doi: 10.12011/SETP2020-1536 中图法分类号: C931; TP18 文献标志码: A

人工智能决策可解释性的研究综述

孔祥维, 唐鑫泽, 王子明

(浙江大学 管理学院, 杭州 310058)

摘 要 人工智能决策的性能在某些特定领域超过了人类能力, 中国、美国等多国都颁布了人工智能发展战略和行动规划, 期望人工智能在多个领域得到落地应用. 但在人工智能决策过程中, 存在着固有算法黑盒和系统信息不透明的问题, 导致其结果正确但不可理解, 阻碍了人工智能的进一步发展. 为了人工智能的商用和普及, 对智能决策可解释性的需求越来越迫切, 需要将黑盒决策转化为透明过程, 建立起人与机器之间的信任. 本文从系统应用视角和决策收益者视角出发, 重点对人工智能决策可解释性的基本概念、模型解释方法、高风险决策应用解释和解释方法评估等四个方面的国内外相关研究进行综述, 并展望了未来研究发展趋势.

关键词 人工智能; 智能决策; 可解释性; 用户信任; 评估

A survey of explainable artificial intelligence decision

KONG Xiangwei, TANG Xinze, WANG Ziming

(School of Management, Zhejiang University, Hangzhou 310058, China)

Abstract The performance of decision making by artificial intelligence has exceeded the capability of the human being in many specific domains. Countries like China and the USA have promulgated artificial intelligence development strategies and action plans to encourage the applications of artificial intelligence. In the artificial intelligence decision-making process, the inherent black-box algorithms and opaque system information lead to highly correct but incomprehensible results, which hinder the further development of artificial intelligence. For the commercialization and popularization of artificial intelligence, the need for explainability of intelligent decision-making is becoming more and more urgent. It is necessary to study the transformation of black-box decision-making into a transparent process and establish trust between humans and machines. From the perspective of system application and decision beneficiaries, this paper focuses on the domestic and foreign-related research on four aspects: The basic concepts of explainable artificial intelligence decision-making, explanation methods of black-box models, applications of explanation methods in high-risk domains and explanation methods evaluation. Meanwhile, we state insights into future research and development trends.

Keywords artificial intelligence; intelligent decision-making; explainable artificial intelligence; user trust; evaluation

1 引言

当前人工智能在特定领域产生了超人的智能, 围绕人工智能建立的系统在医疗、交通、司法、金融、安全

收稿日期: 2020-06-15

作者简介: 通信作者: 孔祥维 (1963–), 女, 汉, 北京人, 教授, 博士生导师, 博士, 研究方向: 人工智能可解释, 非结构数据分析和管理, 数据驱动决策, 电子商务信息管理, E-mail: kongxiangwei@zju.edu.cn; 唐鑫泽 (1997–), 男, 汉, 浙江安吉人, 硕士研究生, 研究方向: 人工智能可解释, 非结构数据分析和管理, E-mail: xinzetang@outlook.com; 王子明 (1997–), 男, 汉, 山东威海人, 博士研究生, 研究方向: 人工智能可解释, 非结构数据分析和管理, E-mail: zimingwang@zju.edu.cn.

基金项目: 国家自然科学基金面上项目 (61772111); 国家自然科学基金国际合作与交流项目 (72010107002)
Foundation item: The General Program of National Natural Science Foundation of China (61772111); NSFC Projects of International Cooperation and Exchanges (72010107002)
中文引用格式: 孔祥维, 唐鑫泽, 王子明. 人工智能决策可解释性的研究综述 [J]. 系统工程理论与实践, 2021, 41(2): 524–536.
英文引用格式: Kong X W, Tang X Z, Wang Z M. A survey of explainable artificial intelligence decision[J]. Systems Engineering — Theory & Practice, 2021, 41(2): 524–536.

盒和社会等许多领域产生了巨大的价值. 人工智能赋能数字经济的需求日益增长, 智能决策具有显著的竞争优势. 2015 年国务院出台了《促进大数据发展的行动纲要》[1], 提出建立 "用数据说话、用数据决策、用数据管理、用数据创新" 的管理机制, 以实现基于数据的科学决策. 2017 年以来, 人工智能已上升为我国的国家战略, 国家《新一代人工智能发展规划》提出 "到 2030 年, 使中国成为世界主要人工智能创新中心" [2]. 与此同时, 以美国为代表的多个先进国家也都相继颁布了人工智能发展战略和行动计划, 竞相期望未来人工智能的发展将会带动经济和社会的变革, 并在引起革命性的变化中占领主导地位. 人们期望人工智能在多个领域得到落地应用, 但当前端到端的智能决策存在着不可解释的黑盒问题, 阻碍了人们对智能系统的理解和应用深化. 从用户视角和决策视角来看, 不可解释的黑盒意味着用户只能看到结果, 无法了解做出决策的原因和过程, 因而难以分辨人工智能系统某个具体行动背后的逻辑. 这样的人工智能系统难以得到决策者的信任和理解, 尤其是自主决策可能存在无法控制的风险, 因为不知道是什么在控制设计、操作和决策. 人工智能系统还可能存在固有偏差, 不定时产生虚假警报, 存在不确定的安全风险, 用户难以根据没有解释的决策而采取行动. 但对端到端的人工智能模型如何做出决策进行解释是一件非常困难的事情. 传统决策模型中的线性模型或决策树等模型直观容易理解, 但性能上存在大偏差或高方差等问题, 如线性模型容易欠拟合, 树模型容易过拟合等. 深度学习具有非常高的预测性能, 但端到端的黑盒模型难以认知, 高性能系统的整体复杂性加剧了人工智能系统自动化决策流程的不可解释性. 怎么决策? 为什么决策? 如何相信决策? 这一系列为什么问题的解决是人工智能大规模推广应用的深层次挑战性难题, 也是人们对人工智能决策信任的当务之急. 2018 年欧盟通用数据保护条例 (The EU General Data Protection Regulation, GDPR) [3] 生效, 强制要求人工智能算法具有可解释性. 只有使人工智能具有可解释其决策的能力, 获得人类信任, 才可以推动人工智能战略的实施, 使其成为新的生产力.

本文针对当前人工智能决策系统的可解释性难题, 对国内外相关研究进展进行综述. 在第二节首先对人工智能决策可解释性的概念进行阐述, 第三节对人工智能决策可解释方法进行了分类, 第四节总结了典型的高风险智能决策应用的解释, 最后对人工智能可解释方法的评估进行了归纳分析, 在此基础之上, 展望了人工智能决策可解释的未来研究趋势.

2 人工智能可解释性的概念和目标

2.1 人工智能可解释性的定义

关于人工智能可解释存有不同的观点 [4], 这是由于研究领域和关心的问题不尽相同导致的, 本研究尝试从智能决策受益者的角度对人工智能可解释性进行定义. 可解释的人工智能来自英文 Explainable Artificial Intelligence, 美国 DARPA 于 2016 年提出 Explainable Artificial Intelligence (XAI)[5] 项目之后, 开始被普遍接受. 更早出现的是可理解的机器学习 Interpretable Machine Learning (IML)[6]. 总体来说, XAI 研究范围包括了 IML. 除此之外, 有两个专业用语经常混淆: 可理解性 (interpretability) 和可解释性 (explainability). 可理解性主要是指向人类提供可理解的能力, 多指代原模型即可理解. 可解释性指使用解释作为人类和智能模型之间的接口, 作为模型代理能够被人类所理解 [7]. 倾向于指代原模型不可理解, 而需要构造事后模型进行解释. 在发表的文章中, 经常会出现这两个概念混用的情况 [8]. Arrieta 等从解释受众的角度给出了 XAI 的定义 [9]: 针对特定的听众用户, XAI 指的是可以提供细节和原因以使模型运转能够被简单、清晰地理解的技术. 宽泛地讲, XAI 技术指所有能够帮助开发者或者用户理解人工智能模型行为的技术 [10]. 包括原模型自身可解释和模型需要事后解释两种类型. 考虑用户为不同领域的决策者, 本文认为 XAI 可以定义为: 针对智能决策受益者, 人工智能可解释指可以给不同背景知识的用户, 以简单、清晰的方式, 对智能决策过程的根据和原因进行解释的方法. 目标是将黑盒人工智能决策转化为可解释的决策推断, 使用户能够理解和相信决策.

