arXiv:2601.15064v1 [cs.HC] 21 Jan 2026

激励调优：理解并设计实证人机决策研究中的激励

SIMRAN KAUR,代尔夫特理工大学，代尔夫特，荷兰SARA SALIMZADEH,代尔夫特理工大学，代尔夫特，荷兰UJWAL GADIRAJU,代尔夫特理工大学，代尔夫特，荷兰

尽管人工智能（AI）已彻底改变多个领域的决策方式，但在人类判断至关重要的高风险决策过程中，人类判断依然举足轻重。这推动了人们探索人类与人工智能系统协同决策的方式，以发挥双方各自的优势。为探究这种动态，研究者开展实证研究，考察人类如何借助人工智能辅助进行决策，以及这种协作如何影响结果。开展此类研究的一个关键方面是参与者的作用，他们通常通过众包平台招募。此类研究的效度取决于参与者的行为，因此，可能影响这些行为的有效激励是研究设计与执行的关键部分。由此，激励方案可以成为连接研究受控环境与现实世界决策复杂性的桥梁。通过精心设计与研究目标及参与者动机相符的激励，研究者能够充分发挥实证研究在探究人机决策方面的真正潜力。本研究旨在探讨激励设计在开展实证人机决策研究中的关键作用，重点关注激励方案的理解、设计与记录。通过对现有研究进行主题综述，我们探讨了人机决策实证研究中激励设计的当前实践、挑战与机遇。我们归纳出若干反复出现的模式或主题，例如激励方案由哪些要素构成、研究者如何操控激励方案，以及激励方案可能对研究成果产生何种影响。基于这些认识，我们整理出一套指南，帮助研究者为研究设计有效的激励方案，称为激励调优框架。该框架阐述了研究者如何开展、反思并记录激励设计流程。我们倡导采用标准化而又灵活的激励设计方法，并提供有价值的见解和实用工具，希望以此推动人机决策领域形成更可靠且更具普适性的知识。

CCS 概念： • **以人为中心的计算** → **人机交互实证研究**； •**信息系统** → **激励方案**。

补充关键词和短语：实证研究、人机决策、众包、激励方案、主题分析

ACM 参考格式：
Simran Kaur、Sara Salimzadeh 和 Ujwal Gadiraju。2026。激励调优：理解并设计实证人机决策研究中的激励。1，1（2026年1月），32页。https://doi.org/10.1145/nnnnnnn.nnnnnnn

作者联系信息：Simran Kaur，代尔夫特理工大学，荷兰代尔夫特，skaur1@tudelft.nl；Sara Salimzadeh，代尔夫特理工大学，荷兰代尔夫特，s.salimzadeh@tudelft.nl；Ujwal Gadiraju，代尔夫特理工大学，荷兰代尔夫特，u.k.gadiraju@ tudelft.nl。

允许为个人或课堂用途免费制作本作品的全部或部分数字副本或纸质副本，前提是这些副本不得用于营利或商业利益，也不得为此目的进行分发，且副本须附带本声明及首页上的完整引文。必须尊重本作品中由作者以外的其他人拥有版权的部分。注明出处后允许摘录。若要以其他方式复制、再版、发布到服务器或分发至列表，须事先获得具体许可和/或支付费用。请向 permissions@acm.org 申请许可。© 2026 版权归所有者/作者所有。出版权已授权给 ACM。提交给 ACM 的稿件

1 引言与背景

为了将人工智能有效融入人类决策过程，理解人类如何与人工智能系统交互、如何将人工智能建议纳入决策，以及这种协作如何影响结果，至关重要。为此，研究者开展了实证研究，考察决策背景下人机交互的动态过程，探讨算法厌恶、信任、依赖、公平性感知、可解释性、认知偏差等多种因素 [7, 8, 17, 28, 44, 57, 84, 94, 98, 116,119]。这些研究通常会模拟现实世界的决策场景，让参与者扮演人类决策者，从而深入了解人类如何与人工智能系统交互，以及人工智能辅助的有效性 [28, 32, 79]。有些研究还会要求参与者评估人工智能生成的建议或解释的质量，或对人工智能系统与界面提供反馈，以确定人工智能模型的改进方向 [8, 19, 67, 102]。

众包市场使不同研究社群能够开展需要人类以各种身份参与的大规模实验 [39, 43, 90]。众包研究的设计涉及若干关键方面，例如设计并向众包工作者分配任务、提供参与激励、实施质量控制措施，以及将个人贡献汇总为可用数据，从而得出有价值的见解[24, 58, 78, 85]。其中，激励方案是激发并维持参与者积极性的关键，也有助于确保积极参与并维护数据质量 [78, 83]。当参与者认为自己获得了公平报酬，并受到激励鼓舞时，他们可能更愿意真正投入任务，并且会模仿他们所扮演角色的人在现实世界中的动机 [15, 104]。激励方案的设计本身就是一个值得关注的研究领域；研究者正在探索如何恰当地奖励众包工作者，同时确保不同领域研究结果的可靠性与效度 [83, 117]。常见做法包括金钱奖励，例如按任务付费或按时薪支付；也包括非货币激励，例如游戏化元素、徽章，或对高质量贡献给予平台认可 [40, 81, 83]。研究者还致力于优化金钱激励结构，在激发参与积极性与控制成本之间取得平衡 [50, 59, 106], ，同时考量公平报酬、防止剥削等伦理影响 [103, 114]。因此，理解并实施有效的激励方案，对于任何众包研究的成功都至关重要，包括人机决策的实证研究。

传统的众包任务通常由数据标注或转录等简单微任务组成[61, 64], ，而决策任务则涉及复杂的认知过程 [62]。人机决策研究的参与者通常会参与高风险决策过程，分析信息、作出判断并评估人工智能能力，最终影响我们对决策中的人机交互的理解基础 [66]。设计能够反映此类任务现实利益攸关性的众包实验可能颇具挑战，因为参与者通常是从主要以金钱作为激励的平台招募而来 [89]。因此，金钱激励——而非决策本身被感知到的利害关系——可能成为决策者的主要动机。研究者如何激励参与者，可能会影响对研究结果的推断 [70]。一些先前研究探讨了参与者动机可能影响某些结果的情况，并承认这可能是观察结果背后的原因 [32, 109]。

尽管存在这些细微差异，当前研究格局仍严重依赖简单的 [16, 22, 23, 29, 36, 63, 68, 98, 105]或临时的 [7, 10, 28, 32, 65, 75, 100, 110] 货币激励方案，而这些方案并未描述或未必能够捕捉具体的背景、研究目标和预期成果。尽管少数研究讨论了激励措施的潜在影响

对其实验的影响 [46, 47, 49, 52, 94, 107, 109], , 但也有一些研究完全没有描述激励方案 [2, 9, 33]。因此，激励方案似乎很少经过战略性定制，以与研究目标保持一致，或应对人机决策领域所面临的独特挑战。这些挑战不仅仅是确定“合适”的报酬金额，还包括确定应予鼓励、评估和激励的期望行为。以往工作中激励方案的设计和描述存在差异与不一致，凸显了一个批判式缺口——当前实践在激励措施的设计和文档记录方面缺乏标准化方法。这种缺乏凝聚力的状况也导致研究群体对激励措施所发挥作用的理解碎片化。这造成了一个重大的知识缺口，因为我们目前甚至不了解激励方案的性质及其可能带来的影响。这种“未知的未知”局面使我们难以评估现有研究发现的影响，也阻碍了未来标准化实践的发展。由于货币激励一直是众包人机决策研究中激励参与者的主要手段，因此我们将研究范围限定在货币激励上。

因此，我们迫切需要采取更细致的方法来理解和设计激励方案。首先，重要的是深入了解当前人机决策研究中的激励设计现状。其次，我们需要确定如何切实设计激励方案，同时应对这一领域特有的挑战。在此，我们还需要推动设计流程的标准化。第三，我们需要明确如何妥善记录激励方案、其设计流程及相关结果，从而为未来研究者基于现有知识开展研究铺平道路。因此，为应对本研究中的这些重要方法论挑战，我们提出以下研究问题：

研究问题1：目前，开展实证人机决策研究时，货币激励方案是如何设计的？**研究问题2**：如何通过标准化流程，恰当地为实证人机决策研究设计货币激励方案？**研究问题3**：如何通过标准化流程记录货币激励方案的设计，以促进人机决策领域的未来研究？

为回答这些研究问题，我们通过半结构化综述和定性分析，对人机决策文献进行了主题式文献综述——即对摘录进行反思性主题分析 [13] ，这些摘录描述了激励方案，或讨论了已发表研究论文中的参与者动机。

通过这项全面分析确定的主题，构成了我们为实现这一目标以及推动人机决策领域发展所作贡献的基础。我们对这些主题进行了批判式讨论与反思，并由此得出了若干见解和未来工作的方向。

我们进一步运用这些见解，开发了激励调优框架，以指导研究者为其人机决策研究调优激励方案。该框架是一项实践性贡献，有助于建立标准化的激励实践，并提高人机决策研究的可靠性。我们的工作补充了标准化众包工作研究设计的持续努力 [87, 93, 113]。我们还为该框架配备了一份报告模板，以及一个公开协作存储库，用于存储和访问

已发表研究所采用的激励方案及激励设计决策背后的理由说明。与本研究相关的所有补充材料和资源均收录于一个公开的 GitHub 存储库中。[1]

2 方法论

2.1 综述范围与纳入标准

在本研究中，我们聚焦于探讨人机协作决策任务的实证、人类受试者研究。这些研究旨在评估、理解或提升决策背景下人类的绩效和体验。

此外，由于研究问题围绕激励方案设计展开，我们也只考察招募众包工作者担任人类决策者的研究。由于众包工作者通常通过金钱奖励获得激励，因此后续分析也仅聚焦于金钱激励。由此，我们将文献综述的纳入标准界定如下：

- 所研究的任务必须围绕人机协作中的决策活动展开。排除关注其他目标任务的论文，例如调试、模型改进、共同创作或游戏。

- 研究必须涉及众包工作背景下的人类决策者。仅关注由内部人员担任人类决策者的论文（例如组织员工或大学学生）不纳入研究范围。

- 研究必须以评估为重点，评估人机协作的有效性、可用性或影响，且众包工作者需要完成任务。因此，形成性研究以及纯粹基于定性访谈或调查的研究不在纳入范围内。

2.2 搜索策略

我们发现， [66] 已分享[2] 一份研究论文清单，该清单编制时所依据的纳入标准涵盖了本研究定义的纳入标准。因此，我们将其中列出的全部 81 项研究纳入筛选，并根据本研究的具体标准重新审查。这涵盖了截至 2021 年可能相关的文献。为确保覆盖近期研究，我们针对 2021 年至 2023年11月期间发表的文献进行了搜索。我们的搜索针对该领域主要会议的论文集，包括 ACM智能用户界面会议（ACM IUI）、AAAI/ACM人工智能、伦理与社会会议（AIES）、ACM计算系统中的人为因素会议（CHI）、ACM公平、问责与透明度会议（ACM FAccT）、ACM SIGCHI计算机支持的协同工作与社会计算会议（CSCW），以及 AAAI人类计算与众包会议（HCOMP）。搜索关键词为“人机协作”和“人机决策”。

通过标题和摘要进行初步筛选后，确定了 86 篇可能相关的论文，因此共有 160 多篇论文需要进一步评估。初步筛选后，我们依据既定的纳入标准，对每篇论文进行了细致的重新评估。经过严格评估，最终选出 97 篇符合本研究范围的论文。图 1 展示了为开展主题分析而搜索、筛选和纳入研究论文的各个步骤。

[1] https://github.com/simrankaur1509/IncentiveTuning [2] https://haidecisionmaking.github.io/ 提交给 ACM 的稿件

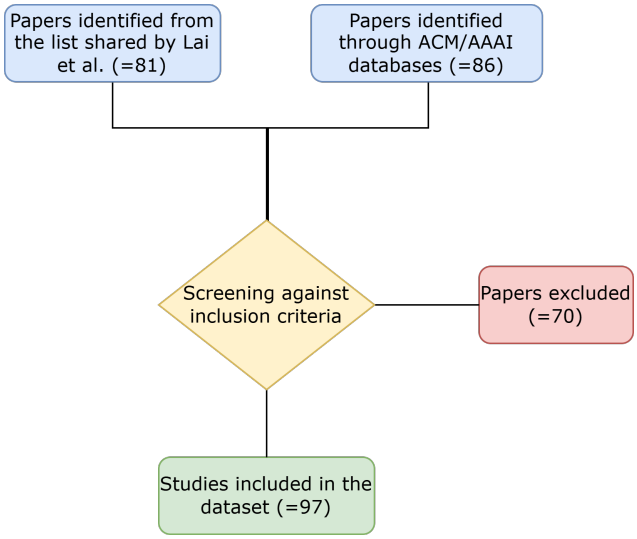


图1. 展示为主题分析创建数据集时论文筛选流程的流程图

2.3 数据集开发

我们首先将入选论文的书目信息汇总到电子表格中，记录了每篇入选论文的标题、作者、出版年份、发表场所及可访问链接等信息。随后，我们从每篇论文中提取了以下信息：研究的目标和研究目标、研究（及任务）领域、任务相关的感知风险或利害程度、研究设置的详细信息（如目标受众和参与者信息，包括参与者人数、平台、筛选标准等）、参与者的角色及其需要完成的任务、所采用激励方案的概要和任务完成时间、描述激励方案的摘录，以及讨论参与者动机或激励设计的摘录。

描述激励设计及任何围绕激励的讨论的摘录构成了主题分析的数据集。其余信息被认为与激励设计相关，因此也一并汇编，以期有助于未来的分析。该数据集已在此公开，供公众访问。

