arXiv:2603.01457v1 [cs.HC] 2 Mar 2026

权力回响：探究在线权力不对称冲突中的审核偏见

Yaqiong Li
liyq22@m.fudan.edu.cn 复旦大学，中国上海

Peng Zhang*
zhangpeng_@fudan.edu.cn 复旦大学，中国上海

Peixu Hou*
peixu.hou@meituan.com 美团，中国上海

Kainan Tu
kntu25@m.fudan.edu.cn 复旦大学，中国上海

Guangping Zhang
gpzhang20@fudan.edu.cn 复旦大学，中国上海

Shan Qu
qushan@meituan.com 美团 中国上海

Wenshi Chen
wenshi.chen@meituan.com 美团 中国上海

Yan Chen
ych@vt.edu
弗吉尼亚理工大学
美国弗吉尼亚州

Ning Gu
ninggu@fudan.edu.cn
复旦大学
上海，中国

Tun Lu*
lutun@fudan.edu.cn
复旦大学 上海，中国

摘要

在线权力不对称冲突十分普遍，目前大多数平台依赖人工审核员开展内容审核。以往研究持续关注不同场景下的人工审核偏见，而权力不对称冲突中的审核偏见仍有待探索。因此，我们旨在探究人工审核员在权力不对称冲突审核中表现出哪些类型的权力相关偏见（研究问题1），并进一步探讨人工智能建议对这些偏见的影响力（RQ2）。为此，我们以消费者与商家之间的真实冲突为场景，开展了一项包含50名参与者的混合设计实验。结果表明，在这两种审核模式中存在若干偏向支持强势方的偏见。人工智能辅助缓解了人工审核中的大多数偏见，但也放大了少数偏见。基于这些结果，我们为未来关于权力不对称冲突的人工审核及人机协作审核系统研究提出了若干见解。

CCS概念

• **以人为中心的计算 → 协作与社会计算**.

关键词

权力不对称冲突、内容审核、偏见、人机协作、绿野仙踪法

1 引言

Twitter 和 Yelp 等各种数字平台一直是支持社会互动和活动的重要媒介。随着这些平台的用户群不断扩大，不同角色之间的冲突日益普遍，给平台监管带来了挑战。例如，在 Yelp 等电子商务平台上，消费者与商家因产品质量或服务问题发生冲突的情况屡见不鲜。根据《全球数字信任与安全指数》 [48]，2023年第三季度，全球报告了数百万起此类冲突。在中国，消费者协会报告称，2024年上半年消费者投诉超过78万起，同比增长27.21% [116]。Amazon Mechanical Turk 和 Upwork 等数字劳动平台上也存在类似冲突，劳动者与需求方经常就补偿和任务质量等问题产生分歧 [4, 18, 77]。

不同角色之间的这些冲突通常具有一个显著的结构性特征，即往往发生在权力不对称的关系中 [3, 65, 88]。一方通常拥有更多专业知识和资源（例如商家或雇主），而另一方则缺乏冲突解决经验，并严重依赖平台协调（例如消费者或劳动者）。以网购场景为例，当消费者投诉商品质量并给出负面评价时，商家可能会援引平台的“恶意评价处理规则”提出申诉，并提供细致解释[40]。相比之下，消费者的谈判技巧、规则解读能力和申诉渠道通常有限，且