2.2 决策者视角的 XAI 目标

人工智能决策是从数据设计面向任务的端到端模型, 测试样本从模型产生决策结果, 但智能决策的成功应用和商业实施要有大规模的采纳. 人工智能决策在应用中, 若作为辅助决策, 需要和人的决策进行耦合, 若为自主决策, 必须让人完全信任. 针对人工智能可解释的发展, 不同的利益相关者对 XAI 的目标是不同的. 这些利益相关者大致分为 4 类: 学术研究者, 开发工程师, 社会管理者和终端用户 [8,11]. 学术研究者包括研究

XAI 技术和方法的计算机和人工智能相关的研究人员, 以及应用的医学专家、生物学专家等 [9]. 学术研究者提出 XAI 方法, 促进人工智能学科发展, 拓宽人类的认知边界. 开发工程师是指在工业界应用 XAI 技术的人员. 他们需要用 XAI 方法辅助进行系统的调试、监测、改进以及安全性审查等 [12,13], 成为 XAI 技术和终端用户之间的桥梁. 社会管理者包括政府、公共安全、法律等和社会大众生活息息相关领域的管理者. 他们对于 XAI 技术的需求主要从道德、法律、规章等角度出发, 关注 AI 技术可能带来的不公平、不透明、追责困难、偏见等社会问题, 希冀能够通过 XAI 阻止黑盒模型产生上述问题 [8]. 终端用户即 AI 模型和 XAI 的使用者, 终端用户包括了各行各业的使用者, 如医生、银行和保险公司、法官和警察等等. XAI 的目标是保证智能决策高性能的同时给出合理的可解释模型, 使人类用户能够理解、信任并有效地管理人工智能. 不同的利益相关者关注问题的维度不同, 对于 XAI 的目标还没有达成一致. 大多数可解释人工智能的研究只使用了研究人员对 "好" 解释的直觉 [14], 本文从决策者视角, 对 XAI 目标的多样性总结如下:

1) 可信任度 (trustworthiness). 可信任度是指, 当面对一个决策时, 人类认为模型性能的置信度高, 具有鲁棒性和稳定性, 还能产生可靠的、可信的解释.

2) 公平性 (fairness). 从社会角度出发, 提供的解释要具有保证模型决策公平的能力 [15]. XAI 可使模型能够进行道德分析, 可识别出模型中存在的偏见 [16], 这个目标对社会管理者是十分重要的.

3) 可达性 (accessibility) 和交互性 (interactivity). 可达性指构建模型时, 终端用户能够更多地参与到过程当中 [17], 交互性指终端用户可以对模型施加影响, 获得一定的控制权 [4]. 由于智能决策的终端用户受益者大部分是非计算机技术人员, XAI 能帮助他们理解和交互参与智能决策模型的运作.

4) 因果关系 (causality). 当前的机器学习模型大多揭示了数据之间的相关关系而没有揭示因果关系, 但人们更希望看到因果性解释而不是相关性解释 [9].

3 人工智能决策可解释方法研究现状

人工智能决策可解释方向在近年受到持续的关注, 从多个角度研究的文献不断涌出, 对黑盒决策解释的方式多种多样. 本文从用户需求解释的角度对当前的研究现状进行分类, 主要归纳为直观探测内部的视觉解释、从外部扰动的探索解释、用户易于理解的知识解释和因果解释四类. 从用户需求角度对解释的分类, 期望能回答人们对黑盒决策的探究, 如视觉解释用可视化显示什么特征对决策有效, 外部扰动的解释探索训练数据对决策的影响, 基于知识的解释连接人的知识和 AI 决策, 因果解释反映出决策的前因后果关系等, 本节将进行具体阐述.

3.1 视觉解释研究方法进展

对于人工智能黑盒模型的解释, 人们首先要打开黑盒, 看看黑盒里什么内容决定了模型的结果. 这种 CT 式的视觉解释方法是探寻神经元内部运行规律和工作原理最直接的途径, 将特征重要性可视化展示是目前研究深入的解释技术. 由于视觉解释的形式符合人类的认知习惯, 不熟悉人工智能原理的用户也可以直观地理解复杂人工智能系统的内部工作, 因此学术界首先开展了视觉解释方法的研究.

3.1.1 基于反向传播的解释

基于反向传播的解释方法利用深度神经网络的反向传播机制, 将模型决策的重要性信号从模型的输出层神经元逐层传播到模型的输入层, 以获得输入样本的特征重要性. 基于梯度的方法通常能够生成细粒度的相关性映射, 得到显著图的视觉解释. Zeiler 等使用网络各层的特征图作为输入, 通过反卷积技术得到可视化结果 [18]. Simonyan 等利用反向传播算法计算模型的梯度, 以输入层的梯度作为像素重要性, 得到感兴趣的解释性区域 (gradient)[19]. 之后 Springenberg 提出了通过 ReLU 非线性反向传播时操纵梯度的视觉解释方法 GuideBP 方法 [20], 反向传播时保留了梯度与激活值均为正的部分. 由于大部分深度神经网络的非线性映射采用 ReLU 函数, 它的负半轴为饱和区, 这些区域的梯度均为零, 无法揭示任何有效信息. 因此 Sundararajan 等 [21] 提出了积分梯度 (integrated gradients) 方法, 利用输入图像在一幅基准图像上的相对梯度信息反映特征重要性, 解决由于梯度消失造成的误导解释. 另外针对上述方法生成的视觉显著图通常质量较低, 解释性较差, 并存在比较多的随机噪声问题 [22], VarGrad[23] 方法利用噪声采样对图像多幅加噪副本的解释进行算术平均、方差分析, 进一步降低了随机视觉噪声. 非梯度的视觉解释方法, 如分层相关传播方

法 LRP[24]、DeepLift[25] 利用了自上而下的相关传播规则, 提高了显著图的视觉质量.

3.1.2 基于激活映射的解释

这类方法通过激活映射的线性加权组合生成显著映射, 以突出图像空间中的重要区域. 主要有类激活映射解释方法 CAM[26] 及其改进方法 Grad-CAM[27] 和 Grad-CAM++[28] 等. CAM 需要改变模型结构并重新训练, 因此局限性较大. Grad-CAM 和 Grad-CAM++ 利用模型输出对于激活映射 (特征层) 的梯度信息计算权重, 因此比较灵活, 适用于多种 CNN 模型族, 增强了空间位置信息, 关注了对预测结果起到重要影响的类. 但随机噪声多, 激活图不突出, 解释精细度不高. 继而 Guided Feature Inversion 方法 [29] 利用两步优化策略优化权重, 生成低噪的视觉解释, 但优化过程速度不如 Grad-CAM 等方法. Huang 等进一步将分割、注意力机制和模型结合, 提高了模型的解释性 [30].

3.2 基于扰动的解释研究方法进展

上述的解释方法将复杂黑盒模型内部展开, 寻求对预测结果重要的特征进行视觉解释, 需要了解具体的模型结构和参数, 称为白盒解释. 另一类方法通过扰动输入观察输出变化进行解释, 无需了解模型的参数和结构, 也称为黑盒解释.

