《心理科学进展》审稿意见与作者回应

题目：引入 AI“新队友”：人-AI 协同自主性对人-AI 协作的影响机理研究
作者：樊子立, 程垦(通讯作者), 朱梦霆

第一轮

审稿人意见：

这篇研究构想聚焦“技术自主性对人-AI 代理人团队有效性的影响机理”，从“工具→队友”的关系转变切入，并尝试基于角色理论构建一个多路径的理论模型。从选题上来说，这一议题具有一定的前沿性和现实意义，在当前生成式 AI 和 AI 代理快速发展的背景下也具有研究价值。尽管如此，通读全文之后可以发现，该研究构想在概念界定、理论逻辑以及模型建构等方面还存在较为突出的不足，尚需进一步系统梳理和深化。以下列出几个关键问题，供作者参考：

意见 1：技术自主性的概念界定尚不清晰，且对“自主性”的把握不够聚焦

本文将“适应性学习、不确定性、不透明性和开放性”并列为技术自主性的四个维度，但这些维度在理论属性上并不属于同一层面。其中，适应性学习和开放性更多体现的是技术系统本身的能力或功能特征，而不确定性和不透明性则更接近于个体在面对技术时所产生的认知结果或主观感知。这种将“技术属性”与“认知后果”混合在同一构念中的处理，会导致技术自主性的构念边界模糊，其内部结构也缺乏一致性。

更为关键的是，从“技术自主性”这一概念本身出发，其理论核心理应落在“自主”二字之上，即技术系统在多大程度上能够自主决策、自主行动以及自主学习。这一点在相关人机协作与自动化/自主性研究中通常是较为清晰的，也是区分“自动化”与“自主性”的关键所在。但在当前的概念化中，这一核心反而没有被突出，取而代之的是一些与“自主”本身并非直接对应的维度（如不透明性、不确定性），这在一定程度上削弱了概念的内在逻辑，也使得“技术自主性”这一核心变量显得不够聚焦。

回应：感谢审稿专家对本文的认可，诚如专家所言，本文对技术自主性四大维度的界定和拆解不够全面，以下我们将根据专家意见进行更正和补充：

第一，本文中技术自主性的四大维度是引用了 Simmler 和 Frischknecht（2021）发表在 AI & Society 上《A taxonomy of human-machine collaboration: capturing automation and technical autonomy》中的结论，他们将技术自主性分为了四大维度，具体如下表 1 所示。

表 1 技术自主性的四大维度

	四大维度	释义
技术自主性	1.不透明性（Non-transparency）	用户或观察者是否能够根据系统来跟踪输入 A 到输出 B 的全过程？
	2.不确定性（Indetermination）	输入 A 是否一定能够导致输出 B？
	3.适应性学习（Adaptability）	该系统是否能够从其经验中吸取教训，完成自我更新？
	4.开放性（Openness）	该系统是否能够扩大其原始数据，特别是与其他系统合作时表现出的在收集源数据方面的灵活性？

资料来源：Simmler, M., & Frischknecht, R. (2021). A taxonomy of human-machine collaboration: Capturing automation and technical autonomy. *AI and Society*, 8(12), 78-98.

Simmler 和 Frischknecht（2021）倾向于将独立的，不完全确定的和无法追踪的东西称之为具有“自主性”，因此，技术自主性涉及系统性能的可追溯性和确定性的某种缺乏，并且涉及两个方面：透明性和确定性。如表 1 所示，我们得知当一个系统不确定时意味着 A 并不一定会生成 B；而当一个系统不透明时，则意味着 A 是如何生成 B 的，其过程无法追踪，这两项特征都会影响人们对该系统的感知。因此，不透明性和不确定性更偏向于技术属性，但它的确会影响人们的认知后果，Simmler 和 Frischknecht（2021）也将不透明性和不确定性定为了技术自主性的第一个和第二个维度。

同时，“自主性”的技术特征也让其无需人工干预，可以开放性地广泛与其他智能体（包括人类）进行合作来扩展和更新自身的原始数据，自主进行学习和不断适应变化的外部环境，因而这种开放性和适应性学习能帮助其构建“人-AI 协作”等等多智能体协作系统，Simmler 和 Frischknecht（2021）也将开放性和适应性学习定为技术自主性的第三个和第四个维度。

总而言之，**由于技术自主性原因，使得 AI 代理人将会与谁交互（开放性）是不可预见的，而且，它可以根据经验和学习（适应性学习）改变自己的行为，它的编程可能会使其不可预测（不确定性）和不可追踪（不透明性）。**

据此，本文沿用了 Simmler 和 Frischknecht（2021）的观点，将适应性学习、不确定性、不透明性和开放性定为技术自主性的四大维度。我们已在原文中做了进一步的补充和解释，**并用红字标出**。

第二，审稿专家认为技术自主性的理论核心应落在“自主”二字之上，即技术系统在多大程度上能够自主决策、自主行动以及自主学习，而本文没有突出其核心，取而代之的是与之相关的维度（如不透明性、不确定性）来进行。我们依据审稿专家的观点，先对“技术自主性”的不同等级以及与“技术自动化”的区别进行进一步地阐述。Simmler 和 Frischknecht（2021）将“技术自主性”由低到高区分为 5 个等级，具体如表 2 所示。

表 2 技术自主性的五等级

	五等级	释义
技术自主性	1.确定性系统（Deterministic system）	技术系统是确定的、透明的、不适应的、封闭的
	2.非透明性系统（Non-transparent system）	技术体系是确定的、不透明的、不适应的、封闭的
	3.不确定性系统（Indetermined system）	技术体系是不确定的、不透明的、不适应的、封闭的
	4.适应性系统（Adaptable system）	技术体系是不确定的、不透明的、适应性强的、封闭的
	5.开放性系统（Open system）	技术体系是不确定的、不透明的、适应性强的、开放的

资料来源：Simmler, M., & Frischknecht, R. (2021). A taxonomy of human-machine collaboration: Capturing automation and technical autonomy. *AI and Society*, 8(12), 78-98.

如表 2 所示，“技术自主性”由低到高的 5 个等级分别为：确定性系统（Deterministic system）、不透明性系统（Non-transparent system）、不确定性系统（Indetermined system）、适应性系统（Adaptable system）、开放性系统（Open system）。由此可见，对技术自主性 5 个等级的划分也是根据其四大维度来进行界定的。

与此同时，为了将“技术自动化”和“技术自主性”区分开来，Simmler 和 Frischknecht（2021）进一步地将技术自动化也分为了 5 个等级（如表 3 所示）。

表 3 技术自动化的五等级

	五等级	释义
技术自动化	1.提供决策（offer decision）	技术系统只提供建议的选项，最终决定权在人
	2.执行人类批准的决策（Executes with human approval）	在人类批准后，技术系统才会采取行动
	3.执行人类没有否决的决策（Executes if no human vetoes）	如果人类没有否决，则技术系统就会采取行动
	4.采取行动，然后通知人类（Executes and then informs）	技术系统独立人类布置的运行，并将执行的操作通知给人类
	5.完全自动执行（Executes fully automated）	技术系统独立人类布置的运行，不通知给人类

资料来源：Simmler, M., & Frischknecht, R. (2021). A taxonomy of human-machine collaboration: Capturing automation and technical autonomy. *AI and Society*, 8(12), 78-98.

如表 3 所示，“技术自动化”具体分为 5 个等级：提供决策（Offers decision）、执行人类批准的事（Executes with human approval）、执行人类没有否决的事（Executes if no human vetoes）、直接执行分配好的任务并通知人类（Executes and then inform）、直接执行分配好的任务无需通知人类（Executes fully automated）。尽管如此，但 Simmler 和 Frischknecht（2021）认为“技术自动化”更侧重于强调任务分配，技术系统在已分配好的任务中实现由低到高的自动化工作，这本质上还是需要人为参与，并没有脱离人而独立采取行动。而“技术自主性”则更侧重于强调技术特征，强调技术系统可以在完全不需要人为参与的情况下进行自我学习和更新，独立作出判断并采取行为，且过程难以追踪。**这也是技术自动化和技术自主性最本质的区别**。我们依据专家的意见，已在原文中加入了对技术自主性 5 个等级的具体划分以及技术自主性和技术自动化的区别的阐述，并将修改的内容**用红字标出**。

