不确定性的层次：人工智能不确定性可视化如何影响人们对阿尔茨海默病预测的信任

Jonatan Reyes*
jonatan.reyes01@utrgv.edu 得克萨斯大学里奥格兰德河谷分校 得克萨斯州爱丁堡，美国

Mina Massoumi 康考迪亚大学 蒙特利尔，魁北克，加拿大

Anil Ufuk Batmaz
康考迪亚大学 蒙特利尔，魁北克，加拿大

Marta Kersten-Oertel 康考迪亚大学 蒙特利尔，魁北克，加拿大

arXiv:2602.01264v1 [cs.HC] 1 Feb 2026

摘要

人工智能（AI）越来越多地用于支持阿尔茨海默病（AD）的预后判断，但由于缺乏透明度和可解释性，其应用仍然有限；对于长期预测而言，这一问题尤为突出，因为不确定性本身不可避免，而且结果可能多年后才会揭晓。我们将不确定性可视化 视为一种可解释人工智能（XAI）技术，并考察用户解读人工智能生成的未来认知衰退转变预测时，它如何影响用户的信任、信心和依赖。我们开展了两项研究，一项面向普通参与者（N=37），另一项面向神经影像学和神经病学专家（N=10），比较二元（存在/不存在）与连续（饱和度）不确定性编码。连续编码提升了感知可靠性，并帮助用户认识到模型局限性；而二元编码则增强了用户当下的信心。这揭示了在高不确定性下解读未来预测时，存在因专业水平而异的权衡。这些发现揭示了为预后人工智能设计不确定性表征所面临的关键挑战，并最终总结出一套基于实证的指南，用于创建值得信赖且适合用户的临床决策支持工具。

CCS 概念

•**以人为中心的计算** → **用户研究**；• **计算方法** → 《人工智能》。

关键词

人工智能不确定性、人机协同决策、阿尔茨海默病、可视化、对人工智能的信任

ACM 参考格式：

Jonatan Reyes、Mina Massoumi、Anil Ufuk Batmaz 和 Marta Kersten- Oertel。2018年。《不确定性的层次：人工智能不确定性可视化如何影响对阿尔茨海默病预测的信任》。载于请确保填写

特此授予许可，允许个人或课堂使用本作品的全部或部分内容，并可制作数字或纸质副本，且无需支付费用，但前提是不得为营利或商业利益而制作或分发副本，并且副本必须附有本声明及首页上的完整引文。对于本作品中由非作者拥有版权的组成部分，必须尊重其版权。允许在注明出处的情况下进行摘要。若要以其他方式复制、再版、发布到服务器或重新分发至列表，则需事先获得具体许可和/或支付费用。请通过 permissions@acm.org 联系我们申请许可。会议缩写 *'XX*，纽约州伍德斯托克© 2018年 版权所有归所有者/作者所有。出版权已授权给 ACM。ACM ISBN 978-1-4503-XXXX-X/2018/06 https://doi.org/XXXXXXX.XXXXXXX

从版权确认邮件中获取的正确会议名称（会议缩写 *'XX*）。ACM，美国纽约州纽约，13页。https://doi.org/ XXXXXXX.XXXXXXX

1 引言

阿尔茨海默病（AD）影响着全球超过 5,000 万人，预计这一数字将增加至原来的三倍，直至 2050 [51, 72]。由于目前尚无治愈方法，早期检测和准确的预后对于及时干预、患者咨询和临床试验招募至关重要 [8]。一个核心临床挑战是理解不同生物标志物如何相互作用并与疾病发展相关，从而能够根据前驱期准确地区分和预测 疾病进展轨迹 [22, 45, 62]。预测这些疾病进展轨迹，例如从认知正常（CN）转变为轻度认知障碍（MCI），或从 MCI 转变为阿尔茨海默病，是一项艰巨的任务，因为这些变化进展缓慢、具有异质性，并且通常发生在明显症状出现前数年 [21]。然而，这些转变会对治疗规划和风险沟通产生重大影响。

人工智能（AI）模型通过整合包括磁共振成像（MRI）和正电子发射断层扫描（PET）成像、脑脊液（CSF）生物标志物、遗传风险因素以及认知评估在内的多模态数据，在建模阿尔茨海默病进展方面展现出良好前景 [30, 63, 76]。此类模型能够检测细微的神经退行性变模式，并生成可支持临床决策的长期预后预测。然而，尽管技术不断进步，临床应用仍然有限，很大程度上是因为存在误解人工智能输出及其相关不确定性的风险。当不同疾病严重程度下的预测在视觉上相似时，临床医生可能会误判模型的实际置信度，从而过度或不足依赖其预测。

人工智能系统应用于阿尔茨海默病的一大障碍在于，临床与影像数据具有异质性、采样不规则且常常不完整，原因包括就诊缺失、随访不一致以及影像数据可用性不一，这些因素导致长期预测存在显著不确定性 [49, 58]。因此，偶然不确定性（生物标志物和认知测量中的噪声与变异性）以及认知不确定性（纵向记录缺失、群体分布不均衡，以及对非典型发展轨迹的代表性有限） [34, 53] 都很高。临床医生和研究人员通常只能获得人工智能模型给出的类别预测或置信度分数，却不了解模型为何不确定，也不清楚预测的可靠性如何。

对发展轨迹的预测可能 [44]如何可靠。这种不透明性使信任校准更加困难，在高风险情境中可能导致过度依赖或依赖不足 [5, 17]。这造成了模型呈现不确定性的方式与临床医生解读不确定性的方式之间长期存在差距，因此难以判断对人工智能预测的信任是否恰当。

尽管人工智能研究在阿尔茨海默病领域日益强调预测性能，但相对较少有研究关注提供可信、可解释且易于理解的输出，以切实支持临床应用。理想情况下，人工智能系统应帮助临床医生理解：从异质性临床数据中提取的领域特定特征，如何促成所建议的决策或诊断。近期关于阿尔茨海默病研究中可解释人工智能的系统综述指出，模型通常依赖归因式和遮挡式解释技术，以及可解释模型架构，来揭示其决策逻辑的相关方面 [20, 67]。然而，这些研究大多聚焦于短期任务或诊断任务，而长期预后中的不确定性更为显著且影响更为深远，因此相关研究仍相对不足。此外，既有研究对于不同利益相关者如何理解这些解释所提供的洞见有限，包括临床医生和通才如何在高风险医疗情境中理解不确定性可视化 [3, 23, 24, 36]。

与此同时，关于不确定性沟通的研究凸显了设计和实现有效不确定性可视化所面临的重大障碍 [65]，其中包括方法学方法的限制，以及终端用户在理解方面长期面临的挑战。其他研究指出，尽管可视化技术有助于支持用户理解，但在许多现有人工智能系统中仍未得到充分应用 [22, 35, 64]。近期一项调查进一步指出，当前文献未能充分探讨可信度 [20]，并呼吁将可解释人工智能方法与适当的数据表示相结合，以更有效地传达模型复杂性、冲突和不确定性。基于这些见解，我们认为，将不确定性可视化用作可解释人工智能技术，为向临床医生及其他利益相关者清晰呈现模型置信度和局限性提供了契机，尤其适用于不确定性较高的预后任务 [13, 18]。

本文通过探讨不确定性可视化如何影响人工智能支持的阿尔茨海默病进展预测中的信任、信心和决策，弥补了这些研究空白。阿尔茨海默病预后为不确定性沟通带来了独特的挑战。预测必须推断数月乃至数年后的情况，使临床医生无法获得即时反馈来重新校准信任 [57]。轻度认知障碍等诊断类别仍存在争议，患者的病程轨迹差异很大，数据缺失或不规律也很常见。错误可能会显著影响护理规划和试验资格。这些因素共同构成了脆弱的决策环境，因此，要实现校准后的依赖，不确定性沟通至关重要；这也促使我们研究不确定性可视化如何影响信任和决策。通过两项互为补充的用户研究，我们作出以下贡献：

- **可解释人工智能：** 我们将不确定性可视化定位为一种可解释人工智能技术。
- **可解释人工智能/人机交互：** 我们探索性地分析了：当用户事先获得有关模型训练和验证步骤的信息时，他们对人工智能的信任如何得到校准。
- **阿尔茨海默病：** 我们通过考察通才用户和领域专家如何看待包含不确定性信息的人工智能生成输出，以及这些信息在解读预测的认知正常/轻度认知障碍/阿尔茨海默病转变时如何影响信任校准、感知可靠性和信心，从而拓展现有的阿尔茨海默病研究。
- **人机交互：** 我们基于两项研究中的定量模式和定性见解，为具备不确定性感知能力的临床人工智能工具提供设计指南。

这些贡献共同深化了我们对不确定性可视化如何在高风险、长期医疗人工智能中发挥可解释性作用的理解，并为设计值得信赖的阿尔茨海默病预后决策支持工具提供了实用指导。

2 相关工作

医疗保健领域的人机协作涉及系统设计、任务特征与用户专业知识之间的复杂互动。这些因素会影响临床医生采纳、遵从或推翻算法建议，进而直接关系到患者安全 [1, 7, 16, 48, 70]。本节回顾与本研究最相关的三个既有研究领域：（1）基于人工智能的临床决策支持；（2）人机协作中的信任与依赖；（3）人工智能系统中的不确定性可视化，尤其关注阿尔茨海默病预后。

2.1 基于人工智能的临床决策支持

临床决策支持系统（CDSS）将患者数据整合为算法建议，以提高诊断效率、安全性和工作流程 [32, 60]。然而，这些益处也伴随着风险，例如自动化偏见：临床医生可能会过度依赖人工智能输出，即使其内容有误，尤其是在时间压力或认知负荷较高时 [61]。实证研究表明，在计算病理学 [55], 、新冠肺炎分诊 [73], 和临床药学 [31],领域，专家有时也会采纳错误的人工智能建议，这说明仅凭训练和领域专业知识并不能保证对人工智能的使用恰到好处。

