用户如何感知混合主动式人工智能：对问题求解中辅助的态度

Yunhao Luo
yunhaoluo@ucsb.edu
计算机科学 加利福尼亚大学圣塔芭芭拉分校，美国加利福尼亚州

Arthur Caetano
caetano@ucsb.edu 计算机科学 加利福尼亚大学圣塔芭芭拉分校，美国加利福尼亚州

Avinash Ajit Nargund
anargund@ucsb.edu 电气与计算机工程 加利福尼亚大学圣塔芭芭拉分校，美国加利福尼亚州

Tobias Höllerer
holl@cs.ucsb.edu 计算机科学 加利福尼亚大学圣塔芭芭拉分校，美国加利福尼亚州

Misha Sra
sra@cs.ucsb.edu 计算机科学 加利福尼亚大学圣塔芭芭拉分校，美国加利福尼亚州

arXiv:2602.01481v1 [cs.HC] 1 Feb 2026

摘要

在混合主动系统中，人工智能辅助的提供模式可能与辅助本身同样重要。我们研究了两种协助提供模式：按需帮助（用户通过按钮请求）和预定式帮助（按照用户选择的间隔提供辅助，且用户操作会重置计时器）。为评估这两种模式，我们选择 Rush Hour 谜题作为人机协作任务，因为它涵盖了现实问题解决中的若干要素，例如分析、资源管理，以及在约束条件下进行决策。为提高生态效度，我们对时间和人工智能辅助均设置了金钱成本，以模拟人们必须权衡隐性或显性取舍的情境，例如时间压力、经济限制或机会成本。尽管两种模式下的任务表现相近，使用预定式（计时器）模式的参与者对人工智能的感知却更为积极，即使他们最终预算较低也是如此。这表明，协助提供模式能够独立于任务结果影响用户体验；因此，人机系统在提升任务表现的同时，或许还需考虑人工智能辅助的提供方式。

CCS 概念

• **以人为中心的计算 → 交互范式**

范式。

关键词

人机交互、人工智能辅助模态、协作式问题解决

1 引言

在人类与智能系统共同控制任务执行的混合主动系统中，一个核心挑战在于：当用户意图和注意状态存在不确定性时，如何确定提供辅助的恰当时机 [28]。这种时机选择问题在多个领域都有体现。导航系统可能在行程中途为驾驶者重新规划路线，却将沿途观景绕行误判为低效。教育平台则可能因提供过早或过于直接解决问题的提示而助长过度依赖 [34]。医疗应用可能会在不合时宜的时刻发送提醒，例如用户工作期间或深夜，从而降低用户采取行动的可能性。在每种情况下，时机不当的辅助都可能削弱用户能动性、干扰认知过程，或错过用户真正需要帮助的时刻 [47, 64]。

Horvitz关于混合主动界面的框架提出，应在用户发起的交互与系统发起的交互之间取得平衡，并主张采用动态分配主动权的方法 [28]。在此框架基础上，当代系统通常实现两种辅助模式：按需（反应式）帮助，即用户通过明确请求辅助来保持完全控制；以及主动（预期式）帮助，即系统根据情境线索预判需求 [43, 45]。按需辅助能够维护用户自主性，但要求用户中断工作流程，并识别自己何时需要帮助。主动辅助可以提高及时性并减轻用户负担，但面临一个根本挑战：人工智能系统难以准确推断用户意图和注意状态 [7]。这可能导致在不恰当的时刻进行打扰，从而削弱用户的控制感 [5, 38]。

为解决完全自动化主动介入时机的局限性，许多现实中的系统如今采用混合方法：将时机决策交由人工智能，同时让用户控制决定何时以及如何提供辅助的参数。这类混合方法主要有两种形式。定时辅助会在固定的、由用户设定的时间间隔触发，例如每隔几小时提醒用户的用药提醒 [30, 53, 54] ，或 ChatGPT 中的 OpenAI Tasks [1] 等生产力工具，它们会在预设时间执行重复操作或发送提醒。另一方面，不活动触发协助或空闲时间辅助会响应行为线索，在系统检测到长时间停顿或缺乏参与度时提供帮助。例如，辅导系统可以在用户不活动一段时间后显示提示 [20], ，而苹果公司的 Check In [2] 等安全功能则会在检测不到回应时升级警报。尽管这两种方法都旨在兼顾及时性与用户控制，但不活动触发模式尤其值得期待，因为它们能够实时适应用户行为，而非按照固定日程运行。

本作品采用知识共享署名4.0国际许可协议进行许可。

[1]https://help.openai.com/en/articles/10291617-tasks-in-chatgpt[2] https://support.apple.com/guide/personal-safety/use-check-in-for-messages-ips56b5bc469/web

然而，尽管不活动触发协助在实践中十分普遍，与纯粹的按需帮助相比，它是否以及如何影响用户体验，我们仍然缺乏了解。

理解这些动态对于设计既有效又尊重用户自主性的自适应系统至关重要，但有两个因素使这一问题变得复杂。首先，在现实情境中，提供帮助的时机和提供的帮助形式往往相互交织，因此难以单独考察时机机制的影响 [63]。其次，大多数实证研究考察的是只提供一次辅助的单次交互，而非用户反复接受人工智能辅助并形成自适应策略的序列式问题解决场景 [62]。

为探究这些研究空白，我们开展了一项对照研究，考察在资源受限的情况下，不同的辅助时机机制如何影响人们对有用性的感知及问题解决结果。我们选用Rush Hour谜题[3] 作为测试平台，设计了一项序列式问题解决任务：玩家需要滑动阻挡车辆，为目标汽车腾出通道。Rush Hour具备若干特征，这些特征是广泛存在于现实问题解决场景中的基本要素，包括时间压力、多步骤规划所需的人类推理，以及寻求帮助的机会。这些特征在越来越多人与人工智能系统协作的各个领域中十分常见。例如，软件调试通常需要在时间限制下逐步形成并检验假设，开发人员必须决定何时检查程序的中间状态、使用诊断工具或查阅文档 [39]。类似地，诊断和修理家用电器也涉及逐步排查故障，用户需要决定何时查阅外部指导、检验假设或亲自动手操作 [11]。通过将这些共有特征抽象到受控的谜题环境中，Rush Hour使我们能够独立操控辅助时机与辅助内容，从而为一大类人机协作式问题解决场景提供可能具有普遍性的见解。

为模拟现实中寻求帮助时面临的资源约束，我们设计了一个交互式预算受限环境：不断递减的货币预算代表时间成本（每秒 1 美分），每次人工智能干预还会额外扣除 5 美分，以模拟寻求帮助所需的资源成本。在研究过程中，人工智能根据当前谜题状态决定移动哪些车辆，而参与者则通过两种协助机制控制何时获得帮助以及获得多少帮助。

在按钮条件下，参与者会在游戏过程中主动请求人工智能帮助，并指定所需的辅助时长（例如，连续辅助 3 步或 5 步）。在计时器条件下，参与者通过设置不活动阈值来隐式发起辅助；达到用户设定的不活动时长后，人工智能便会自动提供帮助（例如，“如果我闲置 10 秒，就提供 5 步辅助”）。这两种条件分别体现了主动权分配的典型形式：用户主动权和系统主动权。通过按钮按需获取辅助，体现了用户主动权：用户明确决定何时调用自动化，并指定辅助范围，这与强调手动调用和直接控制的框架一致 [9, 50]。基于计时器的辅助则体现了受约束的系统主动权，在这种模式下，系统会根据行为线索自动触发干预，同时受用户定义的参数限制。这一设计符合混合式和可调自主性方法：它们将主动权转移给系统，同时保留用户的监督权 [9]。我们选择这两种模式，是因为它们有坚实的理论基础[9, 50]，并广泛应用于 ChatGPT [4] 和苹果公司的 Check In [5]等已部署系统。先前研究已阐述人机交互的宏观范式（例如对话式、持续式和主动式系统） [61]，；而我们的研究聚焦于这些范式中的一个基本设计维度，即辅助何时启动这一主动权如何在用户与系统之间分配。我们在相同的预算约束下比较这两种模式，以考察不同的主动权分配如何影响用户体验、对人工智能有用性的感知以及问题解决表现。

我们的研究结果（$N = 66$）表明，协助提供模式会影响用户体验，而这一影响独立于任务结果。这说明，人工智能以何种方式提供帮助，可能与提供什么帮助同样重要。这对协作式人工智能系统的设计和评估都有启示。从设计角度来看，交互方式——包括时机、主动权和控制机制——应与能力提升一样，作为首要考量。从评估角度来看，仅凭绩效度量指标可能无法全面反映情况，因此评估应纳入体验测量指标，以捕捉不同交互范式对用户满意度和能动感的影响。

本研究有三项贡献：

- 我们提出了一种实验框架，将帮助何时提供与提供何种帮助解耦，从而可以独立研究辅助的时机与内容。
- 研究提供的实证证据表明，在影响人们对人工智能辅助的偏好时，用户体验（包括感知自主性、人工智能能力和有用性）与任务表现同样重要。
- 我们的研究结果表明，在多步推理和问题解决任务中，不活动触发协助是协作型人工智能的一种可行设计方案。这一发现揭示了更广泛的设计启示：在人机协作中，应如何平衡时机、对照组与系统主动权。

2 相关工作

我们的研究建立在混合主动权系统以及人机协作中的人类能动性这两个交叉领域的既有研究之上。

2.1 混合主动性交互

长期以来，人们一直将人机协作视为一种发挥人类与人工智能互补优势的途径，在某些情况下，其效果优于单独工作 [60]。人机交互中的一个关键范式是混合主动性交互，即人类和人工智能都可以发起操作，并根据情境和能力动态共享对照组，而非预先分配 [3]。在这一研究领域，Horvitz [28] 提出了“优雅耦合”的概念，认为辅助

[3]https://en.wikipedia.org/wiki/Rush_Hour_（谜题）

[4]https://chat.openai.com [5] https://support.apple.com/guide/personal-safety/use-check-in-for-messages-ips56b5bc469/web

只有与用户的情境和注意力同步提供时，才能发挥最大效用。LookOut 项目在一个日历系统中体现了这一原则：该系统利用文本解析和概率推断，预测用户何时可能想要安排事件 [28]。LookOut 支持多种交互模式，包括一种自动化辅助模式；在该模式下，系统会根据推断出的用户意图发起操作。

要在实践中实现优雅耦合，就必须解决一个关键的设计问题：系统操作应在何时以及如何被触发？关于注意力敏感警报的研究提出了相关模型，依据情境线索 [29], , 并借鉴情境感知计算领域的基础性工作 [1] 以及隐式与显式交互模式之间的区别 [56]，以确定系统进行干预的恰当时机。然而，实证证据表明，用户对于自动控制与手动控制有着不同偏好，而最佳平衡取决于任务特征和个体差异 [31]。这些围绕触发机制的设计考量，对于用户如何在现实世界的人工智能系统中体验并行使能动性，具有重要影响。

当代人工智能系统实现了日益增强的智能体能力，能够在更广泛的情境中运行，并支持比早期混合主动系统更复杂的工作流 [4]。例如，GitHub Copilot [6] 可以根据当前文件、光标位置和最近的编辑内容生成代码建议。除了基本的代码补全之外，Chen 等人 [12] 提出了一种主动式人工智能编码辅助工具，它可以利用开发者的代码及其周围情境（如聊天历史）来提供及时且易于整合的建议。他们的用户研究表明，主动式辅助工具能够同时提升程序员生产力和用户体验。在心理健康支持领域，Liu 等人 [40] 推出了 ComPeer，这是一种对话式同伴支持智能体，能够根据先前的用户交互自主发起消息，并调节主动联系的频率。他们的研究表明，与完全由用户发起的对话式智能体相比，这种主动式设计能够改善压力管理，并提高用户满意度。类似地，Kuang 等人 [35] 研究了主动式对话人工智能在用户体验评估领域的影响，考察建议出现的时机（即是在可用性问题出现之前、期间还是之后）如何影响评估者的体验。在一项混合式 Wizard-of-Oz 研究中，他们发现，尽管不同时间点下的总体分析表现没有显著差异，但参与者更偏好在潜在可用性问题出现后立即呈现的建议。

混合主动式研究长期以来强调动态协商控制的价值，并指出辅助时机是关键的设计因素 [13]。此前关于注意力敏感型提醒的研究展示了一种判断系统何时应主动介入的方法，即借助用户注意力和期望效用模型 [29]。然而，对于如何在持续进行的问题解决任务中安排辅助时机，以及用户如何体验辅助时机，目前所知仍然有限。前文讨论的系统依赖设计者设定的基于规则或启发式方法来决定何时启动辅助。与此不同，我们通过研究多步骤问题求解中的辅助时机，探讨由用户触发的触发机制。通过实证比较按需辅助和不活动触发协助，我们的研究补充了此前的混合主动式交互研究 [28, 29] ，提供了有关辅助时机机制如何影响用户对人工智能的态度和感知的证据。

