arXiv:2301.05809v1 [cs.HC] 14 Jan 2023

我该相信谁：人工智能还是自己？利用人类与人工智能正确可能性，促进人工智能辅助决策中的适当信任

SHUAI MA,香港科技大学，中国YING LEI, 华东师范大学，中国XINRU WANG, 普渡大学，美国CHENGBO ZHENG,香港科技大学，中国
CHUHAN SHI,香港科技大学，中国MING YIN,普渡大学，美国XIAOJUAN MA,香港科技大学，中国

在人工智能辅助决策中，人类决策者必须知道何时应信任人工智能，何时应相信自己。然而，先前研究仅依据人工智能置信度（即人工智能正确的正确可能性，CL）来校准人类信任，却忽略了人类自身的正确可能性，从而妨碍了团队决策达到最佳效果。为弥补这一不足，我们提出在任务实例层面，根据双方的正确可能性促进人类的适当信任。我们首先通过近似人类的决策模型，并计算其在相似实例中的潜在表现，对人类的正确可能性进行建模。我们通过两项初步研究证明了该模型的可行性和有效性。随后，我们提出三种CL利用策略，在人工智能辅助决策过程中显式或隐式地校准用户的信任。一项被试间实验（N=293）的结果表明，与仅使用人工智能置信度相比，我们的CL利用策略能促进人类对人工智能形成更适当的信任。我们进一步提出了对打造更符合人类需求的人工智能辅助决策的实践启示。

ACM 概念分类：• **以人为中心的计算 → 人机交互实证研究**。

附加关键词与短语: P 、对人工智能的信任、信任校准 校准、对人工智能的信任、信任校准

ACM 参考格式：
Shuai Ma、Ying Lei、Xinru Wang、CHENGBO ZHENG、CHUHAN SHI、MING YIN 和 XIAOJUAN MA。2023。《我该相信谁：人工智能还是自己？运用人类与人工智能正确性似然，促进人工智能辅助决策中的适当信任》。1，1（2023年1月），27页。https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 介绍

人工智能（AI）系统日益广泛地应用于各种决策场景 [24, 27, 49, 96, 101]。然而，在许多现实应用中，人工智能的准确率仍远未达到 100% [3, 14]。此外，由于法律与伦理问题，人工智能自主作出决策仍存在风险，尤其是在医学等高风险领域，

作者地址：Shuai Ma，香港科技大学，中国香港，shuai.ma@connect.ust.hk；Ying Lei，华东师范大学，中国上海；Xinru Wang，普渡大学，美国西拉法叶；Chengbo Zheng，香港科技大学，中国香港；Chuhan Shi，香港科技大学，中国香港；Ming Yin，普渡大学，美国西拉法叶；Xiaojuan Ma，香港科技大学，中国香港。

允许出于个人或课堂用途免费制作本作品的全部或部分内容的数字副本或纸质副本，但前提是不得出于营利或商业利益制作或分发副本，且副本须附有本声明及首页的完整引文。对于本作品中归其他方而非 ACM 所有的组件，必须尊重其版权。注明出处后允许摘录。如需以其他方式复制、再版、发布到服务器或分发至列表，必须事先获得明确许可和/或支付费用。请通过 permissions@acm.org 申请许可。© 2023 计算机协会。提交给 ACM 的稿件

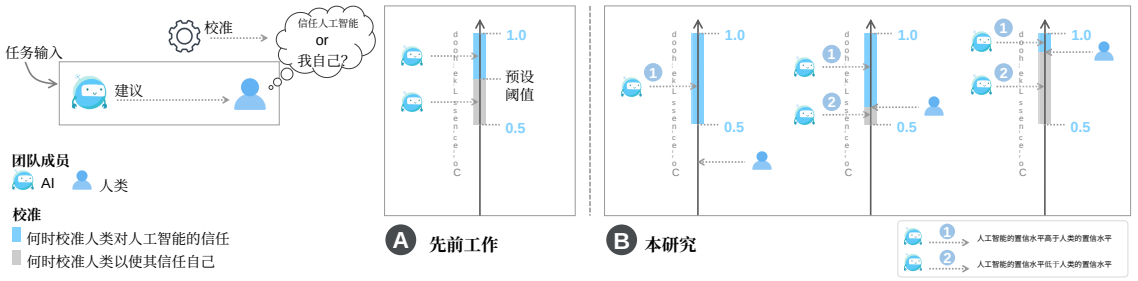

图1. 先前工作与本工作的区别。(A) 在先前工作中，通常使用人工智能校准后的置信度来表示人工智能的正确可能性（置信水平），其取值范围为 0.5 至 1.0。他们根据经验设定的阈值校准人类的信任，即当人工智能的置信度超过该阈值时，引导人类信任人工智能；当人工智能的置信度低于该阈值时，则引导人类不信任人工智能（信任自己）。(B) 在本工作中，除了考虑人工智能的置信水平外，我们还估计人类在每个任务实例中的置信水平。我们提出根据双方相对的置信水平来校准人类的信任，而非仅仅依赖人工智能的置信度是否高于预设阈值。例如，如果人工智能的置信水平高于人类的置信水平，我们就会引导人类信任人工智能；否则，就会引导人类信任自己。

刑事司法等领域 [10, 18, 59]。因此，一种名为人工智能辅助决策 [6, 14, 97, 103] 的范式被提出，并在人机交互和人工智能领域得到了广泛研究。在这一范式中，人工智能通过提供建议发挥辅助作用，而人类决策者可以在最终决策中选择接受或拒绝人工智能的建议。

人工智能辅助决策的一项关键挑战是，人机团队能否实现互补性能，即协作决策结果优于人类或人工智能单独决策的结果 [6, 54, 103]。实现互补性能的关键一步是，人类决策者能够恰当地判断何时考虑人工智能的建议，何时对其持怀疑态度 [14, 81, 103]。由于经过良好校准的人工智能置信度分数能够体现模型实际的正确可能性（CL） [5, 6, 41],，近期有多项研究提出不同的设计，帮助人们依据这一信息对人工智能给予适当信任 [6, 81, 103]。例如，Zhang 等人 [103] 直接向人类决策者显示人工智能的置信度分数。Bansal 等人 [6] 则在人工智能的置信度低于某个阈值时，展示人工智能针对替代预测给出的解释，使人们对人工智能产生怀疑。Rastogi 等人[81] 提出，当人工智能的置信度低于某个阈值时，为人们留出更多决策时间，以减少锚定偏差。然而，这些研究的实证结果充其量也只是喜忧参半 [6, 52, 81, 103]。这可能有两个原因。首先，这些研究假设，人们在了解人工智能的 CL 后，能够恰当地评估自己在某个任务实例中的能力（CL），进而做出合理决策。然而，人们的自信通常校准不佳，无法可靠地反映其 CL [47, 66–68, 99]。其次，这些方法仅依据人工智能的正确可能性来引导人们对人工智能建议的重视程度（如图1（a）所示），却基本忽视了人们在每种情况下的正确可能性。这引出了一个问题：当人工智能的正确可能性较低（较高），但人类的正确可能性更低（更高）时，我们是否仍应鼓励人们怀疑（信任）人工智能？

为了探究这个问题的答案，本文提出了一个框架，旨在根据任务实例层面预测的人类-人工智能正确性可能性（CL），促进人类对人工智能的适当信任和互补团队绩效。在该框架中，如图1（b）所示，我们不再需要仅依据人工智能的置信度是否超过预设阈值来校准人类信任。相反，我们会将人类和人工智能在给定任务实例上的 CL 都纳入考量。为验证所提出框架在促进适当信任和互补性能方面的可行性与有效性，我们的研究分为两个阶段：1）如何 提交给 ACM 的稿件

如何对人类在给定任务上的能力（CL）进行建模？以及，2）如何利用人机协作能力（CL），促进人工智能辅助决策中的适当信任？

在第一 阶段，基于认知科学中的理论——人类通常会采用相似的解决方案来处理相似的问题 [16, 40, 47, 70], ——我们提出根据人们在相似任务中的表现，估计他们在新任务上的 CL。例如，如果某人在相似任务中表现良好，那么她在当前任务上的 CL 也可能较高。然而，通常很难获得足够的决策数据来计算人类在相似任务中的表现。为了解决这一问题，我们提出一种方法：首先近似人类的决策模型（即从任务输入到人类决策的映射），再应用该模型预测人类在相似任务实例中可能作出的决策。在此过程中，我们提出一个问题：**研究问题1：如何有效地近似人类的决策模型？** 为探究答案，我们提出结合数据驱动初始化与交互式修改，推导出每个人可能采用的决策规则。我们还设计了一个名为交互式规则集的界面，供用户修改初始模型，使其更符合自身的决策过程。我们通过初步研究（N=20），验证了所设计界面的适用性，并将其与另一种界面交互式决策树进行了比较。我们将由系统生成并经人类修订的决策规则集视为近似的人类决策模型。对于每个新任务案例，我们从现有任务数据集（用于训练人工智能）中检索最相似的案例，再应用推导出的人类决策者个体模型，获得他们对这些案例可能作出的预测。之后，根据估计的预测结果和真实值，我们可以计算个人可能达到的性能，并进一步利用这些信息估计此人在当前新任务案例上的正确可能性（CL）。结合人类 CL 和人工智能 CL，我们可以识别每个任务实例中谁的能力更强。我们通过一项众包研究（N=30）验证了，与传统的基于 AI 置信度的方法相比，我们的方法在识别互补任务实例（团队中只有一人能够正确完成任务）方面的有效性。

在我们工作的第二 阶段，在获得输入任务案例的人机 CL 估计值后，我们进一步探究如何利用这些信息来促进适当的人类信任，并最终在人工智能辅助决策中实现互补性能。具体而言，当人类的置信水平高于人工智能时，我们尝试降低人类对人工智能的信任；反之则提高这种信任。基于有关人类认知过程的相关文献 [5, 14, 32, 75], ，我们提出三种 CL 利用策略，以显式或隐式地向相关人类决策者传达 双方的置信水平，分别为直接显示、自适应工作流和自适应推荐。围绕这三种 CL 利用策略，我们提出了两个相关研究问题：**研究问题2：不同策略如何影响人类信任适当性和团队表现？** 以及**研究问题3：不同策略如何影响人们在决策过程中的感知和经历？** 通过一项有293名参与者参加的被试间众包实验，我们发现，与基线条件相比，所提出的三种 CL 利用策略均能使用户对人工智能产生更适当的信任，尤其是在人工智能给出错误建议时。这三种策略也提升了团队表现。然而，在大多数主观测量中，不同条件并未导致人类感知或经历出现显著差异。

我们的工作为促进人工智能辅助决策中的适当人类信任提供了新的视角。总之，我们的主要贡献包括：

- 我们提出了一个框架，旨在基于双方的能力，在任务实例层面促进人类对人工智能辅助决策的适当信任。

• 因此，我们设计了一种方法，通过数据驱动初始化和交互式修改来估计人类在新任务实例上的置信水平，并推导出用于估计用户决策模型的决策规则。
• 我们开展了两项初步研究，以验证交互式决策规则创建界面的适当性，并验证人类正确性可能性建模方法的有效性。
• 基于人机 CL 及有关人类认知过程的相关理论，我们提出三种 CL 利用策略，以显性或隐性的方式促进人类对人工智能的适当信任。
• 我们开展了一项用户研究，分析不同 CL 利用策略对用户信任适当性、团队表现和用户体验的影响。根据主要发现，我们提出了相关设计启示，以促进更有效的人机协作决策。

2 相关工作

2.1 人工智能辅助决策中的信任校准

信任校准是指人们对人工智能的信任与人工智能实际能力之间的对应关系 [57]。当信任超出人工智能的能力时，过度信任会导致误用，即人们在不应该信任人工智能时仍然信任它[58, 76]。信任不足是指信任低于人工智能的能力，这会导致弃用，即人们在应该使用人工智能时却没有使用它 [58]。这种有缺陷的人机合作关系可能会造成高昂代价，甚至酿成灾难性后果。成功的决策要求人类针对具体情境校准对人工智能的信任 [4–6, 92, 103]。

校准人类信任的一个关键方法，是向人们传达人工智能的能力（也称为可靠性或可信度） [6, 80, 92, 97]。有多种线索可以反映人工智能的能力，例如人工智能的准确率（包括声明准确率 [82, 102] 和观察准确率 [64, 82, 102]）、解释 [53, 54, 80], 、实际行为/输出[3, 4, 38], 以及置信度 [6, 81, 103], 等。例如，一些研究通过让人们观察人工智能的输出，帮助他们建立关于人工智能错误边界的心理模型 [4]。此外，若干研究认为，如果向人们展示人工智能决策的解释 [6, 80, 92, 97], ，他们就能够判断预测背后的可信度。

最常用的能力指标之一是人工智能的校准置信度分数，因为经过良好校准的置信度能够准确反映人工智能在特定任务实例中的实际正确可能性（CL） [5, 6]。因此，近期许多研究都基于人工智能置信度来校准人类信任。一类研究直接向人们显示校准后的置信度分数。例如，Zhang 等人 [103] 比较了显示与不显示人工智能置信度对人们的信任校准和任务表现的影响。另一类研究则将人工智能置信度融入界面设计。例如，Rastogi 等人 [81] 发现，如果给予人们更长的思考时间，他们便有更多认知资源投入分析性思维，并减少受人工智能锚定效应的影响。因此，他们根据人工智能的置信度为人类分配不同长度的决策时间。此外，Bansal 等人 [6] 开发了一种自适应解释策略：当人工智能置信度低于阈值时，该策略会解释其他预测类别；否则，只解释最高置信度预测。

