无以信，何以立：人机交互中的可持续信任机制

向安玲
（中央民族大学新闻与传播学院，北京 100081）

摘 要：可持续信任是保障人机高频交互和高效协作的关键要素。基于计算扎根理论，文章针对 7235 条 ChatGPT 用户反馈进行编码分析，并由此提炼可持续信任的影响因素。研究发现，机器因素（可用性、易用性、可供性和安全性等）占比最大，其次为用户要素（技术恐惧、需求适配、媒介素养、心理预期），而任务因素（关键性失误、任务复杂度、违规成本）占比偏低。关键任务失败、机器安全性、用户需求适配度对用户信任水平影响最为显著。在人机之间建立清晰而互补的职责界限，用算法的可解释性对冲输出的不确定性，同时通过奖惩机制的“界面化”，引导用户调整心理预期，有助于可持续信任的动态校准。

关键词：生成式人工智能；人机交互；可持续信任；影响因素
中图分类号：TP18-02 **文献标识码：**A **文章编号：**2096-8418（2024）02-0029-13

从蒸汽时代、电气时代，到信息时代，再到智能时代，人类历史上每一次技术革命和社会范式变革都伴随着人机交互关系的演化和人机交互场景的跃迁。机器从替代人力劳动、辅助工业生产，也逐步进化到延伸人的感知、激发创意思维。在这个过程中，人机交互逐步趋向人机协同，甚至人机共生。伴随着智能体的指数级增长与全行业渗透，人机交互场景将进一步拓展，人与智能系统之间的交互关系也将重新被定义。机器是否会替代人？多大程度上能取代人力工作？如何在挖掘技术红利的基础上规避技术滥用？如何建构可持续的人机协作关系？这一系列问题都有待学界和业界回应。如果说技术能力迭代是回应以上问题的“硬变量”，那么人机信任关系就是影响以上问题的“软变量”。

信任程度直接影响人机交互强度和可持续性，进一步影响人机协作效能和生产效率。在智能技术加速发展的当下，过度信任带来的技术滥用和信任不足衍生的技术弃用普遍存在。尤其是对于人工智能生成内容（Artificial Intelligence Generated Content，AIGC）这类大众级智能交互产品，机器能力的通用化和实用场景的泛化，使得人机交互成为一个常态化的需求。在持续性、高频次的交互中，人机信任也在不断演变，从信任产生到信任波动，再到信任校准，人对机器的依赖性或排斥性也在动态变化。可持续交互中的人机信任在很大程度上影响了生产力输出，同时也滋生了错误信息、虚假信息、学术伦理、技术偏见等一系列社会问题。

在 AIGC 火爆出圈，迅速渗透各行各业的当下，我们应该如何理性地去看待和使用它？如何调整信任水平以避免滥用或弃用？有哪些因素可以干预信任水平？信任又如何影响到人机协作结果？这些问题都有待学界和业界进一步探讨和实践。

基金项目：国家自然科学基金青年项目“面向人工智能生成内容的风险识别与治理策略研究”（72304290）。
作者简介：向安玲，女，讲师，博士。

一、文献回顾

（一）从感性到理性：人机协作中的信任生成

哲学视角下的传统信任研究强调信任的情感内核，认为信任是一种基于对方品德而产生的心理依赖。[1] 这种感性驱动的信任侧重于建构人与人之间的深层次连接。相比于基于道德准则的人际信任，人机交互过程中的电子信任（e-trust）更多基于实践结果，也即机器完成某项既定任务的能力和价值产出。[2] 从这个角度来看，人机信任是一种从（机）客观能力到（人）主观感知的生成机制，是人对机器执行特定目标可控性的主观评估。具体而言，当人们认为机器的行动能力在可控范围内，其执行任务有益或至少不会对人产生危害时，人们会选择信任它们。[3] 可以说，信任的生成是一种感性判断和理性标准的融合，既包含了对诚信和仁爱的心理期望[4]，对机器道德和行为规范的伦理期待[5]，又包含了对机器可靠性和可替性的理性判断[6]，对机器决策和产出结果的现实评估[7]。

从信任生成与演化阶段来看，人机信任可划分为初始信任（initial trust）和可持续信任（sustainable trust）。[8] 其中，初始信任是指双方初次互动或协作时，在没有直接经验的前提下所产生的信任。[9] 而可持续信任则是用户在反复交互和使用过程中对机器产生的信任，通过对其行为的熟悉和理解来减少不确定性，进一步调节信任程度。[10] 可持续信任是人机协作过程中的一种适应性过程。人类对人工智能的信任水平可根据情境和任务的变化而自适应地调整，相比于先验经验或既有判断，这种信任是一种动态评估与调节机制[11]。

如果说初始信任反映的是人类对机器的先验印象，可持续信任则反映了人对人机协同过程和结果的现实评估。前者更多指向感性判断，是人在信息不充分的情形下基于自身既有认知衍生的信任。而后者更多依赖于理性决策，是将合作行为的结果加入到信任的反馈环节中所形成的判断。[12] 从初始信任到可持续信任的演化，既是人机交互频次深化的结果，也是人机协作目标趋于一致的过程。[13]

（二）从相似到差异：人机信任的影响因素

从传统信任生成机制来看，熟悉度、期望值和风险共同决定了信任的强弱。其中，熟悉度指向了信任主体之间的关系构建，期望值指向了任务完成的预期程度，风险则指向了决策失败的机会成本。[14] 甚至可以说，信任就是一种风险投资。[15] 对“投资对象”的熟悉度、对“投资结果”的预期值以及“投资环境”的风险度都会影响信任的生成。