[6] https://github.com/features/copilot

2.2 人机协作中的能动性与控制

在人机交互中，由谁掌握主动权和控制权，长期以来一直被视为一项关键的设计选择。早期理论框架奠定了基础概念，帮助人们理解这种控制权的分配。Parasuraman 等人 [50]描述了人与自动化交互的类型和层级，阐明了感知、决策和行动如何分配给人类和机器智能体。在此基础上，关于可调节自主系统的研究确定了人类与自主智能体之间动态协商控制权的关键维度 [9]。这些理论见解已转化为人机交互的具体设计指南，其中包括高效调用和撤销等原则，以及全局控制机制，使用户能够自定义人工智能监测的内容及其行为方式 [4]。

实证研究展示了这些动态在实践中的表现。Kang 和 Lou 针对 TikTok 用户开展的一项研究 [32] 发现，用户既欢迎人工智能驱动的个性化，也会主动影响人工智能算法，使其更好地满足自身需求。这种人类与人工智能之间的双向能动性塑造了用户参与模式，也表明由平台促成的共同能动性如何改变用户体验。然而，要维持人类能动性，仅有双向影响还不够。Fanni 等人 [19]认为，可争议性和纠正机制（允许用户质疑或撤销人工智能的决策）对于保障自主性至关重要。

随着全面的人工智能自动化仍受到限制，近期研究已转向刻画人机协作。一种框架提出了一个统一模型，以能动性、交互和适应为核心，通过综合文献和专家访谈开发而成，用于描述和分析复杂的人机系统 [27]。然而，一些关键设计问题仍未得到充分探讨，尤其是人工智能辅助的时间维度。基于这项先前研究，我们探究这些动态如何在协作式问题解决情境中展开，在这一情境下，人类与人工智能共同促成结果。我们专门将辅助时机与辅助的数量和内容区分开来，以应对混合主动式研究中反复出现的挑战：这些维度往往被混为一谈。通过单独考察时机的影响，我们能够得出更清晰的见解，为问题解决中的人机协作设计提供指导。

3 任务设计

为系统地研究不同形式的人工智能辅助如何影响协作式问题解决，我们设计了一项使用Rush Hour谜题的对照研究。该任务既体现了协作的关键特征，又能够精确操控人工智能何时以及如何提供帮助。

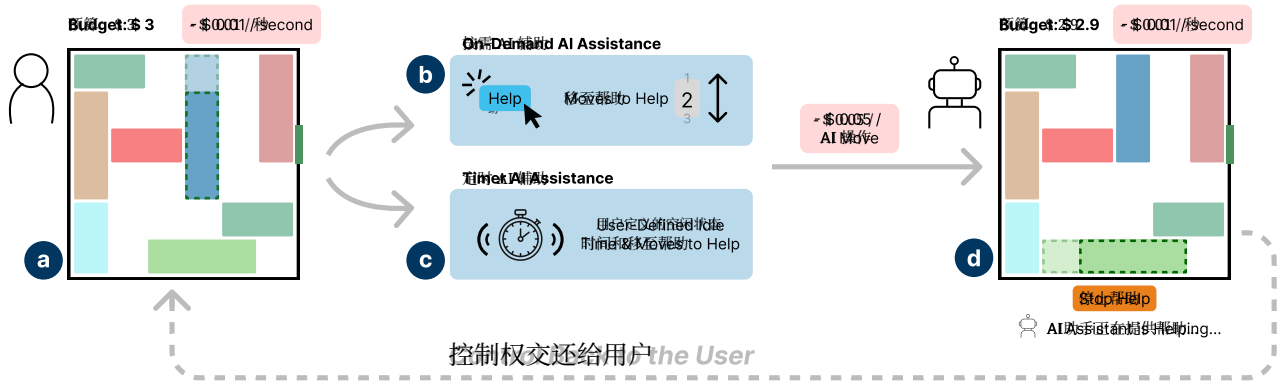


图 1：研究系统设计。(a) 用户可以使用鼠标选择一辆汽车（任意颜色的矩形），选中的汽车将以虚线轮廓高亮显示；然后使用键盘方向键移动选定的汽车。人工智能辅助有两种模式：(b)按钮模式：在解谜过程中，用户可以随时点击帮助按钮并指定希望人工智能辅助的步数来请求帮助。所选步数会保留以供后续请求使用，但用户可以随时调整。(c) 计时器模式**：在启动谜题之前，用户需要设定空闲时间阈值以及希望人工智能辅助的步数。如果用户在超过空闲时间阈值后仍处于非活动状态（既未选择也未移动汽车），则会触发人工智能辅助。如果用户在达到空闲时间阈值之前选择或移动拼图片，计时器将重置。这些参数在解谜过程中无法更改。(d) 当人工智能辅助被触发时，人工智能将接管棋盘并执行所请求的步数；在两种条件下，每一步均产生$0.05的成本，并在激活时从预算中扣除。在AI助手处于活动状态时，用户可以点击两种条件下均提供的停止帮助按钮来中断辅助。在两种辅助条件下，每秒均产生$0.01的时间成本。**

Rush Hour谜题提供了一个定义明确、目标导向的问题空间，既易于上手，又具有认知挑战性。玩家在一个 6×6 网格上解谜，网格中有宽度为一格、长度为两格或三格的汽车，方向为水平或垂直。汽车只能沿其朝向每次移动一格，但不能相互重叠，也不能移出网格。玩家的目标是移动阻挡道路的车辆，为红色汽车开辟一条从网格右侧驶出的通道。

Rush Hour需要顺序推理和空间推理，提供了一种结构化挑战，使寻求帮助的模式能够被观察到。该任务具有明确、客观的指标，例如完成谜题所需的最少步数和进度，有助于评估表现。同时，规则简单易学，无需任何先验专业知识，确保参与者能够迅速学会如何解谜题。

3.1 人工智能辅助

我们在两种模式中都采用确定性广度优先搜索（BFS）算法 [42] 实现人工智能辅助。广度优先搜索会系统地探索谜题状态空间，并保证从当前谜题状态出发找到最短解（步数最少）。这样做是为了确保各轮试验中的辅助保持一致且最优，从而使观察到的差异能够归因于辅助提供模式的不同。

3.1.1 辅助模式。 基于 van Berkel 等人提出的混合主动权系统设计框架 [61]，该框架以用户与系统之间主动权的不同分配方式来刻画人机交互，我们实现了两种辅助模式，如图 1 所示：

• 按钮模式。用户可按需点击帮助按钮并指定希望人工智能辅助的步数。系统会立即执行指定步数，并从预算中扣除相应费用。• 计时器模式。在开始每道谜题之前，用户需设置空闲时间阈值以及希望获得辅助的步数。如果用户在超过阈值的时间内没有操作，人工智能便会自动介入并执行指定步数。如果用户在达到阈值之前移动拼图片，计时器就会重置。谜题进行期间，用户无法调整这些参数。

这两种模式体现了人类指令下由用户发起与由系统发起的支持。按钮模式让用户直接掌控，用户可以决定何时请求帮助以及获得多少帮助。这与传统的求助行为相似，用户会有意识地选择何时咨询 AI 资源。

在计时器模式下，辅助会在用户设定的不活动时段结束后触发，从而引入主动性，并将部分主动权转移给人工智能。例如，如果参与者将阈值设为 10 秒，系统便会监测活动，并在谜题步数停止增加 10 秒后自动提供帮助；用户的任何操作都会重置计时器。这种设计应对了主动式人工智能系统面临的一项关键挑战：自主判断介入的最佳时机，即“金发姑娘时间窗口” [18]，对人工智能而言仍然很困难，而且还可能在不合时宜的时候打断用户。计时器模式允许用户定义触发辅助的不活动阈值，将这一判断权交给用户，同时

保持系统的主动性。这种做法符合相关建议，即人工智能系统应根据用户偏好调整主动程度 [13, 41], ；同时也反映了现实场景：自主智能体会在人为设定的边界内响应情境线索。

3.2 预算结构

为了模拟时间压力下的现实问题解决情境，并激励参与者积极投入 [37],，我们采用了结构化成本系统，这种方法已用于先前的研究[17, 33]。每道谜题开始时，参与者会获得 3 美元，金额显示在谜题面板顶部。预算按两种成本类型递减：

- **时间成本：**每经过一秒扣除 0.01 美元。- **人工智能辅助成本：**每次 AI 操作扣除 0.05 美元。

因此，在没有辅助的情况下，参与者最多有 300 秒来解开一道谜题。当参与者请求或触发 AI 帮助时，相应成本会立即扣除；AI 执行操作期间，预算仍会随时间流逝而递减。AI 选择一辆车需要半秒，移动该车也需要半秒。连续移动同一辆车时，AI 只需选择该车一次。研究记录了谜题完成时的剩余预算。如果预算在谜题完成前降至零，研究便继续进行下一道谜题。

预算系统经过试点测试加以完善，以平衡公平性与挑战性。参与者可以通过独立解题，或在卡住时策略性地使用辅助，来最大化收益。仅依赖人工智能并不足够，因为鉴于人工智能的移动速度，有些谜题无法及时解出，而且谜题难度和解题长度事先未知。这种不确定性促使人们在节省预算与利用人工智能辅助之间谨慎权衡。

4 用户研究

我们开展了一项受控的在线用户研究，以考察不同的辅助提供模式如何影响问题解决进度、问题解决效率以及对人工智能辅助的感知。

在本研究中，我们探讨了两个主要研究问题：

RQ1：按钮模式与计时器模式如何在时间与预算限制下影响问题解决效率、资源使用和进度?

RQ2：这些辅助模式如何影响用户对人工智能的能力、有用性和协作的感知?

通过探讨这些问题，本研究旨在增进我们对混合主动系统中时机控制的理解，具体而言，是了解当辅助需要付出明确成本时，用户如何平衡自主性、效率与支持。

4.1 参与者

为确定所需样本量，我们针对包含两个实验条件（按钮条件和计时器条件）及一个协变量的协方差分析，进行了事前效能分析。假设效应量较大，为 $f = 0.45$、 $\alpha = 0.05$、 $1 - \beta = 0.80$，分析结果表明，最低样本量为41。为确保两个AI辅助条件和控制组的组别人数相等，我们共招募了66名参与者（每个条件22名；35名男性，30名女性，1名非二元性别者）。参与者还报告了自己使用人工智能的频率，分为四类：每天（$N = 25$）、每周几次（$N = 26$）、每月几次（$N = 13$）和从不（$N = 2$）。我们通过 Prolific[7] 招募参与者，开展了一项通过定制网页界面进行的在线实验。参与者须满足批准率至少为99%，并且英语流利，因为所有研究说明以及研究前／研究后问卷均以英语进行。每位参与者均获得4美元的基础报酬，以补偿约30分钟的研究环节，另有一笔基于任务的奖金，金额根据第3.2节所述的预算系统确定。参与者获得的奖金等于其在七道谜题（两道基线谜题和五道主要研究任务谜题）中的平均剩余预算。

4.2 研究流程

参与者被随机分配至三种条件之一：（1）**对照组** 条件，不提供人工智能辅助；（2）**按钮** 条件，可通过按钮请求帮助；（3）**计时器**条件，在达到用户定义的空闲阈值后提供辅助。研究流程示意图见图2。研究开始时，参与者先提供知情同意（IRB方案编号 #38-25-0458），随后阅读介绍页面，了解 Rush Hour 谜题规则、键盘控制方法、研究目标和预算系统。然后，参与者填写研究前问卷，内容包括人口统计信息、AI工具使用经验以及技术亲和力量表（ATI） [22]。

接下来，参与者完成一项包含两道谜题的练习任务。所有参与者都先在没有人工智能辅助的情况下完成第一道谜题，以熟悉界面。第二道谜题则按照参与者被分配的条件完成，即按钮、计时器或无协助的对照组条件。练习结束后，所有参与者都在没有人工智能辅助的情况下解决两道谜题，以确定其基线问题解决能力。所有参与者完成的谜题相同，且顺序固定。

在主研究阶段，参与者使用分配给各自条件的辅助模式解决了五道难度各异的谜题。我们使用平衡拉丁方对参与者的谜题顺序进行了平衡 [8]。开始每道谜题前，参与者会看到一个信息面板，提醒他们当前所处的研究阶段，以及如何请求人工智能辅助。在计时器辅助条件下，该面板还包含一个配置模块，用于设置空闲时间阈值和获取帮助所需的步数。只有在参与者关闭信息面板后，谜题才会显示，此时相关成本开始生效。在所有条件下，参与者都可以随时使用重置按钮，将谜题重置为初始状态。完成全部五道谜题后，参与者会进入研究后问卷。研究界面的截图见图 3。

