息时，解释内容的基本单元是用户可理解的高阶变量，如像素块，此时，$h(x_i)$ 为由高阶变量组成的样本。

3.2.2　敏感性

敏感性是指 XAI 对输入样本变化的敏感程度[19]。低敏感性的 XAI 通常更受欢迎，因为其具有较强的抗干扰性，当输入样本受到与模型预测无关的微小扰动时，XAI 的解释不会产生明显变化。A. Ghorbani 等[48]

的研究表明,具有高敏感性的解释可能更容易受到对抗攻击。文献[23]提出最大敏感性指标 $SENS_{MAX}$(Max-sensitivity)来评估解释的敏感性,该指标计算邻近样本解释间的最大距离作为敏感性

$$SENS_{MAX}(\boldsymbol{\Phi},\boldsymbol{f},\boldsymbol{x},r)=\max_{\|\boldsymbol{y}-\boldsymbol{x}\|\leqslant r}\|\boldsymbol{\Phi}(\boldsymbol{f},\boldsymbol{y})-\boldsymbol{\Phi}(\boldsymbol{f},\boldsymbol{x})\| \tag{2}$$

其中,r 是一个预定义的参数,表示扰动范围;$\boldsymbol{x}$ 为输入样本;$\boldsymbol{f}$ 表示黑盒模型;$\boldsymbol{\Phi}$ 表示解释方法。文献[49]提出最大敏感性和平均敏感性两个指标来计算敏感性,这两个指标选取与输入样本预测结果相同的邻近样本。最大敏感性按式(3)计算

$$\mu_M(\boldsymbol{f},\boldsymbol{g},r;\boldsymbol{x})=\max_{\boldsymbol{z}\in N_r}D(\boldsymbol{g}(\boldsymbol{f},\boldsymbol{x}),\boldsymbol{g}(\boldsymbol{f},\boldsymbol{z})) \tag{3}$$

其中,D 为距离函数;$\boldsymbol{f}$ 表示黑盒模型;$\boldsymbol{g}$ 表示解释方法;$\boldsymbol{x}$ 为输入样本;N_r 表示与 $\boldsymbol{x}$ 距离为 r 的样本集合中与 $\boldsymbol{x}$ 预测结果相同的所有样本。

一些解释方法关注于解释的高敏感性。例如,基于显著图的解释方法,通过计算输入特征对模型输出影响的重要性分数来解释黑盒模型[14],文献[51]提出 Sensitivity-n 指标,通过扰动来量化具有相同重要性水平的不同特征被移除时对模型预测结果的影响,以此分析解释对重要特征的敏感性。文献[52]和文献[53]利用 Spearman 秩相关(Spearman Rank Correlation)、梯度直方图的 Pearson 相关(Pearson Correlation of the Histogram of Gradients)、结构相似指数(Structural Similarity Index)指标分别评估解释方法对模型参数和超参数的敏感性。

3.2.3 保真度

保真度是指解释描述模型行为的准确程度,即解释与黑盒模型的一致程度。保真的解释一方面应能提供足够的信息来描述从样本输入到模型预测过程中模型的完整行为,另一方面应能真实反映模型行为[17]。一些研究工作通过计算解释的预测结果与黑盒模型预测结果间的偏差来评估保真度。例如:文献[46]在解释样本的邻近样本集上,计算黑盒模型预测与解释预测的 F1 分数来评估解释的保真度。除 F1 分数外,Accuracy[54]、AUC 分数[55]也是常用的指标。更多的研究工作基于样本的扰动来评估解释的保真度,如文献[23]、[42]、[49]等。文献[56]利用均方根误差(Root Mean Square Error, RMSE)计算预测偏差来评估保真度,且基于样本扰动进一步评估保真度,从 3 方面测试解释所包含的特征是否真实地影响黑盒模型的行为:1)特征推断测试(Feature Deduction Test):通过抹去测试样本中解释所包含的特征对应的特征值来构造新样本,观察新样本的模型预测结果是否改变,若改变,则通过测试;2)特征增强测试(Feature Augmentation Test):从与测试样本 $\boldsymbol{x}$ 不同类别的样本集中随机挑选一个样本 $\boldsymbol{y}$,将测试样本中解释所包含的特征对应的特征值替换 $\boldsymbol{y}$ 的特征值来构建新样本,观察新样本的预测结果是否与测试样本的预测结果相同,若相同,则通过测试;3)综合测试(Synthetic Test):保留测试样本中解释所包含的特征对应的特征值,并将其他特征进行随机赋值来构建新样本,观察新样本的预测结果是否与测试样本相同,若相同,则通过测试。在测试集上分别进行以上三种测试,计算各测试中通过测试的样本所占的比例来评估解释的保真度。与上述特征推断测试类似,文献[5]和文献[57]通过对图像进行特征遮挡来计算解释的保真度。

