DOI: 10.16538/j.cnki.fem.20240622.301

人工智能反馈：文献述评与研究展望

关　健[1]，李文朴[1]，何国华[2]，张新安[1]
（1. 上海交通大学 安泰经济与管理学院，上海 200030；2. 深圳大学 管理学院，广东 深圳 518060）

摘　要：绩效反馈是激励并促进个体进步最重要的方式之一。随着人工智能技术的进步，由AI提供的反馈在实践中得到日益广泛的应用，在质量上也开始逐步超越人类管理者所提供的绩效反馈，已经成为组织管理研究中的一个重要话题。然而，现有文献对AI反馈的研究分散在组织管理、教育和医疗等不同学科领域，这导致AI反馈相关研究在研究范式、理论视角和实证方法等方面具有较大差异，且现有文献对AI反馈带来的不同效应在理论机制方面尚未形成一致的认识。鉴于此，本文首先厘清了AI反馈的概念内涵和外延；然后，系统地归纳和总结了AI反馈产生影响的配置效应和披露效应机制，构建了AI反馈的研究框架；接下来，介绍和总结了现有AI反馈研究常用或具有深远启示意义的理论机制，并讨论了理论的使用方式；最后，在此基础上，提出了五个具有科学价值和现实意义的未来研究方向。

关键词：人工智能；人工智能反馈；配置效应；披露效应
中图分类号：F270　**文献标识码：**A　**文章编号：**1001-4950（2025）03-0083-18

一、引　言

绩效反馈是激励并促进个体进步最重要的方式之一，其有效性很大程度上取决于反馈的特征以及反馈接收者的主观心理感受（王永丽和时勘，2004；李璨等，2019；Lechermeier和Fassnacht，2018；Panadero和Lipnevich，2022）。近年来，人工智能（artificial intelligence，AI）技术的进步使机器执行任务的形式愈加复杂和丰富（Glikson和Woolley，2020；Pereira等，2023），基于自身强大的大数据分析和自我学习能力，AI已经能够对人类的任务执行进行有效评估，并为个体提供任务反馈（Luo等，2021；Parent-Rocheleau和Parker，2022；Qin等，2023；Tong等，

收稿日期：2023-12-19
基金项目：国家自然科学基金项目（72232005，72472100）；国家留学基金“国家建设高水平大学公派研究生项目（所在单位或个人合作渠道）”（CSC202306230250）；广东省自然科学基金面上项目（2024A1515012634）；深圳大学高水平大学三期建设项目（000001032226）
作者简介：关　健（1993—），男，上海交通大学安泰经济与管理学院博士研究生；
李文朴（2000—），女，上海交通大学安泰经济与管理学院博士研究生；
何国华（1990—），男，深圳大学管理学院助理教授；
张新安（1977—），男，上海交通大学安泰经济与管理学院教授，博士生导师（通信作者，xinanzhang@sjtu.edu.cn）。

2021；Vrontis等，2022）。AI反馈（AI feedback）是指在组织管理、教育和医疗等情境下，综合应用AI技术来追踪个体行为、自动化评估个体的任务表现，并通过文字、语音、动作等方式推荐任务改善策略（Tong等，2021）[①]。例如，AI能够将员工与顾客之间的语音对话转化成文字，在进行智能分析后针对员工的错误提供改进反馈，从而提升员工绩效（Tong等，2021）。

使用AI反馈有诸多好处：首先，AI能随时响应人类需求，能大幅降低组织的人力成本（Oranburg和Palagashvili，2021；Vrontis等，2022）；其次，AI能整合大量信息，同时避免人类的主观偏差，使反馈更全面和客观（蒋路远等，2022；魏昕等，2021；Blair和Saffidine，2019）；最后，AI作为反馈提供者不容易出现认知疲劳或情绪失控等问题（Barne等，2015）。因此，相比于人类反馈，AI反馈能以更全面、一致、客观的方式追踪和评估个体在任务中的行为表现，并生成个性化的改进建议（Heaven，2020），进而帮助个体显著提高任务表现（Colangelo，2020；Parent-Rocheleau和Parker，2022）。基于AI反馈的诸多积极影响，AI反馈在实践中得到日益广泛的应用。例如，联合利华采用AI为新员工提供反馈，帮助他们更好地适应工作（Marr，2018）；大量平台企业利用AI算法对零工从业者的工作过程进行监控，根据任务完成情况和质量给出及时的、有针对性的反馈，以更好地对其行为进行强化或纠偏（刘善仕等，2022；De Cremer，2020）。

然而，采用AI提供绩效反馈也可能带来负面影响。首先，由于AI算法并不是透明的，组织成员在与AI的交互中往往处于权利劣势地位，只能被动地接受AI给予的反馈，这会让组织成员对AI产生负面看法，从而阻碍其绩效的提升（陈龙，2020；李胜蓝和江立华，2020；Wood等，2019）。其次，人们在许多情况下对AI存在厌恶情绪，即使人们知道AI的建议优于人类，人们还是经常听从人类而非AI的建议（罗映宇等，2023；Castelo和Ward，2021；Dietvorst等，2015；Glikson和Woolley，2020）。最后，人们倾向于将AI视为没有感情的工具，认为AI不具备思维能力（Lanz等，2024），进而会认为AI缺乏情绪或道德感（Bigman和Gray，2018；Gray等，2007），从而对AI提供的反馈没有足够的信任。例如，当员工知道反馈由AI提供时，他们对反馈的信任度会下降，从而影响绩效改善（Tong等，2021）。

上述研究表明，AI反馈对组织的影响是研究者亟须关注的重要问题。Tong等（2021）的研究指出，采用AI进行绩效反馈可能产生两种不同的效应：配置效应和披露效应。配置效应（deployment effect）认为，采用AI进行数据分析能够提高反馈的质量，从而提升个体的绩效表现；而披露效应（disclosure effect）则认为，一旦反馈由AI提供被披露出来，员工可能对反馈信息产生负面看法，从而损害其工作绩效。近年来，越来越多的研究开始关注AI反馈的影响及其作用机制，相关研究的数量呈加速增长的趋势，并吸引了越来越多社会科学、医学、计算机科学等领域的研究者投入其中。然而，关于AI反馈的两种效应带来的影响是好是坏以及相关机制是怎样的，现有研究尚不明确。此外，AI反馈相关研究分散在管理学、教育学、医学等多个领域，存在研究范式不统一、理论不清晰和机制未厘清等问题。这既不利于现有研究之间的跨学科对话，也不利于后续研究对AI反馈形成系统性理解和借鉴。

鉴于此，本文系统地梳理了AI反馈相关研究，总结了AI反馈带来的影响及其中介机制和边界条件。具体来说，本研究将“人工智能/AI反馈”和“AI / artificial intelligence feedback”作为检索词，在Web of Science、PsycINFO、Springer、Scopus和中国知网等核心数据库中进行主题检索，并通过阅读标题对检索到的文献进行初步筛选，获得102篇文献。我们进一步阅读了文章的

①人工智能（AI）、算法（algorithm）和机器人（robot）是三个相关但不同的概念。AI是一种基于机器的系统，可以针对人类给定的一组目标，做出影响真实或虚拟环境的预测、建议或决策（OECD，2024）；算法是将输入数据转换为所需输出的计算机编程过程（Kellogg等，2020），是AI的重要模块和组成部分（Scott，2021）；而机器人是一种能自主控制、可编程、能够执行任务的物理实体（IFR, 2024），由AI所驱动的机器人才能提供反馈。本文重点关注AI反馈。

摘要、关键词和正文，排除与本文的研究对象、主题和领域无关的文献，截至2024年5月20日，获得目标文献共69篇，主要聚焦于组织管理[①]（24篇）、教育（33篇）和医疗（12篇）等研究情境。参考Tong等（2021）提出的AI反馈具有配置效应和披露效应两个不同的效应视角，本文将AI反馈相关研究依据这两种效应视角进行了机制拓展，并分成两类：第一，AI反馈的配置效应主要关注AI反馈带来的相对客观的技术层面的影响，主要聚焦于探讨AI反馈的质量、时间和多样化及其对结果变量产生的影响；第二，AI反馈的披露效应主要关注AI反馈带来的相对主观的心理层面的影响，主要聚焦于探讨人类对于AI反馈的不同主观心理感知变化及其对结果变量产生的影响。基于上述分类，本文系统地梳理了AI反馈影响个体工作结果的结果变量、中介机制和边界条件，以期为未来研究提供有价值的建议和指引。

二、AI反馈的配置效应相关研究

AI反馈的配置效应主要探讨的是AI作为反馈提供者，能够提供相对客观的技术层面反馈，进而对反馈接收者产生影响（Tong等，2021）。AI能够利用强大的数据分析和自我学习能力，全面跟踪个体的任务活动，准确、可靠地评估个体的任务表现。由于AI能够分析更多的个体表现数据，响应速度更快，因此能够生成高质量、即时和多样化的任务改进策略，对个体产生的直接影响也更大（Kiyasseh等，2023；Parent-Rocheleau和Parker，2022；Ryan等，2019；Tolsgaard等，2023）。结合现有文献，本文从AI反馈的质量、时间和多样化三个方面对AI反馈的配置效应研究文献进行总结（见表1）。