相关文献指出，安全关键型人工智能有一项核心要求：系统必须支持<风格 id='1'>恰当校准的信任，而非一味提升信任 [11]。提出的机制包括不确定性沟通 [46], 、透明解释 [54], 、传达系统局限性的交互式工具 [26, 40], , 以及以决策为重点的反馈循环 [2]。然而，对于预后任务而言，结果可能多年后才可观察到，因此很少能获得反馈，这使得前瞻性不确定性沟通尤为重要。

2.2 人机协作中的信任、依赖与任务特征

信任体现了用户对人工智能系统能力的信念，而依赖则反映了他们实际使用系统建议的情况 [38, 57]。这两个概念往往并不一致：用户可能信任某个系统，却不愿依赖它；也可能在信任有限的情况下仍依赖该系统 [19]。衡量信任和依赖的方法包括自我报告、行为性让步，以及对信任随时间变化的动态追踪

[15, 29]。多维信任模型进一步区分感知能力、可预测性、透明度和正直性 [4], ，从而更深入地揭示信任如何形成和校准。

任务特征会显著影响依赖程度。随着任务复杂度增加，用户往往会更多地听从人工智能，即使这种额外依赖并无充分依据 [6, 50]。预后任务涉及长期预测，因此尤其容易出现依赖失准，因为用户难以轻易验证结果 [12]。这一模式与阿尔茨海默病进展预测尤其相关，因为这类预测本质上具有不确定性且风险很高。了解用户如何解读此类任务中的不确定性，对于设计能够促进适当依赖、而非盲目信任或过度怀疑的人工智能系统至关重要。

2.3 人工智能系统中的不确定性可视化

不确定性可视化是帮助用户理解模型局限性，并在复杂决策任务中校准依赖程度的一种重要策略 [46, 47]。颜色渐变、透明度和数值不确定性标签等视觉编码，能够提供有关模型置信度和模糊性的感知线索 [9, 10]。这些线索可以引导用户注意力，影响感知可靠性，并改变用户对人工智能建议的行为依赖 [13, 14, 18, 41]。

研究表明，任务情境和不确定性类型会显著影响可视化对用户的作用。例如，在医院床位预测中，醒目的不确定性显示降低了用户遵循人工智能预测的意愿 [39]。在游戏场景中，连续不确定性编码提升了信任和信心，尤其是在起初对人工智能持怀疑态度的用户中 [52]。在大学录取预测等预测任务中，有序不确定性表达改善了依赖校准 [77]。数理能力、过往经验和决策风格等个体差异也会进一步调节这些影响 [33]。

在阿尔茨海默病研究中，我们发现少数既有研究强调，先进的可视化方法对于应对阿尔茨海默病生物标志物及其相互作用动态的复杂性至关重要。其中一项研究引入了基于图的人工智能模型，整合多组学数据，以识别对于理解阿尔茨海默病进展至关重要的生物标志物相互作用 [74, 75]。其他人工智能驱动的研究则探讨了卷积神经网络（CNN）及其解读用于阿尔茨海默病相关生物标志物的磁共振成像数据的效能，展现了计算方法在可视化生物标志物关系方面的价值 [68]。值得注意的是，临床研究强调，需要能够整合多项临床评估的工具，以可视化阿尔茨海默病生物标志物网络，并支持可靠的患者个体化决策。此类工具被视为全面理解阿尔茨海默病机制的关键 [25, 56]。鉴于该领域研究有限，现有文献指出，需要更先进、更具整合性的可视化框架，为长期预后建模和轨迹分析提供更有力的支持，并构建以人为本、可解释且负责任的人工智能系统 [20, 67]。此外，不确定性的可视化仍缺乏充分研究，尤其是针对长期预后预测。Eslami等人 [22] 和 Lai 等人 [37] 已开始将不确定性纳入阿尔茨海默病预测工具，凸显了更清晰地传达模型局限性的必要性。然而，先前的研究尚未考察不同用户群体（如普通参与者和临床专家）如何解读阿尔茨海默病进展预测中的可视化不确定性，以及如何据此采取行动。

3 研究目标

本节概述了促成本研究开展的研究空白、本研究目标，以及指导我们开展研究的研究问题与假设。

3.1 研究空白与目标

在临床决策支持系统（CDSS）、人类与人工智能信任研究以及不确定性可视化领域，既有研究强调了校准后的依赖的重要性，以及清晰传达人工智能局限性的必要性。然而，对于不确定性可视化如何影响长期、高不确定性的医学预后中的信任与决策，我们仍缺乏实证认知；在这种情境下，结果会在数月或数年间逐渐显现，不确定性是内在存在的，而且用户很少能获得反馈来重新校准信任。对于阿尔茨海默病进展预测而言，这一研究空白尤为突出，因为临床决策事关重大，且模型的置信区间可能很宽。

因此，我们旨在考察不同的不确定性编码如何塑造用户对人工智能生成的阿尔茨海默病进展轨迹的解读，以及这些编码如何影响信任、信心和决策。 不同于以往主要关注分类或短期诊断任务的研究，我们在预后场景中研究不确定性；在这种场景下，模糊性不可避免，解读要求也很高。 由于不确定性并非孤立存在，我们还考察不同程度的模型透明度如何影响用户对预后输出的解读和信任。在临床人工智能领域，利益相关者对于适当依赖所需的信息量仍存在分歧，这仍是一个尚待解答的问题。为了解这些影响如何体现在不同利益相关者身上，我们分别针对广泛招募的参与者和领域专家研究这些不确定性编码，从而描述不同群体解读、评估和依赖人工智能生成的预后信息时呈现的不同模式。

3.2 研究问题与假设

基于这一研究动机，我们探讨不确定性可视化和模型信息如何塑造用户对预测的认知正常/轻度认知障碍/阿尔茨海默病转变的解读，以及这些因素如何影响信任、信心和决策。我们提出以下研究问题与假设。

研究问题1：模型信息与信任。关于人工智能模型的信息量（最少与适中）如何影响用户对人工智能辅助的阿尔茨海默病进展预测的信任？

假设1： 提供适中的模型信息将带来更恰当校准的信任，相比之下，提供最少信息的效果较差。

研究问题2：不确定性可视化与信任。二元与连续不确定性编码如何影响用户对人工智能预测的信任，以及他们对特定信任维度的评估？

假设2： 连续不确定性可视化将带来 更适当校准的任务中的信任 而非二元可视化。 **假设3：** 连续不确定性可视化将在具体方面获得更高评价

信任维度，尤其是感知到的可靠性，高于二元可视化。

研究问题3：不确定性、模型信息与信心。不确定性呈现形式和模型信息如何影响用户基于人工智能预测进行决策时的信心？分别考察普通用户和领域专家时，这些影响有何不同？由于先前研究对于不确定性会提高还是降低决策信心的证据不一，因此信心分析属于探索性分析。

研究问题4：专家的信任与决策。与以二元形式呈现不确定性相比，以连续形式呈现不确定性时，领域专家的信任、信心和决策判断有何不同？

假设4：连续不确定性可视化将带来更加恰当校准的信任无论在试次层面还是在多维信任维度上，领域专家均会如此。

4 方法论

4.1 数据与模型设置

我们实现了机器学习可视化阿尔茨海默病（ML4VisAD）模型的一个实例 [22]，将多模态输入数据转换为颜色编码的预测结果，以表示阿尔茨海默病随时间的进展。ML4VisAD 模型的架构由两大部分组成：特征提取阶段和图像生成阶段。网络的第一部分通过全连接层和特征拼接，从各种模态（磁共振成像、正电子发射断层扫描等）及认知评估中提取特征。模型的第二部分涉及张量化，即将特征表示转换为形状适当的张量；随后将这些张量重塑为二维格式，并应用转置卷积操作，根据提取的特征重建二维图像。ML4VisAD 模型的架构可在公开代码仓库中获取 [1]。

4.2 刺激材料：人工智能生成的阿尔茨海默病进展轨迹

本研究的核心是由 ML4VisAD 模型生成的视觉刺激。我们选择这种可视化方案，是因为它采用了新颖的估计器设计，将多模态输入特征编码为表示阿尔茨海默病进展的二维 RGB 图像。在图像生成阶段，模型会生成一幅 $23 \times 23 \times 3$ 图像，并将其与一幅尺寸相同的真实值图像进行比较。在训练之前，根据 ADNI QT-PAD 数据库 [2]中记录的每次就诊时的临床状态生成真实值图像。这些图像使用颜色编码的片段表示患者的疾病阶段。在训练期间，模型会学习尽量减小预测图像与这些真实值表示之间的误差。

真实值图像的尺寸定义了空间布局，即模板，模型通过该模板在训练过程中学习特征交互。该模板由五个部分组成。前四个部分对应不同的疾病阶段，每个部分沿**水平轴**跨越5个像素，并代表一个临床时间点（基线 = 0、6个月、12个月、24个月）。第五个部分具有独特的语义作用：它编码了 ML4VisAD 模型在特征提取和多模态数据可视化过程中可能引入的不确定性程度。在这一部分中，较低的像素强度表示较高的不确定性，说明模型的预测反映了将输入特征映射到所建模特定区域时存在的模糊性。 **垂直轴**反映了二维张量的一个空间维度；它不编码临床信息，而是作为一种空间结构，使转置卷积层能够学习视觉模式。x 轴和 y 轴共同描述多模态信号如何在二维空间中组织，以生成疾病轨迹的最终颜色编码表示。默认情况下，图像尺寸设为 23×23 像素；若增至 45×45 像素，则会产生更平滑的过渡 [22]。我们选择保持默认尺寸，以便更好地根据图像像素值所反映的不确定性程度对图像进行分类。 颜色表示每次就诊时预测的诊断结果（绿色 = 认知正常，蓝色 = 轻度认知障碍，红色 = 阿尔茨海默病）。所有就诊时均为单一颜色的图像代表稳定病例（例如，稳定的认知正常、稳定的轻度认知障碍、稳定的阿尔茨海默病），而包含两种或更多颜色的图像则表示疾病进展或转化（例如，认知正常→轻度认知障碍，轻度认知障碍→阿尔茨海默病；图1）。