第三，专家认为“在当前的概念化中，技术自主性这一核心没有被突出，取而代之的是一些与‘自主’本身并非直接对应的维度（如不透明性、不确定性），这在一定程度上削弱了概念的内在逻辑，也使得‘技术自主性’这一核心变量显得不够聚焦”。我们非常认可专家的观点，以往的研究也是依据审稿专家的研究思路展开研究，例如，已有研究将技术自主性看做一个整体，探讨技术自主性等级的高低的后续影响（Devisser et al.，2018；Simmler & Frischknecht，2021；O'Neill et al.，2022），结果发现技术自主性的后续影响效应并不稳定，既会给团队带来益处，也会给团队带来风险，但却并未探索其缘由，也未对技术自主性做更细致的分解。随之，Frischknecht（2022）提出了可以从社会认知视角来解构“技术自主性”，并借鉴了 Simmler 和 Frischknecht（2021）的研究，将技术自主性分为四大维度（如上表 1 所示），以更精细化地观察技术自主性会如何影响人机协作。同时，Alexandra 等（2023）也提出可以从社会感知视角来探索人与 AI 团队的协作。因此，本文响应了 Frischknecht（2022）的号召，在以往研究的基础上将技术自主性进一步解构为四个维度，并依据 Alexandra 等（2023）的观点，更多从社会感知的推理中观察四个维度分别如何影响人-AI 协作。

我们在原文中没表述全面，给专家造成了误解我们深表歉意。我们已在原文中进行了进一步地补充和修正，并**用红字标出**。

参考文献：

【1】Devisser, E. J., Pak, R., & Shaw, T. H. (2018). From 'automation' to 'autonomy': The importance of trust repair in human-machine interaction. *Ergonomics, 14*(3), 156-178.

【2】Simmler, M., & Frischknecht, R. (2021). A taxonomy of human-machine collaboration: Capturing automation and technical autonomy. *AI and Society, 8*(12), 78-98.

【3】O'Neill, T., Mcneese, N., & Barron, A. (2022). Human-autonomy teaming: A review and analysis of the empirical literature.*Human Factors The Journal of the Human Factors and Ergonomics Society, 15*(5), 131-145.

【4】Frischknecht, R. (2022). A social cognition perspective on autonomous technology. *Computers in Human Behavior, 122*(1), 14-22.

【5】Alexandra, M., Harris-Watson, A., & Lindsay, E. (2023). Social perception in human-AI teams: Warmth and competence predict receptivity to AI teammates.*computers in human behavior, 8*(3), 145-162.

意 2：从技术属性到感知变量的转向是否必要？

文中多处引入“对技术（自主性）的感知”来界定或延展技术自主性的内涵，这在理论层面上也需要更加谨慎。需要注意的是，一旦将构念从“技术本身的属性”转向“个体的感知”，实际上是发生了分析层级的转换（从技术特征转向心理感知变量）。如果这一转换缺乏清晰界定，就容易导致理论层级混杂，使得原本应作为前因变量的技术特征，与作为中介或结果的个体感知混合在一起，从而削弱整体理论建构的精度。进一步来看，这里还涉及一个更根本的问题：在当前的研究设定中，为什么不直接围绕“技术自主性”这一技术属性本身展开理论建构，而是转而聚焦于“对技术自主性的感知”？如果这一转向是出于理论上的必要性（例如强调主观认知在其中的关键作用），则需要在文中加以明确论证；否则，这种处理容易给人一种在核心构念尚未充分界定的情况下，过早引入感知层变量的印象。尤其考虑到本文是一篇研究构想而非实证研究，本应有更大的空间去从概念层面对技术自主性本身进行充分刻画和理论化，而不是过早地退回到“感知”层面进行展开。因此，作者有必要进一步思考：这一层从“技术属性”到“感知变量”的转化是否是必要的，其理论收益是什么，以及这一处理是否会在无形中降低理论建构的聚焦性与解释力度。

回应：感谢审稿专家的建议。专家提出“为什么不直接围绕‘技术自主性’这一技术属性本身展开理论建构，而是转而聚焦于‘对技术自主性的感知’”。

首先，我们在已有研究的基础上响应了 Alexandra 等（2023）的倡导，更多从社会感知的推理中来观察和探讨人-AI 协作的问题，因此，在“技术自主性→人-AI 协作”的逻辑推理中我们引用了感知到技术特征的观点，但我们本意并不是将“技术特征”转向“心理感知变量”。

其次，我们是从技术自主性的四个维度（适应性学习、不确定性、不透明性和开放性）出发，基于角色理论来展开理论构建的。我们基于角色理论认为关键在于 AI 在团队中角色的转变（工具→队友）（Dai et al., 2025），在 AI 作为“新队友”融入团队的社会化过程中，人类成员对 AI 队友的社会感知起到了关键性的作用。因此，我们在技术自主性影响人-AI 协作的过程中运用了感知的相关逻辑进行了推理，给专家造成了误解是我们思虑不周。

最后，本文的理论基础是角色理论，角色理论指出个体对自我角色的认知转变与适应很大程度上是由个体对客观环境的认识和感知所决定的（Biddle & Thomas,1979; Turner, 1990；Dai et al., 2025），因此我们虽然围绕技术自主性展开研究，但论证过程中更多聚焦于对技术自主性的感知也是由角色理论的理论内涵所决定的。

我们没有交代清楚这一问题给专家造成了误解我们深表歉意，我们已在文中进行了更进一步的补充和更正，**并用红字标出**。

参考文献：

【1】Biddle, B. J., & Thomas, E. (1979). Role theory: concepts and research. *Huntington, NY:Krieger*

【2】Turner R H. (1990).Role Change.*Annual Review of Sociol*ogy, 16(1):87-110.

【3】Dai, R., Thomas, M. K. E., & Rawolle, S. (2025). The roles of AI and educational leaders in AI-assisted administrative decision-making: a proposed framework for symbiotic collaboration. *The Australian Educational Researcher, 52*(2), 1471-1487.

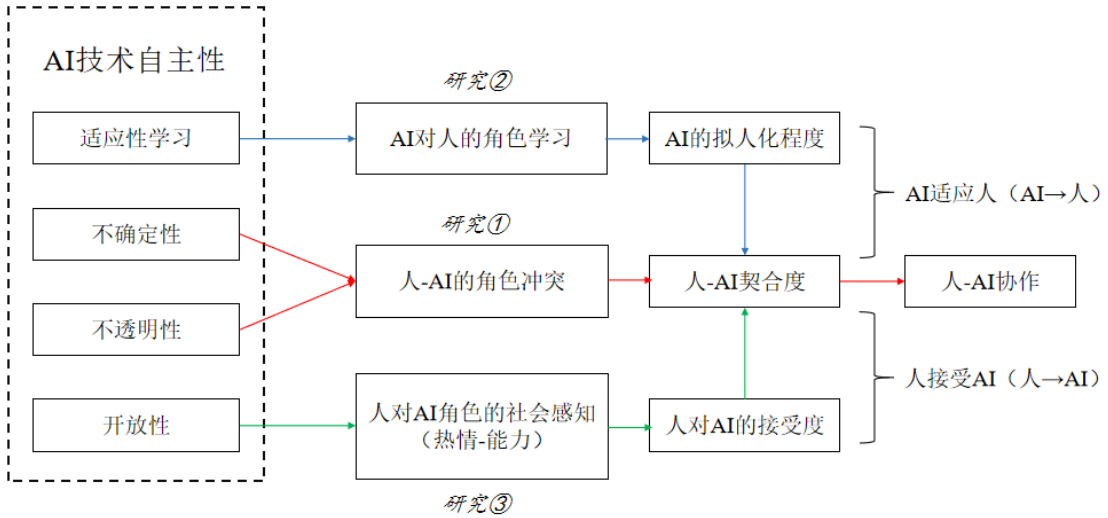
意见 3：理论模型的逻辑结构存在混乱，部分关系设定不够严谨

当前模型在变量设置和路径关系上存在较为明显的逻辑问题。例如，模型中直接设定“人-AI 协作→人-AI 团队有效性”的路径，但“协作”本身已经在很大程度上蕴含了效率、协调和绩效等结果性内涵，这种设定容易导致概念上的重叠甚至同义反复，使得模型解释变成一种循环论证。同时，模型中混合了过程变量（如角色学习）、状态变量（如契合度）和结果变量（团队有效性），但缺乏清晰的层级划分，使得整体结构显得较为松散。此外，三项子研究之间更多呈现为并列关系，而非围绕同一核心机制展开的系统性分解，这也在一定程度上削弱了理论模型的整体性和连贯性。因此，有必要对模型进行整体重构，使其在变量层级和作用路径上更加清晰、紧凑和一致。

回应：感谢审稿专家的建议，我们根据以上意见进行了逐一修改。

首先，诚如专家所言，“协作”本身的确蕴含了效率、协调和绩效等结果性内涵，与文中的“人-AI 团队有效性”之间存在内涵上的重叠。我们依据专家意见，将“人-AI 团队有效性”这一变量删除，仅保留人-AI 协作，以防止出现概念上的重叠和同义反复。