这些研究存在两个缺陷。一方面，它们假设人们在了解 AI 的 CL 后，能够适当感知自己在某个任务实例中的能力（CL），从而做出合理决策。然而，人们的主观自信通常无法准确代表其实际能力（CL） [47, 66–68]。另一方面，这些方法仅根据 AI 的 CL 来校准人类的信任，而忽略了人类的 CL。例如，现有方法会在人工智能的置信度（CL）较低时使人们怀疑人工智能。但如果人的 CL 更低呢？需要注意的是，人工智能的置信度仅代表一种“可能性”；因此，低置信度的预测仍然可能是正确的，而高置信度的预测

也可能出错。为解决这些问题，我们提出了一种基于人类和人工智能能力（CL）的新方法，用于校准人类的信任。

2.2 人机协作中的心理模型

心理模型是人们用来与周围世界互动的外部现实表征 [46, 74]。在人机协作中，一些研究探讨如何构建人类对人工智能伙伴的心理模型 [4, 38, 62, 75],，以便人类了解是否以及何时应将任务交给人工智能。例如，Gero 等人 [38]发现，在合作博弈场景中，获胜次数较多的人对人工智能代理能力的估计也更准确。Bansal 等人 [6] 帮助人类构建人工智能系统错误边界的心理模型，并发现良好的心理模型有助于提升人类的性能。除了构建人工智能如何运作的心理模型，准确反映人类如何运作的心理模型也至关重要。例如，在人机交互中，一些研究通过建模人们在不同环境下的行为方式，来近似人类的决策策略 [25, 72]。然而，在人工智能辅助决策中，如何利用人类决策方式的模型却鲜少受到关注。本文在实例层面近似人类的决策（心理）模型（即给定一个任务实例，人们会作出怎样的预测），进而基于该模型估计人类对新任务实例的 CL。

构建人类心理模型的一种方法是采用数据驱动方法。例如，在贷款审批任务中，Wang 等人 [95] 利用众包数据，通过神经网络构建通用的人类预测模型。另一种方法是采用基于规则的方法。例如，Bansal 等人 [4] 使用简单规则来构建人类对人工智能错误边界的心理模型，例如“（年龄 = 偏大 & 血压 = 偏高）”。Mozannar 等人 [71]要求人类在完成一组精选示例后，通过编写描述错误区域的规则，将自己对人工智能错误区域的心理模型形式化。尤其是，基于规则的方法具有可解释性优势 [2, 51, 55, 60, 61]。在本研究中，我们提出将数据驱动初始化与交互式规则修改相结合，以推导出个人可能采用的决策心理模型。这种方法有两个优点。首先，它通过少量用户决策数据训练初始模型，节省人们的时间，因此无需从头构建模型。其次，还可以将模型展示给用户，以便用户手动进行交互式改进。

2.3 认知偏差与人类对人工智能的依赖

在人机交互中，由于人们通常倾向于进行系统1思维 [47], ，往往会出现各种认知偏差，包括常见的锚定偏差 [75], 、确认偏误 [73], 、自动化偏差/厌恶 [22],、可得性偏差 [94], 、效度错觉 [88], 等。这些认知偏差可能会对人们对人工智能的信任产生负面影响。例如，在早期观察模型行为后，人们往往会对人工智能的建议产生锚定偏差 [75],，从而过度依赖人工智能的建议。人们也常常会受到人工智能所显示信息的效度错觉影响 [29, 48, 54]。例如，Kaur 等人 [48] 发现，解释的存在可能会使数据科学家误以为模型已准备好部署，因而产生过度自信。Eiband 等人 [29] 发现，即使是无法传递有用信息的安慰剂式解释，也会引发与真实解释相近程度的信任。

为了减少认知偏差对人机协作的不利影响，现有研究提出了一些缓解方法。一种方法是提供干预措施，促使人们投入更多精力进行系统二思维[47]。例如，关于“认知强制”的研究探讨了如何促使人类决策者花更多时间审慎思考问题的方法 [14, 77, 81], ，比如让人类在看到人工智能的建议之前先做出独立预测 [14] ，或采用“慢速算法” [77]。研究发现，这些认知强制功能能够**降低**人们对人工智能的依赖。其他缓解方法还包括让人们主动探索数据 M

[88, 94], 清楚地说明并培训用户如何使用解释/人工智能 [53], 针对未预测的结果提供论据 [15], 监测用户的锚定状态 [28], 展示结果的先验概率 [94], 等等。

在这项工作中，我们“利用”人类的认知偏差，通过将认知偏差融入自适应交互设计，帮助校准人类的信任。具体而言，我们不会盲目地提高或降低人们对人工智能的信任，而是根据双方的置信水平调节人们信任的分布。当人工智能的置信水平较高时，我们利用锚定偏差来提升人们对人工智能的信任；当人的置信水平较高时，我们则采用认知强制，促进不依赖于人类的分析性思维。

3 第一阶段：针对给定任务实例的人类正确可能性建模 E

3.1 人类正确性可能性建模的整体流程

我们在一个典型的人工智能辅助决策场景中研究人类正确性可能性（置信水平，CL）建模方法：对于任务数据集，我们拥有真实标注数据，但对于当前案例则没有。为了在人工智能辅助决策过程中促进更恰当的人类信任，这项工作的第一阶段旨在从任务实例层面估计人类的能力。受认知科学研究启发，该研究指出，人类会通过权衡相似的过往经历来作出决策 [11, 16, 40, 47, 70], 因此，我们提出基于人类在相似任务实例中的过往性能，估计人类在给定任务实例 [68] 上的置信水平。然而，要获取足够的决策数据，以计算人类在相似任务实例中的性能往往很困难，尤其是对于新任务。为解决这一问题，我们设计了一种方法，首先近似人类的决策模型（即建立从任务输入到人类决策的映射）。随后，我们应用该模型预测人类在相似任务实例中可能作出的决策，并将这些决策与真实值进行比较，计算其在这些实例中的潜在准确率。为回答**研究问题1：如何有效地近似人类的决策模型?**，我们提出结合数据驱动初始化和交互式修改，以推导每个人可能采用的决策规则。我们提出的人类能力建模方法包含四个步骤，如图2所示。

- 步骤1：收集用户预测：我们收集人们在少量抽样任务实例上的决策数据。
- 步骤2：生成初始化的决策模型：我们拟合一个经典决策树模型，以推断人类的决策模型，并生成初始决策规则 [37, 90]。
- 步骤3：交互式修改决策模型：由于训练数据有限，生成的初始模型可能无法充分反映人们实际的决策模型。因此，我们设计了一个交互式界面，供用户修改初始模型，使其更符合他们内在的决策过程。

- 步骤4：应用决策模型估计新案例的正确可能性：我们将近似的人类决策模型应用于当前任务案例的邻近案例，并计算人类可能达到的性能。然后，基于距离加权方法（见公式1），我们可以估计人类在当前案例中的置信水平。

在接下来的小节中，我们将介绍任务设置，并通过两项小规模研究详细说明这四个步骤（Study I.1 聚焦于步骤1-3，Study I.2 聚焦于步骤4）。

3.2 任务设置

3.2.1 任务选择。我们选择收入预测作为测试任务，该任务已用于此前多项关于人工智能辅助决策的研究 [39, 43, 86, 103]。在这项任务中，参与者需要根据某人的一些人口统计信息和工作信息，预测其年收入是否超过$50K。任务所用数据来自 UCI机器学习知识库中的 Adult Income 数据集 [50]。该数据集共有 48,842 条受访者记录，提交给 ACM 的稿件

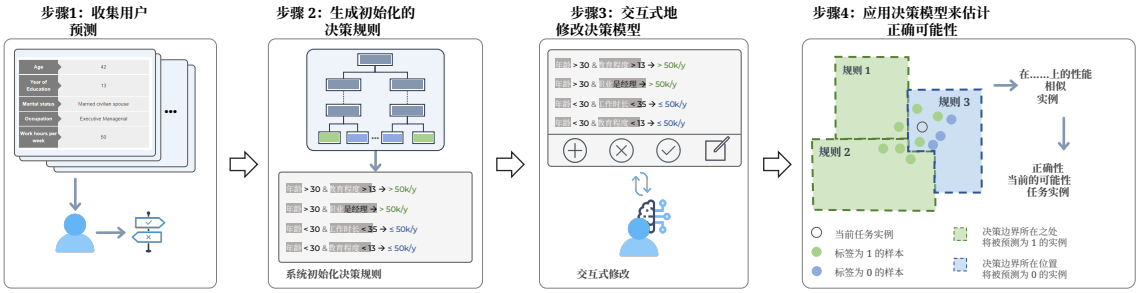

图2。人类能力建模过程。整个过程分为四个步骤。步骤1：收集用户预测。步骤2：生成初始化决策模型。步骤3：交互式修改决策模型。步骤4：应用决策模型估计正确可能性。

每条记录都包含年龄、职业等 14 个属性。这些人的实际年收入均有记录，并被二值化（高于/低于 50K），作为评估参与者预测准确率的真实值。这项任务适合我们的研究，因为它对领域专业知识的要求较低，风险也相对有限，因此适合由非专业参与者完成 [39]。

为确保任务具有适当的复杂度，使普通人能够建立决策模型，我们按照[39, 103], 的做法，根据基于特征置换法 [1]得出的特征重要性值，从 14 个属性中选出最重要的五个，作为呈现给参与者的最终属性。这些属性包括年龄、受教育年限、职业、婚姻状况和每周工作时长。研究表明，对于非专家用户而言，通过体验若干任务样本来构建决策模型时，采用这一数量的特征较为合适（例如，Bansal 等人 [4, 5] 通过一个包含三个特征的任务，建立了用户关于人工智能错误边界的心智模型）。未来的工作可以扩展至模拟人类在更复杂任务中的决策模型。

3.2.2 人工智能模型。与 [39], 相同，我们选择逻辑回归模型（使用*sklearn*的默认设置）作为人工智能模型，协助人类完成所选的收入预测任务中的决策。由于逻辑回归模型直接优化对数损失，因此能够返回校准良好的置信度分数 [79]。模型的校准置信度可以为其预测提供准确的正确概率。例如，如果模型对某个样本做出的预测具有 0.6 的置信度（经过校准），那么该预测有 60% 的概率是正确的；等价地，如果模型对 M 个样本做出的预测均具有 0.6 的置信度，那么其中约有 0.6 * M 个样本会被正确预测。请注意，一些机器学习分类器（例如支持向量机和神经网络）无法直接生成经过校准的置信度分数 [6, 41, 103], ，因此需要进行事后校准（例如 Platt 缩放或保序回归 [41, 79]）。

我们的模型基于原始数据集的 70% 随机划分进行训练，而实验中提供给参与者的预测试验则取自剩余的 30%。对于测试集中的任何新任务案例，我们的人类能力估计方法会从训练集中检索相似案例，以预测人类的置信水平。

3.2.3 任务案例选择。 为用户研究选取的任务案例满足若干标准。首先，为使人类与人工智能的协作设置更适合实现互补性能，这些样本上人类的独立准确率应与人工智能的独立准确率相当 [6, 103]。其次，这些案例应遵循测试集中的数据分布 [97]。第三，这些样本中人工智能的置信度分数应经过良好校准，以反映其实际能力（CL） [97, 103]。

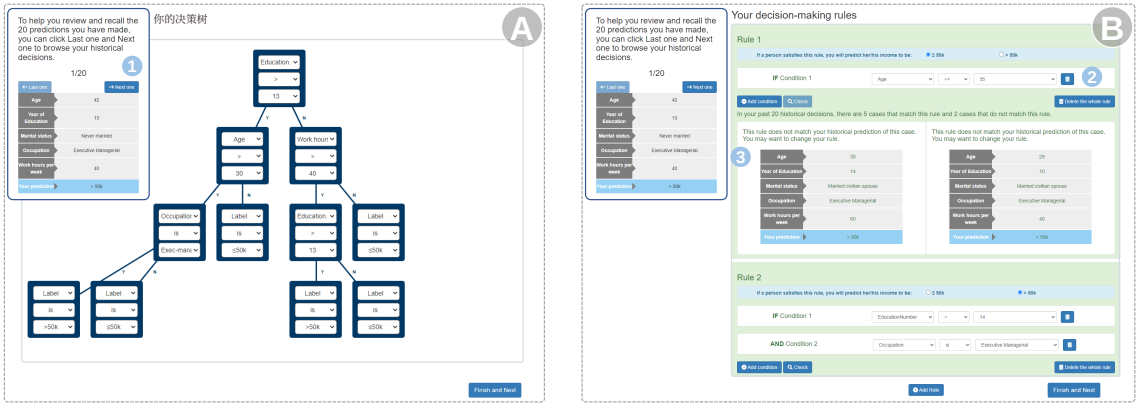

图 3. 两种决策模型创建界面。(A) 交互式决策树。(B) 交互式规则集 .

