·医学高新技术伦理·

可解释人工智能医疗应用的伦理问题*

江　婧[1]，曹　东[2]，鄢来均[3]**

（1　广州中医药大学马克思主义学院，广东　广州　510006；

2　广州中医药大学医学信息工程学院，广东　广州　510006；3　广州中医药大学研究生院，广东　广州　510006）

〔摘要〕可解释人工智能推动医学发展的同时引发了不同以往的伦理挑战，这种新挑战表现在事前解释、事后解释两个方面。导致可解释人工智能医疗应用伦理问题的技术性根源是可解释性技术的负面效应，其社会性根源是相关道德原则和立法的缺失，主体性根源在于各利益相关方的利益多样性。因此，强化技术创新、可解释人工智能医疗应用伦理教育、对医护人员的培训和跨学科合作，完善相关道德原则和法律法规，实现个性化的交互式解释等对策是问题的解决之道。

〔关键词〕可解释性；人工智能；医疗应用；交互式解释

〔中图分类号〕R-052　　〔文献标志码〕A　　〔文章编号〕1001-8565(2022)12-1322-07

DOI：10.12026/j.issn.1001－8565.2022.12.06

Ethical Issues of Explainable Artificial Intelligence for Medical Applications

JIANG Jing[1], *CAO Dong*[2], *YAN Laijun*[3]

(1　*School of Marxism, Guangzhou University of Chinese Medicine, Guangzhou* 510006, *China*;

2　*School of Medical Information Engineering, Guangzhou University of Chinese Medicine, Guangzhou* 510006, *China*;

3　*School of Graduate, Guangzhou University of Chinese Medicine, Guangzhou* 510006, *China*)

Abstract: Explainable artificial intelligence (AI) promotes the development of medicine and raises different ethical challenges. This new challenge is manifested in two aspects, including pre-interpretation and post-interpretation. The technical root of the ethical issues leading to the medical application of explainable AI is the negative effects of explainable technology, the social root is the lack of relevant moral principles and legislation, and the root of subjectivity lies in the diversity of interests of various stakeholders. Therefore, the solutions to the problem are to strengthen technological innovation, explainable AI medical application ethics education, training of medical staff and interdisciplinary cooperation, improve relevant moral principles, laws and regulations, and realize personalized interactive interpretation.

Keywords: Explainability; Artificial Intelligence; Medical Applications; Interactive Interpretation

0　引言

可解释人工智能已成为学术界、产业界和政府部门研究的热门话题。全世界范围内都在积极探索可解释性的人工智能，例如2016年美国国防部高级计划署(DARPA)就启动了一项名为“可解释人工智能”的大型项目(explainable artificial intelligence，XAI)。2019年欧盟委员会发布《人工智能道德准则》(Ethics Guidelines for Trustworthy AI)，强调要提高人工智能的透明度和实现可追溯性。2017年我国也在《新一代人工智能发展规划》中，明确将“实

*基金项目：广东省普通高校重点领域专项“中医临床多模态异构大数据智能获取与决策支持”(2020ZDZX3080)

**通信作者

现具备高可解释性、强泛化能力的人工智能”作为未来我国人工智能布局前沿基础理论研究。由于医学领域的高风险性，人工智能在医疗领域的可解释性受到了更加广泛的关注。在医疗领域，人工智能的可解释性有助于提高模型的安全性和可靠性，增强目标用户的信任度等，但是，可解释性方法可能引发的医疗伦理问题也日益凸显。

探究可解释人工智能医疗应用伦理问题的意义关乎用户的安全和公平问题。伦理学界从不同的角度对其进行了伦理辩护。其中有两种代表性观点：一种观点认为，在医疗领域中，事后可解释人工智能有助于增强用户信任、提高系统的安全性等。另一种观点认为，人们应该警惕甚至是避免在医疗领域中使用事后可解释人工智能，因为事后解释可能会引发潜在的风险。尽管伦理学界对此不能取得一致意见，但他们都认为探讨可解释人工智能医疗应用伦理问题对患者的安全和公平具有重要意义。本文试图从科技伦理视角审视可解释人工智能医疗应用伦理问题的新挑战，以期寻求合理的解决之道。

