arXiv:2601.13376v1 [cs.ET] 19 Jan 2026

有限心智，生成式机器

Envi构建能与人类启发式协同并降低偏差风险的对话式人工智能 k

JIQUN LIU,俄克拉荷马大学，美国

对话式人工智能正迅速成为信息寻求和决策的主要界面，但大多数系统仍假设用户是理想化的。实际上，人类推理受到注意力有限、知识不均衡以及依赖启发式的制约；这些启发式虽具有适应性，却也容易产生偏差。本文提出了一条以有限理性为基础的研究路径，并指出，对话式人工智能的设计应与人类启发式协同，而非与之对抗。文章还提出了若干关键方向：识别认知脆弱性、支持不确定情境下的判断，以及不再局限于评估事实准确性，而是进一步评估对话系统的决策质量和认知稳健性。

CCS 概念：**·信息系统 → 用户与交互式检索**。

其他 关键词和短语：有限理性、启发式、对话式人工智能、生成式人工智能、评估 n

ACM 参考文献格式：

Jiqun Liu. 2025. 有限心智，生成式机器：构想与人类启发式协作并降低偏差风险的对话式人工智能。*J. ACM* 37, 4，第111篇文章（2025年8月），5页。https://doi.org/XXXXXXX。XXXXXXX

对话式人工智能中的有限理性

对话式人工智能系统正逐渐进入主流，成为信息寻求和决策的主要界面。然而，大多数系统、评估方式以及隐含的用户模型，仍然假设用户是理想化的，能够仔细权衡证据并识别不可信的论断。实际上，人类认知受到有限时间和注意力、不均衡的领域知识，以及依赖于启发式的制约；这些启发式经过演化，旨在支持不确定性下的快速判断 [9]。锚定效应、可得性启发、确认偏误、权威启发和稀缺性启发等启发式，并非推理失败，而是具有适应性的思维捷径。然而，当这些捷径与流畅、个性化且自信的生成式回复结合时，便会成为一把双刃剑，可能系统性地使判断产生偏差。

人机交互、认知科学和行为经济学领域的研究人员长期以来一直在研究这些偏差，人工智能伦理和安全社区也日益认识到这一点 [2]。然而，主流设计实践和系统评估仍然强调正确性、相关性以及狭义定义的效用，同时对认知影响的探讨仍然不足 [5]。在多轮对话中，一次自信的回答就可能锚定用户，生动轶事可能压过基准率，反复的认同可能强化预期和错误信念，甚至中性的布局或排序选择也可能引入框架效应，从而改变信任和偏好。

风险不仅在于对话式人工智能无法缓解这些偏差，更在于系统可能利用这些偏差，大规模实施行为操纵；这与欧盟人工智能法案 [1] 对不可接受风险的界定相一致。在广告、说服和政治微定向领域，系统已经

[1]参见 https://www.europarl.europa.eu

作者的 联系信息：Jiqun Liu，jiqunliu@ou.edu，俄克拉荷马大学，诺曼，俄克拉荷马州，美国。

特此免费授予出于个人或课堂使用目的制作本作品全部或部分内容的数字副本或纸质副本的许可，但前提是不得以营利或商业利益为目的制作或分发副本，并且副本的第一页须附有本声明及完整引文。本作品中由作者以外的其他人拥有版权的组成部分，必须尊重其版权。经注明出处后，可以进行摘要。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 申请许可。
© 2025 版权归所有者/作者所有。出版权授权予 ACM。ACM 1557-735X/2025/8-ART111
https://doi.org/XXXXXXX.XXXXXXX