为使用户研究的时长适中，同时避免参与者感到疲劳，我们选取了 40 个任务案例，并将其分为两批。前 20 个样本用于收集人类的决策数据，并通过计算构建其决策模型。其余 20 个样本用于主要的人工智能辅助决策任务。尽管所有参与者所使用的两批样本均固定不变，但每批样本的呈现顺序是随机的。为了使人工智能的性能能够与人类的独立准确率相比较，我们按照 [6], 的做法，首先开展了一项额外的先导研究，以确定人类在未获得辅助的情况下对随机抽取的 20 个任务实例进行预测的平均准确率。根据研究结果，该准确率约为 70%。随后，我们选取了 40 个任务样本，使人工智能模型在这些样本上的准确率为 70%，且正负标签数量相等，假阳性率和假阳性率也相等（设置与 [6] 类似）。为确保所选样本具有代表性，我们确保这 40 个样本涵盖了每个特征的大多数常见取值。我们还仔细控制了人工智能在这些实例上的置信度，使其与人工智能的实际 CL 相符。具体而言，每批 20 个样本中，一半样本的置信度分数低于 0.7（代表低置信度样本，平均值为 0.6），其中人工智能正确预测了 6 个样本（CL 为 6/10 = 60%）。另一半样本的置信度分数高于 0.7（代表高置信度样本，平均值为 0.8），其中人工智能正确预测了 8 个样本（CL 为 8/10 = 80%）。

人工智能模型和任务案例准备就绪后，我们开展了一项实验室研究，探索适合非专家用户交互式修订其决策模型的界面。请注意，可能有人会问：为什么需要对人类正确性可能性（置信水平，*CL*）进行建模？正如人工智能的置信度可以反映其正确可能性，人们的自信能否代表其正确可能性？我们开展了一项小规模用户研究，发现参与者的主观置信度校准得并不好。也就是说，他们的实际准确率与自我报告的置信度之间在统计上没有相关性，这表明自我报告的置信度并不是可靠的人类置信水平指标。详情请参阅补充材料。

3.3 研究 I.1：比较用户指定其决策模型的界面 s

根据现有研究，规则被认为是近似人类决策过程的一种适当机制 [34, 35, 37]。另一方面，人类往往会基于类似决策树的结构做出判断 [23]。因此，为了比较这两种表示方式的有效性，我们设计了两个界面，用于显示并交互式更新人类的决策模型（即**交互式决策树** 与**交互式规则集**）。提交给 ACM 的稿件

两个界面采用相同的初始化方法：使用前 20 个任务案例中的人类决策数据拟合决策树模型（*sklearn* 的默认设置）。我们选择决策树模型而非黑箱模型，因为即使是不具备机器学习知识的人也能轻松理解它 [23]。

这两个界面如图3所示。交互式决策树 界面（图3 (a)）直接显示后端生成的决策树模型。在该界面中，首先向人类展示一段教程，介绍如何解读和修改决策树（图中未展示）。接着，他们可以浏览前 20 个实例的历史决策数据，以回顾自己的决策依据（图3 (1)）。最后，他们可以添加、删除或修改任意树节点，以反映实际的决策过程。交互式规则集 界面（图3 (b)）展示由决策树转换而来的如果-那么规则集 [61]。使用该界面时，人类同样会先查看教程，接着回顾历史决策数据，最后可以添加、删除或修改任意规则，或修改规则中的条件（图3 (2)）。

3.3.1 研究流程与参与者。我们开展了一项组间研究，从当地研究型大学招募了20名参与者（8名女性、12名男性，平均年龄：27岁），让他们使用分配到的界面构建自己的决策模型（每种条件各10人）。在知情同意后，他们通过教程熟悉收入预测任务。随后，他们在没有人工智能帮助的情况下完成了20项预测任务（第一批）。完成后，他们使用分配到的界面创建自己的决策模型。由于我们主要关注他们对界面的定性感知，因此在研究结束时对他们进行了结束访谈。

3.3.2 结果。根据访谈结果，使用交互式决策树界面的10名参与者中，有7人表示，决策树无法充分呈现他们实际的决策过程。例如，P3（男性，30岁，对人工智能了解甚少）指出：“我实际的决策过程并不是一棵单一的（决策）树。有时，我会把‘年龄’作为首要标准，但有时，我会把‘受教育年限’作为主要因素。”此外，10名参与者中有3人认为决策树在视觉上较为复杂。例如，P9（男性，26岁，对人工智能了解甚少）提到：“我很难在短时间内看懂这棵树。”相比之下，该条件下的参与者认为交互式规则集界面在视觉上更易理解，也更符合他们的决策过程。因此，在最终版本中，我们采用交互式规则集界面，让参与者能够交互式地修改自己的决策模型。

根据参与者的反馈，我们还改进了交互式规则集界面。对于每条规则，我们都提供了一个“检查”按钮，点击后，用户可以查看自己的历史决策中有多少与该规则相冲突，以及该规则是否与其他已创建的规则相冲突（图3 (3)）。

3.3.3 讨论。虽然决策规则集更适合模拟人类决策模型，但也存在一些局限性。首先，有时会出现难以用有限的决策规则涵盖的边缘情况 [35]。对于这些情况，我们现在使用系统初始化模型来进行覆盖。其次，有些用户会基于直觉做出决策，而这无法被表述为一套明确的规则。第三，非专家用户可能很难仅通过体验少量任务样本来制定准确的决策规则。我们将在第 6.6 节中对此进行更详细的讨论。

3.4 Study I.2：我们的人类正确性可能性估计方法的性能测试

基于用户修订的决策模型，我们可以获得其对从训练集中检索出的 N 个相似任务实例的可能预测。通过比较这些可能的预测与已知的真实标签，我们可以计算人类在这些任务实例上的潜在表现，从而得到当前任务实例的估计置信水平（CL）。在本研究中，我们将数量 N 经验性地设为 10，以兼顾充分的相似性与

覆盖率。如果数量设得过大，就会纳入许多不相似的样本；如果设得过小，样本量则不足以获得稳定的准确率。请注意，对于属性不同的其他任务，该数量可能有所不同。我们根据以下公式计算当前任务实例 I_c 上的人类正确性可能性 CL_c。

$$CL_c = \frac{\sum_{i=1}^{N} w^i \cdot IF(\hat{y}^i = y^i, 1, 0) + (1 - w^i) \cdot 50\%}{N}, \quad where\ w^i = \frac{\alpha}{\alpha + d(\mathbf{x}_c, \mathbf{x}_n^i)}. \tag{1}$$

其中，$\hat{y}^i$ 是第 i个邻居实例 I_n^i中人类可能做出的预测，而 y^i 是该实例的真实标签。 $IF(\hat{y}^i = y^i, 1, 0)$表示如果$\hat{y}^i = y^i$，则返回1，否则返回0。 w^i是每个邻居实例的权重， $d(\mathbf{x}_c, \mathbf{x}_n^i)$是当前任务实例 I_c的特征向量$\mathbf{x}_c$与其邻居实例 I_n^i的特征向量$\mathbf{x}_n^i$之间的欧几里得距离。可以看出，权重与距离呈负相关。越相似的邻居实例对性能计算的影响越大。例如，如果人类能够正确预测一个非常近的邻居实例（$d(\mathbf{x}_c, \mathbf{x}_n^i) \to 0$），那么它将为置信水平贡献$1/N$。如果人类能够正确预测一个（极其）遥远的任务实例（$d(\mathbf{x}_c, \mathbf{x}_n^i) \to \infty$），距离因子将折减其贡献，使 w^i 趋近于0（即，它对置信水平的贡献仅为$0.5/N$，相当于随机猜测）。我们根据训练集中任意两个实例之间的欧几里得距离中位数，将参数 α 设为2。尽管其他取值可能更合适，但我们将其留待未来工作探讨。

结合人类置信水平和人工智能置信水平（校准后的 AI 置信度），对于新的任务实例，我们可以估计人机团队中哪一方有更高的正确可能性。接下来，我们从两个目标出发验证方法的有效性。首先，估计的人类置信水平应与实际的人类准确率显著相关。其次，回顾一下，我们对人类能力进行建模的一个关键目的是更好地区分何时应信任人工智能，何时应信任自己。因此，我们聚焦于互补区域，在该区域中，每个案例只有人机团队中的一方能够做出正确预测。如果估计人类在某个案例中的置信水平较高，该案例将标记为“人类更优”；否则，标记为“人工智能更优”。相比之下，在基于 AI 置信度的方法中，与以往研究相同 [103], ，当人工智能的置信度超过设定阈值（0.7）时，我们将该案例视为“人工智能更优”，否则视为“人类更优”。我们将有效性量化为 *recall* 互补案例的比例，即在整个互补区域中，被我们的方法正确预测的互补案例所占的比例。我们没有关注 *precision*，因为当人类和人工智能都可能做出正确或错误的预测时，谁的可能性更高都不会导致显著不同的后果。

*3.4.1 研究流程与参与者。*在与 Study I.1（第 3.3 节）相同的设置下，我们开展了一项众包研究，以比较我们的方法与基于 AI 置信度的方法的有效性。我们从 Prolific 招募了 30 名参与者[1] （18 名女性、11 名男性、1 名非二元性别者，年龄为 21 至 61 岁，平均年龄 35 岁，均居住在美国）。研究流程相同，但我们还要求参与者在制定决策规则后完成剩余的 20 项任务（同样不借助 AI，以便衡量人类独立判断的正确性）。

3.4.2 结果。 我们发现，基于自动生成的人类决策模型，参与者在最后 20 个任务实例中的决策预测准确率为 77.5%。相比之下，基于经人类修订的决策模型，准确率为 80.7%。这表明准确率略有提升，但提升并不显著。我们推测，这是因为默认的决策树模型已经与人类决策过程较为接近，因此参与者可以

[1]www.prolific.co 提交给
ACM 的稿件

仅对初始化的规则进行少量调整。随后，我们计算了估计得到的人类平均 CL 与其在后20项任务中的实际准确率之间的皮尔逊相关。结果显示二者呈显著正相关（r=0.482， p<.01）。此外，我们发现，人机正确性可能性方法平均能够召回 76.4% 的互补区域，而基于 AI 置信度的方法平均能够召回 66.7% 的互补区域。配对 t 检验显示差异显著（p<.05）。结果验证了，与传统的仅基于人工智能置信度的方法相比，我们的方法在互补案例中更能准确估计人机 CL。

3.4.3 讨论。我们注意到，我们的方法高度依赖于近似人类决策模型的准确性。此外，尽管我们的方法在识别互补案例方面优于基于 AI 置信度的方法，但提升幅度并不大。我们推测，这可能是由于人类与人工智能之间的互补性有限，从而影响了我们方法的优势。我们将在最后的讨论部分探讨这一问题。

4 第二阶段：传达人类和人工智能的正确可能性，以促进适当信任

本研究的第二阶段旨在探索如何整合建模得到的人类-人工智能正确性可能性（CL），以增强人工智能辅助决策过程。具体而言，我们提出三种利用CL的不同策略，即直接显示、自适应工作流和自适应推荐。随后，我们通过一项被试间实验，旨在探讨两个研究问题：**研究问题2：不同的CL利用策略如何影响人类信任适当性和团队表现？**以及**RQ3：不同的CL利用策略如何影响人们在决策过程中的感知和用户体验？**

4.1 实验条件与界面设计

为了帮助人们判断何时参考人工智能的建议、何时依靠自己，一种直观的机制是向人类决策者明确地显示人类和人工智能的CL信息。

• **直接显示**：在此条件下，我们直接向人类呈现估计得出的人类和人工智能CL，以及人工智能的建议（图4 C）。具体而言，在实验网站上，系统在个人资料区域（待预测对象的五项属性，图4 A1）旁并列展示估计的人类和人工智能CL（图4 C3）。该区域顶部有一条总结性说明：“根据系统估计，在此任务案例中，人工智能（你）做出正确决策的概率可能高于你（人工智能）。”下方是两个仪表盘图，分别显示人类和人工智能的CL值，随后呈现人工智能的建议（即预测收入）。最后，人们需要输入自己的最终决策（图4 B4）。

在这一条件下，人类用户需要自行决定如何解读置信水平（CL）信息，以及是否信任人工智能。不过，我们承认，对人类和人工智能能力的估计远非完美，而依赖这些信息来评估人工智能的建议可能会造成严重后果，尤其是在高风险领域。例如，在临床决策场景中，如果我们的模型高估了医生的能力，他们可能会对自己的诊断产生虚假的自信。为缓解这一问题，我们基于认知科学理论提出另外两种隐式CL利用策略。一方面，根据决策中的锚定偏差理论[14, 30, 31, 33, 81]，如果人类决策者能够接触到锚点（例如人工智能的观点），他们可能会减少对替代假设的进一步探索，并**增加**人类对人工智能的依赖。另一方面，关于“认知强制”的研究探讨了促使人类决策者花更多时间审慎思考问题的方法