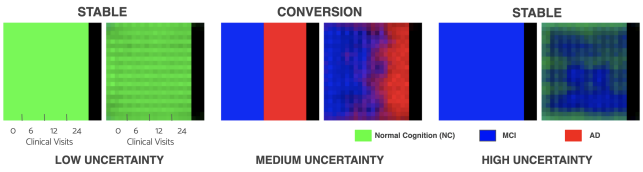


图1：示例刺激材料。每行展示一个病例（认知正常且病情稳定、由轻度认知障碍转为→阿尔茨海默病、轻度认知障碍且病情稳定）。各列分别展示二元（左）和连续（右）不确定性格式。

不确定性可视化格式。 我们操纵了不确定性的呈现方式，同时保持预测不变：

- **二元不确定性**：最右侧的三行像素作为指示条。填充（黑色）的行表示高不确定性；空白行表示低不确定性。
- **连续不确定性**：除了最右侧三行像素指示条外，还通过颜色饱和度编码不确定性。图像被转换到 CIELAB色彩空间，并使用亮度（L）作为确定性代理指标——亮度越高 L，颜色越鲜艳（确定性越高）；亮度越低 L，颜色越暗、饱和度越低（不确定性越高） [59]

在所有条件下，不确定性始终都会显示；操纵的重点是如何呈现不确定性。

5 用户研究

我们开展了两项用户研究，以考察不确定性可视化如何影响人们对人工智能支持的阿尔茨海默病进展预测的信任、信心和决策。研究1是一项面向广泛参与者样本的混合方法在线研究，调查普通用户如何在不同不确定性编码下解读人工智能生成的轨迹。研究1还操纵了第二个因素：所提供的关于人工智能模型及其

[1]https://github.com/mohaEs/ML4VisAD [2] https://ida.loni.usc.edu

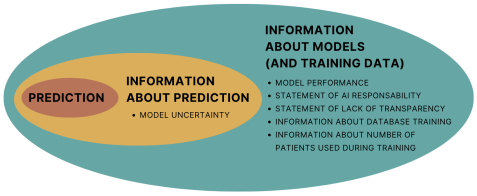


图2："中等信息"条件下向参与者展示的信息元素，改编自 Lai 等人 [37]。

训练过程的信息量（少量与中等）。研究1的见解为第二项研究的设计修订提供了依据。研究2聚焦于神经病学、神经影像学和痴呆研究领域的领域专家，评估受过训练的专业人员如何解读相同的不确定性编码。为了在尊重专家时间有限的同时分离不确定性可视化的效应，研究2并未改变模型信息的水平；相反，所有专家都收到了关于该模型的标准化中等信息 描述，包括数据集详情、队列规模以及训练流程的简化摘要。

在两项研究中，因变量包括参与者对人工智能预测的试次级信任、决策信心；在研究2中，还包括专家在查看具有不同不确定性编码的同一患者病例时，其决策是否发生变化。正式实验前开展了一项试点研究，以完善任务界面、教程和模型信息材料。 所有研究均获得本机构研究伦理委员会的批准。

5.1 试点研究

我们首先开展了一项试点实验验证研究，共有5名参与者；他们独立于招募来参与研究1的队列。试点研究的发现为用户研究1的设计修订提供了依据。通过出声思维协议，我们发现了多项可用性挑战，包括：(1) 修改教程页面上呈现的信息；(2) 将分散在多个页面的教程整合到一个页面；(3) 统一信任量表，将所有量表标准化为10分制；(4) 调整"中等信息"条件下提供的信息量，补充有关模型训练过程的细节；以及(5) 确保恰当选择能够呈现不同不确定性水平（低、中、高）的图像。

5.2 研究1：广泛参与者研究

我们采用了一组间（人工智能模型信息）×组内（不确定性可视化形式：连续型与二元）混合设计，所有参与者都看到了两种不确定性可视化，但被随机分配到两种人工智能信息条件之一。针对组间比较进行的先验统计功效分析假设效应量较大（$d \approx 0.8$）、显著性水平为 $\alpha = .05$且统计功效为 80%，结果表明约需 42 名参与者（每组 21 人）。参与者通过机构邮件列表和社交媒体平台（LinkedIn 和 X）招募。研究采用在线工作流程，包括知情同意、教程、前测问卷、主要预测任务和后测问卷。参与者被随机分配到两种人工智能模型信息条件之一：最少细节（不提供模型背景）或中等细节（数据集描述、训练队列规模、纳入/排除标准，以及一则说明超参数可按需提供的备注）。

5.2.1 教程。 教程向用户介绍了阿尔茨海默病以及早期检测的重要性，随后通过示例引导他们了解人工智能生成的预测，展示疾病可能如何随时间保持稳定或进展。用户查看了疾病稳定阶段的可视化：认知正常（CN）、轻度认知障碍（MCI）和阿尔茨海默病（AD），并查看了二元不确定性指标，这些指标标示出模型缺乏信心之处。随后，他们探索了阶段转换，例如 CN → MCI 或 MCI → AD，并使用通过颜色饱和度编码的连续不确定性可视化。最后，教程展示了以人为中心的设计原则如何提升人工智能预测的可解释性，帮助临床医生在确认或验证阿尔茨海默病预后时，区分高确定性与高不确定性的输出。

5.2.2 前测问卷。 参与者提供了人口统计信息，并以 10 分制评价了自己对人工智能开发、数据可视化和阿尔茨海默病的熟悉程度。

5.2.3 任务：阿尔茨海默病进展的人工智能不确定性预测。参与者查看了针对稳定和进展病例的12项人工智能生成的患者病程轨迹预测。每位参与者看到六项采用二元不确定性编码的刺激材料和六项采用连续编码的刺激材料，这些材料以区组间被试设计呈现，区组顺序随机安排；自变量为所提供的有关人工智能模型及其训练过程的信息量。对于每项刺激材料，参与者判断患者在6、12或24个月时是保持稳定还是出现转变（例如，从认知正常 → 转为轻度认知障碍，或从轻度认知障碍转为 → 阿尔茨海默病）。刺激材料涵盖三个不确定性水平：低（0–10%）、中（11–20%）和高（≥21%）。

5.2.4 测量指标。每项刺激材料之后，参与者报告了（1）感知到的转变时间点、（2）决策信心（10分制）以及（3）对人机决策过程的信任程度（10分制）。每个区组结束后，参与者填写经改编的 Ashoori 和 Weisz 的 [4] 信任量表，评估可信度、可靠性、技术能力和个人依恋。

5.2.5 后测问卷。参与者就不确定性如何影响他们的决策、他们如何解读可视化，以及哪些因素影响了他们对人工智能预测的信任，提供了开放式反思。

5.3 研究2：专家研究

本研究邀请了神经病学、神经影像学和痴呆症研究领域的专家（从事阿尔茨海默病研究的人工智能研究人员），以评估经验丰富的从业者如何解读基于人工智能的阿尔茨海默病进展预测中的不确定性。 专家通过定向电子邮件招募和滚雪球抽样相结合的方式招募：研究人员邀请初始联系人参与研究，并鼓励他们将研究信息转发给符合条件的同事。当符合条件的人群规模较小

且高度专业化时，医疗人工智能研究通常会采用滚雪球抽样。招募开始前，研究人员针对被试内比较进行了先验效能分析，采用较大的效应量（基于研究1的结果）（$d \approx 1.1$）、$\alpha = .05$ 和80%的统计效能，结果表明约需8–10名专家。因此，我们计划招募8–12名参与者，并成功招募了 $N = 10$ 名专家。

流程与研究1相同，包括知情同意、前测问卷、主要任务和后测问卷。专家还会收到有关人工智能模型的背景信息，包括数据集详情、队列规模、纳入/排除标准，以及对训练流程的简要说明。

5.3.1 前测问卷。专家报告了人口统计信息，并在10分制量表上评价了自己对人工智能开发、数据可视化和阿尔茨海默病的了解程度。

5.3.2 任务：阿尔茨海默病进展的人工智能不确定性预测。为尊重专家有限的时间，我们精简了任务，减少了刺激材料总数。每位专家查看六条患者病程轨迹，每条呈现两次，分别采用二元不确定性和连续不确定性，共计12项刺激材料。不确定性分为两个级别：低（<30%）和高（>30%）。专家完成与研究1相同的核心判断任务，指出每位患者在6、12、24个月或[原文缺失]时是否保持稳定或发生转变。刺激材料以区组被试内设计呈现，不确定性呈现形式的顺序随机安排。

5.3.3 测量指标。收集的测量指标与研究1相同：转变时间点、决策信心，以及对每个刺激的信任评分。

5.3.4 后测问卷。专家反思了不确定性可视化是否提升了可解释性、阐明了模型局限性、传达了可靠性，或支持在研究或临床实践中采用这些方法。开放式回答记录了关于可视化的优势与局限性的详细推理。

6 结果

6.1 广泛参与者研究（研究1）

我们收集了43名参与者的回答，其中21人被随机分配至最少信息条件，22人被分配至中等信息条件。然而，部分回答不完整，因此不得不予以排除。最终，我们分析了37名参与者的数据，第一组19人，第二组18人。（59%为男性，41%为女性；$M_{age} = 27.8$，范围 = 21–38），他们主要是来自虚拟现实/增强现实、人机交互、机器学习/深度学习、神经影像学、心理学、计算机视觉、光子学和形式化方法领域的研究生。参与者来自多个国家和地区（加拿大、美国、墨西哥、伊朗、中国、土耳其），带来了多元视角，有助于减少对可视化进行解读时的文化偏见。在10分制的自我报告量表中，参与者对自身在人工智能开发（$M = 6.1$，标准差 = 2.63）、人工智能使用（$M = 7.3$，标准差 = 1.77）、数据可视化（$M = 6.1$，标准差 = 2.09）和阿尔茨海默病（$M = 3.1$，标准差 = 2.33）的经验进行了评分。[3]