2.4 流程

针对摘录数据集的定性分析，我们采用了反思性主题分析。反思性是指有意识地承认研究者所扮演的角色，以及其对数据解读方式可能产生的影响。此外，[13] 强调，反思性不止于自我反省，还包括批判性地审视研究所产生的知识，以及我们如何产生这些知识。因此，为了践行反思性，我们明确将主题分析的方法界定为归纳式（允许主题直接从数据中浮现）、语义式（聚焦于数据的显性内容，以分析摘录中如何落实和讨论激励设计），以及批判式（旨在剖析激励设计更广泛的含义及其影响）。此外，我们按照 Braun 和 Clarke [13] 所述，执行了反思性主题分析的六个阶段——“熟悉数据”“数据编码”“初步生成主题”“主题发展与审查”“主题完善、定义与命名”以及“撰写报告”。生成的编码手册可在此处查阅。在整个过程中，我们还遵循了《15点主题分析检查清单》

由 [12] 提供，概述了开展高质量主题分析的关键标准。此清单应用于本研究所开展的主题分析，点击此处查看。

2.5 立场性

主题分析取决于研究者能否从定性数据中系统地提取意义。然而，研究者自身并非一张白纸。他们的背景、经历和偏差——这些因素构成了他们的立场性——可能会影响他们对数据的解读与分析 [13]。为了成功开展反思式 主题分析，研究者承认自己的立场十分重要。这样，研究者便能意识到自身可能存在的个人视角，同时仍以批判式的方式进行分析。因此，我们在下文通过立场性陈述来说明自己的立场。

2.5.1 立场性陈述。 我们这些开展本项分析的作者来自不同群体，在性别、国籍以及学术界和业界的专业背景方面各不相同。我们的背景主要涉及人机交互（HCI）和人工智能（AI）的不同领域。通过参与人机决策项目，我们亲身经历了针对本研究所分析的此类研究，开展激励方案研究与设计的过程。以往的经验使我们直接接触到这一领域中人类行为的复杂性，以及在人机协作研究中协调参与者动机所面临的挑战。目睹这些动态变化，让我们得以了解设计良好或欠佳的激励会如何影响这些研究。

在对人机决策研究中的激励设计进行主题分析时，我们承认自身的立场性及其可能造成的影响。尽管我们认为，兼具该领域的理论知识与实践经验，使我们能够以全面的视角开展这项研究，但这也可能使我们对这类研究中激励设计是如何开展或应当如何开展形成先入之见。我们的研究设计，尤其是纳入分析的研究之选取，受到既有知识与专业经验的影响。这种做法虽然高效，却可能无意间排除了与我们最初的研究问题或理论框架不一致的研究。此外，我们所分析研究中研究者与参与者之间的权力关系，可能影响了激励方案的报告方式，进而导致对相关实践的呈现有所偏颇。此外，我们对人机协作的切身关注，可能影响了对研究结果的解读。我们可能更倾向于强调设计良好的激励对参与者参与度和数据质量的积极影响，同时淡化潜在的负面后果，例如伦理问题或意外偏见。为减轻这些潜在偏差，我们严格遵循反思性主题分析方法，积极寻求识别不同观点，并检视自身假设。通过承认自身局限并积极努力加以解决，我们力求为人机决策研究中的激励设计提供平衡且有洞见的分析。

3 主题分析

对描述人机决策文献中奖励方案的摘录数据集进行严格的反思性主题分析后，我们确定了以下主题。我们还根据 [13] 反思性主题分析检查清单，结合文献中的具体示例对这些主题进行了说明。

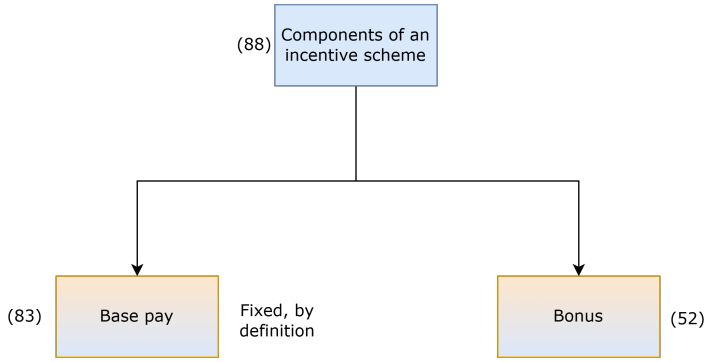


图2. 主题1：激励方案组成部分的顶层主题图

3.1 主题1：激励方案的组成部分

激励方案的组成部分出现在 88/97 篇论文中，是最常见的主题。我们进一步在其中识别出两个子主题，即基本报酬 和奖金。如图 2 所示。

88篇论文中有47篇同时提到了基本报酬和奖金。例如，[71] 提到，“每位参与者完成研究后可获得*3*美元（*1*美元基本报酬和*2*美元奖金，以激励其绩效）.....”同样，[27] 写道，“参与者完成研究后可获得*1*美元，并且准确预测的绩效最高还可额外获得*1*美元。”

88篇论文中有41篇只提到了基本报酬，没有提到奖金。例如，[91] 写道，“每位参与者均获得*2.50*美元的固定报酬。”与此同时，88篇论文中有5篇只提到了奖金，没有提到基本报酬，例如 [11]：“.....我们向赚取最多虚拟货币的两位标注员提供了（相对）丰厚的奖金。”

3.1.1 基本报酬。83/97 篇论文提到了基本报酬。图 3 展示了这一主题及其子主题。

基本报酬被定义为参与者完成分配给他们的任务后获得的核心或基本支付。为了描述基本报酬，研究者要么直接说明一个报酬金额，要么说明工资率，即每小时的具体金额，或者两者都说明。

例如，[16] 写道：“每名工人获得了 *2* 美元的报酬。”与此同时，[118] 提到：“参与者按照每小时 *12* 美元的工资率获得奖励.....”。[22] 则同时描述了两者：“每完成一次获得 *3.75* 英镑（=每小时 *7.09* 英镑）”

报酬率和金额最常通过启发式方式确定，且没有解释金额是如何决定的。在少数情况下，工资率明确表示受到最低工资或平台建议的影响。例如，[73] 提到：“为了向参与者提供公平报酬，*Mturkers* 获得了相当于美国联邦最低工资的报酬.....”而 [48] 提到：“所有参与者都获得了 *[.....]* 被平台认为是‘不错的’报酬.....”。

研究者通常只会提供一笔固定金额。少数情况下，研究者会在说明报酬金额或费率的同时，提及完成任务所需的预计时间。时间估算大多通过试点研究进行，也有少数情况是凭启发式估算。例如，[55] 提到："根据我们研究小组开展的试点研究，我们确定完成任务的平均时间约为一小时。因此，我们将该任务的报酬定为 *17* 美元。"

3.1.2 奖金。97 篇论文中有 52 篇提到了奖金。每项任务或整个研究的最高可能奖金金额是通过启发式方法确定的。例如，[67] 提到："每位参与者获得 *2.50* 美元报酬，另有 *0.05* 美元

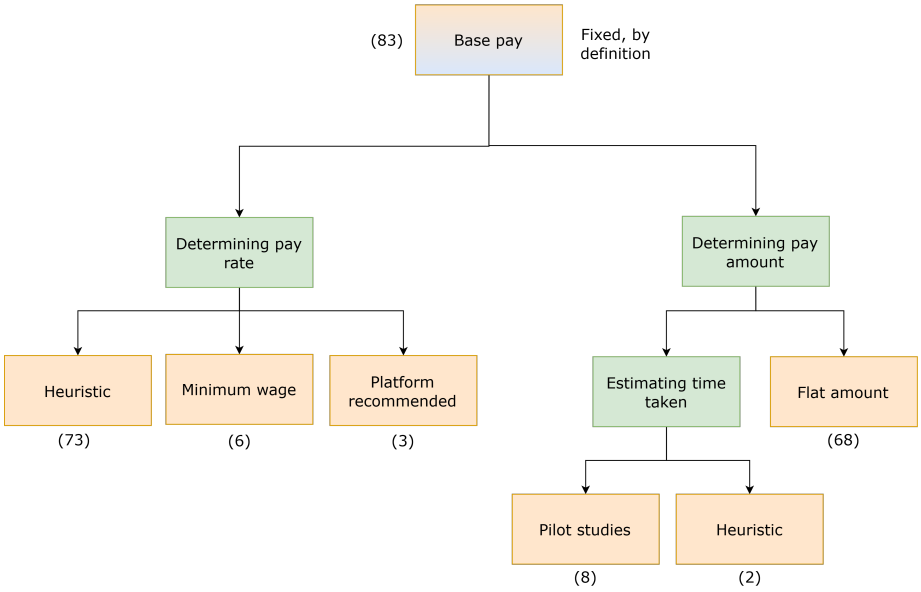


图3。主题1.1的主题图：基本报酬及其子主题

每份正确标注的测试评审可获得的奖金。”与此同时 [52]提到“所有参与者均获得了 *7.50* 美元的报酬，其中包括 *5.50* 美元的基础金额和 *2.00* 美元的奖金支付*.....*”。

支付方案说明了如何计算最终支付给参与者的奖金金额。根据用于确定支付的标准，这些方案可分为不同的类型，分别为基于表现、基于完成情况和基于运气。

图4展示了奖金的高层主题及其子主题。

基于表现的方案最为常用。参与者根据其表现获得报酬，而表现通过特定的绩效指标进行评估。通常还会说明将表现映射为报酬的政策。例如， [32] 指出：“参与者完成任务后获得 *1.00* 美元，如果他们在该任务中的总体准确率高于 *65%*，则可获得 *5.00* 美元的奖金*.....*”。

对于基于完成情况的奖金，参与者完成特定任务（例如回答可选调查）后即可获得固定金额。例如， [7] 提到：“参与者获得了 *[.....]* 完成调查后可获得 *0.25* 美元的固定奖金*.....*”。

对于基于运气 的奖金，通常会随机 决定参与者是否获得奖金。随机化可以通过不同方式引入设计中，例如设定参与者获得奖金的特定概率，或为每位参与者随机挑选一项任务，并根据某些标准进行评估。例如， [74] 提到，“*.....*我们随机选取了序列中的一项预测任务，检查受试者在该任务中的最终预测是否正确。如果正确，受试者除了基础报酬外，还会获得 *1* 美元奖金。”

此外，基于表现的奖金中还出现了一些特有的子主题。这些内容见图5，并将在下文讨论：

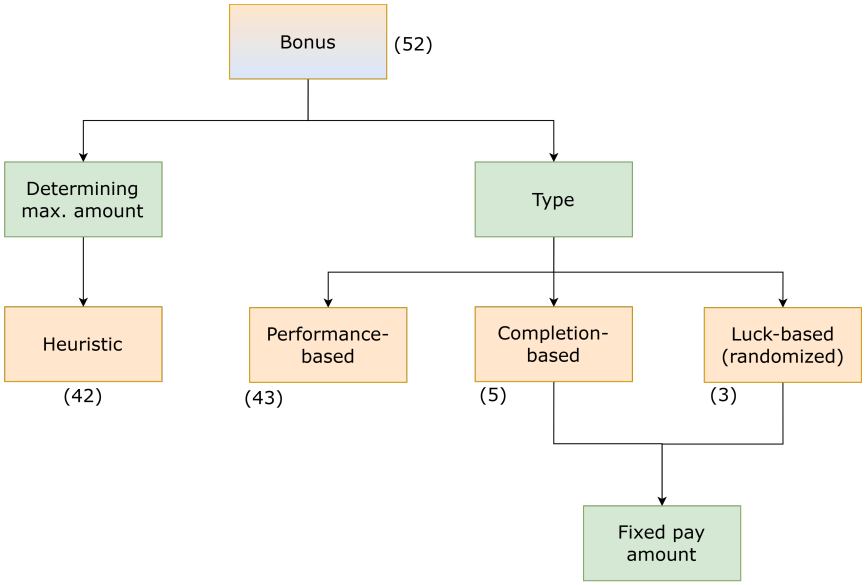


图 4. 主题1.2：奖金及其顶层子主题的主题图

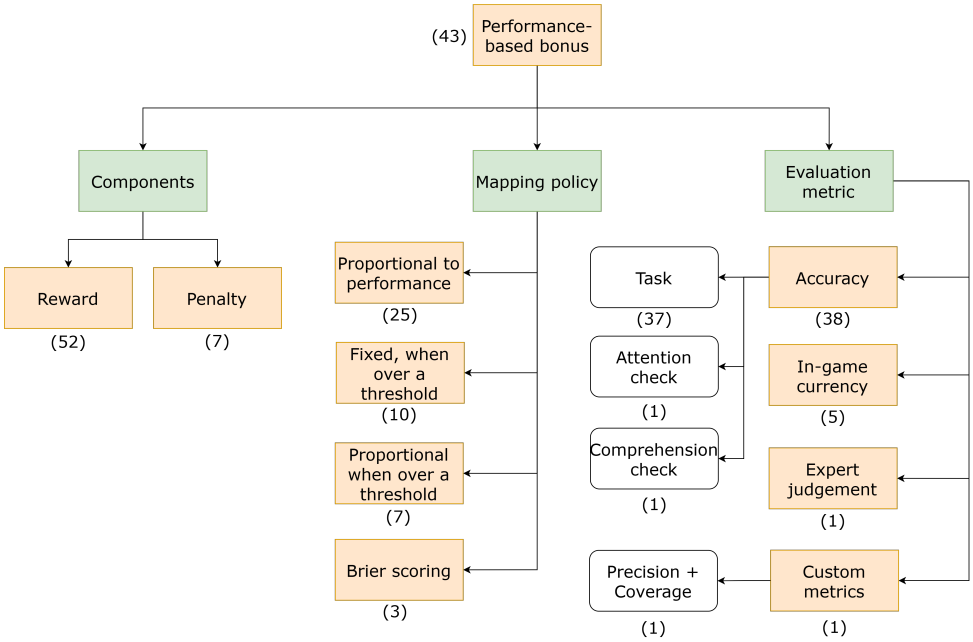


图 5. 基于表现的奖金 及其子主题的主题图

首先，绩效评估指标用于评估参与者绩效，其中最常见的是准确率。例如， [4] 提到，“.....参与者每答对一题，便可获得额外的 *0.5* 英镑基于表现的奖金.....”