人机信任相比于人际信任，其生成与演化具备更强的不确定性和不稳定性。因为机器本身的行动存在不确定性，限于人们在经验和认知上的局限，无法对机器行动进行预测和评估。这就使得信任处于可控和不可控的双重关系中。[16] 一方面，人们希望机器更接近人类行为逻辑，具备一定拟人性，以增强其可控性和可解释性；另一方面，人们也希望机器在任务执行和决策过程中具备超人性，尤其是在复杂任务执行中突破既有模式。具体而言，对“相似”的期待主要体现在：（1）透明度（transparency）。机器（AI）算法模型和内部规则越清晰，与人的预期越相似，人们对其信任度往往越高；[17]（2）可解释性（interpretability）。“黑箱算法”所衍生的非线性和复杂决策过程往往很难被人为理解，机器逻辑和人类逻辑的差异阻碍了信任生成；[18]（3）拟人性（anthropomorphism）。类似人的说话方式、声音、面部特征、自我意识和情感表达往往会增加人对机器的信任感[19]，但也需警惕拟人化过程中“恐怖谷”（uncanny valley）效应对信任的削弱；[20]（4）即时行为（immediacy behaviors）。类似于人类的即时反应会激发情感上的信任，此外会犯错的机器人比完美无缺的机器人往往更受欢迎。[21]

相较而言，信任生成过程中的人机“差异”体现在：（1）智能性（intelligence）。机器在决策效率、产出规模、产出效能等方面均具备一定“超人”性，尤其是在复杂任务解决过程中人往往更加依赖于算法；[22]（2）接触方式。人际信任往往受到身体、视觉层面的物理接触影响，而人机信任基于数字界面交互，介质差异也会干预信任程度；[23]（3）伦理规范。数字环境与现实社会伦理规约差异也会影响信任生成[24]，规范人工智能向“善”发展是建立人机之间信任关系的前提条件。[25]

无论是初始信任，还是可持续信任，其影响因素均可划分为三个类别：用户特性、机器特性和环境特性。综合相关学者[26]-[31] 对人机信任的归因研究，将相关因素梳理如表 1 所示。

表 1　人机信任影响因素

	初始信任	可持续信任
用户特性	技术认知、文化背景、个性特征、经验知识、自信心	协作/交互频次、媒介素养、专业能力、心理预期、注意力控制、情绪、风险承受力
机器特性	可测试性、透明度、感知易用性、感知可用性	可靠性、鲁棒性、系统故障、意图说明、可预测性、可理解性/可解释性、可替代性、隐私保护、社交能力、目标一致性
环境特性	社会评价、社会规范、组织设置、平台背书	任务难度、风险承担、工作负荷、违规成本

（三）从抑制到修复：适度信任下的效能激发

人机合作信任是一个动态演化过程，人对机器表现的心理预期和机器自身实际效能不断混合交互。[32] 期望水平、实际表现和交互程度三方面原因共同导致了人机之间信任不足或信任过度的问题，通过自适应校准实现最优信任水平有利于提升人机协作安全性和效率。[33] 一方面，信任过度会导致技术的误用和滥用，从而衍生安全、质量、伦理等问题；[34] 另一方面，信任不足又可能导致技术的低效能甚至是弃用，在带来负面情感体验的同时影响团队绩效，进一步影响人机持续性合作。[35] 而要充分挖掘技术红利，实现人机协作效能最优化，就需要建构信任再定位（trust re-orientation）和信任校准机制（trust calibration）。

从“人”的角度来看，在持续合作中不断调整自身的认知兼容性、情感兼容性、价值兼容性，可实现信任水平校准。[36] 从“机”的角度来看，通过引入可解释的“白箱”系统，采用持续性强化学习模型，可实现协作优化与信任修复。[37] 从“交互”的角度来看，可通过动态修改信息呈现方式、重新分配任务、增加操作权限、明确操作边界等方式来调整人机合作中的信任水平。[38]

综合而言，适度信任是保障人机可持续协作的关键要素。当下包括 ChatGPT 内在的 AI 交互产品逐渐渗透信息生产与传播各个环节，过度信任导致的技术滥用和信任不足导致的技术弃用普遍存在。挖掘影响可持续信任的影响要素，优化信任校准机制，在适度信任下才能充分激活人机协作效能和技术应用红利。在此背景下，本文从 ChatGPT 用户的使用体验切入，通过对“失望”“失信”与“弃用”等评价进行计算扎根分析，由此提炼影响可持续信任生成和演变的关键因素，进一步探讨信任校准机制及其对人机协作的影响。具体来看，主要针对以下三大问题展开研究：

研究问题 1：影响人机可持续信任的因素有哪些？具体可划分为哪些维度和指标？

研究问题 2：不同影响因素的重要性如何？哪些因素对用户持续使用行为影响更大？

研究问题 3：如何实现人机可持续信任的动态校准？如何避免用户对技术的过度信任和信任不足？

二、研究设计

（一）研究方法：计算扎根理论

计算扎根理论（computational grounded theory）是一种利用计算机技术和大数据进行扎根理论研究的方法，旨在通过大规模数据计算提取出关键概念、关系和主题，从而建构新的理论模式。[39] 其核心思想和应用逻辑延承于传统的扎根理论，但不同于传统扎根的定性研究范式与主观解读取向[40]，计算扎根理论更多地应用计算机辅助手段对数据资料进行量化处理，并从大规模数据中提取出客观的模式和关系，进一步生成具有理论性的解释和结论。