。研究发现，这些认知强制功能能够**降低**人类对人工智能的依赖。基于这些理论依据，我们提出以下条件。

• **自适应工作流**：在此条件下（图 4 D），我们根据估算的人类和人工智能的置信水平，动态调整人类与人工智能决策的顺序。如果预测的人类置信水平高于人工智能，我们的界面会先请人类用户输入初始决策，然后再显示 AI 的建议（图 4 D5）。同样，如果估算出的人工智能置信水平高于人类，界面会直接向人类呈现 AI 建议（不显示图 4 D5）。在这两种情况下，人类都需要在查看人工智能建议后做出最终决策。

另一种隐式运用认知强制来促进适当信任的方法是：如果人们的置信水平高于人工智能，就不向他们显示 AI 建议。然而，这会使人们在这种情况下无法利用人工智能的帮助。一种折衷方案是，当人类的置信水平高于人工智能时，向用户提供人工智能的解释，但不提供预测结果，从而促使他们必须自行做出决策。Garhos 等人 [36]发现，如果只展示人工智能的解释而不提供具体推荐，人们会更多地依据自身知识进行分析性思维。因此，我们提出以下条件。

• **自适应推荐**：在此条件下，人工智能默认显示其预测的解释（由 LIME[85], 这一广泛使用的可解释人工智能方法生成）。我们根据估算的人类与人工智能置信水平的比较，控制人工智能建议的显示。如果人工智能的置信水平较高，我们的界面将显示人工智能的建议（图 4 E6）及其解释（图 4 E7）。如果人类的置信水平较高，我们的界面将不向用户展示人工智能的建议（图 4 E6 中隐藏）。

除上述三种条件外，我们还按照 [6, 81] 纳入了两种基线条件。

• **仅人类**：在此条件下，人类必须在没有任何人工智能辅助的情况下自行做出决策（图 4 A）。
• **人工智能置信度**：在此条件下，向人类展示人工智能的建议及其校准后的置信度
但不提供人类认知负荷信息（图 4 B2）。我们纳入这一基线，是因为这是一种广受认可的
用于校准人们对人工智能辅助决策的信任的设计 [6, 103]。

我们通过先导研究测试了这些界面，以确保参与者能够清晰地遵循工作流程。

4.2 研究设计

我们采用 被试间研究设计，设置了五个条件。该研究已获大学机构审查委员会批准 .

4.2.1 任务与流程。 我们采用了与第一阶段相同的任务（第 3.2 节），即预测某人的年收入是否超过 5 万美元。参与者在研究期间经历了五个阶段（如图 5 所示）：（1）介绍：取得参与者的知情同意后，我们提供了教程演示，帮助他们熟悉任务，并向参与者详细说明个人资料表中每项属性的含义及其取值范围。对于每项属性，我们都展示了一张图表，呈现整个数据集中相应收入的分布，让参与者对薪资情况有基本了解。我们在教程末尾加入了两个注意力检查问题，以帮助筛除未认真阅读介绍的参与者。教程结束后，我们向参与者提供了两个带有真实值的训练示例。（2）第一批 20 个任务：接下来，参与者独立完成前 20 个任务案例（不显示人工智能建议或真实值信息）。（3）交互式决策规则创建：参与者进入决策规则创建页面（详情见第 3.3 节）。（4）第二批 20 个任务：完成自定义决策规则后，参与者继续完成最后 20 个任务案例，此时，提交给 ACM 的稿件

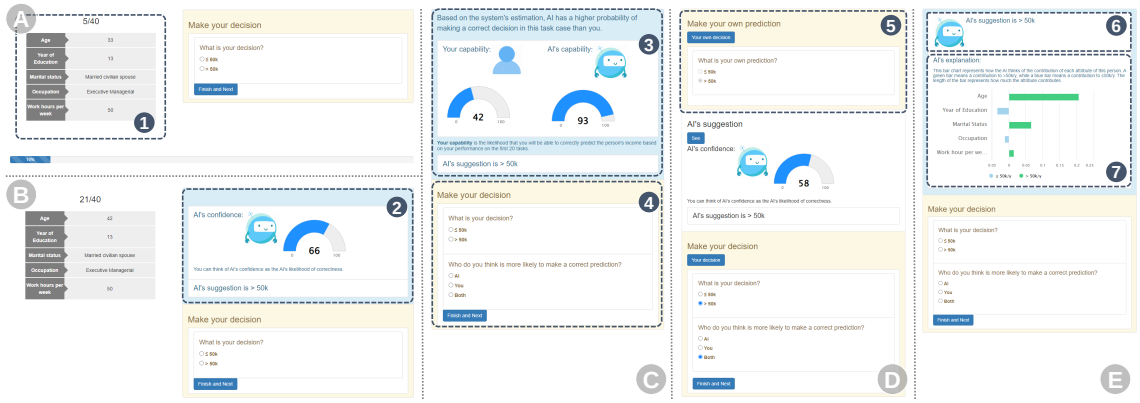

图4. 不同条件下的界面。(A) **仅人类**。(B) **人工智能置信度**。(C) **直接显示**。(D) **自适应工作流**。(E) **自适应推荐**。所有条件下的界面布局相似：左侧为个人资料区域，右侧为决策区域。为节省空间，图4(C)、(D)、(E)中不再重复绘制个人资料区域。

他们将在人工智能的辅助下完成任务（“仅人类”条件除外）。根据分配的条件，参与者会看到不同的界面（如图 4 所示）。（5）退出问卷：最后，我们请参与者填写退出问卷，收集他们的基本人口统计信息、主观测量结果，以及对决策过程中自身感知的开放式反馈，相关内容将在第 4.3 节中介绍。

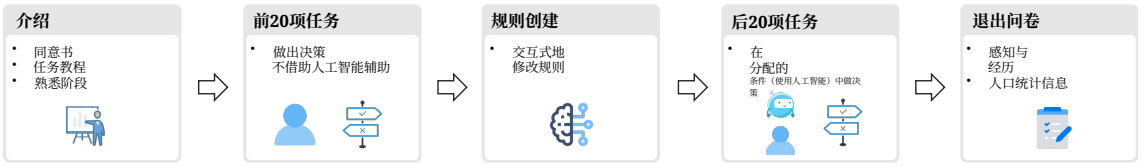

图5. 实验流程。参与者在整个研究中依次经历五个阶段。

4.2.2 参与者。我们从 Prolific[1] 招募了 300 名参与者（每种条件 60 人）。为确保回答质量，所有参与者均符合以下标准：(1) 居住在美国（因为任务是预测美国成年人的收入）；(2) 以往提交内容的获批率至少为 99%；(3) 英语为第一语言；(4) 拥有学士学位或更高学历；(5) 使用台式电脑参与实验。本研究采用被试间设计，因此不允许重复参与。我们共收到 293 份完整提交。根据注意力检查问题筛选后，我们获得了 289 份有效回答（仅人类：59 人，人工智能置信度：59 人，直接显示：59 人，自适应工作流：56 人，自适应推荐：56 人）。最终参与者中，有 174 人自报为男性，110 人为女性，5 人为非二元性别。18 至 29 岁的参与者共 77 人，30 至 39 岁的有 116 人，40 至 49 岁的有 45 人，50 至 59 岁的有 28 人，60 岁及以上的有 23 人。参与者还评估了自己对人工智能的了解程度：40 人表示不了解，205 人了解人工智能的基本概念，43 人使用过人工智能算法，1 人是人工智能专家。为激励参与者认真完成任务，除基础报酬外，如果整体准确率超过 80%，我们还会发放 0.50 美元奖金。整个研究持续约 20 分钟。参与者的平均时薪约为 9.34 美元。

4.3 评估指标

4.3.1 研究问题*2*的测量指标。 我们通过两项主要测量指标，考察不同条件对人类信任适当性和人机团队表现的影响。(1) 人机一致性 [6, 103]：参与者的最终决策与人工智能建议一致的任务所占比例，无论该建议正确与否。(2) 团队表现 [6, 81, 97, 103]：最终决策的准确率。我们还收集了参与者的感知正确可能性（置信水平）；在每个任务实例中，我们都会询问参与者认为哪一方（人类、人工智能或双方）的置信水平更高。

4.3.2 RQ3 的测量。 在此，我们关注参与者在不同条件下的经历和感知。具体而言，参考相关研究并加以调整，我们在退出问卷中以七点李克特量表题目考察以下主观测量（1：非常不同意，7：非常同意）：(1) 对人工智能的信任 [14, 39]；(2) 对决策过程的信心 [52, 82]；(3) 感知系统复杂度 [14]；(4) 心理需求 [14, 39, 42, 52]；(5)感知自主性 [44]；(6) 满意度 [39]；(7) 未来使用意愿 [12]；(8) 对人机 *CL* 估计的信任；(9) 对人机协作负荷估计的感知有用性 [56]；(10) 感知到的在决定何时信任人工智能方面的帮助性 [56]）；以及 (11)对其 *CL* 估计的接受度。除了这些问题，我们还向参与者提出了开放式问题，询问他们如何使用和看待所传达的人机 CL，以及不同界面设计如何影响他们的决策过程。详细问题见补充材料。

4.3.3 分析方法。 我们针对上述指标进行了混合方法分析。对于研究问题2中的客观数据以及研究问题3中参与者的主观数据，由于大部分数据不符合正态分布，我们采用了非参数检验。具体而言，对于两两比较，我们根据样本是否来自同一组参与者，采用曼-惠特尼U检验或威尔科克森符号秩检验。对于两个以上参与者组之间的分析，我们采用克鲁斯卡尔-沃利斯检验，并进行邦费罗尼校正的事后分析。定性分析方面，两位作者通过归纳式主题分析对开放式反馈进行了编码 [45]。经过多轮讨论和协调，最终确定了各项主题，并从原文中选取具体示例，以便在本文中展示。

5 结果

5.1 置信水平利用策略对团队表现和人类对人工智能的信任的影响

我们将结果分为两部分。第一部分分析参与者的总体团队表现以及他们对人工智能的信任。第二部分则深入探究数据，并根据不同情境（例如，不同的人类－人工智能置信水平）分析结果。为便于阅读，所有结果均以“研究发现”的形式呈现。

第 1 部分

发现 1：在提出的三种条件下，人类信任人工智能的趋势与估计的人机CL的趋势一致。由于我们三种设计的基本目的，是让人们在人机团队中更多地依赖置信水平较高的成员，因此我们想了解这些方法是否让人们在人工智能的置信水平较高时更加信任人工智能，而在其他情况下较少信任人工智能。图6展示了三种条件下不同人机CL对应的人机一致性。结果显示，在三种条件下，当人工智能的置信水平高于人的置信水平时，人们与人工智能的一致性均显著更高（所有三种条件下均为 p<.001）。

发现2：所提出的三种置信水平利用条件实现了互补性能，而人工智能置信度条件则没有。图7（a）展示了所有条件下的总体团队绩效（即人类最终决策的准确率）。直接显示（M=0.758， SD=0.078）、自适应工作流



图6. 不同人类与人工智能正确可能性下的人机一致性（含均值和95%置信区间）。我们可以看到，当人工智能的置信水平高于人类时，参与者在最终决策中更倾向于认同人工智能。

（M=0.767， SD=0.085）、自适应推荐（M=0.754， SD=0.072）的团队表现均优于单独使用人工智能（0.7）和仅人类（M=0.721， SD=0.113）。然而，人工智能置信度（M=0.720， SD=0.099）的团队表现并未优于仅人类。从图7（b）可以看出，尽管人工智能置信度使人类在人工智能建议正确时更倾向于认同人工智能，但也使人类更倾向于认同人工智能的错误建议。这一发现与先前研究一致：显示人工智能置信度未必能提升团队表现[81, 103]。与人工智能置信度相比，所提出的置信水平利用方法使团队表现略有提高，最高可显著提高（直接显示： p=.078；自适应工作流： p=.027；自适应推荐： p=.078）。

发现3：在所提出的三种置信水平利用条件下，当人工智能的建议错误时，人类对人工智能的信任更加适当。如图7(b)（红色柱形）所示，人工智能置信度条件下的人机一致性显著高于自适应工作流条件（p=.017）和自适应推荐条件（p<.001），且略高于直接显示条件（p=.07）。请注意，人工智能出错时，与人工智能的一致性越低越好。这些结果表明，在我们提出的置信水平利用条件下，人类对人工智能的**过度信任更少**。通过参与者的定性反馈，我们发现，我们的三种设计促使人们在人工智能的建议错误时（人工智能的置信水平通常低于人类）更多依靠自己的思考。具体而言，在直接显示条件下，显示较高的人类置信水平使他们对自己的判断更有信心。例如，P26（49岁，女性，了解人工智能基础知识）表示：“有时人工智能的观点与我的不同。当我看到自己的（置信水平）值高于人工智能时，这让我更加确信自己的观点。”而在自适应工作流条件下，大多数情况下，参与者先自行判断，之后即使人工智能的建议与自己的判断相反，也不会改变决定。例如，P16（31岁，男性，了解或使用过人工智能算法）提到：“在人工智能给出建议之前，我已经仔细思考过了，所以我会坚持自己的看法。”在自适应推荐条件下也能发现类似现象。P4（40岁，女性，不了解人工智能）说：“人工智能没有告诉我任何答案，我只能根据自己的想法做决定。”