此外,复杂性、因果性、有效性等也是客观评估中普遍关注的特性。还有一些客观评估工作只针对特定解释方法或黑盒模型。例如,文献[58]利用决策树代理模型来解释卷积神经网络(Convolutional Neural Network, CNN),通过控制和调节 CNN 来计算代理模型的特征信息增益、特征稀疏性、特征完整性、决策树的预测准确性、完整性以量化评估可解释性,该评估方法涉及 CNN 的调节和控制,是一种针对特定黑盒模型的评估方法。类似的评估工作见文献[57]、[59]、[60],不再赘述。

3.2.4 总结和分析

客观评估方法量化了可解释性的特性,能快捷地评估 XAI 可解释性。然而,由于解释的特性通常是概念性的,且解释方法、解释形式、评估目标具有多样性,因此即使针对同一特性,其量化方法也不尽相同。此外,一些评估方法受限于特定黑盒模型和应用场景,不具有通用性。对于一些重要特性,诸如解释确定性、公平性及隐私意识等,仍缺乏可靠的量化评估方法。

4 可解释性评估框架

XAI 系统整个生命周期中,从最初需求确定到设计和开发,再到系统使用,都需要解释。将可解释性评估与 XAI 设计和开发过程结合,研究 XAI 可解释性评估的框架具有重要意义[71]。

XAI 系统在不同阶段具有不同设计目标,一个观点是考虑 XAI 设计目标和评价方法之间的依赖

关系[26]，根据 XAI 设计过程和解释目标来选择合适的评估方法，从而对 XAI 可解释性进行整体评估。文献[26]构建了一个 XAI 系统设计与评估的嵌套框架，如图 1 所示。XAI 系统设计中，需根据 XAI 设计目标来确定每个框架层的可解释性要求。这些要求根据用户需求确定，包括法规、法律、安全标准等，随后选择合适的评价方法来评估可解释性是否达到预期要求。该框架结构包括：

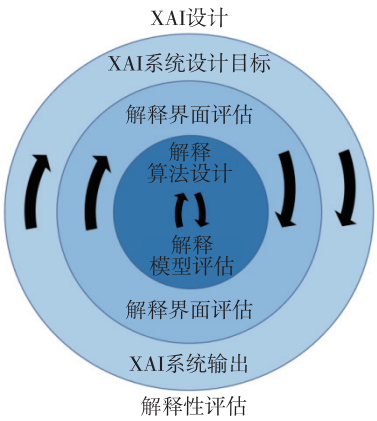

图 1　XAI 设计与评估框架[26]

Fig. 1　Design and evaluation framework of XAI [26]

1）外层：XAI 系统级设计目标层，其解释的要求包括：a）确定解释的目的；b）结合应用场景和目标用户类型，选择合适的解释内容；c）利用XAI系统的输出来定性和定量评估 XAI 系统目标的实现情况。具体评估方法取决于设计目标、应用范围和目标用户，例如：用户信任和依赖[72-73]、人机任务性能[74]、用户意识[75]等。