4.3 谜题选择

所有参与者都收到了一组相同的谜题，这些谜题取自在线 Rush Hour 平台[8]。该平台将难度分为四个类别：初学者、中级、高级和专家级，我们从中

[7] https://www.prolific.com [8] http://www.mathsonline.org/game/jam.html

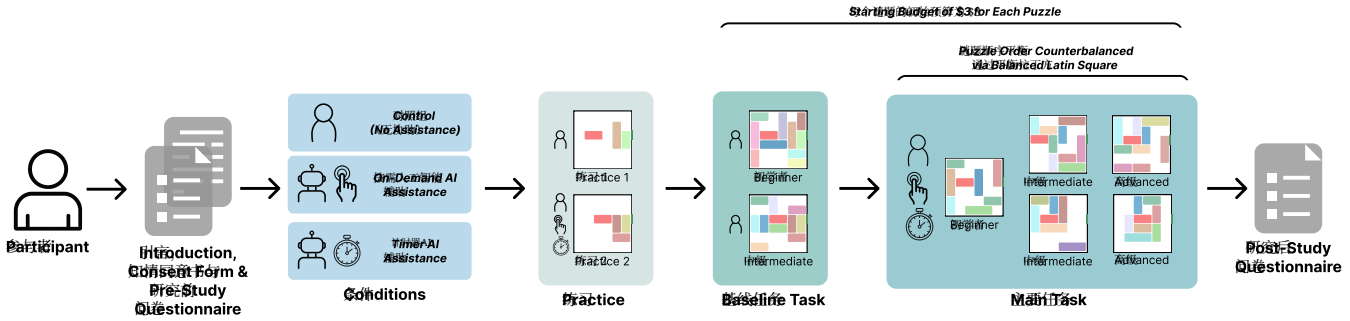


图 2：研究流程。参与者首先阅读 Rush Hour 谜题和研究目标的介绍，随后填写研究前问卷。接着，他们完成两道练习谜题：第一道不使用人工智能辅助，以熟悉谜题和操作方式；第二道则使用其所分配条件下的 AI助手。练习结束后，所有参与者在不使用辅助的情况下解答两道难度级别不同的谜题，以确定其基线解谜能力。随后，他们进入主研究阶段，解答五道难度各异的谜题，并使用与其所分配条件相对应的人工智能。最后，参与者填写研究后问卷。



图 3：用户研究中使用的用户界面。(a1) 按钮条件下的说明面板。(a2) 按钮条件下由用户控制的界面。用户可以调整下一次人工智能辅助将提供多少步帮助。(a3) 由人工智能控制；用户可以点击停止帮助按钮来终止辅助。(b1) 计时器条件下的说明面板，下方设有配置模块，供用户设置空闲时间阈值和帮助步数。(b2) 计时器条件下由用户控制的界面。(b3) 由人工智能控制；与按钮条件下相同，用户可以点击停止帮助按钮来终止辅助。(c) 对照条件和基线阶段使用的界面。用户可以点击“重置谜题”按钮，将谜题恢复到初始状态。按钮条件和计时器条件下也提供此重置按钮。

选取了涵盖三个类别（初学者、中级和高级）的谜题，以覆盖不同难度级别，同时避免在固定时间限制下出现天花板效应和地板效应。两道基线谜题分别为一道初学者谜题和一道中级谜题，而五道主研究谜题包括一道初学者（谜题1）、两个中级（谜题 2和3），以及两个高级谜题（谜题4和5）。每个谜题都有已知的求解所需最优移动次数；然而，这一信息并未向参与者透露。

4.4 度量指标

为评估人工智能辅助对问题解决的影响，既与仅由人类完成（对照条件）进行比较，也在两种AI辅助模式（按钮和计时器）之间进行比较，我们收集了客观绩效指标，以及参与者对其辅助感知的主观反馈。

4.4.1 绩效指标。我们从三个维度评估表现：（1）谜题解决进度，（2）谜题解决精确度，以及（3）预算效率。为量化这些维度，我们引入绩效度量指标。对于给定的谜题，我们定义：

- O = 从初始状态解决该谜题所需的最优移动次数。
- M = 移动总次数，包括人类和人工智能所进行的移动。
- R_i = 在步骤 i时解决谜题所剩的步数，其中 $i \in [1, M]$。根据定义，在谜题开始时或其初始状态（$i = 1$）下，$R_1 = O$。• B = 已花费的预算总额。

据此，我们得出以下绩效度量指标：

（1）谜题步数精确度：对于每个谜题，精确度 P 定义为：

$$\frac{O - min_{1 \le j \le M} R_j}{M}$$

根据定义，$P = 1$ 当谜题以最优移动次数得到解决时，并且 P 会随着额外不必要的步数增加而降低。

（2）最大进度：首先，对于每个谜题，进度 G 定义为：

$$1 - \frac{min_{1 \le j \le M} R_j}{O}$$

根据定义，$G = 0$ 在谜题开始时，$G = 1$ 在完成时。每个谜题的最大进度是所有已进行步数中 G 的最高值。

（3）剩余预算：每个谜题的剩余预算。记录于谜题被解决时，或预算被完全耗尽时；在后一种情况下，记录值为0。

（4）预算效率：对于每道谜题，预算效率定义如下：

$$\frac{O - min_{1 \le j \le M} R_j}{B}$$

该指标表示朝解决方案取得的有效进展与预算支出（包括时间成本和人工智能成本）之间的比率，用于量化每花费一美元预算所取得的有效步数。

4.4.2 主观测量。使用六项简版 NASA-TLX问卷 [24]评估感知工作负荷。为了解参与者对AI辅助模式的态度，我们设计了一组研究后问题，将改编自成熟问卷的题目（$N = 11$）与针对解谜任务定制的问题（$N = 7$）相结合。这些问题改编自：（1）人机交互信任感知量表 [55], 中的两个问题。该量表广泛应用于人机协作研究，重点考察感知能力、可靠性及AI可信度的其他维度；（2）受技术接受模型 [15],启发的三个问题，尤其是其中的“感知有用性”构念；该构念评估一个人认为使用某个系统能在多大程度上提升其任务表现；（3）源自团队合作质量量表 [26], 的两个问题。该框架用于衡量协作，涵盖交流、协调和相互支持等维度；以及（4）基于人机交互研究中用于衡量感知协作的题目 [46], 编写的四个问题，涉及满意度和控制感。这些题目使我们不仅能够评估任务的认知负荷，也能了解参与者对人工智能辅助的感知。所有问题均采用七点李克特量表呈现（完整问题列表见附录A）。我们还加入了自拟的开放式问题，以进一步了解参与者的体验。对于处于AI辅助条件下的参与者，这些问题重点关注他们所接触的具体协助提供模式。对于处于对照条件的参与者，问题则探讨他们对接受辅助的一般态度，以及他们偏好的人工智能辅助形式（完整问题列表见附录B）。

5 定量结果

我们共收集了66名参与者的数据。每位参与者完成了两道基线谜题和五道主要研究谜题。由于缺少日志数据，P60（计时器，男性）的一道谜题被排除在分析之外，最终用于分析的数据集包含329次谜题试验。

为评估三个条件下的表现差异，我们采用了在 R 中通过 glmmTMB 包实现的广义线性混合模型（GLMMs） [10],，并根据数据分布和模型拟合情况选择族和链接函数。对于每项绩效指标，我们将条件和谜题难度（通过最优移动次数衡量，O）设为主要预测变量，并将其作为固定效应。此外，我们还将每位参与者在两道基线谜题中该指标的平均得分作为协变量，以控制个体技能水平差异。为考虑参与者间的变异，我们将参与者ID建模为随机截距。为保持模型的可解释性，我们仅关注具有生态学重要性的效应 [2]。我们检查了所有拟合模型的收敛情况，并使用 DHARMa [25]进行残差诊断，以验证分布假设。

为评估显著性，我们使用 R 中的 car 包，对每个固定效应进行 II 型 Wald 卡方检验 [21]。如果条件是显著预测变量，我们便使用估计边际均值进行事后成对比较，并采用 Tukey 校正处理多重比较。

5.1 任务表现

我们评估了参与者在五个主要任务谜题中的表现（见图 2），并考察了不同条件之间的差异。

5.1.1 移动精度。鉴于谜题步数精确度是一个取值范围为 0 到 1 的连续指标，我们采用了带有 logit 链接函数的有序贝塔回归。这类模型专为有界连续数据或比例数据而设计 [36]。

结果显示，条件（$\chi^2 = 17.654$, $p < 0.001$ $\chi^2 = 9.610$ $p = 0.002$ ）和谜题难度（ ,）具有显著的主效应。基线移动精度不是显著预测变量（$\chi^2 = 2.742$, $p =$

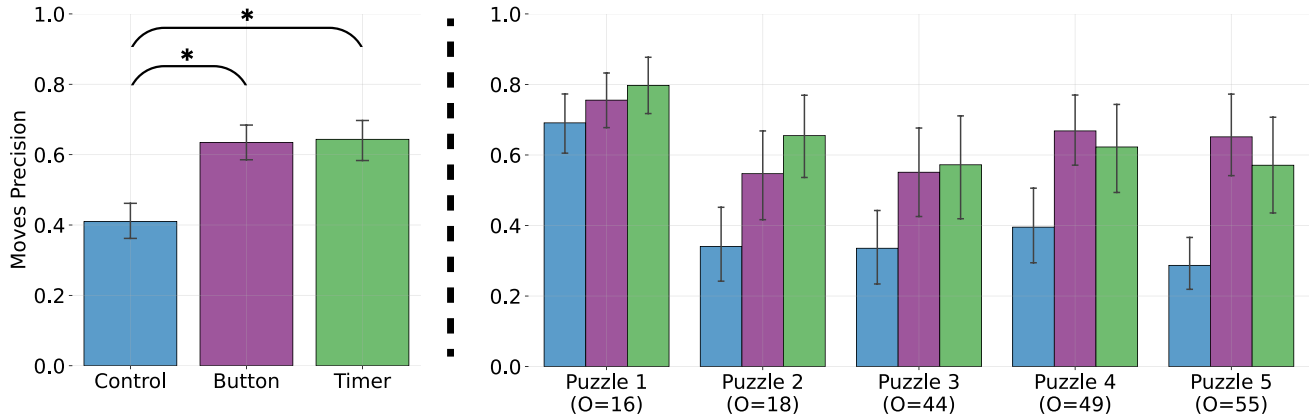


图 4：各条件下主研究谜题的移动精度（左）。星号表示按钮条件与对照条件之间，以及计时器条件与对照条件之间存在显著差异。主研究阶段中各谜题的条件间比较（右）。 *O* **表示从初始状态解出各谜题所需的最优移动次数。条形表示均值，误差条表示 95% 置信区间；这些约定适用于后续所有条形图。**

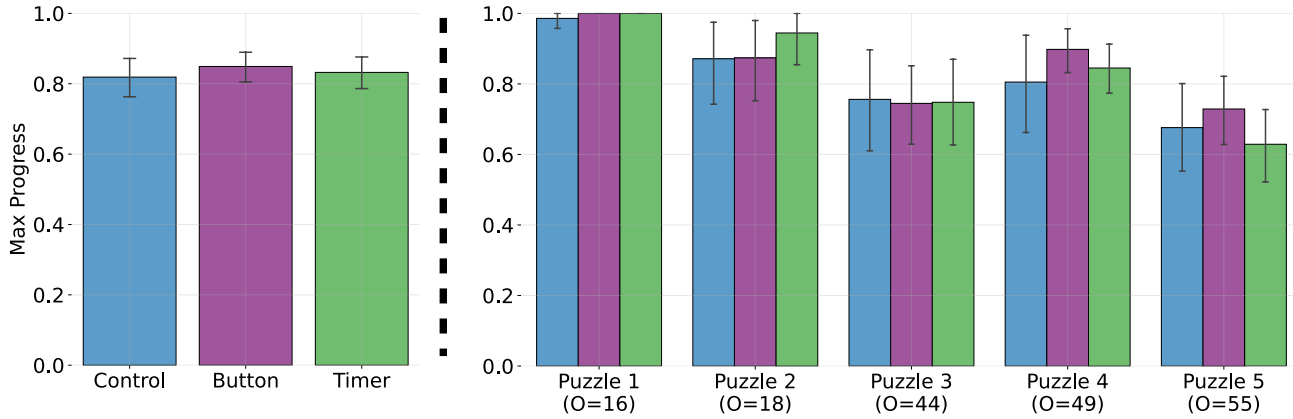
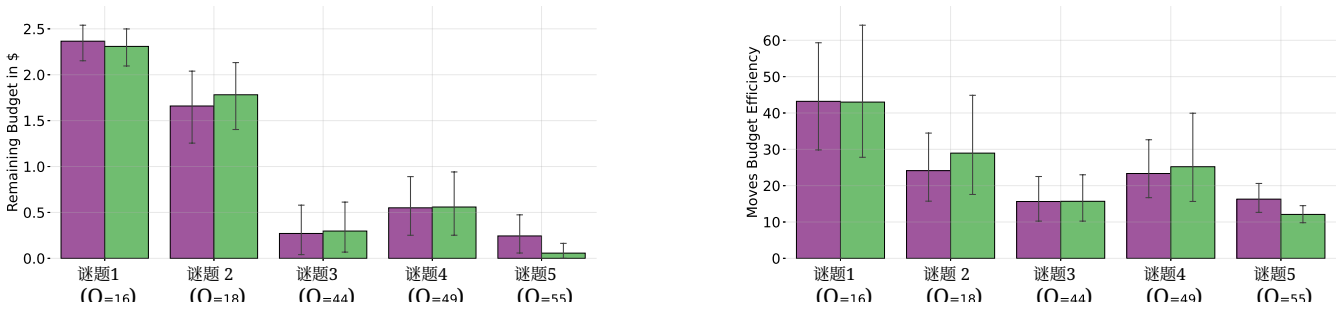


