武漢大學學報(哲学社会科学版)
WUHAN UNIVERSITY JOURNAL (Philosophy & Social Science)
第74卷第5期 2021年9月
Vol. 74 No. 5 Sept. 2021 016~028
DOI:10.14086/j.cnki.wujss.2021.05.002

迈向可解释的交互式人工智能:动因、途径及研究趋势

吴 丹 孙国烨

摘 要 近年来,人工智能的功能愈发强大,应用场景也越来越广泛。在人机交互协作成为常态的背景下,人们对人工智能可信任度及交互体验的要求也随之增高,机器的可解释性受到广大用户的高度重视,可解释的人工智能正在成为相关领域的重要议题。要迈向可解释的交互式人工智能,应当从设计指导性的框架准则、开发良好的算法模型、深入研究用户需求、构建个性化的交互式解释系统、展开有效的可解释性评估等方面发力。

关键词 人工智能;人机交互;可解释性;系统透明度;用户信任
中图分类号 TP18 **文献标识码** A **文章编号** 1672-7320(2021)05-0016-13
基金项目 武汉大学"人工智能问题"融通研究专项课题(2020AI020)

随着新一代信息技术的快速发展,人工智能技术已经在多个领域得到了应用,并对人类社会产生了巨大的影响,小到产品推荐、广告定制、交友建议,大到无人驾驶、疾病诊断、司法裁判,人工智能不断渗透到人们生活的方方面面,许多人已经习惯参考并采纳人工智能提供的决策建议。人正在被无处不在的人工智能所深入影响。基于此,近几年,多所机构或企业陆续围绕以人为中心的人工智能展开了研究,如斯坦福大学成立了"以人为本的人工智能研究所",清华大学与阿里巴巴集团宣布共同建设"自然交互体验联合实验室"。这些研究强调,人工智能的未来不仅仅在于技术,人工智能必须是关乎人类的,其落脚点在于增强人的能力,而不是将人类取而代之。与此同时,在实践应用层面,支持人类与之进行交互的人工智能系统层出不穷。这表明,人工智能的发展越来越需要纳入"人"的因素,只有将人工智能(Artificial Intelligence,简称AI)与人机交互(Human-Computer Interaction,简称HCI)结合起来考虑,才能实现人类对人工智能的更佳运用。

然而,在人类与人工智能的交互成为常态的背景下,不可忽视的是人工智能存在的"黑匣子"问题阻碍了人们对相应系统的理解和运用。为此,建立可解释的人工智能,开始被许多学者、机构乃至政府组织所呼唤,成为以人为中心的人工智能的重要研究板块。2017年4月,美国国防部高级研究计划署(Defense Advanced Research Projects Agency,简称DARPA)启动了"可解释的人工智能"(Explainable Artificial Intelligence,简称XAI)项目,旨在提高AI系统的可解释性[1]。2018年5月,欧洲联盟(European Union,简称EU)发布了《通用数据保护条例》(General Data Protection Regulation,简称GDPR),强制要求人工智能算法具有可解释性[2]。2019年6月,我国发布的《新一代人工智能治理原则——发展负责任的人工智能》指出,人工智能系统应不断提升可解释性[3]。

与此同时,在建立可解释人工智能的过程中,我们应当认识到人机交互的重要性,人工智能的可解释性是在人工智能与用户的交互中得以实现的。因此,区别于以往研究中常被提及的可解释人工智能,本文侧重于交互这一要素,试图讨论可解释的交互式人工智能——在与人类进行交互时,其行动背后的具体逻辑能够被用户所理解的人工智能。

由“黑匣子”转向可解释的交互式人工智能是相关领域的发展趋势，也是实现以人为中心的人工智能的前瞻性挑战。本文即着眼于可解释的交互式人工智能，主要从其发展动因、实现途径两方面对国内外的相关研究进展进行归纳与阐述，在此基础上展望可解释交互式人工智能未来的研究趋势。

一、可解释交互式人工智能的基本概念及发展背景

人工智能是当前诸多科技领域的核心所在，可解释性是面向人工智能的一个研究话题，旨在使人类更好地理解人工智能系统的行动与决策。可解释的人工智能并不属于全新的研究问题。最初针对人工智能的研究大多认为，系统得出决策结果的依据应当被解释，尤其在专家系统解释的相关研究出现之后，可解释性这一研究方向便一直存在。与可解释的人工智能相关的研究工作甚至出现在20世纪80年代左右的文献中，当时人工智能系统的解释主要是基于对应用规则的利用实现的，因为这些规则和信息是由专家来拟定的，再加上最初的系统较为容易解释，所以理解人工智能的行动相对来说难度较低[4]（P815-823）。

在相当长的一段时间内，随着人工智能的不断发展，相关研究重点开始转向算法与模型的改进，更多强调其技术方面，特别是预测方面的能力。与之相对应的是，解释方面的能力在人工智能领域的研究地位下降，研究人员对其重视度稍显不足。近年来，在深度学习技术取得惊人进步的背景下，人工智能拥有了超强的学习与决策能力，能够解决的任务也前所未有的复杂，人们越来越多地运用人工智能系统帮助自己在重要场景下做出决定，人工智能逐渐成为影响社会发展的重要存在。

在人工智能领域，有两个与可解释性相关的关键概念：透明度与准确性。透明度（Transparency）强调人工智能系统能够给出自身工作原理的程度，是用户理解机器的基础[5]（P673-705）。准确性（Accuracy）则强调模型的拟合能力以及在某种程度上准确预测未知样本的能力。当人工智能生成的决策极大影响到人类生活的多个方面时，人类对人工智能可解释性的要求也随之攀升，相对于此前对系统准确性、使用便捷性的单一关注，如今的利益相关人员对透明度的要求越来越高[6]（P1-8）。

人工智能的准确性与透明度之间存在一定的矛盾，如果仅仅在意系统的性能水平，那么其可解释性将难以得到提升，权衡两者的关系是一直以来的难题[7]（P210-215）。但如果不提高人工智能的可解释性，其决策就难以得到解释，很可能产生不合理甚至危险的结果，而提升可解释性也许就能够修正系统在准确度方面的缺陷。因此在预测能力与解释能力的平衡方面，至少可以确定的一点在于人类需要具有解释能力的人工智能。

