以人为本的可解释智能医疗综述

宋淑超[1,2)], 陈益强[1,2)*], 于汉超[3)], 张迎伟[1,2)], 杨晓东[1,2)]

[1)] (中国科学院计算技术研究所移动终端与新型计算北京市重点实验室 北京 100190)
[2)] (中国科学院大学 北京 100190)
[3)] (中国科学院前沿科学与教育局 北京 100864)
(yqchen@ict.ac.cn)

摘 要: 随着人工智能(artificial intelligence, AI)的高速发展, “黑盒”模型已逐渐展示出逼近甚至超越人类的能力, 尤其在智能医疗等高风险应用场景中, 其可解释性是用户在应用中信任和理解 AI 的关键基础. 虽然已有工作提供了大量事前和事后的 AI 可解释方法, 但大都采用通用型解决思路, 未考虑不同用户在不同场景下多维度的理解和信任需求. 以人为本的 AI 可解释方法能够针对用户实际需求对 AI 模型进行可解释分析, 近年来逐渐受到国内外学者的关注. 文中聚焦智能医疗应用, 对近 5 年人机交互国际顶级会议的文献进行分析, 回顾现有辅助诊断、辅助用药、未病预警方面以人为本的 AI 可解释方法及系统, 从决策时间花费、用户专业度和诊疗工作流程 3 个维度梳理和定位可解释需求的系统性方法, 得出 4 类典型用户画像和对应案例; 并从考虑资源受限、不同用户的多样需求、与现有流程结合 3 个方面, 为如何设计可解释的医疗辅助诊断系统提出建议.

关键词: 可解释人工智能; 临床决策支持系统; 智能医疗; 人智交互; 用户体验
中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2024.2024-00052

Review of Human-Centered Explainable AI in Healthcare

Song Shuchao[1,2)], Chen Yiqiang[1,2)*], Yu Hanchao[3)], Zhang Yingwei[1,2)], and Yang Xiaodong[1,2)]

[1)] (*Key Laboratory of Mobile Computing and Pervasive Devices, Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190)
[2)] (*University of Chinese Academy of Sciences, Beijing* 100190)
[3)] (*Bureau of Frontier Sciences and Education, Chinese Academy of Sciences, Beijing* 100864)

Abstract: With the development of Artificial Intelligence (AI), “black box” models have demonstrated significant capabilities that now approach, or even surpass, human performance. However, ensuring the explainability of AI is crucial for users to trust and understand its applications in their daily lives, particularly in high-risk scenarios like healthcare. Although previous research has introduced numerous direct and post-hoc explainable AI methods, many of them adhere to a “one-fits-all” approach, disregarding the multidimensional understanding and trust requirements of diverse users in different contexts. In recent years, there has been growing attention from researchers worldwide towards human-centered explainable AI, which aims to provide explainable analyses of AI models based on the specific needs of users. This article examines literature reviews published over the last five years at top-tier global conferences in the field of human-computer interaction, with a specific emphasis on healthcare. It reviews existing human-centered, explainable AI methods and systems used for computer-aided diagnosis, com-

收稿日期: 2024-01-23; 修回日期: 2024-03-28. **基金项目:** 国家重点研发计划(2021YFC2501202); 国家自然科学基金(62202455, 61972383, 62302487); 北京市科学技术委员会资助项目(Z221100002722009). **宋淑超**(1995—), 女, 博士研究生, CCF 会员, 主要研究方向为人机交互、智能医疗; **陈益强**(1973—), 男, 博士, 研究员, 博士生导师, CCF 会员, 论文通信作者, 主要研究方向为普适计算、人机交互; **于汉超**(1985—), 男, 博士, 副研究员, CCF 会员, 主要研究方向为人机交互、人工智能与交叉科学; **张迎伟**(1992—), 女, 博士, 副研究员, 硕士生导师, CCF 会员, 主要研究方向为脑机接口、人机交互; **杨晓东**(1991—), 男, 博士, 副研究员, 硕士生导师, CCF 会员, 主要研究方向为普适计算、人机交互.

puter-aided treatment, and preventive disease warning. Based on this review, it explores and identifies explainability needs from three perspectives: decision time constraints, user expertise levels, and diagnosis workflow processes. Additionally, the article lists four classic user persona types along with respective examples and provides suggestions for designing explainable medical diagnostic systems, considering resource constraints, varying user needs across different stakeholders, and integration with existing clinical workflows.

Key words: explainable artificial intelligence; clinical decision support system; healthcare; human-AI interaction; user experience

随着算力、算据、算法及机器学习理论的高速发展，临床决策支持系统(clinical decision support system, CDSS)等人工智能(artificial intelligence, AI)产品的应用极大地提升了临床工作效率与人们生活质量. 然而，机器学习算法的性能通常与其可解释性成反比关系[1]. 特别是目前普遍采用的深度学习方法虽然能够大幅提升模型性能，但其不可解释的黑盒性质，导致AI决策结果存在透明度低、无法回溯等问题[2-3]. AI 决策在医疗领域应用具有决策风险高、专业要求高和用户种类多等特点，用户对 AI 决策易产生不理解、不信任和缺乏掌控感等困难，使得 AI 决策日益增长的精度需求与高精度 AI 难解释间的矛盾尤为突出. 此外，国内外相关法律法规也提出对AI相关产品可解释性的要求. 例如，我国在 2021 年 6 月 10 日颁布的《个人信息保护法》[4]要求个人信息处理者进行自动化决策时，应当保证决策的透明度；欧盟的《通用数据保护条例》[5]与美国的《美国数据隐私和保护法》(草案)[6]也对算法透明度和决策可回溯性提出对应要求. 因此，可解释 AI(explainable AI, XAI)已成为国内外研究的热点方向之一[7].

近年来，虽然已有大量针对黑盒 AI 的可解释分析方法[8]，但是实际应用中的 CDSS 等 AI 决策系统仍存在透明度低的问题，难以真正应用于临床[9]. 这是由于传统的“一套解决所有”的方法无法满足不同场景中不同用户的需求[10-11]，需要通过以人为本的思想对 AI 的临床决策提供适度且适量的解释，解决用户需求千人千面的问题. 由此，以人为本的 XAI 系统设计及应用逐渐成为近年来人机交互领域中 ACM CHI Conference on Human Factors in Computing Systems (CHI)和 ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)等顶级学术会议广泛探讨的问题之一，相关工作涉及不同的疾病(如帕金森病[12]、急性心肌梗死[13]等)、不同的数据模态(如文字[14]、图像[15]、语音[16]等)，以及不同的应用场景(如图像辅助分析[17]、病情辅助诊断[10]、用药建议[18]等).

目前，已有工作主要针对特定场景中的智能医疗 AI 应用进行优化改进，但缺少可梳理医疗场景 AI 可解释需求的系统性方法，即在何时何地提供解释、提供何种解释以及如何提供解释. 因此，本文收集、整理并分析了近 5 年发表于 CHI 和 CSCW 上与 CDSS 相关的文献，系统地梳理了不同用户在智能医疗应用的可解释需求；从决策时间花费、用户专业度和诊疗工作流程 3 个维度进行展开相关综述，三者分别决定了“2H2W”问题，即提供解释的数量(how much)、类别(what)、地点(where)和方式(how)；介绍了如何将该方法应用于 CDSS 的 AI 可解释设计，并针对可解释需求和典型用户画像进行象限分析；最后为后续设计可解释 CDSS 提出建议.

1 相关工作

1.1 XAI 技术

XAI 技术最早可以追溯到 20 世纪 70 年代的专家系统[19]，从 20 世纪 80 年代的贝叶斯网络[20]到 21 世纪的推荐系统[21]，它一直对如何理解 AI 起着举足轻重的作用[22]. 当前，随着机器学习和深度学习技术的快速发展，虽然模型精度提升到前所未有的高度，但通常情况下精度与可解释性成反比关系[1]，两者陷入“性能-可解释性交易”的关系[11]，而深度学习处于性能的最高点和可解释性的最低点. 因此，伴随深度学习方向一起发展的还有 XAI 技术，如图 1 所示，在 https://dblp.org 网站搜索题目中包含“explainable AI”或“explainable artificial intelligence”的文献，其数量从 2017 年的 11 篇一路上升至 2023 年的 565 篇，近年来，XAI 在医疗领域中的研究也有大幅上涨[8].