其次，专家认为“模型上整体结构较为松散，缺乏整体性和连贯性，而且三个子研究并非围绕同一核心机制展开分解”。我们遵从专家的意见，对原模型进行整体的重构，使其在变量层级和作用路径上更加清晰、紧凑和一致，具体如图 1 所示。



注：红色箭头（→）代表不确定性和不透明性对人-AI 协作的影响（研究①）

蓝色箭头（→）代表适应性学习对人-AI 契合度的影响（研究②）

绿色箭头（→）代表开放性对人-AI 契合度的影响（研究③）

图 1 技术自主性对人-AI 协作的影响机制模型

如图 1 所示，首先，我们解决专家提出的“原模型混合了过程变量（如角色学习）、状态变量（如契合度）和结果变量（团队有效性），但缺乏清晰的层级划分，使得整体结构显得较为松散”的问题，在新模型中，我们从技术自主性出发，探讨其对过程变量（角色学习、角色冲突、角色的社会感知）的影响，然后观察其对之后状态变量（AI 的拟人化程度、人

-AI 契合度、人对 AI 的接受度）的影响，最终考察其对结果变量（人-AI 协作）的影响。将相同类型的变量进行了清晰的层级划分，让整体结构更紧凑。

最后，我们也回应了专家提出的“三项子研究之间更多呈现为并列关系，而非围绕同一核心机制展开的系统性分解，这也在一定程度上削弱了理论模型的整体性和连贯性”的问题。在新模型中，我们从研究 1 出发，发现其技术自主性的不透明性和不确定性两个维度对人-AI 协作具有阻碍作用（**研究①**），为了缓解这一问题，我们从人-AI 协作的内涵入手，要加强人-AI 的协作需满足两个方面的要求：一方面，AI 要能主动学习和适应人类角色；另一方面，人类也愿意接受 AI 作为“新队友”（O'Neill et al.，2022；Alexandra.，2023）。因此，在“AI 要能主动学习和适应人类角色”方面，我们从技术自主性的适应性学习维度出发，探索其如何推动 AI 主动适应人类角色（**研究②**），然后在“人类也愿意接受 AI 作为‘新队友’”方面，我们从技术自主性的开放性维度出发，探索其如何推动人接受 AI 作为“新队友”（**研究③**）。三项研究层层递进，相互呼应，而且我们明显界定了三项研究虽然从不同的维度出发，但依然是围绕技术自主性这一核心机制展开，探讨其对最终人-AI 协作的影响。

我们已按照专家的建议对模型进行整体重构，并在原文中进行了修正，**并用红字标出**。

参考文献：

【1】O'Neill, T., Mcneese, N., & Barron, A. (2022). Human-autonomy teaming: A review and analysis of the empirical literature.*Human Factors The Journal of the Human Factors and Ergonomics Society, 15*(5), 131-145.

【2】Alexandra, M., Harris-Watson, A., & Lindsay, E. (2023). Social perception in human-AI teams: Warmth and competence predict receptivity to AI teammates.*computers in human behavior*, 8(3), 145-162.

意见 4：理论动机尚不充分：为何必须从“技术自主性”出发仍有待论证

尽管本文强调技术自主性的重要性，但目前的论述尚未充分回答一个关键问题：为什么在人-AI 团队有效性这一议题中，技术自主性是一个“关键且不可替代”的解释变量。从现有模型来看，诸如角色学习、角色冲突、契合度等机制，在既有团队或人机协作研究中已有较多探讨，即使不引入技术自主性，这些关系在理论上依然可以成立。因此，如果无法进一步说明技术自主性相较于其他技术特征（如自动化程度、智能水平等）所带来的独特作用机制，那么本文的理论贡献就会受到限制。作者需要更加清晰地界定技术自主性的独特性，并论证其在解释人-AI 协作问题中的必要性。

回应：感谢审稿专家的建议，我们根据专家之前提出的意见，将结果变量“人-AI 团队有效性”删去，目前的结果变量是“人-AI 协作”，因此我将根据要求继续论证“为什么在‘人-AI 协作’这一议题中，技术自主性是一个‘关键且不可替代’的解释变量？”

O’Neill 等（2022）在研究人-人协作团队中指出，在团队中引入新队友是会给团队带来新鲜血液，但同时也是会给原有团队流程和规则带有破坏的事件，因而引入新队友后要达到“人-人协作”的良好状态就必须要满足两个条件：1.新队友能够主动适应新的角色；2.原有成员也愿意接纳这位新队友。Alexandra 等（2023）将这一概念从**“人-人协作”**引入到**“人-AI 协作”**，这就要求，**一方面，AI 能够主动适应人类角色；另一方面，人类也愿意接受 AI 作为“新队友”**。

首先，以.往的“技术自动化”难以满足上述要求，“技术自动化”本质上还是停留在任务分配层面上，将技术系统完成任务的自动化程度由低到高分为 5 个等级（详见上表 3），但这本质上还是需要人工来分配任务，技术系统依然没有脱离人的干预，也就不具备“主体性”。而“技术自主性”强调了技术系统是难以追踪的，可以开放地与其他智能体进行合作、学习和自我更新，无需人工干预的不确定、不透明的智能实体，这正是“主体性”的典型特征，而想要 AI 技术系统作为“新队友”的角色进行人-AI 协作，具备“主体性”特征的技

术自主性更能满足其两个必要条件：1.AI 作为独立个体主动去适应人类角色；2.人类也愿意接受 AI 作为“新队友”，因此，“技术自主性”相比于其他技术特征而言具有不可替代性。

其次，Guo 等（2026）的相关研究也提出关于“人-AI 协作”这一主题的研究近年来已不再是将 AI 当做辅助“工具”的角色，更倾向于将其当做一个与人类员工平等的“队友”角色来展开探讨。随着 AI 在协作中扮演的角色的转变，以往基于“工具”角色，从技术自动化或智能水平出发探索“角色学习、角色冲突、契合度等机制如何影响人机协作”的解释力度逐渐受限，因而基于队友角色的技术自主性也逐渐取代了传统的智能水平、技术自动化等技术特征，在研究当 AI 的角色发生转变（工具→队友）的背景下，为如何更好地推进人-AI 协作提供了新的理论支持（见图 1）。

综上所述，我们认为在人-AI 协作这一主题中，“技术自主性”是一个“关键且不可替代”的变量。同时，非常感谢专家对这一问题的指出，让我们意识到因为自己思虑不周全所带来的失误，我们已在原文中进行了补充和详述，**并用红字标出**

参考文献：

【1】O'Neill, T., Mcneese, N., & Barron, A. (2022). Human-autonomy teaming: A review and analysis of the empirical literature.*Human Factors The Journal of the Human Factors and Ergonomics Society*, 15(5), 131-145.

【2】Alexandra, M., Harris-Watson, A., & Lindsay, E. (2023). Social perception in human-AI teams: Warmth and competence predict receptivity to AI teammates.*comput*ers in human behavior, 8(3), 145-162

【3】Gao, B., Zuo, P., & Xia, C. (2026). Impact of interaction diversity and interlayer coupling on the evolution of human-AI cooperation.*Applied Mathematics and Computation*, 518(08):129-138.

意见 5：文献综述存在明显滞后，未能反映该领域近两年的快速发展

本文在对 human-agent teams 及相关领域文献的梳理上，整体上停留在两到三年前的研究进展，未能充分纳入近两年来生成式 AI 和 AI 代理快速发展背景下的大量新研究成果。需要指出的是，人-AI 协作尤其是 human-agent teams 是一个高度动态、快速演进的研究领域，近两年的发展并非边际性的补充，而是在技术形态（如生成式 AI 驱动的多智能体系统、自主代理）、协作模式以及理论视角上都出现了实质性变化。在这样的研究背景下，如果文献综述仍主要依赖两三年前的研究积累，是难以支撑一篇以“研究构想”为定位的论文的。进一步来看，这一问题不仅体现在文献更新上，也体现在对研究现象本身的描绘上。当前引言部分对人-AI 协作情境的刻画仍较为笼统，缺乏对近两年来典型技术形态和应用场景的具体呈现。例如，围绕新一代 AI 代理（如 Manus、OpenClaw 等所代表的自主任务执行与多代理协同系统）所引发的实践探索与学术讨论，已经在很大程度上重塑了“AI 作为队友”的内涵与边界。如果这些具有代表性的技术进展与应用叙事没有被纳入研究背景之中，那么理论问题的提出就容易显得抽象甚至略显滞后，也难以充分体现研究的现实紧迫性与前沿性。