表 1　AI反馈的配置效应研究总结

类别	亚类别	主要结果变量	文献数量	参考文献
反馈质量	准确性	认知：动机、自我效能感等 态度：信任、反馈接受度、学习意愿、学习投入度等 行为：反馈接受行为、纠偏行为等 绩效：工作绩效、任务绩效、学习效果、工作技能等	13	Banihashem等，2024†；Darvishi等，2022†；Fu等，2020†；Kooli和Yusuf，in press†；Lee等，2022†；Lu等，2023‡；Nagaraj等，2023‡；Nazari等，2021†；Qin等，2023†；Ryan等，2019‡；Tong等，2021*；Vodrahalli等，2023‡；Zhao等，2023†
	可靠性		7	Bulten等，2021‡；Butow和Hoque，2020‡；Hwang等，2023†；Kiyasseh等，2023‡；Tolsgaard等，2023†；Tong等，2021*；Yang等，2023‡
	可解释性		6	Adnan等，2022†；Afzaal等，2021†，2024†；Conati等，2021†；Das等，2023‡；Mirchi等，2020‡
反馈时间	即时性	认知：动机、认知负荷等 态度：学习投入度/满意度/意愿等 绩效：学习效果、任务绩效、任务质量等	11	Al Hakim等，2022†；Chen等，2022a*，2022b†；Chien等，in press†；de Laat等，2020*；Fidan和Gencel，2022†；Godwin-Jones，2022†；Kang等，2023†；Lee等，2022†，2023†；Ouyang等，2023†
反馈多样化	个性化	认知：自我效能感、认知负荷等 态度：学习满意度/投入度、情绪（焦虑）等 绩效：工作绩效、学习效果、任务质量/绩效等	11	Adnan等，2022†；Belcadhi，2016†；Chen，2024†；Conati等，2021†；Jacobs等，in press†；Hsia等，in press†；Hwang等，2023†；Kim等，2022*；Liu等，2023†；Marafie等，2021*；Suraworachet等，2023†
	丰富性		7	魏爽和李璐遥，2023†；Cold等，2024‡；Drewery等，2022*；Ghafouri等，in press†；Lee，2023†；Luo等，2021*；Ryan等，2019‡

注：*组织管理情境，†教育情境，‡医疗情境。

①部分心理学实验研究未指明场景，由于心理学实验与组织管理紧密相关，因此我们将这些研究也纳入组织管理情境。

（一）AI反馈质量

根据现有文献，AI反馈质量包含三个核心维度：准确性（accuracy）、可靠性（reliability）以及可解释性（interpretability），这三个维度并非孤立存在，而是相互影响、相互制约的关系（Alvarez-Melis和Jaakkola，2018；Li等，2022；Nussberger等，2022）。基于此，本文结合现有文献对AI反馈的准确性、可靠性和可解释性三个方面进行深入的探讨和回顾。

1.AI反馈准确性

AI反馈准确性是指AI提供的反馈在多大程度上能正确反映事实（Nussberger等，2022）。目前，AI反馈的准确性已经达到能够与人类专家相媲美的水平，甚至优于人类专家（Das等，2023；Tong等，2021；Zhao等，2023；Lu等，2023）。因此，AI反馈被广泛应用于不同行业。例如，在教育领域，相比于人类教师提供的反馈，使用AI进行反馈能让学生获得更多关于课程的详细、准确的反馈信息（Banihashem等，2024；Kooli和Yusuf，in press），这有助于学生进行自主学习，为学生的认知强化提供了机会，从而能够提高学生的学习投入度（Nazari等，2021）、学习意愿（Fu等，2020）和反馈接受度（Darvishi等，2022）。更重要的是，AI反馈准确性能够对学生的学习效果产生直接影响，提升学生的成绩和技能（Darvishi等，2022；Nagaraj等，2023；Lee，2023；Qin等，2023）。例如，Lee等（2022）发现，与采用常规的课后复习方式相比，使用AI进行课后复习能够为学生提供更高质量的反馈，提升学习活跃度，从而提高学生的学习成绩、自我效能感和学习意愿。在医学领域，AI能够为人类医生提供详细和准确的诊断反馈，有助于医生更加准确地判断病情，从而能够提高医生的工作效率和绩效（Lu等，2023；Ryan等，2019；Vodrahalli等，2023；Nagaraj等，2023）。例如，在远程医疗场景中，患者需要提供病情相关图片以便医生做出医疗诊断和治疗决策，AI能够在准确识别患者所提供照片的清晰程度后给患者提供反馈，将图像质量差的患者数量减少了68.0%，进而提升了医生的诊断准确性。

2.AI反馈可靠性

AI反馈可靠性是指AI在连续或多次应用时所给出反馈的稳定性和一致性（Glikson和Woolley，2020）。这种可靠性意味着在相似或重复的任务中，AI能够提供一致的输出，从而有助于提高个体的工作绩效（Tong等，2021）。在医学领域，AI已经展现出在图像解读上的出色可靠性。例如，Tolsgaard等（2023）发现，AI能够为医生提供可靠的医学图像解释反馈，从而提升医生的诊断绩效。此外，AI还能够通过深入分析患者的沟通记录，为医务工作者提供更为准确和可靠的病症反馈，进而简化工作流程并提升工作效率（Butow和Hoque，2020）。这种高效、可靠的信息反馈对于患者的治疗及健康行业的决策制定具有深远影响（Hwang等，2022；Kiyasseh等，2023；Yang等，2023）。例如，Bulten等（2021）研究发现，基于深度学习的AI系统能够在医疗诊断中提供可靠的评分结果，在对160例活组织检查进行评分后，AI组与国际专家组的诊断标准一致性得到显著提高，有AI辅助的病理学家的表现优于无AI辅助的病理学家的表现。

3.AI反馈可解释性

AI反馈可解释性是指AI提供的反馈在多大程度上能被人类所解释（Cadario等，2021）。许多研究指出，个体对AI的抵制往往源于其无法理解AI如何产生特定的输出。因此，提高AI可解释性能够帮助组织更好地应用AI（Bauer等，2023；Cadario等，2021；Das等，2023）。在教育领域，结合AI与学习分析技术，能够为学习者提供基于数据的可解释反馈以及智能化的课程建议，这种结合不仅有助于学习者对其学习进度进行自我监管，还能显著提高学习效果、激发学习热情，以及增强学习者对AI技术的信赖和使用意愿（Adnan等，2022；Afzaal等，2021，2024；Conati等，2021）。具体来说，Afzaal等（2021）的研究使用了一种基于可解释AI的方法，为学习者

自动提供有关其学习表现的反馈和建议，并深入剖析学习成绩变化背后的原因。实验结果表明，这种方法促进了学习者的自我调节，进而显著提高了其学习成绩。在医疗领域，AI反馈可解释性显得尤为关键，因为在该领域，难以解释的AI反馈可能会导致医生和患者的不信任或不接受，从而影响医疗诊断的准确性（Cadario等，2021）。例如，Das等（2023）的研究发现，肺科专家在可解释AI反馈的帮助下，在解释肺功能测试时显著提高了诊断准确率；更重要的是，肺科专家还会根据可解释AI的反馈更新他们的决策，从而进一步提升诊断准确率。

（二）AI反馈时间

AI反馈时间，或称为“AI反馈延迟”，是指个体进行某项行为与其接收到AI提供的关于此行为的反馈之间的时间间隔（Ilgen等，1979）。与传统的人类反馈相比，AI的一大优势在于其持续运行的能力，AI不受时间和地点的限制，从而能够实时、自动地为用户提供所需的反馈。这一即时性使得AI在各种应用场景中都具有巨大的潜力。随着数字化的深入推进，现代组织中的工作和团队沟通的方式都在发生变化。在此背景下，AI的即时反馈能力为解决组织中的复杂问题提供了关键支持，能够促进组织成员的学习、职业发展和绩效提升（Chen等，2022b；de Laat等，2020；Tong等，2021）。以教育领域为例，Fidan和Gencel（2022）通过为期六周的实验发现，相比于人类教师反馈，得到AI即时反馈的学生能够获得更多关于课程的详细反馈，这有助于刺激他们的认知活动，促进他们的自主学习，从而提升他们的学习成绩。在远程和异步学习环境中，实时反馈尤为关键。在传统的教学中，教师可能难以与学生进行频繁、实时的互动，但AI技术的介入有助于填补这一空白。AI技术不仅有助于提高教师的授课质量，还能提升学生的学习效果（Al Hakim等，2022；Chien等，in press；Kang等，2023；Lee等，2023）。例如，Kang等（2023）通过研究11名学生和4名老师为期两周的在线舞蹈教学发现，AI导师能够给老师和学生提供及时、充分的反馈，有助于促进学生的学习反思，并通过多模态反馈资源帮助学生提高学习表现。类似地，AI即时反馈在书面文本创作上也展现出巨大潜力。如Godwin-Jones（2022）所述，基于自动写作评估AI辅助系统的即时反馈能够显著地帮助学习者提高书面文本写作质量。更为重要的是，AI即时反馈能够帮助学生规避因人类教师反馈而产生的焦虑和认知负荷，进而提高学生的学习意愿和学习投入以及最终的学习效果（Chen等，2022a；Lee等，2022；Ouyang等，2023）。