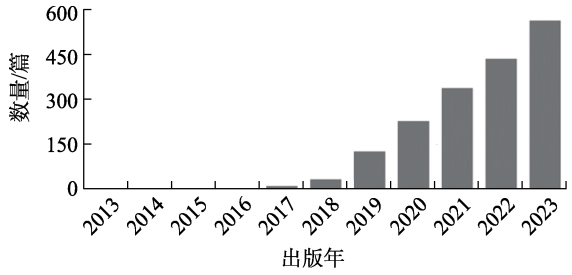


图 1　2013—2023 年 XAI 论文发表数量

早期, XAI 有 2 个主流的英文术语: Explainable AI 和 Interpretable AI. 学者一直未能统一这 2 个概念的差别与意义[23], 但两者的共识部分是 XAI 是为了让人们更好地理解模型或算法[11]. Yarmand 等[24]对 Explainability 和 Interpretability 这 2 个词给出定义, Interpretability 是将抽象概念(如分类类别)映射到人类可以理解的领域, 而 Explainability 是将模型决策链接到人类理解[25]. 基于此, 较为主流的观点认为: Interpretable AI 多指本身就可以被解释的白盒模型, 而 Explainable AI 指用于解释黑盒模型的方法[26]. 近年来, XAI 已经更多地采用 Explainable AI 的表述, 而 Interpretable AI 则成为其中的一个子集. 因此, XAI 可根据解释的类型分为“直接解释的模型”和“建模后解释的模型”2 类.

1.1.1　直接解释的模型

自身可被解释的白盒模型被称为直接解释的模型, 如决策树、线性回归等[8]. 该类模型可以通过统计学或可视化方式归纳出其规律, 并以直观的方式帮助人们理解其作出决策的原理. 因此, 该类模型也是模型相关的可解释方法, 即一种解释方法对应一种特定模型. 直接解释的模型可用于一些规律简单的场景中, 如决策树可用于心理测试评分中, 但通常情况下无法适用于较复杂的场景[25].

1.1.2　建模后解释的模型

建模后解释的模型指对本身不可解释的黑盒模型进行解释的一类方法, 通常是模型无关的. 由于其解耦了可解释模型与原模型, 因此具有更好的普适性. 该方法分为局部解释和全局解释 2 类. 局部解释旨在揭示一个具体推理例子时, 模型为何会做此推断, 帮助人们判断该预测是否值得信任, 如不同特征的贡献值等; 全局解释则从算法整体的角度解释算法习得的知识和能力, 帮助人们理解算法的性能和不足, 如分类问题中模型认为某一类别有什么特征.

(1) 局部解释

局部解释可以帮助人们了解不同特征对结果的影响程度. 如在金融场景中, 信用卡申请失败的用户, 可以通过局部可解释得知他的哪一项数据(如银行卡数量)对该结果影响最大, 以及改变后是否会改变申请结果. 局部解释设计的基本思想为: 若要知道模型在具体实例中如何判断, 那么这个问题就可以转变为每个特征对判断结果的影响. 所以该方法一般是通过对特征进行移除或改变, 了解该特征对结果的影响程度. Ribeiro[27]提出的 LIME(local interpretable model-agnostic explanations)是局部解释的一种代表性方法, 它首先通过随机扰动生成扰动样本, 再利用扰动的样本和该样本在模型中预测的结果形成一个新的数据集, 然后在这个新数据集上训练一个局部线性模型来达到可解释[28]. 因为 LIME 可以用于各种不同模型, 并识别其决策影响最大的输入区域, 所以很受欢迎. 另外一种代表性方法是 2017 年 Lundberg 等[29]提出的 SHAP(Shapley additive explanations), 根据博弈论中的沙普利值得出. 沙普利值通过计算合作中各贡献者对集体的贡献度作为利益分配的参考值, SHAP 将模型中各特征模拟为各贡献者来计算它们对集体的贡献度. 因为有严格的数据证明基础, 理论上保证了 SHAP 的有效性, 但因其计算量大被很多学者认为效率较低.

(2) 全局解释

与局部解释不同, 全局解释是解释整个模型的归因思路, 而不是某个样本或部分样本, 如用于说明某区域房价与不同特征(如周围商圈数量)的关系, 帮助理解模型并判断是否正确归因. 全局解释的一种常见思路如下: 用黑盒模型预测的预测标签替换原数据集中的标签, 生成新的数据集, 并对该数据集训练一个自身可解释的代理模型(如决策树、基于规则的模型等). 全局解释可以大致解释模型的决策思路, 但因为学习了黑盒模型的知识而不是原始数据集, 所以存在不准确性. 其中代表性方法是知识蒸馏[30], 通过迁移学习的思路, 将结构化的逻辑规则转入神经网络的权值中[31], 实现将烦琐模型中的知识提取到简单可解释模型中的目的.

1.2　CDSS

CDSS 是帮助使用者更好地做决策的一种计算系统[32], 也是辅助诊断的一种常见形式, 其将医学知识与患者数据融入算法中, 系统通过患者的信息进行推断并给出相应的建议, 以提升诊疗的质量与效率. CDSS 通常用来解决 3 种任务: 辅助医生对患者进行病情诊断[33]、提供用药方案或治疗建议[34]和预警未发生但高可能的病症或用药的预期结果[35].

CDSS 有时作为独立软件出现, 有时为了更好地融入医生已有工作流程, 被嵌入到医院的电子医疗信息(electronic medical record, EMR)系统中[36-37].

CDSS 最初多以基于规则或基于经验的形式出现, 其从医生或医学文献中获得经验和知识[38], 并基于此进行设计. 1986 年问世的 DXplain[39], 基于因特网对患者输入的病情描述与诊断结果进行简单的匹配, 但基于医疗知识与经验的方式难以适应时常更新的医疗知识, 也难以解决经验未涉及的案例. 由此, 基于机器学习和深度学习的 CDSS 在相应技术发展后逐渐面世[40]. 近年来, 随着 ChatGPT 的成功应用, CDSS 开始尝试利用大数据训练大模型以达到更好的模型性能. 然而, 由于 CDSS 的设计工作仍然是以算法人员为主的团队进行, 因此优秀的模型性能很难直接赋能医疗产业, 技术与应用之间的缺口需要由人机交互技术填补[41], 最终达到 Osheroff 等[42]提出的 CDSS 设计的目标“5 个正确”: 对正确的人, 在工作流程正确的时间点里, 通过正确的途径, 以正确的形式, 提供正确的信息.

1.3 CDSS 在人机交互领域的挑战

虽然 XAI 已经得到了很好的发展, 但在实际临床中应用仍面对重重困难. Abdul 等[22]和 Doshi-Velez 等[43]认为, XAI 不再受制于算法的发展, 而是缺少对于人本因素的考量, 从而产生可用性和可解释性差的问题; 并进一步引发透明度低、信任度低、可回溯性低和公平性难考量等问题[5]. 在人机交互领域, 一套适用所有场景的解决方案不再适用于大部分场景[11], 需要考虑场景与用户需求的复杂性, 从而设计以人为本的系统.

CDSS 作为医疗场景中的主要应用类型之一, 自然也不例外[44-45]. (1) 因为需要辅助决策者做决策, 所以如何让用户对系统产生信任[41]是其面临的第 1 个人机交互问题. (2) 由于医疗领域的决策风险很高, 而大部分 AI 都不能保证 100%的正确率, 且不能为人类的决定负责, 因此不能让用户对 CDSS 产生过度依赖, 避免本身可以辨别但因盲目信任 AI 而导致决策错误的情况, 适度的信任才是 CDSS 需要追寻的目标[46]. (3) 医疗场景中涉及很多不同的利益相关者, 如资深医生、新手医生、护士、看护、患者和患者家属等, 他们有些需要专业深入的解释, 有些需要浅显的告知, 又或是简单的二次确认. 由于一款 CDSS 经常要为不同利益相关者提供服务, 因此它应该根据不同用户的不同需求给出对应解释[47-48]. (4) 医疗系统中的人员有固有解决问题的方式和工作流程[10], 给出的解释需要以他们习惯的方式提供, 并保证提供的帮助比造成的应用负担(不符合认知水平的理解负担、需要占用较多额外时间查看的时间成本等)大, 才能真正被应用并提供价值. 除了上面提到的问题, 如何处理医疗人员与 AI 提供的预测发生冲突的情况[49]、如何让用户快速了解其 CDSS 模型的能力和不足[50]、哪些相关信息是可以补充哪些是应该舍弃的[51]等, 都是人机交互领域非常关心的实用问题.