更进一步地，文献与现象的双重滞后会直接影响理论建构的起点：一方面，可能遗漏已有研究已经提出的重要机制或解释路径，使得当前构想与既有前沿研究之间缺乏有效对话；另一方面，也削弱了论文在问题提出阶段应有的“时代感”和解释张力。对于《心理科学进展》这类强调理论整合与前沿引领的期刊而言，这一点尤为关键。因此，作者有必要系统补充近两年的核心文献，并结合最新技术发展与实践情境，对研究问题进行重新凝练和表述，从而提升论文的前沿性与理论价值。

回应：感谢审稿专家的建议，我们根据专家的意见进行了如下修改：

首先，我们对文中的参考文献进行了更新，纳入了近两年来生成式 AI 和 AI 代理快速发展背景下的新研究成果，并据此进一步补充了文中综述的内容，并在原文中用红字标出。

其次，“在引言部分对人-AI 协作情境的刻画仍较为笼统”的问题，我们补充了近两年来典型的技术形态和应用场景，例如：OpenClaw 等自主任务的执行和超级个体的打造等，我们已在原文中进行了补充，并用红字标出。

意见 6：研究构想部分缺乏对未来研究方向的系统呈现

作为一篇研究构想类论文，除了提出具体的理论模型之外，通常还需要在更高层次上对未来研究进行系统性展望，例如凝练若干关键研究议题、提出一个更具开放性的研究框架或研究议程（research agenda）。从目前的呈现来看，本文虽然在最后部分提及了一些研究局限与展望，但整体上仍停留在较为分散和概括性的层面，缺乏一个能够统摄该领域后续研究的框架性思考。这在一定程度上限制了论文作为“研究构想”的价值，也削弱了其对后续研究的引领作用。因此，建议作者在理论模型之外，进一步提炼一个更具整合性和前瞻性的研究方向框架。

回应：感谢审稿专家的建议，我们根据专家的意见凝练若干关键研究议题并提出一个更具开放性的研究框架，并添加进原文中，如图 2 所示。

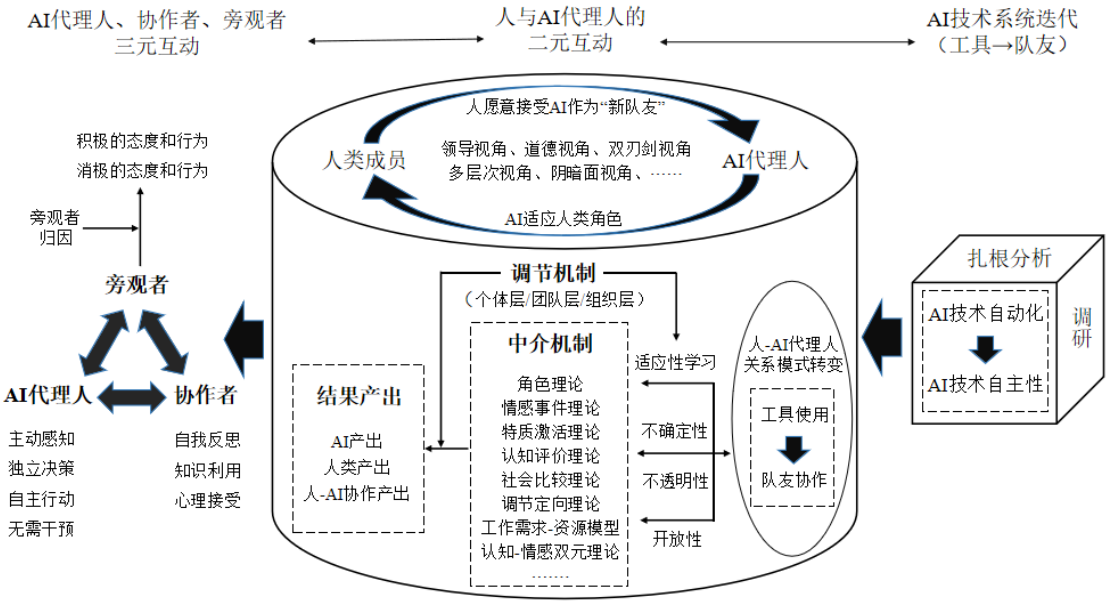


图 2 未来研究展望图

首先，**视角**上可以有更多的创新，本研究从队友视角来进行分析和探讨，但对 AI 拟人化后，随着人际关系的复杂化，我们可以从领导视角、道德视角（唐伟等，2026）、双刃剑视角、多层次视角、以及阴暗面视角等来进行探讨。第二，**理论**上可以采用更多元化的解释机制，本研究运用角色理论，基于人-AI 代理人双方互动关系中角色的演变与磨合这一逻辑主线进行建模研究。未来研究可以尝试从特质激活理论、、情感事件理论、认知评价理论、社会比较理论、认知-情感双元理论等来进行研究，从更多元化的理论视角来剖析技术自主性对人-AI 代理人团队有效性影响的内部机理。第三，可以对技术自主性**四大维度**进行进一步地细化研究，本研究分别从适应性学习、不确定性、不透明性和开放性四个维度来总体说明其技术自主性是如何推进人-AI 代理人团队有效性磨合的，未来可以分别将四个维度单独进行更细致的研究，为未来 AI 技术自主性在企业中的落实和完善提供更丰富的理论支持和实践启示。第四，本研究只考虑人-AI 代理人协作的双方，却忽略了组织中**第三方**的存在，众所周知，在组织的前进过程中，第三方往往起着举足轻重的作用，未来可从第三方旁观者

的角度来展开研究，探索当第三方看见人与 AI 双方协作会如何影响旁观者的工作态度和行为。

我们已将以上内容在原文中进行了补充和修正，并用红字标出

第二轮

审稿人 1 意见：

作者根据审稿意见对稿件进行了较大幅度修改，在概念界定、模型调整、文献补充以及研究展望等方面均进行了补充和修订，也能够看出作者进行了认真修改和思考。尤其是在删除“人-AI 团队有效性”变量以及重新调整理论模型方面，相较初稿已有一定改善。然而，通读修改稿后发现，目前的修改更多体现为对原有观点进行进一步解释和补充，而并未对部分关键理论问题进行实质性回应。当前稿件在核心构念界定、理论层级一致性以及模型逻辑整合等方面仍存在较为明显的问题，建议进一步修改。具体意见如下：

意见 1：作者补充说明了技术自主性的四个维度来源于 Simmler 和 Frischknecht（2021），同时增加了技术自主性的等级划分以及技术自动化与技术自主性的区别。然而，原意见的核心问题并不在于“四维度是否具有文献依据”，而在于“四维度是否属于同一理论层级”。换句话说，Simmler 和 Frischknecht（2021）的那种划分方式并不见得是合理的；作为学者，不应仅仅迷信文献，而忽略独立思考和判断。当前的概念化中，适应性学习和开放性更偏向于技术能力或技术功能属性，而不透明性和不确定性则更接近于系统行为特征，甚至容易被理解为个体面对技术时所产生的认知结果。作者目前主要通过补充文献来源来说明四维度的合理性，但尚未充分解释为什么这些维度能够共同反映“自主性”这一核心构念，其内在统一性基础是什么。因此，技术自主性的构念边界依然较为模糊，建议进一步论证四个维度与“自主性”核心内涵之间的理论联系。

回应：感谢审稿专家对本文修改的认可，诚如专家所言，适应性学习和开放性更偏向于技术功能属性，不透明性和不确定性则更接近于个体面对技术时所产生的认知结果，虽然以往有研究支持四个维度放一起的合理性（Simmler & Frischknecht，2021），但本文的确应该进一步进行思考和判断为什么这些维度能够共同反映“自主性”这一核心构念，其内在统一性基础是什么。

由此，本文结合具体的研究模型，其结果变量是人-AI 协作，而且具体的研究背景也是随着人与 AI 协作这种新型团队协作方式的出现，如何保障人-AI 协作的有效性成为亟待解决的问题。因此，本文在探讨如何促进人-AI 协作的过程中涉及的“自主性”更多的是指“**人-AI 协同自主性**（human-AI collaborative autonomy）”（Zou et al.，2025），而不仅仅只是孤立地谈论 AI 技术的自主性（AI technological autonomy）。因而我们在 Simmler 和 Frischknecht（2021）原有研究的基础上将本文的自变量做了进一步的修正和完善，将视角转向人与 AI 系统作为一个联合认知系统时，所表现出来的整体自主能力，其构念具体表现为：