1 可解释人工智能医疗应用伦理问题的新挑战

由于考察视角的不同，可解释人工智能并没有一个标准的定义。可解释人工智能来自英文Explainable Artificial Intelligence，该术语由Van Lent等于2004年首次创造，用来描述他们的系统在模拟游戏应用中解释人工智能控制实体行为的能力[1]。自2016年DARPA启动关于可解释人工智能的研究之后，该术语开始被普遍接受，DARPA将可解释人工智能定义为一种系统，它可以向人类用户解释其基本原理，描述其优缺点，并传达对其未来行为的理解[2]。Arrieta等[3]认为，可解释人工智能指的是，针对特定的受众，通过给出细节和原因，使模型运转更清楚和更容易，被受众所理解的一项技术。本文认为可解释人工智能可以定义为：针对不同背景知识的目标用户，以其可以理解的方式，对模型决策过程的根据进行解释的一项技术。目的就是帮助人类理解机器为什么会作出这样的决策以及决策是否可靠。

人工智能的可解释问题源于深度学习的“黑盒”属性。所谓“黑盒”属性，指深度学习算法是不透明的。不同于传统算法输入和输出之间的确定性，深度学习模型通过反向传播不断调整自己内部的参数和规则，所以人类无法理解它们的内部工作机制。在某些应用领域，比如说人脸识别、文字翻译，可解释性并不是关键的要求，只要这些系统的整体性能足够好，即使系统在运行过程中出现错误，也不会造成很大的影响，因此，这些领域对人工智能系统可解释性的要求相对比较低。但是，医疗领域则不同，医疗中的许多决策实际上是生死攸关的问题，微小的错误都可能会威胁到患者的生命安全，这时缺乏可解释性就成为人工智能走向临床应用的限制性因素。所以，越来越多的专家学者将目光投向了人工智能在医疗领域的可解释性，各种解释方法应运而生。目前，可解释人工智能在医学影像处理、疾病诊断、风险预测等方面都取得了不错的成绩。例如，Nafisah等[4]利用可解释人工智能处理胸部X光片检测结核病，Thimoteo等[5]研究出通过血液检测诊断COVID-19的可解释人工智能，Curia[6]利用可解释人工智能预测患宫颈癌的风险。通过给出决策依据，为临床医生提供有效的辅助信息，增加了用户的信任，改善了模型的性能。尽管可解释人工智能给医学带来了很多好处，但其在医疗应用中也引发了不同以往的伦理挑战，具体而言，可解释人工智能医疗应用引发的伦理新挑战主要表现在以下两个方面。

1.1 事前解释可能导致的医疗伦理问题

针对深度学习的“黑盒”属性，人工智能专家创建了不同的解释方法。根据不同的标准，这些方法又分为不同的类别。Du等[7]根据获得可解释性的时间，将可解释性方法总体上划分为两类：内在解释和事后解释。这是两种不同类型的解释方法，它们以不同的方式处理不透明问题。内在解释指的是模型本身可解释，因为可解释性发生在模型训练之前，所以也称为事前解释，即无需事后引入另一个解释模型就可以理解预测模型的决策依据。根据解释的实现途径，又可将事前解释进一步分为两种：自解释模型和内置可解释模型。自解释模型就是指传统机器学习中的简单模型，例如决策树、线性回归等。无论是从理论上还是理解上，这些模型都具有比较好的可解释性。内置可解释模型指的

是，利用某些方法构建可解释的深度神经网络模型。目前，关于事前解释性的研究多局限于决策树、线性回归等结构简单的传统机器学习算法，而内置可解释模型由于对技术要求很高，暂时未取得突破性进展。