综合相关学者的理论研究和实证探讨，本文将计算扎根理论的实现梳理为五大步骤：（1）大规模数据收集. 包括社交媒体数据、新闻媒体数据、开源网络数据、图书资料数据、访谈数据等多源异构数据。（2）数据预处理。如去除重复项、过滤无关信息、标准化与结构化等。（3）文本分析。基于计算机程序对数据进行编码和分类，包括词频统计、文本聚类、主题模型等，从而提出数据中的关键概念、关系和主题。（4）模式发现。基于扎根理论的核心思想和操作方法，结合机器计算和经验解释来挖掘数据中的规律和模式。（5）理论生成。根据前期从数据中发现的模式和规律，建构新的理论或验证现有理论，并将其转化为可操作的知识。

（二）研究对象与数据采样

相比于基于先验印象的初始信任，可持续信任来自于用户对人机协作过程和结果的现实评估。受心理预期和使用体验的影响，在采样时需选取有实际使用体验的现实用户。本文以 ChatGPT 作为研究对象，探讨用户对其可持续信任的影响因素。采样对象限定为有产品使用经验并基于使用体验对其发表评价的用户。为了精准筛选样本用户，本文以 AppleStore 和 GooglePlay 应用市场作为采样平台。这两个平台分别为苹果（IOS）系统和安卓（Android）系统应用产品下载的主流渠道，完成下载和安装的用户可在平台上进行产品评分（分五级）和评价。

截至 2023 年 11 月 15 日，本文通过网络爬虫从两大应用市场共采集了 55100 条 ChatGPT 的用户评价数据，包括 6900 条苹果系统评价和 48200 条安卓系统评价。综合来看，评分在 4-5 分的用户对产品持较高评价，而评分在 3 分及以下的用户往往存在负面使用体验，并可能对其后续产品使用造成影响。相比于正面使用体验，负面体验中往往更能反映用户的使用痛点与核心需求。为了探索影响产品可持续信任的因素，本文从用户使用评价中筛选了 7235 条中差评（3 分及以下评论），包括 1565 条（占比 22.68%）苹果系统产品评价和 5670 条（占比 11.76%）安卓系统产品评价。其中包括 6938 条英文评价（95.89%），71 条中文评价（0.98%）和其他语种评价（包括西班牙语、意大利语、法语等）。在此基础上，通过计算扎根理论来提炼影响产品可持续信任及可持续使用的因素及维度。

（三）基于 LDA 的主题聚类

为了更进一步探索影响用户信任及可持续使用的因素，对采样数据进行 LDA 主题聚类，得到 5 个主题在文本中的分布概率及相关高频词。结合实际应用场景可将主题归纳为五方面：

第一，错误响应与内容质量问题（如 wrong、bad、answer）。占比约为 24.19%，主要是针对 AI 输出内容的准确性和响应质量的评价，包括错误回答和不符合预期的内容响应。

其二，交互功能与计算性能（如 version、bug、text）。占比约为 15.77%，主要针对文本交互、插件功能和版本迭代后的功能升级所发表的评价，包括功能缺陷和交互界面相关问题。

其三，内容时效性与数据处理（如 time、update、2021）。占比约为 16.67%，主要针对产品底层数

据库的更新问题及错误数据反馈，包含过时回答、单位时间内交互频次限制和无法响应等情况。

其四，个人信息安全与终端使用体验（如 phone number、email、safe）。占比约为 10.97%，主要是针对产品存在的隐私泄露风险、终端适配性、操作安全性等问题进行讨论。

第五，登录限制与操作门槛（如 login、account、sign、can't）。占比约为 32.40%，主要是针对注册和登录的区域限制、产品及部分功能的使用门槛限制进行讨论。

（四）范畴提炼与模型建构

为了对相关影响因素进行结构化分析，基于聚类结果和细化编码对核心概念、范畴和关系进行提炼。首先是对原始语料进行开放式编码，也即扎根理论中的第一层级编码，侧重于将原始文本抽象为概念并进行范畴化的归纳。在具体操作过程中，剔除部分无效数据（如语意不清的模糊表达）后，针对剩余 7201 条文本进行逐句编码和统计，最终得到 20 个次要范畴。

主轴编码是扎根理论中的第二层编码，也即针对开放式编码结果进行聚类处理，根据其内在逻辑关联构建更高维度的独立范畴，从而形成更具概括性、抽象化、结构性的编码维度。结合前期文献调研，本文将次要范畴进一步归类为 13 项主要范畴。

理论编码是扎根理论中的第三层编码，根据主要范畴之间的归属关系进一步归纳核心维度，从而建构整体的理论框架。本文将前期编码结果抽象为用户因素、任务因素、机器因素三大核心范畴。这三大核心范畴的确定，旨在捕捉影响可持续信任的关键变量，是对前期编码工作的进一步提炼和总结，最终编码结果如表 1 所示。由此回应了本文研究问题 1：影响人机可持续信任的因素包括用户、任务、机器三大维度，具体又涉及需求适配度、任务复杂度、技术安全性等 13 项细分指标。

表 2　编码结果[3]

原始文本示例	次要范畴	主要范畴	核心范畴
Despite the hype around ChatGPT, it turned out to be practically useless for my needs.	基本没用	需求适配度	用户因素
It is a satanic being, a non human life form with this much smarts is devil!	认为技术过于强大需要被控制	技术恐惧	
There didn't seem to be any option to choose plugins or browsing on the app	使用门槛和操作障碍	媒介素养	
Disappointing, it just gives the most pointless responses, essentially telling you to ask a professional.	输出结果没有达到预期效果	心理预期	
Always struggles with complex queries, failing to deliver satisfactory solutions	无法响应较难任务	任务复杂度	任务因素
I could be talking about one thing and it would go to a completely different thing.	无法准确率理解用户需求和意图		
I use this app to edit my novel I'm writing. After the recent update, the usual prompt I give it has been giving me trouble.	在关键任务中表现失败	关键性失误（Critical Misstep）	
It only responds back with its biased information than information that challenges its own assertions.	道德伦理和意识形态问题	违规成本	