*通讯作者。

他们难以提供更多证据，导致其评论最终被判定为不当。

许多平台已引入审核策略，以处理权力不对称冲突。不同于传统的内容审核（如有害内容审核 [72, 73] 和隐私审核 [74], ），权力不对称冲突调解需要在权力不对称以及双方陈述含糊不清的背景下，判断“谁更可信、谁应获得支持”，从而加剧了审核难度。因此，目前大多数平台依靠人类志愿者或众包审核员开展权力不对称冲突调解。然而，现有研究表明，在涉及主观判断的冲突情境中，人工审核员往往会表现出不同的审核偏见，例如在立场或语言风格方面存在偏见。例如，Roberts 等人通过对社会内容审核开展实地研究发现，人工审核员难以始终保持中立，因为他们的观点往往受到政策模糊性、任务压力或个人信念的影响。该研究还指出，这些因素可能导致审核员对某些群体产生偏见性判断，尤其是那些语言表达相对直接、通俗，或缺乏策略性措辞的用户——这类用户通常被称为低语境表达者 [30]。此外，Seering 等人报告称，Reddit 版主在审核用户冲突时，往往会强化既有的权力结构和主导性的社区价值观，导致边缘群体的声音被忽视或删除 [57]。此外，心理学中的“弱者效应”表明，在冲突判断中，人工审核员可能会同情弱势方并偏袒他们；但当弱势方的陈述含糊不清时，他们也可能只是默认将责任归咎于弱势方 [31]。这些发现表明，在权力不对称冲突调解实践中，人工审核员的判断不仅会受到个人立场和偏好的影响，还可能进一步受到强势方的社会权力线索（例如语言风格、规则引用和表达自信度）的放大，从而影响调解过程的公正性。因此，我们提出第一个研究问题（**研究问题1**）：**人工审核员在调解权力不对称冲突时会表现出哪些权力相关偏见?**

随着机器学习和深度学习技术驱动的人工智能工具不断发展，尤其是大语言模型（LLM），AI辅助审核在不同平台上日益普及 [38]。这些工具被广泛用于生成解释、提供决策建议以及判定风险等级，显著提高了内容审核的效率 [72, 82, 100]。然而，现有研究表明，人工审核员接受人工智能生成的建议的程度，也受到人们对该建议是否由人工智能生成的认知影响 [6, 45, 93]。这揭示了算法服从效应：用户面对“这是 *GPT-4* 大模型给出的判断结果。”时可能会表现出不同态度。因此，我们进一步提出第二个研究问题（**RQ2**）：**如何人工智能生成的建议会如何影响人工审核员在权力不对称冲突审核中的权力相关偏见？这些偏见会被放大还是减轻?**

为回答这两个研究问题，我们聚焦于两种判断模式，即人工审核和人机协同审核，并系统探究权力不对称冲突调解中相应的偏见。基于社会权力基础理论所定义的六种权力类型 [1, 28],，我们以消费者与商家之间的冲突作为具有代表性的权力不对称情境，并从大众点评平台（中国领先的本地生活平台）收集真实数据，以调查审核偏见。通过编码，我们提取出与在线冲突情境相匹配的权力表现形式，并进一步提出权力不对称冲突调解中潜在权力相关偏见的分类体系，例如正当主张偏见、惩罚威胁偏见和补偿偏见。随后，我们采用包含被试间和被试内设计的混合设计，并开发了一个名为“I Support”的基于网页的判断程序来开展实验。我们共招募了50名参与者，并将他们随机分配到人工审核组或人机审核组。对于**研究问题1**，人工审核组的参与者独立阅读权力不对称冲突样本，不接收外部建议。每个冲突样本都包含一种特定的权力表现形式（例如正当主张或补偿），参与者需使用5点李克特量表判断自己愿意支持哪一方。在本研究中，“支持一方”是指版主认为该方的陈述更合理或更可信。通过比较参与者在不同情境下的判断，我们可以识别哪些权力表现形式更容易引发偏见判断，从而揭示人工审核中固有的偏见。对于**RQ2**，针对相同的冲突样本，人机审核组的参与者还会获得额外的人工智能生成的建议。为控制不同大语言模型输出可能存在的差异，我们采用绿野仙踪设计。这种方法使我们能够考察参与者对人工智能生成的建议的看法如何影响其审核决策。具体而言，我们预先编写基于真实审核员众包数据的高质量建议，并告知参与者这些建议由“人工智能”生成。通过比较两组参与者在不同权力表现形式下的判断，我们揭示人工智能辅助是否会加剧或缓解不同偏见。定量与定性分析得出了若干发现。对于研究问题1，人工审核表现出五种倾向于支持强势方的权力相关偏见。对于研究问题2，这些偏见在人机协同审核中依然存在。尽管人工智能辅助缓解了人工审核中的大多数偏见（四种得到缓解，一种被消除），但也加剧了少数偏见（一种新出现，另一种有所加剧）。此外，人工智能生成的建议中不支持对方的视角能够促使审核员支持弱势方。这些发现为改进审核过程提供了实证依据