图 5：各条件下主研究谜题所达到的最大进度（左），以及主研究阶段中各谜题的条件间比较（右）。



(a) 剩余预算。　　**(b)** 预算效率。

图6：剩余预算及主研究谜题在AI辅助条件下的预算效率。紫色表示按钮条件，绿色表示计时器条件。

0.098）。事后两两比较表明，对照条件的移动精度显著低于按钮条件（优势比 $= 0.483, p = 0.001$）和计时器条件（优势比 $= 0.463, p < 0.001$）。按钮条件与计时器条件之间未观察到差异（优势比 $= 0.958, p = 0.978$）。见图 4。

5.1.2 剩余预算。对于剩余预算，将数值换算为相对于初始 3 美元预算的比例，并采用带 logit 链接函数的有序 beta 回归进行分析。

结果显示，条件（$\chi^2 = 6.719, p = 0.035$）、$\chi^2 = 283.791$ $p < 0.001$）、谜题难度（ , ）和基线剩余预算（$\chi^2 = 38.032, p < 0.001$）均具有显著的主效应。事后分析表明，对照条件的剩余预算显著高于计时器条件（优势比 $= 1.56, p = 0.049$）。与按钮条件相比，对照条件的剩余预算也呈现较高趋势，但差异不显著（优势比 $= 1.50, p = 0.080$）。按钮条件与计时器条件之间未观察到显著差异（优势比 $= 1.04, p = 0.964$）。见图6a。

5.1.3 预算效率。对于预算效率，由于其呈连续且右偏的分布，因此采用了带对数连接函数的伽马分布。为满足伽马分布要求数值严格为正的条件，在一个预算效率为零的数据点（P17，对照组，谜题3）上加了一个小常数（ϵ）。

结果显示，谜题难度（ $\chi^2 =$ 59.680 $p < 0.001$ $\chi^2 = 40.073$, ）和基线预算效率（, $p < 0.001$ ）具有显著影响。然而，条件没有显著影响（$\chi^2 = 1.521$， $p = 0.468$）。

5.1.4 最大进度。由于最大进度反映了每个谜题的完成比例，我们最初采用了有序贝塔回归。然而，该模型未能收敛，可能是因为取值为一（即完全解出的谜题）的数量较多，因此不适用于有序贝塔回归。为解决这一问题，我们通过计算 1–最大进度对该变量进行了转换，得到一个表示谜题剩余未解比例的新指标。此转换将分布的概率质量从上界移至下界。随后，我们采用了适用于在零处有集中值的连续数据的 Tweedie 分布族和对数连接函数 [57]。

基线剩余进度显示出边缘效应（ $\chi^2 =$ 3.427 $p = 0.064$, ），谜题难度显示出显著效应（$\chi^2 = 53.693$， $p < 0.001$）。条件没有显著效应（$\chi^2 = 0.138, p = 0.934$）。

5.1.5 随时间变化的进度。尽管各条件下的总体进度相似，但探索性的时间进程分析揭示了各个谜题之间细微的差异。在谜题2（中级难度， $O = 18$）中，计时器条件下的参与者在最初的110秒中的大部分时间里，进度都显著高于对照条件下的参与者。在谜题4（高级难度， $O = 49$）中，按钮条件在120至190秒之间优于对照条件。同样，在谜题5（高级难度， $O = 55$）中，按钮条件在约40至250秒的大部分时间里都领先于对照条件，并且在接近结束时也偶尔领先，直到预算耗尽。各谜题随时间变化的进度见图7。

5.2 行为模式

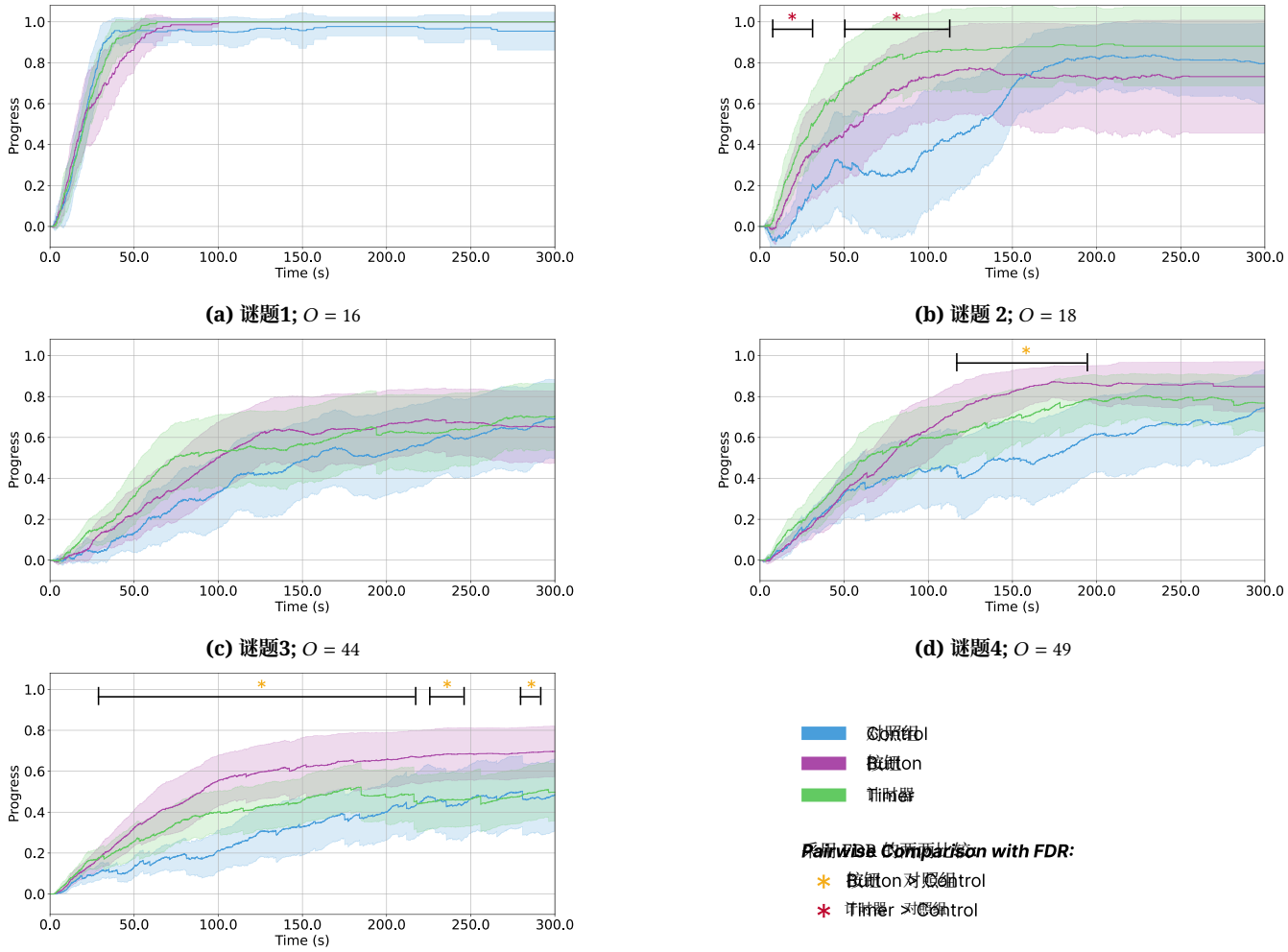
虽然两种AI辅助模式在帮助的触发方式上有所不同，但我们分析了参与者的行为模式，以了解他们如何参与每种模式，以及这些差异可能如何塑造他们的交互策略。

*5.2.1 AI*使用情况。我们使用带对数连接函数的伽马分布，分析了至少出现一次人工智能辅助的谜题中AI辅助操作的平均次数，并将条件作为固定效应、参与者ID作为随机截距。未触发人工智能辅助的谜题被排除在本次分析之外，原因包括：在按钮条件下用户未通过按钮请求帮助，或在计时器条件下用户的空闲时间从未超过预设的空闲阈值。最终，按钮条件中有77个有效谜题，计时器条件中有75个有效谜题。按钮条件（$M = 24.182$， $SD = 14.186$）与计时器条件（$M = 21.053$， $SD = 13.200$）之间未发现显著差异（$p = 0.297$）。同样，每次请求的AI帮助操作平均次数在按钮条件（$M = 6.454$， $SD = 6.490$）与计时器条件（$M = 6.173$， $SD = 6.077$）之间也未观察到差异（$p = 0.556$）。

5.2.2 AI 辅助前后的空闲时间。空闲时间，即参与者未采取任何操作（选择或移动汽车）的时段，是本研究的一项关键行为指标。由于预算会随时间持续减少，参与者因此受到激励，在整个任务过程中保持投入。因此，我们将空闲时间解读为审慎思考的时段，即参与者正在思考或规划下一步操作，而非脱离任务。在此分析中，我们仅纳入至少出现过一次人工智能辅助的谜题（按钮： $N = 77$；计时器： $N = 75$）。

为了了解参与者如何使用按钮模式，我们分析了求助请求之前的空闲行为。在此分析中，我们将空闲时间定义为对选定汽车或谜题状态（包括人工智能所做的更改）的最后一次更改，与启用人工智能辅助的时刻之间的间隔。在按钮条件下，对于主要研究任务中的每个参与者—谜题配对，我们测量了人工智能请求前的空闲时间，发现其长于其他连续参与者操作（即连续谜题操作）之间的空闲时间。这一模式在各个谜题中始终如一。人工智能请求前的空闲时间中位数，平均超过了同一参与者—谜题配对中所有用户操作间停顿的 91.67%（$SD = 11.49\%$）。人工智能求助请求前空闲时间的示意图见图 9。

随后，我们比较了按钮条件和计时器条件下，求助请求之前的空闲时间。具体而言，我们计算了在某个谜题过程中至少触发过一次人工智能辅助的所有参与者-谜题组合在人工智能辅助前的平均空闲时间，并据此进行比较。我们采用伽马分布和对数连接函数，将条件设为固定效应，将参与者ID设为随机截距。结果显示，计时器条件下人工智能辅助前的平均空闲时间（$M = 7.273s, SD = 5.437s$）显著长于按钮条件下的平均空闲时间（$M = 4.661s$, $SD = 3.632s$ $p = 0.009$ ）（）。



(a) 谜题1; $O = 16$

(b) 谜题 2; $O = 18$

(c) 谜题3; $O = 44$

(d) 谜题4; $O = 49$

(e) 谜题5; $O = 55$

图7：主研究阶段中，每种条件下所有参与者在5个谜题上的平均进度随时间变化。 O **表示从初始状态解出每个谜题所需的最优移动次数。进度时间线的生成方法是：首先将每位参与者的个人进度与其开始解谜题的时间对齐，然后在每个对齐的时间点，计算每种条件下参与者进度的平均值。对于在300秒前完成谜题或用尽预算的参与者，将其最终进度值保持不变，以延长其进度轨迹，确保所有参与者的数据都覆盖完整的300秒时间窗口。阴影区域表示95% 置信区间。带星号的水平条表示基于经过错误发现率（FDR）校正的t检验，条件间差异具有统计显著性的时间区间。**

接下来，我们考察参与者在接受人工智能辅助后的反应。我们测量了人工智能辅助后的空闲时间，将其定义为人工智能操作结束到最近一次后续谜题状态变化之间的时长；谜题状态变化既可能由参与者操作引起，也可能由请求或触发的人工智能辅助引起。我们采用与人工智能辅助前空闲时间分析相同的固定效应和随机截距。按钮条件（$M = 4.394s$, $SD = 3.663s$）与计时器条件（$M = 5.272s$, $SD = 3.521s$）下，人工智能辅助后的空闲时间没有差异（$p = 0.164$）。

5.3 感知任务负荷

由于违反正态性，采用曼-惠特尼U检验对NASA-TLX响应进行比较。结果显示，时间需求存在显著差异（$p = 0.048$）。事后两两比较表明，与按钮条件（$M = 5.409$， $SD = 0.269$）和计时器条件（$M = 5.545$， $SD = 0.244$）相比，参与者在对照条件（$M = 4.545$, $SD = 0.300$）下感受到的时间需求较低。然而，这些差异未达到统计显著性（对照组 vs. 按钮： $p = 0.111$；对照组