续表

原始文本示例	次要范畴	主要范畴	核心范畴
Even if I spend money, I still have to endure " You have sent too many messages to the model.	限制交互频次	可供性	机器因素
I have tried to log in through various means but it has proved futile.	使用权限与注册登录问题		
My account was deactivated for no reason and no one told me how to resolve this matter.	服务稳定性问题		
Gives wrong information at times.	生成内容质量	可用性	
Anytime you ask a question it says it "cutoff knowledge" is only up to September 2021.	信息时效性问题		
Problematic scrolling issue when history is turned off which has gone unaddressed for several updates now.	产品功能存在缺陷		
It' s hard to scroll down without it trying to scroll left to pop open the left menu.	功能易用性	易用性	
Overall interface is clean, but on the iPad in landscape orientation it' s basically a phone sized cutout.	不适应或不满意交互设计		
Several dollars for such a few rows of stupid and useless words.	价格太贵或扣费不合理	经济性	
My phone heats up like crazy when using this app.	性能问题和手机发烫	兼容性	
Could you please make the app compatible with older versions?	兼容性问题		
Requires your phone number for use, collects your user data, and is coded by its developers.	隐私数据担忧	安全性	

三、研究发现

(一) 机器因素：从可供可用到可信可控

相比于用户因素（16.43%）与任务因素（4.08%），机器因素（79.49%）成为影响用户对AI可持续信任的核心变量。也就是说，不同于冷启动阶段主观感知因素的主导性影响，在持续性使用过程中客观技术因素始终是关键因子。在诸多技术性因素中（图1），AI的可用性成为当下占比最大的范畴。尤其是AI幻觉、数据时效性、内容准确性上的技术瓶颈，成为人机可持续交互的最大桎梏。其中，内容可信度也成为当下AIGC的核心诟病。功能易用性及交互界面友好度也对用户使用体验带来了较大影响，这在一定程度上受到了用户对既有产品（如搜索引擎、社交媒体等）功能界面的技术固着性影响。当用户对既有技术工具的产生使用惯性与模式依赖，对交互式AI的对话模式、信息筛选模式、提示语结构会形成一定不适应，这种多平台之间的模式转换和迁移成本也会影响用户感知易用性。相较之下，在生成式AI应用门槛逐步降低的当下，关于可供性的负面感知逐渐降低，算力资源对使用频次所构成的限制也成为部分高粘性用户持续性使用的硬性阻碍。此外，虽然用户隐私、数据安全、意识

形态安全问题作为技术底线成为媒体和学界所关注焦点，但对于用户持续性使用而言并未构成重要影响（<5%）。从 AI 的可供到可用，也对应着创新扩散从早期采用者逐步渗透到早期大众，而要从早期大众扩散到稳定大众使用，则还需要进一步在可信和可控性上持续优化。

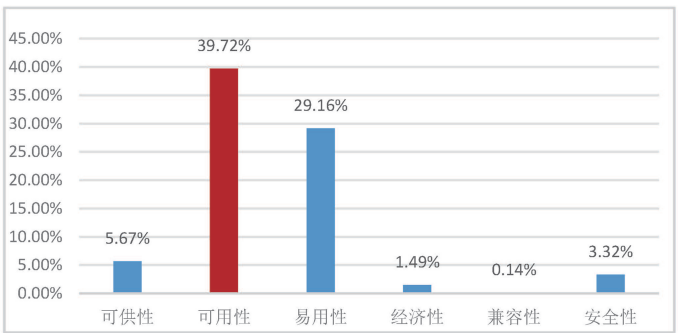

图 1　机器因素编码占比

根据信任阈值理论（Trust Threshold Theory），人机之间的持续性信任建立在特定阈值之上。当机器某一行为或事件超过了该性能阈值，信任就会迅速崩塌。虽然安全性指标在人机交互过程中显著性较低，较少被用户提及，但研究发现（图 2），安全因子对用户可持续使用行为的影响程度大于其他因子，也即信任阈值更低。一旦在人机交互过程中出现了涉及隐私安全和意识形态安全的问题，用户对 AI 的信任度会迅速降低，对后续交互会产生持续性影响。相比之下，虽然可用性、易用性是当下阻碍用户持续信任的主导性因素，但用户对其容错度更高，信任阈值也更高。此外，与终端和其他产品的兼容性虽然也影响了用户的持续性信任，但对用户的使用行为一般不会造成实质性影响。结合现实场景来看，安全性问题可能会导致用户的永久性“弃用”，而技术性问题则更多会形成阶段性“失信”。对低阈值因素的信任修复需要花费更多成本和时间，加强安全类因素的脆弱性和鲁棒性研究是维系人机长期信任的关键。

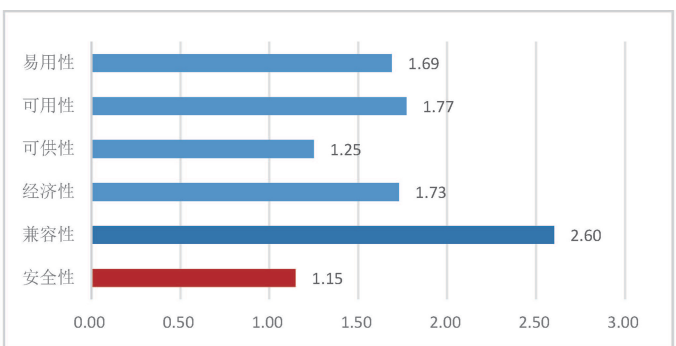

图 2　机器因素影响程度（问题维度 VS 评价得分）

（二）用户因素：从感性技术恐惧到理性需求适配

在初始信任形成过程中，用户对产品的认知多源于他者的间接评价和品牌背书，用户主观感知因素占据主导地位。包括个体既有偏见、媒体负面宣传、低需求动力和信息不对称等都会降低个体对 AI 的初始信任度。而在可持续信任形成过程中，用户既有主观感知会被实际使用经验放大或逆转。尤其是当用户本身对 AI 存在恐惧或抵触等负面情绪时，在确认偏误影响机制下，其会倾向于寻找、解释并记录可以确认自己的先入之见的信息，而忽略或低估与之相反的证据。[41] 研究发现（图 3），技术恐惧成为影响人机可持续信任的核心主观因素。这种感性层面的技术恐惧通常源于对 AI 复杂性、不可预测

