以双向理解促进人智协同
——以人机合作决策为例

马 帅[1] 麻晓娟[2] 石楚涵[3] 等
[1]芬兰人工智能中心及阿尔托大学
[2]香港科技大学
[3]东南大学

人机合作决策面临的挑战

人智协同（human-AI collaboration）指人类与人工智能系统（或模型，简称AI）共同合作完成任务或解决复杂问题的过程。该过程强调通过结合人类智能与机器智能的互补优势来提升整体效能。人机合作决策（human-AI decision-making），通常也称为AI辅助决策（AI-assisted decision-making）[1]，是人智协同领域的一个重要研究方向。随着AI技术的发展，AI模型越来越多地被应用于各类决策场景中，如医学诊断、刑事司法、招聘、金融投资、科学探索等。这一领域在过去几年中得到了学术界和工业界的广泛关注。

人机合作决策的一个典型范式是：针对同一决策任务，人工智能模型或系统作为辅助者提供建议，而人类则作为最终决策者，决定是否采纳AI的建议或在其基础上形成自己的判断。通常，决策的责任和后果由人类决策者承担[2]。

人机合作决策的主要目标之一是实现互补性能（complementary performance），即人与AI组成的团队在决策表现上优于人类或AI单独完成任务时的表现[3]。然而，实现这一目标面临诸多挑战，如合理的任务分工和人对AI建立适当的信任。已有的实证研究指出，人在决策过程中可能会对AI过度信赖（over-trust / reliance），如采纳了AI的错误建议；或信任不足（under-trust / reliance），如忽视了AI的正确建议[4]。这些问题的核心原因之一在于人类与AI之间缺乏有效的相互理解，主要体现在以下四个方面：

1. 决策者对AI的理解不足：许多功能强大的AI基于复杂的神经网络模型，其内部机制对于人类用户而言是不透明的[5]。在典型的AI辅助决策系统中，AI往往只提供一个建议，决策者无法充分理解该建议的可信度、不确定性以及背后的逻辑和原因，从而难以做出有充分信息支持的决策[6]。

2.AI对决策者的理解不足：当前的AI模型在训练过程中通常并未将“理解人类”作为优化目标，更多地专注于针对任务本身的性能提升。由于AI对人类的能力、目标、偏好以及价值观缺乏理解[7]，使AI难以根据人类决策者的个体差异动态调整提供的建议，限制了AI在实际决策场景中的自适应性和灵活性[8]。

3. 决策者对自身的理解不足：决策者对自身能力的认知偏差，同样影响其是否能够合理采纳AI的建议。研究表明，即使是领域专家，在某些决策任务中也可能对自身能力做出欠准确的评估[9~11]。低估自身能力的决策者可能会过度依赖AI的帮助，而高估自己能力的决策者则可能忽视AI的正确建

议，进而做出非理性的判断，影响决策质量。

4. 缺乏促进决策者与 AI 互相理解的交互方式：现有的人机合作决策界面通常仅展示 AI 提供的信息，赋予人类决策者的交互手段大多局限于完成任务本身，缺乏进一步的人机沟通渠道 [12]。相比之下，在人类团队协作决策中，成员之间会相互了解彼此的能力和意图，尤其是面对相互冲突的观点时，能够进行有效的沟通与解释，从而促进思维的深入交流 [13]。但这些在人类协作中至关重要的行为在现有的人机合作决策系统中仍然存在缺失。

针对以上挑战，本文将介绍笔者所在团队在人机合作决策场景中围绕人机互相理解这一关键问题展开的研究与探索，旨在提升人机协同中的决策者与 AI 的相互理解，从而更好地支持人机合作决策系统的调整和优化。

促进决策者理解AI

AI 在辅助决策者开展决策任务时，除了提供具体的建议之外，还应充分理解决策者的信息需求和既有的工作流程，提供适配决策任务流程的辅助与可解释性信息，从而帮助决策者理解和分析 AI 的建议，做出深思熟虑的决策。