正因为如此，可解释性近来得到了重新关注，而且已经成为人工智能研究领域的热点话题。但实现可解释的人工智能并不是一件易事，诸如深度神经网络等被智能系统所运用的不透明模型，使得打开人工智能的黑箱变得更加困难[8]（P20）。可以说，可解释性是人工智能发展到一定阶段后产生的又一个重要难题。人工智能的相关研究人员正在围绕可解释性进行不断的努力，以期达成打开人工智能黑箱的目标。在学者层面，部分国际会议如International Conference on Machine Learning（简称ICML）、Conference and Workshop on Neural Information Processing Systems（简称NeurIPS）等将可解释的人工智能纳入研讨会的讨论主题中。在政府组织层面，由美国国防部高级研究计划署启动的可解释人工智能计划影响较为广泛。在工业界层面，许多企业致力于提升其人工智能相关产品的可解释性。

随着“以人为中心的人工智能”这一术语被广泛接受，越来越多的学者开始给可解释的人工智能下定义。宽泛来讲，可解释的人工智能可以指能够使得人类用户理解并信任其输出的相关技术[9]，它可以解释算法的工作形式，帮助使用者了解为何以及怎样得出决策结果。当重点考虑交互这一要素，放眼可解释的交互式人工智能时，可以发现，可解释性涉及机器与人类两方面的问题，因此对可解释交互式人工智能的讨论不应只集中于技术，人工智能的解释能力与用户这一要素紧密相关。与可解释人工智能相比，可解释交互式人工智能更多探讨用户、交互等与解释对象相关的因素。基于此，在学者们给出的

可解释人工智能的众多概念中,从解释对象角度进行的阐述也许更加贴合可解释交互式人工智能的定义。本文对 Arrieta 关于可解释人工智能的定义加以引申,认为可解释交互式人工智能是指针对特定使用者,在交互过程中能够提供细节和原因,使得系统背后的行动逻辑能够被用户所理解的人工智能[10](P82-115)。

可解释的交互式人工智能不仅局限于文本中的概念,它更多涉及实践中采取的各项举措,深入了解可解释交互式人工智能在现实中的发展非常重要。因此,本文对其发展动因、实现途径、研究趋势进行了细致的梳理与总结,以期描绘出可解释交互式人工智能的概貌。

二、可解释交互式人工智能发展的主要动因

复杂变化的环境提升了研究人员对可解释交互式人工智能的关注度,推动了其在理论与实践层面的发展,可以从人工智能研究通常涉及的两个主体(机器与用户)、一种关系(人机关系)对可解释交互式人工智能的发展原因进行探讨。无论是在技术进步的背景下决策更加复杂、影响更加广泛的机器,还是在大数据时代更加关注人工智能使用体验与安全性的用户,抑或是人与机器更加深入的交互与协作关系,都是可解释交互式人工智能发展的主要原因所在。

(一)AI 在改变——功能的强化与应用领域的增多

随着人工智能功能的强化,用户难以理解其决策逻辑,更容易产生不解与反感,人工智能必须变得可解释以证明其决策结果的合理性。同时,人工智能应用领域的增多也使得其面临的风险更多样、更庞大,实现可解释的交互式人工智能能够在一定程度上防范重大事故。

得益于技术的进步,人工智能的功能越来越强大,包括收集、存储和使用用户数据,以及进行自动推荐与评估等。与人工智能的功能强化相对应的是,其系统和算法的复杂度也在不断提高,要理解人工智能做出决策的依据,需要使用者具有相当丰富的专业知识及较高的技术水平。现实中,大多数用户都是非专业用户,对于他们而言,复杂的人工智能越来越容易被视作难以理解的“黑匣子”,这阻碍了使用者对人工智能的理解,人们在与机器进行交互的过程中,更加难以仅仅凭借自身的努力去理解人工智能的行动逻辑。用户不知道人工智能是依据怎样的行动逻辑得出了现有的决策,这容易对系统的可信任度产生负面影响,在交互过程中也会出现更多问题。比如,有研究发现,Airbnb 的房东对于系统将其房源排在检索结果列表的某个次序,以及针对相应房源给出建议定价等操作的依据感到困惑[11](P1-12)。人工智能的决策能力愈发强大,也意味着它对人们现实权益的影响更加深入。比如,休斯顿的一所学校通过人工智能系统来评估教师的表现,却无法解释该系统生成不利评价的原因,遭到了相关教师的起诉[12](P563-574)。如果人工智能的强大功能是在不能为人类提供对应理由的背景下完成的,那么受到系统决策影响的人们则很可能对其产生强烈的不解与反感。因此,人工智能需要向人类提供必要的解释信息,以证明其决策结果的合理性。欧洲联盟 2018 年 5 月发布的《通用数据保护条例》即规定,公民具有受算法决策影响的“解释权”[2]。

人工智能可解释性的意义不仅在于证明其决策结果的合理性,也在于预防重大问题的发生。近年来,人工智能开始被运用到大量场景中,并涉及自动驾驶、刑事司法、医疗诊断、金融财务等高风险领域。当人工智能被用于开展这些特定工作的时候,如果用户面对的是“黑箱”,则人工智能的自主决策将存在无法控制的巨大风险。比如,在交通领域,优步公司(Uber Technologies,Inc.,简称 Uber)的自动驾驶车辆曾导致一起死亡事故的发生[13](P52138-52160);在司法领域,美国威斯康星州法院曾以使用封闭源代码的系统为依据指导定罪量刑,引发了诸多争议与质疑[14](P327-343);在医学领域,有研究发现预测肺炎的人工智能可能因为系统性偏差得出明显错误的结论[15](P1721-1730)。而这些领域的用户往往具有专业知识,他们对人工智能的理解程度越深,越能够及时发现系统存在的问题。当前,在自动驾驶、医疗临床、刑事司法、金融服务等领域已经逐渐展开了人工智能可解释性的相关研究。

（二）用户在改变——对AI可信任度及交互体验的要求增高

伴随着人工智能的广泛运用，用户对这个插手自身生活的“黑匣子”提出了更高的要求，他们希望人工智能是公平的、透明的、负责任的，拥有较高的可信任度，能够提供良好的用户体验，而可解释性是实现这些目标的共同基础。