然而，当人工智能的建议正确时（图7 (b) 中的红色柱形），我们没有观察到三种拟议条件与人工智能置信度基准之间的人机一致性存在显著差异。我们推测，这可能是因为人类和人工智能都能做出正确预测的任务实例高度重叠。从仅人类条件可以看出，在人工智能给出正确建议的情况下，即使人们没有得到人工智能的任何帮助，其性能也达到了 80%（在 80% 的情况下与人工智能意见一致）。这表明，在这类情境中，人类与人工智能的互补性相对较弱，因此改进空间有限。

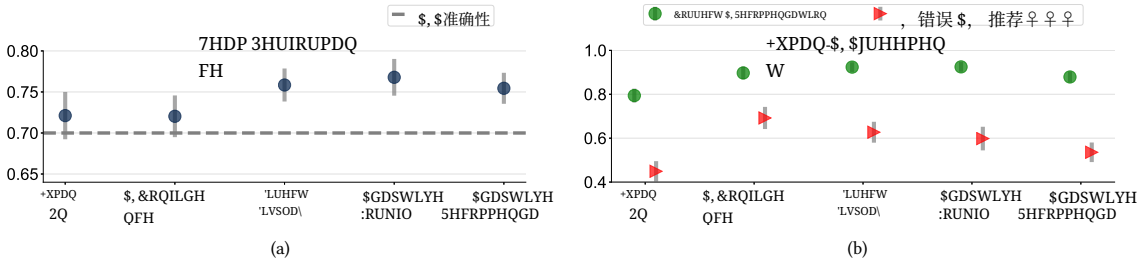

图7. 不同条件下的总体团队绩效和信任适当性（含平均准确率和95%置信区间）。(A) 五种条件下的总体团队绩效。与人工智能准确率（0.7）和仅人类 （0.72）相比，提出的三种沟通策略实现了互补性能。(B) 人类的信任适当性，以人工智能给出正确建议和错误建议时的人机一致性表示。

发现4：人类的置信水平较高时，团队表现更好。在提出的三种置信水平利用条件中，我们根据人类与人工智能的置信水平，向参与者展示不同的信息或调整决策工作流程。因此，我们希望分析在不同置信水平情境和人工智能建议正确性不同的情况下，团队表现有何差异。总体而言，如图8(a)所示，人类的置信水平较高时，团队表现显著优于人工智能的置信水平较高时（所有条件下均为 p<.001）。

具体而言，如图8 (b)所示，(1) 当 AI 的 CL 较高且人工智能的建议正确时，三种条件下的团队表现没有显著差异。(2) 当 AI 的 CL 较高且人工智能的建议错误时，三种条件下的团队表现没有显著差异。(3) 当人的 CL 较高且人工智能的建议正确时，与自适应推荐相比，直接显示 和自适应工作流 条件下的团队表现显著更高（分别为p<.05和 p<.01）。这可能是因为在自适应推荐条件下，当人的 CL 较高时，人工智能的建议不会显示，因此参与者无法获得人工智能正确建议的帮助。(4) 当人的 CL 较高且人工智能的建议错误时，直接显示条件下的团队表现略低于自适应推荐，差异达到边缘显著水平（p<.1），且显著低于自适应推荐（p<.05）。

此外，正如预期，当人工智能的建议错误时，人的置信水平较高时的团队表现显著优于人工智能的置信水平较高时的团队表现（所有条件下均为 p<.001）。但令我们惊讶的是，当人工智能的建议正确时，人的置信水平较高时的团队表现显著优于人工智能的置信水平较高时的团队表现：直接显示 (p<.05)；自适应工作流 (p<.01)；自适应推荐 (p=.058，呈边缘显著）。可能的原因是，当人工智能的置信水平较高时，由于人的能力不如人工智能，人们有时会听从自己错误的判断，因此对人工智能信任不足。

发现5：当人工智能的置信水平与人工智能建议的正确性相“矛盾”时，直接显示 和 自适应工作流 条件的表现更好。仅利用人工智能置信度来校准人类的信任是不够的。具体来说，即使人工智能的置信度高于某个阈值（我们使用了 0.7[97, 103]），人工智能仍可能输出错误预测（称为高且错误区域）。有时，即使人工智能的置信度较低，它也能给出正确建议（称为低且正确区域）。我们将这两种情况称为冲突区域。我们认为，仅显示人工智能的置信度不足以让人们识别这些情况。

总体而言，如图 9 (a) 所示，在冲突区域中，人工智能置信度 条件下的团队表现显著低于直接显示 (p=.002)和自适应工作流 (p=.006)。人工智能置信度与提交给 ACM 的稿件之间未发现显著差异

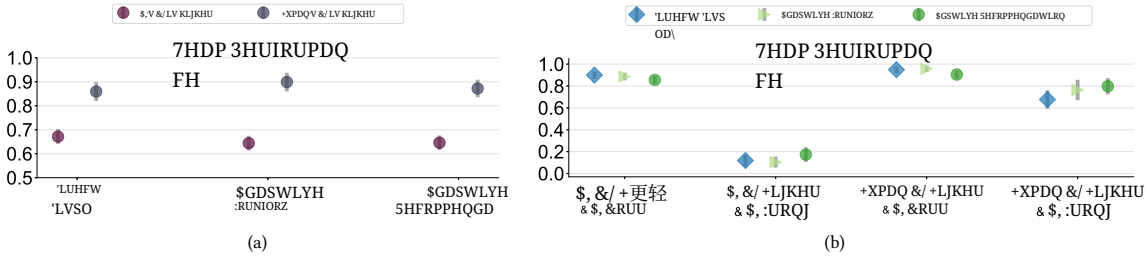

图 8. 不同 CL 情境下的团队表现（含平均准确率和 95%置信区间）。(A) 三种条件下不同人机 CL 情境中的团队表现。(B) 我们结合 AI 建议的正确性与人机 CL，详细分析三种条件在不同情境下的团队表现。

以及自适应推荐。具体而言，如图 9 (b) 所示，在低置信度且正确区域中，直接显示、自适应工作流 和人工智能置信度之间未发现显著差异。可能的原因是提升空间有限（表现已超过 90%）。我们注意到，自适应推荐的表现显著低于人工智能置信度 (p=.032)，这可能是因为在自适应推荐 条件下，人们经常看不到人工智能建议，因此逐渐形成了不依赖人工智能建议的独立思考模式，这可能导致信任不足。在高置信度且错误区域中，人工智能置信度 条件下的团队表现显著低于直接显示 (p=.005)、自适应工作流 (p=.022)，且略低于自适应推荐 (p=.056)。

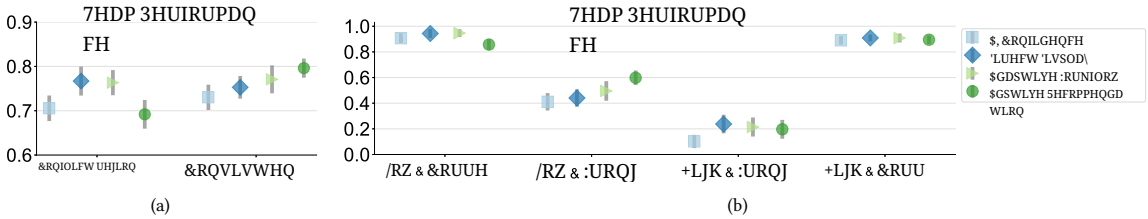

图 9。（A）当人工智能的置信水平与其给出的建议的正确性相冲突（记为冲突区域）或一致（记为一致区域）时的团队表现（平均准确率和 95%置信区间）。（B）具体而言，我们将冲突区域 分为（1）低置信度且正确 （人工智能的置信度低于阈值，但建议正确）和（2）高置信度且错误 （人工智能的置信度高于阈值，但建议错误）；并将一致区域 分为（1）低置信度且错误 （人工智能的置信度低于阈值，且建议错误）和（2）高置信度且正确 （人工智能的置信度高于阈值，且建议正确）。

发现 6：当人工智能的置信水平与其建议的正确性“一致”时，自适应工作流 和自适应推荐条件的效果更好。另一类任务实例称为一致区域，包括（1）低置信度且错误 （人工智能的置信度低于阈值，且给出的建议错误）和（2）高置信度且正确 （人工智能的置信度高于阈值，且给出的建议正确）。

总体而言，如图 9 (a) 所示，人工智能置信度 条件下的团队表现略低于自适应工作流 （p=.074），且显著低于自适应推荐 （p<.001）。但人工智能置信度 与直接显示之间没有显著差异。接下来，我们深入探究这两个子区域。在

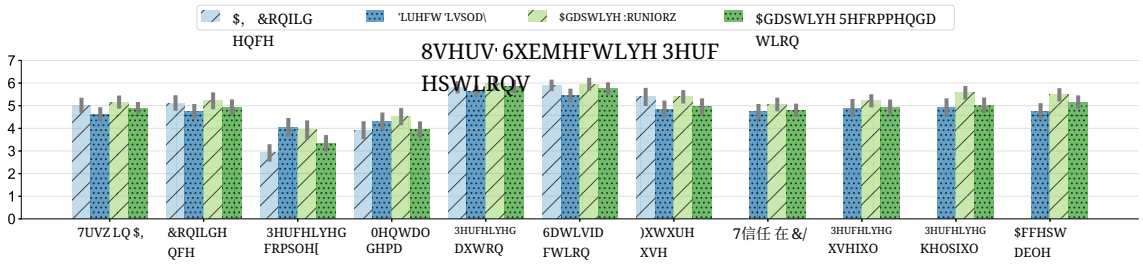

图10. 参与者在退出问卷中的主观评分（含平均值和95%置信区间）。

低置信度且错误区域中，人工智能置信度 条件下的团队表现略低于自适应工作流（p=.088），且显著低于自适应推荐 （p<.001）。但人工智能置信度 与直接显示之间没有显著差异。在高置信度且正确区域中，人工智能置信度 基线与所提出的三种置信水平利用条件之间均无显著差异。

除上述结果外，我们还发现：（1）提出的三种条件有效地向人类传达了估计的 CL信息；（2）参与者在置信水平一致任务实例中表现更好（即他们感知的人机 CL与系统传达的人机 CL一致）。详细结果见补充材料。

总结。总体而言，我们提出的三种置信水平利用方法提升了人类对人工智能的适当信任（尤其是减少了人类的过度信任，且没有导致信任不足），从而带来了更好的团队表现。此外，提出的三种置信水平利用策略能够有效地向人类传达系统估计的人机 CL。此外，当人工智能的置信度与其正确性相矛盾时，我们的方法也优于人工智能置信度 基线。我们还注意到设计中的一些不足，并将在后文讨论。

5.2 置信水平利用策略对人类感知与体验的影响。

为回答 RQ3，我们通过退出问卷，使用七点李克特量表（1：非常不同意，7：非常同意）分析参与者在不同条件下的主观感知。图10展示了结果。

感知系统复杂度。总体而言，在四种条件下，参与者感知到的系统复杂度相对较低至中等。克鲁斯卡尔-沃利斯检验显示，不同条件之间存在显著差异（χ^2=19.223，p<.001）。事后分析表明，与人工智能置信度相比，参与者认为直接显示 （p<.001）和自适应工作流 （p<.01）中的系统复杂度显著更高，这或许是因为这两种条件展示了更多信息，并要求采用更复杂的工作流。

心理需求。总体而言，参与者对于四种条件下的决策过程是否需要较高的心理投入持中立态度。不同条件之间不存在统计学上的显著差异。不过，我们观察到一种趋势：自适应工作流会让参与者感到心理需求较高。

感知到的帮助性：判断何时信任人工智能。总体而言，参与者认为估计的人机 CL 有助于做出信任选择。克鲁斯卡尔-沃利斯检验显示，不同条件之间存在显著差异（χ^2=7.039， p<.05）。事后分析表明，自适应工作流条件下的参与者认为，估计的人机 CL 的帮助性略高于直接显示（p=.056）和自适应推荐（p=.073）。

对其置信水平估计的接受度。总体而言，参与者认为对其置信水平的估计可以接受。克鲁斯卡尔-沃利斯检验显示，不同条件之间存在显著差异（χ^2=9.162， p<.01）。事后分析显示，提交给 ACM 的稿件

表1. 参与者对开放式问题的定性反馈。（请注意，我们排除了仅表达正面反馈、但未提供具体内容的回答，例如“有帮助”“喜欢”等，因此参与者人数总和可能不到 100%。）