2 以人为本的 CDSS

本文检索了近 5 年人机交互领域中 CHI 和 CSCW 上与 CDSS 和 XAI 相关的工作, 以标题中含有关键词“Healthcare”或出现在“Healthcare”相关板块中且摘要与 CDSS 相关为搜索条件, 共收集 100 余篇文献. 其中, 以内容主体是 CDSS 设计或创新, 并包含一定对可解释需求的探索为条件, 筛选相关工作至 21 篇列入综述案例, 剔除了网络社区对心理健康的影响等“泛 Health”主题相关工作; 剔除了不涉及系统开发或设计只有定性调研的工作, 该部分主要集中在 CSCW 会议中, 不被作为案例分析, 但相关工作启示会被包含在可解释需求分析中.

近年来, 人机交互领域的 CDSS 可解释相关工作大部分围绕辅助诊断展开, 其他少量工作是关于辅助用药和未病预警. 在数据类型上, 得益于计算机视觉领域和自然语言处理(natural language processing, NLP)领域相关技术的高速发展, 多数 CDSS 以图像和多模态数据为主, 其他数据形态包含文本和可穿戴数据等. 表 1 所示为将 21 个综述案例[10,13-15,18,24,41,49-50,52-63]按照任务类型进行分类, 并根据数据类型、应用场景和目标用户进行划分, 对其可解释需求进行简单介绍.

3 CDSS 中的可解释需求与典型应用

CDSS 中, 以人为本的可解释问题有很多种类, 单一 CDSS 工作只会涉及其中一部分. 本文归纳其共性并总结出拆解需求的参考方向, 帮助其他学者分类和理解现有工作, 以及梳理和指导未来工作. 最终, 本文提炼出 3 个拆解 CDSS 可解释需求的维

表 1　近 5 年以人为本的可解释临床决策系统简要介绍

任务	数据类型	系统名称	应用场景	目标用户	可解释相关需求
辅助诊断	图像	Prostate cancer grade predictions[50]	前列腺癌	病理学医生	是否患病为无理解偏差的分类问题，但严重程度的可视化表达在不同用户的理解中存在偏差
	图像	AI detection of diabetic retinopathy[52]	糖尿病视网膜病变	眼科医生	当对 CDSS 提供的预测结果形成固化偏见时，对后续接收的信息缺乏吸收和思考
	图像	SMILY[15]	体液和组织分析	病理学医生	如何提供医生认为“相似”的案例图像
	图像	Assertiveness-based BreastScreening-AI[53]	乳腺癌	医生(新手和专家)	不同肯定性语气的解释是否影响不同知识储备的医生的信任与判断
	图像	NaviPath[54]	癌症诊断	病理学医生	对于需要放大上千倍看图像的场景，如何判断哪些区域是有价值的以及如何告知医生
	图像	Eye disease diagnosis[55]	眼部疾病	病理学医生	专家对于困难案例存在意见不同的情况，通过提供专家裁决的讨论是否可以提升对问题的理解
	图像	CheXplain[56]	胸部 X 光片	内科、放射科医生	如何通过 XAI 做出决策，如何通过交互让其更好地理解对诊断结果的解释
	图像	Contouring learning platform[24]	放射肿瘤科	新手医生	实习医生需要执业医生的指导和纠错来快速进步，如何在练习平台融入可解释的交互反馈
	文本	Data-driven DST predictions[14]	神经病学、精神病学、临终关怀	医生	AI 给出错预测时，如何辅助医生根据自己的判断做出正确预测(是否该相信 AI 的决策)
	时序数据	Sleep stage classification[49]	睡眠阶段分析	睡眠专家	很多分类任务的标签存在意见冲突，将这些冲突告知并解释给决策者是否会改变其表现
	可穿戴数据	Remote Gait Assessments[18]	多发性硬化的走路障碍	医生	对于需要长期观察和用药的场景，如何收集可穿戴设备数据并以合适的方式呈现，让医生更好、更快地了解并给出建议
	多模态	Early decision aid to diagnose in an Intensive Care Unit[57]	重症监护室	医生	除给定解释外，希望提供辅助解释帮助判断 AI 的能力，如原始数据、归因是连续或独立等
	多模态	AI-based decision support system[58]	中风运动评估	医生、外行人士	外行人士比医生更容易过度依赖 AI 的结果和解释
	多模态	OralCam[59]	口腔疾病	需要口腔诊断的普通用户	用户进行口腔疾病初筛时，需要了解疾病预测的哪些相关信息和辅助知识
	多模态	Dr.AI & Dr.XAI[13]	急性心肌梗死	医生、护士、护理	提供解释是否会增加信任
	多模态	Brilliant AI Doctor[60]	综合性诊断	临床医生	在快速问诊的场景中，如何提供清晰的解释和相关信息让医生产生适当信任并理解
	多模态	Chatbot-based symptom checker[61]	全病诊断	普通用户	诊断机器人提供医生式的说话方式和非医生式的解释，增加系统的透明度和用户的信任
辅助用药	结构化数据	AT Planner[62]	抗抑郁症停药期用药方案	医生、护士、实习医师	停药期过程长且涉及较多利益相关者，因缺少基于标准的用药方案，难以给出统一的预测和解释
	多模态	AI-generated treatment insights for Type 2 diabetes[41]	糖尿病	医生、护士、医生助手、药剂师、看护	对于机器生成的启示，医疗人员是否认为有价值以及哪些启示有价值，能否产生合适的信任
	多模态	DSTs in antidepressant treatment decisions[63]	抗抑郁症	医生、护士、住院医师	用药建议为非常用药物时，对决策不熟悉不自信，需要更多解释
未病预警	多模态	VAD decision processes[10]	心室辅助装置副作用预测	中级和资深医生	医生需要知道是否用 AI 的决策，分别对后续病情有怎样的影响

度：决策时间花费、用户专业度和诊疗工作流程，并分别影响了 CDSS 可解释中的 2H2W 问题.

3.1　基于决策时间花费

2021 年，在 CHI 上，Jacobs 等[63]在关于抗抑郁症的辅助用药 CDSS 中，根据用户研究提出不要让医生决定看什么、相信什么解释，而是提前为他们筛选好可以相信的解释. 虽然该工作到如今只有 3 年的时间，却似乎与当前的主流看法相悖. 当前，学者认为，CDSS 应该退回到助手的角色[64]，把全局信息和辅助判断信息(如置信度)给到医生，让医生自行了解 AI 的能力，并判断哪种情况可以信任 AI. 这种观点上的冲突使得本文对两者进行进一步探索，发现其前提条件存在差异：在 Jacobs 等[63]研究中，该洞见是出自决策时间非常紧张的决策者口中，而目前的主流观点大部分是未考虑决策紧迫度这一因素的.

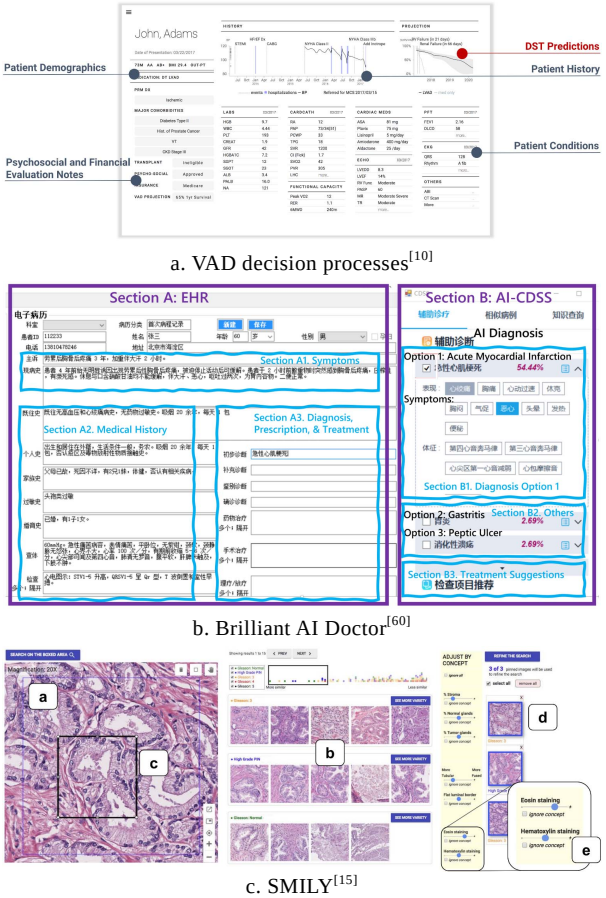
在表 1 其他工作中也得到验证后，本文归纳第 1 个拆解 CDSS 的维度为决策时间花费，即决策者在现有工作流程中，做该决策时花费的时间；并默认可解释 CDSS 的加入不能大幅改变现有流程所

花费时间，否则将会因适应成本过大而失去可用性. 决策时间花费与决策紧迫度和决策难度相关: 一方面，对于急诊科或对生命威胁较大的疾病的患者，需要在紧迫的时间内做决策，而对于一些慢性疾病或对生命威胁较小的疾病的患者，则时间紧迫度相对较低；另一方面，对于诊断较为简单的疾病，可能通过较少的检查和时间就可以确诊，但有些复杂的疾病的诊断，则需要更多的检查和更多轮的讨论，这些都会影响决策时间花费. 决策时间花费会影响解释的数量(how much)，越多的决策时间在设计时可以提供的解释数量就会越多.