3.2.1 局部扰动解释方法

基于局部扰动解释方法的主要思路是通过扰动输入探测模型的预测变化. LIME[31] 基于想要解释的预测值及其附近的样本, 构建局部的线性模型或其他代理模型. 对图像数据而言, LIME 输出的是连接的超像素以及权重, 其权重系数体现了决策中特征重要性, 这个解释方法可以对文本、图像或表格数据分类进行解释.

原始图像　解释电吉他　解释木吉他　解释拉布拉多

来源：Tulio Ribeiro et al（2016）

图 1　LIME 方法解释出图像中的电吉它和拉普拉多狗

Fong 等提出 Mask 方法 [32], 通过对输入图像部分区域遮挡, 找到使其预测值下降最显著的遮挡模板作为解释的显著图. Petsiuk 等用多个随机抽样的遮挡模版与原图相乘, 根据不同模版上的模型预测分数定义重要性 [33]. Ribeiro 等针对 LIME 中线性模型无法确定覆盖度的缺点, 提出锚点解释 (anchor) 的局部解释方法, 将特征和预测结果简化成 if-then 的规则逼近待解释模型的局部边界 [34]. Cui 等提出 CHIP 解释方法 [35], 通过控制网络某层每个通道的输出状态, 优化得到通道重要性矩阵, 进而与特征图加权得到解释的显著图. 针对上述解释方法无法处理特征依赖和非线性局部边界的问题, Guo 等提出了 LEMNA 方法 [36], 不需要访问模型的内部状态或结构, 但解释粗略, 进而 Wagner 等提出了细粒度的视觉解释方法 [37]. SHAP[38] 框架整合了 LIME 等 6 种方法, 利用博弈论中的 Shapley Value 作为预测特征贡献值的重要度测度.

3.2.2 影响函数解释方法

影响函数是稳健统计中的经典技术 [39], 有助于近似估计训练扰动的影响. Wojnowicz 等在广义线性模型中提出近似与影响函数相关的方法 [40]. Koh 等证明二阶优化可有效地近似影响函数进行预测解释 [41]. Anirudh 等提出了利用样本构造图进行影响样本选择的通用方法 [42]. Yeh 等提出了一种选择对模型预测有影响的训练实例代表点的新方法 [43], 在多个大型模型和图像数据集上性能超过了文献 [41] 中的影响函数方法.

3.3 基于知识的解释方法研究进展

当前的解释方法大都面向专家和学者, 对于非专业人员来说, 很难直接理解可解释方法提供的专业术语. 利用容易理解的知识辅助进行解释, 可以使解释更加用户友好. 本文将基于知识的解释方法分为两大类: 提取内部知识的解释方法和引入外部知识的解释方法. 提取内部知识的解释方法是指: 提取原模型学到的或数

据集中的知识建立解释模型; 引入外部的知识解释方法是指: 利用外部输入的知识, 如常识、概念、语料库等辅助建立解释模型.

3.3.1 内部知识提取的解释

1) 提取内部规则的解释

规则提取方法的基本思想是从原模型中提取知识的可解释符号描述, 使提出的规则能够逼近原模型的决策过程, 提供人类可理解的解释. 目前解释方法的原模型主要针对于人工神经网络模型, 解释模型通常是决策树或基于规则的模型, 包括全局规则解释和局部规则解释. 全局规则提取解释方法可分为分解法和教学法. 分解法将人工神经网络模型分离到神经元的层面进行规则提取, 将提取出来的规则整合成一个整体. 如 Sato 等人提出了 CRED 方法 [44], 用决策树从原模型中同时提取出连续型和离散型的规则; Zilke 等进一步提出了专门针对深度神经网络的 DeepRED 方法 [45], 减少了内存占用和计算时间. 教学法将神经网络模型视作一个黑盒, 只利用模型的输入和输出进行规则提取. Augusta 等人基于逆向工程提出了 RxREN 方法 [46], 反向分析输出的输入组成部分; KDRuleEx 方法生成二维的决策规则表格, 同时处理离散的和连续的输入 [47].

2) 基于知识提取的解释

用已有的知识提取方法如知识蒸馏和知识图谱也可以构造解释方法. 知识蒸馏是 Hinton 等提出的一种降低模型复杂度的模型压缩方法 [48], 基本思想是训练简单的学生模型去模仿复杂的教师模型. Liu 等将决策树作为学生模型, 将知识蒸馏的方法扩展到了多输出的回归问题上 [49], Tan 等利用知识蒸馏训练 iGAM 作为学生模型 [50], 在多个风险相关的数据集上做了验证. 知识图谱将数据集中的每个元素看作一个实体, 相邻实体之间有不同的关系, 实体到实体之间存在路径, 实体、关系和路径中蕴含着从数据中提取出来的知识. Wang 等将知识图谱和 LSTM 模型结合起来, 提出结合知识的路径循环网络 (KPRN) 模型 [51] 用于音乐推荐系统, 可以直接利用路径上的实体关系进行解释, 简单易懂且用户友好. Ma 等提出基于知识图谱的规则生成推荐模型 RuleRec[52], 在具有较高推荐召回率的同时, 还可以对推荐行为进行解释.

3.3.2 外部知识引入的解释

外部知识的引入主要考虑到人类已经有很多常识, 形成了固有的概念, 因此可以通过定义已知概念引入到智能决策解释中更容易被用户接受. 基于这一思想, Kim 等引入了概念激活向量 CAV 和重要性衡量指标 TCAV[53], 用来判断某一概念对于当前分类的重要程度, 例如概念 “条纹” 对于分类 “斑马” 的重要性. 概念关联多容易造成混淆, 人为参与引入知识的过程很难标准化和自动化, 可以通过建立知识库来存储更多的概念、语料等领域背景知识, 辅助构建基于知识的解释方法. Bau 等提出了 Network Dissection 方法, 用来识别隐藏层中的语义信息并和人类可理解的概念保持一致, 通过 IoU 分数判定一致性 [54]. Zhou 等提出结合概念激活向量和特征向量分解的方法 [55], 根据提供的概念信息以及 GradCAM 方法, 把模型的每一个预测分解为多种概念的组合解释模型的预测. 以上从视觉、加扰、知识等多个角度对人工智能决策黑盒进行的解释进行了阐述, 主要侧重于统计性解释和相关性解释, 针对的是可解释目标中的可信任性. 从决策受益者角度考虑, 终端用户还需要了解影响决策前因后果的因果解释.

3.4 因果解释方法的研究进展

因果解释的目的是回答与因果推理性和反事实解释性相关的问题, 将因果关系作为解释的目标. 本文根据当前因果解释研究方法的原理, 将其研究进展归纳为基于模型的因果解释和基于实例的因果解释两类.

3.4.1 基于模型的因果解释

基于模型的因果解释其主要思想是解释人工智能模型组件对最终决策的因果影响. 例如将深度神经网络 DNN 的结构建模为结构因果模型 (structural causal model, SCM), 并通过执行因果推理来估计模型中每个组件对输出的因果影响. Zhao 等 [56] 证明了利用部分依赖图 PDP 和独立条件期望 ICE 等, 从黑盒模型中提取因果信息是可行的, 前提是掌握因果结构的领域知识. Parafita 引入了一个因果归因框架来解释基于潜在因素的分类器决策, 包括构造分布因果图, 生成与原始图像相似的反事实图像, 最后通过估计因果效应来估计干预因子的影响.