6.1.1 任务*1*：阿尔茨海默病进展。我们首先评估了人工智能模型信息的影响（图3）。信任评分略高于中等信息量条件（$M = 6.99$，标准差 = 1.64），而非最少信息量条件（$M = 6.44$，标准差 = 1.66），但无论是独立的 t 检验还是线性混合模型，这一差异均未达到显著水平。尽管获得的模型信息最少，参与者仍表示希望了解更多背景信息："如果我能更多地了解这些预测是如何生成和可视化的，我会更加信任它们。"这一发现未能支持假设1，但定性证据表明，若要实质性地影响信任，可能需要更丰富或分层呈现的模型解释。

信任。为检验假设2，我们考察了参与者在测试过程中对人工智能预测的信任评分，这些评分是在每次试次后报告的。与连续编码相比，参与者对不确定性的二元编码给予了显著更高的信任评分（M = 7 19, = 2 52 M = 6 22, = 1 36 。SD .）（. SD .），$t(36) = 2.550$，$p < .05$。考虑重复测量的线性混合模型证实了这一效应（$p < .05$）。参与者解释说，尽管连续格式传达了更丰富的信息，但它们往往显得"模糊"或"嘈杂"，使决策边界不够清晰，并降低了信任："尽管提供了更多信息，但模糊的转变让预测变得含糊，也降低了我的信任。"这些结果仅为假设2提供了部分支持：使用连续不确定性并未提高即时信任，因为参与者当下更信任二元编码。

我们还评估了参与者完成每种可视化条件后收集的信任四个维度的区块级评分 [4],（图 5）。与二元编码相比，连续不确定性可视化显著提高了感知可靠性（$M = 6.92$，标准差 = 1.99；二元编码为 $M = 4.87$，标准差 = 2.83），$t(36) = -3.768$，$p < .001$。其他维度均未出现显著差异。定性反馈也印证了这一差别："如果没有将不确定性可视化 *[二元不确定性]*，就很容易对模型过度自信。可视化得越多，我就越能信任它的决策。"参与者还将二元编码描述为"没有提供有用信息，难以信任。"这些结果支持假设3：连续编码提高了感知可靠性，尽管它们降低了决策时的信任评分。

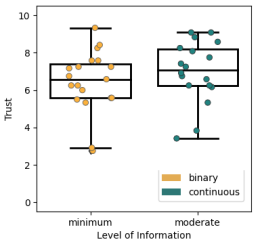


图3：模型信息水平对信任的影响。

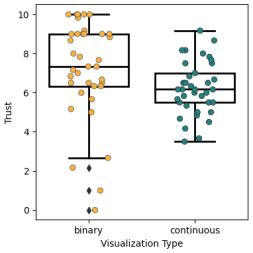


图4：二元与连续不确定性编码的比较及其对信任的影响。

正如观察到的，这些发现表明，参与者区分了当下的整体信任与事后评估的具体信任维度。连续编码提高了可靠性判断（支持假设3），但降低了整体信任（与假设2相悖），

[3]所有统计分析均使用 JASP v.0.17.2.1 完成，https://jasp-stats.org

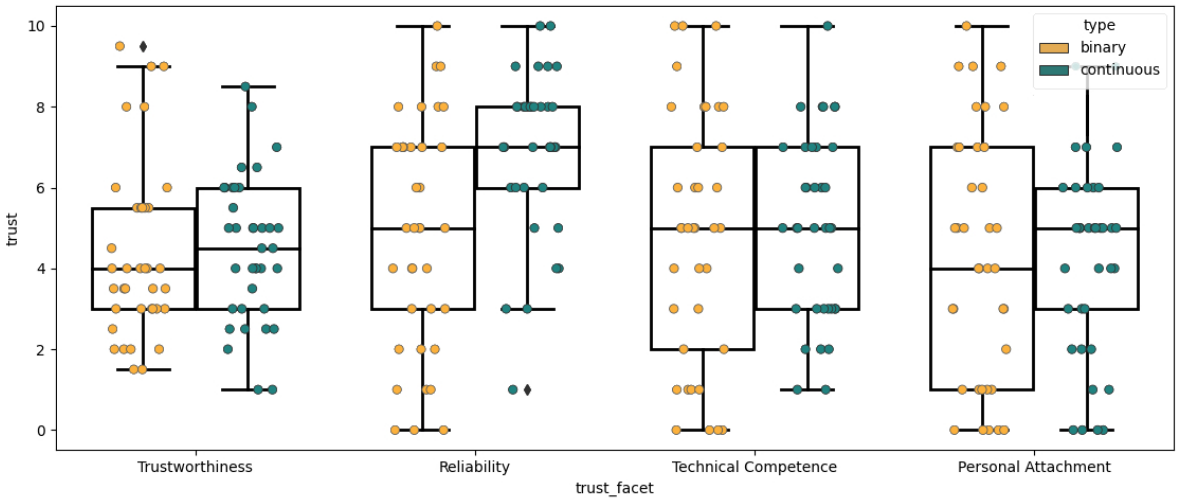


图5：根据 Ashooriand Weisz [4] 的多维信任框架，对人工智能不确定性的可视化表示（二元与连续）进行比较。

而二元编码保留了较高的总体信任，但提供的可靠性信号较弱。模型信息水平的影响较弱，这表明本研究未能支持假设1。

信心。为回答研究问题3，我们分析了参与者在每次试次后填写的信心评分。如图6所示，模型信息的详略程度显著影响了信心：与获得适度信息的参与者相比，获得极少细节的参与者信心较低（M = 6.83，SD = 2.42；适度信息：M = 8.70，SD = 1.54），$F(35)$ = 9.953，$p < .01$。因此，至少提供适度的模型细节，能够增强参与者基于人工智能模型输出做出决策时的信心。此外，人工智能不确定性的编码类型也显著影响了信心（图7）。与连续可视化相比，二元不确定性可视化引发了更高的信心（M = 8.70，SD = 1.34；M = 6.83，SD = 1.36；连续可视化：SD），$t(36)$ = 8.128，$p < .001$。在线性混合模型中控制重复测量和个体间变异后，这一差异仍然显著（$p < .001$）。这些结果明确回答了研究问题3：与连续编码相比，参与者在使用二元编码时信心更强。

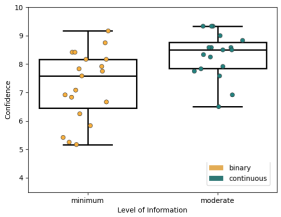


图6：模型信息水平对信心的影响。

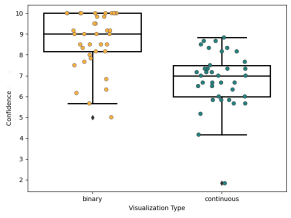


图7：比较二元与连续不确定性编码及其对信心的影响。

定性反馈为这些发现增添了更多细节。几位参与者表示，看到不确定性信息后信心有所增强，因为这帮助他们判断应该在多大程度上依赖模型："如果我不知道不确定性，我根本不会相信自己的判断"。另一位参与者也表示："我喜欢看到不确定性信息，这让我意识到自己能够（也应该）在多大程度上信任数据"。与此同时，也有人认为，当不确定性以更连续的方式呈现时，有时会削弱信心，让选择变得不那么明确："可视化不确定性确实拖慢了决策过程，也降低了信心，但提供的信息更丰富"。

综合来看，这些结果表明，尽管不确定性可视化能够提升信心，但其效果取决于呈现形式。乍看之下，二元编码产生更高信心似乎有悖直觉，因为连续编码能够传达有关模型确定性的更丰富细节，人们可能会认为它更有助于做出知情决策。然而，定性反馈显示，实际情况恰恰相反。连续编码中额外的视觉噪声和模糊边界有时带来的不是清晰度，而是疑虑。正如一位参与者所指出的，连续格式使决策边界变得"模糊"，从而降低了他们对自身决策的信心。相比之下，二元编码虽然较为粗略，却提供了更清晰的类别线索，有助于增强决策确信度，尽管代价是掩盖了模型的不确定性。

6.2 专家研究

我们招募了 10 位在神经影像学和痴呆症研究（从事痴呆症研究的人工智能研究人员）领域拥有 10 年以上经验的领域专家，其中包括专攻磁共振成像、衰老和人工智能的研究人员，以及教授、神经科医生和数据科学家。参与者的平均年龄为 42.8 岁（范围：35–51 岁）；8 位自我认同为男性，2 位自我认同为女性。在 1（毫无经验）至 10（经验非常丰富）的量表上，参与者表示自己在人工智能开发（M = 8.0, SD = 2.16）、人工智能使用（M = 8.3, SD = 1.82）、数据可视化（M = 7.4, SD = 1.50）和阿尔茨海默病（M = 6.9, SD = 2.55）方面具备较高的专业知识。

6.2.1 任务：人工智能不确定性预测阿尔茨海默病进展。

信任。 图 9（左）展示了专家在各项任务中的信任评分，对比了有无可视化不确定性的人工智能预测，并按低（<30%）和高（>30%）不确定性条件进行了区分。总体而言，连续可视化比二元可视化更能提升信任，尤其是在高不确定性条件下。配对样本 t 检验确认，连续编码显著提高了专家对模型在高不确定性试次中预测结果的信任（$p < .05$），这表明以连续形式呈现不确定性，可以让模型局限性更加透明。定性反馈也印证了这一趋势：专家指出，连续显示有助于他们更好地评估预测可靠性。

四个信任维度的区块级评分凸显了连续不确定性可视化的益处。如图8所示，当不确定性以连续方式呈现时，所有维度的信任评分都明显更高。在二元条件下，平均分数较低：个人依恋（$M = 1.8$, $SD = 2.35$）、可靠性（$M = 2.6$, $SD = 3.31$）、技术能力（$M = 1.6$ $SD = 1.84$ $M = 1.8$ $SD = 2.04$ ）和可信度（ ）。相比之下，连续条件下的评分显著更高：个人依恋（$M = 8.2$, $SD = 2.35$）、可靠性（$M = 7.4$, $SD = 3.31$）、技术能力（$M = 8.4$, $SD =$ 1.84 $M = 8.2$ $SD = 2.04$ ）和可信度（ ）。配对样本t-检验确认，所有信任维度的差异均具有统计显著性（可信度 $p < .001$、可靠性 $p = .047$、技术能力 $p < .001$、个人依恋 $p < .001$）。因此，这些发现为**假设4**提供了有力支持，表明连续不确定性可视化显著增强了专家对人工智能预测的信任。