人-AI 协同自主性 = f（系统的技术属性，人对系统的感知属性）

其具体维度包含如下：

适应性学习（技术）：该系统是否能够从其经验中吸取教训，完成自我更新。

开放性（技术）：系统能否广泛与其他智能体开展链接，扩大原始数据，拓展能力边界。

不确定性（感知）：人对系统未来行为不可预测程度的主观感受。

不透明性（感知）：人对系统内部工作原理难以理解的程度。

例如，在人-AI 协作的自动化驾驶任务中，就包含了：自动驾驶+人类感知，这四个维度共同决定了联合系统能否有效、安全地完成自主性任务。

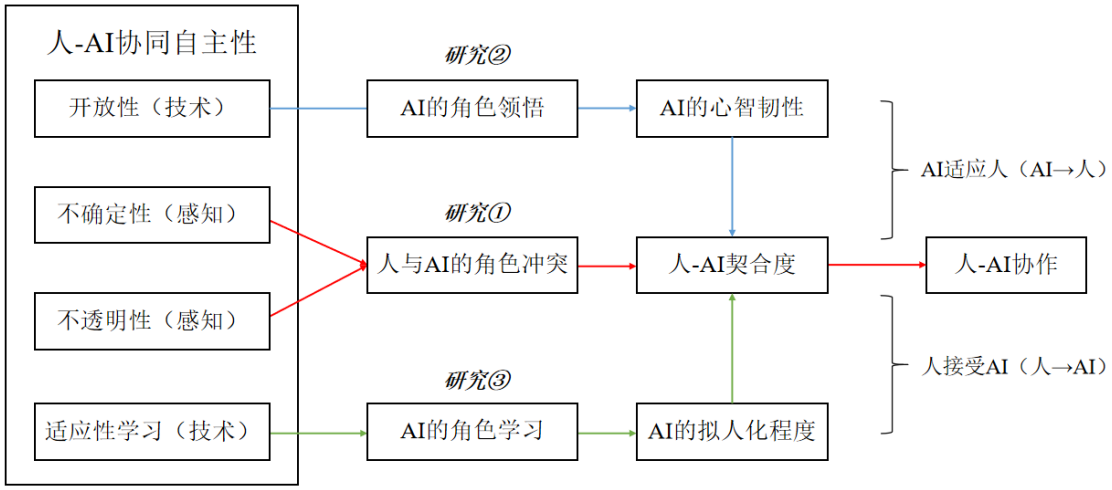
参考文献：

【1】Simmler, M., & Frischknecht, R. (2021). A taxonomy of human-machine collaboration: Capturing automation and technical autonomy. *AI and Society*, 8(12), 78-98.

【2】Zou, H. P., Huang, W. C., Wu, Y., Miao, C., Li, D., Liu, A., ... & Yu, P. S. (2025). A call for collaborative intelligence: Why human-agent systems should precede AI autonomy. computers in human behavior, 12(4), 97-108.

意见 2：作者指出研究本意并非将技术属性转向心理感知变量，而只是借用了社会感知逻辑进行理论推理。但从修改后的论证过程来看，技术自主性、技术自主性的社会感知以及个体对 AI 角色的感知依然同时存在于理论框架中，二者之间的层级关系并不清晰。需要注意的是，一旦引入社会感知相关变量，实际上分析层级已经从技术属性进入个体心理层面。目前尚未明确哪些变量属于客观技术特征，哪些变量属于个体心理感知，以及这些变量之间究竟构成何种理论关系。因此，不同分析层级之间的混合问题仍然存在，也容易削弱整体模型的解释力度，建议进一步厘清技术属性与感知变量之间的边界和关系。

回应：感谢审稿专家的建议，我们将自变量“人-AI 协同自主性”中的适应性学习和开放性两大维度归为客观技术特征，将不透明性和不确定性归为个体心理感知，构建了人-AI 协同自主性对人-AI 协作的影响机制模型，其中，研究 1 中更侧重于心理感知层面，研究 2 和研究 3 更侧重于技术属性层面。同时，**基于角色理论（role theory），按照“角色冲突→角色领悟→角色学习→新的角色冲突→……”的逻辑链条**将三项研究相互联系，相互影响。具体如图 1 所示。



注：红色箭头（→）代表不确定性和不透明性对人-AI 协作的影响（研究①）

蓝色箭头（→）代表开放性对人-AI 契合度的影响（研究②）

绿色箭头（→）代表适应性学习对人-AI 契合度的影响（研究③）

图 1 研究模型图

具体而言，研究 1 中人-AI 协同自主性的不透明性和不确定性属于个体心理感知维度，会让人类感知到与 AI 在协作过程中的摩擦和不可控性，从而产生的角色冲突（role conflict）会降低人-AI 契合度（Yan et al.，2025），进而削弱人-AI 协作；研究 2 中开放性这一技术

特征维度使得 AI 系统可以广泛与其他智能体（包括人类）产生链接，是 AI 系统向外延伸、领悟和拓展的基础，因而在研究 1 产生角色冲突的情况下，开放性维度会通过 AI 的角色领悟（role taking）来让 AI 系统进一步理解和把握与人类协作过程中所需具备的技术特质和角色要求（Derakhshan et al.，2025），从而增强 AI 的心智韧性，以便 AI 更好地理解人类意图并展开协作；研究 3 中适应性学习这一技术特征维度在研究 2 的 AI 领悟角色要求、理解人类意图的基础上，会进一步通过角色学习（role learning）也让人类团队能更好地理解和接受自己（Bauer et al.，2025），即通过海量数据训练和人类反馈来习得扮演该角色所需的技能，增强自身的拟人化程度，最终达到提升人-AI 契合度，促进人-AI 协作的目的。

综上所述，研究 1 从个体心理感知层面论述人-AI 协作过程中的阻碍（角色冲突），研究 2 和研究 3 从技术属性层面分别论述 AI 会如何通过角色领悟来把握角色要求，领悟人类意图，同时又是如何通过角色学习让人类也能理解和接受自己，从而按照“角色冲突→角色领悟→角色学习→新的冲突→……”这一逻辑链条阐述了人-AI 协同自主性对人-AI 协作的影响机制。

参考文献：

【1】Yan, J., Xiao, Q., Zhang, S. X., & Guchait, P. (2025). Artificial intelligence in hospitality: a transactional stress perspective on the mental health of hospitality workers. *International Journal of Hospitality Management, 19*(7), 104-119.

【2】Derakhshan, A., Teo, T., & Khazaie, S. (2025). Investigating the usefulness of artificial intelligence-driven robots in developing empathy for English for medical purposes communication: The role-play of Asian and African students. *Computers in Human Behavior, 16*(2), 121-138.

【3】Bauer, E., Greiff, S., Graesser, A. C., Scheiter, K., & Sailer, M. (2025). Looking beyond the hype: Understanding the effects of AI on learning. *Educational Psychology Review, 37*(2), 45-56.

意见 3：作者根据意见对原模型进行了重构，并删除了“人-AI 团队有效性”这一变量，同时对过程变量、状态变量以及结果变量进行了区分，整体结构相较初稿有所改善。但目前三项研究之间仍更多呈现为不同技术自主性维度分别对应不同作用机制的并列关系，例如不透明性和不确定性对应角色冲突机制，适应性学习对应角色学习机制，开放性对应社会感知机制。这种结构更像多个局部机制的组合，而并非围绕同一核心逻辑展开的系统拆解。此外，角色理论目前同时承担解释角色冲突、角色学习、拟人化以及接受等多个过程，也容易使理论解释范围过宽。建议进一步提炼一个能够贯穿三项研究的统一机制，以增强整体模型的连贯性和理论聚焦性。

回应：感谢审稿专家的建议，本文是基于**角色理论，按照“角色冲突→角色领悟→角色学习→新的角色冲突→……”**的逻辑链条来贯穿三项研究，具体如上图 1 所示。

当团队中引入 AI 作为“新队友”时，其人-AI 协同自主性中的不确定性和不透明性维度会让人类成员感知到与 AI 在协作过程中的摩擦和不可控性，从而滋生角色冲突并影响之后的状态（人-AI 契合度）和行为（人-AI 协作）（**研究 1**）。为了应付双方存在的摩擦和冲突，人-AI 协同自主性中的开放性这一技术特征属性会帮助 AI 系统主动感知和把握融入人类团队需要具备的技术特质和角色要求，即提升 AI 系统自身的角色领悟，让 AI 系统进一步理解和把握与人类协作过程中所需具备的技术特质和角色要求，从而增强 AI 的心智韧性，以便 AI 更好地理解人类意图并展开协作（**研究 2**）。在研究 2 的 AI 领悟角色要求、理解人类意图的基础上，人-AI 协同自主性中的适应性学习这一技术特质属性可以帮助 AI 系统通过角色学习（role learning）让人类团队能更好地理解和接受自己，即通过海量数据训练和人类反馈来习得扮演该角色所需的技能，让自身更加贴合团队成员对协作队友的角色要求，

让人类成员尽可能不把自己当做 AI 系统，而是一个可以自主学习和交流的“类人类成员”，即提升自身的拟人化程度，最终达到加强人-AI 的契合度，提升人-AI 协作的目的（研究 3）。