事前解释可能会导致医疗安全问题，其表现为以下两种情况：一方面，事前解释的人工智能系统预测准确性较低，导致模型自身存在安全隐患。学术界一般认为，人工智能系统的准确性和可解释性之间存在一定的矛盾，即模型的准确性越高，可解释性就越低；相反，模型的准确性越低，可解释性就越高[8]。尽管在 Rudin[9] 看来，在数据结构化并具有良好的特征时，复杂模型和简单模型之间的性能通常并没有显著差异。例如，利用决策树算法构建区分感冒和咳嗽的预测模型也具有较高的准确率，基本达到诊断要求[10]。但是，这是基于医疗数据具有良好特征时，在现实情况中，绝大多数医学数据是以多种模态呈现，每种模态各有所长、相互关联，极大限制了对病症的预测和诊断[11]。因此，虽然自解释模型自身具备良好的可解释性，但在面对多模态的医疗数据时，模型筛查水平和诊断效率与深度学习模型相比还是存在比较大的差异。因此，在医疗应用中使用事前解释人工智能就意味着要以牺牲准确性为代价，从而导致人工智能系统在辅助临床医生治疗中存在安全隐患，进而给患者的生命安全带来致命的威胁。

另一方面，事前解释为对抗攻击提供了有利条件，导致模型自身存在医疗安全隐患。对抗攻击是神经网络模型中常见的攻击方法，它通过输入人类难以察觉的微小扰动，从而使模型进行错误决策，甚至可以根据设计的扰动，输出攻击者想要的分类结果。研究发现，解释方法可以本能地为对抗样本的生成提供特定区域[12]。对于模型的研究者来说，可解释性技术有助于有效评估模型决策行为的安全性和可靠性，通过对抗训练，模型的鲁棒性和安全性能得到有效的提升，从而消除模型实际部署应用中的安全隐患。但是，对于模型的攻击者来说，可解释方法却也为攻击者探测原始模型弱点提供了有利条件，在医学影像处理方面，对原始图像添加人眼不可分辨的扰动，对于输入中产生的微小变化，都会对深模型预测准确性产生很大的影响。在临床治疗中，系统一旦受到对抗攻击的干扰，那么提供的解释结果必然会影响医生的诊断过程，甚至会误导医生的诊断而给患者带来致命的威胁。

1.2 事后解释可能导致的医疗伦理问题

事后解释指的是创建专门的解释模型来解释原始模型，即需事后引入另一个解释模型才可以理解预测模型的决策依据。它往往针对的是复杂度比较高的模型[13]，比如深度神经网络。因为可解释发生在模型训练之后，所以称为事后解释。根据解释的范围，可进一步将事后可解释分为全局可解释和局部可解释。全局可解释意味着用户可以通过检查复杂模型的结构和参数从而理解模型整体是怎样运行的。局部可解释为模型的单个预测提供局部解释。根据解释方法是否依赖具体模型内部参数，又可分为模型无关的解释方法和模型相关的解释方法，模型无关解释方法可针对任何预测模型，而模型相关解释方法则是针对特定的预测模型。目前常见的事后解释方法主要有知识蒸馏、激活最大化、类激活映射、反向传播等。事后解释能够在保持模型性能的基础上实现可解释，因此，目前在医疗领域主要还是依赖基于事后解释的人工智能系统。事后解释主要会出现以下几种伦理问题：

其一，医疗责权问题。将人工智能引入医疗领域本身就提高了医疗责任主体认定的复杂度。在传统医疗模式下，如果发生医疗事故，医疗机构和医护人员是责任主体，而将人工智能引入医疗领域之后，医生和患者之间增加了人工智能和制造商，这就使得医疗责任主体的认定变得更加复杂。即使将深度学习引入医学领域提高了医疗诊断效率和准确性，但是其应对突发情况的反应处置能力还是带来许多有待解决的问题，目前的技术条件还无法保证百分之百的医疗准确率。比如在医学影像处理方面，对原始图像添加人眼不可分辨的扰动对于输入中产生的微小变化，都会对深度神经网络的预测准确性产生很大的影响，如果出现医疗事故，应该由谁负责。而此时事后解释不仅仅是对传统人工智能医疗责任主体问题的放大，并且是一种颠覆。Babic 等[14]认为这种事后合理化的解释往往可

能会成为一些利益相关者逃避医疗责任的手段。因为对于自身可解释的人工智能来说，如果系统出现错误，我们可以通过回溯运行步骤找出错误的环节，事后解释却难以回溯，而且这种事后的合理化有可能还会被一些利益相关者用来转移注意力。例如，人工智能的制造商为了逃避医疗责任，故意设计一些对自身有利的事后解释模型。当出现医疗事故时，如果医疗责任主体难以认定，患者权益将难以得到保障。