3.4.2 基于实例的因果解释

基于实例的因果解释方法旨在为模型生成反事实解释. 反事实是指与原始实例的输入和输出结果不同的实例, 反事实解释希望通过对原始实例的输入特征进行最小的更改, 以获得不同输出结果的新实例. 例如, 在申请被拒绝的信用卡申请人的特征上可以做哪些最小的改变, 以使他们的申请被接受. 为了产生反事实的例子, Wachter 等 [58] 提出了无条件反事实解释的概念和一种新的自动决策解释类型. Goyal 等提出使用干扰图像生成对查询图像的反事实视觉解释方法 [59]. Kanehira 等针对视频分类任务, 提出了利用多模态信息生成反事实解释的方法 [60], 解释模型预测所有负类的反事实得分, 通过最大化正类和负类之间的反事实得分来生成视觉语言解释.

4 人工智能决策应用的可解释研究进展

AI 可解释性最迫切的应用当前在医疗诊断、自主决策、智能金融和智能司法等高风险决策的情景. 在高风险决策领域, 预测错误可能带来无法挽回的严重后果, 在医疗诊断领域可能会造成误诊、危及病人的生命安全; 在无人驾驶领域可能导致交通事故; 在司法领域可能会使不该获得保释的人获得保释; 在金融领域可能会导致可靠的创业者没有得到应有的贷款 [61], 此时缺乏解释的智能决策更加难以令人信任. 因此, XAI 在高风险领域决策的需求和应用价值更大, 本文从应用的角度出发, 聚焦在智能医疗、无人驾驶、智能司法和智能金融四个典型的高风险智能决策, 对其可解释性研究进行分析.

4.1 智能医疗

人工智能在医疗领域突出的成就是辅助医学诊断. 医学影像与人工智能的结合是数字医疗产业的热点, 人工智能读片具有高效率低成本等优势, 可以减少人为操作的误判率 [62]. 常见的 X 光、CT 等医疗检测产生的图像数据、电子医疗记录 (EHR) 中记录的人口统计信息、体温等生命体征信息、各种医疗检测的数据以及病史信息等都是人工智能系统重要的数据来源, 人工智能算法在复杂的数据上训练后给出诊断结果, 需要进行解释才能被医生理解且信任 [63]. 当前利用黑盒模型进行辅助医疗决策并应用可解释模型多以图像数据、表格式电子医疗记录数据为主 [64]. 智能影像诊断是 “人工智能 + 医疗” 较快落地的应用领域, 因此可解释研究开展也相对多. Zhang 等提出了在语义上和视觉上为可解释性的医学图像诊断网络 (MDNet)[65]. Tang 提出识别全视野数字切片 (WSI) 中 AD 相关的淀粉样斑块和淀粉样血管病特征, 利用了 Guided Grad-CAM 等方法 [66]. 文献 [67] 给出图像热力图的同时, 还生成对应于视觉特征的病理诊断报告. 付贵山等用 CAM 和 Grad-CAM 解释基于深度学习的乳腺癌超声图像诊断分类 [68]. Karim 构建了 Deep COVID Explainer 的诊断解释器 [69], 解释了 2020 年 COVID-19(新型冠状病毒肺炎) 的智能诊断. 该研究设计多个深度网络集成对 13808 位病人的 16995 张胸部 X 光图像进行诊断预测, 用 Grad-CAM[33]、Grad-CAM++[34] 以及 LRP[30] 方法对预测模型进行解释. 图 2 所示是两种解释方法对于一个 COVID-19 患者的预测. Kim 等将 TCAV 解释方法应用在糖尿病视网膜病变的诊断分级问题上 [70], TCAV 方法显示的重要概念和医生判断的概念是一致的, 还能纠正智能模型诊断错误. Cai 等还用 CAV 的方法设计了处理医疗决策过程中不完善算法的工具 [71], 基于 CAV 的方法已经被用在包括前列腺癌组织等病理学上 [72].

在表格式数据的 XAI 应用上文献 [73] 进行了系统性的总结. Rafi 等研究 ICU 的再接收概率预测, 获得了更高的效率 [74]. Panigutti 等提出了 Doctor XAI 的事后可解释系统 [75], 采用训练的决策树产生的规则进行解释. Li 等提出了模型的特征重要度, 对局部特征贡献解释和全局一致性分析的临床数据视觉解释系统 [76].

4.2 无人驾驶

工业界和学术界都投入了大量的精力用于无人驾驶系统的开发. 从 DARPA 城市挑战 [77] 开始, 已经有了一系列的的尝试, 奔驰和谷歌等公司正在竞争开发商用全自动汽车 [78]. 目前大多数无人驾驶的视觉系统都基于深度学习算法, 对可解释性有迫切的需求. 无人驾驶特斯拉命案事故、Uber 无人车致死案和谷歌旗下 Waymo 无人驾驶测试车事故, 考验着人们的容忍度, 因此对无人驾驶智能决策系统的可解释需求愈加强烈. 对于端到端无人智能系统的解释, 通常利用视觉注意模型进行自动驾驶决策的解释. Kim 等人使用因果过滤器的视觉注意模型可视化注意力热力图 [79]. 但当前的视觉解释对决策的逻辑缺乏有效的解释. 为此 Kim 等

对驾驶视频数据中的动作进行描述, 用文本解释构建知识库描述驾驶决策, 并给出对应的解释 [79]. 文献 [80] 从模型理解、模型诊断和模型改进等方面对 SAR 图像目标识别的可解释性进行了探讨. 在自主规划的解释方面, Borgo 等提出了面向 AI 规划决策的解释方法 [81]. Anjomshoae 等对可解释的代理和机器人进行了系统性的综述 [82]. 目前距离实现真正安全、可靠的无人驾驶可解释系统, 依然存在许多关键问题亟待解决.