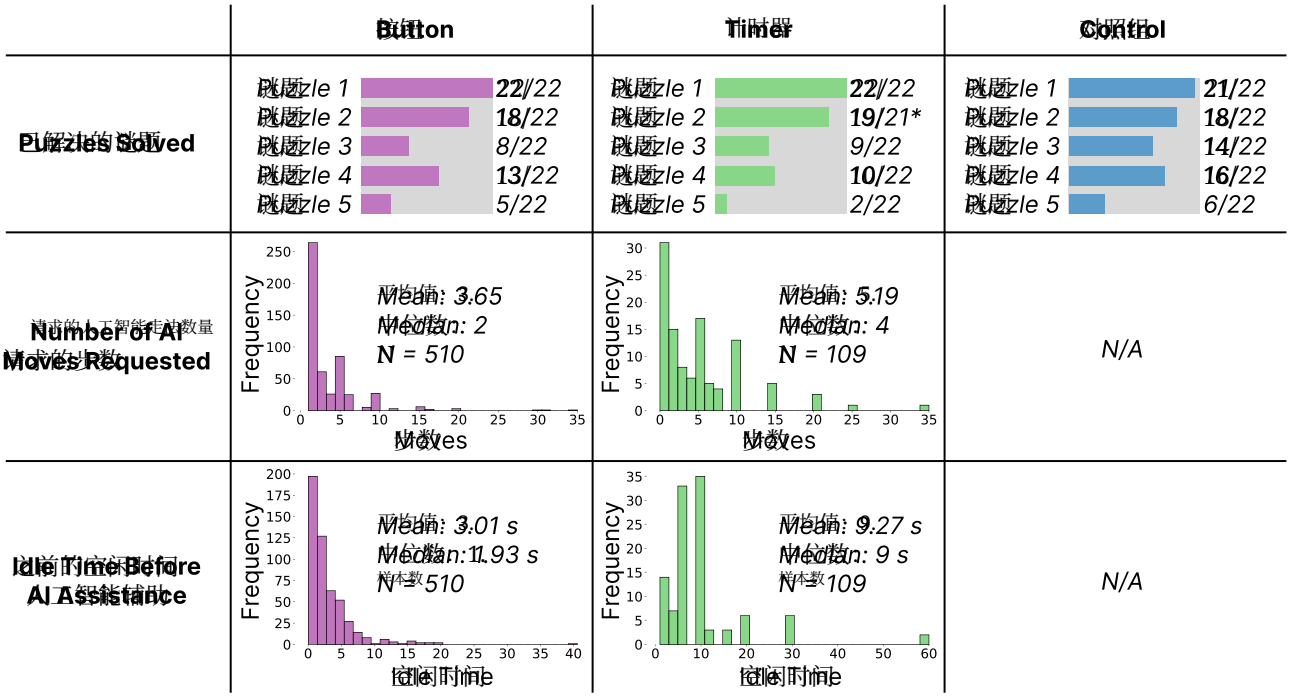


图8：已解决的谜题：在三种条件下解决主要研究任务谜题的参与者人数。请求的人工智能走法数量：在按钮条件下，这反映了每次请求人工智能辅助的情况（$N = 510$）。在计时器条件下，这对应于用户在启动每个谜题前预设的人工智能步数（$N = 109$）。人工智能辅助前的空闲时间：在按钮条件下，这是指每次请求人工智能辅助前测得的空闲时间（$N = 510$）。在计时器条件下，这反映了用户在启动每个谜题前设置的空闲时间。注：此处报告的平均值与第***5.2.2***节中的数值不同；后者仅针对至少触发过一次人工智能辅助的谜题，计算了每个谜题的平均值。***由于数据缺失，舍弃了一组参与者与谜题的配对数据。**

vs. 计时器：$p = 0.069$）。NASA-TLX的其他分量表未观察到显著差异。

5.4 研究后问卷

为了评估参与者对两种人工智能辅助模式下人工智能的看法，我们首先将问题分为五类：人工智能的能力与可靠性、AI的帮助性与有效性、AI所提供帮助的质量与充分性、AI可解释性，以及与AI的团队合作。完整的问题列表请参见附录A。我们使用Cronbach's alpha [14]评估了各子类别的内部一致性：人工智能的能力与可靠性（$\alpha = 0.966$）、AI的帮助性与有效性（$\alpha = 0.906$）、AI所提供帮助的质量与充分性（$\alpha = 0.873$）、AI可解释性（$\alpha = 0.864$）以及与AI的团队合作（$\alpha = 0.933$）。按照通常建议的信度阈值 0.8 [48]，这些数值证实所有子类别均具有较强的内部一致性。

结果显示，两种条件在人工智能的帮助性与有效性（曼-惠特尼检验：$p =$ 0.097 $= 0.293$，秩双列相关）以及人工智能提供的帮助的质量和充分性（独立样本t检验：$p = 0.063$，科恩d值 $= -0.576$）方面存在边缘性差异。在这两种情况下，计时器条件的评分都较高。为进一步探究这些差异，我们进行了分析采用协方差分析，将条件作为固定效应，并将主研究阶段的任务表现指标作为协变量进行控制，每个模型仅纳入一个协变量。通过这种方法，我们能够在控制这些变量的同时，考察不同辅助模式下对人工智能的感知有何差异。

5.4.1 人工智能的能力与可靠性。计时器条件下的参与者报告的评分（$M = 5.273$，$SD = 1.804$）高于按钮条件下的参与者（$M = 4.606$，$SD = 1.711$），但这一差异在统计学上并不显著（$p = 0.210$）。协方差分析显示两个显著的协变量：帮助被使用的总次数（$p = 0.018$）以及主研究阶段的最大进度（$p = 0.013$），二者均对评分有显著的正向影响。

*5.4.2 AI*的帮助性与有效性。与按钮条件相比，计时器条件下的评分呈现较高趋势（$M = 5.045$，$SD =$ 1.616 $M = 4.080$ $SD = 1.909$），但这一差异未达到统计显著性（$p =$ 0.097）。协方差分析显示，在控制帮助使用总次数或主研究阶段的最大进度后，计时器条件下的评分均显著高于按钮条件（$p = 0.028$ 和 $p =$

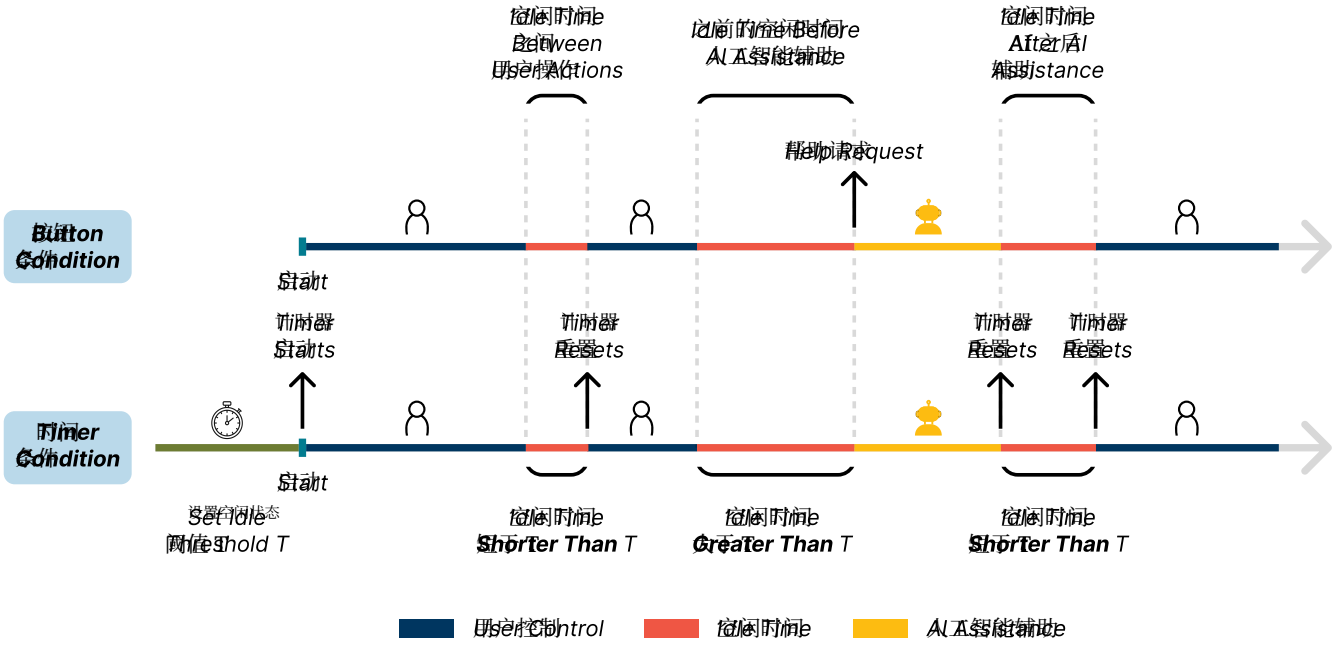


图9：此图展示了两种人工智能辅助条件下模拟的用户流程。按钮条件：用户先自行移动拼图块（用户控制），随后短暂空闲，以思考下一步行动。又进行几步后，用户感到无从下手，并空闲更长时间。之后，用户决定寻求人工智能辅助，于是指定需要辅助的步数并点击帮助按钮，从而触发人工智能辅助。人工智能完成移动后，用户短暂空闲，规划下一步行动，然后继续解谜。计时器条件：开始解谜前，用户设置两个参数：空闲时间阈值（例如，10秒）以及触发时人工智能应提供的步数（例如，5步）。游戏过程中，用户执行选择或移动拼图片等操作时，计时器会重置，且不会提供辅助。但是，如果用户持续空闲的时间超过阈值，就会自动触发人工智能辅助。随后，人工智能执行指定步数，之后计时器重置。

0.039，分别如此）。两个协变量也都对评分产生了显著的正向影响（分别为$p = 0.005$ 和 $p = 0.024$）。

*5.4.3 AI*所提供帮助的质量与充分性。 按钮条件（$M = 3.795$, $SD = 1.867$）与计时器条件（$M = 4.784$, $SD = 1.549$）之间观察到边缘显著差异（$p = 0.063$）。ANCOVA分析显示，在控制帮助使用总次数后，计时器条件的评分显著高于按钮条件（$p = 0.040$）。同样，在控制主研究阶段的最大进度后，计时器条件的评分仍显著高于按钮条件（$p = 0.034$）。最大进度本身也成为显著的协变量（$p = 0.043$）。

*5.4.4 AI*可解释性。 按钮条件（$M = 3.985$, $SD = 1.798$ $M = 4.485$ $SD = 1.778$ ）与计时器条件（，）之间的可解释性评分没有显著差异。不过，帮助使用总次数和主研究阶段达到的最大进度均成为显著的协变量（分别为$p = 0.025$ 和 $p = 0.029$），且二者都与较高评分呈正相关。

5.4.5 与*AI*的团队合作。在按钮条件（$M = 4.055$, $SD = 1.924$）与计时器条件（$M = 4.755$, $SD = 1.608$）之间的团队合作评分未发现显著差异。然而，主研究阶段的帮助使用总次数被发现是一个显著的协变量（$p = 0.001$），且与评分呈正相关。

5.4.6 先前人工智能经验和 *ATI* 的影响。我们考察了先前人工智能经验或 ATI 评分是否会影响参与者对人工智能的主观评分。先前人工智能经验以四级分类变量衡量，并采用方差分析进行分析，结果显示对任何主观评分指标均无显著影响。ATI 评分则采用皮尔逊相关进行分析，同样未发现其与任何指标存在显著关联（均 $p > 0.05$）。

6 定性结果

前三位作者使用框架分析 [23]分析了开放式问题的回答。分析首先通过主动阅读和反复阅读用户回答来熟悉数据。随后采用演绎-归纳混合方法建立初步编码框架；接着通过识别模式、关系和总体主题来解读编码后的数据。编码在语义层面进行，

表 1：协方差分析结果：各类别代表用户在研究后问卷中对人工智能的感知，并将主研究阶段的一项额外绩效指标纳入为协变量。

类别	平均值（标准差）	使用的帮助总量				剩余预算				最大进度			
		条件		协变量		条件		协变量		条件		协变量	
		p	η^2	p	η^2	p	η^2	p	η^2	p	η^2	p	η^2
AI's 能力 and 可靠性	按钮：4.6 (1.7) 计时器：5.3 (1.8)	0.119	0.058	**0.018***	0.128	0.225	0.036	0.767	0.002	0.117	0.059	**0.013***	0.141
AI's 有用性 and 有效性	按钮：4.1 (1.9) 计时器：5.0 (1.6)	**0.028***	0.112	**0.005***	0.173	0.083	0.071	0.559	0.008	**0.039***	0.100	**0.024***	0.118
质量和充分性 人工智能提供的 Help	按钮：3.8 (1.9) 计时器：4.8 (1.5)	**0.040***	0.099	0.102	0.064	0.066	0.080	0.996	< 0.001	**0.034***	0.105	**0.043***	0.097
人工智能的可解释性	按钮：4.0 (1.8) 计时器：4.5 (1.8)	0.228	0.035	**0.025***	0.117	0.371	0.020	0.704	0.004	0.238	0.034	**0.029***	0.110
团队合作 与人工智能	按钮：4.1 (1.9) 计时器：4.8 (1.6)	0.075	0.075	**0.001***	0.220	0.210	0.038	0.251	0.032	0.129	0.055	0.054	0.088