信心。图 9（右）展示了专家对其决策的信心评分。在高不确定性试次（$p = .014$）和低不确定性试次（$p = .003$）中，二元可视化都显著提高了信心。这些结果表明，二元输出可能会将预测呈现为类别化结果，从而助长过度自信；而连续可视化则提供了更细致的线索，有助于促使人们以经过校准且谨慎的方式进行决策。

决策变化。 决策变化定义为：同一位患者的预测以二元可视化和连续可视化呈现时，判断出现差异。专家并未被告知同一患者案例出现了两次，以确保任何决策差异反映的是可视化形式，而非有意进行交叉核对。在所有试次中，专家在66个案例中的23个案例（34.8%）改变了决策，其余65.2%的回答在不同形式下保持一致。按不确定性水平细分后，高不确定性条件下出现了12次决策变化（36.4%），低不确定性条件下出现了11次（33.3%）。卡方检验表明，这种分布并不显著（$\chi^2 = 0.0, p = 1.0$）。考察不确定性水平与可视化模式之间的交互作用时，高不确定性条件下的决策变化（48次试次中有17次；35.4%）比低不确定性条件下（48次试次中有10次；20.8%）更频繁。尽管这表明高不确定性与连续可视化相结合时，可能促使专家更多地重新考虑判断，但卡方检验表明，这一差异在统计学上并不显著（$\chi^2 = 1.86, p = .17$）。

在23次决策变化中，14次（60.9%）发生在使用连续可视化时，9次（39.1%）发生在使用二元编码时。尽管这表明专家在查看连续可视化时更常修正自己的判断，但二项检验证实，这一差异在统计学上并不显著（$p = .40$）。

表 1：按可视化方向和不确定性水平划分的决策变化。

	高不确定性	低不确定性	总计
无变化	21	22	43
二元 < 连续	3	11	14
二元 > 连续	9	0	9
总计	33	33	66

后测问卷。 与采用二元编码时相比，专家表示，当不确定性以连续方式可视化时，他们需要的有关人工智能模型的补充信息较少（$M = 7.5$, $SD = 1.62$；$M = 8.8$, $SD = 7.5$）。开放式反馈也印证了这一趋势：专家认为，连续不确定性显示能够提升人工智能预测的可信度，并有助于解读不同的置信水平。

与此同时，专家一致强调，仅凭不确定性不足以进行验证。信任更多地建立在外部验证、透明度和已展现出的模型性能之上。一位AI神经影像专家解释道“不确定性并不能证明效度”，并指出“在外部数据集上进行评估，是提升我对人工智能预测信任度的更好方式”。其他专家则强调以人为本的解释的价值，认为与单纯的概率可视化相比，“类人的解释”更能有效地增进信任。

尽管不确定性可视化通过明确承认预测不确定性，并区分高概率与低概率结果，增添了一些可信度，但这种可信度是有条件的。专家强调，只有结合与模型训练、假设和局限性相关的信息，可视化才有意义：“将不确定性可视化固然不错，但除非我能获得更多关于模型训练过程的信息，否则这对我来说意义不大。”总体而言，反馈表明，不确定性可视化很有用，但必须融入更广泛的透明度与可解释性框架中。若缺乏充分的背景信息和验证，它们有效影响信任的能力，尤其是在专家群体中的影响力，可能会受到限制。

研究发现综述。在研究2中，连续不确定性可视化提高了专家的信任度，既体现在试次级评分中，也体现在所有任务后信任维度上；而二元可视化则带来了更高的信心，提示了过度自信的潜在风险。专家在看到同一位患者以不同编码呈现时，约有三分之一的情况下改变了决策，不过改变率并未随不确定性水平或可视化模式而变化。任务后反馈表明，连续可视化增强了可信度，并降低了对补充模型信息的需求；但专家强调，信任也取决于能否获取有关模型训练和评估的详细信息。

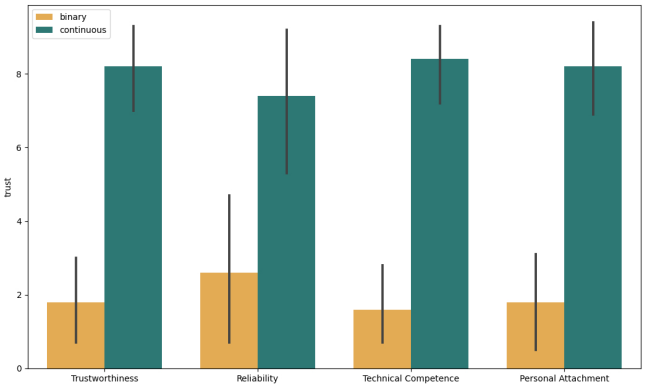


图8：使用 Ashoori 和 Weisz 的多维信任框架 [4] 比较专家对二元和连续人工智能不确定性可视化的信任。

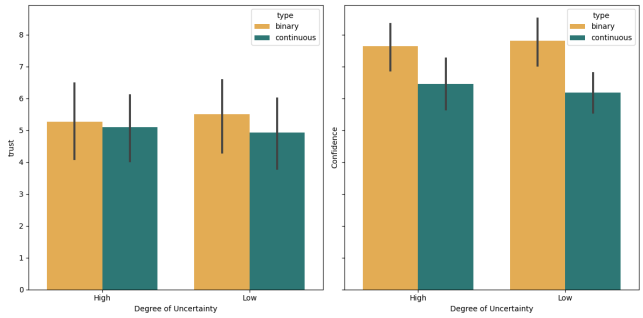


图9：专家群体按可视化类型和不确定性水平划分的信任（左）和信心（右）分布。

6.3 总结

总之，两项研究表明，不确定性可视化形式与用户专业知识共同影响着人们对人工智能预测的信任和信心。普通参与者当下更信任二元编码，但认为连续格式更可靠；而专家在连续可视化下所有维度的信任度都始终较高，尽管二元编码会抬高他们的信心。表2总结了这些发现，重点说明专业知识如何调节清晰度、可信度与过度自信之间的权衡。这些差异不仅表明，需要采用考虑不同专业知识水平的自适应可视化策略，也显示不确定性可视化作为一种可解释人工智能技术，能够阐明人工智能的预测能力，帮助用户认识模型局限性，并在做出复杂决策时校准对模型的依赖。

7 讨论

在近期关于阿尔茨海默病预后中人工智能不确定性估计的研究基础上 [22], ，我们的混合方法研究考察了通才用户和领域专家在不确定性以不同格式可视化时，如何解读并信任人工智能预测。在两个人群中，我们比较了二元（有色/无色）与连续（颜色饱和度）编码，以及多种可视化风格，以评估它们对信任、信心、可解释性和决策稳定性的影响。我们的发现拓展了先前关于不确定性可视化和对人工智能的信任的人机交互研究，强调了专业知识如何调节清晰度与透明度之间的平衡。

我们的结果与先前的发现相呼应，即可解释性本身并不总能提高信任 [50], 而结果反馈和情境线索往往更为重要 [17]。在我们的研究中，通才在当下更信任二元编码，并报告了更高的信心，但连续可视化改善了可靠性判断，并降低了过度自信的风险。 任务中的信任与区块后可靠性评分之间的这种差异表明，参与者可能将试次级任务中的“信任”映射为他们做出决策的感知能力，而非对模型质量的更广泛评估。这与先前工作表明，用户在与决策支持系统交互时，常常将信任与感知到的任务难易程度或决策清晰度混为一谈 [19, 38]。 相比之下，专家在连续显示的所有维度上都表现出持续更高的信任，尽管二元格式夸大了信心。这种差异呼应了信任具有多维性的论点 [4] 以及用户专业知识会影响不确定性的感知方式 [69]。

两项研究中的参与者均强调，仅靠不确定性可视化不足以验证人工智能预测。与信任理论一致 [42], , 参与者期望获得关于模型如何训练和验证的更充分透明度。专家尤其要求了解数据集、纳入/排除标准以及外部验证的细节，这强化了先前的人机交互研究发现，即解释在与情境证据结合时更为有效 [5]。因此，我们的结果拓展了先前工作，表明连续不确定性可以揭示模型局限性，但只有当此类视觉线索与人类可读的模型信息相结合时，信任才会得到增强。

7.1 局限性与未来工作

有必要承认本研究存在若干局限。普通人群样本主要由技术素养相对较高的研究生组成，因此研究结果的普遍适用性受到限制。这种偏差可能反映了首次用户研究招募中的网络效应。不过，由于这是一项旨在评估原型可用性的探索性研究，这一群体提供了恰当的初步反馈。我们承认，在线招募平台有助于招募更多样化的参与者。虽然这些平台可以接触到规模庞大、背景多样的参与者群体，但由于其中存在一些有充分记录的挑战，我们没有采用这些平台。 [71]例如本研究中没有唯一正确答案的任务，尤其容易受到数据质量问题的影响，包括粗心或恶意回答。此外，这些平台也难以确保参与者专注、认真地参与研究材料。出于这些原因，我们选择了更受控的招募方法。虽然这种方法的可扩展性较低，但更适合本次探索性评估的目标。

表2：研究1（广泛参与者）和研究2（专家）的结果汇总。箭头表示得分／效应较高的方向。

结果	操纵 / 比较	主要发现	支持
研究1：广泛参与者研究			
信任（信息水平）	极少与中等模型信息	差异不显著；参与者表示希望获得更多背景信息。	假设1:
信任（逐试验）	二元与连续不确定性	二元 ↑ 信任与连续不确定性（任务内）。	假设2:
信任维度（区块后）	二元与连续不确定性	可靠性 ↑ 与连续性相关（其他维度不显著）。	假设3: ✓
信心（信息水平）	最少模型信息 vs. 中等模型信息	中等信息 ↑ 信心。	研究问题3：中等
信心（可视化类型）	二元与连续不确定性	二元 ↑ 信心与连续不确定性。	研究问题3：二元
研究2：专家研究			
信任（每次试次）	二元与连续不确定性；低与高不确定性	连续 ↑ 信任，尤其是在高不确定性下。	假设4: ✓
信任维度（任务后）	二元与连续不确定性	所有信任维度 ↑ 均采用连续形式（差异较大）。	假设4: ✓
信心	二元与连续；低与高不确定性-水平	对于两种不确定性水平，均采用二元 ↑ 信心（可能存在过度自信）。	研究问题3：二元
决策变化	同一患者，二元与连续不确定性	23/66（34.8%）发生了变化；不同不确定性水平或可视化模式。	研究问题4：探索性分析
模型信息需求（后测）	二元与连续不确定性	使用连续不确定性时，报告称对补充模型信息的需求较少；专家仍要求提供训练/评估详情。	研究问题4：探索性