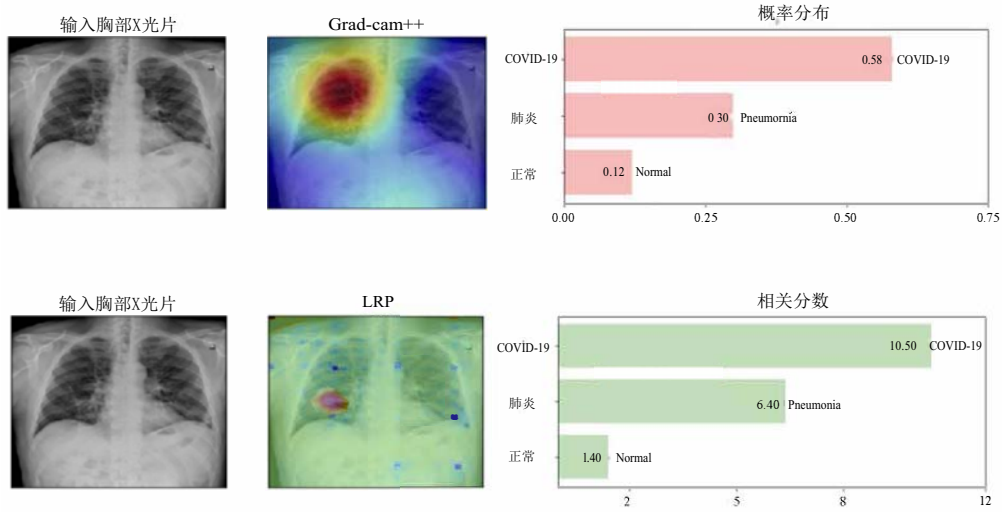

图 2 Deep COVID Explainer 对 COVID-19 智能诊断的胸部 X 光图像解释

4.3 智能司法

智能司法领域面临着高风险的挑战, 若采用人工智能进行辅助判决, 法官面对判决案例需要得到预测模型的解释. XAI 方法促进法官和人工智能决策系统之间的信任 [83], 可以帮助识别算法出现的偏见或者不公平的结果. 目前 AI 在司法领域的应用大多使用浅层透明的模型. Lakkaraju 等针对保释裁决问题, 利用规则集算法建立了考虑裁决成本的模型, 显著降低了裁决成本 [84]. Zeng 等用超稀疏线性整数模型 (SLIM) 建立了预测罪犯的再犯罪概率系统 [85], 在高准确率的同时可以保持透明的可解释性. 辅助裁决算法 COMPAS 是一个商用模型, 可以帮助法官判断一个罪犯是否会二次犯罪. 由于 COMPAS 是商业机密, 一般将其视作黑盒模型进行研究, 文献 [86] 指出此模型存在对非裔美国人的歧视. Dodge 等用逻辑回归模型学习 COMPAS 黑盒模型的知识, 利用特征重要性、敏感性分析等方法对 COMPAS 模型做了解释 [87]. Tan 等对司法相关的模型应用知识蒸馏的方法, 对芝加哥警方逮捕数据集、纽约警方拦截排查数据集和 ProPublica 再犯罪风险评估数据集 [88] 进行了研究 [89], 验证了知识蒸馏解释方法的有效性和芝加哥警方使用模型无道德偏见声明的真实性.

4.4 智能金融

以人工智能为代表的创新技术, 正在推动传统金融变革, 继而向金融科技转变. 但不透明的黑盒金融决策系统可能会导致决策者的不信任、金融监管困难等新的挑战, 因此对智能金融决策的解释更为迫切. Thomas 等对近二十多年来的金融风险评估预测模型进行了全面的回顾 [90]. 英格兰银行试图通过 Shapley 值解释抵押贷款违约模型 [90]. 在文献 [91] 中使用定量输入影响 (QII) 来识别特征影响, 例如通过改变申请人的种族来知道种族改变是否会改变最终结果. Lecue 等提出了结合语义网络的人工智能金融系统 (AIFS)[92], 利用数据可视化方法对异常费用进行预测. 文献 [93] 提出了利用信念规则库 (BRB) 实现贷款承销过程的自动化可解释的人工智能决策支持系统. 系统可以通过激活规则的重要性和规则中先行属性的贡献来解释导致贷款申请决策的事件链. Bussmann 提出了用于衡量使用贷款平台借贷时的信贷风险的可解释人工智能模型 [94], 该模型采用 Shapley 值解释信用评分进行预测, 目前在金融方向的人工智能解释文献仍然较少.

5 人工智能可解释方法的评估

人工智能可解释方法需要设计评估方法去度量, 目前学术界对于 XAI 的评估已经有了初步的研究, 但是

还没有形成统一的体系和评估指标 [9]. 吴俊杰等从算法维度讨论了深度学习的可解释性问题与挑战, 也提到对解释需要保证的性质和定量描述仍缺乏统一的标准, 是一个高度开放的问题, 亟需进一步探索和完善 [95]. 本文将 XAI 的评估指标分为主观评估和客观评估两类.

5.1 主观的评价标准

主观评估关注人类对于 XAI 解释的主观认知, 包括解释的良好度, 对解释的满意度, 对解释的理解程度, 对解释的信任程度, 以及用户的体验等 [73,96,97]. 这些指标大部分可以通过李克特量表来衡量, 特别地在评估信任度时 Jian 等人 [98] 提出的无人系统信任度问卷被广泛使用 [73,99,100]. 然而, Buçinca 等人 [101] 研究发现主观评估存在很多的误导, 即用户使用主观评价好的系统时不一定会有好的表现和绩效, 所以对 XAI 主观评估的结果可能无法预测使用 XAI 系统进行实际决策任务的效果. 目前针对主观评估研究的正式成果还比较少, 可以看出 XAI 领域的研究需要心理学、人机交互等领域的研究者共同参与.

5.2 客观的评价指标

目前的研究主要聚焦在 XAI 评估研究的客观指标上. 常用的指标如保真度 [6] (衡量解释是否真实反映原模型)、复杂度 [102] (衡量解释自身长短和个数) 和鲁棒性 [103] (衡量解释是否容易被输入扰动或者对抗样本所攻击) 等. 针对同一指标, 不同的学者也提出了不同的评估方法, 例如文献 [104] 基于扰动提出 (in)fidelity 与 sensitivity 两个衡量指标, 文献 [105] 提出 AIC 和 SIC 两个指标以衡量保真度和鲁棒性. 本文将可解释性客观评估方法分为以下四类: 第一类是通过构建基准数据集建立评估, 例如采用 TCAV[53] 提出了概念激活向量, 并构建了特殊的数据集, 利用解释关注图像概念来评估解释方法, 衡量解释的准确性. 文献 [106] 构建了客观评价视觉解释方法的数据集, 基于物体的颜色和部位区分类别, 利用解释得到的可视化区域和 Ground Truth masks 的交并比衡量解释的准确性. 第二类是建立无基准数据集的评估. 构建数据集的方法限制了评估的适用性, 难以推广, 因此无数据集评估的指标有所不同, 主要有评估解释的复杂性、相关性和完整性 [107]. 文献 [6] 评估解释的泛化性, 保真度和说服度. 文献 [108] 提出了类敏感性的评估指标, 通过计算预测最大类别和最小类别的显著图的相关性, 衡量解释的准确性. 文献 [109] 通过对网络权重破坏并重新训练未被破坏的权重, 衡量解释的 “恢复力”. 第三类通过移除或扰动解释确定的重要像素来评估对分类的影响. 文献 [110] 提出了 ROAR 指标, 将移除像素的图像重新训练模型, 在相同分布下进行评估和训练, 文献 [111] 提出了像素扰动指标, 删除 k 个最不突出的像素之后衡量对分类器输出的变化. 第四类是对因果解释的评估. 现有的因果解释方法大多是基于反事实的解释, 对于这类方法的评估一般是通过生成反事实解释的好坏来衡量. Mothilal 等 [112] 通过在原始样本和反事实样本上训练一个辅助分类器来预测新输入的类别, 将该模型的精度与原始模型的精度进行比较, 认为精度越高则反事实解释的效果越好. 对于反事实解释的因果关系评估还涉及到了对公平性的考量. Kusner 等提出一个衡量公平决策如何基于反事实的标准 [113]. Zhang 等提出了一种指标来定量计算算法的公平性, 其衡量标准包含反事实直接效应、间接效应和虚假效应三种因果传递的衡量标准 [114]. 综上所述, 由于决策者的理解和固有的主观不一致性, XAI 评价指标的侧重点不同, 对应的定义和指标体系也不尽相同, 应根据决策任务进行合理定义和选取.

6 人工智能决策可解释的未来研究趋势

可解释人工智能 XAI 近年吸引了众多学者参与到 XAI 的研究中, 成为近年的研究热点 [115], 带来了不同学科的视角, 也出现了商业化部署 XAI 的系统 [116]. 但是, 这个方向的研究还处于初级阶段, 不管是在理论层面还是在应用层面, 都需要深入研究, 本文对人工智能决策可解释的未来研究方向进行了展望:

1) 人工智能解释应从感知走向认知. 在数据 - 预测 - 决策这个模式下, 人工智能决策要做出更加智能的决策, 也要更好的决策解释. 对高风险人工智能决策的合理解释, 应从视觉感知走向推理认知, 需要和认知科学、心理学科、人机交互等学科相结合.

2) 人工智能解释需和人类决策协作互补. 决策涉及到人的经验和行为, 深入研究智能决策和人的决策、智能决策和 AI 采纳之间的关系和交互作用, 才能确保人工智能系统不会侵蚀人类的自主性. 人工智能决策可解释和系统运行相结合, 确保人工适当参与到人工智能高风险应用当中, 才能实现可信赖、合伦理和以人为核心的人工智能目标. 如自动驾驶中驾驶员感觉不安全, 可当即使用停止按钮控制决策权.