2）中间层：解释形式和界面设计，目的是以用户可理解的、满意的方式呈现解释内容。采用用户对解释的理解、用户对解释的满意度、用户心理模型等主观评估方法，以改善解释界面设计。

3）内层：解释算法设计层。XAI 利用解释技术来解释黑盒模型，而各种解释技术具有各自优缺点和应用范围。因此，只有选取合适的解释技术，才能向用户提供有用且值得信赖的解释。可以通过定量评估 XAI 的可信性[76]、保真度等指标来评估内层解释算法的有效性。

DARPA 的 XAI 项目[2]基于 XAI 解释过程的概念模型来评估 XAI 的可解释性，如图 2 所示。概念模型包括：用户、XAI 的解释、用户心理模型（User's Mental Model）及用户系统任务表现（User-System Task Performance）4 个模块。模块之间的关系为：用户收到 XAI 提供的解释，解释用于建立和完善用户的心理模型，完善的心理模型可提高用户系统任务表现。优良的解释可帮助用户构建良好的心理模型，而良好的解释与心理模型能使用户信任与依赖 XAI 的决策。针对概念模型中 4 个模块，将可解释评估划分为以下五方面：

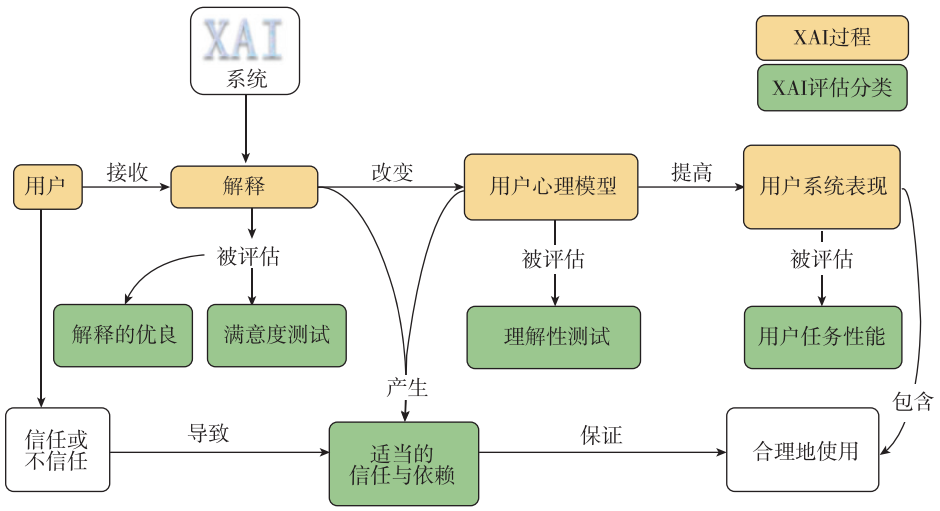

图 2　解释过程和解释效果测量类别的初步模型[2]

Fig. 2　Initial model of the explanation process and explanation effectiveness measurement categories[2]

1）解释优良性（Explanation Goodness）：评估解释是否满足优良的解释应具备的特性；

2）满意度测试（Test of Satisfaction）：用户对解释的主观评价，包括解释的完整性、有用性、准确性

和满意度等；

3)理解性测试(Test of Understanding)：测试用户理解 XAI 系统的程度以及用户在新场景下预测系统决策/行为的能力；

4)用户任务性能(User Task Performance)：用户能成功地执行 XAI 系统所支持的任务；

5)合理的信任与信赖(Appropriate Trust and Reliance)：用户能合理地判断 XAI 系统提供的解释和预测，并适当地信任与依赖该系统。

不同类型用户对解释需求和理解程度存在差异，因此，多数主观评估工作从用户类型角度设计评估目标，而文献[77]从用户所需求信息的角度出发，根据解释中包含信息的必要性，构建一个三层框架来分析 XAI 的设计与评估：解释对 XAI 系统当前行为的感知、解释对 XAI 行为或决策原因的感知、解释对 XAI 行为的反事实分析或预测。