提供以决策者为中心的多维度决策指标

我们针对科学发现中的人机合作决策进行了探索 [14]。具体来说，多步逆合成路线规划（Multi-step Retrosynthetic Route Planning, MRRP）是合成化学中的核心任务，化学家需要递归地解构目标分子，寻找一组可实现目标分子的反应物 [15]。由于搜索空间巨大，化学家仅凭人力往往难以找到最佳路径。现有的 AI 模型可以快速实现目标简单的自动 MRRP，但对复杂分子的规划仍依赖化学家的专业知识。因此，人机合作决策在该场景中尤为重要。

MRRP 过程主要分为两个步骤：第一步是对目标分子进行解构（deconstruction）；第二步是修订（revision），当解构过程中出现不可行或难以获取的分子时，对这些问题分子进行回溯和修改。通过 6 名化学家参与式设计，我们深入理解了决策专家的信息需求和关注点，提出了一种人机协作决策系统 RetroLens（如图 1 所示）。在这个系统中，AI 通过联合行动（joint action）和算法在回路（algorithm-

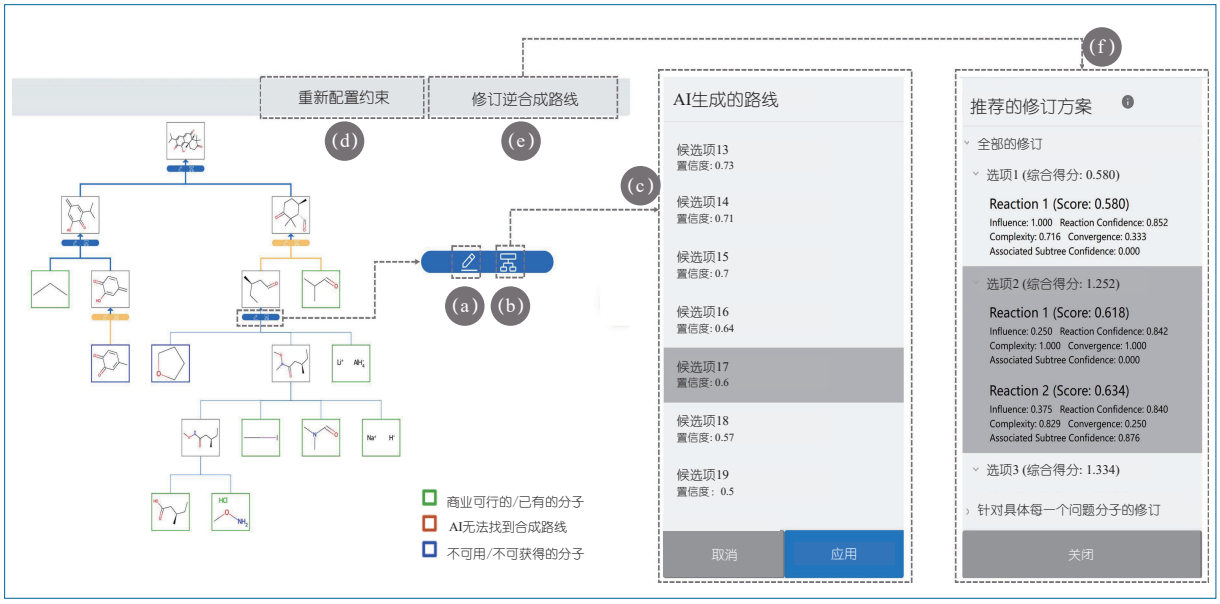


图1 RetroLens界面。（a）编辑化学反应按钮；（b）AI生成路线按钮；（c）AI生成的候选路线推荐页面，呈现针对每一条路线的置信度；（d）重新配置约束按钮；（e）修改逆合成路线按钮；（f）AI逆合成路线推荐页面，呈现专家所需的多维度决策信息

in-the-loop）两种方式辅助化学家进行决策。针对解构步骤，RetroLens不仅提供AI生成的候选解构路线，还为每条路线提供AI的置信度（confidence）信息，帮助决策者更好地理解AI建议中的不确定性，使其在结合自身领域知识的基础上对候选路线进行权衡。对于修订步骤，RetroLens除了提供AI推荐的修改方案外，还基于分子合成相关文献和专家反馈，展示了化学家关心的五个关键维度指标：修订量（amount of revision）、反应可信度（reaction confidence）、复杂性降低程度（complexity reduction）、收敛性（convergence）和相关分支可信度（associated branch confidence）。这些指标有助于决策者全面理解AI建议的依据，从而做出更加优化的决策。