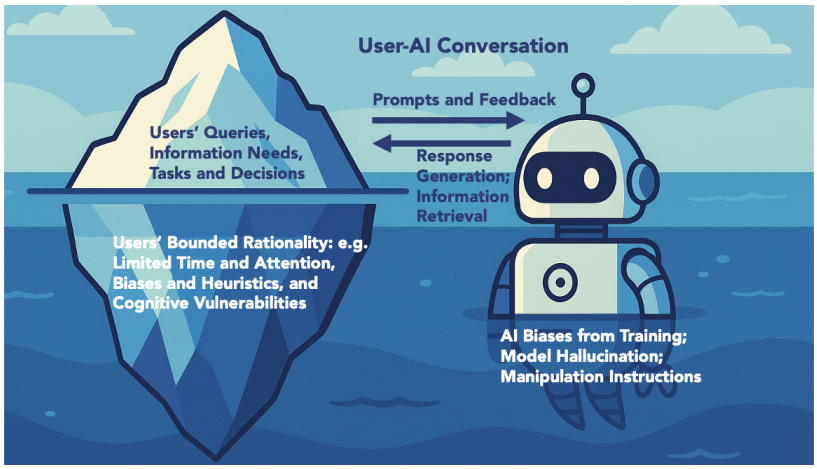


图 1. 用户与人工智能对话中的挑战：系统会响应明确提出的提示，但无法察觉潜在的有限理性，也无法支持“水面之下”的隐性人类启发式。

通过迎合可预测的脆弱点来优化用户参与度 [1]。随着生成式模型变得更加自适应和个性化，细微的引导也更容易实现，更难以察觉。

因此，挑战并不只是提高事实准确性，而是要构建能够识别用户实际如何借助启发式进行推理的对话系统——这些启发式会因情境、个体差异和系统选择而发挥适应性作用，也可能造成扭曲。一项新兴研究方向有时被称为机器心理学，它探讨生成式模型如何表现出认知捷径、如何与之互动，以及系统如何支持更好的推理，而不是放大偏差 [8]。这促使我们开辟一条研究路径，将设计、评估和治理联系起来，以在实践中减少人工智能危害和偏差风险 [6]。

为何如今这成为一条研究路径？有限理性早已得到研究，但随着大语言模型（LLM）和智能体对话系统得到广泛部署，其影响变得迫切得多。与早期的信息系统不同，对话式人工智能能够在每一轮对话中调整自然语言，并在更长的时间跨度内影响推理，从而放大框架效应、锚定效应和启发式强化的累积影响。未来十年，认知影响将因此成为首要的设计考量，而非次要问题。

为启发式而设计，而非与之对抗

本文并非主张采用某一种设计方案，而是强调一些反复出现的认知动态；随着启发式在多轮交互中展开，这些动态为对话式人工智能勾勒出更广阔的研究路径。认识到人们依赖的是启发式，而非数学上理性的效用计算后，研究的挑战便在于理解：当系统针对不同任务，以流畅、适应性强且可规模化的方式生成回应时，这些捷径会如何发挥作用。曾经看似微妙的实验室效应，可能在日常交流中逐渐累积；措辞、顺序和强调重点的细微变化，都会重塑判断、比较和偏好。

根据前景理论 [9],，人们会相对于参照点 评估结果，并且对损失的看重程度高于同等收益。在对话情境中，参照点往往通过语言、排序和语气隐含地确立，而早期的框架效应可能在多轮对话中持续存在。随着时间推移，即使没有引入事实错误，路径依赖性认同也可能限制探索。

选项集的结构也会改变结果。当比较内容是即时生成时，可能会出现被支配的选项，并在不知不觉中左右选择。例如，一名学生向大语言模型询问学习资源时，可能会看到一本免费的开放式教科书、一门价格合理的课程，以及一个价格过高、却几乎没有额外价值的“白金套餐”；诱饵选项会让付费课程看起来像是稳妥且均衡的选择。在对话式信息检索和推荐 [5]中，锚定效应、生动轶事、权威语气、稀缺性线索和寻求确认的语言也会产生类似影响；这些因素可能同时出现，并相互强化。

这些动态解释了为什么表层准确性不能成为评判对话式人工智能的唯一标准。一条回复可能是正确的，却仍会固化损失厌恶、放大过度自信，或通过隐性框架引导决策。研究问题由“答案是否正确”转向“互动是否在有限认知能力范围内支持健全的推理”：它是否拓展了视角、校准了不确定性，并降低而非加剧脆弱性？由于互动是逐轮展开且依赖语境的，这一路径要求除了关注质量导向的有效性之外，还应以认知影响为核心开展设计与评估。