可解释性评估框架为 XAI 可解释性评估提供指导思路，能够从多方位整体评估 XAI 的可解释性，发现 XAI 可解释性的缺陷，有助于设计解释性更好的 XAI，使用户能够系统、全面地理解 XAI。当前虽然已有一些可解释性评估框架，但这些框架的合理性和实用性还有待于进一步的实际应用验证。此外，这些框架只提供了可解释性评估的指导思路，没有提供具体的评估方法，因此在使用中需要在框架指导下，结合具体的 XAI 系统，选取合适的评估方法与指标。

表 2 总结和比较了主观评估方法、客观评估方法以及可解释性评估框架的优缺点。

表 2 XAI 可解释性评估方法的对比

Tab. 2 Comparisons of evaluation methods for XAI's interpretability

评估方法	优点	缺点
主观评估方法	借助合理的用户实验，可有效地评估解释方法的实际应用效果	缺乏用户实验设计的统一标准；具有主观性，受用户认知程度的影响；用户实验所基于的假设的成立条件有待进一步探究
客观评估方法	无需用户参与，具有客观性，能快速地、自动地评估 XAI 的可解释性	受限于特定黑盒模型和应用场景，不具有通用性；一些可解释性的特性难以客观量化
可解释性评估框架	多方位整体评估 XAI 的可解释性，发现 XAI 可解释性的缺陷，有助于设计解释性更好的 XAI，使用户能够系统、全面地理解 XAI	框架的合理性和实用性还有待于进一步实际应用验证；只提供了可解释性评估的指导思路，没有提供具体的评估方法

5 总结和展望

随着 XAI 的快速发展，XAI 可解释性评估研究得到越来越多的关注。本文综述了 XAI 可解释性评估的研究进展。首先，分析了可解释性应具备的特性，这些特性可用于评估和比较 XAI 的解释性。然后，从主观和客观评估两方面总结了当前可解释评估方法。最后，综述了一些可解释性评估框架。

XAI 可解释评估研究仍然处于早期发展阶段，一些研究工作有待进一步开展，未来的研究方向包括：

1)可解释性的客观评估方法。相较于主观评估方法，客观评估方法较少，这是因为：a)有些可解释的特性是概念性的，与用户主观感受相关(如满意度)，难以客观量化；b) 有些可解释的特性目前还缺乏可靠的量化方法。客观评估可以实现 XAI 的快速、自动评估，避免主观评估成本高的不足，是可解释性评估的未来发展方向。

2)可解释性评估的统一标准。可解释性评估标准需考虑多方面因素。一方面，不同领域 XAI 的评估目标不同，不同类型的用户具有不同的解释需求，因此需结合具体应用领域和用户类型来划分可解释评估工作；另一方面，XAI 设计者或用户可能不清楚需要何种类型、何种程度的解释，因此，需提供可解释性评估列表，引导 XAI 的可解释性评估向着规范化方向发展。

3)可解释性评估方法比较研究。目前已存在多种可解释性评估方法，这些方法各有优缺点，但鲜有研究比较这些评估方法的评估效果和适用场景。可解释评估方法的比较研究，对于 XAI 设计者和用户选取合适的评估方法来评估 XAI 的可解释性具有重要意义。

4)可解释性的系统性评估方法。可解释性评估需要融入 XAI 系统整个生命周期中，从多角度评估 XAI 系统的可解释性。虽然已有一些可解释性评估框架，但这些框架是概念性的，缺乏具体的评

估细节和应用案例。深入研究和完善可解释性评估框架,对系统评估 XAI 的可解释性具有重要意义。

5)可解释性在安全方面的评估。解释可能会给 XAI 和用户带来安全隐患:a)解释方法往往会揭示底层模型和训练数据信息,其展示的信息可能包含模型和用户信息,由此导致隐私泄露,因此,需要评估 XAI 解释的隐私性;b)解释中包含的信息可能会被恶意利用,以此发现模型漏洞和脆弱点,实施对 XAI 的恶意攻击,因此解释需要考虑安全性因素,需要评估 XAI 解释的安全性。

参考文献

[1] VanLent M, Fisher W, Mancuso M. An explainable artificial intelligence system for small-unit tactical behavior[C]// Proceedings of 16th Conference on Innovative Applications of Artifical Intelligence. AAAI Press, 2004: 900-907.