就人工智能的公平性而言，由于机器与人类不同，不具有复杂的情感特征，因此从表面来看，它的行动与决策应当是更为理性公平的，但前提是人工智能并未受到部分企业或开发者的利用，变成小部分群体获取利益的工具。当前，越来越多的个人或群体用户认识到，人工智能的算法不仅要在计算方面保持公平，还应当摒弃潜在的偏见与歧视，否则现实生活中的不平等也许会因为人工智能而加剧。可解释性在一定程度上能够用于保障人工智能的公平性，部分可解释的机器学习模型基于可视化的方法，显示系统得出决策结果的关系图，以分析人工智能是否受到偏见的影响[16]（P153-163）。

就人工智能的透明度而言，在当今信息爆炸、互联网迅速发展的时代，个人数据极易被获取并滥用，用户的隐私意识也相应越来越强，透明度较高的人工智能更加符合人们的心理需求。具体而言，在人机交互的过程中，用户往往需要输入部分涉及个人隐私的数据，他们通常希望了解这些数据是如何被系统处理的[17]（P137-141）。可解释性越高的系统，其透明度越能够得到保证，透明的解释能够为用户提供心理层面的保障，因此可解释的交互式人工智能往往更受人们青睐。

就负责任的人工智能而言，在人工智能运行逻辑得以被理解的情况下，人们才更加容易识别应当承担算法后果责任的实体，实现负责任的人工智能[18]（P1-28）。总而言之，用户感知到的人工智能可解释性会影响其对机器的信任，当人工智能的可解释性越高时，用户更可能认为它是公平的、透明的、负责任的。

另外，对于人机交互中的用户体验感来说，可解释性也是非常重要的。从“人”的视角来看，如果无法了解机器的运行逻辑，交互的体验则会大打折扣。其一，当机器总是能够向用户解释其行动的依据，且用户认为人工智能给出的解释足够时，更可能认为系统具有较高的有用性及易用性，也更容易接受机器在交互过程中提供的相关建议[19]（P277-284）；其二，高质量的解释能够在人机交互过程中使人类与机器之间的关系更为融洽、和谐，促进两者间的相互理解，尤其是在人工智能出现意外行动或失误时，减轻人类对机器的厌恶感等负面情绪。因此，用户与可解释性较好的人工智能通常有着更为自然的交互，在使用系统的过程中容易产生亲密感和舒适感。总之，用户对解释能力出色的系统具有较高的满意度[20]（P2390-2395）。

（三）人机关系在改变——从机器的单方面输出到人机交互下的任务协作

随着人类与人工智能的交互越来越频繁且深入，人机交互开始被运用于多种任务协作的场景下，而实现更好的协作效果需要人机互相理解。要让用户充分地参与人机协作，达成利用机器增强人类能力的目的，人工智能的可解释性是必不可少的要素之一。

一方面，建立可解释的人工智能是使用者在人机协作中得以优化交互的基础。当面对直接决策、对其行动毫无解释与说明的机器时，使用者无从得知如何优化自身与机器之间的交互协作，便难以采取相应的举措。用户只有理解了人工智能的运行逻辑，才能够做到通过修改输入机器的内容，来优化自身接收的内容[21]（P102078）。另一方面，在人机关系中，随着机器自主权的提升，英特尔自适应机器人等人工智能已经能够实现主动交互与自主修正，如何在机器发起主动交互时让用户充分理解与信任人工智能，并愿意参与人机交互，帮助人工智能进行动态学习与修正，都对系统的可解释性提出了新的要求[22]（P326-331）。此外，有研究认为，人工智能在实现决策的过程中涉及许多人类并不知晓的知识，因此，如果人工智能在人机交互中是能够提供相应解释的存在，则其可以主动向人类传递或说明新信息，促成人机协作质量的进一步提高[13]（P52138-52160）。

三、可解释交互式人工智能的实现途径

在实现可解释的交互式人工智能的过程中,采取何种途径与方法是我们无法绕过的重要问题。事实上,在实践中认识到可解释交互式人工智能现实意义的部分企业、学者已经从理论及技术方面出发,进行了相关探索与尝试。在对已有研究进行梳理后,本文认为,实现可解释的交互式人工智能需要在指导层面牢牢把握框架准则的构建,在技术层面持续致力算法模型的开发,在设计层面着重关注用户需求的分析,在优化层面有效开展可解释性的评估。这些要素是在实现可解释交互式人工智能的过程中不可或缺的角色。

(一)以框架准则构建为指导

实现可解释的交互式人工智能,绝不是一项平地起高楼的简单工作,系统开发者在设计时需要参考可解释交互式人工智能的相关准则及指导性框架,综合考虑多方因素。微型计算机软件公司(Microsoft Corporation,简称微软)、谷歌公司(Google Inc.,简称Google)、国际商业机器公司(International Business Machines Corporation,简称IBM)及部分学者提出了构建可解释交互式人工智能的参考准则;部分研究则以人类认知或系统交互为核心,构建了指导可解释交互式人工智能设计的概念或理论框架。

在准则方面,微软从交互阶段出发,提出了18种普遍适用的人机交互指南,涵盖人工智能系统在与用户初次接触、互动时期、发生错误时期以及深入协作四个阶段的设计建议,为生成便于用户理解的系统提供了参考准则[23](P1-13)。Google基于对部分可解释人工智能的工作示例的收集展示,针对可解释交互式人工智能的设计提供了推荐指南[24]。IBM给出了设计可解释交互式人工智能的措施建议,包括实现系统与用户之间纽带的途径[25]。另有学者立足于降低解释复杂度,提出了“人类自治系统设计指南”,为设计者提升系统可解释性,将其复杂性降至最低提供了设计准则[26](P29-34)。

在指导性框架方面,有研究从人类认知的角度出发,将人工智能的解释工具同人类认知模式联系起来,构建了用于指导设计以人为中心的可解释交互式人工智能系统的概念框架[27](P1-15);另有研究从系统交互的角度入手,开发了涵盖系统交互协议、用于指导实现可解释交互式人工智能的理论框架[28]。

无论是企业根据已有实践提炼的设计准则,还是涵盖人类认知与系统交互两方面的理论框架,它们对于实现可解释的交互式人工智能均具有指导性意义。系统开发者不应在设计过程中忽略这些框架准则,要在充分理解的基础上运用它们,助力可解释式交互人工智能的开发工作。

(二)以算法模型设计为基石

在增强人工智能的可解释性方面,不可或缺的是算法、模型等解释方法的运用及改进。模型复杂度是一个在人工智能领域常常被提及的概念,它强调模型在结构上的复杂程度。模型复杂度与模型的准确性密切相关,一般情况下,模型的复杂度越低,其拟合能力越差,准确性越弱,但透明度较高,可解释性相对较好;模型的复杂度越高,其准确性越强,但透明度较低,可解释性相对较差。