并有助于缓解审核偏见。总而言之，本研究作出了以下贡献。

- 据我们所知，这是首项研究权力不对称冲突中人工审核和人机协同审核偏见的研究。
- 基于真实数据，我们构建了一套理论驱动的分类体系，总结了权力不对称冲突审核中可能存在的权力相关偏见及其对应的权力表现形式。通过一项有50名参与者参与的实验，我们揭示了人工审核中的偏见以及人工智能辅助的影响。
- 我们针对权力不对称冲突审核机制和人工智能辅助审核系统的设计提出了若干启示。

本文其余部分安排如下：第2节回顾相关工作。第3节介绍权力表现形式。第4节和第5节介绍实验方法与结果。第6节和第7节讨论启示与局限性。最后，第8节总结全文。

2 相关工作

本节回顾权力不对称的理论基础及其在数字平台上的表现形式，并借鉴人机交互领域关于冲突和审核偏差的研究。

2.1 权力不对称冲突

权力不对称（或权力不平等）指的是“一种状态，在这种状态下，个人与群体之间存在地位差异，而这些差异导致他们采取行动或促使他人采取行动的能力有所不同”[27, 88]。在社会心理学和组织研究中，权力不对称体现为资源和决策权方面的差异。关于社会关系和资源依赖的早期研究为理解权力奠定了基础。Dahl 提出了权力作为人与人之间关系的概念，指出一方影响另一方行为的程度[15]。Emerson权力依赖模型进一步指出，权力并非个人属性，而是相互依赖的函数。如果一方依赖另一方获取关键资源，那么被依赖的一方就拥有更大的权力。这种依赖关系由资源的重要性和可替代性共同决定 [23]。为系统地描述权力，研究人员还尝试对社会互动中的权力来源进行分类。“社会权力基础”理论得到广泛认可，并将权力来源分为六种类型：合法权力（职位或角色赋予的权威）、奖赏权力（通过提供奖励影响他人行为的能力）、强制权力（通过施加惩罚影响他人行为的能力）、专家权力（基于专业知识或技能的影响力）、参照权力（因尊重或钦佩他人而产生的追随效应），以及信息权力（影响他人行为的能力这种能力通过控制、传播或操纵信息来实现） [1, 27, 28]。该理论为理解不同情境下的权力不对称提供了一个通用框架。

随着互联网平台的快速发展，人们的社交活动日益从线下转向线上。社会心理学和人机交互领域的研究也开始将权力理论延伸至数字情境 [42, 108, 115]。既有研究表明，数字技术通过界面设计、个性化推荐、内容排序和过滤等方式，改变了用户获取和处理信息的方式 [64, 85, 95], , 进而影响用户能看到什么、能看到多少，以及按什么顺序看到。这些变化可能导致信息分配失衡，重构甚至加剧用户之间的资源差距，并在不同的在线场景中催生新的权力不对称形式 [64, 76]。在在线社区中，新用户通常依赖老用户提供的内容和指导；因此，老用户凭借其知识和专业能力，拥有较强的专业权力和信息权力 [94]。在社交媒体上，名人的帖子通常因其社会吸引力而获得更高的曝光度，而普通用户的内容则难以获得相当的人气 [12, 112]。这些资源差距可能限制互动中弱势当事方的行动空间，甚至导致对抗性行为或冲突行为，即权力不对称冲突。尽管权力不对称研究源自线下互动场景，但在线环境中的权力不对称冲突，本质上仍源于用户之间在信息、影响力及其他资源方面的不平等。不过，数字环境引入了更多资源（例如帖子、关注者、账号声誉），这些资源都可能增加资源不对称的发生概率。因此，将现有权力理论应用于数字情境时，可能需要扩展其所反映的表现形式。

与线下场景相比，资源结构发生了显著变化的电子商务和数字劳动平台，是容易发生权力不对称冲突的典型数字情境 [17, 47, 77, 108]。在电子商务平台上，商家往往掌控着产品信息、定价策略和售后服务等关键资源 [40, 87], , 这使他们在交易中处于信息优势地位。相反，消费者高度依赖商家通过平台提供的信息，由此造成信息不对称。这种失衡意味着，一旦出现问题，消费者通常难以找到有效渠道来维护自身权利。受商家或平台规则的限制，他们难以表达自身需求，因此会采取公开评价等替代策略来维护权利 [17]。数字劳动平台上也能观察到类似现象 [18, 77]。劳动者通常在信息获取和资源方面处于劣势。由于缺乏足够的谈判技巧，他们容易接受低工资或不平等的任务要求。相比之下，需求方可以凭借对平台规则的熟悉程度、发布任务的权力以及