RetroLens采用以用户为中心的设计理念，聚焦决策者在任务执行和可解释性方面的实际需求，促进决策者更深入地理解AI建议的内在逻辑。在一项有18位化学从业者参与的用户研究中，相较于基线系统，使用RetroLens的化学家在MRRP任务中的完成速度更快，探索设计空间的广度更大，规划信心更高，认知负荷更低。

提供符合决策者工作思路的AI辅助

我们的第二项工作[16]是开发了一个交互式可视化系统MedChemLens（如图2所示），帮助药物化学家探索和发现具有研究价值和潜力的药物靶点，决定“朝哪个方向前进”。MedChemLens对药物化学家的心智模型进行计算并建模，从而设计了一个可以整合和分析药物设计相关学科（化学、药理学和临床医学）的大规模文献和临床数据的AI决策助手。该系统模拟化学家的决策思路，从文献中检索与给定药物靶点相关的药物化合物，并分别从文献的文本、图片和表格中提取这些化合物的关键分子特征。基于这些数据，系统从成药性、研究潜力、研究热度、研究难度、研究可行性五个维度对药物靶点进行分析。为了让决策者更好地理解AI的分析结果，MedChemLens进一步提供了交互式可视化

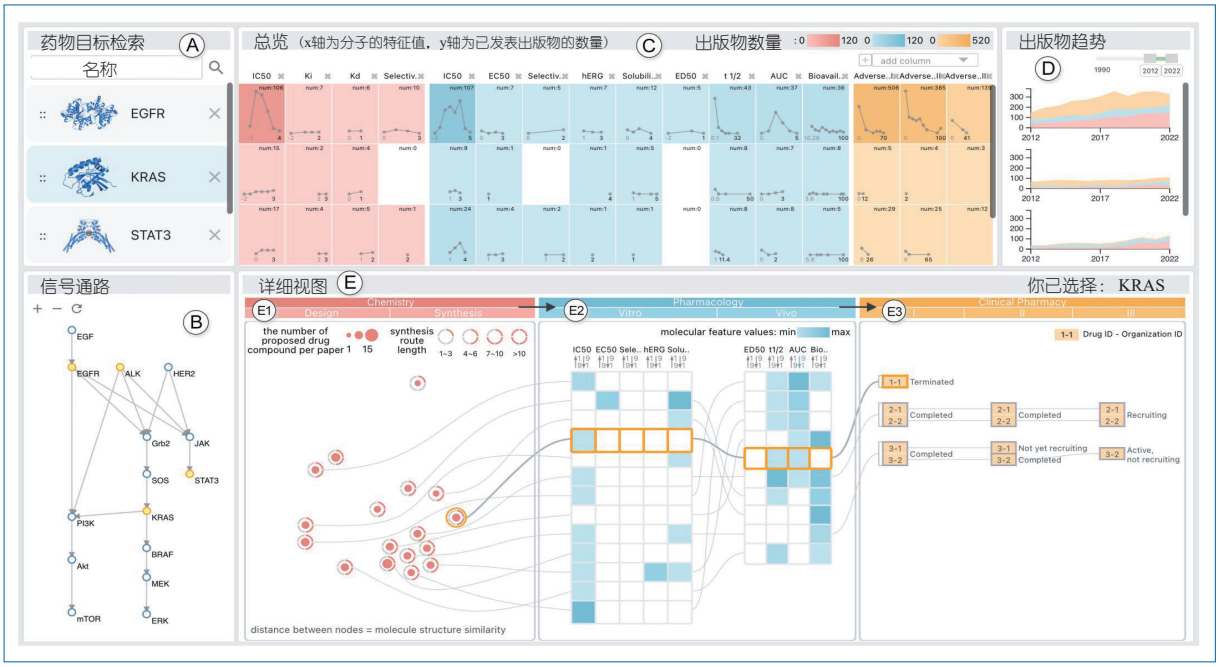


图2 交互式可视化系统MedChemLens。（A）药物目标检索视图允许用户按名称搜索药物靶点；（B）信号通路视图显示所搜索靶点的信号通路；（C）总览视图显示现有药物化合物研究的总体分布；（D）出版物趋势视图显示与所搜索靶点相关的出版物数量随时间的变化；（E）详细视图包括：（E1）Chemistry面板总结了化学文献中提出的药物化合物，（E2）Pharmacology面板显示在体外和体内药理学试验中测试的药物化合物的分子特征值，以及（E3）Clinical Pharmacy面板对药物化合物的临床试验进展进行可视化