其他指标的例子包括游戏内货币（游戏类任务的游戏化得分） [41], 自定义指标，例如“精确率 + 覆盖率” [65], 以及专家评估（专家评估参与者绩效） [42]。

其次，将绩效映射为报酬有多种方式。这些方式包括：绩效超过阈值时支付固定报酬的政策 [32], 与绩效成正比的报酬 [67], 绩效超过阈值时按绩效比例支付报酬 [21], 以及通过布里尔评分 [14] 函数计算报酬 [45]。例如， [28] 将其政策描述为：“如果参与者的官方预测与学生的实际百分位数相差不超过五个百分点，就会获得 *0.50* 美元的奖金。误差每多五个百分点，奖金就减少 *0.10* 美元 *[.....]*。因此，预测偏差超过 *25* 个百分点的参与者不会获得奖金。”

此外，基于表现的奖金也可以同时包含惩罚 （负向激励）和奖励 （正向激励）。例如， [119] 提到，“.....如果最终预测正确，则奖励 *5* 美分；否则损失 *2* 美分.....”。此类政策的设计规范通常基于启发式，只有少数案例以既有文献为依据 [45, 119]。

在相当大比例的研究（47/88）中，基本报酬和奖金并用，这表明混合方法是激励参与者的常见做法。这说明研究者往往倾向于采用以基准薪酬为基础的薪酬结构，同时也试图通过基于表现的奖励来激励参与者。此外，在确定基本时薪和金额时，高度依赖启发式方法（83/97），表明该领域缺乏标准化实践。这说明有必要采用更加系统且循证的方法来设定薪酬水平。

3.2 主题二：激励操纵

下一个发现的主题是，激励会因各种目的而受到操控。图6展示了这一主题及其子主题的图表。

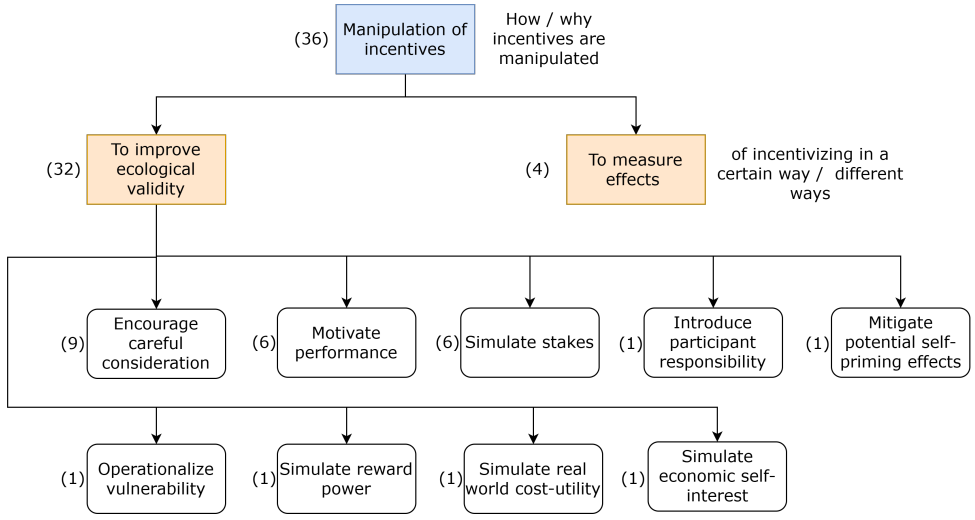


图 6. 主题 2：激励操纵主题图

措施

97篇论文中有36篇提到，为了某些目的而使用或操控激励。大多数情况下，这些目的被确定为提高实验的生态效度。例如，[119] 说道：“我们采取了两项措施来提高生态效度。首先，将决策绩效与金钱奖励挂钩……”。

研究者提到使用激励（主要是奖金）来：“激发绩效”[17], “鼓励参与者专心”[32], “模拟风险”[6], “模拟奖励的作用力”[52], 等等。我们将这些视为提升生态效度这一更广泛主题的例子（而非子主题）。

最后，奖金方案也会经过有意调整（例如设置高额或低额奖金），以衡量不同方案对参与者绩效或研究成果的影响。例如，[116] 提到：“我们还提出并预先注册了另外两个假设：*[H3]* 风险金额会显著影响人们在看到反馈界面之前对模型的信任。*[H4]* 风险金额会显著影响人们在看到反馈界面之后对模型的信任 *[……]* 以检验当人们投入更多‘真金白银’时，声明准确率对信任的影响是否会有所不同。”

这一主题表明，研究者意识到实验室环境可能存在局限，并试图营造更真实的实验条件。然而，这种做法也引发了一个问题：研究结果能否推广到现实场景。

3.3 主题 3：激励的影响

第三个主题是激励方案对研究结果的影响。图 7 展示了该主题及其子主题的主题图。

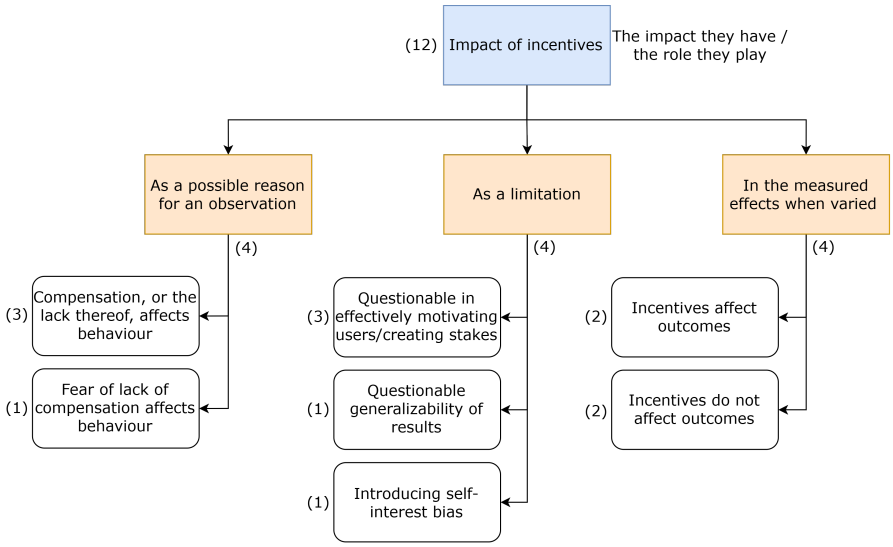


图 7. 主题 3：激励的主题图

97 篇论文中有 12 篇提到了激励的影响，并反思了激励在其研究中可能发挥的作用。一些研究者讨论了激励方案的局限性，质疑其是否能够有效再现真实生活场景或激励众包工作者，并承认研究结果的可推广性尚不确定。例如，[28] 指出，“由于现实生活中的最终用户在人口特征方面可能存在很大差异

以及非货币激励，并在高风险环境中运作，我们无法可靠地将这一发现推广到真实工作场所……"。

在少数情况下，激励也被视为某一观察结果背后的潜在原因，例如，观察到某种特定趋势，可能是因为参与者受到报酬的激励而努力完成任务，也可能是因为缺乏报酬而未受激励。例如，[41] 指出，“分析有两个重要的注意事项。首先，分析依赖于实验接近尾声时收集的非激励性自我报告数据。因此，我们无法确认受试者是否认真思考了自己的回答……”。

最后，少数有意操纵激励以衡量其影响的论文，也会描述研究结果，进而讨论激励的影响。对于这些研究，不同研究者发现，在不同条件下，激励可能会影响、也可能不会影响绩效或结果。例如，[109] 描述道，“……我们发现奖励会影响过度依赖……”。与此同时，[116] 发现，“……至少在我们的实验所采用的有限赌注范围内，赌注金额不会影响普通公众对模型的信任。”我们注意到，这些研究针对不同任务和领域，在不同条件下开展，并衡量奖励对不同变量的影响。

明确讨论激励影响的研究数量相对较少（12/97），这表明研究者往往忽略对不同激励方案因果效应的调查。这进一步说明，该领域尚未全面理解激励如何影响参与者行为和研究成果。

3.4 主题 4：激励的沟通

该主题突出了向参与者进行激励方案沟通的不同趋势。图 8 展示了该主题及其子主题的主题图。

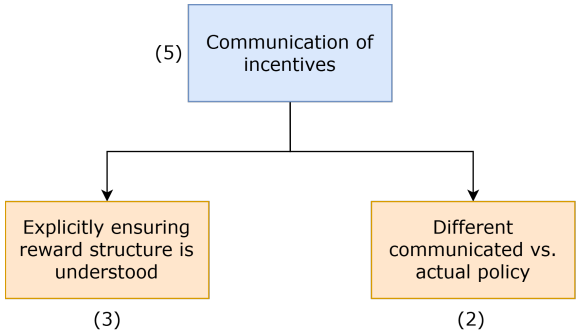


图 8. 主题 4：向参与者进行激励沟通的主题图

97 篇论文中有 5 篇讨论了其激励方案的沟通。在一些情况下，研究者会明确确保参与者在开始任务前理解薪酬结构。例如，[45] 描述道：“……参与者受到激励，要报告他们对风险的真实估计。我们在教程期间向参与者阐明了这一点，并在理解测试中加入了一个关于奖励结构的问题，以确保他们理解。”

此外，还有少数案例中，向参与者传达的奖金方案与实际用于计算参与者最终付款的方法有所不同。例如，[25] 提到，“……研究者告诉参与者，至少排名前 *50%* 的参与者将根据其评估的详尽程度获得 *2* 美元奖金；但他们并不知道，最终所有人都获得了奖金。”

对激励沟通的讨论有限（5/97），表明研究流程或其报告可能存在不足。研究者可能没有充分考虑向参与者透明地传达激励方案的重要性；也有可能，问题出在激励沟通的报告，而非其实施。此外，这也凸显出，沟通策略的设计本身可以成为整体激励设计流程中不可或缺的一部分。

3.5 主题5：未提及激励

97篇论文中有8篇未提及任何与激励方案相关的内容 [9, 33]。一些研究（8/97）缺乏对激励的明确讨论，这引发了人们对隐性或未被承认的激励可能对参与者行为和研究成果产生影响的担忧，也表明实验设计和报告实践缺乏透明度。

4 讨论

在此，我们讨论主题分析的主要发现，并反思这些观察结果所带来的影响，提出可指导进一步研究的问题。

4.1 主要发现

4.1.1 支付金额。 主题分析显示，目前确定基本报酬和奖金金额的策略主要依赖启发式方法。此外，在基本报酬和奖金的计算与报告方面也缺乏一致性。

在基本报酬方面，研究者采用了各不相同的方法：有些设定固定金额，有些则采用工资率。此外，各项研究对基本报酬计算方式的解释也不尽相同。有些研究明确说明了计算过程的部分环节，例如估算任务完成时间以确定金额；另一些研究则缺乏此类透明度。对于奖金，研究中很少讨论如何确定分配金额。这些观察结果凸显了采用更标准化的方法来确定并报告基本报酬和奖金金额的必要性。是否存在最优或适当的报酬金额？能否以标准化的方式确定报酬金额？未来工作能否探索数据驱动方法，以确定最优报酬水平？我们能否制定标准化的方法论，以便在不同任务和平台上计算基本报酬和奖金金额？

先前研究中曾将类似问题作为“预算约束任务调度”问题进行研究。研究者提出了用于任务分配和推断方案的算法，旨在固定预算 [50, 59, 106] 内尽可能提高结果质量。然而，这些算法尚未在研究者中得到广泛应用。这类方法可作为基础，帮助研究者设计自己的方案，或为未来研究开发更复杂、能够满足人机决策任务特定需求的算法。

4.1.2 奖金方案的各个方面。对奖金方案的考察进一步揭示了奖金各个方面存在许多不同做法，例如奖金类型多样（基于表现、基于完成情况、基于运气）、奖金计算政策（固定／按比例）以及绩效评估指标。值得注意的是，并非所有研究都全面探讨了我们所确定的各个方面。未来工作应致力于使奖金设计过程更加全面。还需进一步研究在不同任务条件下，哪些做法最为合适。我们能否制定一种标准化方法论，全面推荐具体的

基于研究条件的奖金计算方法？一种奖金方案是否比另一种更恰当或更合适？哪些指标最适合用于评估绩效？将绩效映射为报酬的最佳方式是什么？这些政策能否得到优化？在什么条件下？

通过主题分析识别出的不同政策，可作为研究者进一步完善并试验各种激励方案的起点。通过开展试点研究和受控实验，研究者可以探究不同政策的影响。研究者还可以设计更复杂的算法（类似上一节提到的算法），以优化激励分配并最大限度地提高参与者参与度。

4.1.3 奖励与惩罚的使用。分析显示，研究中除了基本报酬外还会提供奖金，有些研究甚至采用惩罚来模拟高风险情境。要确定这些方法的有效性，还需要进一步研究。使用奖励与惩罚是否有价值？这会营造出“更高”的风险吗？加入奖金和惩罚是否真的会营造出“更高的风险”，还是会引发伦理问题？是否可以在正向激励与负向激励之间取得平衡，在不损害伦理考量的前提下优化参与者动机？

理论框架可用于论证此类策略的合理性。例如，前景理论 [56] 指出，人们对损失比对收益更敏感。将损失厌恶纳入激励方案的设计，研究者或许能够提高动机和参与度。不过，仍需进一步研究来确定这些方法的有效性。

4.1.4 提高生态效度。 分析强调了人们对激励能否营造“利害关系”并提高研究生态效度的担忧。未来研究应探索评估激励结构对参与者行为的影响，以及这些结构在实践中提高生态效度的能力的方法。能否评估和/或确保激励对参与者绩效和行为产生预期效果？如何做到？ 能否制定一些实践方法，客观评估激励是否确实提高了实验的生态效度？