3) 人工智能可解释从事后解释向自解释发展. 事后解释黑盒模型始终难以从内在逻辑上准确直接的解释模型决策的依据, 而只能根据对于机器重要性高的特征进行近似解释. 因此在发展 XAI 事后解释的同时, 需要大力开发自解释的复杂决策模型.

4) 对终端用户的背景和需求进行解释. 针对不同背景知识的用户, 应提供针对决策的不同利益相关者的解释, 发展个性化定制化和常识结合的智能决策可解释.

5) 将 XAI 纳入人工智能监管框架机制中. 人工智能可解释性的相关政策制度尚不完善, 需要尽快研究全流程的监管政策, 定期地评估人工智能系统的生命周期和运转状态.

7 总结

本文分析了人工智能决策可解释性的概念和挑战, 重点从人工智能决策可解释性的基本概念、模型解释方法、高风险应用解释和解释方法评估等 4 个方面等方面对现有相关研究进行综述, 最后展望了人工智能决策可解释性未来的研究方向. 从本文总结的应用来看, 基于医疗图像数据的 XAI 研究在医疗领域占了很大的比重, 而无人驾驶领域的视觉感知在很大程度上依赖于计算机视觉的处理和深度学习的迅速发展, 因此当前多聚焦在计算机视觉领域的 XAI 研究. 对司法领域而言, XAI 技术的作用主要是追求模型的公平和无偏, 希望通过 XAI 技术识别黑盒模型中存在的偏见, 以促进司法的透明和公正. 在金融领域, XAI 技术被广泛应用于信用预测、风险评估等方面, 由于该领域对模型可解释性与可信任性依赖较高, 目前应用集中于简单自解释模型, 对于提高精确度和性能更高的黑盒模型可解释性还有很长的道路要走. 综上所述, 本研究从决策受益角度, 对当前吸引众多计算机学者、管理学者、社会学者等关注的人工智能决策可解释性进行了综述研究. 提出了新的可解释人工智能的定义, 同时根据利益相关者的不同阐述了 XAI 研究的需求, 给出了 XAI 研究和应用需要达成的目标. 在此基础上, 对 XAI 研究方法做了新的分类: 对模型黑盒 “CT 式” 的视觉解释方法、对黑盒施加外在扰动的解释方法, 基于知识的解释方法和因果解释方法. 更有意义的是, 对 XAI 技术在业界的应用实例做了详细的整理, 涉及了医疗、无人驾驶、金融和司法四个高风险的决策领域. 人工智能决策可解释是非常有应用价值和商业价值的研究方向, 相关研究不仅可以促进人工智能理论的发展, 也可以促进人工智能系统在多个领域中的赋能和部署, 扩大智能决策系统的受众, 充分发挥人工智能强大的能力使之成为普惠大众的新生代工具, 对推进国家人工智能决策战略的快速落地和发展, 具有重要的现实意义.

参考文献

[1] 新华社. 国务院印发《促进大数据发展行动纲要》[EB/OL]. [2015-09-05]. http://www.gov.cn/xinwen/2015-09/05/content_2925284.htm.

[2] 国务院. 国务院关于印发新一代人工智能发展规划的通知 [EB/OL]. [2017-07-20]. http://www.gov.cn/home/2017-07/20/content_5212053.htm.

[3] European Union. General data protection regulation (GDPR)[EB/OL]. [2018-05-25]. https://gdpr.eu/tag/gdpr/.

[4] 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[R]. Report on Award No. FA8650-17-2-7711, DARPA XAI Program, 2018.

[5] Gunning D. Explainable artificial intelligence (XAI)[R]. Report for DARPA XAI Program, 2017.

[6] Du M, Liu N, Hu X. Techniques for interpretable machine learning[J]. Communications of the ACM, 2019, 63(1): 68–77.

[7] 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.

[8] Preece A, Harborne D, Braines D, et al. Stakeholders in explainable AI[C]// AAAI FSS-18: Artificial Intelligence in Government and Public Sector Proceedings, 2018.

[9] Arrieta A B, Díaz-Rodríguez N, Del Ser J, et al. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI[J]. Information Fusion, 2020, 58: 82–115.

[10] Gunning D, Stefik M, Choi J, et al. XAI — Explainable artificial intelligence[J]. Science Robotics, 2019, 4(37): eaay 7120. doi: 10.1126/scirobotics.aay7120.

[11] Bhatt U, Xiang A, Sharma S, et al. Explainable machine learning in deployment[C]// Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020: 648–657.

[12] Samek W, Wiegand T, Müller K R. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models[J]. ITU Journal: ICT Discoveries, 2017, 1(S1): Article 5.

[13] 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 S L, Li J F, Du T Y, et al. Survey on techniques, applications and security of machine learning interpretability[J]. Journal of Computer Research and Development, 2019, 56(10): 2071–2096.

[14] Miller T. Explanation in artificial intelligence: Insights from the social sciences[J]. Artificial Intelligence, 2019, 267: 1–38.

[15] Malik N, Singh P V. Deep learning in computer vision: Methods, interpretation, causation, and fairness[M]// Operations Research & Management Science in the Age of Analytics. INFORMS, 2019: 73–100.

[16] Murdoch W J, Singh C, Kumbier K, et al. Interpretable machine learning: Definitions, methods, and applications[J]. PNAS, 2019, 116(44): 22071–22080.

[17] Miller T, Howe P, Sonenberg L. Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioral sciences[C]// IJCAI Workshop Explainable AI (XAI), 2017: 36–42.

[18] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks[C]// European Conference on Computer Vision, 2014: 818–833.

[19] Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps[C]// Workshop at International Conference on Learning Representations, 2014.

[20] Springenberg J T, Dosovitskiy A, Brox T, et al. Striving for simplicity: The all convolutional net[C]// International Conference on Learning Representations, 2015.

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

[22] Du M, Liu N, Song Q, et al. Towards explanation of DNN-based prediction with guided feature inversion[C]// Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018: 1358–1367.

[23] Adebayo J, Gilmer J, Goodfellow I, et al. Local explanation methods for deep neural networks lack sensitivity to parameter values[C]// International Conference on Learning Representations, 2018.

[24] Bach S, Binder A, Montavon G, et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation[J]. PLoS ONE, 2015, 10(7): e0130140.

[25] Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences[C]// Proceedings of the 34th International Conference on Machine Learning, 2017: 3145–3153.

[26] Zhou B, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 2921–2929.

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

[28] Chattopadhay A, Sarkar A, Howlader P, et al. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks[C]// 2018 IEEE Winter Conference on Applications of Computer Vision. IEEE, 2018: 839–847.

[29] Du M, Liu N, Song Q, et al. Towards explanation of dnn-based prediction with guided feature inversion[C]// Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018: 1358–1367.

[30] Huang Z, Li Y. Interpretable and accurate fine-grained recognition via region grouping[C]// IEEE Conference on Computer Vision and Pattern Recognition, 2020.

[31] Ribeiro M T, Singh S, Guestrin C. "Why should I trust you?" Explaining the predictions of any classifier[C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016: 1135–1144.

[32] Fong R, Vedaldi A. Interpretable explanations of black boxes by meaningful perturbation[C]// 2017 IEEE International Conference on Computer Vision. IEEE, 2017.

[33] Petsiuk V, Das A, Saenko K. Rise: Randomized input sampling for explanation of black-box models[C]// British Machine Vision Conference (BMVC), 2018.