其二，医疗安全问题。在疾病预测方面，利用可解释性人工智能取得了不错的成绩。例如，Rajpurkar等[15]基于深度学习开发了诊断肺部疾病的医疗预测模型，准确度已达到专家级诊断精度。同时，通过事后解释提取模型特征，为临床医生提供了有效的辅助信息，使医生不再盲目的依赖黑匣子。但是研究发现[12]，攻击者可以利用事后可解释性技术对医疗预测模型进行对抗攻击。一方面，在不改变解释结果的前提下，攻击者可以利用可解释性技术探测医疗模型的漏洞，诱导模型作出错误的医疗决策；另一方面，在不改变医疗模型决策结果的前提下，攻击者可以利用可解释技术干扰医疗解释过程，诱导解释方法作出错误的医疗解释。相关研究表明[16]，由于事后解释只是对原始模型的一个间接和近似的解释，攻击者可以还利用二者之间的不一致性设计针对系统的新型对抗样本攻击。因此，在临床治疗中，系统一旦受到对抗攻击算法的干扰，那么提供的解释结果必然会影响医生的诊断过程，而错误的诊断可能会对患者生命安全产生严重的后果。此外，由于事后解释的近似性，有时医生甚至可能会被误导引发错误的诊断，进而给患者带来致命的威胁。

其三，歧视性问题。在医疗人工智能的研发设计阶段，由于训练数据偏倚等原因而将种族、性别、阶层偏见等伦理问题带入模型中，从而导致人工智能在医疗应用中出现歧视性问题。例如，一种用于诊断乳腺癌的人工智能系统认为黑人女性患乳腺癌的风险更高[17]。可解释性通过提高模型的透明度而被认为是防范系统偏见的有效手段，但是研究发现，目前的可解释性方法并未达到预期目标，并且对解释的依赖甚至会降低我们对这些歧视行为的警惕[18]。医疗歧视引发的临床风险对患者的生理和心理都造成了严重的伤害，而且这种不平等很可能会加剧社会偏见，从而引发重大的道德问题。

其四，患者医疗自主权问题。所谓患者医疗自主权，是指在医患关系中，以实现患者自由意志为目的，基于患者的自主能力而对生命、健康等具体人格利益进行选择与决定的抽象人格权[19]。也就是说，以人工智能为辅助诊疗时，患者有权知道系统的局限性、决策的合理性等相关方面的问题。对于可解释性的定义，学术界尚没有统一的标准，这是由于可解释性并不是一个单纯的技术问题，它是在人工智能与目标用户交互中才得以实现的。但是研究[20]却发现，医疗领域现有的大多数可解释人工智能设计要么只关注算法本身的性能，要么就是侧重于为临床医生提供解释，往往缺乏面向患者的可解释人工智能设计实践。所以，对于绝大多数没有专业知识背景的患者来说，人工智能医疗决策的根据依旧无法理解。患者不知道输入数据的来源、预测模型的局限性、系统什么时候会出错等相关方面的信息，患者无法在医疗中得到有效的沟通，只能被动地去接受。因此，在这个过程中，患者的医疗自主权就受到了挑战。

2 可解释人工智能医疗应用伦理问题的根源剖析

鉴于以上可解释人工智能医疗应用伦理问题表现，寻求问题的根源尤其重要。将从技术性、社会性、主体性三个方面探寻问题的根源所在。

2.1 技术性根源：可解释性技术的负面效应

可解性技术具有两面性，它在推动医学发展与进步的同时，也难以摆脱其固有缺陷所带来的负面效应。其一，医疗人工智能系统的准确性与可解释性之间存在着一种权衡。即模型的准确性越高，可解释性就越高；相反，模型的准确性越低，可解释性就越低。因此，对于事前解释的医学人工智能系统来说，将自己限制在充分可解释性的算法中，就意味着要以牺牲医疗准确性为代价；其二，可解释技术给对抗攻击带来了有利的条件。无论是事前解释技术还是事后解释技术，都给攻击者提供了可乘之机。一些不法分子会利用可解释技术探测医疗模型的漏洞，进而干扰原始医疗模型和解释模型，严重威胁着医疗人工智能系统自身的安全；其三，