[2] Gunning D, Aha D W. DARPA's explainable artificial intelligence program[J]. AI Magazine, 2019, 40(2): 44-58.

[3] Ribeiro M T, Singh S, Guestrin C. "Why Should I Trust You?": explaining the predictions of any classifier[C]// Proceedings of 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. NY, USA:ACM, 2016: 1135-1144.

[4] Lundberg S M, Lee S-I. A unified approach to interpreting model predictions[C]// Proceedings of 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2017: 4768-4777.

[5] Selvaraju R R, Cogswell M, Das A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization[C]// Proceedings of 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 618-626.

[6] Sutskever I, Vinyals O, Le Q V, et al. Learning phrase representation using RNN encoder-decoder for statistical machine translation[C]// Proceedings of Empherical Methods in Natural Language Processing, 2014: 1724-1734.

[7] 周勇, 王瀚正, 赵佳琦, 等. 基于可解释注意力部件模型的行人重识别方法[J]. 自动化学报, 2020, 41(x): 1-16.
Zhou Yong, Wang Hanzheng, Zhao Jiaqi, et al. Interpretable attention part model for person re-identification[J]. Acta Automatica Sinica, 2020, 41(x): 1-16(in Chinese).

[8] Sokol K, Flach P. Explainability fact sheets: a framework for systematic assessment of explainable approaches[C]// Proceedings of 2020 Conference on Fairness, Accountability, and Transparency, 2020: 56-67.

[9] 纪守领, 李进锋, 杜天宇, 等. 机器学习模型可解释性方法、应用与安全研究综述[J]. 计算机研究与发展, 2019, 56(10): 2071-2096.
Ji Shouling, Li Jinfeng, Du Tianyu, et al. Survey on techniques, applications and security of machine learning interpretability[J]. Journal of Computer Research and Development, 2019, 56(10): 2071-2096(in Chinese).

[10] Jesus S, Belém C, Balayan V, et al. How can I choose an explainer? An application-grounded evaluation of post-hoc explanations[C]// Proceedings of 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021: 805-815.

[11] Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)[J]. IEEE Access, 2018, 6: 52138-52160.

[12] Garson G D. Interpreting neural-network connection weights[J]. AI Expert, 1991, 6(4): 46-51.

[13] Guidotti R, Monreale A, Ruggieri S, et al. A survey of methods for explaining black box models[J]. ACM Computing Surveys, 2018, 51(5): 1-42.

[14] Arrieta A B, Díaz-Rodríguez N, Ser J D, et al. Explainable Artificial Intelligence(XAI): concepts, taxonomies, opportunities and challenges toward responsible AI[J]. Information Fusion, 2020, 58: 82-115.

[15] Hoffman R R, Mueller S T, Klein G, et al. Metrics for explainable AI: challenges and prospects[J]. ArXiv e-prints, arXiv:1812.04608, 2018.

[16] Nunes I, Jannach D. A systematic review and taxonomy of explanations in decision support and recommender systems[J]. User Modeling and User-Adapted Interaction, 2017, 27(3): 393-444.

[17] Markus A F, Kors J A, Rijnbeek P R. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies[J]. Journal of Biomedical Informatics, 2021, 113: 1-11.

[18] Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning[J]. arXiv e-prints, arXiv:1702.08608, 2017.

[19] Vilone G, Longo L. Notions of explainability and evaluation approaches for explainable artificial intelli-

gence[J]. Information Fusion, 2021, 76: 89-106.