注释：对勾（✓）和叉号（）表示假设 假设结果（支持 / 不支持）。箭头表示数值较高。n.s. = 表示不显著。研究问题3和研究问题4标示与信心相关的探索性分析 nce, 决策变化，以及对额外 模型信息的感知需求。

我们承认，普通人群研究的最终样本量（N = 37）小于事先设定的目标（N = 42）。这是因为有若干份未完成的回答必须排除。不过，使用 G*Power 软件进行的敏感性分析表明，凭借最终样本量 $N = 37$，$\alpha = .05$，以及 80% 的统计功效，本研究足以检测到最小效应量 $d = 0.83$ (Cohen's d)。专家研究的样本量较小（N=10），并精简至 12 种刺激材料，这提高了研究的可行性，但限制了统计功效，也使我们无法探索中等程度的不确定性。最后，尽管我们考察了可视化格式，但没有测试其与结果反馈之间的交互作用；先前研究表明，结果反馈是影响信任的重要因素 [2]。

未来研究应在纵向和真实世界场景中考察不确定性可视化，在这些场景中，信任和依赖会随时间变化 [2, 17]。我们的研究结果指出了几个具体方向。首先，提供更丰富的模型上下文，包括训练数据、评估指标和可解释推理，可能进一步提升信任和决策质量。其次，采用结合二元线索与连续线索的自适应策略，或许能更好地适应用户在可视化素养方面的差异；同时，将不确定性与结果反馈相结合，并提供透明的模型文档，也可能进一步校准信任。第三，研究用户专业知识、认知负荷与可视化复杂度之间的相互作用，或可阐明为何缺乏经验的用户有时更偏好复杂的可视化。最后，在临床工作流程中开展纵向研究，可评估不确定性可视化如何影响长期信任、依赖和患者结局。将这些见解与多维信任 [4] 和人工智能透明度 [37] 方面的既有研究相结合，将有助于为可解释、可信赖的人工智能系统提供设计依据，使其能够在医疗保健领域支持安全、以人为本、可解释且负责任的人工智能。

这些发现凸显了对人工智能预测的信任所具有的多维本质，以及根据任务情境和用户专业知识来定制不确定性可视化的重要性。尽管我们的分析揭示了二元编码与连续编码之间的权衡，并表明在缺乏情境化模型信息的情况下，仅靠不确定性提示是不够的，但它们也指向了可付诸实践的设计策略。在下一节中，我们将这些见解提炼为一组具体的设计指南（表3），以支持开发能够促进对人工智能预测进行校准依赖的临床决策支持系统。

7.2 人工智能不确定性可视化的设计指南

基于不确定性可视化 [27, 33] 和人机信任 [4, 38], 方面的既有研究，并以我们从研究1和研究2中获得的实证发现为基础，我们提出了六项指南，用于设计预后人工智能系统中的不确定性表征。为明确这些指南的新贡献，我们明确区分：(a) 既有文献已经提出的建议，(b) 本研究在阿尔茨海默病长期预后情境下的补充发现，以及 (c) 对临床人工智能设计的启示。

G1. 使用连续不确定性支持可靠性判断。既有研究：不确定性显示可以提升校准后的依赖，但也可能增加认知负荷 [27, 33]。我们的发现：连续不确定性提高了研究1中的感知可靠性，并提升了研究2中专家在所有信任维度上的评分，尽管它降低了通才的瞬时信任。启示：当目标是支持对模型可靠性的反思性评估时，尤其是在预后情境中，应使用连续不确定性。

G2. 谨慎使用二元编码，以避免过度自信。既有研究：简化编码可以提升果断性，但可能掩盖不确定性 [50]。我们的发现：二元不确定性提高了通才和专家的信心，却未相应提升可信度或可靠性。启示：将二元显示用于快速分诊，并搭配对剩余不确定性的提醒。

G3. 将不确定性可视化与简明、分层的模型信息相结合。既有研究：轻量级模型文档有助于用户理解预测的背景 [43]。我们的发现：适度的模型信息提升了信心（研究1）；即使展示了连续不确定性，专家仍要求提供模型细节（研究2）。启示：将不确定性可视化图示与简要的模型摘要搭配，并按需提供更深入的文档。

G4. 针对多维信任进行设计与评估。既有研究：信任包含可靠性、能力和可预测性等维度 [4, 38]。我们的发现：研究1中，连续不确定性提高了可靠性，但降低了瞬时信任。尽管如此，研究2中它提高了所有信任维度，而二元编码则夸大了信心。启示：应针对不同的信任维度进行测量与设计，而不是依赖单一的总体信任评分。

G5. 在长期预后中突出呈现不确定性。既有研究：与诊断相比，预后涉及更大的不确定性，反馈机会也更少 [28]。我们的发现：当不确定性呈现形式发生变化时，专家在三分之一的试次中改变了决策，这表明不确定性会影响推理，但不会使决策变得不稳定。启示：对于长期预测（例如阿尔茨海默病进展的24个月预测），必须让不确定性十分醒目。

G6. 将不确定性可视化融入更广泛的透明度与验证之中。既有研究：与单纯提供解释相比，模型性能证据更能提升信任 [5, 66]。我们的发现：专家强调，“不确定性并不能证明效度”，并要求提供外部验证和训练详情。启示：将不确定性显示与验证证据、已知失效模式及简明解释结合起来。

指导原则	从既有研究中获知 Work	本研究的贡献
G1：连续 以提升可靠性	不确定性 显示 能够支持校准后的 依赖	可靠性 ↑但短暂的 信任 ↓对通才而言；专家表现出最高的信任 采用连续
G2：二元，采用 谨慎	更简单 显示 in- 增加果断性	二元↑信心适用于所有 用户；潜在性的证据 专家过度自信
G3: Pair un- 确定性 with 模型上下文	用户受益于简要的 关于……的信息 预测是如何生成的- ated	中等模型信息 息 ↑信心；专家 仍然要求提供训练和 验证详情
G4：多维 的 信任	信任涉及不同的 维度（例如可靠性、 能力）	不同的编码会以不同方式改变 各个维度；conf i- 证据与信任出现分歧
G5：让不确 定性在 预后中凸显	预后任务通常涉 及更高的不确 比诊断更确定	不确定性呈现形式促使 专家重新考虑决策, 占病例的 1/3
G6: 整合 不确定性 with 更广泛 透明度	仅有解释并不能 确保信任	专家：“不确定性并非 效度”；他们需要解- 外部验证与模型 证据

表3：六项指南的要点式总结，区分既有知识与本研究的独特贡献。

8 结论

不确定性可视化正日益被提出，作为使人工智能系统更加透明的一种方式，但我们对具体编码如何影响对人工智能的信任和决策仍然知之甚少在长期、高不确定性的医学预后中。本文在人工智能支持的阿尔茨海默病进展这一背景下填补了这一空白；在该领域，预测本质上具有不确定性，结果会延迟出现，而校准不当的信任可能带来严重的临床后果。

在两项研究中，我们考察了不同编码方式（如二元和连续不确定性）如何影响对预测的CN/MCI/AD转变的信任、信心和决策稳定性。研究1中，参与者表示，相较于连续编码，二元编码带来了校准得更恰当的任务中的信任和信心。连续编码则有选择地提升了感知可靠性。研究2中，尽管二元输出再次带来了更高的信心，但在连续可视化不确定性时，领域专家在所有维度上的信任校准都显著更恰当，这引发了人们对分类显示中专家过度自信的担忧。专家在查看采用不同不确定性编码的同一患者时，约有三分之一的情况下改变了决策；他们还强调，“不确定性并不能证明效度”，并指出其信任建立在外部验证和模型证据之上。

这些发现带来了三项贡献。第一，它们提供了实证证据，表明不确定性可视化对阿尔茨海默病预后的影响并非一概而论：可视化形式与用户专业知识共同决定不确定性是否会提升感知可靠性、抬高信心，或两者兼有。第二，它们表明，对人工智能预测的信任、可靠性判断和信心可能以系统性的方式彼此分离，凸显了将信任视为多维概念，并在评估自我报告的同时考察行为依赖的必要性。第三，它们提出了六项设计指南，区分了既有研究提出的建议与本研究特有的发现：使用连续编码以支持可靠性判断；谨慎采用二元格式；将不确定性与简明的模型上下文相结合；面向多维信任进行设计；在长期预后中突出呈现不确定性；并将不确定性融入更广泛的透明度与验证之中。

总之，我们的结果表明，临床人工智能中的不确定性可视化并非简单地“添加误差条”，而是需要根据用户专业知识、任务时域以及实践中所期望的信任类型，审慎地匹配编码。通过将设计指南建立在人工智能辅助阿尔茨海默病预后中观察到的实证权衡之上，本研究既推进了不确定性可视化研究，也推进了面向神经退行性疾病的可解释、可信赖决策支持系统的设计。未来工作应在纵向及人在回路的临床研究中扩展这些发现；在这些研究中，不确定性提示、结果反馈与模型更新会随时间相互作用，从而塑造信任、依赖，并最终影响患者护理。

9 致谢

参考文献

[1] Ahmad A Abujaber、Abdulqadir J Nashwan 和 Adam Fadlalla。2022年。推动机器学习在临床决策支持中的应用：一种总体解释结构建模方法。《医学信息学解锁》33 (2022)，101090。[2] Daehwan Ahn、Abdullah Almaatouq、Monisha Gulabani 和 Kartik Hosanagar。2024年。模型可解释性和结果反馈对人工智能信任的影响。载于*CHI*计算机系统中的人为因素会议论文集。ACM。