综上所述，我们根据角色理论，旨在说明当团队中引入 AI 时，**随着 AI 角色的转变（工具→队友），应该如何根据“角色冲突→角色领悟→角色学习→新的角色冲突→……”的逻辑闭环来阐述人-AI 协同自主性对人-AI 协作的影响机理**，最终达到有效预测和提升人-AI 协作的目的。

意见 4：作者进一步区分了技术自动化与技术自主性，并强调技术自主性具有“主体性”特征。但目前关于技术自主性独特价值的论证仍然更多停留在概念描述层面，尚未充分说明技术自主性相较于智能水平、自动化程度、能动性或拟人化等其他技术特征所具有的独特解释价值。换言之，目前仍未充分回答为什么必须从“技术自主性”出发，而不能从其他技术属性出发构建类似模型。因此，建议进一步从理论机制层面阐述技术自主性的不可替代性，以进一步突出研究的理论贡献。

回应：感谢审稿专家的建议，本文的自变量“人-AI 协同自主性”相较于“智能水平、自动化程度、能动性或拟人化”等其他技术特征而言，它彻底转换了分析单元——它不再孤立地审视 AI 系统本身，而是将“人-AI 联合认知系统”作为基本分析单位，加之模型的结果变量是人-AI 协作，因此选择人-AI 协同自主性在理论上与人-AI 协作的逻辑关联性更强，具体表现如下：

第一，从“替代逻辑”转向“互补逻辑”（超越自动化程度）

自动化程度本质是“人机分工”的连续谱（从手动到完全自动化）。它隐含的叙事是“AI 逐步取代人类”，核心问题是“机器能做多少，人还剩多少”（Tenakwah & Watson，2025）。而人-AI 协同自主性则强调了“相互依赖”而非“替代”。在本文人-AI 协作的研究框架下，即使某个子任务 AI 能做到 99%自动化，那剩余的 1%人类介入也不是因为技术不够，而是为了注入价值观、进行授权或提供战略方向。

第二，从“静态能力”转向“动态关系”（超越智能水平）

智能水平通常会通过基准测试来度量，是一个静态的、去语境化的智力存量，它回答“AI 懂不懂”，但无法回答“AI 懂我懂不懂”（Dong et al.，2025）。而人-AI 协同自主性具有共享心智模型和交互适应性等特点，它解释的是 AI 能否感知到人类伙伴当前的认知负荷、是否理解人类意图的微妙变化，并根据人类的实时状态主动调整自己的行动节奏。这种“双向校准”的动态过程相较于智能水平这种单向的智力展示而言，显然与本文中强调团队中引入 AI 新队友，加强人与 AI 协作和动态磨合的过程更加契合。

第三，从“独立主体”转向“交互界面”（超越能动性）

能动性强调 AI 主动发起行动、设定子目标的能力，聚焦于 AI 单方面的主动权（Karim et al.，2025）。而人-AI 协同自主性不仅包含 AI 的主动性，还强调“可干预性”与“可授予性”，解释了人如何在必要时接管或重新定向 AI 的自主流程，即 AI 应当具备“何时该主动、何时该请示、何时该沉默”的社交情境判断。这种对“自主性张力”（人类控制欲 vs. AI 自主效率）的解释力，正是本文中人-AI 协作过程中所必须具备的，也是单维度“能动性”所无法提供的。

第四，从“表面拟真”转向“功能信任”（超越拟人化）

拟人化关注 AI 在外观、语气或情感表达上像不像人，更多表现在界面审美层面（Ramaul et al.，2026）。而人-AI 协同自主性解释的是“校准信任”（Calibrated Trust）的形成机制，即 AI 系统的行为可预测性和犯错时的可解释性，这种力图获得人类长期、理性的功能性依赖和信任正是人-AI 协作的基础。

综上所述，人-AI 协同自主性相较于智能水平、自动化程度、能动性或拟人化等技术特征而言更能够作为解释人-AI 协作的前因要素和出发点，其独特价值正是在于它**跳出了以往孤立地从 AI 技术系统单方面出发与人类互动、以相互替代的静态逻辑来看待人-AI 协作的缺陷，转而将“人-AI 联合认知系统”作为基本分析单位，把人也纳入到分析系统中，强调双方的协同互补、交互信任、动态平衡**。因此，人-AI 协同自主性作为本文模型的出发点，对解释人-AI 协作具有独特价值和不可替代性

参考文献：

【1】Tenakwah, E. S., & Watson, C (2025), Embracing the AI/automation age: preparing your workforce for humans and machines working together. Strategy & Leadership, 53(1), 32-48.

【2】Dong, L., Tang, X., & Wang, X. (2025). Examining the effect of artificial intelligence in relation to students' academic achievement: A meta-analysis. Computers and Education, 12(3), 15-24.

【3】Karim, M. M., Van, D. H., & Kholodov, Y. (2025). AI agents meet blockchain: A survey on secure and scalable collaboration for multi-agents. Future Internet, 17(2), 57-68.

【4】Ramaul, L., Ritala, P., & Aaltonen, P. (2026). Rethinking how we theorize AI in organization and management: A problematizing review of rationality and anthropomorphism. Journal of Management Studies, 63(2), 761-807.

意见 5：作者根据意见补充了近年来生成式 AI 以及 AI 代理相关案例，增强了现实背景描述，但新增内容中较多采用了技术产品和应用案例进行论证，例如 OpenClaw 以及“超级个体”等内容。这类内容更多体现为技术发展现象，而非理论研究进展。研究构想类论文更重要的是与已有学术研究进行有效对话，目前仍然缺少近两年 Human-AI teaming、Agentic AI、多智能体协作等方向核心研究成果的系统整合。因此建议进一步补充近两年的重要理论研究，并通过前沿研究中的关键问题来强化研究问题提出的理论价值。

回应：感谢审稿专家的建议，我们已在原文中补充了近两年 Human-AI teaming、Agentic AI、多智能体协作等方向核心研究成果，并在原文中用红字标出，具体如下。

近两年来，人-AI 团队协作（Human-AI Teaming，HAT）、能动性 AI（Agentic AI）与多智能体协作（Multi-agent Systems）三个研究领域呈现出前所未有的交汇与共振。**这一趋势的核心驱动力是大语言模型（LLM）的崛起——它使 AI 从被动工具转变为具备推理、行动与交互能力的自主实体（Abou et al.，2025）**。三个领域虽各有侧重，却共同指向一个根本问题：AI 如何从被使用的“工具”进化为可协作的“伙伴”？

首先，人-AI 团队协作（HAT）的核心议题是当 AI 成为团队中的“成员”而非仅仅是“工具”时，团队协作模式会发生怎样的变化？Lou 等人（2025）将这一演化概括为 AI 从被动工具向主动队友的转变，强调其学习、适应和在复杂环境中自主运行的能力，并提出了一个四维研究框架：团队形成（Formulation）、协调（Coordination）、维护（Maintenance）与训练（Training）。然而，Vaccaro 等（2024）却表明协同并非天然优于独立，打破了“人+AI 必然优于各自为战”的直觉假设，而关键在于人与 AI 如何交互，但是当前人-AI 协作的交互模式仍以简单的协作范式为主导，真正意义上支持交互协同的模式的研究还十分有限。据此，本文认为随着大模型（Large Pre-trained Models，LPTMs）的引入，HAT 研究迎来了新的维度——既需要 AI 主动适应人类角色，也需要人类愿意接受 AI 作为队友，从而提供反馈来改进 AI 系统。正如 Vats 等人（2024）所言，成功的 HAT 系统并非强大模型的自然产物，而是人为精心设计的结果。

其次，近两年对能动性 AI（Agentic AI）的研究逐渐从“智能”过渡到“能动性”，Plaat 等人（2024）将其定义为具备（1）推理（Reasoning）、（2）行动（Acting）、（3）交互

（Interact）三种能力的大语言模型。然而三者之间并不是孤立存在，反而存在显著的互惠关系。随后，Abou Ali 等（2025）在之前研究的基础上，提出了一个关键的概念区分：单智能体（AI Agent）是自我包含的自主系统，旨在独立完成目标，而能动性 AI（Agentic AI）则是指更广泛的领域与架构方法，常涉及多智能体系统的编排。同时，从更宏观的视角指出能动性 AI（Agentic AI）代表了从被动、任务特定的工具向展现真正能动性的自主系统的队友转变。因此，本文中人-AI 协同自主性中的 AI 更多地是指能动性 AI（Agentic AI），以此来探讨当 AI 的角色发生转变（被动工具→主动队友），该如何促进人-AI 协作。