主题	定义和示例	#参与者
参与者如何看待估计的人机 CL（以直接显示的形式）？		
质疑正确性似然度	一些参与者认为，自己的能力并不能被轻易且准确地估算出来。“这只是猜测。人工智能并不了解我，所以我觉得它不太可靠。”(P41)	14（23%）
确认感	如果人类和人工智能的观点一致，且人的 CL 较高，就会让参与者更有信心。“这让我对我们意见一致的那些题目很有信心”（P3）	8（13.5%）
认识到人工智能的缺陷	当人们看到人工智能较低的置信水平时，他们意识到人工智能并非总是值得信任。“这让我意识到，人工智能在这项任务中也可能出错。我比人工智能做得更好！”（P5）	6（10%）
降低置信度	有时，当显示的自身置信水平较低时，人们会变得不那么自信。“在某些情况下，看到系统显示我的能力估计值较低时，我开始对自己的能力产生了一些怀疑。（*P14*）”	5（8%）
参与者如何利用显示的人机 CL 做出决策（在直接显示中）？		
依赖较高的一方	他们首先查看了两个置信水平的值。如果人工智能的置信水平更高，他们就会遵循人工智能的建议——否则，他们会自行做出决定。“这两张图表（显示人机 *CL*）让我知道什么时候该凭直觉做决定，什么时候该依赖人工智能。”(P12)	19（32%）
直接忽略置信水平	一些参与者只相信自己，完全忽略置信水平。“我更相信自己的能力，对显示的数值并不太在意”(P23)	15（25%）
反思	一些参与者在看到估计的置信水平后，反思了自己的决策。“‘看到自己的分数没有我想象的那么高后，它帮助我重新审视了自己的决定。”(P12)	10（17%）
在意见不一致的情况下参考它	一些参与者会先自行做出决定。如果人工智能的意见与他们一致，他们就会忽略置信水平。否则，他们主要听从置信水平较高的一方。“如果我发现自己和人工智能的判断一致，我就会直接做出决定。如果人工智能与我的判断不一致，我会比较双方的能力，并选择能力更强的一方来遵循。”（P49）	9（15%）
在不确定的案例	在人们没有把握的情况下，他们会参考置信水平。而当他们有把握时，他们会忽略它。“它确实帮助我做出了一些自己拿不准的决定。不过，我不会在我觉得自己完全正确时考虑它。”(P39)	6（10%）
自适应工作流如何影响参与者的决策过程?		
投入更多认知资源	当人们被要求先做出决策时，他们会投入更多认知资源来完成任务本身，从而避免受到人工智能判断的影响。“如果人工智能没有先给出建议，我会更加关注自己认为正确的内容。如果它给出了建议，那我或多或少就会听从人工智能的判断。”（*P42*）	30（53%）
影响不大	一些参与者将人工智能视为复核或第二意见。“这对我的影响不大。无论人工智能是否允许我先做决定，我始终把人工智能当作第二意见，再做判断。”（P54）	26（46%）
自适应人工智能推荐如何影响参与者的决策过程?		
独立思考	不展示人工智能的具体推荐，促使人们独立思考。“当我看到人工智能的建议时，我更难相信自己，反而会发现自己默认听从人工智能。”(P33)	35 (62%)
影响甚微	这并未影响他们的决策，因为他们往往依靠自己。“这并未影响我的决策。我仍在思考每张信息卡片所呈现的内容，这些内容是”(P44)	19（33%）

与直接显示条件相比，自适应工作流中的参与者认为对其置信水平的估计明显更能被接受（p<.01）。

然而，就**对人工智能的信任、对决策过程的信心、感知自主性、满意度、未来使用意愿、对人机协作负荷估计的信任**以及**对人机协作负荷估计的感知有用性**而言，不同条件之间没有显著差异。

总而言之，除了感知系统复杂度外，我们提出的三种条件并未导致参与者在其他方面产生显著不同的感知。在我们提出的三种 CL 利用方法中，自适应工作流似乎让人们对其有用性和接受度的评价更高。这也与自适应工作流在所有条件中取得最高团队表现的结果相一致。不过，我们也注意到，这些益处伴随着复杂度和心理需求的增加。因此，建议未来通过有人类参与的实证研究，探索效果与用户体验之间的权衡。

5.3 Qual性分析：参与者如何感知、使用人机 C，以及受到哪些影响 L.

为了更好地理解不同 CL利用条件的影响，我们在退出问卷中设置了开放式问题，询问参与者如何感知和使用人机 CL（在直接显示中），以及自适应决策过程（在自适应工作流 和自适应推荐中）如何影响他们。按照归纳式主题分析流程 [45]（第 4.3.3 节），两位作者编制了编码手册（表1）。我们重点介绍以下发现，以解释前文提及的定量结果。

一些参与者对显示的 CL信息表示怀疑，并在决策过程中忽略了这些信息。在直接显示条件下，23% 的参与者对显示的 CL 表示怀疑，25% 的参与者在决策过程中忽略了 CL。这也与研究结果相呼应：尽管人工智能的 CL 较高，人机一致性仍未达到 100%（图6）；这或许可以解释，为什么在人工智能的 CL 较高且其建议正确时，团队表现仍未达到 100%（图8 (b)）。

大多数参与者会参考显示的置信水平或基于置信水平的自适应机制，或受到其影响。 在直接显示 条件下，74%的参与者会参考置信水平；在自适应工作流 和自适应推荐条件下，分别有53%和62%的参与者受到自适应过程的影响。这与研究结果一致：提出的三种方法都能有效影响人们对人工智能的采纳（图6），并促进人类形成适当信任（图7）。

在自适应工作流 和自适应推荐条件下，参与者被迫独立思考。在自适应工作流 和自适应推荐条件下，当人类的置信水平较高时，大多数参与者受到自适应过程的影响，开始独立思考。这解释了为什么我们的方法有助于在人工智能出错时减少过度信任（图7（b））。具体而言，在自适应推荐 条件下，62%的参与者在看不到人工智能建议时会独立思考，33%的参与者表示自己始终独立思考。这表明参与者形成了不依赖人工智能的行为模式，因此当人工智能的置信水平较高时，人机一致性最低（图6）。这也解释了为什么在低置信度且正确区域中，自适应推荐 的团队表现最低（图 9）。其他发现请参阅我们的编码手册（表1）。

6 讨论

通过探究三个研究问题，我们的研究展现了对人类的置信水平进行建模并加以传达，以促进人类在人工智能辅助决策中形成适当信任的潜力。基于主要发现，我们讨论了改善人机团队决策的若干关键问题，以及本研究的局限性。

6.1 人类对自信的感知及对人工智能置信度的理解

保持恰当的自信，对于人类建立对人工智能的适当信任至关重要。 证据表明，人们对自己的置信度会显著影响他们是否采纳人工智能的建议 [19, 93]。然而，无论是对于提交给 ACM 的稿件，个人对自身能力的置信度都可能与其实际能力不符 [47, 66–68, 99]

专家和普通人，从而导致过度自信（或信心不足） [66, 67, 92, 99]。根据我们的结果，在人工智能置信度条件下，当人工智能给出正确建议（且置信度较高）时，一些参与者仍然遵循了自己的错误判断。这是因为人类难以保持“校准”后的自信 [68],，因而忽视了人工智能的建议。我们认为，如果人们能够准确感知自身能力（例如正确可能性），并据此校准自信，人类与人工智能的协作将更加成功。本文提出的人类正确可能性建模与沟通方法，是朝这一目标迈出的第一步。我们希望本研究能够启发研究者借助以人为中心的计算建模和界面设计，探索有效方法来调整人们对人工智能以及对自身的置信度。

人类对概率的理解会影响信任校准的效果。人工智能的置信度和人的 CL 都是数值概率。然而，先前研究表明，人类难以依据数字（例如置信度、准确率）来协调与人工智能的协作，尤其是在时间紧迫场景中认知资源有限的情况下 [9, 13, 54, 89]。此外，人们，尤其是不擅长运用数学思维的人，往往难以轻松理解概率值实际代表什么 [21, 78, 84]。这或许是向人类显示人工智能置信度分数不足以校准其信任的原因之一，我们的研究和现有研究均揭示了这一点 [81, 103]。因此，仅依靠人们做出理性依赖选择并进行协作（例如由谁先做出预测）可能颇具挑战。我们的研究提出，将计算概率估计／比较任务交由系统处理，并通过自动调整决策过程／界面来校准人类信任。这有助于抵消潜在的人类认知偏差 [8, 94] ，并避免让人们直接处理概率。未来研究可以探索另外两个方向。其一是在保留适当人类自主性的同时，设计更有效的算法在环任务协调方法（例如学习委托 [65]）。其二是设计界面来提升人们对概率的理解，例如添加关于概率与频率的简明教程 [69],，以更易理解的方式呈现概率 [54], 等。

6.2 超越信任校准，实现互补性能

发挥知识互补性有助于提升团队表现。 我们的研究发现，尽管所提出的三种条件下的团队表现均优于仅*AI* 和仅人类，但提升幅度并不“显著”（约为 3-4%）。其中一个关键原因是，人类与人工智能知识空间之间的**互补区域/范围** 相对较小。仅人类 条件下的性能分析也体现了这一点：只有少数情况下，人机团队中仅有一名成员能够正确处理问题，因此仅通过校准人类信任，很难实现显著的互补性能。相比之下，在 Bansal 等人的研究 [6] 中，人类在人工智能无法给出正确建议时的独立准确率，甚至高于人工智能能够给出正确建议时的独立准确率。因此，正如 [6, 103], 所指出的，为了确保实现互补性能，除了校准人们对人工智能的信任之外，还需要利用人类与人工智能智能的互补性来实现最优结果 [5, 100], ，或许可以通过训练能够补足人类知识与错误区域的人工智能来实现。

对人类能力进行建模，可以为更精细的设计提供支持。除了近似估计人们的置信水平（CL）外，我们提出的建模方法还能够估计人们的预测。我们认为，这些信息很有价值，因为它们可以帮助我们提前预测人类是否会与人工智能作出一致判断。我们可以将这些信息与人机 CL 结合起来，采用更复杂的策略，帮助人们作出更好的决策。例如，如果预测人类和人工智能的判断会一致，且双方的 CL 都不高，那么双方很可能都会作出错误预测。在这种情况下，人工智能可以重点鼓励人们进行分析性思考，而不是肯定他们的决定。

例如，正如 [6], 所指出的，人工智能有时可以扮演唱反调者的角色，质疑人类的判断。未来的研究可以探索更先进的方法，充分利用本文提出的人类决策模型和人机协作能力。

6.3 适当置信水平沟通方法的设计

恰当地传达置信水平信息，与正确地对其进行建模同样重要。由于人类才是最终的决策者，人们在决策过程中如何接收、感知和使用这些信息，对决策结果至关重要。尽管本文提出的三种置信水平利用机制提升了人们对人工智能的适当信任，但我们发现，仍有一些参与者固守着自己的错误判断。从开放式反馈中，我们发现，一些参与者认为系统无法轻松、可靠地估算他们的能力，这无疑阻碍了我们方法发挥潜力。尽管我们已向他们提供必要的信息，但对一些参与者而言，背后的过程仍像一个黑箱。因此，我们建议提供更详尽、易于理解的指南，向人类介绍并解释置信水平建模背后的原理，以增进他们的理解和接受度。

其他可能有效的置信水平沟通设计。除了本文提出的自适应设计外，也可以调整其他类型的信息，以促进人类信任的校准。以往研究表明，无论人工智能的解释是否正确，提供人工智能的解释都可能提升人们对人工智能的信任 [6, 54, 80, 97]。因此，我们可以设计一种自适应解释策略，仅当人工智能的置信水平高于人类时，才提供人工智能的解释。此外，一些研究发现，置信度框架可能会影响人们的风险感知 [20]。因此，我们可以在人工智能置信水平较高时使用积极的语气进行描述，例如“人工智能有 75% 的概率做出正确预测”；否则，则使用不确定的语气，例如“人工智能有 25% 的概率做出错误预测”（尽管这与前者等价）。此外，近期研究强调了人们在决策过程中处理信息时所涉及的双重认知过程 [13, 17, 47, 98], : 系统1以快速、自动的方式处理刺激，若使用不当，可能导致认知偏差 [47], ; 而系统2则进行审慎的分析性思维。运用这一理论的一般方法是：当人们需要更多依赖自己的判断时（例如他们的置信水平高于人工智能时），界面应促使人们进行系统二思维。借助这些理论视角，我们可以设计出更有效的置信水平信息使用方式。

6.4 当前 CL建模与利用方法的陷阱

界面的潜在副作用。尽管我们提出的设计在促进人类的适当信任方面效果显著，我们仍建议设计者谨慎对待这些设计可能带来的陷阱。例如，在自适应推荐中，我们发现参与者似乎形成了一种对人工智能的**怀疑**模式，因为他们往往看不到 AI 建议；这可能会阻碍他们在人工智能给出正确建议时利用其帮助。此外，自适应工作流可能会导致人类产生确认偏见 [73]。例如，人们做出初步判断后，如果发现人工智能“确认”了这一判断，就会非常确信答案正确。但事实上，有时人类和人工智能会同时做出错误判断。因此，我们建议，除了以现有认知理论为设计依据之外，还必须通过实证研究验证设计可能产生的影响及不利后果。

人机 CL 的缺点及其引发的伦理和问责问题。 围绕人机 CL 存在两个问题。首先，即使人工智能的置信度经过校准，且人的 CL 得到了准确建模，仍会出现不一致的情况：对于某个特定任务实例，人机团队中 CL 较高的成员做出了错误预测，而 CL 较低的成员却做出了正确预测。这可能导致人类对人工智能产生不当信任。由于 CL 提交给 ACM 的稿件