[3] Abeer Al-Nafjan、Amaal Aljuhani、Arwa Alshebel、Asma Alharbi 和 Atheer Alshehri。2025年。预测性医疗中的人工智能：系统综述。《临床医学杂志》14，19（2025年），6752。[4] Maryam Ashoori 和 Justin D. Weisz。2019年。人工智能可信性与接受度：用户体验的作用。载于第24届国际智能用户界面会议论文集。ACM，263 274 10 1145 3301275 3302310 –。doi：。/。[5] Gagan Bansal 等。整体是否大于部分之和？人工智能解释对互补性团队绩效的影响。载于 CHI计算机系统中的人为因素会议论文集。ACM。[6] Gagan Bansal、Besmira Nushi、Ece Kamar、Walter S. Lasecki、Daniel S. Weld 和 Eric Horvitz。2019年。超越准确性：心理模型在人机团队绩效中的作用。载于AAAI人类计算与众包会议论文集，第7卷。AAAI出版社，2–11。https://ojs.aaai.org/index.php/HCOMP/article/view/5285[7] Renato Antonio Bertão 和 Jaewoo Joo。2021年。UX/UI设计中的人工智能：当前应用与 [未来] 实践调查。《设计研究的安全港》（2021年），1–10。[8] MAPPING A Better。2023年。阿尔茨海默病事实与数据。《阿尔茨海默病与痴呆》19，4（2023年），1598–1695。[9] Nadia Boukhelifa、Anastasia Bezerianos、Tobias Isenberg 和 Jean-Daniel Fekete。2012年。评估草图性作为呈现定性不确定性的视觉变量。《IEEE可视化与计算机图形学汇刊》18，12（2012年），2769–2778。[10] Leonard A Breslow、Raj M Ratwani 和 J Gregory Trafton。2009年。色阶对数据可视化任务影响的认知模型。《人因》51，3（2009年），321–338。[11] Zana Buçinca、Maja Barbara Malaya 和 Krzysztof Z Gajos。2021年。信任还是思考：认知强制功能可减少人工智能辅助决策中对人工智能的过度依赖。《ACM人机交互论文集》5，CSCW1（2021年），1–21。[12] Carrie J. Cai、Emily Reif、Narayan Hegde、Jason Hipp、Been Kim、Daniel Smilkov、Martin Wattenberg、Fernanda Viégas、Greg Corrado、Martin Stumpe 和 Michael Terry。2019年。以人为中心的工具：在医疗决策中应对不完善的算法。载于2019年CHI计算机系统中的人为因素会议论文集（苏格兰格拉斯哥）（CHI '19）。计算机协会，美国纽约州纽约市，1 14 10 1145 3290605 3300234 –。doi：。/。[13] Daniel N Cassenti、Lance M Kaplan 和 Aayushi Roy。2023。以供人类理解的方式呈现人工智能的不确定性信息。载于人因与工效学学会年会论文集，第67卷。SAGE出版社，SAGE，加利福尼亚州洛杉矶：177–182。[14] Mohammad Daradkeh 和 Bilal Abul-Huda。2017年。将不确定性纳入决策：一种信息可视化方法。载于《决策支持系统VII：决策支持系统中的数据、信息与知识可视化：第三届 ICDSST 2017 国际会议》，31，2017年5月29日至31日，比利时那慕尔，论文集。施普林格，–。[15] Munjal Desai、Piyush Kaniarasu、Michael Medvedev、Aaron Steinfeld 和 Holly A. Yanco。2012年。机器人故障与反馈对实时信任的影响。载于第7届ACM/IEEE人机交互国际会议论文集（HRI ' *12* 251 258 10 1145 2157689 2157833 ）。ACM，–。doi：。/。[16] Berkeley J Dietvorst 和 Soaham Bharti。2020。人们会在不确定的决策领域拒绝算法，因为他们对预测误差的敏感度会降低。《心理科学》31，10（2020年），1302–1314。[17] Berkeley J. Dietvorst、Joseph P. Simmons 和 Cade Massey。2015年。算法厌恶：人们看到算法出错后，会错误地避开算法。《实验心理学杂志：总论》144, 1 2015 114 126 10 1037 0000033 （），–。doi：。/xge。[18] Achref Doula、Lennart Schmidt、Max Mühlhäuser 和 Alejandro Sanchez Guinea。2022年。基于增强现实的透视应用中的机器学习不确定性可视化。载于2022年IEEE人工智能与虚拟现实国际会议（AIVR）。IEEE，109–113。[19] Mary T. Dzindolet、L. Greg Pierce、Heather P. Beck、Laurie A. Dawe 和 Stephen E. Anderson。2003年。信任在自动化依赖中的作用。《国际人机研究杂志》58，6（2003年），697–718。doi:10.1016/S1071-5819(03)00038-7[20] Shaker El-Sappagh、Jose M Alonso-Moral、Tamer Abuhmed、Farman Ali 和 Alberto Bugarín-Diz。2023年。阿尔茨海默病领域的可信人工智能：现状、机遇与挑战。《人工智能评论》56，10（2023年），11149–11296。[21] Simon F Eskildsen、Pierrick Coupé、Daniel García-Lorenzo、Vladimir Fonov、Jens C Pruessner、D Louis Collins、阿尔茨海默病神经影像学计划等。2013年。利用皮层变薄模式预测ADNI队列中轻度认知障碍受试者的阿尔茨海默病。《神经影像》65（2013年），511–521。[22] Mohammad Eslami、Solale Tabarestani 和 Malek Adjouadi。2023年。在阿尔茨海默病纵向研究中，结合多模态信息融合与深度学习的独特颜色编码可视化系统。《医学人工智能》140（2023年），102543。[23] Marzyeh Ghassemi、Luke Oakden-Rayner 和 Andrew L Beam。2021年。当前医疗保健领域可解释人工智能方法所带来的虚假希望。《柳叶刀·数字健康》3, 11（2021），e745–e750。[24] Raquel González-Alday、Esteban García-Cuesta、Casimir A Kulikowski 和 Victor Maojo。2023。医学领域可解释人工智能的进展、适用性与未来：范围综述。《应用科学》13, 19（2023），10778。[25] Zhenyu Guo、Matthew O Ward 和 Elke A Rundensteiner。2009。用于多变量线性趋势发现的模型空间可视化。载于《2009年IEEE可视分析科学与技术研讨会》。IEEE，75–82。[26] Xin He、Yeyi Hong、Xi Zheng 和 Yong Zhang。2023。用户有哪些需求？以用户为中心的可解释人工智能诊断系统设计。《国际人机交互杂志》39, 7（2023），1519–1542。[27] Jessica Hullman。2019。作者为何不对不确定性进行可视化。载于《IEEE可视化会议论文集（VIS）》。IEEE，240–244。doi:10.1109/VISUAL. 2019.8933545[28] Clifford R Jack Jr、David A Bennett、Kaj Blennow、Maria C Carrillo、Billy Dunn、Samantha Budd Haeberlein、David M Holtzman、William Jagust、Frank Jessen、Jason Karlawish 等。2018。NIA-AA研究框架：迈向阿尔茨海默病的生物学定义。《阿尔茨海默病与痴呆》14, 4（2018），535–562。[29] Jieyu-Ann Jian、Ann M. Bisantz 和 Colin G. Drury。2000。实证确定的自动化系统信任量表的基础。载于《人因与工效学学会年会论文集》，第44卷。SAGE出版社，501 505 10 1177 154193120004400717 –。doi:。/ Taeho Jo、Kwangsik Nho 和 Andrew J. Saykin。2019。阿尔茨海默病领域的深度学习：利用神经影像数据进行诊断分类和预后预测。《衰老神经科学前沿》11（2019），220。doi:10.3389/ fnagi. 2019 00220 。Ally Salim Jr、Megan Allen、Kelvin Mariki、Kevin James Masoy 和 Jafary Liana。2023。了解人工智能决策支持工具的使用如何影响坦桑尼亚药品发放人员的批判性思维及其对技术的过度依赖。arXiv: 2302 09487 [ ] 2302 09487 。cs.HC https://arxiv.org/abs/。Kensaku Kawamoto、Caitlin A Houlihan、E Andrew Balas 和 David F Lobach。2005。利用临床决策支持系统改善临床实践：旨在确定成功关键特征的试次系统综述。《英国医学杂志》330, 7494（2005），765。[33] Matthew Kay、Tessa Kola、Jessica Hullman 和 Sean A. Munson。2016。超越"怪异、乏味且困难"：信息可视化的定性方法。载于《2016年CHI计算机系统中的人为因素会议论文集》。ACM，4953 4964 10 1145 2858036 2858384 –。doi:。/。Alex Kendall 和 Yarin Gal。。贝叶斯深度学习与计算机视觉中，我们需要哪些不确定性？《神经信息处理系统进展》30（2017）。[35] M Khojaste-Sarakhsi、Seyedhamidreza Shahabi Haghighi、SMT Fatemi Ghomi 和 Elena Marchiori。2022。阿尔茨海默病诊断的深度学习：一项调查。《医学人工智能》130（2022），102332。[36] Krishan Kumar 和 Kiran Jyoti。2025。提升脑肿瘤诊断的透明度与信任：对深度学习和可解释人工智能技术的深入分析。（2025）。[37] Vivian Lai、Q. Vera Liao 和 Chenhao Tan。2023。迈向可信人工智能辅助：在人机协同决策中设计适当依赖。载于《2023年CHI计算机系统中的人为因素会议论文集》。ACM，1 15 10 1145 3544548 3581089 –。doi:。/。John D. Lee 和 Katrina A. See。。对自动化的信任：为适当依赖而设计。《人因》46, 1 2004 50 80 10 1518 46 1 50 30392 ( ), –。doi:。/hfes.。。Dirk Leffrang 和 Oliver Müller。。我应该遵循这个模型吗？不确定性可视化对时间序列预测接受度的影响。载于《2021年IEEE可视分析信任与专业知识研讨会（TREX）》。IEEE，20–26。[40] Q Vera Liao、Daniel Gruen 和 Sarah Miller。2020。质疑人工智能：为可解释人工智能用户体验提供设计实践参考。载于《2020年CHI计算机系统中的人为因素会议论文集》。1–15。[41] Laura R Marusich、Jonathan Z Bakdash、Yan Zhou 和 Murat Kantarcioglu。2023。利用人工智能不确定性量化改善人类决策。arXiv预印本 arXiv:2309.10852（2023）。[42] Roger C Mayer、James H Davis 和 F David Schoorman。1995。组织信任的整合模型。《管理学会评论》20, 3（1995），709–734。[43] Margaret Mitchell、Simone Wu、Andrew Zaldivar、Parker Barnes、Lucy Vasserman、Ben Hutchinson、Elena Spitzer、Inioluwa Deborah Raji 和 Timnit Gebru。2019。用于模型报告的模型卡。载于《公平、问责与透明会议论文集》。220–229。[44] Brent Mittelstadt、Chris Russell 和 Sandra Wachter。2019。解释人工智能中的解释。载于《公平、问责与透明会议论文集》。279–288。[45] Graziella Orru、William Pettersson-Yeo、Andre F Marquand、Giuseppe Sartori 和 Andrea Mechelli。2012。利用支持向量机识别神经系统与精神疾病的影像生物标志物：一项批判性综述。《神经科学与生物行为评论》36, 4（2012），1140–1152。[46] Lace Padilla、Matthew Kay 和 Jessica Hullman。2020。不确定性可视化。（2020）。