[20] Robnik-Šikonja M, Bohanec M. Perturbation-based explanations of prediction models[M]. Human and Machine Learning, Springer, Cham, 2018: 159-175.

[21] Rosenfeld A. Better metrics for evaluating explainable artificial intelligence[C]// Proceedings of 20th International Conference on Autonomous Agents and Multiagent Systems, 2021: 45-50.

[22] Alvarez-Melis D, Jaakkola T S. On the robustness of interpretability methods[C]// Proceedings of 2018 ICML Workshop on Human Interpretability in Machine Learning, 2018: 66-71.

[23] Yeh C K, Hsieh C-Y, Suggala A, et al. On the (In) fidelity and sensitivity of explanations[J]. Advances in Neural Information Processing Systems, 2019, 32.

[24] Miller T. Explanation in artificial intelligence: insights from the social sciences[J]. Artificial Intelligence, 2019, 267: 1-38.

[25] Wang X, Yin M. Are explanations helpful? A comparative study of the effects of explanations in AI-assisted decision-making[C]// Proceedings of 26th International Conference on Intelligent User Interfaces. College Station TX USA: ACM, 2021: 318-328.

[26] Mohseni S, Zarei N, Ragan E D. A multidisciplinary survey and framework for design and evaluation of explainable AI systems[J]. ACM Transactions on Intelligent Systems and Technology, 2021, 11(3-4): 1-45.

[27] Chromik M, Schuessler M. A taxonomy for human subject evaluation of black-box explanations in XAI [C]// Proceedings of ExSS-ATEC @ IUI, 2020, 94.

[28] Balog K, Radlinski F. Measuring recommendation explanation quality: the conflicting goals of explanations[C]// Proceedings of 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. Virtual Event China: ACM, 2020: 329-338.

[29] Lakkaraju H, Bach S H, Leskovec J. Interpretable decision sets: a joint framework for description and prediction[C]// Proceedings of 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: Association for Computing Machinery, 2016: 1675-1684.

[30] Wang T. Multi-value rule sets for interpretable classification with feature-efficient representations[C]//Proceedings of Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2018: 10835-10845.

[31] Barr K N, Blomberg T, Liu J, et al. Evaluating local interpretable model-agnostic explanations on clinical machine learning classification models[C]// Proceedings of 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems. Rochester, MN, USA: IEEE, 2020: 7-12.

[32] Lage I, Chen E, He J, et al. Human evaluation of models built for interpretability[C]// Proceedings of AAAI Conference on Human Computation and Crowdsourcing, 2019, 7: 59-67.

[33] Polley S, Koparde R R, Gowri A B, et al. Towards trustworthiness in the context of explainable search[C]// Proceedings of 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY: ACM, 2021: 2580-2584.

[34] Dey A, Radhakrishna C, Lima N N, et al. Evaluating reliability in explainable search[C]// Proceedings of 2021 IEEE 2nd International Conference on Human-Machine Systems. Piscataway, NJ: IEEE, 2021:1-4.

[35] Van der Waa J, Nieuwburg E, Cremers A, et al. Evaluating XAI: a comparison of rule-based and example-based explanations[J]. Artificial Intelligence, 2021, 291: 103404.

[36] Mueller S T, Hoffman R R, Clancey W, et al. Explanation in human-AI systems: a literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI[J]. arXiv preprint arXiv:1902.01876, 2019.

[37] Mueller S T, Veinott E S, Hoffman R R, et al. Principles of explanation in human-AI systems[C]// Proceedings of WS Explainable Agency in Artificial Intelligence. AAAI, 2021: 153-162.

[38] Anjomshoae S, Najjar A, Calvaresi D, et al. Explainable agents and robots: results from a systematic literature review[C]//Proceedings of 18th International Conference on Autonomous Agents and Multiagent Systems, 2019: 1078-1088.

[39] Poursabzi-Sangdeh F, Goldstein D G, Hofman J M, et al. Manipulating and measuring model interpretability[C]// Proceedings of 2021 CHI Conference on Human Factors in Computing Systems. Yokohama Japan: ACM, 2021: 1-52.