界面，支持决策者直观地对多个药物靶点进行多角度比较，从而做出最终决策。一项对 16 名不同专业水平的药物化学研究人员开展的用户研究显示，与传统的药物靶点选择方法（即在线搜索相关文献和资料）相比，MedChemLens 可以加快研究者对所需资料的搜索速度，降低他们在药物靶点选择过程中的认知负荷，并帮助他们做出更加合理的选择。

促进AI理解决策者

在人类协作决策中，提供建议的一方通常会根据接收者的能力、认知水平和偏好，灵活调整建议的内容和呈现方式。然而，现有的 AI 决策助手往往仅基于算法推测提供最优的建议，而忽略了对决策者的理解。通过对决策者进行建模可以有效促进 AI 对决策者的理解。

人机合作决策的关键之一在于人类决策者能够明确掌握何时应更加信任 AI 的建议，何时应依赖自己的判断。然而，现有的人机合作研究常采取的方案是基于 AI 的置信度校准人类对 AI 的信任程度 [17~19]。尽管经过良好校准的 AI 置信度（well-calibrated confidence）在一定程度上能够反映 AI 在特定决策样本上的正确可能性（Correctness Likelihood, CL）[20]，但这些方法通常忽视了人类自身的 CL，导致团队决策的整体优化受限。如图 3（a）所示，现有的方法通常是在 AI 的 CL 超过或低于某一阈值时，通过干预手段调整人类对 AI 的信任。然而，更合理的校准方式应基于人类和 AI 的相对 CL，而不是仅与一个固定阈值进行比较（如图 3（b）所示）。为解决这一问题，我们提出了一种基于人类与 AI 能力的动态信任校准方法 [21]，分为两个阶段：

第一阶段：对人类决策者的 CL 建模。我们针对基于多属性表格数据（multi-attribute tabular data）的决策任务提出了一种用户能力建模方法，结合了数据驱动和用户交互式反馈。首先，从决策任务数据集中采集部分样本并由决策者进行标注，以此训练一个基础的决策模型；然后，将该模型转换为决策者易于理解的决策树或规则集形式，并设计了一个易用的交互界面，使决策者能够根据自己对任务的理解，交互地完善自己的决策模型；最后，针对新的决策案例，利用该模型对相似样本进行模拟决策，并通过动态权重调整方法，近似计算出决策者在新决策案例上的 CL。

第二阶段：基于双方 CL 的动态干预。在此阶段，我们基于认知科学理论设计了三种干预方式来校准人类对 AI 的信任程度：直观对比 CL、自适应工作流和自适应推荐。在直观对比 CL 中，系统直接向决策者展示 AI 和人类的 CL 对比，显式校准人类的信任。自适应工作流基于锚定误差（anchoring bias）[22] 理论：当决策者的 CL 优于 AI 时，系统会在 AI 建议之前让决策者独立做出判断；否则直接呈现 AI 建议。自适应推荐基于认知强制策略（Cognitive Forcing Strategies）[23]：当决策者的 CL 优于 AI 时，仅向其提供 AI 对决策问题的分析，而不会展示具体建议；否则直接提供 AI 的建议。