为评估激励对参与者行为和生态效度的影响，研究者可以采用多种策略，包括研究前后比较、受控实验和纵向分析，以衡量参与度和绩效的变化。客观评估可以通过行为验证研究、与现实场景进行比较，以及兼顾任务策略和结果的多维评分系统来实现。

4.1.5 激励的影响。分析凸显了众包研究中激励方案与研究成果之间复杂的关系。哪些类型的激励方案会影响结果？哪些不会？在什么条件下？

我们发现，有些研究表明激励方案会影响结果，而另一些研究则没有发现这种影响。这说明，激励设计并不存在一个通用的“公式”。激励方案的有效性可能取决于多个因素，例如实验条件和研究设计。由此，我们需要围绕这一重点展开更深入的探索。

4.1.6 激励的局限性。分析表明，研究者担忧在众包研究中使用激励的局限性，尤其是研究结果的可推广性，以及众包工作者动机可能造成的偏差。更完善的激励设计实践能否缓解与其使用相关的局限性提交给 ACM 的稿件

？如何设计激励，才能更好地了解其影响结果的方式？能否制定实践，有意设计能够克服局限性的激励？

更完善的激励设计实践有望通过使激励与预期成果相一致，并减少意外影响，帮助缓解局限性。有意识的设计可以包括使激励与参与者动机相一致、平衡内在奖励与外在奖励，并确保公平性。为了更好地理解其影响，需要开展严谨探索，系统地改变激励措施，并利用行为指标和定性反馈评估其效果。设计激励措施以促进有意义的参与，而非单纯追求自身利益，有助于提高结果的可靠性和生态效度，同时应对固有局限性。

4.1.7 沟通策略。明确确保参与者理解激励方案及绩效反馈，可以增进参与者的理解，并有望改善结果，但也可能引入偏差或打击参与者的积极性。针对激励措施采用不同类型的沟通策略是否有价值？按所传达的绩效水平支付报酬更合适，还是向所有人支付相同报酬？

透明地沟通激励措施与绩效之间的关联，有助于增进理解并提升动机，但必须谨慎设计，以避免引入偏差或令人气馁。按照所传达的绩效水平支付报酬，符合预期并强化问责，但可能使绩效较低的参与者处于不利地位。另一种做法是支付相同报酬，这有助于促进公平性和包容性，但可能降低高强度努力的动机。要平衡这两种做法，需要仔细权衡，并在具体背景下开展实验，以找出既能优化参与度、又能维护公平性和信任的策略。

4.1.8 缺失的激励方案。 最后，我们注意到，有些研究者选择不讨论其研究中的激励方案。是否应鼓励研究者始终报告众包研究中采用的激励方案，无论他们认为这些方案是否相关？如果应该，如何做到？我们注意到，没有相关描述并不一定意味着研究者没有考虑或采用激励方案。研究者可能认为，提及这些方案与其具体目标无关。尽管如此，透明度有助于更好地复现研究、提升可比性并理解结果。研究者可以在研究设计中说明激励的类型、沟通策略，以及这些激励与研究目标的一致性。标准化的报告框架或指南有助于一致地记录相关信息，确保将激励视为研究设计和分析中不可或缺的一部分。

4.2 反思

上述问题主要分为两类：方法论挑战 和探索方向。方法论挑战涉及激励方案的实际实施，例如报酬金额和结构的设计 。探索方向则涵盖更广泛的研究方向，包括研究操纵的影响 以及不同沟通策略的效果。我们注意到，方法论挑战与探索方向密切交织。例如，绩效指标的选择和奖金方案的设计会影响我们研究激励设计对参与者动机和任务表现的影响。因此，深入理解这些因素如何相互作用，对于设计有效的激励方案至关重要。

尽管全面探讨所有这些问题超出了本研究的范围，但我们优先解决了围绕激励设计的方法论挑战，因为这些问题与我们确定的研究议程相符。主题分析的结果进一步证实，鉴于人机决策任务固有的挑战，标准化十分必要，也因此令人质疑“适当”激励设计这一概念本身。因此，为回答研究问题2和研究问题3，我们着手建立一种标准化方案，以便

研究者能够为各自的人机决策研究调整“适当”的激励方案。为回答研究问题2和研究问题3，我们发现，设计完善的框架能够促进一种标准化、系统且全面的方法，引导研究者设计并记录有效的激励方案。此外，该框架还可防止人们无意间忽视激励设计的重要方面。未来还应进一步深入探讨其他方面，例如激励对研究成果的影响，以及高风险设计的伦理影响，并将这些探讨与我们的框架结合起来。

5 激励调优框架

(1) **明确采用激励方案的目的**：我们希望通过激励参与者实现什么目标？我们期望参与者表现出哪些行为？(2) **确定基本报酬**：为确保实现已确定的目标，同时保证报酬公平，参与者必须获得多少基本报酬？任务的哪些特征会影响这一金额？(3) **设计奖金结构**：提供奖金是否有助于实现已确定的目标？(a) **确定奖金金额**：应为奖金分配多少总金额？(b) **确定奖金类型**：哪种奖金最适合实现已确定的目标？对于基于表现的奖金：(c) **确定绩效评估指标**：如何评估参与者绩效？我们希望奖励什么行为？(d) **确定将奖励与绩效对应的政策**：如何将绩效对应到奖励？(4) **收集参与者反馈**：我们希望从参与者的角度了解什么？能否在实验流程中加入收集此类反馈的环节？(5) **反思设计启示**：该设计是否实现了预期目标？是否产生了任何意料之外的后果？

5.1 步骤 1：明确目的

明确采用激励方案的目的，有助于有意识地进行设计，确保其与研究目标相一致。可能的目标包括满足伦理标准、提高研究的生态效度[5] 、创造或模拟利害关系、激励参与者提升绩效，或鼓励参与者在完成任务或达到绩效要求之外展现期望行为。在此步骤中，研究者必须认识到，确保提供公平报酬是其伦理义务，并将参与者福祉纳入研究目标。

5.2 步骤 2：确定基本报酬

已确定的目标以及任务的特征（如任务复杂度和所需专业技能）都会影响基本报酬的金额。例如，认知负荷较高的复杂任务可能会因为参与者认为投入时间所得回报较低 [99]而降低其参与意愿。因此，为吸引参与者并激励他们取得良好绩效，这类任务可能需要提供较高报酬。同样，与受过少量培训的普通人就能完成的任务相比，需要专业技能或深入知识的任务通常应提供更高的基本报酬 [35]。

最低工资和平台建议报酬标准等其他因素，可以为报酬的伦理标准提供指引。类似过往研究的定价也可能相关，因为先前研究表明，锚定效应[78, 115] 会影响参与者对公平报酬的感知。

最后，开展试点研究正日益成为人类-AI研究中受鼓励的做法，因为试点研究能让研究者有机会发现任务设计中可能存在的问题，并有助于校准任务参数，以便最终在更大规模上开展实验 [88, 108]。以往的研究曾通过试点研究估算完成任务所需的时间。这种做法还可以进一步拓展，例如评估拟提供的基本报酬是否充足，并帮助调整报酬，使其更符合参与者对公平报酬的期望。先前的研究提出了快速原型设计和数据驱动的努力度量指标，用于衡量参与者完成任务所需的努力 [20]。作为试点研究的替代或补充，这可以成为估算基本报酬的合适方法。

5.3 步骤3：设计奖金结构

公平的基本报酬至关重要，且先前的研究已对此进行过明确探讨 [54, 113],，但它可能无法充分体现任务复杂度的细微差异或期望达到的绩效水平。奖金能够激励众包工作者，鼓励他们表现出特定行为，例如追求绩效，而不仅仅是完成任务。尤其是在高风险 场景中，奖金可以让参与者感到利害关系更加重大，而他们往往主要受金钱目标驱动 [72]。与此同时，我们也认识到，奖金需要经过审慎考量。仅仅提供高额奖金之类的策略，并不一定能直接带来更高质量的工作 [78, 111]。因此，重要的是在报酬金额及其分配方式之间取得平衡，同时确保其与具体研究目标一致。因此，我们建议研究者在设计奖金结构时，批判式地评估他们希望激励的行为。下面我们将讨论奖金方案的不同方面。

确定奖金金额。与基本报酬一样，类似任务过去的定价可以作为最高可获得奖金金额的起点，并可通过试点研究进行调整，例如加入有关参与者满意度以及他们对额外奖励公平性的看法的问题。

此外，我们通过主题分析确定了三种可能的奖金方案：基于表现、基于完成情况或随机发放。随机奖金可能会产生一种权衡：既可能打击高绩效参与者，也可能鼓励低绩效参与者更深入地参与，因此应谨慎使用。基于完成情况的奖金有助于鼓励参与者完成可选任务（例如问卷调查），此类任务的主要目标是收集足量数据。最后，基于表现的奖金是最常用于激励期望行为的方案，其有效性取决于能否选择恰当的绩效评估指标，以及将绩效映射为奖励的政策。绩效可以指准确率或速度，但所选指标应真正体现任务目标，而不是促使参与者以不合乎期望且带有偏差的方式行事。此外，绩效也不只是一个指标，还可以指参与任何形式的期望行为。例如，如果解释决策背后的推理很重要，就可以为深入解释提供奖金。如果仔细阅读所提供的信息很重要，就可以通过激励来鼓励参与者认真完成注意力或理解检查。决策主观性等任务特征也会影响这些因素。例如，如果决策主观性较高，可以使用开放式问题来评估绩效。我们还发现，过去有些研究会采用惩罚来营造高风险情境。这或许能利用参与者厌恶损失的倾向，从而发挥作用 [1]。然而与此同时，过度惩罚可能会带来权衡，

抑制参与意愿。因此，在将其纳入设计时必须谨慎。此外，将绩效映射为报酬的方式有多种，例如：绩效达到阈值后给予奖励，或随着绩效提升而逐步增加奖励。需要结合研究目标，确定哪种方式最适合该任务。深入研究激励设计，以找出可能适用的最优政策设计方法，或许会有所助益。最后，还可以再次利用试点研究，观察参与者对不同奖金方案的反应。

5.4 步骤 4：收集参与者反馈

任务结束后进行的调查或退出调查，有助于了解参与者体验和动机。调查中可以询问参与者对报酬公平性的感知以及参与者动机，还可以让研究者比较回答，并评估试点研究中有关类似问题的发现是否适用于正式实验。例如，如果结果显示准确率低于预期，关于报酬不公平的调查回答可能表明，参与者的动机不足，未能尽力完成任务。这些信息对于解读结果及其可推广性至关重要。

归根结底，这种做法可以帮助我们评估所选激励方案的有效性。研究者可以询问参与者是否认为报酬公平、奖金结构是否能激励他们，以此判断激励方案是否实现了目标，例如吸引合格参与者或鼓励期望行为。此外，这也能通过提供参与者的视角，助力下一步反思设计启示。这些做法有助于完善未来研究中的激励方案。例如，如果参与者表示报酬不足，或奖金结构未能激发他们的动机，就可以进行调整，更好地满足参与者对未来研究的期望。

5.5 步骤 5：反思设计启示

至此，我们建议研究者应开展有意识的设计。我们鼓励研究者理解其中的权衡、说明决策理由，并收集参与者的反馈，以便最终能够反思其设计所带来的影响。

为了让设计流程取得成果，框架的最后一步要求研究者也反思其设计对研究成果可能产生的影响。这样一来，研究者便能更有把握地看待自己的发现，并为该领域提供有价值的见解。此外，这也有助于营造透明度与问责文化，从而改进现有方法，推动未来研究。

通过概述上述步骤，该框架旨在促使研究者明确考虑并界定其研究目标，包括那些可能不太容易具体衡量的目标，并以系统的方式开展激励设计流程，突出其中的关键方面。这样一来，该框架有助于确保激励方案的设计直接服务于具体的研究目标，并能得到有效评估。归根结底，研究者如何将该框架付诸实践，可能因研究背景而异，但核心原则始终不变：激励设计应体现研究者希望实现的目标。

5.6 激励方案设计文档编制

尽管激励调优框架系统而全面地概述了研究者必须报告的项目，但在报告各项内容时采用的格式和数据呈现方式仍存在一些模糊之处。例如，主题分析揭示了基本报酬的报告方式不一致，包括使用“固定金额”还是“工资率”，以及是否提及最终总报酬。另一个发现存在歧义的方面是平均值的使用提交给 ACM 的稿件

与中位数来报告最终报酬。尽管两者都能有效概括数据，但在不同研究中采用一致的方法至关重要。

5.6.1 模板。基于激励调优框架各项内容所奠定的基础，标准化的报告模板有望应对上一节指出的挑战。

以下模板可用于在研究论文的实验设计部分记录有关激励方案的详情：

> 我们的激励设计旨在确保 [明确主要目标]。参与者获得的最终报酬为 [最终报酬的平均值和中位数]，报酬基于 [每小时基本报酬] 的基本时薪，以及 [奖金发放额的平均值和中位数]。
>
> 完成任务后，基本报酬设为 [金额]，以确保报酬公平，并考虑了 [理由，例如平台建议／最低工资／以往定价／试点研究反馈／特定任务特征]。[可根据需要进一步说明]。
>
> 为进一步 [激励／确保／促使] [已确定的目标]，实施了基于 [绩效]的奖金结构。 [根据任务的具体特点作出的决定（例如，因感知风险较高而采用惩罚]）。奖金发放额上限设为 [金额]。参与者绩效根据 [评估指标（例如准确率）], 进行评估，并按 [奖励映射策略]计算。 [可根据需要进一步详细说明]。
>
> [可选] 参与者因 [任务具体考量] 而获得额外奖金，以鼓励 []期望行为]。
>
> [可选]更多详情、调查反馈等。

注意：需要指出的是，模板中的建议措辞并非要求严格遵循。其主要作用是示范报告激励设计决策时推荐采用的结构。研究者可以根据研究的具体背景调整这些措辞，同时确保报告整体清晰、一致。

我们注意到，该模板仅涵盖框架中的前三项。鉴于其余各项的描述和讨论高度依赖具体背景，是否补充这些内容由研究者自行决定。为此，我们提出以下建议：

研究设计应明确描述在实验期间收集反馈所采用的方法（例如，问卷调查、开放式问题、访谈）。在结果或观察结果部分，研究者应呈现所收集到的反馈数据。

讨论部分应包括对设计选择和反馈的反思。这可以涵盖从反馈数据中获得的见解、反馈数据对整体研究发现的影响，以及未来研究中可能的改进方向。

5.6.2 一个存储库。为开展文献综述而构建的数据集，是一个记录现有文献中激励设计的重要语料库。我们发现，通过整理该数据集中的相关条目并将其发布到公共仓库中，可以为未来研究者创建一项宝贵资源。因此，我们在 GitHub[3] 上建立了一个公共仓库，以促进人机决策研究中激励设计的透明度与协作。

[3]www.github.com

可在此访问该存储库，以查看源代码或提出拉取请求：GitHub Repository。

截至目前从已发表文章中整理出的激励方案数据表可直接在此处查看：GitHub Pages。

该存储库旨在作为一个中心平台，供研究者分享和获取以往已发表研究中的激励设计知识。它向公众开放，目前正持续收录从既有工作中提取的设计决策。我们也邀请研究者今后将其设计的激励方案贡献到该存储库中。

5.7 案例研究：应用激励调节框架

我们 प्रस्तुत两个案例研究，以展示所提出框架的应用。为此，我们在获得通讯作者同意后，选取了两篇研究论文。

案例研究 I

在第一个案例研究中，我们考察了 [69]撰写的论文《你感觉如何？衡量拒绝仇恨言论检测中的机器决策时用户感知的价值》。论文提到了一种简单的激励方案：

“每 位参与者的时薪为 9 英镑，高于该研究开展时的英国最低工资 ."