对于决策时间花费这一维度，通过列举决策紧迫度不同的 3 个典型案例，以直观地感受该维度对 CDSS 中解释部分的影响.

(1) Yang 等[10]指出，在美国，关于人工心脏安装建议的 CDSS 在实际应用中几乎都失败了，通过调研发现，在做该决策的资深医生的工作流程里几乎没有机会使用电脑，更无从谈起使用 CDSS；于是他们通过用户研究探寻适合融入该 CDSS 的机会，发现这些关乎生命安全的重大医疗决策通常会召开专门的会议来确认方案，而会上聚集几乎所有与该决策相关的医护人员，从主治医师到住院医师、从护士到护工等. 最重要的是，在 Yang 等调研的 3 家医院中，因为会议上一般以患者在 EMR 系统中的信息作为辅助供大家讨论，所以这次会议是唯一一个资深医生离电脑很近的场景，将人工心脏的安装建议嵌入该系统似乎是最可能让资深医生在已有工作流程中看到的方法；但该会议涉及人员众多且时间短，如何提供精简的解释成为了新的挑战. 在与医护人员进行访谈后发现，他们似乎都对接触过的 CDSS 持消极态度，这是因为过去的 CDSS 缺少使他们信服的有价值的信息，他们希望该系统能够借助丰富的数据库提供不同于医护人员已有的角度的信息. 基于对工作流程和时间稀缺的考量，人工心脏的安装建议被嵌入到 EMR 系统的右上角(如图 2a 红色字体所示)，以与系统其他基础信息统一的形式降低医护人员对 CDSS 已有的抵触情绪和对自己权威被质疑的担心. 通过可视化手段，简化了信息接收的认知成本，并通过曲线的变化直观地体现了安装人工心脏的相对风险. 该工作针对既定工作流程难以撼动并时间稀缺的情况，参照泛在计算的概念提出“Unremarkable AI”设计方法，并通过用户访谈证明了其有效性.

(2) 与美国对比，中国的问诊似乎普遍存在着



a. VAD decision processes[10]

b. Brilliant AI Doctor[60]

c. SMILY[15]

图 2 将决策时间花费作为设计考量的 3 个典型 CDSS

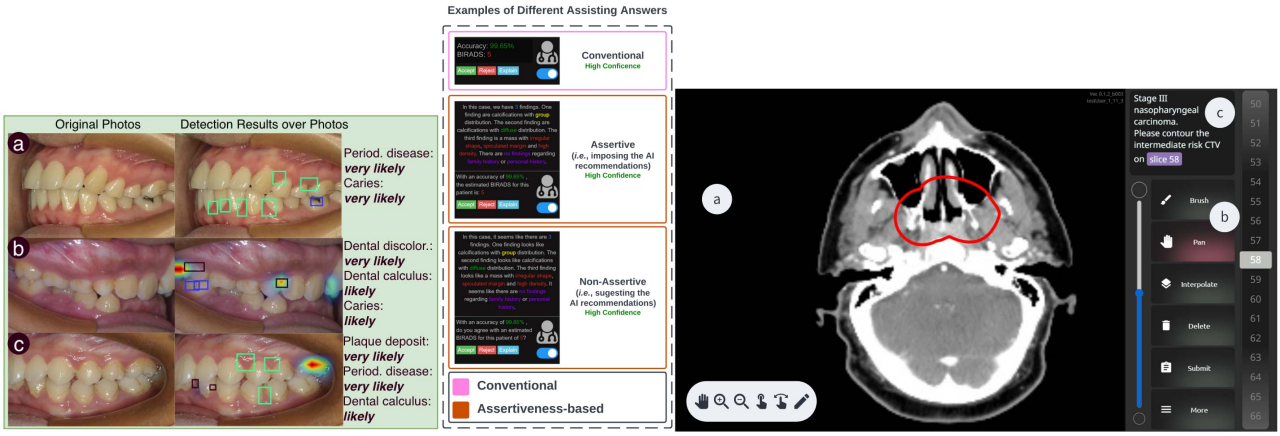
更高的决策紧迫度，中国农村因为医疗资源更加紧张，决策紧迫度比大环境只增不减. 2019 年，面向农村医疗场景的 CDSS 智能助手 Brilliant AI Doctor 问世，但却没有真正赋能农村诊疗；在其问世半年后，Wang 等[60]对首批 6 家使用该系统的医院中的 22 名医生进行调研，发现其中存在的巨大冲突：农村诊疗中医患比极低，同时还存在很多需要开证明转到上级医院的需求，医生每天需要以极快的速度进行患者的面诊；但在设计 CDSS 时，学者预期了时长约 10 min 的会诊，以及多项步骤以期形成闭环，包括对患者问询、电子健康记录(electronic health record, EHR)系统的信息录入、做出诊断决策、将诊断决策录入系统、做出治疗决策、将治疗决策录入系统和再次与患者沟通. 而现实情况是，医生完成诊断每位患者的平均用时只有 2~4 min，因此他们不得不多任务并行，将录入系统与询问或做决策同时进行，甚至将询问与做决策也压缩到同一时间进行. 在这样的多任务并行且高度集中在 EHR 系统的情况下，医生很难仔细查看或思考 CDSS 在同界面给出的辅助信息(如图 2b 的 Section B 所示)，吸收辅助信息的耗时已远远

高出自己做决策的时间成本，所以几乎没有医生使用．受此启发，本文将决策的紧迫度列为设计 CDSS 的首要考虑点，当医生没有时间查看时，其他设计再精巧也无用武之地．

(3) 与前 2 个决策时间紧张的案例相反，医生思维的相似图像(similar images like yours, SMILY)[15]是针对病理学医生设计的一款基于交互的图像内容检索(content-based image retrieval, CBIR)系统．与问诊医生相比，病理学医生的时间相对自由，所以节制的解释数量不是该应用的设计重点；相反，对于图像检索类的 CDSS，计算机天然具备比人类更快、更精细的检索能力和更大的对比数据库，所以在该应用中，如何提供翔实且扎实的解释成为首要目标．实际诊疗中，病理学医生常需要相似病例图像做参考，但大部分 CDSS 提供的相似图像都是以模型认为相似的方式(基于视觉相似度)提供，与医生认为的“相似”存在出入．因此，Cai 等[15]通过医生对相似的认知和需求发现，医生一般会锁定图像局部的图案结构(对应某些病理特征)来查找相似的参考图像；但在查找的过程中需求也会发生变化，有时会在全局信息和局部信息之间切换；最关键的是，医生认为的相似大多基于语义和病理，而图案结构只是这些知识的外化形式．基于此，SMILY 系统通过深度学习技术将图案结构与病理知识进行映射，并设计了 3 种交互方式帮助医生搜索需要的相似图像：圈出关注区域进行搜索、选择结果中认为有用的例子进行二次搜索、通过对病理学相关知识关联度的调节搜索(如图 2c 所示)．该工作从可用性、理解度、使用意愿等多方面，验证了其有效性和价值．

3.2　基于用户专业度

医疗场景中，CDSS 涉及很多不同的利益相关者，如资深医生、新手医生、支撑部门医生(如药剂师和麻醉师)、护士、看护、患者、患者家属等，不同用户由于专业知识背景不同，对解释的理解也会有差异，OralCam[59]是一个面向普通用户的口腔病症检测软件，并提供了可视化的诊断解释(如图 3a 所示)．在可用性测试中发现，因为提供的解释内容并非针对外行用户，而是需要一定医学背景知识作为支撑，所以用户虽然建立了信任但表示难以理解解释内容．相对地，Cai 等[50]在前列腺诊断预测系统中发现，病理学医生希望提供更多关于模型的全局信息，如其优劣势和整体设计目标，以更好地了解模型的能力，便于做出更好的判断，该工作中的用户专业度较高，平均问诊年限为 10.8 年，并且要具备一定计算机基础．



a. OralCam[59]　　b. Assertiveness-based BreastScreening-AI[53]　　c. Contouring learning platform[24]