性以及潜在影响的不确定感。作为一种基于大量数据和复杂算法运作的系统，生成式 AI 的决策过程和内部逻辑对于普通用户而言，往往是“黑盒子”。这种不透明性可能引发对其可能产生的未知后果的担忧，包括 AI 对人类工作的替代性担忧也加剧了部分用户的技术恐惧。而要减轻这种主观层面的恐惧，增强系统的透明度和可解释性并确保以人类福祉为核心导向的技术发展观都必不可少。

除了主观认知性因素，客观层面的需求匹配度也是影响可持续信任的重要因素。当产品不能满足用户核心业务需求，或者不能有效解决用户切实问题时，低需求动力和低适配度均会降低用户的使用频次和粘性，进一步弱化主观期待和可持续性信任。此外，用户媒介素养也会限制可持续信任的生成，包括操作层面的障碍和使用熟练度都会降低用户使用体验，影响人机信任维系。同时，过高的心理预期则会对用户实际使用体验带来负面影响。预期确认理论（Expectation Confirmation Theory）表明，用户的满意度是由他们的预期与实际使用体验之间的差异决定的。如果用户对 AI 技术抱有过高的期望，而实际使用体验未能达到这些期望，这种不匹配可能导致信任的削弱。

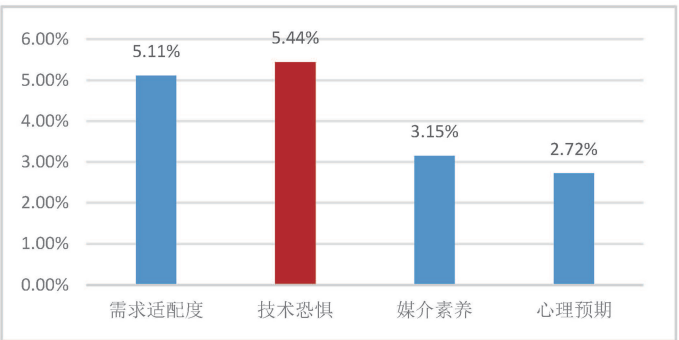

图 3　用户因素编码占比

进一步对不同因素的影响程度进行分析（图 4），可见需求适配度对用户信任及持续使用的影响程度最大。在人机交互过程中实用主义导向仍占据主导，当 AI 能高度契合现实需求场景时，依赖度和信任度都会显著增高。其次为心理预期，一方面过高的预期往往会强化负面感知，另一方面当 AI 表现超出预期时也能强化正向评价与后续信任度。相比之下，用户媒介素养和技术恐惧虽然占比较高，但在影响产品评价和持续使用行为上的作用有限。综合来看，用户的需求场景、心理预期、媒介素养伴随着人机交互频次的深化而变化。且相比于机器因素，用户层面的变动更具可控性，这也使得可持续信任成为一个可形塑、可引导的动态过程。在技术变量达到阶段性瓶颈时，用户心理感知、认知态度和使用行为层面的干预和引导就成为提升人机协作效率的关键抓手。

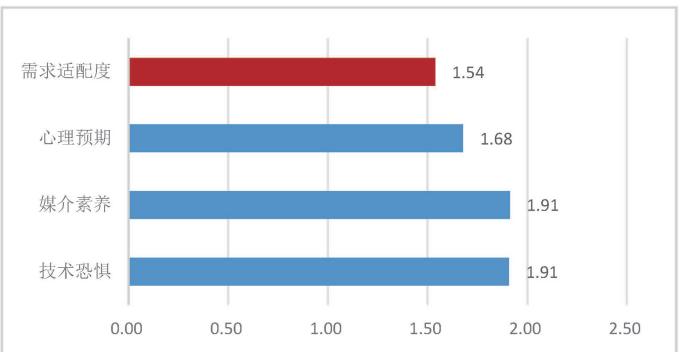

图 4　用户因素影响程度（问题维度 VS 评价得分）

（三）任务因素：从软性情境边界到硬性合规约束

除了机器和用户因素，任务作为人机交互的内置目标与外部情境，更多地关注用户与机器之间具体的交互内容，其复杂度、意图和理解的清晰度、任务失败的机会成本、潜在损失等都会直接影响可持续信任程度。任务因素可被视作一种外部调节变量，可动态干预用户对机器的信任程度。例如，当面对低复杂度、低风险、低试错成本的常规性质任务时，用户往往会展现出较高的信任度；相反，在面对那些要求精确控制、专业知识和高度判断力的复杂任务时，用户的信任水平可能会显著降低。研究发现（图 5），在任务因素中关键性失误出现频次最高，违规成本其次，而任务复杂度被提及频次相对较低。尤其是在关键性任务中技术失误所带来的后果超出用户承受阈值时，用户对机器的有效性普遍会持保留态度。一次关键任务失败所带来的信任损失，往往需要众多成功任务经验来弥补，甚至会带来永久性的“失信”。当然，任务因素的调节效应也与用户的风险承受能力和容错度直接相关。在技术创新扩散初期，包容性试错是提升技术接受度的关键。而随着用户对新技术的使用经验和累积信任的增加，其会动态调整自己的期望偏差，抗风险能力和对复杂任务的尝试意愿一般也会提升。[42]