（三）AI反馈多样化

与人类相比，AI凭借其先进的智能算法和卓越的数据处理能力，能够更为深入和全面地分析各种任务活动。这使得AI不仅可以提供更加精确的反馈，还能给出多种不同的、具有针对性的解读和建议（Kiyasseh等，2023；Parent-Rocheleau和Parker，2022；Tong等，2021）。这种多维度和多层次的分析为用户提供了更为宽广和深入的洞察，从而有助于他们更好地理解和执行任务。基于此，本文主要从个性化（personalization）和丰富性（richness）这两个关键维度来探讨和总结AI反馈多样化相关文献。

1.AI反馈个性化

AI反馈个性化是指AI能够根据个体的实际情况和具体情境提供定制化的反馈回应。相比于标准化反馈，这种有针对性的反馈通常更具相关性和实用性（Marafie等，2021）。得益于大数据处理和算法训练，AI能够为行动者生成精准、具体的行动建议。这种定制化的反馈不仅有助于行动者更加清晰地了解自己的表现（Adnan等，2022；Hwang等，2023；Suraworachet等，2023），还可以协助组织更高效地满足客户需求，从而提升绩效水平（Kim等，2022）。例如，在教学场景中，由于教师的时间和精力有限，尤其在学生人数众多的情况下，教师难以完全根据学生的具体情况提供有针对性的反馈，而AI可以辅助教师实现个性化反馈（Conati等，2021；

Chen，2024；Belcadhi，2016；Jacobs等，in press；Hsia等，in press）。在写作教学中，由于涉及自我调节、创意发展和深入反思，个性化反馈显得尤为重要。Liu等（2023）的研究表明，与传统AI提供的标准化反馈相比，个性化AI反馈能有效增强学生的自我效能感，降低认知负荷，并显著提高其英语写作水平。类似地，Suraworachet等（2023）的研究也证实，AI个性化反馈有助于提升学生的学习投入和表现。Hwang等（2023）发现，针对学生原创文稿，AI提供的个性化修改建议能够显著提高文稿质量，从而帮助学生获得更高的成绩。

2.AI反馈丰富性

AI反馈丰富性是指AI能够从多个维度深入地为用户提供全面、详尽的信息。这种多维度的详尽反馈源于AI的技术能力，它可以利用数据分析和算法等先进技术，对任务的多个层面进行全方位的解读，从而为用户提供有深度和针对性的信息和建议。相较于人类反馈，AI反馈通常更为系统和细致，能够在短时间内对大量信息进行分析，从而给出更为全面的建议（Cold等，2024）。在有人类反馈的场景下，AI不仅可以补充或验证人类的判断，还能在必要时为现有的反馈提供进一步的细节和标准，从而使其更为完备和具体（Ghafouri等，in press；Lee，2023）。例如，魏爽和李璐遥（2023）发现，相比于人类教师，AI能够为学生提供更为全面的写作反馈，帮助学生快速发现写作问题并及时修正。然而，过于依赖AI的丰富反馈也可能带来一些问题。Drewery等（2022）发现，现有的AI提供了大量的战术反馈和较少的战略反馈，导致使用AI反馈的学生在简历修改中处于劣势，进而阻碍了个体对AI的接受。

值得注意的是，反馈的丰富性并不总是一个优势。由于人类的认知处理能力有限，过于复杂或详细的反馈可能会导致信息过载，反而会对接收者产生不利影响。Luo等（2021）的研究印证了该观点，他们发现，尽管AI能为员工提供丰富的工作表现反馈，但这种反馈对于绩效中等的员工可能更有益，而对于绩效较低的员工来说，过多的信息容易给他们带来信息过载问题，导致他们压力过大，从而影响其绩效提升潜能。

（四）小结

综上所述，AI反馈的配置效应研究主要从反馈质量（即反馈准确性、可靠性和可解释性）、反馈时间、反馈多样化（即反馈个性化和丰富性）三个方面进行探索。这些研究旨在揭示AI在客观的技术层面如何对反馈对象产生影响，其中主要的结果变量包括认知（例如，动机、自我效能感、认知负荷等）、态度（例如，信任、反馈接受度、学习投入度等）、行为（例如，反馈接受行为、纠偏行为等）以及绩效（例如，工作绩效、任务绩效、学习效果等）（Ötting等，2022；Pereira等，2023）。可以看出，尽管大量研究关注该问题，但学界对于从配置效应角度解释AI反馈的影响仍未达成共识。一些研究发现，得益于AI的自动数据处理和自我学习功能，在组织中部署AI能对组织和个体产生积极影响（例如，Das等，2023；Lee等，2022；Tong等，2021）。但也有研究指出，依赖AI反馈可能加重个体的认知负荷，给个体带来更多压力，从而对组织和个体造成消极影响（例如，Luo等，2021）。这些矛盾的结论可能源于人类的有限认知能力和AI反馈的丰富性之间的天然冲突，即使AI能提供高质量的反馈，人们也可能因为信息过载而难以充分适应和吸收AI反馈。可以看出，这类研究把AI这一新兴工具纳入传统反馈研究体系，并深入探讨了AI独特性对反馈有效性的影响，从而揭示了不同AI反馈特性对个体的认知、态度和行为的影响机制。此外，其他如AI特征（例如，拟人化：Yam等，2021）、反馈特征（例如，丰富性：Luo等，2021）、人类特征（例如，任期：Tong等，2021）和情境特征（例如，竞争vs.合作：Lei和Rau，2023）都可能作为调节变量影响AI反馈的效果。总的来说，该领域的研究从技术视角拓展了我们对AI反馈如何影响组织和个体绩效的理解。

三、AI反馈的披露效应相关研究

AI反馈的披露效应相关研究主要关注AI作为反馈提供者，如何引起接收者的主观心理层面变化，进而对接收者产生影响（Luo等，2019；Tong等，2021）。根据心智感知理论（theory of mind perception）（Gray等，2007），个体对各类实体（例如，人、动物、AI等）所具有的心理能力的感知可以分为能动性（即有意识地思考、计划和行动的能力）和感受性（即感受到积极和消极情绪、感知到痛苦和愉快的能力）两个独立的维度。这种心智感知会进一步影响个体对它们的评估和反应（Gray和Wegner，2012；Roesler等，2021；Schein和Gray，2018；Yam等，2022）。基于此，现有研究主要从能动性和感受性两个方面分析个体对AI反馈和人类反馈的主观感受差异，并发现这种感受会进一步影响他们对AI及其反馈的认知和态度，进而带来特定的行为反应。相比于人类反馈，AI反馈往往被认为更加客观，因为它基于历史数据并且遵循固定的算法和规则（Lindebaum和Ashraf，2024）。值得注意的是，尽管AI在某些场合可能提供更优的反馈，但一些研究发现人们在许多情况下对AI存在厌恶情绪（即AI厌恶，artificial intelligence aversion）（例如，罗映宇等，2023；Dietvorst等，2015；Luo等，2021）。然而，另一些研究则指出，在某些情况下（例如，反馈接收者是外行人或非专家），个体可能会更偏好AI的建议（即AI偏好，artificial intelligence appreciation）（Logg等，2019）。基于此，为了对该领域进行系统的回顾，本文将结合三个主要方面："能动性vs.感受性""客观性vs.主观性"和"AI厌恶vs.AI偏好"，来梳理AI反馈披露效应的中介机制及其影响（见表2）。

表 2　AI反馈的披露效应研究总结

类别	亚类别	结果变量	文献数量	参考文献
能动性 vs. 感受性	感知能动性	认知：自我评价、AI评价、自我效能感等 态度：工作满意度、反馈接受度、同理心、情绪等 行为：报复行为、环保行为等 绩效：工作绩效等	7	Ghazali等，2019*；Hall等，2022*；Lee等，2019*；Selenko等，2022*；Thuillard等，2022*；Tussyadiah和Miller，2019*；Yam等，2022*
	感知感受性		5	Akalin等，2019*；Lee等，2019*；Lei和Rau，2023*；Min等，in press*；Sharma等，2023*；
客观性 vs. 主观性	责任归因	态度：反馈接受度、情绪等 行为：报复行为等	2	Horstmann等，2021*；Lei和Rau，2021†
	意图		3	Garvey等，2023*；Pei等，in press*；Yam等，2022*
AI厌恶 vs. AI偏好	AI厌恶	态度：信任、算法态度等 绩效：工作绩效、服务绩效等	2	Parenti等，2023*；Tong等，2021*
	信任		3	Abendschein等，in press†；Filiz等，2021*；Kim等，2022*