图 3　将用户专业度作为设计考量的 3 个典型 CDSS

基于此，本文归纳出第 2 个拆解 CDSS 的维度——用户专业度，即决策者本身的专业知识储备；并默认当所提供的解释超出决策者知识储备时，会出现理解成本过高而难以理解，而提供的解释远小于决策者专业储备时，会出现无法满足其解释需求的情况．用户专业度与用户的职业及其职业经验年限有关，职业背景与该决策匹配度越高、职业年限和职位越高时，用户专业度相对较高；反之，则相对较低．用户专业度决定了解释的类别(what)，应选择与决策者背景知识相匹配的解释．

对于用户专业度这一维度，本文列举根据专业度不同调节解释口吻和为专业度较低的医护人员设计 CDSS 的典型应用．

(1) 资深医生和新手医生因为知识和经验的

差异，在工作中得到其他医生的意见往往存在不同. 资深医生一般是与其他资深医生讨论病情，得到的都是建议性语气的意见；而新手医生往往是向资深医生请教，得到的建议一般是肯定的、教育性的语气. 归因理论认为，当行为与期望的结果一致时，人们更容易关注自身的需求；否则，会去外界寻找更多信息来决定行为. Calisto 等[53]从该理论得到灵感，提出当 CDSS 对医生提供建议的语气不符合他们日常的预期时，医生是否会因此减少信任从而向外寻找解释? 基于此，通过 NLP 技术将中性的解释转变成带有不同肯定程度语气的解释，再根据医生的专业度给出对应的解释(如图 3b 所示)；通过定性与定量验证发现，对不同专业度的医生给出对应肯定语气的解释，使医生对系统在信任度、理解度和吸引力等方面都有显著提升.

(2) 来自加利福尼亚大学圣迭戈分校的 Yarmand 等[24]受到认知心理学的启发，通过模拟学徒制学习模式提升了实习医生的技能锻炼效率. 学徒制学习模式是认知心理学中的一个理论，认为学习者通过密切监督专家的行为方式获得相关技能和知识是一种效率很高的学习模式，也是实习医生学习高度专业化的项目中常见的课程形式；实习医生往往通过观察教员制定的工作流程获得医疗技能，然后重新构建自己的流程. Yarmand 等还发现，放射肿瘤病的实习医生需要通过学习软件，练习用等高线轮廓绘制准备放疗的肿瘤，但在调研的场景中，实习医生很难获得资深医生在学习软件上的指导，使得学习效率很低；于是，他们通过收集医生数据，对该学习软件进行优化(如图 3c 所示): 对比实习医生绘制的等高线轮廓与专家轮廓的重叠程度，对绘制过度和欠缺的轮廓区域提供具体指导，以及对肿瘤的长期毒性和附近器官的风险提供深度解释. 该工作通过 3 种验证方法，证明了其可提升实习医生独立使用学习软件的效率.

除了上述典型应用外，针对医疗场景中不同利益相关者应该提供不同理解背景的解释信息，经常作为一条启示或未来工作出现在相关研究上[48].

3.3 基于诊疗工作流程

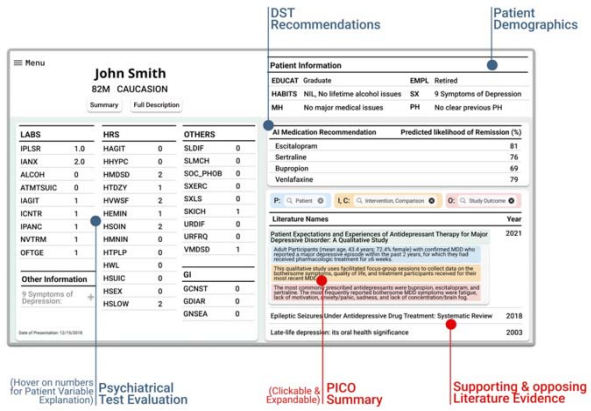
人机交互领域主张泛在的交互，并认为只有融入生活和工作本身才能更好地被应用，这在时间稀缺的医疗领域更是如此，所以要以决策者现有的诊疗工作流程做参考，设计 CDSS 与其可解释部分. 在 Yang 等[14]探索如何让医生根据自己的判断决定是否要听取 AI 意见，并做出正确预测的研究工作中，他们参考医生日常听取他人建议时的处理流程——通过查阅相关文献辅助决策，基于 GPT-3 搜索相关文献并高亮相关论据，并以如图 4a 所示融入日常 EHR 系统的方式呈现给医生，验证了该方法的有效性和接受度.

对此，本文归纳第 3 个拆解 CDSS 的维度: 诊疗工作流程，即根据决策者本身工作中处理该情况的方法，将其融入诊疗的工作流程；并默认当 CDSS 提供的可解释未结合本身的诊疗流程时，则会因适应新工作流程产生的时间精力成本而被放弃使用. 需要注意的是，诊疗工作流程不是像前 2 个维度一样是“多”或“少”的衡量值，而是定性的考虑项. 梳理该维度时，可以请求对目标用户的工作流程进行实际观察，或通过访谈等形式，了解其本身如何做决策以及相关的工作流程. 诊疗工作流程决定了解释的地点(where)和方式(how)，即在工作流程的哪个节点、用什么方式提供解释.

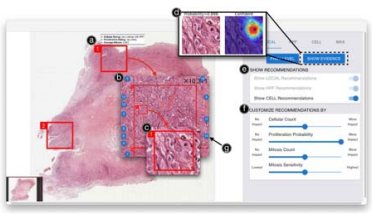
通过基于不同诊疗工作流程的典型案例，展示研究者如何根据应用场景的诊疗流程进行个性化设计.

(1) Yang 等[10]关于“Unremarkable AI”的设计方法在 CHI 2019 上获得最佳论文奖项，该研究引起人机交互领域对医疗场景中工作流程的关注，在何时何地(when and where)提供预测和解释也成为高频讨论的话题. 除了提出让新设计贴近原有工作流程外，他们也在验证环节提出了新问题，有些信息在融入原有流程时遇到了很大困难，但又确实有展示的必要性. 那么如何衡量最佳的融入程度，什么条件下可以在原有流程外创建新的流程，都是目前还未解决的挑战.

得益于计算机强大的计算能力和速度，对于原本多用显微镜观察的病理学场景，也可以很好地通过计算机呈现并替代其中的重复性工作. 如用于癌症诊断的肿瘤切片图像，有着与 4K 屏幕分辨率一个数量级的超高分辨率(4 096×2 160 像素)，怎样在如此高分辨率的图像中寻找有价值的区域来辅助诊断是病理学医生面对的一个难题. Gu 等[54]通过以人为本的设计方法，探究医生本身面临相同问题时的工作流程，发现了他们处理对应问题的几种共性方式——先整体查看图像再放大看细节，然后根据图案结构先整体寻找一遍再细致过筛避免遗漏；模仿医生实际的工作流程，他们设计出 NaviPath 系统(如图 4b 所示). 该系统对搜索的问题区域提供可视化导航与跳转，每部分的判断依据(概率、置信度、显著图等)，以及可根据属性调节的定制化搜索等，为如何以人为本地解决图



a. Data-driven DST predictions[14]



b. NaviPath[54]



c. CheXplain[56]

图 4　将诊疗工作流程作为设计考量的 3 个典型 CDSS

像导航和大分辨率图像搜索问题提供了参考.

(2) 胸部 X 光片分析系统 CheXplain[56](如图 4c 所示)中，以医生理解的方式提供了可视化的判断依据、相似相反病例等，大幅提升了医生对解释的理解度. 但通过用户研究发现，只有专家会在现有工作流程中包含看该类胸部 X 光片的步骤，而初级医生和中级医生的工作流程中则不会有，所以在面向后者的实际应用中也存在困难. 另外，在糖尿病 CDSS 的研究中，医生反馈机器生成的诊断解释对他们是有帮助的，但很难在正式问诊有限的时间内参考这么多信息，建议将其嵌入到预问诊阶段(阅读患者病例等相关信息)，然后进行正式面诊.

4　CDSS 可解释的设计建议

以人为本的 XAI 思想提供了解决 AI 模型信任与应用难题的可行性，但其具体实践仍面临严峻挑战. 突出问题是：与以算法为中心的工作相比，以人为中心的研究缺乏严谨可遵循的逻辑框架，研究人员只能在技术发展的同时探索人类更多的可能和边界. 因此，本文根据已有的相关工作，从 3 个维度分析可解释 CDSS 需求，旨在梳理出一套系统、全面的指导建议，尽可能覆盖目前研究面临的潜在挑战问题，帮助研究人员在设计 CDSS 时可以快速定位需求，并为关键性因素提供保障. 基于第 3 节的典型案例和前沿观点可以发现，Osheroff 等[42]于 2012 年提出的 CDSS 设计目标的 5 个正确至今仍具有重要价值. 本文针对可解释 CDSS 的 3 个分析维度的考量，与 5 个正确目标具有很强的一致性.