不过，这只是对正确可能性的概率估计，因此置信水平与正确性不一致不可避免。其次，我们对人的 CL 的估计可能并不准确，而人工智能模型的置信度有时也可能校准不佳。因此，不恰当地使用人机 CL 可能会导致负面结果。例如，如果我们错误地将某人的 CL 估计得较低，我们的方法可能会导致该人接受人工智能给出的错误建议，而实际上她／他本可以独立做出正确决策。这可能造成严重后果，在高风险场景中甚至会带来危险。因此，要让人机 CL 发挥积极作用，在部署前确认人机正确性可能性的可靠性至关重要。此外，在适当情况下传达 CL 背后的不确定性，提醒人类决策者注意此类信息带来的风险，可能也有益处。缓解 CL 估计值负面影响的另一种可能方式，是避免营造信息已获确认且精确无误的印象，例如避免使用具体百分比或带有判断意味的词语 [83]。研究人员可以改为隐含地传达这些信息，通过类似于我们所提出的自适应方法的设计，将其融入决策过程。

6.5 关于我们的方法和结果的可推广性

将我们的方法和结果推广到不同的任务领域和受试者群体时，应当谨慎。首先，我们选择了基于规则的方法，帮助用户理解和修改自动生成的决策模型。然而，这种方法可能不适用于更复杂的决策任务，例如涉及文本或图像数据的任务。因此，我们需要根据任务和数据的具体特点，设计恰当的知识表示和建模算法。例如，在文本情感分析任务中，用户可以指定关键词或示例句来表示其决策模型 [52]。其次，我们的研究对象是参与低风险决策任务的非专家用户。尽管这为探索人类在人工智能辅助决策中的信任适当性提供了合适的试验环境 [39, 103], ，我们仍提醒读者，谨慎将我们的结果推广到其他群体或任务。例如，目前尚不清楚，当我们设计的界面应用于高风险任务时（决策者可能采用不同的认知路径 [91]），我们的结果是否仍然成立。此外，我们的方法能否充分建模领域专家的能力，也尚不得而知。尽管如此，我们相信，基于双方能力校准人类信任的框架，可以推广到需要协作的不同人工智能辅助决策场景。未来的工作可以将我们的人类正确可能性建模及沟通方法应用于涉及不同利益相关者的其他决策任务 [60]。

6.6 局限性与未来工作

我们提出的方法和实验设置存在若干局限性。首先，我们使用决策规则来近似人类的决策模型。然而，规则只能提供一个通用模型，无法涵盖所有边缘情况。未来的解决方案可以考虑融入“行为测试”方法 [7] ，系统可使用测试用例“检验”用户的能力，就像测试软件或自然语言处理模型一样 [87]。其次，由于本文侧重研究我们的方法对人类的影响，因此我们没有在后20项任务中更新人类的决策模型。我们假设，在没有正确性反馈（例如无法获取真实值）的情况下，用户的决策模型在短期内相对固定，这一假设对于我们的实验而言是合理的。然而，在现实世界的决策中，用户的决策模型可能会随着用户与人工智能服务的互动以及接触更多任务实例而发生变化 [63], ，因此静态模型并不足够。未来，我们计划探索如何在长期的人工智能辅助决策过程中维护一个实时更新的用户决策模型。第三，我们通过人机一致性来衡量人类对人工智能的信任。尽管这种方法已被广泛使用 [6, 97, 103], ，但其一个明显的缺点是，当人们的最终判断是

与人工智能的建议一致，但我们无法区分这是因为他们听从了人工智能的建议，还是因为他们自己的决策与人工智能的建议一致。未来的研究可以探索更合适的测量方法。

7 结论

人类对人工智能的适当信任是人工智能辅助决策中的一项根本挑战，而我们的工作有助于根据人类和人工智能双方的能力校准人类的信任。我们的研究分为连续的两个阶段。第一阶段，我们探讨如何对人类在给定任务实例上的能力（正确可能性）进行建模。我们提出了一种人类决策模型近似方法，并配有交互式决策规则修改界面。第二阶段，我们探讨如何利用人机协作能力，促进人工智能辅助决策中的适当信任。基于人们认知过程的理论，我们提出了三种置信水平利用方法，并研究其对人类信任适当性、任务表现和用户体验的影响。研究结果凸显了所提出的人类正确可能性建模与利用方法的有效性：与传统的基于人工智能置信度的方法相比，该方法能够促进人类形成更适当的信任。基于主要发现得出的实践启示，我们希望这项工作能够通过考量双方的能力信息，推动更适当的人机决策。

参考文献

[1] André Altmann、Laura Toloşi、Oliver Sander 和 Thomas Lengauer。2010年。置换重要性：一种经过修正的特征重要性度量。《生物信息学》26卷，10期（2010年），1340–1347页。[2] Alejandro Barredo Arrieta、Natalia Díaz-Rodríguez、Javier Del Ser、Adrien Bennetot、Siham Tabik、Alberto Barbado、Salvador García、Sergio Gil-López、Daniel Molina、Richard Benjamins 等。2020年。可解释人工智能（XAI）：面向负责任人工智能的概念、分类体系、机遇与挑战。《信息融合》58（2020年），82–115页。[3] Gagan Bansal、Besmira Nushi、Ece Kamar、Eric Horvitz 和 Daniel S Weld。2020年。优化人工智能以促进团队协作。*arXiv*预印本 *arXiv:2004.13102*（2020年）。[4] Gagan Bansal、Besmira Nushi、Ece Kamar、Walter S Lasecki、Daniel S Weld 和 Eric Horvitz。2019年。超越准确率：心理模型在人机团队表现中的作用。载于*AAAI* 人类计算与众包会议论文集，第7卷。2–11页。[5] Gagan Bansal、Besmira Nushi、Ece Kamar、Daniel S Weld、Walter S Lasecki 和 Eric Horvitz。2019年。人机团队的进展：理解并解决性能与兼容性权衡问题。载于*AAAI* 人工智能大会论文集，第33卷。2429–2437页。[6] Gagan Bansal、Tongshuang Wu、Joyce Zhou、Raymond Fok、Besmira Nushi、Ece Kamar、Marco Tulio Ribeiro 和 Daniel Weld。2021年。整体是否大于部分之和？人工智能解释对互补团队绩效的影响。载于*2021*年 *CHI* 人机交互大会论文集。1–16页。[7] Boris Beizer。1995年。黑盒测试：软件和系统功能测试技术。John Wiley & Sons, Inc.。[8] Astrid Bertrand、Rafik Belloum、James R Eagan 和 Winston Maxwell。2022年。认知偏差如何影响可解释人工智能辅助决策：系统综述。载于*2022*年 *AAAI/ACM* 人工智能、伦理与社会会议论文集。78–91页。[9] Donald M Berwick、Harvey V Fineberg 和 Milton C Weinstein。1981年。当医生遇上数字。《美国医学杂志》71卷，6期（1981年），991–998页。[10] Reuben Binns、Max Van Kleek、Michael Veale、Ulrik Lyngs、Jun Zhao 和 Nigel Shadbolt。2018年。“这等于把一个人简化成一个百分比”：人们对算法决策中公正性的感知。载于*2018*年 *CHI* 人机交互大会论文集。1–14页。[11] Aaron M Bornstein、Mel W Khaw、Daphna Shohamy 和 Nathaniel D Daw。2017年。对过去选择的回忆会使人类的奖励决策产生偏差。《自然·通讯》8卷，1期（2017年），1–9页。[12] John Brooke 等。1996年。系统可用性量表——一种快速而粗略的可用性量表。《工业中的可用性评估》189、194（1996年），4–7页。[13] Zana Buçinca、Phoebe Lin、Krzysztof Z Gajos 和 Elena L Glassman。2020年。代理任务和主观测量可能会误导对可解释人工智能系统的评估。载于第*25*届国际智能用户界面会议论文集。454–464页。[14] Zana Buçinca、Maja Barbara Malaya 和 Krzysztof Z Gajos。2021年。是信任还是思考：认知强制功能可以减少人工智能辅助决策中对人工智能的过度依赖。《*ACM* 人机交互会议论文集》5卷，CSCW1（2021年），1–21页。[15] Adrian Bussone、Simone Stumpf 和 Dympna O’Sullivan。2015年。解释在临床决策支持系统中的信任与依赖所发挥的作用。载于*2015*年国际医疗信息学会议。电气电子工程师学会，160–169页。[16] Pietro Carlo Cacciabue、Françoise Decortis、Bartolome Drozdowicz、Michel Masson 和 J-P Nordvik。1992年。COSIMO：用于复杂工业设施事故管理中人类决策与行为的认知仿真模型。《*IEEE* 系统、人类与控制论汇刊》22卷，5期（1992年），1058–1074页。M

[17] John T Cacioppo 和 Richard E Petty。1984。说服的精细加工可能性模型。《ACR 北美进展》（1984）。[18] Carrie J Cai、Samantha Winter、David Steiner、Lauren Wilcox 和 Michael Terry。2019。“你好，人工智能”：揭示医疗从业者在人机协作决策中的入职引导需求。《ACM 人机交互会议论文集》3，CSCW（2019），1–24。[19] Leah Chong、Guanglu Zhang、Kosa Goucher-Lambert、Kenneth Kotovsky 和 Jonathan Cagan。2022。人类对人工智能和自身的信心：信心的演变及其对采纳人工智能建议的影响。《计算机与人类行为》127（2022），107018。[20] George I Christopoulos、Philippe N Tobler、Peter Bossaerts、Raymond J Dolan 和 Wolfram Schultz。2009。参与风险决策的价值、风险与风险厌恶的神经相关机制。《神经科学杂志》29，40（2009），12574–12583。[21] Leda Cosmides 和 John Tooby。1996。归根结底，人类是否擅长直觉统计？重新审视不确定性判断文献中的一些结论。《认知》58，1（1996），1–73。[22] Mary Cummings。2004。智能、时间关键型决策支持系统中的自动化偏差。载于《AIAA第一届智能系统技术会议》。6313。[23] Marc Damez、Thanh Ha Dang、Christophe Marsala 和 Bernadette Bouchon-Meunier。2005。基于人机交互的用户建模模糊决策树。载于《第五届人类系统学习国际会议（ICHSL）论文集》，第5卷。287–302。[24] Jeffrey Dastin。2018。亚马逊弃用显示出对女性存在偏见的秘密人工智能招聘工具。载于《数据与分析的伦理》。Auerbach Publications，296–299。[25] Qi Deng、Jiao Wang 和 Dirk Soffker。2018。基于改进隐马尔可夫模型方法的人类驾驶员行为预测。载于《2018年IEEE智能车辆研讨会（IV）》。IEEE，2066–2071。[26] Bella M DePaulo、Kelly Charlton、Harris Cooper、James J Lindsay 和 Laura Muhlenbruck。1997。欺骗检测中的准确性与信心相关性。《人格与社会心理学评论》1，4（1997），346–357。[27] Steven E Dilsizian 和 Eliot L Siegel。2014。医学人工智能与心脏成像：利用大数据和先进计算提供个性化医疗诊断与治疗。《当代心脏病学报告》16，1（2014），1–8。[28] Jessica Maria Echterhoff、Matin Yarmand 和 Julian McAuley。2022。人工智能辅助决策：捕捉并平衡序贯决策任务中的锚定偏差。载于《CHI人机交互大会》。1–9。[29] Malin Eiband、Daniel Buschek、Alexander Kremer 和 Heinrich Hussmann。2019。安慰剂式解释对智能系统信任的影响。载于《2019年CHI人机交互大会扩展摘要集》。1–6。[30] Nicholas Epley 和 Thomas Gilovich。2001。让调整回归锚定与调整启发法：对自生成锚点和实验者提供锚点的差异化加工。《心理科学》12，5（2001），391–396。[31] Nicholas Epley 和 Thomas Gilovich。2006。锚定与调整启发法：为何调整幅度不足。《心理科学》17，4（2006），311–318。[32] Riccardo Fogliato、Shreya Chappidi、Matthew Lungren、Paul Fisher、Diane Wilson、Michael Fitzke、Mark Parkinson、Eric Horvitz、Kori Inkpen 和 Besmira Nushi。2022。谁先开始？人机工作流程对临床成像决策的影响。载于《2022年ACM公平、问责与透明大会》。1362–1374。[33] Adrian Furnham 和 Hua Chu Boo。2011。锚定效应的文献综述。《社会经济学杂志》40，1（2011），35–42。[34] Johannes Fürnkranz、Dragan Gamberger 和 Nada Lavrač。2012。《规则学习基础》。施普林格科学与商业媒体。[35] Johannes Fürnkranz、Tomáš Kliegr 和 Heiko Paulheim。2020。认知偏好与基于规则模型的合理性。《机器学习》109，4（2020），853–898。[36] Krzysztof Z Gajos 和 Lena Mamykina。2022。人们会在认知层面参与人工智能互动吗？人工智能辅助对附带学习的影响。载于《第27届国际智能用户界面会议》。794–806。[37] Efstathios D Gennatas、Jerome H Friedman、Lyle H Ungar、Romain Pirracchio、Eric Eaton、Lara G Reichmann、Yannet Interian、José Marcio Luna、Charles B Simone、Andrew Auerbach 等。2020。专家增强机器学习。《美国国家科学院院刊》117，9（2020），4571–4577。[38] Katy Ilonka Gero、Zahra Ashktorab、Casey Dugan、Qian Pan、James Johnson、Werner Geyer、Maria Ruiz、Sarah Miller、David R Millen、Murray Campbell 等。2020。合作博弈情境下对人工智能代理的心理模型。载于《2020年CHI人机交互大会论文集》。1–12。[39] Bhavya Ghai、Q Vera Liao、Yunfeng Zhang、Rachel Bellamy 和 Klaus Mueller。2021。可解释主动学习（XAL）：将面向机器教师的人工智能解释作为界面。《ACM 人机交互会议论文集》4，CSCW3（2021），1–28。[40] E Bruce Goldstein。2014。《认知心理学：连接心智、研究与日常经验》。圣智学习。[41] Chuan Guo、Geoff Pleiss、Yu Sun 和 Kilian Q Weinberger。2017。现代神经网络的校准。载于《国际机器学习会议》。PMLR，1321–1330。[42] Sandra G Hart。2006。NASA任务负荷指数（NASA-TLX）：20年之后。载于《人因与工效学学会年会论文集》，第50卷。Sage Publications，Sage CA：洛杉矶，904–908。[43] Peter Hase 和 Mohit Bansal。2020。评估可解释人工智能：哪些算法解释能帮助用户预测模型行为？arXiv预印本 arXiv:2005.01831（2020）。[44] Joo-Wha Hong 和 Dmitri Williams。2019。人机交互中的种族主义、责任与自主性：对人工智能代理的感知进行测试。《计算机与人类行为》100（2019），79–84。[45] Hsiu-Fang Hsieh 和 Sarah E Shannon。2005。定性内容分析的三种方法。《定性健康研究》15，9（2005），1277–1288。