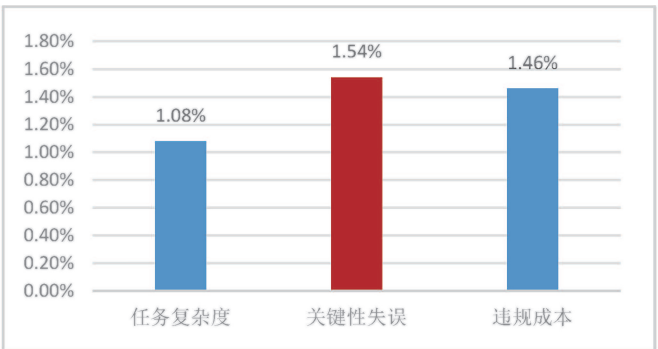

图 5　任务因素编码占比

用户、机器、任务三大影响因素共同构成了“人”“机”“交互”中可持续信任机制。虽然任务因素当下被提及频次较低，但其对用户负面感知的作用（M = 1. 40）却强于机器特性（M = 1. 67）和用户因素（M = 1. 76）。也即，交互任务所带来的负面体验和现实损失，对人机持续信任会带来更大影响。部分关键任务失败甚至可能导致用户较长时期内对技术的回避行为或过度谨慎。而在任务因素中（图 6），涉道德伦理和意识形态问题上的违规成本成为用户的重要考量，其对信任水平的影响略高于任务复杂度和关键性失误。也即一旦机器在特定任务上可能出现有违“政治正确”的表现，用户对其信任程度和使用意愿会明显降低。在强化意图理解和任务拆解的基础上探寻人机价值对齐，也是从人机高频交互到人机有效协作转化的动力。

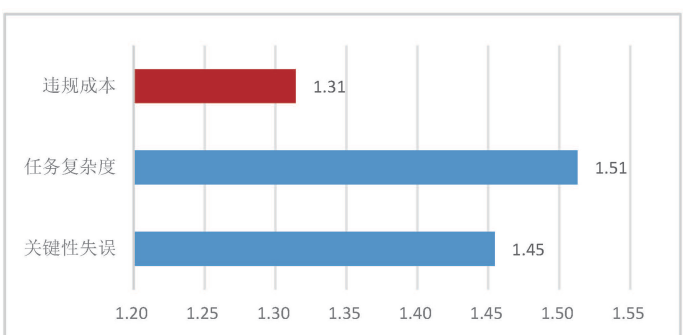

图 6　任务因素影响程度（问题维度 VS 评价得分）

综上，回应了本文研究问题2：在人机可持续信任的影响因素中，机器因素占比最大，其次为用户要素，任务因素占比偏低。而在对用户持续使用行为的影响程度上，关键任务失败、机器安全性、用户需求适配度影响最为显著。

四、研究结论与启示

信任是一种涉及期望、风险和不确定性的复杂社会心理现象，而人机信任更是凸显了人类与机器交互与协作的内在驱动力。作为主观感知变量，对信任的测量与研究通常依赖于量表设定，对相关因素的陈列也受到研究者既有理论框架的局限。不同于传统定量和定性分析方法的既定框架，本研究基于计算扎根理论探讨人机交互中可持续信任的影响因素，综合开源数据挖掘和机器编码，将人机信任置于一种开放性框架中进行探讨。一方面，在研究样本的规模性和精准性上做了有效控制，通过对AIGC深度用户的评价与评分进行多维筛选，提升了扎根文本的信息价值密度。另一方面，通过开展系统性的内容分析和分层范畴提炼，实现了从开放数据到规范模型的理论建构。这种基于海量数据的计算扎根，可以被视作一种综合定量与定性要素的混合研究范式，对于多维理解人机可持续信任的形成和演变过程具备重要价值，也为未来人机交互系统的设计和优化提供实质性的参考和指导。

整体来看，在生成式人工智能的交互应用中，机器因素（可用性、易用性、可供性和安全性等）带来的负面感知相对明显，其次为用户要素（技术恐惧、需求适配、媒介素养、心理预期），而任务因素（关键性失误、任务复杂度、违规成本）的感知显著性相对最低。人、机与交互任务三方面的因素共同作用于信任水平，其中任务因素作为一种外部调节因素，对持续信任的影响相对最大。尤其是关键任务失败所带来的现实损失或机会成本会显著降低用户信任水平，甚至在后续交互使用中形成应激障碍。此外，机器安全性（会不会损害自身现实利益）和用户需求适配度（现实场景中是否用的上）都会对用户使用频次和持续信任产生较大影响。可见，虽然人机交互是一种跨越虚实的对话场景和协作模式，但其对信任的影响仍根植于用户的现实利益，现实增益所带来的信任增强和现实损失所带来的信任损耗均普遍存在。

适度信任是保障人机可持续协作的关键要素。面对根植于外部技术环境的“硬性”因素和源自于内部主观感知的“软性”因素，有效调节信任水平、规避滥用和弃用，既需要从宏观层面创建包容审慎的技术发展环境，在推进人机价值对齐的过程中寻找技术的可持续发展之路，又需要从微观层面提升个体媒介素养和AGI时代的生存技能，探索并从人机交互到人机协同再到人机共生的进程。