注：*组织管理情境，†教育情境，‡医疗情境。

（一）能动性vs.感受性

根据心智感知理论，个体对其他实体的能动性和感受性的感知会影响其对该实体的态度和行为（Gray等，2007；Gray和Wegner，2012；Yam等，2021，2022）。人们在感受到一个实体有较高的能动性时，更倾向于认为该实体具有较强的自主性，能够有意识地思考、计划和行动（Schein和Gray，2018）。对于被感知为具有较高感受性的实体，人们往往会表现出更多的同情和关心，认为伤害它是不道德的（Gray和Wegner，2009；Schein和Gray，2018）。因此，从心智感知理论视角来看，面对人类和AI提供的关于同一行为的反馈，由于对能动性和感受性的不同感知，个体可能会产生不同的反应。基于此，我们从感知能动性（perceived agency）和感知感受性（perceived experience）两个角度来总结现有文献。

1.感知能动性

AI能通过补充任务、替代任务或生成新任务等部署方式来增加或抑制个体从任务中获得的认同感，从而影响个体对AI反馈的能动性感知（王欣等，2021；Selenko等，2022）。在组织中，仅仅感受到AI存在已足以引导个体的行为，并促进他们的团队协作意愿（Tussyadiah和Miller，2019）。AI通过提供社交线索可以增强人类对其能动性的感知，因此当AI反馈包含更多的互动性和社交性时，人们的抗拒感将降低，接受度将提高（Ghazali等，2019）。Hall等（2022）发现，当销售人员对AI反馈有更高的感知准确性时，他们使用AI反馈的意愿会提高。然而，AI反馈的感知能动性并不是越高越好（Lee等，2019）。例如，Yam等（2022）发现，在面对AI负面反馈时，AI的拟人化程度越高，个体对其能动性感知就越高，就越倾向于对AI做出报复行为，进而造成不良影响（王海忠等，2021）。类似地，Thuillard等（2022）的研究表明，相比于人类反馈，消极的AI反馈被认为更不公平，这可能激发工作场所的反生产行为，从而降低工作绩效。

2.感知感受性

感知感受性包括个体对其他实体的个性、情绪和同理心等方面的认知（Gray等，2007）。现有研究发现，AI拟人化有助于增强个体对AI感受性的感知（Yam等，2021），并通过各种视觉和语言线索提升其拟人化特征（Lee等，2019；Min等，in press；Yam等，2021，2022）。例如，Lee等（2021）的研究表明，人们会根据AI的外在行为表现形成对AI人格特质的主观感知。具体地说，慢速的视觉反馈可能被看作是抑郁、焦虑或神经质的表现，而快速的视觉反馈则可能被认为是外向、有创造性和乐观的表现。类似地，Akalin等（2019）发现，提供赞扬的AI被认为更具感受性，因此更容易得到个体喜爱，相反，提供负面反馈的AI往往不被喜爱。同时，AI反馈带来的感知感受性差异也能引起个体情绪、公平感和同理心的变化。例如，Lei和Rau（2023）使用功能性近红外光谱扫描仪监测参与者的前额皮质活动来调查在合作和竞争两种情境下，AI反馈对教育游戏中玩家客观情绪反应的影响，结果发现，积极的、合作性的任务中的AI反馈更容易激发玩家情绪，进而增加个体对AI反馈的积极评价。类似地，Min等（in press）发现，AI表现出热情开朗的特征，并以情感表达的方式进行反馈，能够显著增强个体的人际公平和信息公平感知。此外，Sharma等（2023）在健康咨询中也得出了相似的结论，他们发现使用AI提供感同身受的反馈信息能够提高心理工作者的同理心，并提高其自我效能感。

（二）客观性vs.主观性

个体对AI反馈客观性的认知是影响其主观心理感受的核心机制之一。相比于人类反馈，AI反馈往往有更高的客观性（Blair和Saffidine，2019），反馈的客观性往往导致个体对反馈产生不同的归因（Yam等，2022），并推断反馈提供者的意图（Garvey等，2023）。AI反馈是AI基于行动者的实际情况而提供的行动反馈，基于归因理论，人们对结果（例如行动反馈）会进行内部和外部归因。内部归因（internal attribution）指的是将原因归于某些内在特征，例如，能力、智力和努力等个人内部因素，而外部归因（external attribution）指的是将原因归于某些外部情境，例如，运气、天气和环境等外部因素。归因能够直接影响个体的认知和情感反应，这些归因方式能够直接引起个体对AI反馈的不同理解和反应。因此，部分研究从归因视角，特别是责任归因和意图两个方面，对AI反馈进行深入研究。

1.责任归因

反馈的类型会影响个体的归因方式，特别是当遭遇不利反馈时，个体更可能将不利反馈归咎于反馈提供者而不是自己（Dello Russo等，2023；Lechermeier和Fassnacht，2018），这种现象在AI反馈中同样普遍存在（Yalcin等，2022）。例如，Horstmann等（2021）在一项实验室研究中发

现，在接收到负面反馈时，个体更容易将责任归因于AI代理。类似地，Lei和Rau（2021）在将AI设计成能够为个体提供舞蹈练习反馈的舞蹈导师后发现，个体与内归因AI导师（AI将负面反馈结果归因于自身）的关系比与外归因AI导师（AI将负面反馈结果归因于人类）的关系更亲密，前者也更容易被视为培训教练而非培训工具。

2.意图

由于AI本身的技术属性，其反馈相较于人类反馈更可能被视为是客观无意图的（Garvey等，2023；Gray等，2007）。尽管如此，人们仍然倾向于去推断行为主体的意图，因为这有助于个体评估并解释他们的行为，进而影响个体的态度和行为反应。例如，Garvey等（2023）发现，对于正面反馈，个体更倾向于接受来自人类的而非AI的，因为人类正面反馈的意图被推测为仁慈的，而AI是无意图的；然而，对于负面反馈，个体更倾向于接受来自AI的而非人类的，因为人类的负面反馈可能被推测为出于自私的动机，而AI则被认为是公正的、无私的。类似地，Yam等（2022）的研究也发现，与非拟人化的AI主管相比，拟人化的AI主管提供负面反馈时更可能被视为拥有主动性（或意图），进而导致个体对反馈提供者做出更多的报复行为。因此，Pei等（in press）提出，基于AI的反馈系统可以作为一种“补救”工具，有效地减轻员工对接受负面反馈的担忧，从而提升员工绩效。

（三）AI厌恶vs.AI偏好

现有研究表明，一旦向个体披露反馈由AI而非人类提供，就很可能导致个体的AI厌恶，这不仅可能降低个体对AI反馈的信任度，还可能对其后续态度和行为产生影响（Luo等，2019；Tong等，2021）。基于此，我们从AI厌恶和信任两个方面来回顾现有文献。

1.AI厌恶

AI厌恶并不是指个体对AI技术的直接反感，而是他们选择不采纳或不使用已有的AI技术和建议（Castelo和Ward，2021；Dietvorst等，2015；Glikson和Woolley，2020）。这种情境下AI厌恶可能导致个体对AI所给出的反馈持怀疑或抗拒的态度。对组织来说，这意味着即使有了大量的技术投入，也可能无法获得预期收益或带来所期望的价值（Kim等，2022）。因此，现有研究主要集中于探索相关策略和方法，以缓解AI厌恶所带来的负面影响（Dietvorst等，2018；Filiz等，2021）。例如，Filiz等（2021）发现，通过有针对性的学习和反馈机制，人们可以逐步提高对自己预测能力的估计，从而有效减轻对AI反馈的厌恶感。在实验中，当参与者在连续的股价预测挑战（总共40轮）中每次都得到明确的反馈和与之相关的金钱激励时，他们对AI的厌恶程度显著降低。

2.信任

将AI整合和运用到组织中的关键取决于组织成员对AI的信任（Glikson和Woolley，2020）。由于人们倾向于将AI视为没有感情和自主思维的工具（Lanz等，2024），因而更倾向于认为AI缺乏情绪或道德感（Bigman和Gray，2018；Gray等，2007），从而对AI反馈缺乏足够的信任。例如，当个体知道反馈由AI提供时，个体对反馈的信任度就会下降，从而阻碍绩效改善（Tong等，2021）。为了解决这一问题，目前许多研究致力于解决如何提高人们对AI反馈的信任问题（Abendschein等，in press）。例如，Parenti等（2023）的研究发现，AI的表现越类似于人类，AI反馈就越能得到个体的信任。

（四）小结

综上所述，现有研究对AI反馈披露效应的研究主要集中在三个维度：“能动性vs.感受性”“客观性vs.主观性”以及“AI厌恶vs.AI偏好”。相关研究主要从这三个方面探索了AI反馈对个体