人工智能的解释方法可以根据解释对象的模型复杂度分成两大类,一类是事前解释(ante-hoc),另一类是事后解释(post-hoc)。事前解释主要针对复杂度较低的模型,事后解释主要针对复杂度较高的模型[29](P31-57)。本文基于这两类方法的相关文献,梳理了其对应的具体模型或算法,如表1所示。

事前解释即使得模型本身可解释,事前解释根据解释的实现途径又可分为两种:采用自解释模型和构建具有内置可解释性的模型。

第一种方法是直接采用传统机器学习中的自解释模型。例如,线性模型、决策树、广义加性模型、K最近邻分类算法、基于规则的模型、朴素贝叶斯模型等,这些模型结构简单,自身就具有可解释性,主要体现在能够给出要素对决策的重要性度量。

第二种方法是实现模型的内置可解释性。根据郭炜炜等人的总结,目前构建具有内置可解释性的模型有几种主要方法[30](P462-476)。一是引入注意力机制。注意力机制起源于认知神经学的研究,指

表1　人工智能解释方法

类型		方法	
事前解释	自解释模型	线性模型、决策树、广义加性模型、K最近邻分类算法、基于规则的模型、朴素贝叶斯模型	
	构建具有内置可解释性的模型	引入注意力机制、深化统计模型、基于物理模型	
事后解释	全局性解释	激活最大化、概念激活矢量测试、知识蒸馏	
	局部解释	反向传播	基于梯度的方法、导向反向传播法、积分梯度法 、平滑梯度法
		类激活映射	类激活映射、梯度加权类激活映射
		局部近似	局部线性近似法、非线性逼近的局部解释法
		沙普利解释模型	

人脑在信息过量的背景下可以重点处理某些信息而忽略其他信息，本质是对信息进行加权。已有研究将注意力机制引入文本分类任务，对不同词语进行权重量化，帮助人类理解每个词对分类结果的贡献[31]（P1480-1489）。二是深化统计模型。和现有的深度神经网络相比，统计模型的可解释较强，因此，已有研究尝试以统计学习模型为基础，构建神经网络模型，如基于稀疏编码方法ISTA，生成LISTA（Leaned ISTA）模型[32]（P399-406）。三是基于物理模型。已有研究尝试参考物质世界的规则，进行神经网络建模，如根据雾的生成原理，构建端到端的神经网络模型，使得模型中的各个模块都有清晰的物理意义[33]（P1234-1240）。

事后解释指开发解释技术对已经训练好的学习模型进行后验解释，根据解释对象的不同，事后解释又可以分为全局性解释方法和局部解释方法。全局性解释方法主要针对模型，帮助人们理解模型内在的工作机制；局部解释方法主要针对每一个输入样本，帮助人们理解模型对于输入样本的决策逻辑和依据。

全局性解释方法主要有激活最大化（AM）、概念激活矢量测试（TCAV）、知识蒸馏。激活最大化即寻找一个最大化特定层神经元激活值的输入模式，是可视化DNN神经单元计算内容的典型方法[34]。概念激活矢量测试方法能够用来判断某一概念对于分类的重要程度，常被用于医疗问题的诊断分级[35]（P1-14）。知识蒸馏是用结构简单的模型来模拟结构复杂的模型，提取复杂模型的决策规则，被广泛应用于模型诊断与验证[36]（P905-912）。

局部解释方法主要有反向传播、类激活映射、局部近似、沙普利解释模型等方法。基于梯度的方法（Gradient-based，简称Grad）、导向反向传播法（Guided Back Propagation，简称GuidedBP）、积分梯度法（Integrated Gradient，简称IntegratedGrad）和平滑梯度法（Smooth Gradient，简称SmoothGrad）等反向传播方法是从模型的输出层推导模型输入层样本的重要性，能够有效定位重要特征[34][37][38][39]。类激活映射（Class Activation Mapping，简称CAM）、梯度加权类激活映射（Gradient-weighted Class Activation Mapping，简称Grad-CAM）方法能够针对卷积神经网络模型生成视觉效果较好的解释，被用于定位决策的重要区域[40]（P2921-2929）[41]（P618-626）；局部线性近似法（Local Interpretable Model-Agnostic Explanation，简称LIME）、非线性逼近的局部解释法（Local Explanation Method-using Nonlinear Approximation，简称LEMNA）等局部近似方法通过捕捉模型的局部特征，在文本和图像的解释方面取得了良好的效果[42]（P1135-1144）[43]（P364-379）；基于Shapley值的沙普利解释模型（Shapley Additive Explanations，简称SHAP）通过计算个体的贡献来确定其重要程度，被英格兰银行尝试用于解释抵押贷款违约模型[44]（P4765-4774）；Google还将Tensorflow与SHAP相结合，以进一步提升可解释性[45]。总体而言，当前人工智能的解释方法数量较多，但每种方法都存在着或多或少的不足。比如：自解释模型准确性较低，预测性能与可解释性之间的矛盾较大，受到多种因素的限制；激活最大化方法只能用于连续型数据，无法应用于图像等离散型数据，且容易受到噪音的影响；反向传播方法无法量化特征的重要程度等。因此，在

面对不同情境时,应当根据各个方法的特点及优势,选取合适的算法或模型来实现人工智能系统的解释。同时,可解释交互式人工智能也在呼吁更加优秀的算法。

(三)以用户需求纳入为桥梁

人工智能的解释对象是用户,要实现可解释的交互式人工智能,就需要更加关心人机交互中的人这一要素。在设计可解释的交互式人工智能的过程中,需要充分开展用户研究,调查用户所需的解释内容、解释过程中相关要素对用户的影响等,基于研究结果提升系统的解释质量,使解释更加贴合人类认知。

已有研究探讨了个体的不同对于人工智能解释的影响。如将可解释交互式人工智能的解释对象分为设计者、领域专家、终端用户三类,得出不同类别用户对解释的独特需求[46];通过对比在有无解释的情况下不同用户的注视模式差异,总结个性特征对用户处理解释方式的影响[47](P397-407)。鉴于解释界面是可解释交互式人工智能中不可忽视的因素之一,也有研究探讨了解释界面的不同对于用户理解的影响,如通过用户的阶段参与过程与心理模型研究,探讨用户需要具有何种透明度的解释界面[48](P211-223);测量当解释界面不同时,用户对算法的理解程度差异等[49](P1-12)。