首先，在交互任务层面，需要建立清晰的职责界限，确保人机之间存在清晰而明确的任务和责任分配，通过“互补式”协作规避机器局限性和关键任务风险。一方面，对于不同类型任务需要给出对应风险提示，并对用户的风险承受能力进行量化分级，从而实现精细化的风险管理和选择性响应。当然在对高违规成本采取提示和回避策略的同时，也需要避免“一刀切”式的粗放化管理，平衡任务的响应率和风险控制。另一方面，在面向专业领域的人机协作任务中，可引入透明化的运作流程和决策机制，除了输出最终结果，提供任务中间过程要素（包括代码实现逻辑、计算流程、推理过程等），为用户提供任务核查和排错的渠道，通过提升协作过程的信任度来增强任务结果的可控性。

其次，在机器性能层面，限于生成式人工智能的概率性输出机制，很难完全规避随机性所带来的幻觉问题，如何用可解释性对冲不确定性，在人机价值对齐中实现信任动态校准，还需要在底层算法和交互应用层面做多重优化。除了在产品可供性、内容可用性、功能易用性上不断迭代，补足技术短板、优化用户体验，提升机器的“实际可信度”。在既有技术条件约束下，也可通过运作机制优化来提升机

器的“感知可信度”。一方面，在传播层面提升算法的透明度和可解释性，让用户认知系统的运作方式和决策过程，可在一定程度上校准过高期待及过度信任。另一方面，根据失误恢复理论，当系统在发生错误后能迅速并有效地恢复时，用户的信任可能不仅不会下降，反而可能因为对恢复能力的认可而提高。错误发现和修复的及时性也是信任校准的关键因素，实证研究发现，机器人在发生错误后立即道歉比完成任务后再道歉跟更有利于信任修复。[43] 特别是对于“黑箱”计算的人工智能模型而言，当算法可解释性和生成随机性很难受控时，及时的错误发现、校准响应与迭代输出对于可持续信任维系更为重要。

最后，在用户操作层面，除了长期视角下的媒介素养提升和使用经验积累，就短期内的优化而言，通过在交互场景中引导用户动态调整心理预期、技术恐惧并找到供需连接点，也有助于可持续信任的平衡与校准。对于生成式人工智能而言，基于用户反馈的强化学习机制——RLHF（Reinforcement Learning from Human Feedback）使得 AI 与人类在预期目标和价值观点上趋于对齐，在具体交互场景下的功能性缺陷也能被较好的纠偏与“驯化”。但多数情况下这种奖惩机制需要由用户主动反馈，对于能动性相对较低的用户未能充分发挥效应，这就使得奖惩机制的“前台化”“界面化”展示成为必要。也即让机器引导用户进行反馈，进一步根据用户正向或负向反馈进行后续交互的优化。目前部分 AIGC 产品已引入相关机制，通过同时提供用户多个答案进行比较，来推测用户意图、优化响应策略。这种双向交互的反馈机制可降低信息熵和不确定性，帮助用户调整心理预期和增强能动性。

诚然，可持续信任是一个包含多维变量的动态过程，其生成、维系与演变均涉及各方面复杂因子的影响。本文虽力图对其影响因素进行扎根分析与实证测量，但仍有部分理论局限与技术操作缺陷需进一步完善。首先，从理论层面，可持续信任的阶段与生成机制需进一步细化剖析。从冷启动到高频交互再到稳定协作，在人机交互的不同阶段可持续信任的影响因素差异分化，不同因子之间的相互作用与重要性也在不断演变。限于操作复杂性，本文未对信任的演变过程进行划分与追踪，理论框架的适配性难免存在缺漏。其次，从测量方法来看，本文针对 AIGC 用户的评论数据进行扎根编码，将“低评价分数”等同于“低信任水平”，两者在一定程度上存在关联映射，但后续研究中对信任水平的操作定义仍需精细化。最后，由于扎根样本均源自于网络渠道和公开表达，忽略了其他“沉默用户”的信任水平及其影响因素，高表达意愿的用户群体其信任生成机制与其他群体是否存在差异？这也有待后续测量样本的扩充与论证，以提升结论的代表性。除此之外，当下 AIGC 产品迭代与功能优化不断加速，影响用户信任的因素也不断在发生变化，本文分析样本中的部分因素（如时效性、服务稳定性等）在成文时均得到了优化。伴随着技术的普及扩散，用户的认知、需求和持续使用行为也在转变，这也为后续阶段性、跟踪性、对比性研究提供了方向。

参考文献：

[1] Baier, A. (1986). Trust and antitrust. *Ethics*, 96 (2): 231-260.

[2] Taddeo, M. (2011). Defining trust and e-trust. *International Journal of Technology and Human Interaction*, 5 (2): 23-35.

[3] Gambetta, D. (2000). Can we trust trust. *Trust: Making and Breaking Cooperative Relations*, 13 (2000): 213-237.

[4] Bedué, P. & Fritzsche, A. (2022). Can we trust AI? An empirical investigation of trust requirements and guide to successful AI adoption. *Journal of Enterprise Information Management*, 35 (2): 530-549.

[5] Ryan , M. (2020) In AI we trust: Ethics, artificial intelligence, and reliability. *Science and Engineering Ethics*, 26 (5): 2749-2767.

[6] Hurlburt , G. (2017). How much to trust artificial intelligence? *It Professional*, 19 (4): 7-11.

[7] Glikson , E. & Woolley , A . W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Manage-*

ment Annals, 14 (2): 627-660.

[8] Siau, K. & Wang, W. (2018). Building trust in artificial intelligence, machine learning, and robotics. *Cutter Business Technology Gournal*, 31 (2): 47-53.

[9] Gao, L. & Waechter, K. A. (2017) Examining the role of initial trust in user adoption of mobile payment services: An empirical investigation. *Information Systems Frontiers*, 19: 525-548.

[10] Hoehle, H., Huff, S. & Goode, S. (2012). The role of continuous trust in information systems continuance. *Journal of Computer Information Systems*, 52 (4): 1-9.