[47] Lace MK Padilla、Maia Powell、Matthew Kay 和 Jessica Hullman。2021。对不确定性的不确定：预测者信心的定性表达如何影响不确定性可视化下的决策。《心理学前沿》11（2021），579267。[48] Eun Hee Park、Karl Werder、Lan Cao 和 Balasubramaniam Ramesh。2022。家庭成员为何拒绝医疗保健中的人工智能？情绪的相互竞争影响。《管理信息系统杂志》39, 3（2022），765–792。[49] Ronald C. Petersen。2010。轻度认知障碍：临床特征描述与结果。《神经病学档案》67, 3（2010），303–308。doi:10.1001/archneurol. 2010.27[50] Forough Poursabzi-Sangdeh、Daniel G. Goldstein、Jake M. Hofman、Jennifer Wort-man Vaughan 和 Hanna Wallach。2021。操纵和测量模型可解释性。载于《2021年CHI计算机系统中的人为因素会议论文集（CHI ' *21* 1 52 10 1145 3411764 3445315 ）》。[51] ACM，–。doi:。/。Martin Prince、Anders Wimo、Maëlenn Guerchet、Gemma-Claire Ali、Yu-Tzu Wu 和 Matthew Prina。2015。《2015年世界阿尔茨海默病报告》。痴呆症的全球影响：对患病率、发病率、成本和趋势的分析。博士论文。国际阿尔茨海默病协会。[52] Jonatan Reyes、Anil Ufuk Batmaz 和 Marta Kersten-Oertel。2025。信任人工智能：不确定性可视化会影响决策吗？《计算机科学前沿》7（2025），1464348。[53] Mauricio Reyes、Raphael Meier、Sérgio Pereira、Carlos A Silva、Fried-Michael Dahlweid、Hendrik von Tengg-Kobligk、Ronald M Summers 和 Roland Wiest。2020。放射学中人工智能的可解释性：挑战与机遇。《放射学：人工智能》2, 3（2020），e190043。[54] Marco Tulio Ribeiro、Sameer Singh 和 Carlos Guestrin。2016。"我为什么要相信你？"解释任意分类器的预测。载于《第22届ACM SIGKDD知识发现与数据挖掘国际会议论文集》。1135–1144。[55] Emely Rosbach、Jonathan Ganz、Jonas Ammeling、Andreas Riener 和 Marc Aubreville。2024。时间压力下计算病理学中人工智能辅助医疗决策的自动化偏见。arXiv:2411.00998 [cs.HC] https://arxiv.org/abs/2411.00998[56] Shirin Salimi、Muireann Irish、David Foxe、John R Hodges、Olivier Piguet 和 James R Burrell。2018。视空间测量能否改善阿尔茨海默病的诊断？《阿尔茨海默病与痴呆：诊断、评估与疾病监测》10（2018），66–74。[57] Sara Salimzadeh、Gaole He 和 Ujwal Gadiraju。2024。应对不确定性：了解预后任务与诊断任务对人机决策中的信任和依赖的影响。载于《2024年CHI计算机系统中的人为因素会议论文集》。1–17。[58] Martin E. Schmidt、Martin Dichgans 和 Hilal Ünal。2018。缺失数据及其对阿尔茨海默病临床试验的影响。《阿尔茨海默病杂志》61, 3（ 2018 1125 1133 10 3233 170641 ），–。doi:。/JAD- John Seymour。。[59] 2020为什么CIELAB的a*轴指向品红色而不是红色？《色彩研究与应用》45, 6（2020），1040–1054。[60] Reed T Sutton、David Pincock、Daniel C Baumgart、Daniel C Sadowski、Richard N Fedorak 和 Karen I Kroeker。2020。临床决策支持系统概述：优势、风险与成功策略。《NPJ数字医学》3, 1（2020），17。[61] Steve G Sutton、Vicky Arnold 和 Matthew Holt。2018。自动化到什么程度才算过度？让知识工作中的人类保持相关性。《会计新兴技术杂志》15, 2（2018），15–25。[62] Meng-Shan Tan、Xi Ji、Jie-Qiong Li、Wei Xu、Hui-Fu Wang、Chen-Chen Tan、Qiang Dong、Chuan-Tao Zuo、Lan Tan、John Suckling 等。2020。无痴呆老年人阿尔茨海默病ATN生物标志物的纵向变化轨迹。《阿尔茨海默病研究与治疗》12, 1（2020），55。[63] Muhammad Tanveer、Bharat Richhariya、Riyaj Uddin Khan、Ashraf Haroon Rashid、Pritee Khanna、Mukesh Prasad 和 Chin-Teng Lin。2020。阿尔茨海默病诊断的机器学习技术：综述。《ACM多媒体计算、通信与应用汇刊（TOMM）》16, 1 2020 1 35 s（），–。[64] Alexandra-Maria Tăuţan、Bogdan Ionescu 和 Emiliano Santarnecchi。2021。神经退行性疾病中的人工智能：聚焦机器学习技术的现有工具综述。《医学人工智能》117（2021），102081。[65] Andrea Vázquez-Ingelmo、Francisco J García-Peñalvo 和 Roberto Therón。2020。聚合偏差：一项旨在提高对可视分析中包容性认识的提案。载于《世界信息系统与技术大会》。施普林格，409–417。[66] Oleksandra Vereschak 和合作者。2024。谁需要信任？利益相关者对人工智能系统信任的看法。载于《CHI计算机系统中的人为因素会议论文集》。ACM。[67] Vimbi Viswan、Noushath Shaffi、Mufti Mahmud、Karthikeyan Subramanian 和 Faizal Hajamohideen。2024。阿尔茨海默病分类中的可解释人工智能：系统综述。《认知计算》16, 1（2024），1–44。[68] Chu Wang、Zhengyi Wang、María Trinidad Herrero、Tao Xu、Fengfeng Chu、Hong Zeng 和 Ming Tao。2025年。利用基于多模态特征的深度学习模型对中国老年人群中的轻度认知障碍和阿尔茨海默病进行早期诊断。《亚洲精神病学杂志》（2025），104632。[69] Lixu Wang、Shichao Xu、Xiao Wang 和 Qi Zhu。2021年。解决联邦学习中的类别不平衡问题。载于*AAA*人工智能会议论文集，第35卷。10165–10173。[70] Youmei Wang、Chenchen Liu 和 Yun-Fang Tu。2021年。影响高等教育中基于人工智能的应用采用的因素。《教育技术与社会》24，3（2021），116–129。[71] Kerri Wazny。2017年。"众包"十年：一篇综述。《全球卫生杂志》7，2（2017），020602。[72] Bengt Winblad、Philippe Amouyel、Sandrine Andrieu、Clive Ballard、Carol Brayne、Henry Brodaty、Angel Cedazo-Minguez、Bruno Dubois、David Edvards-son、Howard Feldman 等。2016年。战胜阿尔茨海默病及其他痴呆：欧洲科学与社会的优先事项。《柳叶刀·神经病学》15，5（2016），455–532。[73] Oskar Wysocki、Jessica Katharine Davies、Markel Vigo、Anne Caroline Armstrong、Dónal Landers、Rebecca Lee 和 André Freitas。2023年。评估人工智能模型与医疗专业人员之间的沟通鸿沟：人工智能驱动的临床决策中的可解释性、效用和对人工智能的信任。《人工智能》316（2023），103839。[74] Heming Zhang、Dekang Cao、Zirui Chen、Xiuyuan Zhang、Yixin Chen、Cole Sessions、Carlos Cruchaga、Philip Payne、Guangfu Li、Michael Province 等。2024年。mosGraphGen：一种生成多组学信号图的新工具，以促进集成且可解释的图人工智能模型开发。《生物信息学进展》4，1（2024），vbae151。[75] Heming Zhang、Dekang Cao、Tim Xu、Emily Chen、Guangfu Li、Yixin Chen、Philip Payne、Michael Province 和 Fuhai Li。2025年。MosGraphFlow：一种从多组学数据中挖掘信号靶点的新型集成图人工智能模型。《*BMC*方法》2，1（2025），23。[76] Qing Zhao、Hanrui Xu、Jianqiang Li、Faheem Akhtar Rajput 和 Liyan Qiao。2023年。人工智能在阿尔茨海默病研究中的应用。《清华大学学报（自然科学版）》29，1（2023），13–33。[77] Xujiang Zhao、Feng Chen、Shu Hu 和 Jin-Hee Cho。2020年。图数据上的不确定性感知半监督学习。《神经信息处理系统进展》33（2020），12827–12836。