另有部分研究着眼于用户需求与现有解释水平之间的差距。如通过开发用户需求方面的可解释交互式人工智能问题库,并据此采访从事人工智能用户体验设计与优化的工作人员,得出可解释交互式人工智能算法实践与用户需求之间的差异[50](P1-15);基于人机协作实验的开展,研究用户对人工智能给予解释的时机及详细程度的实际要求[51](P1-13)。

此外,针对不同情境下的人机交互,加利福尼亚大学洛杉矶分校计算机视觉、认知、学习与自主机器人中心及清华大学人工智能研究院智能信息获取研究中心、武汉大学人机交互与用户行为研究中心等机构开展了一系列针对用户的研究,以大规模、多情境的用户研究助力可解释交互式人工智能的发展。其中,武汉大学人机交互与用户行为研究中心基于海量用户数据,着眼语音交互、手势交互、眼动交互、认知等多通道,研究移动端及桌面端的用户行为,进而分析个体或群体需求[52](P102073)[53](P109-128)。

对用户需求的研究,可以用于支持系统解释在两方面的提升。其一是解释系统的交互性,其二是解释系统的个性化。在交互性方面,当前一次性的解释仍然是最为普遍的,在提供给用户初始解释后,应允许用户调整系统,通过持续的交互深化用户理解。已有研究通过多种途径提高人工智能解释系统的交互性,如构建对话驱动的解释系统,允许人工智能根据用户的信息反馈优化调整其解释[54](P87-90);建立提供交互式界面的解释系统,允许用户修改单个数据点的特征,获取修改前后的对比解释等[55](P5686-5697)。在个性化方面,有研究发现解释并不是越多越好,有时过量的解释信息会造成信息过载,进而影响人机协作的表现,机器应允许不同知识水平、背景的用户通过交互调整解释的复杂度,避免"一刀切"[28]。部分研究已经构建了支持用户设置解释中个性化功能的系统,允许用户根据自己的喜好个性化集群结果和解释形式[56](P131-138)。

当解释系统的交互性得以提升时,用户能够获取更加持续的解释,深入了解系统背后的运行逻辑,进而使得人机交互更加流畅和谐。同时,只有当解释系统的个性化程度得以加强后,解释系统才能针对不同的个体、不同的情境提供恰当的解释。而这些都建立在对用户需求的调研上。因此,只有将用户需求研究作为中间桥梁,才能实现对人与机器的综合考虑,迈向可解释的交互式人工智能。

(四)以可解释性评估为辅助

如果不知道怎样的解释才是好的解释,开发者在设计时就会缺乏依据及目标,可解释交互式人工智能的发展也会遭受阻碍。因此,在迈向可解释交互式人工智能的道路上,人工智能解释的开发是一方面,其可解释性评估也同样是不可忽略的一环。目前,交互式人工智能的可解释性评估具有解释本身、开发者、用户三个视角。

一是从系统解释本身的质量出发,衡量的指标包括保真度、复杂度、忠诚度、鲁棒性等。保真度主要

评估解释是否真实反映人工智能系统，通常还涉及解释的说服度[57]（P68-77）；复杂度主要评估解释的数量、长短等，越复杂的解释对用户而言可能越难理解[58]；忠诚度主要衡量机器在与用户交互的过程中，输出的解释是否能够按照用户的调试而改变[59]（P126-137）；鲁棒性则更多关注解释是否容易被干扰[60]（P267-280）。

二是从开发者的角度出发，主要衡量解释是否实现了其设计目标。DARPA的可解释人工智能项目就将评估要素与系统开发时的解释目标相连接，希望通过两者之间的对照，进行可解释性的评估[1]。部分学者也通过测量系统解释设计的各个目标实现程度，评估解释的质量[61]（P329-338）。

三是从用户的角度出发，专注于评估用户接收解释的效果和体验，已有研究大多集中于主观指标的测量上，如用户满意度、解释的良好程度、对系统的信任度等[62]（P81-87）[63]（P3-19）[64]（P1-6）。这样的测量可以得到使用者对解释的实际评价，但是使用者评价的主观性会导致评估结果不一定如实反映解释的质量，因此，还需要衡量用户根据解释所做推断或完成任务的情况[65]（P103404）。可以将用户对系统的信任度、满意度等指标与其工作绩效、心理模型等相结合，以此评估系统的可解释性，或者通过面向用户、基于特定情境的任务测试，来实现更准确的评估。有学者即提出通过采用真实的用户及任务、结合真实用户与简化任务、无人状态下的代理任务三类情境进行人工智能的可解释性评估[66]。

可以看出，当前交互式人工智能的可解释性评估方法较为多样，涉及的指标也较为复杂。虽然由这些方法得出的评估结果是否能够切实反映解释水平还未可定，且解释本身、开发者、用户三个视角的评估方法较为分散，其结合程度不足，但不可否认的是，可解释性评估是在指导性框架准则、算法模型、用户需求研究之外，能够为开发者在设计方面带来更多帮助的环节。想要实现可解释的交互式人工智能，可解释性评估是必需的辅助工作。

四、可解释交互式人工智能的研究趋势

在机器、用户、人机关系均发生变化的背景下，可解释交互式人工智能得到了较多关注，学者、政府、工业界从指导性框架构建、算法模型设计、用户需求分析、可解释性评估等方面发力，以期实现可解释的交互式人工智能，如图1所示。

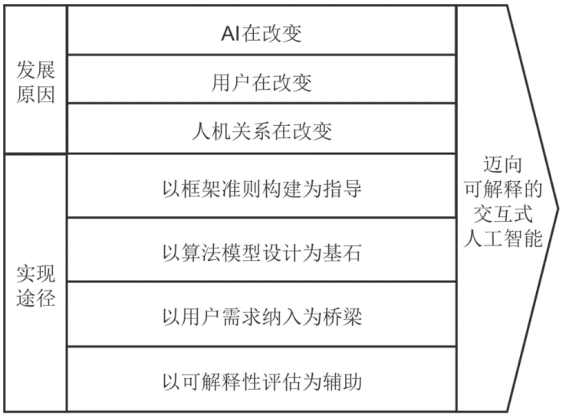

图1　可解释交互式人工智能的发展原因与实现途径

然而，通过对现有研究的梳理，可以发现，它们还存在着一定的不足。如在算法模型设计方面，当前常用的解释方法不一定能够真实有效地反映模型的决策；在用户需求分析方面，对不同情境、不同个体的需求调研不够深入，对交互性解释的支持不足；在可解释性评估方面，已有的评估指标不成体系，难以开展科学全面的评价；在研究宽度方面，结合的学科领域较少，可解释交互式人工智能的研究视角还有