我们通过一项涉及 293 人的众包实验验证了该

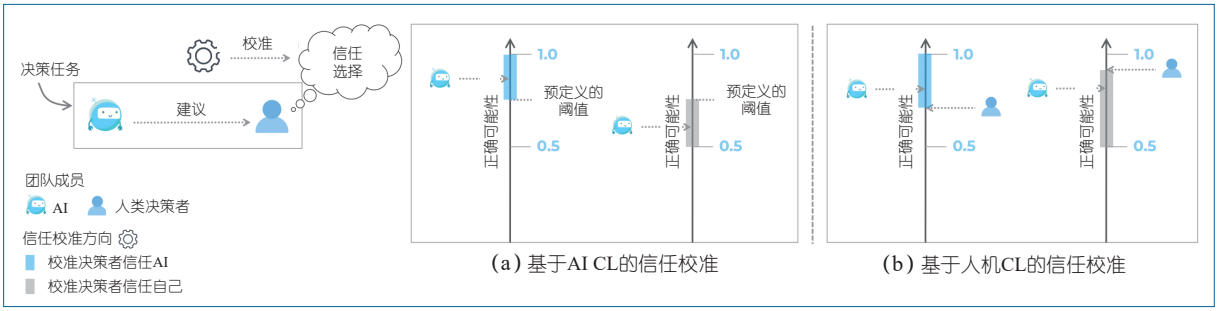


图3 （a）已有的决策系统通常通过将AI的CL与预设阈值进行比较，决定对人类信任的校准；（b）我们的工作通过对决策者的CL建模，提出了一种基于人类与AI能力的动态信任校准方法，使信任校准更加灵活和个性化

方法的有效性。结果表明，与AI单独决策、决策者单独决策以及传统的仅展示AI置信度的基线方法相比，基于双方CL的三种信任校准方式显著提升了人机合作决策的表现。更重要的是，这些方法在显著减少人类对AI过度信任的同时，显著提升了人类对AI的不信任。

提升决策者的自我理解

在人机合作决策任务中，决策者在权衡自身与AI的观点的可信度后做出决策。因此，决策者除了需要对AI的建议有正确的理解，还必须对自身决策能力有准确认知。然而，人类决策者往往受认知偏差影响，导致自我认知不足，进而影响人机合作的有效性。因此，从提升决策者对自身的理解入手，是优化人机协作决策的关键路径之一。

研究表明，决策者在一些决策任务中的自信心（self-confidence）常常与其实际决策能力不匹配[9~11]。这种不适当的自信心可能导致决策者对AI建议的采纳偏差。例如，过度自信的决策者（即对自身错误判断抱有过高信心）可能会忽视AI的正确建议，而过度不自信的决策者（即对自身正确判断缺乏信心）则更可能盲目采纳AI的错误建议。

为解决这一问题，我们基于决策优化的相关方法，提出了“决策者自信心校准”的概念，旨在改善决策者对AI建议的合理采纳[24]。首先，我们构建了一个分析框架，指出决策者的不合理自信是导致AI信任偏差的重要潜在原因。围绕这一问题，我们开展了三项用户实验研究。

第一项用户实验探讨了人类自信心合理性与对AI信赖合理性之间的关系。通过一项94名用户参与的实验，我们发现二者之间存在显著相关性，且在决策者自信心失调的任务中，决策错误率明显增加。

第二项用户实验基于决策理论和认知科学中的自信校准策略[25]，引入了三种自信校准机制：反向思考、下注式思维和校准状态反馈。一项241名用户参与的实验结果表明，反向思考和校准状态反馈机制能够有效调整决策者的自信水平，使其与实际决策能力更为一致。

第三项用户实验考察了自信心校准对人机合作决策的实际影响。在一项117名用户参与的实验中，我们发现相比未校准的情况，自信心校准促使决策者在采纳AI建议时表现出更强的理性选择，减少了对AI的信任不足，并显著提升了任务表现。

基于我们的发现，下一步的人机合作决策系统需要在优化AI建议准确率的同时，关注对决策者自我认知的校准，从而提升整体人机团队的决策质量。

促进决策者与AI双向理解

在构建决策者与AI的心智模型过程中，如何促进双方的双向理解是至关重要的一环。除通过算法定量优化心智模型的匹配外，我们进一步关注沟通机制的设计，以增强决策者与AI之间的相互理解。这种设计不仅有助于改善合作效果，还为提升人机协作系统的整体效能提供了新思路。