仔细审视论文中描述的实验设计和研究目标后，我们通过应用框架，提出了针对该研究激励方案设计的发现。

- **步骤 1：明确目的**：我们发现，作者希望参与者认真处理信息并专注于危害评估。此外，我们推测作者的目标还包括提供公平报酬、提高生态效度，以及模拟现实利益攸关性。

- **步骤 2：确定基本报酬**：所选的每小时 9 英镑基本报酬高于最低工资，符合公平报酬的要求。在尝试了解毒性分类领域研究过去的定价情况时，我们没有找到多少探讨类似任务的相关文章。由于任务不要求特定技能，基本报酬似乎是合适的。然而，考虑到每位参与者需要完成大量任务（40 项），我们必须考虑如何帮助参与者在整个过程中保持动机，并补偿他们投入的时间。一种简单的策略是提高基本报酬本身。我们注意到，作者开展了试点研究。我们建议通过加入以下问题来扩充试点研究，以衡量参与度和满意度：“完成任务时，你是否觉得需要休息？如果需要，休息了几次、每次多久？”以及“你觉得报酬对于所需的时间和精力而言是否公平？”

- **步骤3：设计奖金结构**：我们注意到，作者没有提供奖金。提供奖金有助于提高生态效度、模拟利害关系，并鼓励参与者展现期望行为。我们建议，提供少量基于表现的奖金可能会有效。作者选择不对正确答案提供激励，并表示其主要目标不是鼓励准确率，而是评估用户对价值的感知。相反，他们在任务中加入了“冗长的描述”

通过奖励引导参与者将注意力集中在评估伤害上，并设置注意力检查以筛除注意力不集中的参与者。为更好地实现这一目标，我们建议加入侧重于处理任务信息的理解检查，并将其准确率作为基于表现的奖金的评估指标。对于通过这些检查的参与者，发放固定奖金，可以在“冗长的描述”这一情况下鼓励他们保持专注，同时符合预算限制。因此，此处的绩效指标将是理解检查的准确率，奖励映射策略则为准确率达到 100% 时发放 100% 的奖励。

此外，我们注意到这项研究涉及高风险场景（仇恨言论分类）。正如作者强调的，用户理解错误决策所带来后果十分重要，因此施加少量惩罚是一种可行选择。然而，为避免参与者可能因此丧失积极性，也应提供足够的正向奖励。但我们发现，提升任务准确率并非研究目标，因此决定不建议采用基于任务准确率的奖励。综上，我们决定不采用惩罚。

- **步骤 4：收集参与者反馈**：尤其是由于我们没有采取任何明确措施通过引入后果（例如惩罚）来营造高风险情境，因此我们建议使用任务后调查，更好地了解参与者是否理解决策情境的后果和风险。如果预算约束限制了建议的实施，这一点尤其重要。了解参与者观点有助于解读结果。

- **步骤 5：反思设计启示**：当前设计优先考虑公平性，但缺乏足够的策略来维持参与者在冗长任务列表中的动机。此外，缺少后果可能会导致人们低估评估伤害的重要性。除非重复开展研究，否则我们无法真正衡量该研究在这方面的结果；不过，参与者评估伤害时可能存在的不一致，或许可以归因于这些因素。

案例研究二

在第二个案例研究中，我们考察了 [97] 撰写的论文《应对不确定性：理解预后任务与诊断任务对人机决策中的信任和依赖的影响》。该研究对激励方案进行了更详细的描述：

> “所有参与者在研究期间均按每小时 8 英镑的固定标准获得报酬，不受其绩效影响。此外，参与者在研究期间每提供一个准确的回答，即可获得 0.2 英镑的奖金。总体而言，参与者平均每小时获得 8.44 英镑，远高于 Prolific 平台认定为‘良好’并推荐的工资水平。”

步骤 1：明确目的：我们将目的确定为确保参与者获得公平报酬、提升实验的生态效度、鼓励参与者在行程规划任务中准确决策，以及模拟现实世界决策。

步骤 2：确定基本报酬：所选的基本报酬符合平台建议，且没有足够的先前研究使用相同的任务领域或结构来提供更多见解。

我们注意到，作者通过调整呈现给参与者的特征和约束数量来操控任务复杂度。考虑到复杂度等级（低、中、高）以及任务所带来的认知负荷，我们认为可以根据任务复杂度调整基本报酬。经过校准的基本报酬可以设置为：低复杂度任务的报酬与作者所述相同，中等复杂度任务的报酬略高，而高复杂度任务的报酬则显著

更高。通过恰当地激励参与者，并以激励应对不同的任务复杂度等级，研究者可以在分析结果时考察参与者动机对绩效的影响。例如，作者观察到，中高复杂度任务的绩效有所下降。这也可能是由于认知努力增加，导致参与者动机降低。如果研究者在激励参与者时考虑到额外的认知努力，就可以说明他们已采取有针对性的措施来应对这一因素，从而更有把握地将绩效下降归因于其他因素（例如不确定性或复杂度本身），进而提高研究发现的生态效度。同时，我们也注意到实施此类支付方案存在权衡。根据参与者所接到的任务区别支付，可能会被认为不公平；也有人会认为，所有参与者的时间都应获得同等报酬，无论任务复杂度如何。

作者没有提及试点研究。可以通过开展包含针对性反馈问题的试点研究来验证不同的方法，并衡量参与者对不同薪酬结构的动机和感知公平性。试点研究还可以帮助评估不同复杂度等级所需的实际时间和努力，从而有助于客观地校准薪酬标准。此外，我们可以加入认知努力问卷，以及开放式问题，例如“你是否有动力投入真正的努力来完成任务？”和“这项奖励是否公平、令人满意，并符合你的预期？”根据回答，还可以进一步调整金额和结构。

步骤3：设计奖金结构：作者采用了基于绩效的奖金，并以准确率 作为评估指标，以鼓励准确地规划行程，从而实现研究目标，即鼓励做出正确的决策。此外，由于决策主观性 较低，准确率可能是合适的评估指标。另一种设计方案是，在人工智能出错时奖励准确率，以激励参与者付出真正的努力，并避免盲目依赖人工智能。这有助于让参与者重新聚焦于准确规划行程这一目标，而非最大化奖励。还可以考虑奖励“accuracy-wid”（最终决策正确，且最初不同意人工智能的判断）或恰当依赖。直接奖励恰当依赖可能会影响行为，但也可能让参与者对结果的解读更加自信。

归根结底，在选择表现评估指标时，我们需要在两种方案之间权衡。做出决定时，我们会考虑每种方案的影响：

(1) 鼓励恰当依赖：使用准确率指标，或在人工智能判断错误时，为正确答案提供更高的奖金：这一选择意味着，我们鼓励的正是希望衡量的内容，因为研究者的目标是衡量恰当依赖本身。(2) 在不影响依赖的情况下鼓励提高总体准确率：在此，我们需要考虑参与者可能会过度依赖人工智能。

不过，我们注意到，研究者希望在自然情境中衡量恰当依赖。因此，我们得出的结论是，不应通过奖励来鼓励这种行为。至于将奖励与绩效挂钩，如果预算允许，随着绩效提升而增加奖励，可以作为进一步激励参与者做出正确决策的策略。最后，我们发现感知风险相对较低，因此不考虑采用惩罚等负向奖励。

步骤 4：收集参与者反馈：我们建议在任务后调查中加入有关参与者动机、认知负荷和薪酬公平性感知的问题，因为这些因素对于理解参与者的观点和解读结果仍然至关重要。

步骤5：反思设计启示：当前的激励方案优先考虑公平性，并鼓励准确决策。然而，潜在的对人工智能的过度依赖可能会影响生态效度

，对此必须明确予以承认。我们已在步骤3中详细讨论选择各项不同评估指标时所涉及的影响。在呈现研究结果时，也应纳入此类讨论。

我们承认，我们提出的建议和最终决策未必是唯一恰当的设计选择。介绍这些案例研究的主要目的是强调，借助激励调优框架，我们可以围绕激励方案的各个方面展开更广泛的讨论。我们旨在展示研究者如何在框架的指导下，审慎权衡并为决策选择提供理由。

6 启示

6.1 对方法论和人机交互社区的启示

我们的工作对实证人机决策研究的设计与分析方法具有重要启示。我们提出的激励调优框架为研究者设计有效且有意识的激励方案提供了宝贵基础。该框架系统地促使研究者考量研究目标、激励设计的各个方面、参与者行为以及潜在权衡，从而帮助研究者设计能够促进期望行为的激励方案。该框架具有可扩展性，未来可以更新，以纳入该领域可能取得的进一步进展。框架为激励设计提供的标准化方法不仅有望提升数据质量，还可以促进不同研究项目之间的比较，推动合作并加速人机决策领域的科学进步。同时，该框架也承认，在为每项独特的研究设计激励措施时，“一刀切”的方法可能并非最佳选择。它为研究者提供指导，同时允许根据具体的研究要求和研究问题进行定制。这将有助于开展更加稳健的人机决策研究，为这一快速发展的领域取得进步铺平道路。

此外，我们为记录激励方案所提供的模板采用了标准化格式，具有实际意义。它可以确保研究者在应用该框架后，记录所有相关信息。这将使读者更容易理解设计激励时的意图，以及激励对参与者行为和研究结果可能产生的影响。我们建立的公共访问在线存储库还使研究者能够访问和分享有关激励方案的详细信息，包括设计决策背后的理由说明。存储库的协作特性带来了多项益处。它促进了知识共享，因为研究者可以轻松找到与其工作相关的现有激励方案，从彼此的激励设计选择中学习，从而加快该领域的发展。它还通过允许社区审查不同研究的激励设计决策，促进了透明度。

6.2 对理论的影响

我们开展的主题分析为研究者采用的各种激励设计方法，以及围绕激励的作用和影响展开的多样化讨论，提供了有价值的见解。我们的探索还揭示了一些有待进一步研究的未解问题，例如激励操纵和沟通。

我们提供的概述可以为研究者深入探究激励设计的具体方面奠定坚实基础。我们的见解可以成为研究者审视某个主题的有力视角，并借此启动或塑造自己的研究，进一步推动该领域的发展。通过深入探究仍未解答的问题，研究者可以进一步完善这一方法，设计出更有效的激励方案，并为我们当前对激励设计的理解所依据的理论基础作出贡献。未来研究可以尝试

采用不同的激励结构，直接比较其对人机决策任务中的参与者行为和数据质量的影响。这还可以扩展到纵向研究 [101] ，追踪参与者行为随时间的变化，以了解不同激励方案的长期影响；也可以开展定性研究，深入了解参与者对激励及其对动机和参与度影响的感知。此外，除了专门聚焦于激励设计的研究之外，研究者还可以将关于激励方案影响的假设纳入探讨人机决策中其他因素的研究，例如信任、依赖或公平性。

推进这些方向，有助于更全面地理解激励、参与者行为与人机决策研究成果之间复杂的相互作用。这些知识可用于设计更有效且合乎伦理的激励方案，从而优化参与者参与度、数据质量和整体研究成果。

6.3 注意事项、局限性及其他考量

6.3.1 非货币激励。 尽管我们的工作仅限于考察众包人机决策研究中的金钱激励，但需要注意的是，替代性激励设计的范围更加广泛，例如基于声誉的激励 [112, 120],基于游戏化的激励（例如使用徽章、排行榜和竞赛） [82, 96], 社交奖励 [26, 34, 37], 绩效反馈 [30], 或学习激励 [3, 77]。人类计算和众包领域的大量文献已经探讨了此类激励在塑造众包工作方面所发挥的作用 [38, 76, 92]。既有研究还证明，以个性化反馈作为参与行为研究“真实环境中”研究的参与回报，能够有效激励参与者 [53, 95]。未来工作可以探讨如何有效利用替代性的非货币激励设计，以支持对人机决策研究进行生态效度良好的组织实施。这些替代方案究竟如何与金钱激励进行比较，是一个值得探索的开放性问题。

6.3.2 框架的不完备性与迭代改进。 人机决策研究格局以及众包中的激励设计都在不断演变，新的挑战或意想不到的情境也可能出现。新的任务类型和研究问题可能会涌现，需要考虑主题分析和框架尚未涵盖的因素。此外，研究设计、参与者行为或平台功能中出现的意外复杂情况，也可能要求我们调整建议，或采取超出建议范围的做法。鉴于这一固有局限，我们建议将框架视为指南，而非僵化的公式。我们应做好准备，根据可能出现的情境调整和完善这些建议。此外，我们认为框架本身应是一份不断更新的文件。随着研究者积累使用经验并遇到新的挑战，框架应通过迭代不断改进，以纳入最佳实践并应对新出现的问题。