主观心理感知的影响机制，主要结果变量聚焦在认知（例如，自我评价、自我效能感等）、态度（例如，工作满意度、反馈接受度、情绪等）、行为（例如，报复行为、环保行为等）和绩效（例如，工作绩效、服务绩效等）四个方面（Ötting等，2022；Pereira等，2023）。可以看出，对于AI反馈的披露效应目前仍没有得出一致的结论。部分研究发现，披露AI作为反馈提供者会对个体产生负面影响：一方面，个体的AI厌恶使其更偏好人类反馈而非AI反馈（例如，Tong等，2021；Yam等，2022）；另一方面，个体倾向于将AI视为类人的实体，当接收到负面反馈时，更容易将负面反馈归因于AI而非自己，从而导致负面影响（例如，Horstmann等，2021；Lee等，2019；Thuillard等，2022）。另一部分研究则发现，披露效应也可能带来积极的影响，例如，AI采用拟人化的反馈方式能促进个体对反馈的接受（例如，Lee等，2019；Sharma等，2023），甚至在某些情况下，个体更愿意接受AI而非人类的消极反馈（Garvey等，2023）。上述不一致研究发现的产生原因可能在于个体对AI的偏好（例如，AI厌恶vs.AI偏好）有差异，不同类型的AI反馈导致个体不同的归因（例如，内部归因vs.外部归因），而个体的主观心理感受又受到多种因素的影响（例如，蒋路远等，2022；Glikson和Woolley，2020；Roesler等，2021）。可以看出，这一系列研究从AI反馈如何影响个体的主观心理感受这一角度探索了AI反馈的有效性话题，进而解释了AI反馈给个体带来不同主观心理感受的过程和机制，以及对个体绩效和行为结果的影响。不同的AI特征（例如，AI类型：Horstmann等，2021）、反馈特征（例如，正面vs.负面反馈：Akalin等，2019）、人类特征（例如，爱面子：Pei等，in press）和情境特征（例如，竞争vs.合作：Lei和Rau，2023）在上述不同的披露效应机制中起调节作用，并对结果变量产生影响。该类研究拓展了我们对AI反馈如何通过个体主观心理机制来影响组织和个体产出这一问题的理解。

结合以上研究分类和文献梳理，本文总结了AI反馈配置效应和披露效应的作用机制、边界条件及影响，如图1所示。

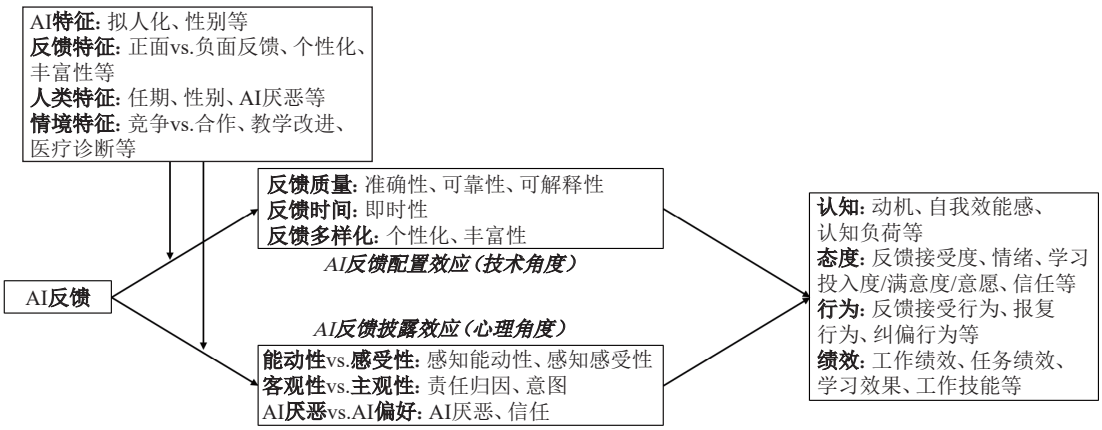

图 1　AI反馈配置效应和披露效应的作用机制、边界条件及影响

四、理论机制

在前文文献梳理的基础上，我们进一步提炼了与AI反馈配置效应和披露效应相关的理论机制。值得注意的是，由于目前对AI反馈的研究较为零散，且不同学科对理论的应用程度有较大差异，现有研究中明确使用理论建立和阐释的假设仍相对较少。因此，本文重点梳理了AI反馈两个效应相关研究中的共同逻辑及其背后的理论机制，并探索了潜在的未来研究方向。基于此，本文从认知和学习两个视角出发，重点梳理了常用或具有深远启示意义的理论机制（见表3），以期对未来研究有所启发。

表 3　AI反馈的主要理论总结

理论视角	理论名称	理论阐释	理论应用举例
认知视角	反馈过程理论	反馈是一种重要的组织和个人资源，反馈过程由一系列连续的认知步骤组成，包括反馈刺激、感知、接受和回应（Ilgen等，1979）	Parenti等（2023）发现，机器人表现出类人特征时能够影响个体的决策过程，并提升个体对机器人的信任度
	社会认知理论	个体的认知、行为及所处环境是一个动态的系统，个体的活动是认知、行为和环境三个变量不断相互作用的函数（Bandura，1986）	Nazari等（2021）发现，AI反馈能提升个体的认知投入、情感投入和行为投入，从而提升个体的学习效果
	社会信息加工理论	人的认知过程就是对信息的加工过程，遵循“输入—内部信息加工—输出”的模式（Salancik和Pfeffer，1978）	Luo等（2021）发现，低绩效销售人员难以处理全面而丰富的AI反馈信息，容易出现信息过载问题，进而阻碍其通过学习来改善绩效
	归因理论	人们如何对自身及他人的行为进行因果解释（Kelley，1967，1973；Weiner，1985，2010）	Horstmann等（2021）发现，人类面对AI负面反馈时，对AI的感知代理权和感知意图越低，对AI的责任归咎就越少
学习视角	自我调节学习理论	通过让学习者了解自己的学习状态、控制自己的学习过程以及应用认知策略来支持学习者实现预期学习目标（Zimmerman，1990）	Afzaal等（2021）发现，AI能够通过提供自动和智能的反馈和建议来支持学习者以数据驱动的方式进行自我调节学习，进而提升学习者的学习表现
	双循环学习理论	学习者经历获取知识、接受学习任务、接受反馈、重新考虑学习行为，并根据学习过程中的反馈调整学习策略的过程（Argyris，1990）	Liu等（2023）发现，AI反馈能够促进反思性思维机制的学习，不仅能够有效提升学习者的学习成绩，还能提升学习者的自我效能感，降低认知负荷
	经验学习理论	学习由具体体验、抽象概念化、反思观察和积极实践等基本结构构成（Kolb，2015）	Kang等（2023）发现，AI舞蹈教师能够帮助学习者熟悉给定的舞蹈动作，通过多模态反馈促进学习者的反思，进而提升其学习效果

（一）认知视角

1.反馈过程理论

反馈过程理论（feedback processing theory）（Ilgen等，1979）认为反馈是一种重要的组织和个人资源，反馈过程由一系列连续的认知步骤组成，包括反馈刺激、感知、接受和回应。个体对反馈的认知取决于反馈的性质、来源、传递方式、接收者、外部情境等诸多因素的共同影响（Hattie和Timperley，2007；Mercer和Gulseren，2024；Parenti等，2023）。与人类不同，AI基于机器学习和数据分析等技术生成反馈。

基于反馈过程理论，当AI作为反馈来源时，AI以及AI反馈的特性会直接作用于反馈过程，从而影响反馈接收者（即人类）（梁肖梅等，2023）。因此，反馈过程理论可以广泛解释并应用于AI反馈对个体认知和行为的影响（Thuillard等，2022；Tong等，2021）。例如，Hall等（2021）的研究发现，AI反馈特性以及人类个体差异可以影响个体对AI反馈的感知准确性，进而影响对AI反馈的使用意愿。具体来说，当AI提供频繁、具体和积极的反馈，或者当个体的反馈导向（即个体对反馈的整体接受度）（London和Smither，2002）较高时，个体对AI反馈的感知准确性更高，进而更有意愿使用AI反馈（Parenti等，2023）。

2.社会认知理论

社会认知理论（social cognition theory）（Bandura，1986）强调将个体的认知、行为及所处环境放在一个动态的系统中进行考察，认为个体的活动是认知、行为和环境三个变量不断相互作

用的函数(Bandura,2001)。其中任意两个因素之间都存在双向关系,这种双向关系的强度和模式会随着个体认知、行为和环境的不同而发生变化。其中,自我效能感是社会认知理论的核心概念,是指人们对自己组织和实施某种行为进而达到期望效果的能力的自我判断(Bandura,1986)。

基于社会认知理论,AI作为社会发展和技术创新的新兴产物,对组织和社会环境产生了深刻的影响,AI对环境的改变又会作用于人的认知和行为(Friedenberg和Silverman,2012;Yam等,2023;Yoganathan等,2021)。因此,社会认知理论可以用于解释AI反馈作为一种新的环境信息对个体认知和行为的影响,尤其是在AI反馈配置效应的机制解释中,能够解释配置AI反馈对于个体认知和行为的影响机制。例如,Nazari等(2021)的研究发现,AI的即时反馈为学习者提供了准确且有意义的学习评估,不仅增强了学习者的自我效能感,而且在积极反馈情境下,进一步提升了学习者的主动性。