检测并干预偏差风险

对话是人类与机器之间推理发展的场域，也是推理可能悄然偏转的地方。这一路径要求监测交流如何展开，并识别偏差风险信号：狭隘框架、对收益与成本的强调失衡、在没有表达不确定性的情况下提出过度自信的主张、掩盖基线或包含冗余选项的比较，以及用户未经审视证据便迅速接受结论。这些信号表明，未来系统可以追踪一些可测量的特征，包括通过来源熵衡量的观点多样性 [4], 表达的概率中的校准失误 [3], 以及支持事后审计的轮次级轨迹 [6]。在某些情境下，其他行为和生理信号也可能有所帮助，但具备可泛化性的、注重隐私的信号仍是核心研究挑战。

一旦检测到这些信号，系统就可以在呈现最终回答之前探索轻量级干预：明确展示基线，以对称方式呈现收益与损失表述中的结果，并将不确定性表示为范围，而非单一数值 [3]。检索可以提供基准率和反例 [4]。支配性检查可以标记或移除诱饵选项；在展示基线后，锚定防护机制可以引入第二个估计值，从而避免第一个数字主导判断 [10]。这些干预措施可以通过结构化输出、函数调用和轻量级评判器来实现，由它们在返回答案前审核草稿 [7]。

示例场景有助于明确这一目标。在健康建议中，如果先呈现一个生动的患者故事，它可能会压过来自临床试验的证据；将轶事与患病率配对，可以重新平衡判断。在电子商务中，高价诱饵可以让中档产品显得合理；检测并标记受支配的选项，可以减少隐藏式引导。在法律指导和新闻摘要中，也可能因冗余选项、问题性的框架效应或锚定顺序而出现类似失误。这些案例表明，启发式如何从捷径转变为脆弱点，也说明了为什么在有限理性下，小幅干预可能有助于让对话保持更加平衡。

评估也必须不断发展。准确性和主题相关性仍然必要，但已不再足够。应测试系统在呈现不确定性后，互动是否能提升校准水平 [3], 、拓宽视角，并降低用户受劣势选项和诱饵选项影响的倾向 [6]。通过改变框架效应、先验和时间压力构建的情境集，可用于对干预措施进行压力测试；而在线实验则可通过选择稳定性和重复查询减少情况来比较不同策略 [7]。反事实日志记录和交错可用于安全地评估替代基线；调优时还可加入惩罚项，以应对高风险轮次未实施干预的情况 [6, 7]。

尚待解决的挑战。实时检测认知脆弱性十分微妙，取决于具体情境，且在不同用户和领域之间存在差异。干预措施必须在有效性与侵入性之间取得平衡，因为过度的保护措施可能增加认知负荷，并降低信任或可用性。将检测、干预、评估和治理整合到连贯的对话流程管线中，会引发关于透明度和泛化能力的未决问题。这些挑战是研究路径中的核心问题，而非已经解决的工程任务。

从原则到实践

将这些理念付诸实践，并不需要重新发明对话式人工智能，但确实需要重新调整优先事项。许多基础模块已经存在，但如何有效地整合并评估它们，仍是开放的研究问题。函数调用、结构化输出和长上下文建模可以支持基准呈现和不确定性表示；检索增强生成和推理可以基于有依据的来源提供多元视角 [4]；轻量级评判器则可以标记缺失的基线或过度自信的语气 [7]。近期模型的发展使这些能力日益可行，但仅有可行性并不能保证带来认知益处和安全性。[2]

开放研究问题与路线图。 这一路径提出了若干可付诸行动的问题：(1) 系统如何才能在不同任务和人群中可靠地检测认知脆弱时刻？(2) 对比式框架、基准呈现或不确定性表示等干预措施，在不同启发式和情境下哪种效果最好？(3) 除准确性外，尤其是在校准、判断稳定性和抵御操纵方面，应如何评估成效？(4) 要让此类系统大规模承担责任，需要哪些治理机制和审计跟踪记录？

评估可以围绕这些目标重新构建。我们不应只问答案是否正确，还可以问交互是否改善了校准、减少了锚定效应，并拓宽了视角 [3, 6]。通过改变框架效应、先验和时间压力的情境集，可以测试稳健性；在线实验则可以通过选择稳定性和对诱饵选项的易感性降低情况来比较策略 [7]。随着时间推移，这些指标可以与准确性和延迟指标相辅相成，为决策质量和认知稳健性提供基准。