待拓展。在实现可解释的交互式人工智能的道路上,这些已有研究的不足正是我们亟待解决的现实问题。因此,在未来,多种解释方法的有效结合与优化、对用户需求的大规模与多通道分析、可解释性科学评估体系的构建、研究视角的合理拓宽,都是进一步研究的方向所在。

(一)改进与强化解释方法

在可解释交互式人工智能的实现过程中,主要涉及的矛盾是准确性与透明度之间的矛盾,而透明度又与可解释性息息相关,因此,准确性与可解释性往往是相对立的。如自解释模型,虽然可以直接进行解释,但通常相伴的是模型的精准度较低。而更为准确的人工智能模型往往结构复杂,因此对其一般采用事后解释,在事后解释过程中存在的多种阻碍因素会导致解释结果不能有效地反映模型的真实决策行为或运行逻辑。当前事后解释相关研究面临的主要挑战就是设计能够真实反映模型行动逻辑的解释方法,保证解释结果的可靠性,使得用户更加全面准确地理解人工智能的内在运行机制,未来的潜在研究方向包括从数学层面入手设计等价的解释模型等。

模型的预测能力与解释能力之间并不是完全矛盾的,我们不应当局限于对模型可解释性的研究,需要放眼可解释的模型。目前来看,自解释模型与事后解释都存在较为明显的缺点,而构建具有内置可解释性的模型很可能是实现可解释交互式人工智能的突破口所在。在未来的研究中,应当从引入注意力机制、深化统计模型、参考物理模型等多方面着手,构建更优的可解释模型,使其兼具传统机器学习模型可解释性强、深度学习模型预测性能优的特点。

综上,未来应当充分发挥各种解释方法的优势,通过多种方法的有效结合或优化,致力于平衡模型的准确性与可解释性,重点构建具有内置可解释性的模型,实现高质量、易理解、程度恰当的解释。

(二)实现个性化的交互式解释

运用人工智能的落脚点在于增强人类能力,在交互式人工智能的解释过程中,应当充分考虑人这一重要主体。现有的解释系统很少考虑到终端用户的期望,未来我们应当转而建立以人为中心的可解释交互式人工智能,其设计应由用户需求来驱动,DARPA 的可解释人工智能项目即鼓励将算法技术与用户实验相结合。开展更为广泛实际的用户调查与实验,深入研究参与者对系统解释的需求及影响因素,是优化解释系统所必需的。

仅有良好的解释方法与用户研究是不够的,交互式人工智能的可解释性是在系统与用户的交互过程中得以实现的,用户在接收到系统提供的初始解释后,应当拥有调整及反馈的渠道,以此来深化对系统的理解。避免一次性的解释,使交互真正从理论探究到实践运用,构建具有更强交互性的解释系统,如运用智能代理(对话、可视化)等工具,将交互操作集成于解释界面,允许用户在接收解释的过程中与机器产生更为深入的交互,以此提升解释质量。

此外,随着社会发展与技术的进步,未来人工智能将被运用到越来越多的领域,在更加丰富的情境下被更为广泛的人群所使用,不同人群、不同应用情境对人工智能解释形式与内容的期望都是不同的,因此可定制的解释是非常有必要的。在进行充分的用户研究后,研究人员应当根据使用者的背景与需求,开发面向不同群体、不同应用情境的解释系统,有针对性地提供特定解释,同时允许使用者通过交互来个性化解释。

(三)构建涵盖多方指标的评估体系

由于解释范围或原理不同,可解释性研究领域仍然没有形成一个较为全面的科学评估体系。基于解释本身视角的评估忽略了用户这一可解释交互式人工智能中的重要因素,缺乏用户对解释质量的评价;基于开发者视角的评估从解释的设计目标出发,在评价指标的选取上具有一定的局限性;基于用户视角的评估则极度依赖人类认知,多属于定性评估,主观性较强,难以对系统解释的水平进行量化。此外,在同一情境下不同模型的可解释性、在不同情境下同一模型的可解释性,以及采用不同方法解释同一模型的可解释性,都存在一定的差异。目前还没有能够较好衡量与比较这些可解释性的评估方法。

因此，在未来的研究中，应当结合解释本身、开发者、用户三个视角，纳入应用情境、解释方法差异等指标，构建覆盖多层次、多角度的评估体系，实现全面有效的可解释性评估。

（四）拓展涉及更多领域的研究视角

目前，在可解释交互式人工智能的相关研究中，其他领域的参与仍然较少，但也有部分研究已经迈出了前进的步伐，如DARPA就计划利用心理学理论助力研发解释模型，构建可解释交互式人工智能的测评框架。

在算法模型的设计方面，应当更加重视人类认知神经学、物理、数学等领域知识的嵌入与融合，如基于注意力机制或物理模型构建具有内置可解释性的模型；利用知识图谱，将人类知识引进人工智能模型中，帮助用户理解模型特征；引入语义概念和关联等信息，使模型能够更好地进行特征学习，具有更高的可解释性。

实现更具可解释性的交互式人工智能，还需要结合社会科学，发挥其学科优势。在用户研究方面，可以纳入心理学与认知科学，引入经典的行为科学实验方法、心理学理论模型、社会技术系统研究方法等[67]；在应用情境方面，可以顺应“十四五”时期我国将人工智能等新技术应用于文化领域、以人为本实现精准服务的发展方向，结合数字文化等特定情境，以用户为中心探讨人工智能的可解释性，为拓宽可解释交互式人工智能的应用领域打下基础[68]（P14-26）；在解释程度方面，可以结合法学、管理学等学科知识，对人工智能解释的适用范围等进行探索。