最后，对多智能体协作（Multi-agent Systems，MAS）的研究近两年也逐渐从单兵作战过渡到群体智能。基于 LLM 的多智能体系统被认为是实现集体智能、迈向通用人工智能的关键路径。Gomez 等（2025）在最近的研究中提炼了一个五组件通用结构，即 Profile（画像）、Perception（感知）、Self-action（自主行动）、Mutual interaction（相互交互）、Evolution（演化）。这一框架将多智能体协作的运行逻辑类比为“人类社会”，并对多个专业化智能体赋予不同的角色身份，通过沟通与协作达成任务目标。随后，Mu 等（2024）发表于《Neurocomputing》的一项综述在已有研究的基础上从社会困境（Social Dilemmas）的视角切入，进一步考察了“多智能体协作、人-智能体协作、AI 增强人类间协作”三个层次的协作，而本文更聚焦于考察人-智能体这一层次的协作，并探究其人-AI 协同自主性会如何有效促进人-AI 的协作。

综上所述，我们可知人-AI 团队协作（HAT）、能动性 AI（Agentic AI）、多智能体协作（Multi-Agent Systems）三个领域并非孤立演进，而是相互关联、层层递进的关系，如图 2 所示：



图 2　人-AI 团队协作、能动性 AI、多智能体协作的逻辑整合图

如图所示，能动性 AI 提供了“能动性”的基础能力（推理、行动、交互）；多智能体协作解决了“多个能动体如何协同”的问题；人-AI 团队协作则回答了人与能动性 AI 如何组成有效团队这一终极问题。**然而，通过对以往研究的整理和梳理，我们发现有以下几个新问题亟待解决**：

第一，**角色分配**，即当 AI 具备能动性时，人与 AI 在团队中的角色会如何变化？

第二，**双方耦合**，如何既保证 AI 能主动融入团队，同时也要确保人类成员能够接受 AI 作为新队友？

第三，**交互设计**，即什么样的交互模式能真正支持人-AI 的协作，而非仅仅只是将 AI 当做工具的使用？

据此，我们引入了人-AI 协同自主性，基于角色理论来构建人-AI 协同自主性对人-AI 协作影响机制模型，旨在探讨当 AI 在团队中的角色发生转变时（工具→队友），人-AI 协同自主性的四大维度（适应性学习、不透明性、不确定性、开放性）会如何以“角色冲突→角色领悟→角色学习→新的角色冲突”为逻辑主线来促进 AI 能主动融入团队的同时也保障人类成员愿意接受 AI 作为新队友，最终达到促进人-AI 协作的目的。

参考文献：

【1】Abou Ali, M., Dornaika, F., & Charafeddine, J. (2025). Agentic AI: a comprehensive survey of architectures, applications, and future directions. *Artificial Intelligence Review, 59*(1), 11-20.

【2】Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human* Behaviour, 8(12), 2293-2303.

【3】Gomez, C., Cho, S. M., Ke, S., Huang, C. M., & Unberath, M. (2025). Human-AI collaboration is not very collaborative yet: A taxonomy of interaction patterns in AI-assisted decision making from a systematic review. *Frontiers in Computer Science, 6*(5), 121-132.

【4】Mu, C., Guo, H., Chen, Y., Shen, C., Hu, D., Hu, S., & Wang, Z. (2024). Multi-agent, human‐agent and beyond: a survey on cooperation in social dilemmas. Neurocomputing, 610, 128514.

意见 6：作者增加了未来研究展望部分，并补充了领导视角、道德视角、阴暗面视角以及多层次视角等未来研究方向。但当前内容仍然更多表现为多个可能研究方向的简单列举（需要指出，这其实是个比较明显的问题），而尚未形成具有整合性和前瞻性的研究议程框架。对于研究构想类论文而言，未来研究部分更重要的是提炼核心研究主题，并形成能够统摄后续研究的开放性框架。因此建议进一步围绕技术自主性与人-AI 协作这一核心问题提炼未来关键研究议题，从而增强论文对后续研究的引领作用。

总体而言，作者进行了较认真且较大幅度的修改，稿件较初稿已有一定改善，但当前修改主要体现为对原有逻辑的进一步解释和补充，而非关键理论问题的根本解决。技术自主性的概念边界、理论层级以及整体模型逻辑等问题仍然存在，建议作者进一步聚焦核心构念，加强理论逻辑的一致性，并进一步完善整体理论框架。

回应：感谢审稿专家的建议，如前所述，**人-AI 协同自主性并不是让 AI 替代人类决策，而是在任务情境中，人与 AI 形成联合认知系统**，通过相互授权和互惠的认知校准，共同涌现出超越任一单体的复合智能体。因此，未来的研究可以聚焦于五个核心研究主题，并通过宏观、中观、微观来分别展开研究，**如图 3 所示**。

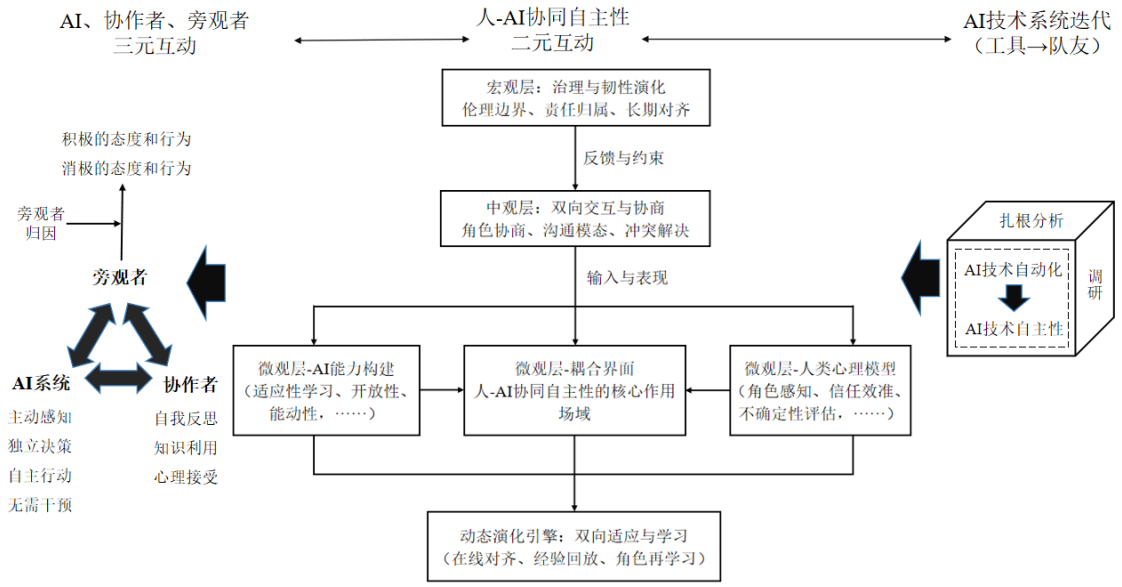


图 3 未来研究框架图

第一，**如何对 AI 进行协同能力的构建**，即 AI 需要具备哪些特定的技术属性，才能成为合格的“队友”，而不仅仅是“工具”？未来的研究可以聚焦于 AI 的交互智能（Interaction Intelligence），包括：适应性学习（如何从人类实时反馈中微调）、开放性（如何接纳人类临时变更的模糊目标）、可解释性（如何生成人类能理解的行为依据）等，打破以往研究认为的“智能越高协同就一定越好”的直观假设，探索人-AI 协作团队中 AI 的协同专用能力（如，主动示弱：AI 在不确定时主动请求人类确认等）的量化标准。

第二，**如何构建人类的心理感知模型**，即人类如何感知 AI 的自主性，这种感知如何在过度信任与信任不足之间动态校准？未来研究可以将人-AI 协同自主性中的不确定性和不透明性作为关键中介变量，研究“感知自主性”与“客观自主性”的偏差及其对协作绩效的影响，可通过研究 AI 的行为模式（如，犯错后的解释风格等）如何影响人类对其心智模型的构建。

第三，**人和 AI 如何进行双向交互与角色协商**，即人与 AI 如何通过交互，动态地“协商”彼此在团队中的角色边界？未来的研究可以将角色理论进一步深度嵌入，研究角色冲突在 AI 场景下的特殊表现（如指令矛盾、安全护栏与用户需求的冲突等），也可以探索交互式角色学习，包括人类如何通过自然语言指令临时重塑 AI 的角色，AI 又如何通过主动提问来澄清角色模糊等。

第四，当人-AI 协同系统面临高风险、长周期任务时，整体系统的稳定性和适应性如何保证？未来研究可从单一的人与 AI 协作扩展到多智能体+人类的混合生态，研究“责任鸿沟”，包括当协同决策出错时，责任在 AI 的设计者、使用者还是部署者，或者在面对环境剧变和 AI 性能骤降时，能否通过人类接管或 AI 降级实现安全退化。