[46] Philip Nicholas Johnson-Laird。1983。《心理模型：迈向语言、推理与意识的认知科学》。第6卷。哈佛大学出版社。[47] Daniel Kahneman。2011。《思考，快与慢》。麦克米伦。[48] Harmanpreet Kaur、Harsha Nori、Samuel Jenkins、Rich Caruana、Hanna Wallach 和 Jennifer Wortman Vaughan。2020。解读可解释性：了解数据科学家如何使用机器学习可解释性工具。载于《2020年CHI人机交互大会论文集》。1–14。[49] Amir E Khandani、Adlar J Kim 和 Andrew W Lo。2010。基于机器学习算法的消费信贷风险模型。《银行与金融杂志》34，11（2010），2767–2787。[50] Ronny Kohavi 和 Barry Becker。1996。Adult Income 数据集（UCI机器学习知识库）。https://archive.ics.uci.edu/ml/datasets/Adult/。[51] Todd Kulesza、Simone Stumpf、Margaret Burnett、Sherry Yang、Irwin Kwan 和 Weng-Keen Wong。2013。太多、太少，还是恰到好处？解释影响终端用户心理模型的方式。载于《2013年IEEE可视化语言与以人为中心的计算研讨会》。IEEE，3–10。[52] Vivian Lai、Samuel Carton、Rajat Bhatnagar、Q Vera Liao、Yunfeng Zhang 和 Chenhao Tan。2022。通过条件式委托实现人机协作：内容审核案例研究。载于《CHI人机交互大会》。1–18。[53] Vivian Lai、Han Liu 和 Chenhao Tan。2020。“为什么‘芝加哥’具有欺骗性？”迈向面向人类的模型驱动教程。载于《2020年CHI人机交互大会论文集》。1–13。[54] Vivian Lai 和 Chenhao Tan。2019。关于人类在获得机器学习模型解释与预测后的预测：欺骗检测案例研究。载于《公平、问责与透明大会论文集》。29–38。[55] Himabindu Lakkaraju、Stephen H Bach 和 Jure Leskovec。2016。可解释决策集：描述与预测的联合框架。载于《第22届ACM SIGKDD知识发现与数据挖掘国际会议论文集》。1675–1684。[56] Bettina Laugwitz、Theo Held 和 Martin Schrepp。2008。用户体验问卷的构建与评估。载于《奥地利人机交互与可用性工程组研讨会》。施普林格，63–76。[57] John D Lee 和 Neville Moray。1994。信任、自信与操作员对自动化的适应。《国际人机交互研究杂志》40，1（1994），153–184。[58] John D Lee 和 Katrina A See。2004。对自动化的信任：以恰当依赖为目标进行设计。《人因》46，1（2004），50–80。[59] Min Hun Lee、Daniel P Siewiorek、Asim Smailagic、Alexandre Bernardino 和 Sergi Bermúdez i Badia。2021。用于康复评估临床决策的人机协作方法。载于《2021年CHI人机交互大会论文集》。1–14。[60] Q Vera Liao 和 Kush R Varshney。2021。以人为中心的可解释人工智能（XAI）：从算法到用户体验。arXiv预印本 arXiv:2110.10790（2021）。[61] Brian Y Lim、Anind K Dey 和 Daniel Avrahami。2009。为何及何时解释能提升情境感知智能系统的可理解性。载于《SIGCHI计算机系统人因大会论文集》。2119–2128。[62] Shuai Ma、Mingfei Sun 和 Xiaojuan Ma。 [不详]。机器人学习投入度的自适应表达建模及其对人类教师的影响探索。《ACM计算机与人类交互汇刊》（[不详]）。[63] Shuai Ma、Zijun Wei、Feng Tian、Xiangmin Fan、Jianming Zhang、Xiaohui Shen、Zhe Lin、Jin Huang、Radomír Měch、Dimitris Samaras 等。2019。SmartEye：通过整合用户偏好与深度视图提议网络辅助即时拍照。载于《2019年CHI人机交互大会论文集》。1–12。[64] Shuai Ma、Taichang Zhou、Fei Nie 和 Xiaojuan Ma。2022。Glancee：帮助教师掌握学生在同步在线课堂中学习状态的自适应系统。载于《CHI人机交互大会》。1–25。[65] David Madras、Toni Pitassi 和 Richard Zemel。2018。负责任地进行预测：通过学习委托提升公平性和准确率。《神经信息处理系统进展》31（2018）。[66] Ashley ND Meyer、Velma L Payne、Derek W Meeks、Radha Rao 和 Hardeep Singh。2013。医生的诊断准确率、信心与资源请求：一项情景短文研究。《美国医学会杂志·内科学》173，21（2013），1952–1958。[67] Deborah J Miller、Elliot S Spengler 和 Paul M Spengler。2015。临床决策中的信心与判断准确性荟萃分析。《咨询心理学杂志》62，4（2015），553。[68] Don A Moore。2020。《信心十足：如何明智地校准决策》。哈珀柯林斯。[69] Don A Moore、Samuel A Swift、Angela Minster、Barbara Mellers、Lyle Ungar、Philip Tetlock、Heather HJ Yang 和 Elizabeth R Tenney。2017。多年地缘政治预测竞赛中的信心校准。《管理科学》63，11（2017），3552–3565。[70] Robert S Moyer 和 Richard H Bayer。1976。心理比较与符号距离效应。《认知心理学》8，2（1976），228–246。[71] Hussein Mozannar、Arvind Satyanarayan 和 David Sontag。2022。通过示例教导人类何时将决策委托给分类器。载于《AAAI人工智能大会论文集》，第36卷。5323–5331。[72] Andrew Y Ng、Stuart Russell 等。2000。逆向强化学习算法。载于《国际机器学习会议》，第1卷。2。[73] Raymond S Nickerson。1998。确认偏误：一种以多种形式普遍存在的现象。《普通心理学评论》2，2（1998），175–220。[74] Donald A Norman。2014。关于心理模型的一些观察。心理学出版社。[75] Mahsan Nourani、Chiradeep Roy、Jeremy E Block、Donald R Honeycutt、Tahrima Rahman、Eric Ragan 和 Vibhav Gogate。2021。锚定偏差影响可解释人工智能系统中的心理模型形成与用户依赖。载于《第26届智能用户界面国际会议》。340–350。M

[76] Raja Parasuraman 和 Victor Riley。1997年。人类与自动化：使用、误用、弃用与滥用。《人因》39，2（1997年），230–253。[77] Joon Sung Park、Rick Barber、Alex Kirlik 和 Karrie Karahalios。2019年。慢速算法提升了用户对算法准确性的评估。《ACM 人机交互会议论文集》3，CSCW（2019年），1–15。[78] Ellen Peters、Daniel Västfjäll、Paul Slovic、CK Mertz、Ketti Mazzocco 和 Stephan Dickert。2006年。数理能力与决策。《心理科学》17，5（2006年），407–413。[79] John Platt 等。1999年。支持向量机的概率输出及其与正则化似然方法的比较。《大间隔分类器进展》10，3（1999年），61–74。[80] Forough Poursabzi-Sangdeh、Daniel G Goldstein、Jake M Hofman、Jennifer Wortman Wortman Vaughan 和 Hanna Wallach。2021年。操控和测量模型可解释性。载于《2021年 CHI 人机交互大会论文集》。1–52。[81] Charvi Rastogi、Yunfeng Zhang、Dennis Wei、Kush R Varshney、Amit Dhurandhar 和 Richard Tomsett。2020年。快思与慢想：认知偏差在人工智能辅助决策中的作用。arXiv 预印本 arXiv:2010.07938（2020年）。[82] Amy Rechkemmer 和 Ming Yin。2022年。当信心遇上准确率：探究多种绩效指标对机器学习模型信任度的影响。载于《CHI 人机交互大会论文集》。1–14。[83] Robert M Reyes、William C Thompson 和 Gordon H Bower。1980年。论据可得性差异导致的判断偏差。《人格与社会心理学杂志》39，1（1980年），2。[84] Valerie F Reyna 和 Charles J Brainerd。2008年。风险与概率判断中的数理能力、比率偏差和分母忽视。《学习与个体差异》18，1（2008年），89–107。[85] Marco Tulio Ribeiro、Sameer Singh 和 Carlos Guestrin。2016年。“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。[86] Marco Tulio Ribeiro、Sameer Singh 和 Carlos Guestrin。2018年。Anchors：高精确率的模型无关解释。载于《AAAI 人工智能大会论文集》，第32卷。[87] Marco Tulio Ribeiro、Tongshuang Wu、Carlos Guestrin 和 Sameer Singh。2020年。超越准确率：使用 CheckList 对自然语言处理模型进行行为测试。arXiv 预印本 arXiv:2005.04118（2020年）。[88] Auste Simkute、Ewa Luger、Mike Evans 和 Rhianne Jones。2020年。算法系统阴影下的专家：探索决策情境中的可理解性。载于《2020年 ACM交互系统设计会议伴随出版物》。263–268。[89] Paul Slovic 和 Ellen Peters。2006年。风险感知与情感。《心理科学当前方向》15，6（2006年），322–325。[90] Yan-Yan Song 和 LU Ying。2015年。决策树方法：分类与预测应用。《上海精神病学档案》27，2（2015年），130。[91] Harini Suresh、Steven R Gomez、Kevin K Nam 和 Arvind Satyanarayan。2021年。超越专业知识与角色：用于描述可解释机器学习利益相关者及其需求的框架。载于《2021年 CHI 人机交互大会论文集》。1–16。[92] Amy Turner、Meena Kaushik、Mu-Ti Huang 和 Srikar Varanasi。 [无日期。]。人工智能辅助决策中的信任校准。（[无日期。]）。[93] Kailas Vodrahalli、Roxana Daneshjou、Tobias Gerstenberg 和 James Zou。2022年。如果建议来自人工智能，人类会更加信任吗？人机交互分析。载于《2022年 AAAI/ACM 人工智能、伦理与社会会议论文集》。763–777。[94] Danding Wang、Qian Yang、Ashraf Abdul 和 Brian Y Lim。2019年。设计理论驱动、以用户为中心的可解释人工智能。载于《2019年 CHI 人机交互大会论文集》。1–15。[95] Xinru Wang、Zhuoran Lu 和 Ming Yin。2022年。你会接受人工智能建议吗？预测人工智能辅助决策中的人类行为。载于《2022年 ACM网络会议论文集》。1697–1708。[96] Xinxi Wang 和 Ye Wang。2014年。利用深度学习改进基于内容的音乐推荐与混合音乐推荐。载于《第22届 ACM多媒体国际会议论文集》。627–636。[97] Xinru Wang 和 Ming Yin。2021年。解释有帮助吗？人工智能辅助决策中解释效果的比较研究。载于《第26届智能用户界面国际会议》。318–328。[98] Peter C Wason 和 J St BT Evans。1974年。推理中的双重过程？《认知》3，2（1974年），141–154。[99] Nathan Weber 和 Neil Brewer。2004年。绝对与相对人脸识别判断中的信心—准确性校准。《实验心理学杂志：应用》10，3（2004年），156。[100] Bryan Wilder、Eric Horvitz 和 Ece Kamar。2020年。学习如何与人类互补。arXiv 预印本 arXiv:2005.00582（2020年）。[101] Yi Yang、Wei Qian 和 Hui Zou。2018年。通过梯度树提升 Tweedie 复合泊松模型预测保险费。《商业与经济统计杂志》36，3（2018年），456–470。[102] Ming Yin、Jennifer Wortman Vaughan 和 Hanna Wallach。2019年。理解准确性对机器学习模型信任度的影响。载于《2019年 CHI 人机交互大会论文集》。1–12。[103] Yunfeng Zhang、Q Vera Liao 和 Rachel KE Bellamy。2020年。信心与解释对人工智能辅助决策中的准确性和信任校准的影响。载于《2020年公平、问责与透明会议论文集》。295–305。