[11] Okamura, K. (2020). *Adaptive trust calibration in human-AI cooperation*, Ph. D. Dissertation. Kanagawa: The Graduate University of Advanced Studies.

[12] 朱翼. 行为科学视角下人机信任的影响因素初探 [J]. 国防科技, 2021 (4): 4-9.

[13] Cabiddu, F., Moi, L., Patriotta, G. & Allen, D. G. (2022). Why do users trust algorithms? A review and conceptualization of initial trust and trust over time. *European Management Journal*, 40 (5), 685-706.

[14] Luhmann, N. (1982). Trust and power. *Studies in Soviet Thought*. 23 (3): 266-270.

[15] Luhmann, N. (1979). *Trust and power*. Chichester: John Wiley.

[16] Durante, M. (2010). What is the model of trust for multi-agent systems? Whether or not e-trust applies to autonomous agents. *Knowledge, Technology & Policy*, 23: 347-366.

[17] Hoff, K. A. & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors*, 57 (3): 407-434.

[18] von Eschenbach, W. J. (2021) Transparency and the black box problem: Why we do not trust AI. *Philosophy & Technology*, 34 (4): 1607-1622.

[19] Gursoy, D., Chi, O. H., Lu, L., et al. (2019). Consumers acceptance of artificially intelligent (AI) device use in service delivery. *International Journal of Information Management*, 49: 157-169.

[20] Troshani, I., Rao, H. S., Sherman, C., et al. Do we trust in AI? Role of anthropomorphism and intelligence. *Journal of Computer Information Systems*, 61 (5): 481-491.

[21] Glikson E. & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14 (2): 627-660.

[22] Bogert, E., Schecter, A. & Watson, R. T. (2021). Humans rely more on algorithms than social influence as a task becomes more difficult. *Scientific Reports*, 11 (1): 1-9.

[23] Taddeo, M. (2009). Defining trust and e-trust: From old theories to new problems. *International Journal of Technology and Human Interaction* (*IJTHI*), 5 (2): 23-35.

[24] Nissenbaum, H. (2001). Securing trust online: Wisdom or oxymoron? *Boston University Law Review*, 81 (3): 635-664.

[25] 何江新, 张萍萍. 从“算法信任”到“人机信任”路径研究 [J]. 自然辩证法研究, 2020 (11): 81-85.

[26] Siau, K. & Wang, W. (2018). Building trust in artificial intelligence, machine learning, and robotics. *Cutter Business Technology Gournal*, 31 (2): 47-53.

[27] Dorton, S. L. & Harper, S. B. (2022). A naturalistic investigation of trust, AI, and intelligence work. *Journal of Cognitive Engineering and Decision Making*, 16 (4): 222-236.

[28] Sheridan, T. B. (2019). Individual differences in attributes of trust in automation: Measurement and application to system design. *Frontiers in Psychology*, 10: 1117.

[29] Kim, J., Giroux, M. & Lee, J. C. (2021). When do you trust AI? The effect of number presentation detail on consumer trust and acceptance of AI recommendations. *Psychology & Marketing*, 38 (7): 1140-1155.

[30] 杨子莹. 人机交互关系中的信任问题研究 [D]. 大连理工大学, 2022.

[31] 董文莉, 方卫宁. 自动化信任的研究综述与展望 [J]. 自动化学报, 2021 (6): 1183-1200.

[32] Hoffman, R. R. (2017). A taxonomy of emergent trusting in the human - machine relationship. In Philip, J. & Robert, R.

(eds.). *Cognitive systems engineering*: *The future for a changing world*. Leiden: CRC Press, 137-164.

[33] Okamura, K. & Yamada, S. (2020). Adaptive trust calibration for human-AI collaboration. *Plos One*, 15 (2): e0229132.

[34] Jacovi, A., Marasović, A., Miller, T., et al. (2021). Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the* 2021 *ACM conference on fairness*, *accountability*, *and transparency*. New York: Association for Computing Machinery, 624-635.

[35] 齐佳音，张亚．人—机器人信任修复与信任再校准研究［J］．机器人产业，2021（4）：26-38.

[36] 何贵兵，陈诚，何泽桐等．智能组织中的人机协同决策：基于人机内部兼容性的研究探索［J］．心理科学进展，2022（12）：2619-2627.

[37] 于雪．基于机器能动性的人机交互信任建构［J］．自然辩证法研究，2022（10）：43-49.

[38] Ezer, N., Bruni, S., Cai, Y., et al. (2019). Trust engineering for human-AI teams. In *Proceedings of the human factors and ergonomics society annual meeting*. Los Angeles: Sage Publications, 322-326.

[39] Berente, N. & Seidel, S. Big data & inductive theory development: Towards computational Grounded Theory? Retrieved March 5, 2024, from https://core.ac.uk/reader/301361940.

[40] Glaser, B. & Strauss, A. (1967). Grounded theory: The discovery of grounded theory. *Sociology the Journal of the British Sociological Association*, 12 (1): 27-49.

[41] Çalikli, G. & Bener, A. (2013). Influence of confirmation biases of developers on software quality: An empirical study. *Software Quality Journal*, 21: 377-416.

[42] Hu, W.-L., Akash, K., Reid, T. & Jain, N. (2019). Computational modeling of the dynamics of humantrust during human - Machine interactions. *IEEE Transactions on Human-Machine Systems*, 49: 485-497.

[43] Robinette, P., Howard, A. M. & Wagner, A. R. (2015). Timing is key for robot trust repair. In *Proceedings of* 7*th international conference on social robotics*. Paris: Springer International Publishing, 574-583.

［责任编辑：高辛凡］