人工智能在日常生活中的广泛应用加速了社会向更具算法性的方向转型，然而无论技术上的进步多么空前，人工智能的最终目的始终是增强人的能力，而非取代人。在人机协作愈发常见的背景下，结合人工智能（AI）与人机交互（HCI），实现以人为中心的人工智能是大势所趋。当前阻碍人类有效使用人工智能的障碍之一即在于其透明度的缺乏，机器能够给出有力的预测，却无法提供通俗的解释。可解释交互式人工智能的发展正是为了改善这一问题，人工智能可解释性的研究对学界、政府组织、企业、个人用户而言，都具有极高的理论与实践价值。本文梳理了该领域的部分工作，探讨了可解释交互式人工智能的发展动因、实现途径及研究趋势，希望能够为相关研究人员提供参考。人工智能的强大功能不应为“黑匣子”所困。我们相信可解释的交互式人工智能将在未来突破这层阻碍，助力我国新一代人工智能产业的发展。

参考文献

[1] M. Turek. DARPA-Explainable Artificial Intelligence (XAI) Program. DARPA's Official Website, 2017-04. [2020-12-23] https://*www.darpa.mil/program/explainable-artificial-intelligence.*

[2] European Union. General Data Protection Regulation (GDPR). GDPR. EU, 2018-05-25. [2020-12-26]*https://gdpr. eu/tag/gdpr/.*

[3] 中华人民共和国科学技术部. 发展负责任的人工智能：新一代人工智能治理原则发布. 中国政府网，2019-06-17. [2020-12-26]*http://www.gov.cn/xinwen/2019-06/17/content_5401006.htm*.

[4] W.R. Swartout. Explaining and Justifying Expert Consulting Programs//Proceedings of the 7th International Joint Conference on Artificial Intelligence, 1981.

[5] A. Rosenfeld, A. Richardson. Explainability in Human-agent Systems. *Autonomous Agents and Multi-Agent Systems*, 2019, 33(6).

[6] J. Zhu, A. Liapis, S. Risi, et al. Explainable AI for Designers: A Human-centered Perspective on Mixed-initiative Co-creation//2018 IEEE Conference on Computational Intelligence and Games, 2018.

[7] F.K. Došilović, M. Brčić, N. Hlupić. Explainable Artificial Intelligence: A Survey//201841st International Convention on Information and Communication Technology, Electronics and Microelectronics, 2018.

[8] C. Davide. Can We Open the Black Box of AI?.*Nature*, 2016, 538(7623).

[9] D. Gunning, M. Stefik, J. Choi, et al. XAI—Explainable Artificial Intelligence. *Science Robotics*, 2019, 4(37).

[10] A.B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, et al. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Information Fusion*, 2020, (58).

[11] S. Jhaver, Y. Karpfen, J. Antin. Algorithmic Anxiety and Coping Strategies of Airbnb hosts//Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 2018.

[12] F. Xu, H. Uszkoreit, Y. Du, et al. Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges// CCF International Conference on Natural Language Processing and Chinese Computing, 2019.

[13] A. Adadi, M. Berrada. Peeking inside the Black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 2018, (6).

[14] J. Lightbourne. Damned Lies & Criminal Sentencing Using Evidence-based Tools. *Duke L. & Tech. Rev.*, 2017, (15).

[15] R. Caruana, Y. Lou, J. Gehrke, et al. Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-day Readmission//Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015.

[16] A. Chouldechova. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, 2017, 5(2).

[17] A. Rai. Explainable AI: From Black Box to Glass Box. *Journal of the Academy of Marketing Science*, 2020, 48(1).

[18] T. Jiya. Ethical Implications of Predictive Risk Intelligence. *ORBIT Journal*, 2019, 2(2).

[19] D. Shin, Y.J. Park. Role of Fairness, Accountability, and Transparency in Algorithmic Affordance. *Computers in Human Behavior*, 2019, (98).

[20] R.F. Kizilcec. How Much Information? Effects of Transparency on Trust in an Algorithmic Interface//Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, 2016.

[21] S. Renjith, A. Sreekumar, M. Jathavedan. An Extensive Study on the Evolution of Context-aware Personalized Travel Recommender Systems. *Information Processing & Management*, 2020, 57(1).

[22] 葛亚特,叶露. 面向自适应机器人交互的类人反应研究. 工业设计研究,2018,(6).

[23] S. Amershi, D. Weld, M. Vorvoreanu, et al. Guidelines for Human-AI Interaction//Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 2019.

[24] Google. Google AI: Responsible AI Practices. Google AI's Official Website,2020-12.[2021-01-01]*https://ai.google/responsibilities/responsible-ai-practices/?category=interpretability.*

[25] IBM. IBM Design for AI:Explainability. IBM's Official Website,2019-05.[2021-01-01]*https://www.ibm.com/design/ai/ethics/explainability.*

[26] M.R. Endsley. Level of Automation Forms a Key Aspect of Autonomy Design. *Journal of Cognitive Engineering and Decision Making*, 2018, 12(1).

[27] D. Wang, Q. Yang, A. Abdul, et al. Designing Theory-driven User-centric Explainable AI//Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 2019.

[28] J. Schneider, J. Handali. Personalized Explanation in Machine Learning: A Conceptualization. European Conference on Information Systems, 2019.

[29] Z.C. Lipton. The Mythos of Model Interpretability. *Queue*, 2018, 16(3).

[30] 郭炜炜,张增辉,郁文贤,孙效华. SAR图像目标识别的可解释性问题探讨. 雷达学报,2020,9(3).

[31] Z. Yang, D. Yang, C. Dyer, et al. Hierarchical Attention Networks for Document Classification//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016.

[32] K.Gregor, Y.LeCun. Learning Fast Approximations of Sparse Coding//Proceedings of the 27th International Conference on International Conference on Machine Learning, 2010.

[33] H. Zhu, X. Peng, V. Chandrasekhar, et al. DehazeGAN: When Image Dehazing Meets Differential Programming//IJCAI, 2018.

[34] K.Simonyan, A.Vedaldi, A. Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and

Saliency Maps. Cornell University's Official Website,2013-12-20.[2021-06-17]https://arxiv.org/pdf/1312.6034.pdf.

[35] C.J. Cai, E. Reif, N. Hegde, et al. Human-centered Tools for Coping with Imperfect Algorithms During Medical Decision-making//Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 2019.

[36] X. Liu, X. Wang, S. Matwin. Improving the Interpretability of Deep Neural Networks with Knowledge Distillation//2018 IEEE International Conference on Data Mining Workshops, 2018.