[34] Ribeiro M T, Singh S, Guestrin C. Anchors: High-precision model-agnostic explanations[C]// Proceedings of the 32nd AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2018.

[35] Cui X, Wang D, Wang Z J. CHIP: Channel-wise disentangled interpretation of deep convolutional neural networks[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 31(10): 4143–4156.

[36] Guo W, Mu D, Xu J, et al. LEMNA: Explaining deep learning based security applications[C]// Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, 2018: 364–379.

[37] Wagner J, Kohler J M, Gindele T, et al. Interpretable and fine-grained visual explanations for convolutional neural networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019: 9097–9107.

[38] Lundberg S M, Lee S I. A unified approach to interpreting model predictions[C]// Advances in Neural Information Processing Systems, 2017: 4765–4774.

[39] Cook R D, Weisberg S. Characterizations of an empirical influence function for detecting influential cases in regression[J]. Technometrics, 1980, 22(4): 495–508.

[40] Wojnowicz M, Cruz B, Zhao X, et al. "Influence sketching": Finding influential samples in large-scale regressions[C]// 2016 IEEE International Conference on Big Data, IEEE, 2016: 3601–3612.

[41] Koh P W, Liang P. Understanding black-box predictions via influence functions[C]// Proceedings of the 34th International Conference on Machine Learning, 2017: 1885–1894.

[42] Anirudh R, Thiagarajan J J, Sridhar R, et al. Influential sample selection: A graph signal processing approach[R]. 2017. https://dblp.org/rec/journals/corr/abs-1711-05407.

[43] Yeh C K, Kim J, Yen I E H, et al. Representer point selection for explaining deep neural networks[C]// Advances in Neural Information Processing Systems, 2018: 9291–9301.

[44] Sato M, Tsukimoto H. Rule extraction from neural networks via decision tree induction[C]// International Joint Conference on Neural Networks, IEEE, 2001, 3: 1870–1875.

[45] Zilke J R, Mencía E L, Janssen F. DeepRED — Rule extraction from deep neural networks[C]// International Conference on Discovery Science, 2016: 457–473.

[46] Augasta M G, Kathirvalavakumar T. Reverse engineering the neural networks for rule extraction in classification problems[J]. Neural Processing Letters, 2012, 35(2): 131–150.

[47] Sethi K K, Mishra D K, Mishra B. KDRuleEx: A novel approach for enhancing user comprehensibility using rule extraction[C]// 2012 Third International Conference on Intelligent Systems Modelling and Simulation. IEEE, 2012: 55–60.

[48] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network[C]// NIPS Deep Learning and Representation Learning Workshop, 2015: 1–9.

[49] Liu X, Wang X, Matwin S. Improving the interpretability of deep neural networks with knowledge distillation[C]// 2018 IEEE International Conference on Data Mining Workshops (ICDMW). IEEE, 2018: 905–912.

[50] Tan S, Caruana R, Hooker G, et al. Distill-and-compare: Auditing black-box models using transparent model distillation[C]// Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 2018: 303–310.

[51] Wang X, Wang D, Xu C, et al. Explainable reasoning over knowledge graphs for recommendation[C]// Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33: 5329–5336.

[52] Ma W, Zhang M, Cao Y, et al. Jointly learning explainable rules for recommendation with knowledge graph[C]// The World Wide Web Conference, 2019: 1210–1221.

[53] Kim B, Wattenberg M, Gilmer J, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)[C]// International Conference on Machine Learning, 2018: 2668–2677.

[54] Bau D, Zhou B, Khosla A, et al. Network dissection: Quantifying interpretability of deep visual representations[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 6541–6549.

[55] Zhou B, Sun Y, Bau D, et al. Interpretable basis decomposition for visual explanation[C]// Proceedings of the European Conference on Computer Vision (ECCV), 2018: 119–134.

[56] Zhao Q, Hastie T. Causal interpretations of black-box models[J]. Journal of Business & Economic Statistics, 2021, 39(1): 272–281.

[57] Parafita Á, Vitrià J. Explaining visual models by causal attribution[C]// XAIC Workshop at International Conference on Computer Vision, 2019.

[58] Wachter S, Mittelstadt B, Russell C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR[J]. Harvard Journal of Law & Technology, 2017, 31: 841.

[59] Goyal Y, Wu Z, Ernst J, et al. Counterfactual visual explanations[C]// International Conference on Machine Learning (ICML), 2019: 2376–2384.

[60] Kanehira A, Takemoto K, Inayoshi S, et al. Multimodal explanations by predicting counterfactuality in videos[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019: 8594–8602.

[61] Rudin C, Carlson D. The secrets of machine learning: Ten things you wish you had known earlier to be more effective at data analysis[M]// Operations Research & Management Science in the Age of Analytics. INFORMS, 2019: 44–72.

[62] Yu K H, Beam A L, Kohane I S. Artificial intelligence in healthcare[J]. Nature Biomedical Engineering, 2018, 2(10): 719–731.

[63] Che Z, Purushotham S, Khemani R, et al. Interpretable deep models for ICU outcome prediction[C]// AMIA Annual Symposium Proceedings. American Medical Informatics Association, 2016: 371.

[64] Karatekin T, Sancak S, Celik G, et al. Interpretable machine learning in healthcare through generalized additive model with pairwise interactions (GA2M): Predicting severe retinopathy of prematurity[C]// 2019 International Conference on Deep Learning and Machine Learning in Emerging Applications (Deep-ML). IEEE, 2019: 61–66.
[65] Zhang Z, Xie Y, Xing F, et al. Mdnet: A semantically and visually interpretable medical image diagnosis network[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 6428–6436.
[66] Tang Z, Chuang K V, DeCarli C, et al. Interpretable classification of Alzheimer's disease pathologies with a convolutional neural network pipeline[J]. Nature Communications, 2019, 10(1): 2173.
[67] Zhang Z, Chen P, McGough M, et al. Pathologist-level interpretable whole-slide cancer diagnosis with deep learning[J]. Nature Machine Intelligence, 2019, 1(5): 236–245.
[68] 付贵山. 深度学习乳腺超声图像分类器及其可解释性研究 [D]. 哈尔滨: 哈尔滨工业大学, 2019.
[69] Karim M R, Döhmen T, Cochez M, et al. DeepCOVIDExplainer: Explainable COVID-19 Diagnosis from Chest X-ray Images[C]// 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2020: 1034–1037.
[70] Kim B, Wattenberg M, Gilmer J, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)[C]// International Conference on Machine Learning. PMLR, 2018: 2668–2677.
[71] Cai C J, Reif E, Hegde N, et al. Human-centered tools for coping with imperfect algorithms during medical decision-making[C]// Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 2019: 1–14.
[72] Arvaniti E, Fricker K S, Moret M, et al. Automated gleason grading of prostate cancer tissue microarrays via deep learning[J]. Scientific Reports, 2018, 8(1): 1–11.
[73] Payrovnaziri S N, Chen Z, Rengifo-Moreno P, et al. Explainable artificial intelligence models using real-world electronic health record data: A systematic scoping review[J]. Journal of the American Medical Informatics Association, 2020, 27(7): 1173–1185.
[74] Rafi P, Pakbin A, Pentyala S K. Interpretable deep learning framework for predicting all-cause 30-day ICU readmissions[R]. Tech. Rep, 2018.
[75] Panigutti C, Perotti A, Pedreschi D. Doctor XAI: An ontology-based approach to black-box sequential data classification explanations[C]// Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020: 629–639.
[76] Li Y, Fujiwara T, Choi Y K, et al. A visual analytics system for multi-model comparison on clinical data predictions[J]. Visual Informatics, 2020, 4(2): 122–131.
[77] Montemerlo M, Becker J, Bhat S, et al. Junior: The stanford entry in the urban challenge[J]. Journal of field Robotics, 2008, 25(9): 569–597.
[78] Ziegler J, Bender P, Schreiber M, et al. Making bertha drive-an autonomous journey on a historic route[J]. IEEE Intelligent Transportation Systems Magazine, 2014, 6(2): 8–20.
[79] Kim J, Rohrbach A, Darrell T, et al. Textual explanations for self-driving vehicles[C]// Proceedings of the European Conference on Computer Vision (ECCV), 2018: 563–578.
[80] 郭炜炜, 张增辉, 郁文贤, 等. SAR 图像目标识别的可解释性问题探讨 [J]. 雷达学报, 2020, 9(3): 462–476.
Guo W W, Zhang Z H, Yu W H, et al. Perspective on explainable SAR target recongnition[J]. Journal of Radars, 2020, 9(3): 462–476.
[81] Borgo R, Cashmore M, Magazzeni D. Towards providing explanations for AI planner decisions[C]// IJCAI/ECAI 2018 Workshop on Explainable Artificial Intelligence (XAI), 2018.
[82] Anjomshoae S, Najjar A, Calvaresi D, et al. Explainable agents and robots: Results from a systematic literature review[C]// 18th International Conference on Autonomous Agents and Multiagent Systems. International Foundation for Autonomous Agents and Multiagent Systems, 2019: 1078–1088.
[83] Deeks A. The judicial demand for explainable artificial intelligence[J]. Columbia Law Review, 2019, 119(7): 1829–1850.
[84] Lakkaraju H, Rudin C. Learning cost-effective and interpretable treatment regimes[C]// Artificial Intelligence and Statistics, 2017: 166–175.
[85] Zeng J, Ustun B, Rudin C. Interpretable classification models for recidivism prediction[J]. Journal of the Royal Statistical Society: Series A (Statistics in Society), 2017, 180(3): 689–722.
[86] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead[J]. Nature Machine Intelligence, 2019, 1(5): 206–215.
[87] Dodge J, Liao Q V, Zhang Y, et al. Explaining models: An empirical study of how explanations impact fairness judgment[C]// Proceedings of the 24th International Conference on Intelligent User Interfaces, 2019: 275–285.
[88] ProPublica. Data and analysis for "Machine Bias"[EB/OL]. https://github.com/propublica/compas-analysis, 2017.