由于目前事后解释技术都是尝试采用近似的方法来模拟模型的决策行为，以从全局的角度解释模型的整体决策逻辑或者从局部的角度解释模型的单个决策结果。所以，解释过程往往不够精确，解释方法给出的解释结果无法准确地反映待解释模型的实际运行状态和真实决策行为，而且据最新研究表明[16]，攻击者还可以利用二者之间的不一致性设计针对系统的新型对抗样本攻击，从而严重威胁着可解释系统在实际医疗应用中的安全。

2.2 社会性根源：相关道德原则和立法的缺失

有关可解释人工智能医疗应用方面的道德原则与立法的缺失是伦理问题产生的社会原因。首先，道德原则的不完善。在可解释人工智能医疗应用与管理方面还没有形成一套统一而又完善的道德规范体系，因此可解释人工智能医疗应用处于一种无规范可依据的失范状态。各国对可解释人工智能医疗应用的认识不同，从而造成了对可解释人工智能医疗应用的道德评价标准的多样性，这些不同的冲突导致了一系列的伦理问题；其次，有关可解释人工智能医疗应用的立法的滞后同样造成了一系列的道德伦理问题。面对层出不穷的可解释人工智能医疗应用伦理问题，法律显现出滞后性。在我国，可解释人工智能医疗应用并未形成一套完善的法律体系，如果可解释人工智能在医学诊疗中发生医疗事故，那么谁该为此负责？如果医院收集和存储的患者信息遭遇泄露，又该由谁负责？

2.3 主体性根源：各利益相关者的利益多样性

在复杂的技术系统中，可解释技术的负面效应是造成可解释人工智能医疗应用伦理问题的直接原因，但是技术离不开社会，可解释技术的开发和应用是基于社会中的各个组织与用户来实现的，在这个新型系统中的各个主体才是带来医疗伦理问题的内在根源所在。首先，一些道德感不强的个体为了谋取个人私利，利用可解释性技术的缺陷攻击人工智能医疗模型，践踏医疗伦理道德规范；其次，可解释性技术设计主体为了获得某些利益，或者为了表达自己的一些主观观点而设计带有偏见性的解释方法。比如，根据 Linardatos 等[21]的研究，目前现存的可解释性方法无论它们应用到医疗过程的哪个步骤，都过于关注特定群体层面的公平，而忽视了其他群体；最后，由于可解释性与人的认知密切相关，导致对于可解释的评估还未形成一个科学的评价标准。开发人员往往是站在开发者的角度创建解释，因此，对于非专业用户来说，解释常常难以被理解。

3 可解释人工智能医疗应用伦理问题反思与应对

尽管可解释人工智能医疗应用的伦理问题不可避免，但我们不能止步不前，我们需要探讨应对之法。

3.1 强化技术创新

可解释性技术的负面效应是可解释性 AI 医疗应用伦理问题产生的技术根源，但其伦理问题的解决也需要强化可解释性技术的创新发展，借助技术手段消解伦理之困。首先，对于事前解释来说，其面临的最大挑战是消除模型可解释性和准确性之间的矛盾。因此，未来在事前解释方法的设计上，应该重点平衡二者之间的矛盾。其中，一种直观的方法是构建基于因果结构的机器学习模型，使模型具备发现数据间因果关系的能力；其次，对于事后解释来说，其面临的最大挑战是近似的解释往往无法正确反映待解释模型的实际运行状态和真实决策行为。未来一个有前景的潜在研究方向是设计数学上与待解释模型等价的解释方法或解释模型[16]；最后，二者同时面临的挑战是来自外部的对抗攻击。对于研究人员来说，未来同样可以利用对抗训练增强模型的鲁棒性。