3.社会信息加工理论

社会信息加工理论(social information processing theory)(Salancik和Pfeffer,1978)认为,人的认知过程就是对信息的加工过程,遵循"输入—内部信息加工—输出"的模式,力图建立心理活动的计算机模型。该理论涉及人如何注意、选择和接收信息,如何对信息进行编码和组织,以及如何利用信息做出决策和行动等。在社会交往和社会认知领域,个体通过对特定的社会信息进行加工和解读,来决定采取怎样的态度和行为(Salancik和Pfeffer,1978)。

根据社会信息加工理论,个体的认知过程和认知特征能直接影响其对AI反馈的注意、接收和处理;同时,AI反馈作为一种环境信息,会直接或间接地改变个体的信息加工过程。因此,社会信息加工理论可以用于解释个体如何接收、处理和理解AI反馈的信息,以及个体对AI反馈的解读又如何影响其态度和行为(Bauer等,2023)。例如,Luo等(2021)的研究发现,由于人类信息加工能力的有限性,低绩效销售人员在面对全面而丰富的AI反馈信息时会面临信息过载问题,这会阻碍其通过学习来改善绩效。

4.归因理论

归因理论(attribution theory)阐明了人们如何对自身及他人的行为进行因果解释(Kelley,1967,1973;Weiner,1985,2010)。根据归因理论,人们将结果归因于内部因素(例如,能力和努力)或外部因素(例如,运气和环境),从而影响其认知和情感。AI反馈提供基于行动者实际情况的信息,个体会根据自我感知和推断对AI反馈进行归因和解释,进而影响其态度和行为。

根据归因理论,AI反馈作为一种行为结果信息,能够直接影响个体对AI反馈的归因过程。因此,归因理论能够用于阐述个体如何解读AI反馈的原因,以及个体对AI反馈的主观性或客观性的感知,从而解释个体相应的态度和行为反应。例如,受自我服务归因偏差(self-serving bias)(Miller和Ross,1975)的影响,个体倾向于将积极结果归功于自身的良好表现,而将消极结果归咎于AI(王海忠等,2021;Dello Russo等,2023;Lei和Rau,2021;Yalcin等,2022)。此外,对AI反馈的归因与行动者的代理权、自愿性、意图等有关(Garvey等,2023;Yam等,2022)。例如,Horstmann等(2021)的研究发现,人类面对AI负面反馈时,对AI的感知代理权和感知意图越低,对AI的责任归咎就越少。

(二)学习视角

1.自我调节学习理论

自我调节学习理论(self-regulated learning theory)是指通过让学习者了解自己的学习状态、控制自己的学习过程以及应用认知策略来支持学习者实现预期学习目标(Zimmerman,1990)。基于该理论,学习者需要了解他们的当前状态,并充分了解如何应用认知策略来帮助他

们实现预期的学习目标。一项元分析表明，应用自我调节学习策略有助于学习者实现成果学习的预期目标（Hattie，2008）。

现有研究探索了通过AI生成即时和个性化的反馈，以数据驱动的智能方式支持学习者的自我调节学习，提升其学习表现（Afzaal等，2021；Hsia等，in press；Suraworachet等，2023）。因此，该理论能够用于解释AI反馈如何影响个体的自我调节学习过程。例如，Afzaal等（2021）提出了一种可解释的基于人工智能的方法来提供自动和智能的反馈和建议，用于支持学习者以数据驱动的方式进行自我调节学习，进而促进了学习者的学习表现。

2.双循环学习理论

双循环学习理论（double loop learning theory）指的是学习者经历获取知识、接受学习任务、接受反馈、重新考虑学习行为，并根据学习过程中的反馈调整学习策略的过程（Argyris，1990）。双循环学习理论为学习者提供了一个额外的学习阶段，不仅可以有效提高学习者的表现，而且能够促进学习者的学习反思（Liu等，2023）。

根据双循环学习理论，AI反馈支持学习循环中的反思性思维促进机制，为学习者提供了更多的学习和反思机会，提升了学习者的自我效能感和自我调节学习能力，最终提升了学习绩效（例如，写作质量、学习成绩）（Liu等，2023；Suraworachet等，2023）。因此，该理论能够用于解释AI反馈如何影响个体的双循环学习过程。例如，Liu等（2023）提出一种基于反思性思维促进机制的学习方法，不仅能够有效提升学习者的学习成绩，还能提升学习者的自我效能感，降低认知负荷。

3.经验学习理论

经验学习理论（experimental learning theory）指出，学习由具体体验、抽象概念化、反思观察和积极实践等基本结构构成（Kolb，2015）。基于该理论，学习过程中需要先向学习者提供以观察和模仿为基础的教育体验（即具体经验），然后学习者尝试命名各个学习部分（即抽象概念化）并回顾引导者或同行者的动作（即反思观察）来消化知识点，进而进行积极实践。

基于经验学习理论，通过将AI—人类合作的即时反馈用于定量评价学习者和引导者表现之间的相似性，能够有效引起学习者的自我反思，促进经验学习过程。因此，该理论能够用于解释AI反馈如何通过经验学习的方式来促进学习者的反思过程。例如，Kang等（2023）使用AI舞蹈教师辅助远程教学，发现AI导师能够帮助学习者熟悉给定的舞蹈动作，通过多模态反馈促进学习者的反思，进而提升其学习效果。

（三）小结

综上所述，AI反馈的研究主要基于认知和学习两种理论视角。其中，认知理论视角通过研究个体对AI以及AI反馈的认知，来解释AI反馈对个体的影响机制，主要涉及的理论包括反馈过程理论、社会认知理论、社会信息加工理论和归因理论等。具体来说，反馈过程理论描述了反馈的连续认知步骤，并分析了AI作为反馈来源的角色及其影响；社会认知理论强调AI反馈与个体认知和行为之间的动态交互关系；社会信息加工理论通过信息加工的过程来解释认知，指出AI反馈的影响会受到个体信息处理能力的制约；归因理论则通过原因推断机制探讨了AI反馈对个体的影响和个体对其的主观评价。学习理论视角将AI视为一种学习辅助工具，探讨AI反馈对个体自我调节学习、循环学习以及经验学习的影响，认为AI反馈可以通过提升个体的认知能力和实践能力来辅助个体学习。具体来说，自我调节学习理论强调了学习者自我调节的重要性，AI反馈能提升学习者的自我调节能力；双循环学习理论指出AI反馈能够有效帮助学习者在不同学习阶段取得进步；经验学习理论认为，AI能够促进学习者经验学习的全过程。

五、AI反馈未来研究方向

通过对现有AI反馈文献的全面回顾和对理论机制的梳理，可以发现围绕AI反馈主题在不同的学科领域已经产出了一些有意义的研究成果，并得到了学术界和实践界的广泛关注。然而，现有AI反馈相关研究仍较为松散，缺乏系统性，也存在较多研究不足。本文在梳理现有文献的基础上，总结并提出亟待探讨的未来研究方向，旨在为探索AI反馈相关研究提供有价值的启发。

（一）探索复杂情境中AI—人类协同反馈的影响机制

从研究对象来看，现有AI反馈研究虽然关注了AI作为单一反馈主体以及AI和人类提供反馈的差异等问题，但是仍缺乏对AI—人类协同反馈等复杂反馈情境的探讨。在实际工作场景中，反馈来源具有多样性（例如，领导、同事、顾客等），而AI在组织中的引入进一步增加了这种复杂性，反馈实际上是AI和人类共同合作、相互补充的结果（王欣等，2021）。事实上，AI和人类协同反馈的现象在实践中已经不断地涌现（张志学等，2024；Jia等，2024；Sharma等，2023），因此，考虑AI—人类协同反馈对反馈接收者产生的不同影响显得尤为重要。例如，Bulten等（2021）发现，在医疗反馈方面，AI与病理学家协作诊断的表现优于无AI辅助的病理学家和独立的AI系统。此外，反馈不仅仅是单次的，也可能是多次反馈，这种反馈可能对个体造成长期、动态的影响，不同反馈者的反馈顺序也可能产生不同的影响（Lechermeier和Fassnacht，2018）。例如，Jiang等（2023）发现，在招聘决策过程中，相比于AI—AI决策顺序，AI—人类决策顺序会使被试感受到招聘决策具有更高的程序公平性。因此，探讨AI—人类协同反馈对反馈接收者的影响，目前仍需要更多的研究从不同的角度进行。综上，通过对AI—人类协同反馈的探索，有助于更加全面地理解AI—人类协同反馈对反馈接收者的影响机制，进而推动AI反馈的理论研究和实践发展。