4.1　4 类典型用户及可解释建议

在可解释 CDSS 的 3 个关键维度中，不同用户的决策时间花费和用户专业度可进行相对量比较，图 5 所示为根据这 2 个维度信息将用户划分为 4 类(即 4 个象限)，并分别提供了针对性的建议.

第 1 象限. 适用于决策时间多，用户专业度也高的情况. 该类用户一般为资深的病理科医生或药剂师等时间自由度相对高的科室，他们多有丰富的历史资料和数据库，可以利用强大的计算能力帮助其完成重复性工作，并发现更多关键信息. 模型需要提供翔实的解释信息帮助他们快速梳理和判断，提供类似或相反病例等都是帮助他们提高效率的方式.

第 2 象限. 适用于决策时间短，但用户专业度高的情况. 该类用户多为决策紧迫科室的资深医生，CDSS 应退回到辅助者的位置帮决策者查漏补缺、二次确认、提醒易遗漏信息等. 提供的内容需以他们熟悉的表达方式提供，以减少学习成本，精简内容，以免造成无法承受的时间负担.

第 3 象限. 适用于决策时间短，同时用户专业度也低的情况. 此类用户多为经验较少的医生，或相关性较弱的医护人员，对于此类场景，应提供易于理解的关键性判断信息，使用较为肯定的语气帮助他们理解，同时也要精简解释数量，匹配他们的决策时间.

第 4 象限. 适用于决策时间多但用户专业度低的情况. 此类用户多为实习医生或普通(潜在)患者，需要提供大量的背景知识来解释决策，并用对应教育或科普的语气融入认知心理学中对学习过程的理解，匹配他们的较强的求知欲和较充裕的时间.

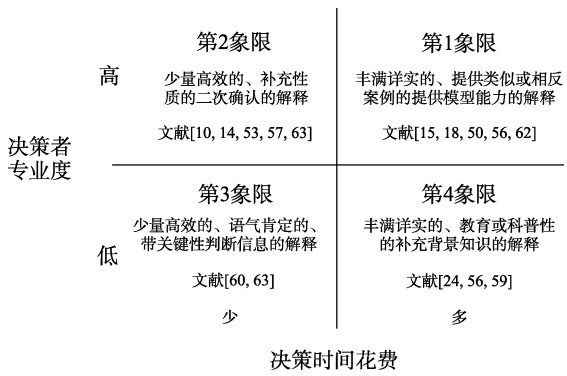


图 5　基于决策时间花费和用户专业度的 4 个象限

4.2 设计 CDSS 的实用建议

4.2.1 考虑受限资源

由于我国的医患比例较低、医疗资源紧张，因此医生的问诊用时往往很短. 医生时间的稀缺性，是设计 CDSS 产品时需要重点考虑的因素[65]. 在时间受限的情况下，需要为医生提供精简的解释信息而不是翔实的数据[10]. 为了让医生在有限的时间内可获取更多的信息，可视化是一种有效手段. 与文字学习相比，可视化学习具有更强的经验积累效应: 将不同模态的解释以不同形式呈现在相同位置或融入同一图例中时，医生经过多次查阅和理解后会形成对图案结构的潜意识判断，从而节省后续的查看时间; 迫于时间压力，医生倾向关注全局的解释信息而非个例的局部解释，通过把握整体知识，可以为后续问诊节省时间，而过于关注局部的解释信息，则需要付出更高的时间成本[63]; 遵循医学问诊逻辑构建的模型和系统，可以节省医生在不同问诊流程之间切换的成本，为医生提供更好的用户使用体验[66]; 过度追求高效也可能带来不良影响，CDSS 需谨慎避免因解释不足而给医生带来错误诱导，严格限制提供低置信度信息.

除了时间资源受限这一普遍问题外，由于医疗资源分配不均衡导致的大型设备紧张、资深医生标注数据紧张等问题[67]，也需纳入 CDSS 设计的考虑范围，避免因此带来的低可用性问题.

4.2.2 考虑不同利益相关者的不同需求

由于医疗场景涉及不同的相关人员，需要考虑各个参与人员的专业背景，以及非直接参与人员可能对系统结果的影响. 因此对于很多同时为不同用户提供预测和解释的 CDSS，可以通过对用户身份进行区分提供自适应的解释. 区分方法既可以是简单的主动选择，也可以融入自然的交互[68]，如通过一些选项的选择偏好来判断，也可以通过提供解释后收集自然的反馈来实时判断用户的背景与解释之间的差距[69]. 对于多模态解释，除了提供模型从多模态输入得出诊断结果的关键决策路径外，也需根据决策者的专业度动态地调整适合展示的解释类型[53]. 此外，可以采取联合设计的方法达到对用户更高的适配度[70]. 当决策者只是单方用户如医生时，联合设计的方法也可以很好地发挥价值，因为医生虽是决策的第 1 个接收者，但决策及提供的解释和建议会通过医生触达更多利益相关者，如患者、患者家属、合作医生、护士等. 有研究表明，让患者在设计初始阶段便参与系统设计，可以增加患者对诊疗建议的依从度[71]，也可能增加患者家属的理解和配合度[72]. 该设计方法可以在初始阶段就充分考虑患者使用偏好，将系统非直接接收者的意愿也纳入考量因素.

4.2.3 与现有诊疗流程相结合

因医疗行业的知识门槛高，医护人员时间紧张，导致固有流程一旦形成，更新将会变得非常困难且缓慢[60]. CDSS 的设计也从最初开发全新的系统，发展为慢慢融入已有诊疗系统中. 但是，受限于医院的隐私保护需求，新的设计很难直接被加入已有流程中[73]，极端情况下甚至难以对已有系统进行实际调研. 对于此类情况，可以考虑跳出已有的流程，通过对医护人员的工作流程进行观察找到融入的时间点和地点; 再针对 CDSS 要解决的问题，重点调研医生在现实中遇到相同问题的解决方法并进行模拟[14]，这种对现实的模仿方法也是人机交互最早的实现方式; 同时，可以结合认知心理学的学科知识，在了解医生的处理方式后深入了解其动机和认知过程. 对于可懂度非常重要的 CDSS 来说，模拟认知心理学中的一些学习模型[24]，应用决策模型进行用户理解度的量化，可以帮助更好地了解和传递知识. 对于多模态的解释信息，可以通过交互的方式精细化提供[54]，如根据医生诊疗流程的推进，提供每步操作所需关注的信息，在提升信息接收效率的同时也避免了信息过载.

5 结 语

CDSS 是医疗场景中重要的提质增效工具，决策的可解释是影响用户理解与信任的重要途径. 多年来，通过对医疗场景的探索和对医护人员需求的挖掘，研究人员遵循以人为本的设计思路设计了众多 CDSS 应用. 本文通过调研和综述近 5 年人交互领域顶会的研究成果，总结归纳出分析可解释 CDSS 需求的 3 个维度，帮助研究人员快速定位和梳理问题，并为设计以人为本的 CDSS 系统提出实用建议. 本文工作总结如下:

(1) 归纳出 3 个拆解可解释需求的维度: 决策时间花费、用户专业度和诊疗工作流程，并对每种维度的定义和依据进行详细说明. 决策时间花费决定医护人员可以用于理解解释信息的时间成本，影响提供解释的多少(how much)问题; 用户专业度决定医护人员理解解释信息的认知成本，影响提供什么解释(what)的问题; 诊疗工作流程决定医护人员接收解释信息的途径和方式，影响什么地

点(where)以什么形式(how)提供解释. 3 个维度对应拆解了解释中的 2H2W 问题，并以此为线索，对与之对应的案例进行详细介绍.

(2) 基于可量化的 2 个维度：决策时间花费和用户专业度，绘制了包括 4 个象限的坐标并提出每个象限的可解释偏好、设计建议和典型用户画像. 此外，从案例和目前前沿的研究工作中，总结了未来 CDSS 设计的 3 条实用建议：资源受限的情况下，如何通过设计提高 CDSS 的可用性以及需要警惕的误区；对于不同利益相关者，如何通过共同设计的方式提升决策对整件事件的正面影响；如何了解诊疗流程并使设计尽量贴近的方法.