3.2 强化可解释人工智能医疗应用伦理教育

既然可解释人工智能医疗应用中的相关主体是导致伦理问题的深层根源，那么加强相关主体的伦理教育是解决问题的必由之路。因为技术人员参与了人工智能的全程开发，数据收集、算法设计、性能测试等都是由技术人员操作，所以技术人员的道德品质将直接影响到目标用户的相关权益。因此，在可解释 AI 设计的过程中，加强对设计人员的伦理教育对减少医疗伦理失序问题尤为重要。第一，鼓励技术人员参加可解释人工智能相关的伦理研讨会。在系统开发的过程中，研究人员往往都是把重点放在算法本身，但是，可解释人工智能并不是简单的技术问题，它还涉及一系列伦理方面的问题。参加相关的伦理研讨会有助于提高设计人员

在开发过程中的伦理意识，减少歧视偏见等问题的产生。第二，将伦理教育融入开发人员的培养之中。例如，在大学阶段开设可解释人工智能伦理教育的相关课程，举办多种形式的人工智能伦理专题讲座。

3.3 强化对医护人员的培训和跨学科合作

加强对医护人员的培训和跨学科合作有助于可解释人工智能医疗应用伦理问题的产生。调查显示[22]，大多数患者认为人工智能辅助诊断是值得信赖的并且可以提高医生的诊断能力，然而所有的调查者都坚持让医生来解释人工智能辅助诊断。作为医生，不一定需要精通计算机科学，但是他们应该充分了解与AI医疗相关的概念、技术和伦理挑战，这些知识能够使他们在工作场所批判地检视人工智能，可以向患者清楚地传达人工智能的局限性，明确需要注意的风险和危害，从而赋予患者更多的自主权和决策权。从长远来看，我们还应该将人工智能知识作为专业医疗课程的一部分，作为住院医师培训计划的一部分，教授医生医疗人工智能相关的知识，鼓励他们参加关于人工智能伦理的研讨会。此外，还需加强跨学科合作，因为人工智能最终面对的用户是人类，基于此，人工智能不是一个单纯的技术问题，它涉及心理等社会学科，因此我们需要优先考虑跨学科合作。例如，人工智能科学家需要医疗专家的指导来选择医疗上重要并且生物学上可信的保健应用，而医疗专家则需要人工智能科学家的指导来选择概念上精心设计和技术上可解决的预测问题。

3.4 完善相关道德原则和法律法规

相关道德原则和立法的缺失是可解释人工智能医疗应用伦理问题产生的重要原因之一，而创建完善相关道德原则和法律法规是解决问题的必由之路。这就需要创建一个通用的可解释人工智能医疗应用的道德标准，确保应用主体中的组织与个人对可解释人工智能医疗应用原则有着共同的认识，有助于减少在数据采集和算法设计过程中因利益多样性而产生的歧视与偏见。另外在相关法律法规中，一方面，要明确医疗责任主体具体应该如何认定，防止相关人员躲避医疗责任；另一方面，要制定针对患者隐私保护方面的法律法规，避免泄露患者个人隐私，进一步增强人们对人工智能的信任。此外，对于为了谋取私利，利用可解释技术攻击预测模型，肆意践踏医疗伦理道德规范的不法分子，法律要进行严厉的打击。国家以及政府部门应该尽快制定出有关可解释人工智能医疗应用的发展政策，只有由政府及其相关部门为可解释人工智能医疗应用的发展提供政策保证和法律监督，才能够保证我国在可解释人工智能医疗应用伦理监管方面处于领先位置。随着可解释人工智能医疗应用方面有关道德原则和法律法规的不断健全，可解释人工智能将会迎来更好的生存与发展契机。

3.5 实现个性化的交互式解释

实现个性化的交互式解释有助于提升用户对人工智能系统的信任。随着社会发展与技术的进步，未来可解释人工智能将在更加丰富的医疗情境下被更为广泛的人群所使用。不同的人群、不同的医疗情境对人工智能解释形式与内容的期望都是不同的[23]，因此个性化的交互式解释是非常有必要的。这就要求在开发可解释人工智能时，应当充分考虑人这一主体，充分调研目标用户的背景与需求，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，用户才能被赋予更多的自主权和决策权，进而增强对可解释人工智能的信任。