并且在整个分析过程中反复完善该框架。下一节将介绍我们的研究发现。

6.1 人工智能的能力与有用性

两种条件下的参与者普遍认为人工智能能力出众且可靠，许多人指出它提供的步数合乎逻辑且有效。例如，P32（按钮）表示“它知道下一步该怎么走，”而 P54（计时器）则评论道“它反应迅速，而且在我无法推进谜题时，能可靠地继续解题。”参与者认可人工智能辅助的高效性。P45（计时器）称人工智能“效率很高”，P60（计时器）则指出“它所走的步数都很高效。”

然而，在按钮（N = 5）和计时器（N = 3）条件下，都有少数参与者对人工智能在解谜中的实用性表示保留意见，指出人工智能的一些步数似乎毫无意义或只是随机操作。例如，P33（按钮）认为，人工智能“并没有真正帮上忙，因为它只是移动方块，却没有提供解决方案。”同样，P24（按钮）评论道：“在我看来，这些步数似乎只是随机的，背后没什么意义。”另一位参与者 P27（按钮）指出，人工智能的干预往往是多余的：“它通常只是重复我已经完成的步数我在它介入之前就已经完成了，所以不太有用。”计时器条件下也有人表达了类似的担忧。P64（计时器）表示：“在我看来，它似乎是在浪费步数，因为这些操作看起来毫无意义。”P50（计时器）总结道，人工智能“完全没有帮助。”

6.2 辅助的对照组与及时性

在两种人工智能辅助模式下，参与者报告的可控性水平相近，但他们对人工智能介入方式的感知略有不同。在按钮条件下，参与者强调了控制权的价值。P41（按钮条件）表示，“我喜欢自己能够控制它走多少步，”P23（按钮条件）也提到，“它让我可以选择希望人工智能辅助多少步。”计时器条件下的参与者也很看重能够调节帮助何时以及如何提供。P64（计时器条件）解释说，他们很看重“你可以选择它以什么间隔 *[AI]* 来帮助你，”P66（计时器条件）也喜欢“能够自定义它何时介入，以及它会走多少步。”除了这种控制感之外，计时器条件下的一些参与者还觉得这种辅助积极主动、及时有效。P61（计时器条件）很欣赏“人工智能会在我卡住时介入，并提供帮助我继续前进的合理移动。这感觉就像

一种温和的推动，而不是完全接管。”同样，P58（计时器条件）提到，他们喜欢“它会在我卡住时介入并提供帮助。”

6.3 对解释的需求

多位参与者(N = 5)强调，解释是需要重点改进的方面。当被问及“你不喜欢人工智能的哪些地方？”时，难以理解人工智能步数背后的依据是一个普遍问题。P44（按钮）回答道，“我无法理解它的逻辑。”同样，P51（计时器）指出，“有时你不明白它为什么要那样做，”而P42（按钮）表示，他们“无法判断它在某些步数中到底想做什么。”缺少解释常常使参与者感到困惑，偶尔还会让他们怀疑人工智能是否故意做出糟糕的选择。例如，P41（按钮）评论道，“有时我觉得它花的时间太长，而且会故意选择错误的步数。”同样，P56（计时器）表示，“我不喜欢自己无法始终判断它是否在帮助我。”

当被问及人工智能应该做出哪些不同之处时，参与者反复要求提高可解释性。P45（计时器）建议“在其步数中加入解说——也就是说，解释它为什么那样做，这样你就能从它停下的地方继续。”同样，P34（按钮）提议，“人工智能可以为其建议的步数提供清晰的解释。”因为这“将有助于建立信任和理解。”

6.4 期望的人工智能行为

参与者提出了各种建议，希望人工智能的行为能更符合自己的喜好。两种条件中反复出现的主题是速度，许多人建议人工智能应当移动得更快。P26（按钮）和 P28（按钮）都批评人工智能“速度不够快”和“太慢了。”类似地，P65（计时器）评论道：“我原以为人工智能的动作会很快。”

除了速度之外，参与者还表达了对另一种条件的功能的偏好，不过他们并不知道这种条件的存在。在按钮条件下，参与者建议人工智能应根据情境采取更主动的行动。P23（按钮）建议，人工智能应该“在你一段时间内没有做出决定后，主动提出建议。”P41（按钮）期望人工智能“知道什么时候我需要它启动 *[它]*，并且能自主判断。所以，如果我遇到困难，人工智能就可以介入。”相反，一位参与者（P59，计时器）希望拥有更多手动控制，并指出“很难界定什么时候是请求帮助，什么时候是让它替你完成——如果通过按钮启动会更好。”

6.5 无辅助解谜行为

对照条件（N = 22）的参与者报告了两种主要的解谜方法：试错（N = 9）和系统规划（N = 12）。试错方法是指几乎不做规划、快速移动，通过测试不同的局面来看看哪种有效。系统规划方法则包括回溯红色汽车的位置、先移动较大的阻挡车辆，并提前规划几步，有时还会从目标状态开始倒推。一位参与者（P17）表示自己没有系统性的策略，并对未能解开谜题感到沮丧。

对照条件参与者对辅助的偏好。当被问及偏好的帮助类型时，大多数参与者（N = 15）希望在关键节点获得轻微提示或情境线索，通常只涉及一步。具体请求包括查看第一步的选项（P3），或在“较长车辆的末端会阻挡路径”时收到提示（P21）。P1希望提示数量有限，并表示提示“会剥夺解开*[谜题]*的满足感。”相反，一位参与者（P12）表示，卡住时希望将整个谜题交给人工智能解决；另一位则请求放宽谜题约束，以减轻难度和挫败感。一位采用试错策略的参与者（P19）表示，如果能获得高层次的策略指导，会很有帮助。

理想的人工智能角色。 大多数参与者（N = 17）表示，希望人工智能充当工具。P16解释这一选择时强调，他们不希望受到“队友身份带来的干扰”，因为他们认为这会导致过多交流；他们更喜欢一种只提供步数的基础实用工具。另一位参与者（P11）更希望人工智能充当工具，因为他们不希望人工智能“表现得像个人类”，通过“对 *[他们]*的 *[表现]* 大喊大叫。”

相比之下，只有少数参与者表示，如果人工智能像教练一样发挥作用，会帮助他们学习和进步。P6表示，“人工智能应该像一位教练，努力让玩家下次成为更好的规划者，从而顺利完成游戏”，P19则希望人工智能“像一位教练，因为我可以不断成长。”

无论偏好哪种人工智能角色，参与者都强烈希望掌控解谜过程，并估计自己应保持对步数80%以上的控制权，以“获得成就感”（P3）。

辅助时机。 在被问及辅助时机时，参与者根据各自的解题策略表达了不同偏好。大多数采用系统化规划的参与者（N = 10/12）表示，他们更倾向于按需帮助。相比之下，依赖试错的参与者（N = 5/9）表示，他们更偏好在失败尝试后触发的情境化帮助。一位参与者（P19）更倾向于不接受辅助，并表示：“只要有解，我就不需要人工智能帮助。”

7 讨论

我们的研究比较了两种人工智能辅助提供模式：一种通过按下按钮显式触发辅助（按钮模式），另一种则在用户设定的无操作时段后自动触发辅助（计时器模式）；此外，研究还设置了一个对照条件。参与者在两种人工智能辅助模式和控制组中的任务进展相似。尽管表现相近，计时器条件下的参与者对人工智能的印象更积极。在下文中，我们将讨论研究发现及其潜在影响。

7.1 人工智能辅助与无协助

资源分配。 与无辅助对照条件相比，人工智能支持条件显著提升了移动精度。这一效果归功于算法生成的精准人工智能移动，它们降低了人类在无辅助情况下表现固有的波动性。尽管获得人工智能支持的参与者

在进展方面曾短暂领先，但人工智能辅助并未转化为统计意义上更优的整体任务表现，这一模式与先前研究结果一致 [65]。虽然这可能是具体实验成本结构所致，但它凸显了在人力投入（时间）与成本高昂的人工智能辅助之间实现实时资源最优分配所面临的挑战。从理论上讲，要最大限度地提升进展，就必须持续评估独立解决问题与人工智能辅助之间的权衡。只有当独立解决问题的进展速度超过成本高昂的人工智能辅助所带来的边际收益时，独立解决才是有效的。未能优化资源分配的参与者付出了更高成本，却未能相较于对照条件取得相应的整体进展。此类次优资源分配或许反映了人机协作中更广泛的挑战：决策者未能准确判断何时应请求辅助、何时应依靠独立努力，以最大化整体任务结果 [59, 60]。

用户感知与自动化偏见。 我们的主观结果显示，人工智能使用频率与用户对人工智能能力和有用性的正面评价呈正相关。这些积极印象似乎强化了用户持续使用人工智能的意愿：参与者更看重感知到的进展，而非因此付出的代价，从而增加了总体成本。这种行为模式表明，存在一个自我强化的循环：最初从人工智能辅助中取得的成功可能增强了信任和满意度，进而促使用户进一步依赖人工智能。虽然这种行为并不属于典型的自动化偏见——即用户不加批判地服从自动化决策 [49], ——但它反映了一种由信任驱动的过度依赖动态，与自动化偏见背后的机制相符。

人工智能作为工具还是伙伴。 定性数据表明，许多用户将人工智能视为一种智能、可靠的工具，而非协作伙伴，并期望它能够审慎地结合情境参与任务。这一发现表明，人们对人工智能关系定位的理解取决于情境。在追求高效率、以任务为导向的场景中，实用性优先于关系支持，这与在关系性和情感性更强的情境中得出的发现形成对比 [51, 52]。然而，即使将人工智能视为工具，用户对人工智能辅助的接受程度也并不一致。在我们的研究中，主观反馈各不相同：有些人认为人工智能的行为合乎逻辑且有效，有些人则对其能力和意图感到困惑与不信任。部分不信任可能归因于脱离循环绩效问题等因素 [16], ；人工智能行为缺乏清晰解释，可能加剧了用户的疏离感，抵消了较短干预间隔的缓解作用。这表明，要让人工智能成为“可靠的工具”，除了及时干预，还必须具备可解释性。

7.2 人工智能辅助的交付机制与用户感知

尽管人工智能辅助的触发方式有所不同，两种条件下的参与者最终呈现出相近的任务表现和总体人工智能使用情况。这种趋同或可用有限理性和满意化 [58], 来解释：有限理性和满意化认为，个体会受到认知和信息约束，因此依靠简单的“足够好”规则做决策，而非进行穷尽式优化。在时间压力和明确的金钱成本下，参与者不太可能计算出最优的求助策略，而是依靠简单的启发式方法，例如在进展停滞时请求辅助。鉴于参与者的技能水平相近，他们很可能在任务的相近阶段陷入僵局。在按钮条件下，这些僵局会触发明确的帮助请求；而在计时器条件下，则表现为更长时间的停顿，从而触发不活动触发协助。因此，两种条件下的参与者可能都依赖同一种基于“卡住感”的底层决策规则，进而满意化地趋向人工智能使用的成本高效均衡，使不同条件下的表现相近。

尽管两种人工智能辅助模式带来了相近的表现和预算效率，但它们却产生了截然不同的用户体验。在控制协助频率和取得的最大进度后，计时器条件下的参与者认为人工智能的帮助作用和支持性显著更强。这种差异表明，交付模式本身会影响参与者满意度，且独立于他们实际获得的辅助程度。计时器条件下的参与者认为人工智能的干预恰逢其时，指出它会在需要时“介入”，让人感觉有一个支持性的存在。这些发现表明，交付模式会调节用户对人工智能协作意图的感知：时机恰当时，相同的辅助可能被视为恰到好处的支持；时机不当时，则可能显得不足或具有干扰性。因此，协助提供模式是影响用户体验混合主动系统的一项关键设计因素。

尽管这两种人工智能辅助提供模式旨在实现相同的功能结果，但它们引发了不同的认知与体验权衡。从可调节自主性的角度来看， [9], 这种自主性描述了用户与自动化系统如何灵活分配主动权和控制权，按钮模式与计时器模式代表了在协助激活方面对主动权的不同分配。在按钮模式下，用户必须主动决定何时请求辅助，因此需要额外调动认知资源来监控并自我评估对帮助的需求。相比之下，计时器模式或许能让参与者通过自动激活辅助，在便利性与舒适度之间取得平衡：主动权可以部分委托给系统，但激活时机受用户设定的空闲时间阈值调节。与随时请求帮助相比，这或许让系统采取行动的时机显得更恰当，也更省力；这也可能解释了为何计时器条件下的参与者对人工智能表现出更积极的态度。尽管在计时器条件下，参与者对人工智能辅助即时激活的直接控制有所减少，但他们并未报告能动性有所丧失。这表明，感知自主性可能既取决于明确的控制机制，也取决于用户能否将系统的委托程度设定在自己可以接受的水平。在恰当的时机提供辅助，即使是根据用户设置自动触发，也可能保留用户的自主感。实际控制与感知控制之间的这种区别，为混合主动系统的设计提供了重要启示：系统应在不要求用户持续决策的前提下支持其能动性。