随着以人为本的 XAI 的发展，研究人员通过对场景与用户的挖掘，拓宽了对理解与信任的认知，通过不断优化可解释系统提升可用性，逐渐将 CDSS 应用真正推入实际应用中. 但是，如何进一步量化信任与理解，如何在此之上提升医疗系统的透明度、公平性和可追溯性等问题，还有很长的路要走. 期待未来 AI 真正助力并建成透明高效的医疗系统.

参考文献(References):

[1] 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

[2] Pasquale F. The black box society: the secret algorithms that control money and information[M]. Cambridge: Harvard University Press, 2015

[3] Guidotti R, Monreale A, Ruggieri S, *et al*. A survey of methods for explaining black box models[J]. ACM Computing Surveys, 2019, 51(5): Article No.93

[4] Cheng Xiao. On basic principles of China personal information protection law[J]. Journal of National Prosecutors College, 2021, 29(5): 3-20(in Chinese)
(程啸. 论我国个人信息保护法的基本原则[J]. 国家检察官学院学报, 2021, 29(5): 3-20)

[5] Holzinger A, Biemann C, Pattichis C S, *et al*. What do we need to build explainable AI systems for the medical domain?[OL]. [2024-01-23]. https://arxiv.org/abs/1712.09923

[6] Schmit C, Larson B N, Kum H C, *et al*. Public health informatics depends on engagement with privacy legislation[OL]. [2024-01-23]. https://www.healthaffairs.org/content/forefront/public-health-informatics-depends-engagement-privacy-legislation

[7] Ali S, Abuhmed T, El-Sappagh S, *et al*. Explainable Artificial Intelligence (XAI): what we know and what is left to attain Trustworthy Artificial Intelligence[J]. Information Fusion, 2023, 99: Article No.101805

[8] Chaddad A, Peng J H, Xu J, *et al*. Survey of explainable AI techniques in healthcare[J]. Sensors, 2023, 23(2): Article No. 634

[9] Zhang Shuyan, Pi Tingting. Interpretability dilemma and governance of artificial intelligence application in medical field[J]. Medicine & Philosophy, 2023, 44(3): 25-29+35(in Chinese)
(张姝艳，皮婷婷. 医疗领域中人工智能应用的可解释性困境与治理[J]. 医学与哲学, 2023, 44(3): 25-29+35)

[10] Yang Q, Steinfeld A, Zimmerman J. Unremarkable AI: fitting intelligent decision support into critical, clinical decision-making processes[C] //Proceedings of the CHI Conference on Human Factors in Computing Systems. New York: ACM Press, 2019: Article No.238

[11] Liao Q V, Varshney K R. Human-centered explainable AI (XAI): from algorithms to user experiences[OL]. [2024-01-23]. https://arxiv.org/abs/2110.10790

[12] Kuosmanen E, Kan V, Visuri A, *et al*. Let's draw: detecting and measuring Parkinson’s disease on smartphones[C] //Proceedings of the CHI Conference on Human Factors in Computing Systems. New York: ACM Press, 2020: 1-9

[13] Panigutti C, Beretta A, Giannotti F, *et al*. Understanding the impact of explanations on advice-taking: a user study for AI-based clinical Decision Support Systems[C] //Proceedings of the CHI Conference on Human Factors in Computing Systems. New York: ACM Press, 2022: Article No.568

[14] Yang Q, Hao Y X, Quan K X, *et al*. Harnessing biomedical literature to calibrate clinicians’ trust in AI decision support systems[C] //Proceedings of the CHI Conference on Human Factors in Computing Systems. New York: ACM Press, 2023: Article No.14

[15] 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 CHI Conference on Human Factors in Computing Systems. New York: ACM Press, 2019: Article No.4

[16] Chen Y Q, Yang X D, Chen B, *et al*. PdAssist: objective and quantified symptom assessment of Parkinson's disease via smartphone[C] //Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Los Alamitos: IEEE Computer Society Press, 2017: 939-945

[17] Cronin S, Freeman E, Doherty G. Investigating clutching interactions for touchless medical imaging systems[C] //Proceedings of the CHI Conference on Human Factors in Computing Systems. New York: ACM Press, 2022: Article No.351

[18] Seals A, Pilloni G, Kim J, *et al*. ‘Are they doing better in the clinic or at home?’: understanding clinicians’ needs when visualizing wearable sensor data used in remote gait assessments for people with multiple sclerosis[C] //Proceedings of the CHI Conference on Human Factors in Computing Systems. New York: ACM Press, 2022: Article No.350

[19] Swartout W R. XPLAIN: a system for creating and explaining expert consulting programs[J]. Artificial Intelligence, 1983, 21(3): 285-325

[20] Lacave C, Díez F J. A review of explanation methods for Bayesian networks[J]. The Knowledge Engineering Review, 2002, 17(2): 107-127

[21] Herlocker J L, Konstan J A, Riedl J. Explaining collaborative filtering recommendations[C] //Proceedings of the ACM Conference on Computer Supported Cooperative Work. New York: ACM Press, 2000: 241-250

[22] Abdul A, Vermeulen J, Wang D D, *et al*. Trends and trajectories for explainable, accountable and intelligible systems: an HCI research agenda[C] //Proceedings of the CHI Conference on Human Factors in Computing Systems. New York: ACM Press, 2018: Article No.582

[23] Tjoa E, Guan C T. A survey on explainable artificial intelligence (XAI): toward medical XAI[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(11): 4793- 4813

[24] Yarmand M, Wang B R, Chen C, *et al*. Design and development of a training and immediate feedback tool to support healthcare apprenticeship[C] //Proceedings of the CHI Conference on Human Factors in Computing Systems. New York: ACM Press, 2023: Article No.79

[25] Chen Chong, Chen Jie, Zhang Hui, *et al*. Review on interpretability of deep learning[J]. Computer Science, 2023, 50(5): 52-63(in Chinese)
(陈冲, 陈杰, 张慧, 等. 深度学习可解释性综述[J]. 计算机科学, 2023, 50(5): 52-63)

[26] 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

[27] Ribeiro B. Prediction of the lime availability on an industrial kiln by neural networks[C] //Proceedings of the IEEE International Joint Conference on Neural Networks Proceedings. IEEE World Congress on Computational Intelligence. Los Alamitos: IEEE Computer Society Press, 1998: 1987-1991

[28] Ribeiro M T, Singh S, Guestrin C. Model-agnostic interpretability of machine learning[OL]. [2024-01-23]. https://arxiv.org/abs/1606.05386

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

[30] Hu Z T, Ma X Z, Liu Z Z, *et al*. Harnessing deep neural networks with logic rules [OL]. [2024-01-23]. https://arxiv.org/abs/1603.06318

[31] Cheng Keyang, Wang Ning, Shi Wenxi, *et al*. Research advances in the interpretability of deep learning[J]. Journal of Computer Research and Development, 2020, 57(6): 1208-1217 (in Chinese)
(成科扬, 王宁, 师文喜, 等. 深度学习可解释性研究进展[J]. 计算机研究与发展, 2020, 57(6): 1208-1217)

[32] Osheroff J A, Teich J M, Middleton B, *et al*. A roadmap for national action on clinical decision support[J]. Journal of the American Medical Informatics Association, 2007, 14(2): 141-145

[33] Greenes R A. Clinical decision support: the road to broad adoption[M]. 2nd ed. Boston: Academic Press, 2014

[34] Zamora A, de Bobadilla F F, Carrion C, *et al*. Pilot study to validate a computer-based clinical decision support system for dyslipidemia treatment (HTE-DLP)[J]. Atherosclerosis, 2013, 231(2): 401-404

[35] Yang Q, Zimmerman J, Steinfeld A. Review of medical decision support tools: emerging opportunity for interaction design[C] //Proceedings of the IASDR Interplay. Australia: IASDR, 2015: 2366-2382

[36] Lei J B, Sockolow P, Guan P C, *et al*. A comparison of electronic health records at two major Peking University Hospitals in China to United States meaningful use objectives[J]. BMC Medical Informatics and Decision Making, 2013, 13: Article No.96

[37] Fu J J, Zhou Y F, Wang X. Unveiling high-speed follow-up consultation for chronic disease treatment: a pediatric hospital case in China[J]. Proceedings of the ACM on Human-Computer Interaction, 2022, 6(CSCW2): Article No.347

[38] Papadopoulos P, Soflano M, Chaudy Y, *et al*. A systematic review of technologies and standards used in the development of rule-based clinical decision support systems[J]. Health and Technology, 2022, 12(4): 713-727