对评分系统的影响力获得更强的控制权。因此，在逾期付款或任务相关争议中，劳动者通常难以寻求补偿。他们还面临评分降低甚至账号停用等潜在风险 [77]。不同当事方之间的这种权力不对称不仅会影响个人权利，还可能引发在线冲突。此外，冲突解决流程缺乏透明度或存在偏见等问题，可能进一步加剧当事方之间的信任危机 [85]。

2.2 人机交互中的冲突分析与解决研究

冲突分析与解决是人机交互领域的重要研究课题 [21], , 涵盖团队协作 [11, 43],、社会互动 [63, 71],、虚拟现实 [49, 102],和众包 [7]等场景。以往研究探讨了不同冲突的成因、冲突在不同环境中的具体表现形式，以及冲突对交互质量、用户体验和协作效率的影响。在临时设计团队中，目标不一致、创意方向分歧和任务分配不均往往会引发冲突 [11]。在社会互动中，研究发现，用户属性、对话历史和情绪波动可能与群体冲突的程度有关 [71]。在在线同行评审过程中，如果反馈只包含负面结论或情绪反应，却没有提出可执行的措施，就可能导致贡献者之间发生关系冲突，降低部分成员继续参与社区的意愿 [43]。对 Twitter 上冲突讨论的分析表明，冲突不仅会减少参与者数量，还会削弱用户的热情和协作效率 [9]。

人机交互研究者还在不同情境下进一步研究了冲突解决策略。根据实践者的不同，这些策略可分为三类：人工主导的解决方式 [49, 102], 人工智能主导的解决方式 [99], 以及人机协作 [33]。人工主导的解决方式依赖人工审核员的专业知识、经验和社会影响力。在虚拟游戏环境中，关怀他人的成年人和年龄相仿的导师在化解纠纷方面发挥着关键作用。他们通过提供解释性反馈、提醒参与者遵守规则，以及提出解决问题的策略来实现这一目标[49]。在团队协作方面，研究者提出了多种营造积极团队氛围的方法，例如引入“团队约会”以增进沟通 [14], 、鼓励沉默的成员参与讨论 [61], , 以及促进成员间平衡参与 [104]。对于人工智能主导的冲突解决，机器学习或大语言模型等技术被直接用于调解冲突，这是冲突审核领域一种新兴且前景广阔的方法。Omar 等人开发了一种名为*Rehearsal* 的交互式大语言模型系统，通过角色扮演和实时反馈，帮助用户练习处理人际冲突的技能 [99]。最后，人机协作式冲突解决结合了人类的理解能力与计算优势人工智能。在社交平台上的在线辩论中，研究者采用提示工程技术，将多种冲突解决策略（例如竞争、回避和妥协）融入大语言模型生成的回复中。该方法旨在帮助用户减少敌意语言并缓解两极分化 [33]。总体而言，这些研究加深了人们对冲突成因和解决机制的理解。然而，大多数研究都假定冲突各方地位平等，对权力不对称冲突的探索仍不充分。

2.3 内容审核中的公平性与偏见

内容审核被定义为“构建社区参与机制以促进合作并防止滥用的治理机制” [34]。内容审核实践往往需要在不同的行为规范、沟通风格和价值取向之间权衡，最终决策取决于具体情境和平台规则等多种因素 [53]。当前的内容审核主要有两种方式：人工审核和AI辅助审核。

作为主要方式，人工审核是指由人工审核员识别仇恨言论或骚扰等不当内容 [52, 60]。然而，先前研究表明，人工审核并非总是中立的。审核员可能受到个人特征、情境模糊性以及认知或工作负荷限制等因素的影响 [97, 103]。首先，身份相关因素会影响审核员对用户行为的解读。例如，审核员的决策可能受到人口统计特征、社会地位、政治立场、文化背景等个人特征的影响 [56]。例如，在 Twitter 上，审核员审核女性主播（从事直播的用户）的频率往往高于男性主播 [119]。政治保守派、跨性别用户和黑人用户发布的帖子更容易受到审核并被删除 [36]。其次，当审核员面对含义模糊的叙述或政策不确定性时，可能会产生感知偏见。实地研究表明，当政策含糊不清或任务压力较大时，审核员难以保持中立性，因而会对沟通方式更直接、使用方言或缺少策略性措辞的用户采取严格审核 [30, 31]。此外，工作负荷过重时也可能产生启发式偏见。随着在线社区的内容规模不断扩大，人工审核员往往面临大量工作和沉重负荷，从而影响注意力和判断力 [26, 111]。审核员可能过度关注某些类型的不当内容，同时对其他类型的内容敏感度下降。