[89] Tan S, Caruana R, Hooker G, et al. Distill-and-compare: Auditing black-box models using transparent model distillation[C]// Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 2018: 303–310.
[90] Thomas L, Crook J, Edelman D. Credit scoring and its applications[M]. Society for industrial and Applied Mathematics, 2017.
[91] Bracke P, Datta A, Jung C, et al. Machine learning explainability in finance: An application to default risk analysis[R/OL]. (2019-08-09). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3435104.
[92] Lecue F, Wu J. Explaining and predicting abnormal expenses at large scale using knowledge graph based reasoning[J]. Journal of Web Semantics, 2017, 44: 89–103.
[93] Sachan S, Yang J B, Xu D L, et al. An explainable AI decision-support-system to automate loan underwriting[J]. Expert Systems with Applications, 2020, 144: 113100.
[94] Bussmann N, Giudici P, Marinelli D, et al. Explainable AI in fintech risk management[J]. Frontiers in Artificial Intelligence, 2020, 3: 26.
[95] 吴俊杰, 刘冠男, 王静远, 等. 数据智能: 趋势与挑战 [J]. 系统工程理论与实践, 2020, 40(8): 2116–2149.
Wu J J, Liu G N, Wang J Y, et al. Data intelligence: Trends and challenges[J]. Systems engineering — Therory & Practice, 2020, 40(8): 2116–2149.
[96] Weitz K, Schiller D, Schlagowski R, et al. "Do you trust me?" Increasing user-trust by integrating virtual agents in explainable AI interaction design[C]// Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents, 2019: 7–9.
[97] Zhou B, Sun Y, Bau D, et al. Interpretable basis decomposition for visual explanation[C]// Proceedings of the European Conference on Computer Vision (ECCV), 2018: 119–134.
[98] Jian J Y, Bisantz A M, Drury C G. Foundations for an empirically determined scale of trust in automated systems[J]. International Journal of Cognitive Ergonomics, 2000, 4(1): 53–71.
[99] Kulesza T, Burnett M, Wong W K, et al. Principles of explanatory debugging to personalize interactive machine learning[C]// Proceedings of the 20th International Conference on Intelligent User Interfaces, 2015: 126–137.
[100] Häuslschmid R, von Buelow M, Pfleging B, et al. Supportingtrust in autonomous driving[C]// Proceedings of the 22nd International Conference on Intelligent User Interfaces, 2017: 319–329.
[101] Cui X, Lee J M, Hsieh J. An integrative 3C evaluation framework for explainable artificial intelligence[C]// The Annual Americas Conference on Information Systems (AMCIS), 2019.
[102] Cui X, Lee J M, Hsieh J. An Integrative 3C evaluation framework for explainable artificial intelligence[J]. The Annual Americas Conference on Information Systems (AMCIS), 2019.
[103] Kindermans P J, Hooker S, Adebayo J, et al. The (un)reliability of saliency methods[M]// Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Springer, Cham, 2019: 267–280.
[104] Yeh C K, Hsieh C Y, Suggala A, et al. On the (in)fidelity and sensitivity of explanations[C]// Advances in Neural Information Processing Systems, 2019: 10965–10976.
[105] Kapishnikov A, Bolukbasi T, Viégas F, et al. XRAI: Better Attributions Through Regions[C]// Proceedings of the IEEE International Conference on Computer Vision, 2019: 4948-4957.
[106] Mogrovejo J A, Wang K, Tuytelaars T, et al. Visual explanation by interpretation: Improving visual feedback capabilities of deep neural networks[C]// International Conference on Learning Representations, 2019.
[107] Dodge J, Liao Q V, Zhang Y, et al. Explaining models: An empirical study of how explanations impact fairness judgment[C]// Proceedings of the 24th International Conference on Intelligent User Interfaces, 2019: 275–285.
[108] Rebuffi S A, Fong R, Ji X, et al. There and back again: Revisiting backpropagation saliency methods[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 8839–8848.
[109] Vasu B, Savakis A. Visualizing the resilience of deep convolutional network interpretations[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019: 107–110.
[110] Hooker S, Erhan D, Kindermans P J, et al. A benchmark for interpretability methods in deep neural networks[C]// Advances in Neural Information Processing Systems, 2019: 9734–9745.
[111] Srinivas S, Fleuret F. Full-gradient representation for neural network visualization[C]// Advances in Neural Information Processing Systems, 2019: 4126–4135.
[112] Mothilal R K, Sharma A, Tan C. Explaining machine learning classifiers through diverse counterfactual explanations[C]// Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020: 607–617.
[113] Zhang J, Bareinboim E. Fairness in decision-making-the causal explanation formula[C]// Proceedings of the AAAI Conference on Artificial Intelligence, 2018.
[114] Zhang J, Bareinboim E. Fairness in decision-making-the causal explanation formula[C]// Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
[115] 吴飞, 廖彬兵, 韩亚洪. 深度学习的可解释性 [J]. 航空兵器, 2019, 26(1): 39–46.
Wu F, Liao B B, Han Y H. Interpretability for deep learning[J]. Aero Weapony, 2019, 26(1): 39–46.
[116] Lecue F, Gade K, Geyik S C, et al. AAAI 2020 Tutorial[EB/OL].[2020-02-07].
https://xaitutorial2020.github.io/raw/master/slides/aaai_2020_xai_tutorial.pdf.