[39] Barnett G O, Cimino J J, Hupp J A, *et al*. DXplain: an evolving diagnostic decision-support system[J]. JAMA, 1987, 258(1): 67-74

[40] Lindsey R, Daluiski A, Chopra S, *et al*. Deep neural network improves fracture detection by clinicians[J]. Proceedings of the National Academy of Sciences of the United States of America, 2018, 115(45): 11591-11596

[41] Burgess E R, Jankovic I, Austin M, *et al*. Healthcare AI treatment decision support: design principles to enhance clinician adoption and trust[C] //Proceedings of the CHI Conference on Human Factors in Computing Systems. New York: ACM Press, 2023: Article No.15

[42] Osheroff J A, Teich J, Levick D, *et al*. Improving outcomes with clinical decision support: an implementer's guide[M]. 2nd ed. New York: HIMSS Publishing, 2012

[43] Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning[OL]. [2024-01-23]. https://arxiv.org/abs/1702.08608

[44] Elwyn G, Scholl I, Tietbohl C, *et al*. "Many miles to go …": a systematic review of the implementation of patient decision support interventions into routine clinical practice[J]. BMC Medical Informatics and Decision Making, 2013, 13(S2): Article No.S14

[45] Jaspers M W M, Smeulers M, Vermeulen H, *et al*. Effects of clinical decision-support systems on practitioner performance and patient outcomes: a synthesis of high-quality systematic review findings[J]. Journal of the American Medical Informatics Association, 2011, 18(3): 327-334

[46] Ma S, Lei Y, Wang X R, *et al*. Who should I trust: AI or myself? Leveraging human and AI correctness likelihood to promote appropriate trust in AI-assisted decision-making[C] //Proceedings of the CHI Conference on Human Factors in Computing Systems. New York: ACM Press, 2023: Article No.759

[47] Liao Q V, Gruen D, Miller S. Questioning the AI: informing design practices for explainable AI user experiences[C] //Proceedings of the CHI Conference on Human Factors in Computing Systems. New York: ACM Press, 2020: 1-15

[48] Wu Dan, Sun Guoye. Towards explainable interactive artificial intelligence: motivations, approaches, and research trends[J]. Wuhan University Journal (Philosophy & Social Science), 2021, 74(5): 16-28(in Chinese)
(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)

[49] Schaekermann M, Beaton G, Sanoubari E, *et al*. Ambiguity-aware AI assistants for medical data analysis[C] //Proceedings of the CHI Conference on Human Factors in Computing Systems. New York: ACM Press, 2020: 1-14

[50] Cai C J, Winter S, Steiner D, *et al*. “Hello AI”: uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making[J]. Proceedings of the ACM on Human-Computer Interaction, 2019, 3(CSCW): Article No.104

[51] Wang D D, Zhang W C, Lim B Y. Show or suppress? Managing input uncertainty in machine learning model explanations[J]. Artificial Intelligence, 2021, 294: Article No.103456

[52] Bach A K P, Nørgaard T M, Brok J C, *et al*. “If I had all the time in the world”: ophthalmologists’ perceptions of anchoring bias mitigation in clinical AI support[C] //Proceedings of the CHI Conference on Human Factors in Computing Systems. New York: ACM Press, 2023: Article No.16

[53] Calisto F M, Fernandes J, Morais M, *et al*. Assertiveness-based agent communication for a personalized medicine on medical imaging diagnosis[C] //Proceedings of the CHI Conference on Human Factors in Computing Systems. New York: ACM Press, 2023: Article No.13

[54] Gu H Y, Yang C X, Haeri M, *et al*. Augmenting pathologists with NaviPath: design and evaluation of a human-AI collaborative navigation system[C] //Proceedings of the CHI Conference on Human Factors in Computing Systems. New York: ACM Press, 2023: Article No.349

[55] Schaekermann M, Cai C J, Huang A E, *et al*. Expert discussions improve comprehension of difficult cases in medical image assessment[C] //Proceedings of the CHI Conference on Human Factors in Computing Systems. New York: ACM Press, 2020: 1-13

[56] Xie Y, Chen M, Kao D, *et al*. CheXplain: enabling physicians to explore and understand data-driven, AI-enabled medical imaging analysis[C] //Proceedings of the CHI Conference on Human Factors in Computing Systems. New York: ACM Press, 2020: 1-13

[57] Wang D D, Yang Q, Abdul A, *et al*. Designing theory-driven user-centric explainable AI[C] //Proceedings of the CHI Conference on Human Factors in Computing Systems. New York: ACM Press, 2019: Article No.601

[58] Lee M H, Chew C J. Understanding the effect of counterfactual explanations on trust and reliance on AI for human-AI collaborative clinical decision making[J]. Proceedings of the ACM on Human-Computer Interaction, 2023, 7(CSCW2): Article No.369

[59] Liang Y, Fan H W, Fang Z J, *et al*. OralCam: enabling self-examination and awareness of oral health using a smartphone camera[C] //Proceedings of the CHI Conference on Human Factors in Computing Systems. New York: ACM Press, 2020: 1-13

[60] Wang D K, Wang L P, Zhang Z, *et al*. “Brilliant AI doctor” in rural clinics: challenges in AI-powered clinical decision support system deployment[C] //Proceedings of the CHI Conference on Human Factors in Computing Systems. New York: ACM Press, 2021: Article No.697

[61] You Y, Tsai C H, Li Y, *et al*. Beyond self-diagnosis: how a chatbot-based symptom checker should respond[J]. ACM Transactions on Computer-Human Interaction, 2023, 30(4): Article No.64

[62] Jo E, Ryu M, Kenderova G, *et al*. Designing flexible longitudinal regimens: supporting clinician planning for discontinuation of psychiatric drugs[C] //Proceedings of the CHI Conference on Human Factors in Computing Systems. New York: ACM Press, 2022: Article No.352

[63] Jacobs M, He J, Pradier M F, *et al*. Designing AI for trust and collaboration in time-constrained medical decisions: a socio-technical lens[C] //Proceedings of the CHI Conference on Human Factors in Computing Systems. New York: ACM Press, 2021: Article No.659

[64] Wischnewski M, Krämer N, Müller E. Measuring and understanding trust calibrations for automated systems: a survey of the state-of-the-art and future directions[C] //Proceedings of the CHI Conference on Human Factors in Computing Systems. New York: ACM Press, 2023: Article No.755

[65] Devaraj S, Sharma S K, Fausto D J, *et al*. Barriers and facilitators to clinical decision support systems adoption: a systematic review[J]. Journal of Business Administration Research, 2014, 3(2): 36-53

[66] Tonekaboni S, Joshi S, McCradden M D, *et al*. What clinicians want: contextualizing explainable machine learning for clinical end use[C] //Proceedings of the 4th Machine Learning for Healthcare Conference. New York: PMLR, 2019: 359-380

[67] Sambasivan N, Kapania S, Highfill H, *et al*. “Everyone wants to do the model work, not the data work”: data cascades in high-stakes AI[C] //Proceedings of the CHI Conference on Human Factors in Computing Systems. New York: ACM Press, 2021: Article No.39

[68] Delitzas A, Chatzidimitriou K C, Symeonidis A L. Calista: a deep learning-based system for understanding and evaluating website aesthetics[J]. International Journal of Human-Computer Studies, 2023, 175: Article No.39103019

[69] Li Yuanfeng, Wu Qun, Zhang Jian, *et al*. Layout design of medical terminal interface combining user-oriented and genetic algorithm[J]. Journal of Computer-Aided Design & Computer Graphics, 2021, 33(5): 662-670(in Chinese)
(李源枫, 吴群, 张剑, 等. 结合用户导向和遗传算法的医疗终端界面布局设计[J]. 计算机辅助设计与图形学学报, 2021, 33(5): 662-670)

[70] Robert G. Participatory action research: using experience-based co-design to improve the quality of healthcare services[M] //Ziebland S, Coulter A, Calabrese J D, *et al*. Understanding and Using Health Experiences: Improving Patient Care. Oxford: Oxford University Press, 2013: 138-149

[71] Schnock K O, Snyder J E, Fuller T E, *et al*. Acute care patient portal intervention: portal use and patient activation[J]. Journal of Medical Internet Research, 2019, 21(7): Article No.e13336

[72] Van de Velde S, Heselmans A, Delvaux N, *et al*. A systematic review of trials evaluating success factors of interventions with computerised clinical decision support[J]. Implementation Science, 2018, 13(1): Article No.114

[73] Bansal G, Nushi B, Kamar E, *et al*. Updates in human-AI teams: understanding and addressing the performance/compatibility tradeoff[C] //Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2019: 2429-2437