6.3.3 采用障碍与个体差异。 研究者可能会认为该框架复杂，或采用和实施起来耗时较长，尤其是对于资源有限的小型研究而言。开展回顾性分析或受控实验，探究使用该框架对研究成果的影响，从而阐明使用该框架的实际效用和价值，是未来工作中一个有意义的方向。这可能有助于开展以人为中心的人工智能研究的研究者和实践者认识到该框架的益处，并帮助他们在自己的工作中采用该框架。

激励调优框架为设计激励方案提供了一种结构化、系统化的方法，同时也能灵活适用于不同类型的研究。不过，必须认识到研究参与者之间的个体差异可能带来的影响。并不存在一种普遍适用于所有参与者、能以相同方式激励每个人的“正确”激励方案。尽管该框架侧重于系统化的逐步方法，研究者也应了解能够适应个体差异的替代方法，例如动态定价或自适应激励 [51, 80]。这些方法会根据参与者技能水平、绩效历史，甚至实时任务复杂度等因素调整激励结构。它们利用算法，在实验过程中根据参与者行为和参与度持续调整激励结构，从而实现更个性化的激励方式。尽管这些方法前景可期，但由于设计复杂度提高，以及需要具备先进的数据分析能力等因素，目前应用仍然有限。

未来，可以对该框架进行改进，将个体差异纳入考量。还可以进一步扩展该框架，加入相关步骤，指导研究者在其具体研究设计的约束范围内探索动态定价，甚至开发基础的自适应激励结构。

6.3.4 潜在偏差与假设。必须承认，激励调优框架所概述的设计流程可能受到认知偏差的影响。其中一种偏差是自利偏差 [31]。这种偏差表明，参与者自然会优先考虑自己的金钱目标，并可能倾向于采取某种行为或表现，以最大化自身奖励。我们通过强调设计基于表现的奖金时需要谨慎，来尝试应对这一偏差。毫无限制的奖金可能会无意中加剧自利偏差。不过，我们并不否定奖金或基于表现给予奖励。相反，我们承认可以利用自利偏差来激励参与者。通过谨慎地使奖励结构与期望行为相契合，研究者可以利用自利偏差，鼓励参与者付出努力、专注于期望指标，并贡献高质量成果。然而，奖励始终可能导致参与者优先考虑最大化报酬，而非提供高质量数据。我们鼓励研究者纳入质量控制措施，以降低这一风险并确保研究结果的效度。

最后，本研究的一个核心假设是，激励能够影响众包工作者行为。尽管这一观点得到了广泛支持 [15, 43, 60, 104], ，以往研究也指出，在某些情况下，激励可能不会产生显著影响 [107]。我们没有深入探讨激励影响较小或较大的具体条件。因此，无法明确确定所设计的激励方案究竟会在多大程度上影响行为。有趣的是，在开展主题分析时，这一难题本身也成为了一个子主题，但对这种动态关系进行更深入的探讨超出了本研究的范围。未来研究可以探讨激励设计的细微之处，以及在不同条件下激励可能对研究参与者产生的影响。我们强调，激励设计仍然是任何众包研究的重要组成部分。本研究开发的、用于辅助这一过程的工具仍然具有价值，至少在践行良好研究实践方面如此。通过遵循严谨的流程，使研究目标与激励相一致，并有意识地设计和记录激励，研究者便能尽到应有的审慎义务。

6.3.5 伦理考量。激励调优框架优先设计既有效又合乎伦理的激励方案。然而，必须承认，我们的一些建议涉及研究者应谨慎处理的伦理考量。

其中一个伦理问题是剥削。我们在建议中强调了合乎伦理的工资的重要性。我们主张，研究者负有一项基本的伦理义务，即确保为研究中的参与者提供公平报酬。提交至 ACM 的稿件

此外，我们不建议使用过高的奖金或惩罚性惩罚。这类做法可能会迫使参与者优先考虑速度或数量，而忽视真实性和幸福感。研究者应努力设计激励方案，在尊重参与者自主权的同时提供公平报酬。

其次，重要的是要认识到研究者与参与者之间存在不平衡的权力动态，以及我们如何通过自身的贡献来缩小这一差距。未来，至关重要的是将参与者意见和反馈纳入框架和存储库本身的设计与评估中。这样一来，我们就能赋予参与者更大的话语权，让他们能够讲述自己的动机和薪酬结构。

最后，我们聚焦于金钱激励，这本身就将本研究局限于外在动机的范畴。大量研究肯定了内在动机在推动众包工作者参与方面的重要性 [18, 70, 86]。未来研究可以探讨内在动机与外在动机对众包工作者行为的相互作用，从而有望更全面地揭示驱动他们的因素。随着人机协作与决策研究领域不断发展，围绕激励设计的伦理考量也将随之演变。通过审慎考量这些伦理影响，研究者可以利用该框架设计激励方案，不仅有效激励参与者，还能维护伦理原则并促进负责任的研究实践。

7 结论

总而言之，本研究围绕激励设计在释放人机协作于决策中的潜力方面所发挥的关键作用，构建了一个引人入胜的叙事。我们首先梳理了现有研究，通过对现有文献进行细致的主题分析，深入考察其中的挑战与机遇。这些认识为制定激励调优框架奠定了基础，该框架旨在帮助研究者为其研究设计恰当的激励方案。最后，我们基于该框架及其理论基础，提供了实用的报告与文档工具，以确保这些工具易于使用，并具备未来改进的潜力。在阐释如何理解、设计和记录 激励方案，并讨论研究结果与建议所带来的影响时，我们倡导采用一种更标准化（同时保持灵活）的整体激励设计流程。这反过来能够为人机决策领域产出更加可靠、可推广的知识铺平道路。归根结底，这一探索旨在帮助研究者建立有效的人机伙伴关系，发挥人类与机器各自的优势，从而在不同领域实现高质量的决策结果。

参考文献

[1] 1979.《前景理论：风险下决策的分析》。*Econometrica* 47, 2 (1979), 263–291。[2] 2018. 解释的可理解性评估。[3] Tahir Abbas 和 Ujwal Gadiraju。2022。众包平台工作者的目标设定行为：对 mturk 和 Prolific 的探索性研究。载于*AAAI*人类计算与众包会议论文集，第 10 卷。2–13。[4] Ahmed Alqaraawi、Martin Schuessler、Philipp Weiß、Enrico Costanza 和 Nadia Berthouze。2020。评估卷积神经网络的显著性图解释：一项用户研究。载于第 *25* 届国际智能用户界面会议论文集 （意大利卡利亚里） （*IUI '20*）。计算机协会，纽约州纽约市，美国，275–285。doi:10.1145/3377325.3377519[5] Chittaranjan Andrade。2018。研究设计、实施与评估中的内部效度、外部效度和生态效度。《印度心理医学杂志》40, 5 (2018), 498–499。doi:10.4103/IJPSYM.IJPSYM_334_18 PMID: 30275631。[6] Gagan Bansal、Besmira Nushi、Ece Kamar、Walter S. Lasecki、Daniel S. Weld 和 Eric Horvitz。2019。超越准确率：心理模型在人机团队绩效中的作用。*AAAI*人类计算与众包会议论文集7, 1 (2019年10月), 2–11。doi:10.1609/hcomp.v7i1.5285

[7] Gagan Bansal、Tongshuang Wu、Joyce Zhou、Raymond Fok、Besmira Nushi、Ece Kamar、Marco Tulio Ribeiro 和 Daniel Weld。2021。整体是否大于部分之和？人工智能解释对互补团队绩效的影响。载于《2021年CHI计算机系统中的人为因素大会论文集》（日本横滨）（CHI '21）。计算机协会，纽约州纽约市，美国，文章 81，共 16 页。doi:10.1145/3411764.3445717[8] Reuben Binns、Max Van Kleek、Michael Veale、Ulrik Lyngs、Jun Zhao 和 Nigel Shadbolt。2018。"这等于把一个人简化为一个百分比"：对算法决策中公平性的看法。载于《2018年CHI计算机系统中的人为因素大会论文集》（CHI '18）。计算机协会，纽约州纽约市，美国，1–14。doi:10.1145/3173574.3173951[9] Reuben Binns、Max Van Kleek、Michael Veale、Ulrik Lyngs、Jun Zhao 和 Nigel Shadbolt。2018。"这等于把一个人简化为一个百分比"：对算法决策中公平性的看法。载于《2018年CHI计算机系统中的人为因素大会论文集》（加拿大魁北克省蒙特利尔）（CHI '18）。计算机协会，纽约州纽约市，美国，1–14。doi:10.1145/3173574.3173951[10] Or Biran 和 Kathleen McKeown。2017。以人为本地说明机器学习预测。载于《第 26 届国际人工智能联合会议论文集》（澳大利亚墨尔本）（IJCAI'17）。AAAI出版社，1461–1467。[11] Or Biran 和 Kathleen McKeown。2017。以人为本地说明机器学习预测。载于《第 26 届国际人工智能联合会议论文集》（澳大利亚墨尔本）（IJCAI'17）。AAAI出版社，1461–1467。[12] Virginia Braun 和 Victoria Clarke。2006。在心理学中运用主题分析。《心理学定性研究》3, 2 (2006), 77–101。doi:10.1191/1478088706qp063oa[13] Virginia Braun 和 Victoria Clarke。2021。《主题分析：实践指南》（第一版）。[14] GLENN W. BRIER。1950。以概率形式表达的预报的验证。《月度天气评论》78, 1 (1950), 1–3。doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2[15] Bruna Bruno、Marisa Faggini 和 Anna Parziale。2017。动机、激励与绩效：跨学科综述。《国际商业与管理杂志》12 (2017年11月), 29。doi:10.5539/ijbm.v12n12p29[16] Zana Buçinca、Phoebe Lin、Krzysztof Z. Gajos 和 Elena L. Glassman。2020。在评估可解释人工智能系统时，代理任务和主观测量指标可能会产生误导。载于《第 25 届国际智能用户界面会议论文集》（意大利卡利亚里）（IUI '20）。计算机协会，纽约州纽约市，美国，454–464。[17] Zana Buçinca、Maja Barbara Malaya 和 Krzysztof Z. Gajos。2021。信任还是思考：认知强制功能能够减少对人工智能的过度依赖。Proc. ACM Hum.-Comput. Interact. 5, CSCW1, Article 188 (2021年4月), 21 pages。doi:10.1145/3449287[18] Francesco Cappa、Federica Rosso 和 Darren Hayes。2019。众包的金钱与社交奖励。《可持续性》11, 10 (2019)。doi:10.3390/su11102834[19] Hao-Fei Cheng、Ruotong Wang、Zheng Zhang、Fiona O'Connell、Terrance Gray、F. Maxwell Harper 和 Haiyi Zhu。2019。通过用户界面解释决策算法：帮助非专业利益相关者的策略。载于《2019年CHI计算机系统中的人为因素大会论文集》（苏格兰格拉斯哥）（CHI '19）。计算机协会，纽约州纽约市，美国，1–12。doi:10.1145/3290605.3300789[20] Justin Cheng、Jaime Teevan 和 Michael S Bernstein。2015。使用错误—时间曲线衡量众包工作量。载于《第 33 届 ACM计算机系统中的人为因素年度会议论文集》。1365–1374。[21] Chun-Wei Chiang 和 Ming Yin。2022。探究机器学习素养干预对普通公众依赖机器学习模型的影响。载于《第 27 届国际智能用户界面会议论文集》（芬兰赫尔辛基）（IUI '22）。计算机协会，纽约州纽约市，美国，148–161。doi:10.1145/3490099.3511121[22] Michael Chromik、Malin Eiband、Felicitas Buchner、Adrian Krüger 和 Andreas Butz。2021。我想我明白你的意思了，人工智能！可解释人工智能中的解释深度错觉。载于《第 26 届国际智能用户界面会议论文集》（美国得克萨斯州大学城）（IUI '21）。计算机协会，纽约州纽约市，美国，307–317。doi:10.1145/3397481.3450644[23] Xinyue Dai、Mark T. Keane、Laurence Shalloo、Elodie Ruelle 和 Ruth M.J. Byrne。2022。可解释人工智能中的预测与诊断反事实解释。载于《2022年AAAI/ACM人工智能、伦理与社会会议论文集》（英国牛津）（AIES '22）。计算机协会，纽约州纽约市，美国，215–226。doi:10.1145/3514094.3534144[24] Florian Daniel、Pavel Kucherbaev、Cinzia Cappiello、Boualem Benatallah 和 Mohammad Allahbakhsh。2018。众包中的质量控制：质量属性、评估技术与保障措施综述。ACM Comput. Surv. 51, 1, Article 7 (2018年1月), 40 pages。doi:10.1145/3148148[25] Devleena Das 和 Sonia Chernova。2020。利用理由说明提升人类任务绩效。载于《第 25 届国际智能用户界面会议论文集》（意大利卡利亚里）（IUI '20）。计算机协会，纽约州纽约市，美国，510–518。doi:10.1145/ 3377325.3377512[26] Esra Cemre Su de Groot 和 Ujwal Gadiraju。2024。"我们是否同舟共济？"通过可定制且不断演变的虚拟形象提升在线众包工作中的工作者参与度并增强社区归属感。载于《CHI计算机系统中的人为因素大会论文集》。1–26。[27] Berkeley Dietvorst、Joseph Simmons 和 Cade Massey。2014。算法厌恶：人们看到算法出错后会错误地回避算法。《SSRN电子期刊》(2014年1月)。doi:10.2139/ssrn.2466040[28] Berkeley Dietvorst、Joseph Simmons 和 Cade Massey。2018。克服算法厌恶：如果可以（哪怕稍微）修改不完美的算法，人们就会使用它们。《管理科学》64 (2018年3月), 1155–1170。doi:10.1287/mnsc.2016.2643