算法及AI辅助审核系统中的偏见也日益受到关注。随着技术的快速发展，人工智能在内容审核中的应用日益普遍，尤其是在有毒或有害内容的检测方面 [24, 68, 98]。借助机器学习 [46], 深度学习 [74, 100], 和大语言模型 [10, 72, 82, 114], , AI辅助审核系统能够高效检测在线有毒或有害内容。一个显著的例子是

谷歌的Perspective API，它在Reddit和Faceit等平台上被广泛使用。该API基于大型语料库训练，能够识别有毒内容并给出毒性评分 [55]。然而，研究表明，其判断有时存在偏见。相同内容以德语呈现时，其毒性评分往往高于其他语言 [86]。这种差异凸显了一个更广泛的问题：在大型数据集上训练的模型会继承这些数据集中所蕴含的偏见。实证研究表明，大语言模型在判断任务中经常表现出系统性偏见，对输入长度和输入顺序等多种因素较为敏感 [10, 114]。此外，人机协同审核中还出现了一些新形式的偏见 [2, 91]。已有研究发现，人工审核员往往过度依赖算法建议，即所谓的算法服从效应；这会削弱独立的批判性思维，并使判断与人工智能的输出高度一致 [2]。在社区层面，成员可能会选择性地采纳人工智能生成的建议 [66, 96]。他们倾向于接受符合个人期望或群体认同的内容，同时质疑或拒绝与自身观点相矛盾的内容 [66]。

总体而言，冲突分析与解决是人机交互领域的重要研究课题，尤其涉及人机协同审核的优势与局限性。作为一种新型在线冲突，权力不对称冲突的特征包括用户地位差异、信息不平衡，以及日益增长的审核偏见趋势。然而，现有研究仍缺乏对这类权力不对称审核任务中产生的偏见进行系统性探究。为弥补这一空白，本研究考察人类审核中的潜在权力相关偏见，并进一步引入人工智能辅助，以探究其是否会缓解或加剧这些偏见。通过这一研究，我们旨在为权力不对称冲突中的审核策略和人机协同审核系统设计提供新的见解。

在线权力不对称冲突中的3种权力表现形式

本文旨在探讨审核员在权力不对称冲突调解中表现出的权力相关偏见。我们首先回顾社会互动中的权力类型，并根据现有理论和实证研究推导出相应的潜在偏差类别。具体而言，我们借鉴社会心理学家 French & Raven 提出的“社会权力基础”理论，该理论将社会权力划分为六类：合法权力、强制权力、奖赏权力、专家权力、参照权力和信息权力 [1, 28]。这一框架已广泛应用于线下人际互动、组织传播和谈判研究 [27, 89, 105], , 近年来也被用于解释在线互动中的权力动态 [42, 108, 115]。这些应用为描述权力关系和理解在线冲突中的权力不对称提供了理论视角。因此，本研究以“社会权力基础”理论为参照，开展实证分析，以在线消费者-商家冲突为研究场景，系统考察在线权力不对称冲突中权力类型的表现形式。这为我们后续的偏见分析奠定了基础。

数据收集。 我们选择大众点评平台[1] 作为研究平台，原因有三：1）大众点评是中国领先的本地生活服务平台之一，涵盖餐饮、酒店、娱乐和旅游等领域，拥有超过2.5亿用户。其功能与 Yelp 和 Tripadvisor 类似；2）大众点评评论区中消费者与商家之间存在大量争议和冲突，呈现出信息不对称、角色地位差异等权力不对称特征；3）该平台提供“大众评审”功能，这是一种典型的众包审核方法。该功能公开呈现冲突样本，允许不同用户进行投票，同时展示投票结果及详细理由，为我们的研究提供了丰富的语料。数据收集方面，我们获取了自2025年4月以来约60,000个消费者与商家冲突样本，涵盖17个主题。随后，我们按照数量选取排名前五的主题（占样本总数的92.7%），即“美食”“丽人”“娱乐”“运动”和“酒店”。对于每个主题，我们选取20个样本：其中10个样本的众包标签认为消费者评论合理，另外10个样本的众包标签认为商家回复合理。众包标签源自审核员通过平台“大众评审”功能作出的判断。每个样本包含消费者评论、商家回复、众包标签以及审核员给出的相应理由。最终，我们的数据集由100个样本组成。