4 结论与展望

可解释人工智能在推动医学向前发展的同时也引发了不同以往的伦理挑战。审视未来的可解释人工智能医疗应用，既要关注可解释性技术发展为人工智能医疗应用带来的有利条件，更需要研究者、管理者等认识到可解释人工智能医疗应用所处的现实困境，为可解释人工智能医疗应用伦理问题的解决提供可行路径，共同推动安全可靠的可解释人工智能在医疗领域的发展。

〔参考文献〕

[1] Van Lent M, Fisher W, Mancuso M. An explainable artificial intelligence system for small-unit tactical behavior[C]//AAAI 2004: Proceedings of the National Conference on Artificial Intelli-

gence. AAAI Press, 2004: 900-907.

[2] Gunning D, Aha D W. DARPA's Explainable Artificial Intelligence (XAI) Program [J]. Ai Magazine, 2019, 40(2): 44-58.

[3] 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.

[4] Nafisah S I, Muhammad G. Tuberculosis detection in chest radiograph using convolutional neural network architecture and explainable artificial intelligence [J]. Neural Computing and Applications, 2022: 1-21.

[5] Thimoteo L M, Vellasco M M, Amaral J, et al. Explainable artificial intelligence for COVID-19 diagnosis through blood test variables[J]. Journal of Control, Automation and Electrical Systems, 2022, 33(2): 625-644.

[6] Curia F. Cervical cancer risk prediction with robust ensemble and explainable black boxes method[J]. Health and Technology, 2021, 11(4): 875-885.

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

[8] Holzinger A, Biemann C, Pattichis C S, et al. What do we need to build explainable AI systems for the medical domain? [J]. arXiv preprint arXiv, 2017: 1-28.

[9] 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.

[10] 王星，刘晓燕. 医疗大数据环境下的疾病预测模型研究[J]. 制造业自动化，2022，44(7)：24-27.

[11] 陈园琼，邹北骥，张美华，等. 医学影像处理的深度学习可解释性研究进展[J]. 浙江大学学报(理学版)，2021，48(1)：18-29，40.

[12] 陈权，李莉，陈永乐，等. 面向深度学习可解释性的对抗攻击算法[J]. 计算机应用，2022，42(2)：510-518.

[13] Riccardo G, Anna M, Salvatore R, et al. A Survey Of Methods For Explaining Black Box Models [J]. ACM Computing Surveys, 2018, 51(5): 1-42.

[14] Babic B, Gerke S, Evgeniou T, et al. Beware explanations from AI in health care[J]. Science, 2021, 373(6552): 284-286.

[15] Rajpurkar P, Irvin J, Zhu K, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning[J]. arXiv preprint arXiv, 2017.

[16] 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.

[17] Parikh R B, Teeple S, Navathe A S. Addressing bias in artificial intelligence in health care [J]. Jama, 2019, 322(24): 2377-2378.

[18] Ghassemi M, Oakden-Rayner L, Beam A. The false hope of current approaches to explainable artificial intelligence in health care [J]. The Lancet. Digital health, 2021, 3(11): e745-e750.

[19] 张娟. 患者自主权：内涵、困境及突破——以马克思人学交往理论为分析视角[J]. 福建论坛(人文社会科学版)，2018(3)：75-82.

[20] He X, Hong Y, Zheng X, et al. What Are the Users' Needs? Design of a User-Centered Explainable Artificial Intelligence Diagnostic System [J]. International Journal of Human-Computer Interaction, 2022: 1-24.

[21] Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable ai: A review of machine learning interpretability methods [J]. Entropy, 2020, 23(1): 18.

[22] Fink C, Uhlmann L, Hofmann M, et al. Patient acceptance and trust in automated computer-assisted diagnosis of melanoma with dermatofluoroscopy[J]. JDDG: Journal der Deutschen Dermatologischen Gesellschaft, 2018, 16(7): 854-859.

[23] 吴丹，孙国烨. 迈向可解释的交互式人工智能：动因、途径及研究趋势[J]. 武汉大学学报(哲学社会科学版)，2021，74(5)：16-28.

收稿日期：2022-06-30
修回日期：2022-09-21　　(编　辑　商　丹)