[29] Jonathan Dodge、Q. Vera Liao、Yunfeng Zhang、Rachel K. E. Bellamy 和 Casey Dugan。2019年。解释模型：解释如何影响公平性判断的实证研究。载于《第24届国际智能用户界面会议论文集》（加利福尼亚州 Marina del Ray）（IUI '19）。计算机协会，美国纽约州纽约，275–285。doi:10.1145/3301275.3302310[30] Steven Dow、Anand Kulkarni、Scott Klemmer 和 Björn Hartmann。2012。引导众包群体能带来更好的成果。载于《ACM 2012年计算机支持的协同工作会议论文集》。1013–1022。[31] Tim Draws、Alisa Rieger、Oana Inel、Ujwal Gadiraju 和 Nava Tintarev。2021。消除众包中认知偏差的清单。载于《AAAI人类计算与众包会议论文集》9，1（2021年10月），48–59。doi:10.1609/hcomp.v9i1.18939[32] Julia Dressel 和 Hany Farid。2018。预测再犯的准确性、公平性与局限。《科学进展》4，1（2018）。doi:10.1126/sciadv.aao5580[33] Jessica Maria Echterhoff、Matin Yarmand 和 Julian McAuley。2022。人工智能调节的决策：在序贯决策任务中捕捉并平衡锚定偏差。载于《2022年计算机系统中的人为因素会议论文集》（美国路易斯安那州新奥尔良）（CHI '22）。计算机协会，美国纽约州纽约，文章161，9页。doi:10.1145/3491102.3517443[34] Shaoyang Fan、Ujwal Gadiraju、Alessandro Checco 和 Gianluca Demartini。2020。Crowdco-op：在众包中共享风险与回报。《ACM人机交互汇刊》4，CSCW2（2020），1–24。[35] Alek Felstiner。2011。众包工作：众包行业中的就业与劳工法。《伯克利就业与劳动法期刊》32，1（2011），143–203。[36] Shi Feng 和 Jordan Boyd-Graber。2019。人工智能能为我做什么？评估协作式游戏中的机器学习解释。载于《第24届国际智能用户界面会议论文集》（加利福尼亚州 Marina del Ray）（IUI '19）。计算机协会，美国纽约州纽约，229–239。doi:10.1145/3301275.3302265[37] Oluwaseyi Feyisetan 和 Elena Simperl。2017。付费协作众包中的社会激励。《ACM智能系统与技术汇刊》（TIST）8，6（2017），1–31。[38] Ujwal Gadiraju、Gianluca Demartini、Ricardo Kawase 和 Stefan Dietze。2019。超越好与坏的众包群体构成：用于众包工作者建模和预筛选的行为轨迹。《计算机支持的协同工作》（CSCW）28（2019），815–841。[39] Ujwal Gadiraju、Sebastian Möller、Martin Nöllenburg、Dietmar Saupe、Sebastian Egger-Lampl、Daniel Archambault 和 Brian Fisher。2017。众包与实验室：利用众包开展以人为中心的实验。载于《众包中的评估：众包与以人为中心的实验：Dagstuhl研讨会15481，德国Dagstuhl城堡，2015年11月22日至27日，修订版论文集》。施普林格，6–26。[40] Luca Galli 和 Piero Fraternali。2014。成就系统解析。新加坡施普林格，新加坡，25–50。doi:10.1007/978-981-4560-26-9_3[41] Meric Altug Gemalmaz 和 Ming Yin。2022。理解决策对象在与基于人工智能的决策系统反复交互中的公平性感知与留存。载于《2022年AAAI/ACM人工智能、伦理与社会会议论文集》（英国牛津）（AIES '22）。计算机协会，美国纽约州纽约，295–306。doi:10.1145/3514094.3534201[42] Marjan Ghazvininejad、Xing Shi、Jay Priyadarshi 和 Kevin Knight。2017。Hafez：一个交互式诗歌生成系统。载于《ACL 2017系统演示》，Mohit Bansal 和 Heng Ji（编）。计算语言学协会，加拿大温哥华，43–48。[43] Antonio Ghezzi、Donata Gabelloni、Antonella Martini 和 Angelo Natalicchio。2018。众包：综述与未来研究建议。《国际管理评论期刊》20，2（2018），343–363。doi:10.1111/ijmr.12135[44] Ben Green 和 Yiling Chen。2019。差异化交互：算法在环分析中的风险评估公平性。载于《公平、问责与透明度会议论文集》（美国佐治亚州亚特兰大）（FAT* '19）。计算机协会，美国纽约州纽约，90–99。doi:10.1145/3287560.3287563[45] Ben Green 和 Yiling Chen。2019。算法在环决策的原则与局限。《ACM人机交互汇刊》3，CSCW，文章50（2019年11月），24页。doi:10.1145/3359152[46] Nina Grgić-Hlača、Christoph Engel 和 Krishna P. Gummadi。2019。机器辅助下的人类决策：一项有关保释与监禁的实验。《ACM人机交互汇刊》3，CSCW，文章178（2019年11月），25页。doi:10.1145/3359280[47] Nina Grgić-Hlača、Claude Castelluccia 和 Krishna P. Gummadi。2022。听取（相）似机器的建议：人机相似性对机器辅助决策的影响。载于《AAAI人类计算与众包会议论文集》10，1（2022），74–88。doi:10.1609/hcomp.v10i1.21989[48] Gaole He、Stefan Buijsman 和 Ujwal Gadiraju。2023。人工智能系统声明的准确率及用于解释准确率的类比如何影响人类对系统的依赖。《ACM人机交互汇刊》7，CSCW2，文章276（2023年10月），29页。doi:10.1145/3610067[49] Gaole He、Lucie Kuiper 和 Ujwal Gadiraju。2023。知其所知：人类能力的错觉可能妨碍对人工智能系统的恰当依赖。载于《2023年计算机系统中的人为因素会议论文集》（德国汉堡）（CHI '23）。计算机协会，美国纽约州纽约，文章113，18页。doi:10.1145/3544548.3581025[50] Chien-Ju Ho、Shahin Jabbari 和 Jennifer Wortman Vaughan。2013。用于众包分类的自适应任务分配。载于《第30届国际机器学习会议论文集》第28卷（美国佐治亚州亚特兰大）（ICML'13）。JMLR.org，I–534–I–542。[51] Chien-Ju Ho、Aleksandrs Slivkins 和 Jennifer Wortman Vaughan。2014。面向众包市场的自适应合约设计：用于重复委托代理问题的多臂老虎机算法。载于《第十五届ACM经济学与计算会议论文集》（美国加利福尼亚州帕洛阿尔托）（EC '14）。计算机协会，美国纽约州纽约，359–376。doi:10.1145/2600057.2602880[52] Yoyo Tsung-Yu Hou、Wen-Ying Lee 和 Malte Jung。2023。“我该听人类的，还是听机器人的？”——掌权的机器人对决策的影响可能超过人类。载于《2023年计算机系统中的人为因素会议论文集》（德国汉堡）（CHI '23）。计算机协会，美国纽约州纽约，文章114，13页。doi:10.1145/3544548.3581066 提交给 ACM 的稿件

[53] Bernd Huber 和 Krzysztof Z Gajos。2020。使用未获报酬、未经监督的样本开展在线虚拟环境实验。PLOS ONE 15，1（2020），e0227629。[54] Lilly C Irani 和 M Six Silberman。2013。Turkopticon：打破亚马逊土耳其机器人中的劳动者隐形化。载于《SIGCHI计算机系统中的人为因素会议论文集》。611–620。[55] Farnaz Jahanbakhsh、Yannis Katsis、Dakuo Wang、Lucian Popa 和 Michael Muller。2023。探索利用个性化人工智能识别社交媒体上的虚假信息。载于《2023年计算机系统中的人为因素会议论文集》（德国汉堡）（CHI '23）。计算机协会，美国纽约州纽约，文章105，27页。doi:10.1145/3544548.3581219[56] Daniel Kahneman 和 Amos Tversky。1979。前景理论：风险下决策的分析。《Econometrica》47，2（1979年3月），263–291。https://ideas.repec.org/a/ecm/emetrp/v47y1979i2p263-91.html[57] Patricia K. Kahr、Gerrit Rooks、Martijn C. Willemsen 和 Chris C.P. Snijders。2023。看起来聪明，行动却愚蠢：在重复法律决策任务中建立对人工智能建议的信任。载于《第28届国际智能用户界面会议论文集》（澳大利亚新南威尔士州悉尼）（IUI '23）。计算机协会，美国纽约州纽约，528–539。doi:10.1145/3581641.3584058[58] Rea Karachiwalla 和 Felix Pinkow。2021。理解众包项目：众包倡议关键设计要素综述。《创意与创新管理》30，3（2021），563–584。doi:10.1111/caim.12454[59] David Karger、Sewoong Oh 和 Devavrat Shah。2011。可靠众包系统中预算最优的任务分配。《运筹学》62（2011年10月）。doi:10.1287/opre.2013.1235[60] N. Kaufmann、Thimo Schulze 和 Daniel Veit。2011。不止乐趣与金钱：众包中的劳动者动机——一项关于土耳其机器人的研究。载于《第十七届美洲信息系统会议论文集》。[61] Aniket Kittur、Ed H. Chi 和 Bongwon Suh。2008。利用土耳其机器人开展众包用户研究。载于《SIGCHI计算机系统中的人为因素会议论文集》（意大利佛罗伦萨）（CHI '08）。计算机协会，美国纽约州纽约，453–456。doi:10.1145/ 1357054.1357127[62] Markus Knauff 和 Ann Wolf。2010。编者按：复杂认知：人类推理、问题解决与决策的科学。《认知加工》11（2010年3月），99–102。doi:10.1007/s10339-010-0362-z[63] Rafal Kocielnik、Saleema Amershi 和 Paul N. Bennett。2019。你会接受不完美的人工智能吗？探索调整终端用户对人工智能系统预期的设计。载于《2019年计算机系统中的人为因素会议论文集》（英国苏格兰格拉斯哥）（CHI '19）。计算机协会，美国纽约州纽约，1–14。doi:10.1145/3290605.3300641[64] Adriana Kovashka、Olga Russakovsky、李飞飞 和 Kristen Grauman。2016。计算机视觉中的众包。《计算机图形学与视觉基础与趋势》10（2016年1月），177–243。doi:10.1561/0600000071[65] Vivian Lai、Samuel Carton、Rajat Bhatnagar、Q. Vera Liao、Yunfeng Zhang 和 Chenhao Tan。2022。通过条件式委派实现人机协作：内容审核案例研究。载于《2022年计算机系统中的人为因素会议论文集》（美国路易斯安那州新奥尔良）（CHI '22）。计算机协会，美国纽约州纽约，文章54，18页。doi:10.1145/3491102.3501999[66] Vivian Lai、Chacha Chen、Alison Smith-Renner、Q. Vera Liao 和 Chenhao Tan。2023。迈向人机决策科学：实证人类受试者研究中设计空间概述。载于《2023年ACM公平、问责与透明度会议论文集》（FAccT '23）。计算机协会，美国纽约州纽约，1369–1385。[67] Vivian Lai、Han Liu 和 Chenhao Tan。2020。"为什么'芝加哥'具有迷惑性？"——构建面向人类的模型驱动教程。载于《2020年计算机系统中的人为因素会议论文集》（CHI '20）。计算机协会，美国纽约州纽约，1–13。doi:10.1145/3313831.3376873[68] Philippe Lammerts、Philip Lippmann、Yen-Chia Hsu、Fabio Casati 和 Jie Yang。2023。感觉如何？衡量用户在仇恨言论检测中拒绝机器决策的感知价值。载于《2023年AAAI/ACM人工智能、伦理与社会会议论文集》（加拿大魁北克省蒙特利尔）（AIES '23）。计算机协会，美国纽约州纽约，834–844。doi:10.1145/3600211.3604655[69] Philippe Lammerts、Philip Lippmann、Yen-Chia Hsu、Fabio Casati 和 Jie Yang。2023。感觉如何？衡量用户在仇恨言论检测中拒绝机器决策的感知价值。载于《2023年AAAI/ACM人工智能、伦理与社会会议论文集》。834–844。[70] Huigang Liang、Meng-Meng Wang、Jian-Jun Wang 和 Yajiong Xue。2018。内在动机与外在激励如何影响众包竞赛中的任务投入：中介调节模型。《人类行为中的计算机》81（2018），168–176。doi:10.1016/j.chb.2017.11.040[71] Brian Y. Lim、Anind K. Dey 和 Daniel Avrahami。2009。为什么及为何不的解释能提升情境感知智能系统的可理解性。载于《SIGCHI计算机系统中的人为因素会议论文集》（美国马萨诸塞州波士顿）（CHI '09）。计算机协会，美国纽约州纽约，2119–2128。doi:10.1145/1518701.1519023[72] Leib Litman、Jonathan Robinson 和 Cheskie Rosenzweig。2015。美国与印度工人在土耳其机器人平台上的动机、金钱报酬与数据质量之间的关系。《行为研究方法》47，2（2015），519–528。doi:10.3758/s13428-014-0483-x[73] Yihe Liu、Anushk Mittal、Diyi Yang 和 Amy Bruckman。2022。我失去宠物时，人工智能会安慰我吗？理解人们对人工智能辅助邮件写作的看法。载于《2022年计算机系统中的人为因素会议论文集》（美国路易斯安那州新奥尔良）（CHI '22）。计算机协会，美国纽约州纽约，文章474，13页。doi:10.1145/3491102.3517731[74] Zhuoran Lu 和 Ming Yin。2021。当绩效反馈有限时，人类对机器学习模型的依赖：启发式方法与风险。载于《2021年计算机系统中的人为因素会议论文集》（日本横滨）（CHI '21）。计算机协会，美国纽约州纽约，文章78，16页。doi:10.1145/3411764.3445562