[37] J.T. Springenberg, A. Dosovitskiy, T. Brox, et al. Striving for Simplicity: The All Convolutional Net. Cornell University's Official Website,2014-12-21.[2021-06-17]*https://arxiv.org/pdf/1412.6806.pdf.*

[38] M. Sundararajan, A. Taly, Q. Yan. Gradients of Counterfactuals. Cornell University's Official Website,2016-11-08.[2021-06-17] *https://arxiv.org/abs/1611.02639v2.*

[39] D. Smilkov, N. Thorat, B. Kim, et al. Smoothgrad: Removing Noise by Adding Noise. Cornell University's Official Website,2017-06-12.[2021-06-17]*https://arxiv.org/pdf/1706.03825.pdf.*

[40] B. Zhou, A. Khosla, A. Lapedriza, et al. Learning Deep Features for Discriminative Localization//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[41] R.R. Selvaraju, M. Cogswell, A. Das, et al. Grad-cam: Visual Explanations from Deep Networks via Gradient-based Localization//Proceedings of the IEEE International Conference on Computer Vision, 2017.

[42] M.T. Ribeiro, S. Singh, C. Guestrin. "Why Should I Trust You?" Explaining the Predictions of Any Classifier//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

[43] W. Guo, D. Mu, J. Xu, et al. Lemna: Explaining Deep Learning Based Security Applications//Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, 2018.

[44] S.M. Lundberg, S.I. Lee. A Unified Approach to Interpreting Model Predictions//Advances in Neural Information Processing Systems, 2017.

[45] Google. AI Explanations Whitepaper. Google Cloud's official website,2019. [2021-06-20] *https://cloud.google.com/ml-engine/docs/ai-explanations/overview.*

[46] M. Ribera, A. Lapedriza. Can We Do Better Explanations? A Proposal of User-centered Explainable AI//Explainable Smart Systems 2019, 2019.

[47] M.Millecamp, N.N. Htun, C. Conati, et al. To Explain or Not to Explain: the Effects of Personal Characteristics When Explaining Music Recommendations//Proceedings of the 24th International Conference on Intelligent User Interfaces, 2019.

[48] M. Eiband, H. Schneider, M. Bilandzic, et al. Bringing Transparency Design into Practice//23rd International Conference on Intelligent User Interfaces, 2018.

[49] H.F. Cheng, R. Wang, Z. Zhang, et al. Explaining Decision-making Algorithms through UI: Strategies to Help Non-expert Stakeholders//Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 2019.

[50] Q.V. Liao, D. Gruen, S. Miller. Questioning the AI: Informing Design Practices for Explainable AI User Experiences//Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 2020.

[51] C. Oh, J. Song, J. Choi, et al. I Lead, You Help but Only with Enough Details: Understanding User Experience of Co-creation with Artificial Intelligence//Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 2018.

[52] D. Wu, J. Dong, C. Liu. Exploratory Study of Cross-device Search Tasks. *Information Processing & Management*, 2019, 56(6).

[53] 吴丹,刘春香. 交互式信息检索研究中的眼动追踪分析. 中国图书馆学报,2019,45(2).

[54] A.R. Akula, S. Todorovic, J.Y. Chai, et al. Natural Language Interaction with Explainable AI Models//CVPR Workshops, 2019.

[55] J.Krause, A.Perer, K.Ng. Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models//Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, 2016.

[56] H. Lakkaraju, E. Kamar, R. Caruana, et al. Faithful and Customizable Explanations of Black Box Models//Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 2019.

[57] M.Du, N. Liu, X. Hu. Techniques for Interpretable Machine Learning. *Communications of the ACM*, 2019, 63(1).

[58] X. Cui, J.M. Lee, J. Hsieh. An Integrative 3C Evaluation Framework for Explainable Artificial Intelligence. The Annual

Americas Conference on Information Systems, 2019.
[59] T.Kulesza,M.Burnett, W.K.Wong, et al. Principles of Explanatory Debugging to Personalize Interactive Machine Learning// Proceedings of the 20th International Conference on Intelligent User Interfaces, 2015.
[60] P.J. Kindermans, S. Hooker, J. Adebayo, et al. The (Un)reliability of Saliency Methods//*Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Cham: Springer, 2019.
[61] K. Balog, F. Radlinski. Measuring Recommendation Explanation Quality: The Conflicting Goals of Explanations//Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), 2020.
[62] U. Ehsan, B. Harrison, L. Chan, et al. Rationalization: A Neural Machine Translation Approach to Generating Natural Language Explanations//Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 2018.
[63] L.A. Hendricks, Z. Akata, M. Rohrbach, et al. Generating Visual Explanations//European Conference on Computer Vision, 2016.
[64] J. Zhou, Z. Li, H. Hu, et al. Effects of Influence on User Trust in Predictive Decision Making//Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, 2019.
[65] J. Van der Waa, E. Nieuwburg, A. Cremers, et al. Evaluating XAI: A Comparison of Rule-based and Example-based Explanations. *Artificial Intelligence*, 2021, (291).
[66] F. Doshi-Velez, B. Kim. Towards a Rigorous Science of Interpretable Machine Learning. Cornell University's Official Web site, 2017-02-28.[2021-01-02]*https://arxiv.org/abs/1702.08608.*
[67] A. Abdul, J. Vermeulen, D. Wang, et al. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda//Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 2018.
[68] 吴丹,郭清玥. “十四五”时期图情学科愿景展望——基于全球战略蓝图的分析. 图书情报知识,2021,38(3).

Towards Explainable Interactive Artificial Intelligence: Motivations, Approaches, and Research Trends

Wu Dan, *Sun Guoye* (Wuhan University)

Abstract　The application scenarios for increasingly powerful artificial intelligence are becoming more and more widespread. As human-computer interaction and collaboration have become the norm, the demand for AI transparency and interaction experience is also increasing accordingly. Explainable artificial intelligence (XAI) is becoming an important topic in related research fields. Future research should design frameworks and guidelines, build good algorithmic models, study user needs in conjunction with other subjects, build personalized interactive explanation systems, and improve the evaluation of explainbility explainability. Only through by these means can we develop move towards explainable interactive AI and human-centerd AI.

Key words　artificial intelligence; human-computer interaction; explainbility; system transparency; users' trust

■ **收稿日期**　2021-03-04
■ **作者简介**　吴　丹,管理学博士,武汉大学信息管理学院教授、博士生导师;湖北 武汉 430072;
孙国烨,武汉大学信息管理学院博士研究生。
■ **责任编辑**　杨　敏