[40] Herman B. The promise and peril of human evaluation for model interpretability[C]// Proceedings of 31st Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2017.

[41] Shah S S, Sheppard J W. Evaluating explanations of convolutional neural network image classifications[C]// Proceedings of 2020 International Joint Conference on

Neural Networks. Glasgow, United Kingdom: IEEE, 2020: 1-8.

[42] Li X-H, Shi Y, Li H, et al. An experimental study of quantitative evaluations on saliency methods[C]// Proceedings of 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. Virtual Event Singapore: ACM, 2021: 3200-3208.

[43] Guidotti R, Monreale A, Giannotti F, et al. Factual and counterfactual explanations for black box decision making[J]. IEEE Intelligent Systems, 2019, 34(6): 14-23.

[44] Fan M, Wei W, Xie X, et al. Can we trust your explanations? Sanity checks for interpreters in android malware analysis[J]. IEEE Transactions on Information Forensics and Security, 2021, 16: 838-853.

[45] Warnecke A, Arp D, Wressnegger C, et al. Evaluating explanation methods for deep learning in security[C]// Proceedings of 2020 IEEE European Symposium on Security and Privacy. Genoa, Italy: IEEE, 2020: 158-174.

[46] Amparore E, Perotti A, Bajardi P. To trust or not to trust an explanation: using LEAF to evaluate local linear XAI methods[J]. PeerJ Computer Science, 2021, 7: 1-26.

[47] Alvarez M D, Jaakkola T. Towards robust interpretability with self-explaining neural networks[J]. Advances in Neural Information Processing Systems, 2018, 31.

[48] Ghorbani A, Abid A, Zou J. Interpretation of neural networks is fragile[C]// Proceedings of AAAI Conference on Artificial Intelligence, 2019, 33(1): 3681-3688.

[49] Bhatt U, Weller A, Moura J M F. Evaluating and aggregating feature-based model explanations[C]// Proceedings of 29th International Joint Conference on Artificial Intelligence, 2020: 3016-3022.

[50] Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks[C]// Proceedings of 34th International Conference on Machine Learning, 2017: 3319-3328.

[51] Ancona M, Ceolini E, Oztireli C, et al. Towards better understanding of gradient-based attribution methods for deep neural networks[C]// Proceedings of 6th International Conference on Learning Representations. Vancouver, BC, Canada: 2018.

[52] Adebayo J, Gilmer J, Muelly M, et al. Sanity checks for saliency maps[J]. Advances in Neural Information Processing Systems, 2018, 31.

[53] Bansal N, Agarwal C, Nguyen A. SAM: the sensitivity of attribution methods to hyperparameters[C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 8673-8683.

[54] Burkart N, Faller P M, Peinsipp E, et al. Batch-wise regularization of deep neural networks for interpretability[C]// Proceedings of 2020 IEEE International Conference on Multisensor Fusion and Integration. Piscataway, NJ: IEEE, 2020:216-222.

[55] Schaaf N, Huber M, Maucher J. Enhancing decision tree based interpretation of deep neural networks through L1-orthogonal regularization[C]// Proceedings of 2019 18th IEEE International Conference on Machine Learning and Applications. Piscataway, NJ:IEEE, 2019: 42-49.

[56] Guo W, Mu D, Xu J, et al. LEMNA: explaining deep learning based security applications[C]// Proceedings of 2018 ACM SIGSAC Conference on Computer and Communications Security. Toronto Canada: ACM, 2018: 364-379.

[57] Pope P E, Kolouri S, Rostami M, et al. Explainability methods for graph convolutional neural networks[C]// Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA: IEEE, 2019: 10764-10773.

[58] Fan L, Liu C, Zhou Y, et al. Interpreting and evaluating black box models in a customizable way[C]// Proceedings of 2020 IEEE International Conference on Big Data. Atlanta, GA, USA: IEEE, 2020: 5435-5440.