7.3 设计启示

支持校准式辅助决策。关于剩余预算的定量结果表明，参与者难以在独立问题解决与需要付出成本的人工智能辅助之间优化成本效益权衡。这说明，系统还应帮助用户决定何时启用人工智能辅助这一元决策，以提升整体效率。系统或许可以监测用户的行为线索，例如注视模式、情绪状态和交互行为，并根据与用户内部状态和表现相关的可观察信号，判断最佳辅助时机 [6, 44]。

可配置主动性。 用户对控制权的需求与其对人工智能根据情境采取行动的期待相结合，表明未来的混合主动系统应支持可配置主动性，让用户能够根据任务特定条件和情境线索设定人工智能干预触发条件。重要的是，这些配置机制应基于可观察且易于人类理解的参数，例如不活动时段或特定任务状态，使自动化辅助符合用户对寻求帮助的心智模型。

缓解高成本辅助情境中的自动化偏见。 当人工智能辅助既精准又代价高昂时，即使它未能提升整体任务效率，用户对其能力的正面感知仍可能促使其增加使用。我们的结果表明，对人工智能辅助下任务进展的良好印象会促使用户持续参与，增加成本，却未带来统计意义上更优的表现。因此，设计者应预见到以效率为导向的任务中可能出现自我强化式过度使用，并提供反馈，让用户看清长期成本与表现之间的权衡。

让人工智能操作具有可解释性，以支持无缝交接。 除了辅助时机，用户对人工智能决策依据的理解对于校准信任和促进交接也至关重要。我们的定性研究结果表明，一些参与者认为人工智能的步数在策略上合理，另一些人却觉得这些步数毫无章法，这可能会让他们在辅助结束后更难继续解决问题。这说明，在序列式问题解决任务中，仅展示单个人工智能操作可能不足以实现可解释性。设计者应考虑提供额外线索来传达人工智能的高层意图，例如一系列步数的中间目标或长期策略，以支持意义建构并实现更顺畅的交接。

8 局限性与未来工作

虽然我们的研究结果表明，协助提供模式会影响用户在协作式问题解决中的体验，但在解读和推广这些结果时，应考虑若干局限性。我们选用 Rush Hour谜题作为测试平台，因为它们体现了现实世界协作式问题解决中的若干重要特征，包括时间压力、序列化决策、策略规划以及寻求帮助的机会。然而，谜题本质上是定义明确的问题，具有清晰的目标和结构化的解法。因此，我们的研究结果未必能直接推广到更开放或更具创造性的任务中；在这些任务中，问题定义可能含糊不清，可能存在多种有效方法，或者成功标准具有主观性。未来研究应系统地考察不活动触发协助能否在更广泛的任务类型中保持其体验优势。我们的研究在单次会话中收集用户感知和偏好，参与者对每种辅助模式的体验有限。虽然这种设计有助于我们分离即时的体验差异，却无法反映随着使用时间延长，偏好和行为可能发生的变化。用户可能会逐渐习惯最初偏好的模式，发现改变其协助需求的新策略，或者随着对任务或人工智能能力的了解加深而重新调整预期。需要开展纵向研究，以评估这些偏好是否会随时间持续，以及用户是否会形成更成熟的策略，以有效利用不同的辅助模式。了解这些随时间变化的动态，有助于设计出在长期使用过程中始终有效且令人满意的协助系统。我们比较了两种特定的辅助模式，它们代表了人机协作设计空间中的重要方向。然而，还可以设计许多其他时序机制。例如，自适应系统可以依据习得的用户偏好、生理信号或绩效模式，动态调整干预时机。按固定间隔触发的定时辅助、考虑任务阶段或环境因素的情境感知系统，或结合多种触发机制的混合方法，可能带来不同的体验特性。虽然我们的研究结果表明，时序模态对用户体验很重要，但仍需进一步研究，以刻画完整的设计空间，并确定哪些机制最适合不同的情境和用户群体。

9 结论

在混合主动系统中，确定人工智能何时应介入协作式问题解决，仍是一项长期挑战。现有方法侧重于优化辅助时机，而我们的研究表明，辅助的交付机制本身也会影响用户体验，这一影响独立于其对任务结果的作用。对于协作式人工智能系统的设计者，我们的研究结果表明，不活动触发协助是一种实用的模式，能够在用户控制与减轻认知负担之间取得平衡。设计者不应将辅助时机视为次要的实现细节，而应将交付方式作为首要设计考量，因为它会直接影响用户能动性和参与度。随着人工智能系统的能力日益增强，理解并优化其融入人类工作流程的方式，对于打造以人为本、既有效又令人满意的协作体验至关重要。

10 生成式人工智能使用情况披露

我们使用大语言模型探索潜在的数据分析方法，并提出改善文本措辞和清晰度的建议。大语言模型提供的任何建议仅作为参考，用于辅助我们自身的决策。在研究界面开发过程中，我们也使用了人工智能工具进行代码调试。

参考文献

[1] Gregory D. Abowd、Anind K. Dey、Peter J. Brown、Nigel Davies、Mark Smith 和 Pete Steggles。1999年。《迈向对情境及情境感知的更好理解》。载于手持与普适计算，Hans-W. Gellersen（编）。

施普林格柏林海德堡，柏林，海德堡，304–307。doi:10.1007/3-540- 48157-5 29 [2] Ken A Aho。2013。使用 R 为生物学家讲授基础统计学与应用统计学。CRC出版社。https://doi.org/ 10 1201 16126 。/b J.E. Allen、C.I. Guinn 和 E. Horvtz。1999。混合主动性交互。《IEEE智能系统及其应用》14，5（1999），14–23。doi:10.1109/5254. 796083[4] Saleema Amershi、Dan Weld、Mihaela Vorvoreanu、Adam Fourney、Besmira Nushi、Penny Collisson、Jina Suh、Shamsi Iqbal、Paul N. Bennett、Kori Inkpen、Jaime Teevan、Ruth Kikin-Gil 和 Eric Horvitz。2019。《人机交互指南》。载于《2019年CHI计算系统中的人因会议论文集》（苏格兰格拉斯哥）（CHI ’19）。美国计算机协会，纽约，NY，美国， 1 13 10 1145 3290605 3300233 – 。doi: . / . Bruno Berberian、Jean-Christophe Sarrazin、Patrick Le Blaye 和 Patrick Haggard。2012。《自动化技术与控制感：洞察人类能动性的窗口》。《PLOS ONE》 7, 3 03 2012 1 6 10 1371 0034075 ( ), – 。doi: . / journal.pone. Robert Bixler、Nathaniel Blanchard、Luke Garrison 和 Sidney D’ Mello。2015。《结合注视与生理指标自动检测阅读时的走神》。载于《2015年ACM多模态交互国际会议论文集》（美国华盛顿州西雅图）（ICMI ’15）。美国计算机协会，纽约，NY，美国， 299 306 10 1145 2818346 2820742 – 。doi: . / . Anna Bodonhelyi、Efe Bozkir、Shuo Yang、Enkelejda Kasneci 和 Gjergji Kasneci。2024。《用户意图识别及其对大语言模型的满意度：一项关于ChatGPT的用户研究》。arXiv:2402.02136 [cs.HC] https://arxiv.org/abs/2402. 02136[8] James V. Bradley。1958。《拉丁方设计中即时顺序效应的完全平衡处理》。《美国统计学会杂志》53，282（1958），525–528。http://www.jstor.org/stable/ 2281872 [9] Jeffrey M. Bradshaw、Paul J. Feltovich、Hyuckchul Jung、Shriniwas Kulkarni、William Taysom 和 Andrzej Uszok。2004。《可调节自主性与混合主动性交互的维度》。载于《智能体与计算自主性》，Matthias Nickles、Michael Rovatsos 和 Gerhard Weiss（编）。施普林格柏林海德堡，柏林，海德堡，17–39。[10] Mollie E Brooks、Kasper Kristensen、Koen J Van Benthem、Arni Magnusson、Casper W Berg、Anders Nielsen、Hans J Skaug、Martin Mächler 和 Benjamin M Bolker。2017。glmmTMB在零膨胀广义线性混合模型软件包中兼顾速度与灵活性。《R期刊》9，2（2017），378–400。doi: 10 32614 2017 066 . /RJ- - Teresa Castle-Green、Simon Castle-Green、Joseph Lindley、Neelima Sailaja 和 Susan Lechelt。2025。《Elmo：面向社区维修咖啡馆的具身对话助手》。载于《第七届ACM对话式用户界面会议论文集》（CUI ’25）。美国计算机协会，纽约，NY，美国，文章7，6页。doi:10.1145/3719160.3737627[12] Valerie Chen、Alan Zhu、Sebastian Zhao、Hussein Mozannar、David Sontag 和 Ameet Talwalkar。2025。《需要帮助吗？面向编程的主动式人工智能助手设计》。载于《2025年CHI计算系统中的人因会议论文集》（CHI ’25）。美国计算机协会，纽约，NY，美国，文章 881, 18 10 1145 3706598 3714002 页。doi: . / . Nazli Cila。2022。《设计人与智能体协作：承诺、响应能力与支持》。载于《2022年CHI计算系统中的人因会议论文集》（美国路易斯安那州新奥尔良）（CHI ’22）。美国计算机协会，纽约，NY，美国，文章420，18页。doi:10.1145/3491102.3517500 [14] Lee J Cronbach。1951。《阿尔法系数与测验的内部结构》。《心理计量学》 16, 3 1951 297 334 10 1007 02310555 ( ), – 。doi: . /BF Fred D. Davis。1989。《感知有用性、感知易用性与信息技术的用户接受度》。《MIS季刊》13，3（1989），319–340。http://www.jstor.org/stable/ 249008 [16] Mica R Endsley 和 Esin O Kiris。1995。《自动化中的脱离循环绩效问题与控制水平》。《人为因素》37，2（1995），381–394。doi:10. 1518/001872095779064555[17] Alexander Erlei、Richeek Das、Lukas Meub、Avishek Anand 和 Ujwal Gadiraju。2022。《物有所值：人们宁愿放弃经济自利，也不愿与人工智能系统议价》。载于《2022年CHI计算系统中的人因会议论文集》（美国路易斯安那州新奥尔良）（CHI ’22）。美国计算机协会，纽约，NY，美国，文章113，18页。doi:10.1145/3491102.3517734[18] Cathy Mengying Fang、Wazeer Zulfikar、Yasith Samaradivakara、Suranga Nanayakkara 和 Pattie Maes。2025。《可穿戴人工智能系统主动干预的适中时间窗》。arXiv:2504.09332 [cs.HC] https: // arxiv.org/abs/ 2504 09332 。Rosanna Fanni、Valerie Eveline Steinkogler、Giulia Zampedri 和 Jo Pierson。2023。《通过救济增强人工智能系统中的人类能动性》。《AI与社会》38，2（2023），537–547。doi:10.1007/s00146-022-01454-7[20] Davide Fossati、Barbara Di Eugenio、Stellan Ohlsson、Christopher Brown 和 Lin Chen。2010。《生成主动反馈，帮助学生保持进度》。载于《第十届智能导学系统国际会议论文集——第二卷》（宾夕法尼亚州匹兹堡）（ITS’10）。施普林格，柏林，海德堡， 315 317 10 1007 978 3 642 13437 1_56 – . doi: . / - - - - John Fox 和 Sanford Weisberg。2019。《应用回归》R语言配套指南。Sage，千橡市 CA。https://www.john-fox.ca/Companion/ [22] Thomas Franke、Christiane Attig 和 Daniel Wessel。2019。技术交互的个人资源：技术交互亲和力（ATI）量表的开发与验证。《国际人机交互期刊》 35, 6 2019 456 467 10 1080 10447318 2018 1456150 ( ), – 。doi: . / . . Nicola K Gale、Gemma Heath、Elaine Cameron、Sabina Rashid 和 Sabi Redwood。2013。使用框架法分析多学科健康研究中的定性数据。BMC医学研究方法学 13，1（2013），117。doi:10.1186/1471-2288-13-117[24] Sandra G. Hart 和 Lowell E. Staveland。1988。NASA-TLX（任务负荷指数）的开发：实证与理论研究结果。载于《人的心理工作负荷》，Peter A. Hancock 和 Najmedin Meshkati（编）。《心理学进展》，第 52 139 183 10 1016 0166 4115 08 62386 9 卷。北荷兰， – 。doi: . /S - ( ) - Florian Hartig。2016。. DHARMa：分层（多层/混合）回归模型的残差诊断。CRAN：贡献软件包（2016）。[26] Martin Hoegl 和 Hans Georg Gemuenden。2001。团队合作质量与创新项目的成功：理论概念与经验证据。《组织科学》 12, 4 2001 435 449 10 1287 12 4 435 10635 ( ), – 。doi: . /orsc. . . . Steffen Holter 和 Mennatallah El-Assady。2024。. 解构人机协作：能动性、交互与适应。载于《计算机图形学论坛》，第 43 15107 10 1111 15107 卷。Wiley在线图书馆，e 。doi: . /cgf. Eric Horvitz。1999。. 混合主动用户界面的原则。载于《SIGCHI计算系统中的人因会议论文集》（美国宾夕法尼亚州匹兹堡）（CHI ’99）。美国计算机协会，纽约州纽约， 159 166 10 1145 302979 303030 – 。doi: . / . Eric Horvitz、Andy Jacobs 和 David Hovel。1999。注意力敏感警报。载于第十五届人工智能中的不确定性会议论文集（瑞典斯德哥尔摩）（UAI’99）。Morgan Kaufmann出版社，旧金山，加利福尼亚州，305–313。[30] Hangxing Huang、Lu Zhang、Yongyu Yang、Ling Huang、Xikui Lu、Jingyang Li、Huimin Yu、Shuqiao Cheng 和 Jian Xiao。2024。用药提醒系统的构建与应用：通用用药时间表的智能生成。《生物数据挖掘》17，1（2024），23。doi:10.1186/s13040-024-00376-y [31] Anthony Jameson 和 Eric Schwarzkopf。2002。可控性的利弊：一项实证研究。载于第二届自适应超媒体与自适应网络系统国际会议论文集（AH ’02）。施普林格，柏林，海德堡，193–202。[32] Hyunjin Kang 和 Chen Lou。2022。人工智能能动性与人类能动性：理解TikTok上的人机交互及其对用户参与度的影响。《计算机中介传播期刊》27，5（2022），zmac014。doi:10.1093/ jcmc/zmac014[33] Nils Köbis、Zoe Rahwan、Raluca Rilla、Bramantyo Ibrahim Supriyatno、Clara Bersch、Tamer Ajaj、Jean-François Bonnefon 和 Iyad Rahwan。2025。委托人工智能可能会增加不诚实行为。《自然》（2025），1–9。doi:10.1038/s41586-025-09505-x[34] Kenneth R Koedinger 和 Vincent Aleven。2007。在认知导师实验中探究辅助困境。《教育心理学评论》19，3 2007 239 264 10 1007 10648 007 9049 ( ), – 。0doi: . /s - - - [35] Emily Kuang、Minghao Li、Mingming Fan 和 Kristen Shinohara。2024。通过与对话式人工智能助手协作增强用户体验评估：主动对话与时机的影响。载于2024年CHI计算系统中的人因会议论文集（美国夏威夷州檀香山）（CHI ’24）。美国计算机协会，纽约州纽约，文章3，16页。doi:10.1145/3613904.3642168[36] Robert Kubinec。2023。有序贝塔回归：一种适用于具有上下界连续数据的简约且拟合良好的模型。《政治分析》31，4（ 2023 519 536 10 1017 2022 20 ），– 。doi: . /pan. . Jonathan Lazar、Jinjuan Heidi Feng 和 Harry Hochheiser。2017。《人机交互研究方法》。Morgan Kaufmann。[38] Kevin Le Goff、Arnaud Rey、Patrick Haggard、Olivier Oullier 和 Bruno Berberian。2018。能动性会调节人与自动化技术的交互。《人因工程学》 61, 9 2018 1282 1297 10 1080 00140139 2018 1468493 ( ), – 。doi: . / . . Kyla H. Levin、Nicolas van Kempen、Emery D. Berger 和 Stephen N. Freund。2025。ChatDBG：利用大语言模型增强调试。《美国计算机协会软件工程会议录》 2, 085 2025 22 10 1145 3729355 FSE，文章FSE（6月），页。doi: . / Tianjian Liu、Hongzheng Zhao、Yuheng Liu、Xingbo Wang 和 Zhenhui Peng。2024。ComPeer：用于主动同伴支持的生成式对话代理。载于第37届美国计算机协会用户界面软件与技术年会论文集（美国宾夕法尼亚州匹兹堡）（UIST ’24）。美国计算机协会，纽约州纽约，文章117，22页。doi:10.1145/3654777. 3676430[41] Christian Meurisch、Cristina A. Mihale-Wilson、Adrian Hawlitschek、Florian Giger、Florian Müller、Oliver Hinz 和 Max Mühlhäuser。2020。探索用户对主动式人工智能系统的期望。《美国计算机协会交互、移动、可穿戴与普适技术汇刊》4，4，文章146（2020年12月），22页。doi:10.1145/3432193