在人类协作决策中，讨论是一种常见且有效的方式。讨论不仅帮助团队成员充分表达观点、深入探讨问题，还能促使不同视角的融合，避免盲目决策，提高决策质量[13]。然而，当前的人机合作决策系统大多只向人类决策者提供AI的建议和理由（如基于AI的解释），而没有为决策者和AI之间的互动或协商提供交互支持。决策者只能在阅读完AI的建议后，整体采纳或拒绝AI的观点[1, 2]，系统缺乏对观点分歧的处理机制，也不支持人类与AI之间的讨论与交流。这种沟通机制的缺失限制了人机双方的理解，也阻碍了决策者对问题的深入思考。

针对这一问题，我们开展了两项探索性工作，旨在促成人类决策者与AI之间的讨论与协商。

第一项工作[26]设计了一个帮助英语教师对学生作文进行打分的讨论型AI决策助手。我们首先训练了一个基于BERT和LSTM的评分模型，并通过SHAP可解释（Shapley Additive exPlanations）AI算法，从内容、组织、风格和惯例四个维度提供针对句子层面的评分解释。该模型通过幕后操纵（Wizard-of-Oz）的方式集成到一个AI决策支持对

话机器人 AESER 中，并通过名义小组技术（Nominal Group Technique, NGT），让 AESER 与两名教师组成小组，对学生作文进行独立评分、讨论和投票。我们邀请了 20 名国内中学和高校的英语教师参与实验。参与者认为，与 AI 进行讨论提高了小组决策的客观性和公平性。但他们也指出，AI 的表现较为僵化，无法在提供建议时灵活地融入人类意见或跟上讨论节奏。

为了进一步优化讨论式决策，我们的第二项工作 [12] 提出了一个全新的人机协商（Human-AI Deliberation）框架（如图 4 所示），旨在促进人类在与 AI 意见发生冲突时展开反思与讨论。该框架支持决策者与 AI 在不同维度上进行观点表达、协商，甚至做出必要的观点妥协与更新。为了让 AI 具备协商能力，我们开发了协商型 AI（deliberative AI），它结合了大语言模型（LLM）的语言理解和表达能力，以及领域专属模型的任务能力，弥合了领域模型与人类之间的沟通差距，从而在实现灵活对话的同时确保提供的信息准确，避免 LLM 幻觉（hallucination）的影响。一项关于研究生招生决策的探索性评估结果显示，与传统的可解释 AI（XAI）助手相比，协商型 AI 助手在提升决策准确性、促进对 AI 建议的合理采纳方面表现出显著潜力。人机协商帮助决策者发现自身与 AI 的决策偏差，从而做出更加理性的决策。

未来展望

展望未来，面向人机协同决策中的双向理解，还有许多关键的问题亟待解决。

在人类理解 AI 方面，尽管现有的可解释 AI 算法已经较为丰富，能够提高 AI 的透明性 [5]，但实证研究表明，简单地将这些算法应用于人智协同场景中，并不总能带来积极效果 [3, 18, 27]。原因之一在于，这些解释往往较为复杂，尤其对于非 AI 相关领域的用户来说，理解这些解释存在困难，还可能增加认知负担 [28]，甚至引发“虚幻的解释效应” [29]。此外，目前的可解释 AI 与人类之间的解释方式存在明显差距。人类给出的解释通常具备对比性、选择性、社会性、因果性等特征 [30]，而 AI 的解释则更多只是从算法角度出发。未来的研究需要从社会科学和认知科学中汲取灵感，设计更加符合人类思维方式的 AI 解释机制，使之更具“人类兼容性”。

AI 理解人类是另一个极具挑战性的领域。理解人类的意图，一直是人机交互中的核心研究问题之一 [31]。由于人类的复杂性和行为的不确定性，精准理解人类意图充满挑战。此外，在人智协同场景中，不同的人对于 AI 的心智模型（或 belief）存在着不同程度的差异，同时也会随着与 AI 合作的展开而动态变化 [32]。目前一些算法通过基于大量人类群体的决策数据对“平均人”进行建模 [33, 34]，或将人类