实用的保障措施也取决于治理和透明度。认知影响审查 可以与隐私和安全检查一道，纳入发布流程。政策卡和模型卡可以披露护栏范围及已知局限，而审计跟踪记录则可以记录展示了哪些选项、纳入了哪些基线，以及触发了哪些干预措施 [6]。实用的审计跟踪记录可以包含检索查询、排序选择、引用片段，以及系统针对高风险回答制定的内部计划。[3] 这类文档或许会让人觉得负担沉重，但它能实现可复现的可追溯性，便于问责。

总体而言，这些方向勾勒出了一条可信且及时的研究路径，旨在构建能够与受限的人类心智协作、支持稳健推理，并降低隐性伤害和认知行为操纵风险的对话式人工智能系统与评估流程。

作者简介

Jiqun Liu 是美国俄克拉荷马州诺曼的俄克拉荷马大学副教授 A.

参考文献

[1] Micah Carr oll、Alan Chan、Henry Ashton 和 David Krueger。2023 年。《刻画人工智能系统的操纵行为》。载于会议论文集 第三届*ACM*算法、机制与优化公平性和可及性会议。1–13。

[2]参见 OpenAI，《Introducing o3 and o4-mini》，https://openai.com/index/introducing-o3-and-o4-mini/ [3]参见 Anthropic 文档，《计算机使用工具》，https://docs.anthropic.com/en/docs/agents-and-tools/tool-use/computer-use- tool

[2] Tim Draws、Alisa Rieger、Oana Inel、Ujwal Gadiraju和Nava Tintarev。2021。应对众包中认知偏差的检查清单。载于*AAAI*人类计算与众包会议论文集，第9卷。48–59。[3] Chuan Guo、Geoff Pleiss、Yu Sun和Kilian Q Weinberger。2017。论现代神经网络的校准。载于国际机器学习会议。PMLR，1321–1330。[4] Patrick Lewis、Ethan Perez、Aleksandra Piktus、Fabio Petroni、Vladimir Karpukhin、Naman Goyal、Heinrich Küttler、Mike Lewis、Wen-tau Yih、Tim Rocktäschel等。2020。面向知识密集型自然语言处理任务的检索增强生成。神经信息处理系统进展33（2020），9459–9474。[5] Jiqun Liu。2023。交互式信息检索的行为经济学方法。信息检索丛书48（2023）。[6] Ahmed Magooda、Alec Helyar、Kyle Jackson、David Sullivan、Chad Atalla、Emily Sheng、Dan Vann、Richard Edgar、Hamid Palangi、Roman Lutz等。2023。生成式人工智能应用中负责任人工智能危害自动测量框架。*arXiv*预印本 *arXiv:2310.17750*（2023）。[7] Long Ouyang、Jeffrey Wu、Xu Jiang、Diogo Almeida、Carroll Wainwright、Pamela Mishkin、Chong Zhang、Sandhini Agarwal、Katarina Slama、Alex Ray等。2022。训练语言模型，使其能够遵循指令并利用人类反馈。神经信息处理系统进展35（2022），27730–27744。[8] Iyad Rahwan、Manuel Cebrian、Nick Obradovich、Josh Bongard、Jean-François Bonnefon、Cynthia Breazeal、Jacob W Crandall、Nicholas A Christakis、Iain D Couzin、Matthew O Jackson等。2019。机器行为。《自然》568，7753（2019），477–486。[9] Amos Tversky和Daniel Kahneman。1992。前景理论的进展：不确定性的累积表征。《风险与不确定性杂志》5，4（1992），297–323。[10] Shunyu Yao、Jeffrey Zhao、Dian Yu、Nan Du、Izhak Shafran、Karthik Narasimhan和Yuan Cao。2023。ReAct：在语言模型中协同推理与行动。载于国际学习表征会议（*ICLR*）。

收稿日期：2007年2月20日；修订日期：2009年3月12日；接收日期：2009年6月5日