[59] Poppi S, Cornia M, Baraldi L, et al. Revisiting the evaluation of class activation mapping for explainability: a novel metric and experimental analysis[C]// Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Nashville, TN, USA: IEEE, 2021: 2299-2304.

[60] Bau D, Zhou B, Khosla A, et al. Network dissection: quantifying interpretability of deep visual representations [C]// Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI: IEEE, 2017: 3319-3327.

[61] Tian Y, Liu G. MANE: model-agnostic non-linear explanations for deep learning model[C]// Proceedings of 2020 IEEE World Congress on Services. Piscataway, NJ: IEEE, 2020: 33-36.

[62] Wu M, Hughes M, Parbhoo S, et al. Beyond sparsity: tree regularization of deep models for interpretability[C]// Proceedings of AAAI Conference on Artificial Intelligence. Menlo Park, CA: AAAI, 2018: 1670-1678.

[63] Chang C H, Tan S, Lengerich B, et al. How interpretable and trustworthy are gams? [C]// Proceedings of 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021: 95-105.

[64] Moraffah R, Karami M, Guo R, et al. Causal interpretability for machine learning-problems, methods and evaluation[J]. ACM SIGKDD Explorations Newsletter, 2020, 22(1): 18-33.

[65] Tian L, Alizadeh A A, Gentles A J, et al. A simple method for estimating interactions between a treatment and a large number of covariates[J]. Journal of the American Statistical Association, 2014, 109 (508): 1517-1532.

[66] Hahn P R, Murray J S, Carvalho C M. Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects (with discussion)[J]. Bayesian Analysis, 2020, 15(3): 965-1056.

[67] Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests[J]. Journal of the American Statistical Association, 2018, 113(523): 1228-1242.

[68] Louizos C, Shalit U, Mooij J, et al. Causal effect inference with deep latent-variable models[C]// Proceedings of Advances in Neural Information Processing Systems, 2017: 6447-6457.

[69] Morgan S L, Todd J J. A diagnostic routine for the detection of consequential heterogeneity of causal effects[J]. Sociological Methodology, 2008, 38(1): 231-282.

[70] Künzel S R, Sekhon J S, Bickel P J, et al. Meta-learners for estimating heterogeneous treatment effects using machine learning[J]. Proceedings of the National Academy of Sciences of the United States of America, 2019, 116 (10): 4156-4165.

[71] Wegener R, Cassens J. Intrinsic, dialogic, and impact measures of success for explainable AI[C]// Proceedings of 12th International Workshop Modelling and Reasoning in Context, 2021.

[72] Pu P, Chen L. Trust building with explanation interfaces[C]// Proceedings of International Conference on Intelligent User Interfaces, Proceedings IUI. Sydney, Australia, 2006: 93-100.

[73] Berkovsky S, Taib R, Conway D. How to recommend? User trust factors in movie recommender systems[C]// Proceedings of 22nd International Conference on Intelligent User Interfaces. New York, NY, USA:ACM, 2017: 287-300.

[74] Bansal G, Nushi B, Kamar E, et al. Beyond accuracy: the role of mental models in human-AI team performance [C]// Proceedings of AAAI Conference on Human Computation and Crowdsourcing, 2019, 7: 2-11.

[75] Kay M, Kola T, Hullman J R, et al. When (Ish) is my bus? User-centered visualizations of uncertainty in everyday, mobile predictive systems[C]// Proceedings of 2016 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: ACM, 2016: 5092-5103.

[76] Kim B, Wattenberg M, Gilmer J, et al. Interpretability beyond feature attribution: quantitative testing with concept activation vectors[C]// Proceedings of 35th International Conference on Machine Learning. Stockholm, Sweden: 2018, 6: 4186-4195.

[77] Sanneman L, Shah J A. A situation awareness-based framework for design and evaluation of explainable AI [C]// Proceedings of International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems. Springer, Cham, 2020: 94-110.

(编辑:孟彬)