（二）探索不同类型的AI反馈对反馈接收者的影响

不同类型的AI反馈对反馈接收者的影响有着明显差异，现有研究主要集中于特定类型的反馈，缺乏对多种AI反馈类型的对比研究。例如，目前的研究主要关注AI负面反馈，对AI正面反馈关注不足（例如，Tong等，2021；Yam等，2022）。然而，由于正面反馈能对个体产生行为强化等重要作用，因此，相比于人类，缺乏情感的AI如何为个体提供正面反馈，以及如何激励反馈接收者提升任务表现，是未来研究需要探讨的一个重要话题。此外，反馈类型还体现在反馈风格（信息vs.控制）、反馈形式（文字vs.语音vs.动作）和反馈目标（过程vs.结果）等方面（Lechermeier和Fassnacht，2018），但关于AI如何依据这些分类提供反馈并影响接收者的研究仍然很少。以反馈形式为例，虽然语音反馈有助于提升AI的拟人化程度，进而提升人类对AI的信任程度（Gray等，2007），但是在披露效应的影响下这也可能导致人类难以接受AI反馈（Tong等，2021；Yam等，2022），换言之，在某个具体情境中采取不同反馈形式（例如，文字vs.语音vs.动作等）的影响机制仍有待探索。综上，探索不同类型的AI反馈不仅有助于更加全面地认知和理解AI反馈对个体、组织和社会的影响，还能为管理实践提供更全面和有效的指引，从而促进AI反馈的应用和发展。

（三）探索AI反馈的情绪影响机制

尽管AI反馈研究已深入探索了其如何通过认知过程（例如，可解释性）影响个体的行为，但这些研究对情绪机制的探索仍然非常有限。实际上，反馈不仅可以影响个体的认知，还可以影响其情绪（Lechermeier和Fassnacht，2018）。对AI反馈的情绪影响机制缺乏研究意味着我们对AI反馈的理解还不够全面，也不利于对AI反馈形成更为系统的认识。尽管AI通常被视为缺

乏情感和情绪的存在(Gray等,2007;Gray和Wegner,2012),但情绪(尤其是积极情绪)在实体拟人化过程中扮演着重要角色(Yam等,2022)。更重要的是,随着AI越来越多地作为人类的“同事”而非“工具”运用于工作场所(Traeger等,2020),其对人类情绪的影响预计在未来会变得越来越突出(王海忠等,2021;Scassellati等,2018)。这意味着,AI反馈对反馈接收者的情绪影响机制可能是AI反馈研究的重点和难点,未来研究有必要系统地探讨AI反馈的情绪影响机制。例如,现有研究表明,AI如同人类一样,能使个体产生压力情绪,从而导致从众行为(Qin等,2022;Vollmer等,2018)。类似地,在抑郁症治疗中,AI反馈更能激发病患的情绪反应,提升社交技能,从而达到较好的治疗效果(Scassellati等,2018)。综上,未来研究可以从情绪视角来进一步揭示AI反馈如何对反馈接收者产生影响。

(四)探索影响AI反馈配置效应和披露效应的边界条件

尽管目前对AI反馈的研究主要集中在配置效应的积极影响和披露效应的消极影响上,但这些研究依然存在矛盾性结论,且较少有研究关注配置效应的消极影响和披露效应的积极影响,这种研究的不平衡性和片面性限制了我们对AI反馈的全面理解。究其原因,主要在于现有研究尚未充分厘清两种效应的边界条件。事实上,在多数场景中,AI反馈的配置效应可能是积极的,而披露效应可能是消极的,但在某些边界条件下,这种常规趋势可能会被颠覆。以配置效应为例,不同绩效水平的员工对AI反馈的接受度和反应各不相同。例如,Luo等(2021)发现,尽管AI反馈对中等绩效员工具有积极的绩效改善影响,但对于低绩效员工的绩效改善效果有限。在披露效应方面,虽然个体可能会因为AI被披露为反馈提供者而厌恶所接收的反馈(例如,Tong等,2021),然而,这可能仅限于专家(Dietvorst等,2015),对于非专家而言,披露AI可能会导致AI偏好,从而对反馈接收者产生不同的影响(Logg等,2019)。综上,为了更深入、全面地理解AI反馈的配置效应和披露效应,我们需要进一步探索和明确其在不同情境下的作用,从而为整合现有研究成果并推进该领域的研究进程提供新的视角和方向。

(五)探索多方法、多文化等交叉研究设计来开展研究

由于AI的发展目前仍处于弱AI阶段,AI尚不能完全模拟人类,因此现有研究多依赖于情境实验来探讨AI反馈的影响(例如,Yam等,2022)。然而,情境实验虽然能排除其他干扰因素,证明变量间的因果关系,但是缺乏真实性,忽视了人类的情绪、人际关系等因素,这可能导致AI反馈的实际影响无法被真实地观测到,从而影响研究结论的可靠性和外部效度。这也是跨学科交叉研究的难点之一,即技术如何与心理学、管理学相结合来影响个体的态度和行为。因此,为了更深入地揭示AI反馈的实际效果,我们需要探索并采纳多种交叉学科研究方法,得出更加稳健和可靠的研究结论。例如,使用现场实验法来开展研究(例如,Tong等,2021;Luo等,2021)能够较为真实地还原现实工作情境,较好地排除其他因素的干扰,使结论更加可信;结合计算机科学和心理学来优化AI技术,从而提升AI反馈的能力、效率和接受度。此外,需要加强本土化和跨文化多情境比较研究,由于中西方文化传统和价值观等文化差异,中西方在对待AI的态度、认知等方面也存在较大差异(Yam等,2023)。例如,东方文化认为AI是人类的拓展,而西方文化认为AI是人类的威胁(Yam等,2023),这可能导致AI反馈研究在中西方文化情境下出现不同甚至完全相反的结论。因此,我们也需要关注跨文化研究,通过对比不同文化情境下的研究结论,将AI反馈研究推广到更为普适的情境中。综上,采用交叉研究设计有助于更全面地理解AI反馈的实际影响机制,特别是在多种文化背景下对反馈接收者的影响,从而为AI反馈研究的持续发展提供有力支持。

六、总　结

近年来，AI反馈在实践中得到日益广泛的应用，已经成为组织管理、教育和医疗等学科重要的话题之一。通过对已有文献的梳理和总结，本文首先厘清了AI反馈的概念内涵；然后基于AI反馈的配置效应和披露效应，对现有研究进行了总结归纳，系统梳理了AI反馈产生影响的结果变量、中介机制和边界条件，构建了AI反馈的研究框架（参见图1）；最后，系统介绍和总结了常用或具有深远启示意义的理论机制，并提出了具有科学价值和现实意义的未来研究方向，从而为探索AI反馈相关研究提供了有价值的洞见。

本研究有三个方面的重要理论贡献。第一，系统梳理了AI反馈研究的进展情况，并厘清了AI反馈的概念内涵和外延。第二，虽然现有AI反馈研究呈加速增长的趋势，但仍存在研究范式不统一、理论不清晰和机制未厘清等问题。本研究归纳和总结了AI反馈产生的配置效应和披露效应及其影响机制，丰富了AI反馈的内涵，为未来研究提供了清晰的理论视角和框架，并提出了五个具有科学价值和现实意义的AI反馈重要未来研究方向。第三，本研究系统介绍和总结了现有AI反馈研究常用或具有深远启示意义的理论机制，这些理论不仅能够为后续AI反馈研究提供理论参考，也能够为未来研究构建新理论提供启示。

本研究对管理实践也具有重要的启示价值。第一，组织可以通过识别AI反馈的配置效应，来更好地发挥AI反馈的积极影响。例如，AI反馈的准确性已经能达到甚至优于人类专家的水平（Tong等，2021；Zhao等，2023），因此，组织可以根据实际情况部署AI来评估成员的任务表现，促进组织发展。第二，组织需要关注AI反馈过程中披露效应可能产生的负面影响。例如，尽管AI在某些场合能够提供比人类更优的反馈，但人们却可能因为AI厌恶而选择不采纳AI的建议（Dietvorst等，2015；Glikson和Woolley，2020）。因此，组织需要采取适当措施消除披露效应的负面影响，例如，根据不同的情境来选择向员工披露反馈是由AI还是由人类提供（Garvey等，2023）。第三，组织需要及时关注AI反馈对员工情绪和行为的影响并进行相应改进。例如，绩效能力水平不同的员工对AI反馈的态度和反应有较大的差异（Luo等，2021），因此需要及时调整对不同类型员工的管理策略，确保AI反馈能够为组织和员工带来实质性的积极效益。

主要参考文献

[1]陈龙. “数字控制”下的劳动秩序——外卖骑手的劳动控制研究[J]. 社会学研究, 2020, 35(6): 113-135,244.

[2]蒋路远, 曹李梅, 秦昕, 等. 人工智能决策的公平感知[J]. 心理科学进展, 2022, 30(5): 1078-1092.

[3]李璨, 吕渭星, 周长辉. 绩效反馈与组织响应: 文献综述与展望[J]. 外国经济与管理, 2019, 41(10): 86-108.

[4]李胜蓝, 江立华. 新型劳动时间控制与虚假自由——外卖骑手的劳动过程研究[J]. 社会学研究, 2020, 35(6): 91-112,243-244.