[75] Shuai Ma、Ying Lei、Xinru Wang、Chengbo Zheng、Chuhan Shi、Ming Yin 和 Xiaojuan Ma。2023。该相信谁：人工智能还是自己？利用人类和人工智能的正确性概率，促进对人工智能辅助决策的适当信任。载于《2023年计算机系统中的人为因素会议论文集》（德国汉堡）（CHI '23）。计算机协会，纽约，美国纽约州，论文 759，19 页。doi:10.1145/3544548.3581058[76] Andrew Mao、Ece Kamar、Yiling Chen、Eric Horvitz、Megan Schwamb、Chris Lintott 和 Arfon Smith。2013。志愿服务与有偿工作：众包中的激励与权衡。载于 AAAI人类计算与众包会议论文集，第 1 卷。94–102。[77] Anoush Margaryan、Timothy Charlton-Czaplicki 和 Ujwal Gadiraju。2020。在线平台工作中的学习与技能发展：比较 Microworkers 和在线自由职业者的实践（CrowdLearnPlus）。（2020）。[78] Winter Mason 和 Duncan J. Watts。2009。经济激励与"群体绩效"。载于 ACM SIGKDD 人类计算研讨会论文集（法国巴黎）（HCOMP '09）。计算机协会，纽约，美国纽约州，77–85。doi:10.1145/1600150.1600175[79] Joseph E. Mercado、Michael A. Rupp、Jessie Y. C. Chen、Michael J. Barnes、Daniel Barber 和 Katelyn Procci。2016。面向多无人驾驶飞行器管理的人类—智能体团队协作中的智能体透明度。《Human Factors: The Journal of the Human Factors and Ergonomics Society》58，3（2016），401–415。doi:10.1177/0018720815621206[80] Xiaoye Miao、Huanhuan Peng、Yunjun Gao、Zongfu Zhang 和 Jianwei Yin。2023。众包任务的动态定价。《ACM知识发现与数据挖掘汇刊》17，2，论文 30（2023年2月），27 页。doi:10.1145/3544018[81] Benedikt Morschheuser、Juho Hamari 和 Jonna Koivisto。2016。众包中的游戏化：综述。载于《2016年第49届夏威夷国际系统科学会议》（HICSS）。4375–4384。doi:10.1109/HICSS.2016.543[82] Benedikt Morschheuser、Juho Hamari、Jonna Koivisto 和 Alexander Maedche。2017。游戏化众包：概念化、文献综述与未来议程。《International Journal of Human-Computer Studies》106（2017），26–43。[83] Conor Muldoon、Michael O'Grady 和 Gregory O'Hare。2018。众包激励设计调查。《The Knowledge Engineering Review》33（2018年4月）。doi:10.1017/S0269888918000061[84] Saumik Narayanan、Guanghui Yu、Chien-Ju Ho 和 Ming Yin。2023。价值观相似性如何影响人类对人工智能的依赖？人工智能辅助伦理决策研究。载于《2023年AAAI/ACM人工智能、伦理与社会会议论文集》（加拿大魁北克省蒙特利尔）（AIES '23）。计算机协会，纽约，美国纽约州，49–57。doi:10.1145/3600211.3604709[85] Fábio Neto 和 Celso Saibel Santos。2018。理解众包项目：对趋势、工作流程和质量管理的系统性综述。《Information Processing and Management》54（2018年7月），490–506。doi:10.1016/j.ipm.2018.03.006[86] Oded Nov、Ofer Arazy 和 David Anderson。2014。Scientists@Home：是什么推动了在线公民科学参与的数量与质量？《PLOS ONE》9（2014年4月），e90375。doi:10.1371/journal.pone.0090375[87] Jonas Oppenlaender、Tahir Abbas 和 Ujwal Gadiraju。2024。众包试点研究报告现状：对最佳实践和指南的反思。《ACM人机交互汇刊》8，CSCW1（2024），1–45。[88] Jonas Oppenlaender、Tahir Abbas 和 Ujwal Gadiraju。2024。众包试点研究报告现状：对最佳实践和指南的反思。《ACM人机交互汇刊》8，CSCW1，论文 184（2024年4月），45 页。doi:10.1145/3641023[89] Gabriele Paolacci、Jesse Chandler 和 Panagiotis G. Ipeirotis。2010。在亚马逊土耳其机器人上开展实验。《Judgment and Decision Making》5，5（2010），411–419。doi:10.1017/S1930297500002205[90] Eyal Peer、Laura Brandimarte、Sonam Samat 和 Alessandro Acquisti。2017。超越土耳其机器人：用于众包行为研究的替代平台。《Journal of Experimental Social Psychology》70（2017），153–163。[91] Forough Poursabzi-Sangdeh、Daniel G Goldstein、Jake M Hofman、Jennifer Wortman Wortman Vaughan 和 Hanna Wallach。2021。操控并测量模型的可解释性。载于《2021年计算机系统中的人为因素会议论文集》（日本横滨）（CHI '21）。计算机协会，纽约，美国纽约州，论文 237，52 页。doi:10.1145/3411764.3445315[92] Goran Radanovic、Boi Faltings 和 Radu Jurca。2016。利用同伴真实度检验机制激励众包参与者付出努力。《ACM智能系统与技术汇刊》（TIST）7，4（2016），1–28。[93] Jorge Ramírez、Burcu Sayin、Marcos Baez、Fabio Casati、Luca Cernuzzi、Boualem Benatallah 和 Gianluca Demartini。2021。众包实验报告现状及帮助改进当前实践的清单。《ACM人机交互汇刊》5，CSCW2（2021），1–34。[94] Charvi Rastogi、Yunfeng Zhang、Dennis Wei、Kush R. Varshney、Amit Dhurandhar 和 Richard Tomsett。2022。快速与慢速决策：认知偏差在人工智能辅助决策中的作用。《ACM人机交互汇刊》6，CSCW1，论文 83（2022年4月），22 页。doi:10.1145/3512930[95] Katharina Reinecke 和 Krzysztof Z Gajos。2015。LabintheWild：使用无偿样本开展大规模在线实验。载于第 18 届 ACM 计算机支持的协同工作与社会计算会议论文集。1364–1378。[96] Markus Rokicki 、 Sergej Zerr 和 Stefan Siersdorfer。2015。群组众包：众包中的团队竞赛设计。载于第 24 届国际万维网会议论文集。906–915。[97] Sara Salimzadeh、Gaole He 和 Ujwal Gadiraju。2024。应对不确定性：理解预后任务与诊断任务对人机决策中信任与依赖的影响。载于《CHI人类因素与计算系统会议论文集》（美国夏威夷州檀香山）（CHI '24）。计算机协会，纽约，美国纽约州，论文 25，17 页。doi:10.1145/3613904.3641905[98] Jakob Schoeffer、Niklas Kuehl 和 Yvette Machowski。2022。"信息不足"：解释对自动化决策中信息公平性感知和可信度的影响。载于《2022年ACM公平、问责与透明度会议论文集》（韩国首尔）（FAccT '22）。计算机协会，纽约，美国纽约州，1616–1628。doi:10.1145/

3531146.3533218[99] Alexander Skulmowski 和 Kate Xu。2021。理解数字与在线学习中的认知负荷：外在认知负荷的新视角。《Educational Psychology Review》34（2021年6月），1–26。doi:10.1007/s10648-021-09624-7[100] Alison Smith-Renner、Ron Fan、Melissa Birchfield、Tongshuang Wu、Jordan Boyd-Graber、Daniel S. Weld 和 Leah Findlater。2020。没有问责，就没有可解释性：对交互式机器学习中解释与反馈的实证研究。载于《2020年计算机系统中的人为因素会议论文集》（美国夏威夷州檀香山）（CHI '20）。计算机协会，纽约，美国纽约州，1–13。doi:10.1145/3313831. 3376624[101] Michael Soprano、Kevin Roitero、Ujwal Gadiraju、Eddy Maddalena 和 Gianluca Demartini。2024。纵向忠诚度：理解在众包平台上开展纵向研究的障碍。《ACM社会计算汇刊》7，1-4（2024），1–49。[102] Aaron Springer 和 Steve Whittaker。2019。渐进式披露：以实证为依据的有效透明度设计方法。载于第 24 届国际智能用户界面会议论文集（加利福尼亚州 Marina del Ray）（IUI '19）。计算机协会，纽约，美国纽约州，107–120。doi:10.1145/3301275.3302322[103] Susan Standing 和 Craig Standing。2018。众包的伦理使用。《Business Ethics: A European Review》27，1（2018），72–80。doi:10.1111/ beer.12173[104] James R. Terborg 和 Howard E. Miller。1978。动机、行为与绩效：进一步探讨目标设定与金钱激励。《应用心理学杂志》63，1（1978），29–39。doi:10.1037/0021-9010.63.1.29[105] Suzanne Tolmeijer、Markus Christen、Serhiy Kandul、Markus Kneer 和 Abraham Bernstein。2022。能力出众却不道德？伦理决策中人工智能与人类专家协作的比较。载于《2022年计算机系统中的人为因素会议论文集》（美国路易斯安那州新奥尔良）（CHI '22）。计算机协会，纽约，美国纽约州，论文 160，17 页。doi:10.1145/3491102.3517732[106] Long Tran-Thanh、Trung Dong Huynh、Avi Rosenfeld、Sarvapali D. Ramchurn 和 Nicholas R. Jennings。2014。BudgetFix：用于相互依赖任务分配且提供质量保证的预算受限众包。载于《2014年国际自主智能体与多智能体系统会议论文集》（法国巴黎）（AAMAS '14）。国际自主智能体与多智能体系统基金会，南卡罗来纳州里奇兰，477–484。[107] Lauren S. Treiman、Chien-Ju Ho 和 Wouter Kool。2023。人类放弃奖励，以将公平性植入人工智能。《AAAI人类计算与众包会议论文集》11，1（2023），152–162。doi:10.1609/hcomp.v11i1.27556[108] Edwin van Teijlingen 和 Vanora Hundley。2002。试点研究的重要性。《Nursing Standard》（英国皇家护理学院，英国：1987）16，40（2002年6月），33–36。doi:10.7748/ns2002.06.16.40.33.c3214[109] Helena Vasconcelos、Matthew Jörke、Madeleine Grunde-McLaughlin、Tobias Gerstenberg、Michael S. Bernstein 和 Ranjay Krishna。2023。解释能够减少决策过程中对人工智能系统的过度依赖。《ACM人机交互汇刊》7，CSCW1，论文 129（2023年4月），38 页。doi:10.1145/3579605[110] Kailas Vodrahalli、Roxana Daneshjou、Tobias Gerstenberg 和 James Zou。2022。如果建议来自人工智能，人类会更加信任它吗？人机交互分析。载于《2022年AAAI/ACM人工智能、伦理与社会会议论文集》（英国牛津）（AIES '22）。计算机协会，纽约，美国纽约州，763–777。doi:10.1145/3514094.3534150[111] Lili Wang、Ye Yang 和 Yong Wang。2019。更高的激励会带来更好的绩效吗？——软件众包的探索性研究。载于《2019年ACM/IEEE国际实证软件工程与度量研讨会》（ESEM）。1–11。doi:10.1109/ESEM.2019.8870175[112] Mark E Whiting、Dilrukshi Gamage、Snehalkumar (Neil) S Gaikwad、Aaron Gilbee、Shirish Goyal、Alipta Ballav、Dinesh Majeti、Nalin Chhibber、Angela Richmond-Fuller、Freddie Vargus 等。2017。众包行会：众包平台上工人主导的声誉与反馈。载于《2017年ACM计算机支持的协同工作与社会计算会议论文集》。1902–1913。[113] Mark E Whiting、Grant Hugh 和 Michael S Bernstein。2019。公平工作：只需一行代码即可实现众包工作最低工资。载于 AAAI人类计算与众包会议论文集，第 7 卷。197–206。[114] Vanessa Williamson。2016。论众包研究的伦理。《PS: Political Science & Politics》49（2016年1月），77–81。doi:10.1017/ S104909651500116X[115] Ming Yin、Yiling Chen 和 Yu-An Sun。2013。绩效挂钩的经济激励对在线劳动力市场的影响。《第 27 届 AAAI 人工智能会议论文集》，AAAI 2013 27（2013年6月），1191–1197。doi:10.1609/aaai.v27i1.8461[116] Ming Yin、Jennifer Wortman Vaughan 和 Hanna Wallach。2019。理解准确率对机器学习模型信任度的影响。载于《2019年计算机系统中的人为因素会议论文集》（英国苏格兰格拉斯哥）（CHI '19）。计算机协会，纽约，美国纽约州，1–12。doi:10.1145/3290605.3300509[117] Man-Ching Yuen、Irwin King 和 Kwong-Sak Leung。2011。众包系统调查。载于《2011年IEEE第三届隐私、安全、风险与信任国际会议》及《2011年IEEE第三届社会计算国际会议》。766–773。doi:10.1109/PASSAT/SocialCom.2011.203[118] Mireia Yurrita、Tim Draws、Agathe Balayn、Dave Murray-Rust、Nava Tintarev 和 Alessandro Bozzon。2023。解析算法决策中的公平性感知：解释、人工监督和可争议性的影响。载于《2023年计算机系统中的人为因素会议论文集》（德国汉堡）（CHI '23）。计算机协会，纽约，美国纽约州，论文 134，21 页。doi:10.1145/3544548.3581161[119] Yunfeng Zhang、Q. Vera Liao 和 Rachel K. E. Bellamy。2020。置信度与解释对人工智能辅助决策中准确率和信任校准的影响。载于《2020年公平、问责与透明度会议论文集》（西班牙巴塞罗那）（FAT* '20）。计算机协会，纽约，美国纽约州，295–305。doi:10.1145/3351095.3372852

[120] Yu Zhang 和 Mihaela Van der Schaar。2012。众包应用中的基于声誉的激励协议。载于 *2012 IEEE INFOCOM* 会议论文集。IEEE, 2140–2148。