[42] Edward F Moore。1959。迷宫中的最短路径。载于《开关理论国际研讨会论文集》。哈佛大学出版社，285–292。[43] Pavithra PM Nair、Gilad Gressel、Malavika Nambrath Anand 和 Krishnashree Achuthan。2025。征求还是不征求？人工智能辅助交付机制对决策的影响。《国际人机交互期刊》（2025），1–24。doi:10.1080/10447318.2025.2536617[44] Avinash Ajit Nargund、Arthur Caetano、Kevin Yang、Rose Yiwei Liu、Philip Tezaur、Kriteen Shrestha、Qisen Pan、Tobias Höllerer 和 Misha Sra。2025。理解人机协作中的模式切换：行为洞见与预测建模。arXiv: 2509 20666 [ ] 2509 20666 。cs.HC https://arxiv.org/abs/ 。[45] Benjamin A Newman、Reuben M Aronson、Kris Kitani 和 Henny Admoni。2022。跨越时空助人：将辅助视为人机交互的一种视角。《机器人与人工智能前沿》8（2022），720319。doi:10. 3389/frobt.2021.720319[46] Stefanos Nikolaidis、Yu Xiang Zhu、David Hsu 和 Siddhartha Srinivasa。2017。共享自主中的人机相互适应。载于2017年ACM/IEEE人机交互国际会议论文集（奥地利维也纳）（HRI '17）。美国计算机协会，纽约州纽约，294–302。doi:10.1145/2909824.3020252[47] 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论文集（美国德克萨斯州大学城）（IUI '21）。美国计算机协会，纽约州纽约，340–350。doi:10.1145/3397481.3450639[48] J.C. Nunnally。1994。《心理测量理论》第3版。塔塔麦格劳-希尔教育。[49] Raja Parasuraman 和 Victor Riley。1997。人与自动化：使用、误用、不用和滥用。《人为因素》39，2（1997），230–253。doi:10.1518/ 001872097778543886[50] R. Parasuraman、T.B. Sheridan 和 C.D. Wickens。2000。人与自动化交互类型和级别的模型。《IEEE系统、人类与控制论汇刊A辑：系统与人类》30，3（2000），286–297。doi:10.1109/ 3468.844354[51] Pat Pataranutaporn、Valdemar Danry、Joanne Leong、Parinya Punpongsanon、Dan Novy、Pattie Maes 和 Misha Sra。2021。用于支持个性化学习与幸福感的人工智能生成角色。《自然·机器智能》3，12（2021），1013–1022。doi:10.1038/s42256-021-00417-9[52] Pat Pataranutaporn、Sheer Karny、Chayapatr Archiwaranguprok、Constanze Albrecht、Auren R. Liu 和 Pattie Maes。2025。“我的男友是人工智能”：对Reddit人工智能社区中人机陪伴关系的计算分析。arXiv:2509.11391 [cs.HC] https://arxiv.org/abs/2509.11391[53] Liton Chandra Paul、Sayed Shifat Ahmed、Tithi Rani、Md Ashraful Haque、Tushar Kanti Roy、Md Najmul Hossain 和 Md Azad Hossain。2024。一种面向老年人的智能药物提醒套件，配备手机电话和若干健康监测功能。《Heliyon》 10, 4 2024 10 1016 2024 26308 ( )。doi: . /j.heliyon. .e[54] John Rooney 和 Michael Rooney。2009. 用药提醒系统及方法。美国专利申请 12/186,000。[55] Kristin E Schaefer。2016。测量人机交互中的信任：人机交互信任感知量表的开发。载于《自主系统中的鲁棒智能与信任》。施普林格， 191 218 10 1007 978 1 4899 7668 0_10 – 。doi: . / - - - -[56] Albrecht Schmidt。2000. 基于情境的隐式人机交互。《个人技术》 4, 2 2000 191 199 10 1007 01324126 ( ), – 。doi: . /BF[57] Hiroshi Shono。2008. Tweedie分布在单位努力渔获量分析零捕获数据中的应用。《渔业研究》93，1-2（2008），154–162。doi:10.1016/j.fishres. 2008.03.006[58] Herbert A Simon。1955。理性选择的行为模型。《经济学季刊》（ 1955 99 118 10 2307 1884852 ），– 。doi: . / Mark[59] Steyvers 和 Aakriti Kumar。2024. 人工智能辅助决策的三大挑战。《心理科学展望》19，5（2024），722–734。doi:10.1177/ 17456916231181102[60] Michelle Vaccaro、Abdullah Almaatouq 和 Thomas Malone。2024。人类与人工智能组合何时有用：系统综述与荟萃分析。《自然·人类行为》 8, 12 2024 2293 2303 10 1038 41562 024 02024 1 ( ), – 。doi: . /s - - - -[61] Niels van Berkel、Mikael B. Skov 和 Jesper Kjeldskov。2021. 人机交互：间歇式、持续式与主动式。《Interactions》28，6（2021年11月），67–71。doi:10.1145/3486941[62] Quentin Vantrepotte、Bruno Berberian、Marine Pagliari 和 Valérian Chambon。2022。利用人类能动性提升人机交互中的信心与可接受性。《认知》222（2022），105020。doi:10.1016/j. cognition. 2022 105020 。Oleksandra Vereschak、Gilles Bailly 和 Baptiste[63] Caramiaux。2021。如何评估人工智能辅助决策中的信任？实证方法综述。《美国计算机协会人机交互会议录》5，CSCW2，文章327（2021年10月）， 39 10 1145 3476068 页。doi: . / Wen Wen 和 Hiroshi Imamizu。2022。[64] 感知、行为与人机交互中的能动感。《自然综述：心理学》1，4（2022），211–222。doi:10.1038/s44159-022-00030-6

[65] Suqing Wu、Yukun Liu、Mengqi Ruan、Siyu Chen 和 Xiao-Yun Xie。2025。人类与生成式人工智能的协作提升了任务表现，但削弱了人类的内在动机。科学报告 15, 1 (2025), 15105。doi:10.1038/s41598- 025-98385-2

A 研究后李克特量表问题

A.1 人工智能的能力和可靠性

• 人工智能具备能力。• 人工智能很好地理解了谜题。• 人工智能的表现可靠。

A.2 人工智能的有用性和有效性

• 人工智能帮助我更有效地解决了谜题。• 人工智能在解决谜题方面做出了有帮助的贡献。• 人工智能关心帮助我解决谜题。• 总体而言，人工智能很有帮助。

A.3 人工智能所提供帮助的质量与充分性

• 人工智能提供了我所需的全部步数 .
• 人工智能在恰当的时机帮助了我。• 人工智能的帮助不够。• 我能够向人工智能请求适量的帮助。

A.4 AI可解释性

• 我理解了人工智能为何走出这些步数。
• 我觉得人工智能提出的建议与我的想法一致。• 人工智能的帮助打乱了我的思路。

A.5 与AI的团队合作

• 我觉得我和人工智能像一个团队一样合作。• 我和人工智能有一个共同目标。
• 我认为是我解开了谜题，还是 + 我和人工智能一起解开了谜题。• 我对我和人工智能的合作方式感到满意。

• 人工智能提供的内容与我对它的需求十分契合。

B 研究后开放式问题

B.1 AI辅助条件的问题（按钮和计时器）

• 在谜题任务中，你喜欢人工智能的哪些表现?

• 在谜题任务中，你不喜欢人工智能的哪些表现?

• 解谜时，你使用了哪些策略（如果有的话）来更有效地解开谜题?
• 你希望这个人工智能做出哪些不同的表现，才能成为你理想的解谜助手?
• 人工智能协助的步数是否影响了你解谜的方式?

• 你是如何确定自己需要多少步帮助的?

B.2 对照组的问题

在本节中，你将为设计 AI助手提出建议，帮助用户完成任务并解决问题，例如 Rush Hour。

- 解决谜题时，你使用了哪些策略（如果有的话）来更有效地解决谜题?
- 你是否曾在某个时刻感到陷入困境？如果是，什么能够帮助你摆脱困境?
- 您认为人工智能提供哪种支持最有帮助？（例如：提示、文字说明、视觉演示、逐步指导等）
- 您希望人工智能解释它为什么建议某个移动，还是只展示该移动？为什么?
- 在什么时刻你会希望人工智能为你提供辅助——持续提供、应请求提供、在多次失败尝试后提供，还是在其他条件下提供?
- 你更希望人工智能表现得像教练、队友，还是工具？为什么?
- 你希望人工智能通过提问来帮助你反思，还是直接提供建议?
- 解谜时，你希望保留多少控制权?
- 解开谜题后，你希望收到反馈吗？例如更短的替代解法或策略提示?