第五，**探讨人-AI 协作如何从静态对齐到终身协同**，即人-AI 协同系统如何随时间推移，像人类团队一样越磨合越默契？未来研究可以通过研究人-AI 团队记忆系统来探讨如何存储过往协作中的冲突与成功经验，并在未来任务中调用，并结合能动性 AI（Agentic AI）的演化组件，探索 AI 如何在多轮协作中更新对人类偏好的内部表征，将人类的情境感知与 AI 的数据记忆进行有机融合，防止 AI 遗忘人类曾经纠正过它的行为。

总而言之，人-AI 协同自主性，本质上是一场“没有终局的角色扮演”。未来研究的核心任务，不再是比拼 AI 的智能程度或自动化程度，而是研究如何让 AI 具备“团队情商”，即在正确的时刻，以正确的方式，将正确的控制权交给正确的人，同时保持系统整体的目标一致性。

第三轮

审稿人意见：

作者在这一轮修改中对审稿意见进行了认真、深入的回应，稿件质量较上一版已有实质性提升。作者将自变量从“技术自主性”明确调整为“人-AI 协同自主性”，清晰地区分了技术属性维度（适应性学习、开放性）与感知属性维度（不确定性、不透明性），有效化解了构念层级混杂的问题；同时以角色理论为统摄框架，按照“角色冲突-->角色领悟-->角色学习”的逻辑链条串联三项研究，使整体理论模型从并列结构走向递进结构，连贯性明显增强。此外，文献更新及时，纳入了近两年人-AI 团队协作、能动性 AI、多智能体协作等领域的前沿成果，未来研究框架也从零散列举整合为宏观—中观—微观三层次的系统展望，理论贡献的独特性得到了更充分的论证。目前仍有两点建议供作者参考：

意见 1：“人-AI 协同自主性 ＝f（技术属性，感知属性）”这类数学化表述在构念定义中不够规范，建议以更严谨的学术语言替代

回应：诚如专家所言，“人-AI 协同自主性 ＝ f（技术属性，感知属性）”这类数学化表述在构念定义中不够规范，建议以更严谨的学术语言替代”，我们按照专家的建议，通过更严谨的学术语言表达，并在原文中用红字标出，具体如下：

“人-AI 协同自主性”是指人与 AI 在共同完成复杂任务时，所涌现出的一种动态适应能力，该能力源于 AI 的客观技术属性（如适应性学习、开放性等）与人类的主观感知评估（如不确定性、不透明性等）之间的深度耦合与递归校准，其整体水平取决于二者在特定任务情境下的交互适配度，而非线性叠加。

意见 2：三项研究的整合虽已用角色理论串联，但在论述中仍存在“研究 1 为感知层、研究 2 和 3 为技术层”的二分痕迹，建议在理论建构部分进一步阐明“感知层面的冲突如何触发技术层面的适应机制”，使角色冲突-->角色领悟-->角色学习之间的因果递进逻辑更加紧密。

综上所述，本轮的修改已有效回应了上一轮的核心关切，稿件的理论聚焦度和逻辑一致性均有显著改善。我已经没有别的问题。感谢作者团队的努力。

回应：专家认为在论述中仍存在“研究 1 为感知层、研究 2 和 3 为技术层”的二分痕迹，建议在理论建构部分进一步阐明“感知层面的冲突如何触发技术层面的适应机制”，使角色冲突-->角色领悟-->角色学习之间的因果递进逻辑更加紧密。我们对这一问题进行了回应和修改，并在原文中用红字标出，具体如下：

Alexandra 等（2023）在对人-AI 协作团队展开深入研究中发现，想要人-AI 协同自主性更好地发挥作用，在 AI 融入人类团队过程中需要满足两方面的要求：**一方面，AI 作为独立个体主动去适应人类角色；另一方面，也需要人类团队愿意接受 AI 作为“新队友”**（Gao et al.，2026）。在研究 1 中我们论证了由于人-AI 协同自主性具有不确定性和不透明性的维度，使得 AI 在与人类的协作过程中具有不可追踪和难以确定的特征，*这让人类成员从**感知层面**上对 AI 存在着角色冲突（role conflict），影响双方的契合度和协作效率*（**详见研究 1**），*这无论是对“**AI 作为独立个体主动去适应人类角色**”还是“**人类愿意接受 AI 作为‘新队友’**”都是不利的*。我们依据 Alexandra 等（2023）的观点认为，从以上两方面着手可能是缓解研究 1 中出现的问题是破局的关键。*因此，研究 2 和研究 3 通过**技术层面**的适应机制（角色领悟和角色学习）分别从以上两方面来展开论述*。首先，***从“AI 作为独立个体主动去适应人类角色”的方面来看***，人-AI 协同自主性的开放性维度会通过 AI 的角色领悟（role taking）

来让 AI 系统进一步理解和把握与人类协作过程中所需具备的技术特质和角色要求（Bilal et al.，2025），从而增强 AI 的心智韧性，以便 AI 更好地理解人类意图并展开协作（**详见研究 2**）；其次，***从“人类愿意接受 AI 作为‘新队友’”的方面来看***，人-AI 协同自主性的适应性学习维度会进一步通过角色学习（role learning）也让人类团队能更好地理解和接受自己（Cai et al.，2025），即通过海量数据训练和人类反馈来习得扮演该角色所需的技能，增强自身的拟人化程度，最终达到提升人-AI 契合度，促进人-AI 协作的目的（**详见研究 3**）。

综上所述，我们从人类成员在感知 AI 技术中产生的角色冲突出发，然后通过技术层面来解决冲突中的两大问题：一方面，通过开放性（技术）促进 AI 作为独立个体进行角色领悟，主动去适应人类角色；另一方面，通过适应性学习（技术）促进 AI 拟人化程度来逐渐让人类团队愿意接受 AI 作为“新队友”，从而实现角色冲突→角色领悟→角色学习之间的因果递进，具体如图 1 所示。

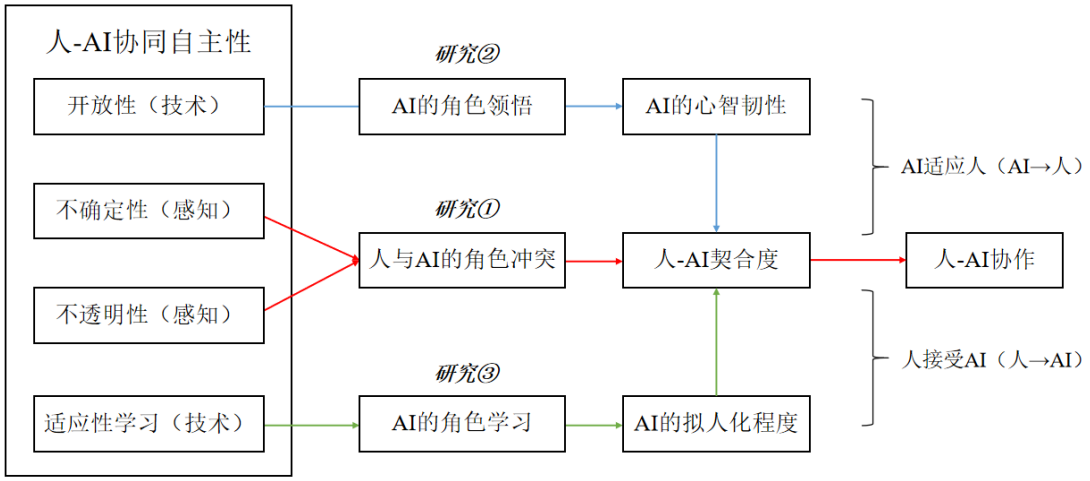


图 1 研究模型图

参考文献：

【1】Alexandra, M., Harris-Watson, A., & Lindsay, E. (2023). Social perception in human-AI teams: Warmth and competence predict receptivity to AI teammates.computers in human behavior, 8(3), 145-162.

【2】Gao, B., Zuo, P., & Xia, C. (2026). Impact of interaction diversity and interlayer coupling on the evolution of human-AI cooperation. Applied Mathematics and Computation, 518(8):129-138.

【3】Bilal, H., Khan, M. N., Khan, S., Shafiq, M., Fang, W., Khan, R. U., ... & Xu, B. (2025). The role of artificial intelligence and machine learning in predicting and combating antimicrobial resistance. Computational and structural biotechnology journal, 27(5), 423-439.

【4】Cai, C., Zhu, G., & Ma, M. (2025). A systemic review of AI for interdisciplinary learning: Application contexts, roles, and influences. Education and Information Technologies, 30(7), 9641-9687.

编委意见：建议录用！