[5]梁肖梅, 汪秀琼, 叶广宇, 等. 绩效反馈理论述评: 知识框架与研究展望[J]. 南开管理评论, 2023, 26(3): 197-212.

[6]刘善仕, 裴嘉良, 葛淳棉, 等. 在线劳动平台算法管理: 理论探索与研究展望[J]. 管理世界, 2022, 38(2): 225-239,14.

[7]罗映宇, 朱国玮, 钱无忌, 等. 人工智能时代的算法厌恶: 研究框架与未来展望[J]. 管理世界, 2023, 39(10): 205-227.

[8]王海忠, 谢涛, 詹纯玉. 服务失败情境下智能客服化身拟人化的负面影响: 厌恶感的中介机制[J]. 南开管理评论, 2021, 24(4): 194-204.

[9]王欣, 朱虹, 姜帝, 等. 人工智能产品“协助者”与“替代者”形象对消费者评价的影响[J]. 南开管理评论, 2021, 24(6): 39-49,139.

[10]王永丽, 时勘. 绩效反馈研究的回顾与展望[J]. 心理科学进展, 2004, 12(2): 282-289.

[11]魏爽, 李璐遥. 人工智能辅助二语写作反馈研究——以ChatGPT为例[J]. 中国外语, 2023, 20(3): 33-40.

[12]魏昕, 黄鸣鹏, 李欣悦. 算法决策、员工公平感与偏差行为: 决策有利性的调节作用[J]. 外国经济与管理, 2021, 43(11): 56-69.

[13]张志学, 华中生, 谢小云. 数智时代人机协同的研究现状与未来方向[J]. 管理工程学报, 2024, 38(1): 1-13.
[14]Bandura A. Social cognitive theory: An agentic perspective[J]. Annual Review of Psychology, 2001, 52: 1-26.
[15]Bauer K, von Zahn M, Hinz O. Expl(AI)ned: The impact of explainable artificial intelligence on users' information processing[J]. Information Systems Research, 2023, 34(4): 1582-1602.
[16]Bigman Y E, Gray K. People are averse to machines making moral decisions[J]. Cognition, 2018, 181: 21-34.
[17]Cadario R, Longoni C, Morewedge C K. Understanding, explaining, and utilizing medical artificial intelligence[J]. Nature Human Behaviour, 2021, 5(12): 1636-1642.
[18]Castelo N, Ward A F. Conservatism predicts aversion to consequential artificial intelligence[J]. PLoS One, 2021, 16(12): e0261467.
[19]Dello Russo S, Mirfakhar A S, Miraglia M. What is the narrative for practice? A review of recommendations on feedback and a guide to writing impactful practical implications[J]. Applied Psychology, 2023, 72(4): 1624-1652.
[20]Dietvorst B J, Simmons J P, Massey C. Algorithm aversion: People erroneously avoid algorithms after seeing them err[J]. Journal of Experimental Psychology: General, 2015, 144(1): 114-126.
[21]Dietvorst B J, Simmons J P, Massey C. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them[J]. Management Science, 2018, 64(3): 1155-1170.
[22]Gray H M, Gray K, Wegner D M. Dimensions of mind perception[J]. Science, 2007, 315(5812): 619-619.
[23]Gray K, Wegner D M. Feeling robots and human zombies: Mind perception and the uncanny valley[J]. Cognition, 2012, 125(1): 125-130.
[24]Ilgen D R, Fisher C D, Taylor M S. Consequences of individual feedback on behavior in organizations[J]. Journal of Applied Psychology, 1979, 64(4): 349-371.
[25]Jia N, Luo X M, Fang Z, et al. When and how artificial intelligence augments employee creativity[J]. Academy of Management Journal, 2024, 67(1): 5-32.
[26]Kellogg K C, Valentine M A, Christin A. Algorithms at work: The new contested terrain of control[J]. Academy of Management Annals, 2020, 14(1): 366-410.
[27]Lechermeier J, Fassnacht M. How do performance feedback characteristics influence recipients' reactions? A state-of-the-art review on feedback source, timing, and valence effects[J]. Management Review Quarterly, 2018, 68(2): 145-193.
[28]Lindebaum D, Ashraf M. The ghost in the machine, or the ghost in organizational theory? A complementary view on the use of machine learning[J]. Academy of Management Review, 2024, 49(2): 445-448.
[29]Logg J M, Minson J A, Moore D A. Algorithm appreciation: People prefer algorithmic to human judgment[J]. Organizational Behavior and Human Decision Processes, 2019, 151: 90-103.
[30]London M, Smither J W. Feedback orientation, feedback culture, and the longitudinal performance management process[J]. Human Resource Management Review, 2002, 12(1): 81-100.
[31]Luo X M, Qin M S, Fang Z, et al. Artificial intelligence coaches for sales agents: Caveats and solutions[J]. Journal of Marketing, 2021, 85(2): 14-32.
[32]Luo X M, Tong S L, Fang Z, et al. Frontiers: Machines vs. humans: The impact of artificial intelligence chatbot disclosure on customer purchases[J]. Marketing Science, 2019, 38(6): 937-947.
[33]Marafie Z, Lin K J, Wang D B, et al. AutoCoach: An intelligent driver behavior feedback agent with personality-based driver models[J]. Electronics, 2021, 10(11): 1361.
[34]Nussberger A M, Luo L, Celis L E, et al. Public attitudes value interpretability but prioritize accuracy in artificial intelligence[J]. Nature Communications, 2022, 13(1): 5821.
[35]Ötting S K, Masjutin L, Steil J J, et al. Let's work together: A meta-analysis on robot design features that enable successful human–robot interaction at work[J]. Human Factors, 2022, 64(6): 1027-1050.
[36]Parent-Rocheleau X, Parker S K. Algorithms as work designers: How algorithmic management influences the design of jobs[J]. Human Resource Management Review, 2022, 32(3): 100838.
[37]Pereira V, Hadjielias E, Christofi M, et al. A systematic literature review on the impact of artificial intelligence on workplace outcomes: A multi-process perspective[J]. Human Resource Management Review, 2023, 33(1): 100857.

[38]Roesler E, Manzey D, Onnasch L. A meta-analysis on the effectiveness of anthropomorphism in human-robot interaction[J]. Science Robotics, 2021, 6(58): eabj5425.

[39]Sharma A, Lin I W, Miner A S, et al. Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support[J]. Nature Machine Intelligence, 2023, 5(1): 46-57.

[40]Tong S L, Jia N, Luo X M, et al. The Janus face of artificial intelligence feedback: Deployment versus disclosure effects on employee performance[J]. Strategic Management Journal, 2021, 42(9): 1600-1631.

[41]Vrontis D, Christofi M, Pereira V, et al. Artificial intelligence, robotics, advanced technologies and human resource management: A systematic review[J]. The International Journal of Human Resource Management, 2022, 33(6): 1237-1266.

[42]Yam K C, Bigman Y E, Tang P M, et al. Robots at work: People prefer—and forgive—service robots with perceived feelings[J]. Journal of Applied Psychology, 2021, 106(10): 1557-1572.

[43]Yam K C, Goh E Y, Fehr R, et al. When your boss is a robot: Workers are more spiteful to robot supervisors that seem more human[J]. Journal of Experimental Social Psychology, 2022, 102: 104360.

[44]Yam K C, Tan T, Jackson J C, et al. Cultural differences in people's reactions and applications of robots, algorithms, and artificial intelligence[J]. Management and Organization Review, 2023, 19(5): 859-875.

Artificial Intelligence Feedback: A Literature Review and Prospects

Guan Jian[1], Li Wenpu[1], He Guohua[2], Zhang Xin'an[1]

(1. *Antai College of Economics and Management*, *Shanghai Jiao Tong University*, *Shanghai* 200030, *China*;
2. *College of Management*, *Shenzhen University*, *Shenzhen* 518060, *China*)

Summary: Performance feedback is one of the most critical ways to motivate and facilitate individual progress. With the advancement of artificial intelligence (AI) technology, feedback provided by AI is increasingly applied in practice, gradually surpassing the quality of performance feedback delivered by human managers. It has become a significant topic in organizational management research. However, existing literature on AI feedback is scattered across different disciplines such as organizational management, education, and healthcare. This dispersion has led to significant differences in research paradigms, theoretical perspectives, and empirical methods within the study of AI feedback. Furthermore, existing literature has not yet formed a unified understanding of the theoretical mechanisms behind the varying effects of AI feedback. In light of this, this paper first clarifies the concept of AI feedback. Next, it systematically summarizes and reviews the deployment effect and disclosure effect through which AI feedback exerts its influence, thereby constructing a research framework for AI feedback. Then, it introduces and summarizes the commonly used or profoundly insightful theoretical mechanisms in existing AI feedback research and discusses the ways these mechanisms are applied. Finally, it proposes five future research directions that hold both scientific value and practical significance.

Key words: AI; AI feedback; deployment effect; disclosure effect

(责任编辑：王舒宁)