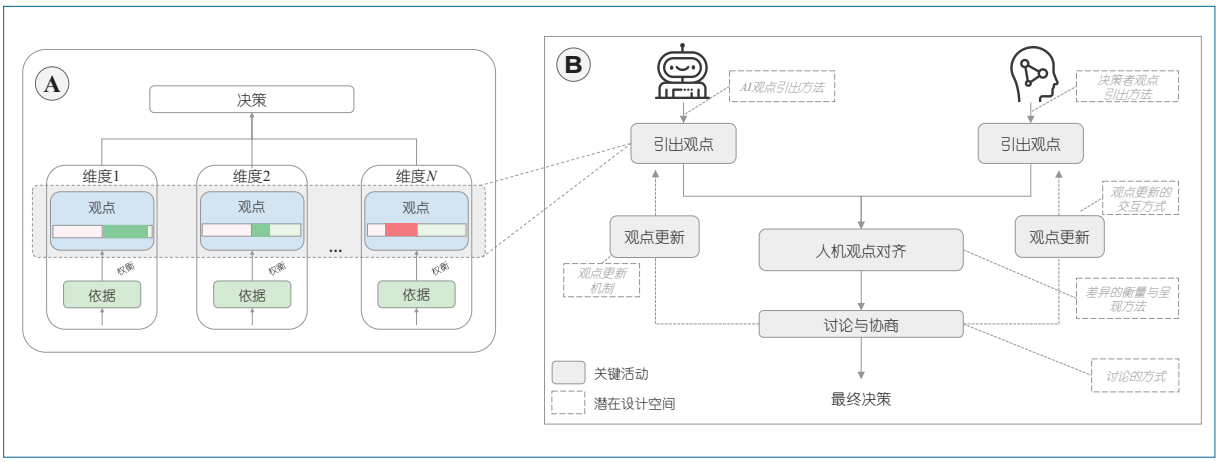
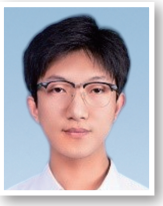


图4 Human-AI Deliberation示意图。该框架基于证据权重（weight of evidence）的方法，支持人类决策者和AI进行细粒度的观点引出、对齐、讨论和更新

决策者当作一个模型驱动的“智能体”[35, 36]来优化AI的行为，然而这些方法忽视了人与人之间的差异。未来的研究应更加关注用户建模方法，一方面将基于认知科学和心理学的理论驱动方法[37, 38]作为建模基础，另一方面利用数据驱动的方法，结合历史交互数据，学习用户的行为模式、偏好和习惯，帮助AI构建关于人的信念、意图、行为的模型[39, 40]，并随着交互的深入动态调整模型。此外，随着生成式模型的发展，可以考虑利用大语言模型或扩散模型等技术构建代表不同用户特征的智能体，在建模初期通过模拟用户行为获得初步启发。

实现双向理解不仅依赖先进的算法支持，还需要设计高效的沟通机制。AI生成的信息应符合人类认知，并以人类接受的形式传达。未来的设计可借鉴“心智理论”（Theory of Mind）[41]，通过优化AI的行为表达，使其更贴合用户的思维模式和推理习惯，从而帮助人类在与AI合作过程中形成对AI的准确认知、感知、预测。

总结来说，未来的人机协作决策研究既要在现有技术上进行突破，也须注重以人为中心的设计理念，将计算机科学与认知科学、社会科学和心理学深度融合，实现人与AI之间的双向理解与高效协作。 ■



马　帅

芬兰人工智能中心及阿尔托大学博士后。主要研究方向为人机交互，聚焦于人智协同、以人为中心的人工智能和智能交互系统，及其在教育、决策等场景的应用。
shuai.ma@aalto.fi

麻晓娟

CCF高级会员、人机交互专委会执行委员。世界华人华侨人机交互协会理事、常务副会长。香港科技大学副教授。主要研究方向为人机交互和人机共谐计算。
mxj@cse.ust.hk

石楚涵

CCF专业会员、人机交互专委会执行委员。东南大学副教授。主要研究方向为数据可视化、可视分析、人机交互及其在自然科学、精准医疗等领域的应用。
chuhanshi@seu.edu.cn

其他作者：郑成博

延伸阅读请登录 http://dl.ccf.org.cn/cccf/list