数据分析与结果。在数据分析阶段，我们基于“社会权力基础”理论，考察在线冲突语料中六种权力类型的具体表现形式。我们采用定性分析方法对冲突语料进行编码和分类。具体而言，三位分别具有人机交互、社会心理学和语言学背景的作者合作，验证语料中六种权力类型的表现形式。例如，消费者评论中的“这是平台的规定”体现了合法权力，而商家的回应“如果您不满意，可以全额退款”则属于奖赏权力。通过这一过程，我们将样本映射到相应的权力类型。最终，所有冲突样本都能归入这六种权力类型，且未出现框架之外的新类型。这验证了该理论在在线权力不对称冲突中的适用性。

我们对冲突语料 [80, 110, 118] 进行定性分析，并细化其中六种权力类型的权力表现形式。三位作者首先独立采用开放式编码，对一个包含 25 个样本的试点子集进行编码（每个主题包含 5 个样本），且不预设子类别。在这一阶段，作者标记任何表明可能体现六种权力类型的文本特征。例如，

[1] https://www.dianping.com

表1：在线权力不对称冲突中的十种权力表现形式。

权力类型	权力类型描述	权力表现形式	权力表现描述
合法性权力	它源自民选、选定或任命的权威职位，并且可能以社会规范为基础或权威引用	正当主张	通过引用规则来巩固自身立场（例如，“这是平台的规则”）
		权威引用	通过提及外部机构来支持一方或来源（例如，“请访问 *http://...*”）
强制权力	通过威胁使用武力来获得他方的服从	惩罚威胁	通过威胁、投诉等迫使另一当事方让步，或造成负面后果（例如，“我要投诉你”）
奖赏权力	它基于提供或拒绝给予物质、社交、情感或精神奖励的权利	补偿	通过提供好处来争取支持（例如，“您可以获得全额退款”）
专家权力	它基于知识、经验，以及专业技能	专家知识	通过提供专业知识来支持一方知识（例如：“质量符合要求的标准”）
参照权力	它建立在尊重或钦佩之上从而促成认同或效仿	群体偏好	通过援引群体共识来支持一方（例如，“大家都说这项服务很差”）
信息权力	它基于对信息的控制被他人需要以达到重要的目标	陈述顺序	通过先发言或后发言来影响认知
		表达语气	通过情感化语言强化自身立场（例如，“太恶心了”）
		选择陷阱	通过比较凸显理性（例如，“其他商店提供更好的服务”）
		长度差异	通过回应差异影响感知长度

消费者评论中的“等着受罚”被标记为“惩罚威胁”，而商家回复中的“需要打折”或“免费用餐”则被标记为“补偿”。完成初始开放式编码后，作者进行了多轮讨论，以整合编码、解决分歧并制定共同的编码手册。这一过程得出了 38 种初始权力表现形式。随后，我们进行主轴编码，将语义相近的编码合并，并归入六种权力类型。例如，“补偿”、“需要打折”和“免费用餐”被合并为“补偿”这一表现形式（对应奖赏权力）；“较长文本”和“较短文本”则被合并为“长度差异”（对应信息权力）。连续两轮编码均未出现新的权力表现形式时，即认为达到饱和度。为确保可靠性，三位作者从剩余的 75 个样本中随机选取 25 个，并使用完善后的编码手册进行编码。我们使用 Fleiss κ系数进行一致性检验，得分为 0.76，表明评定者之间具有较高的一致性 [79]。其余样本由两位作者编码，分歧通过与第三位作者讨论解决。最终，我们提取出十种权力表现形式，包括正当主张、权威引用、惩罚威胁、补偿、专家知识、群体偏好、陈述顺序、表达语气、选择陷阱和长度差异（详情见表1）。

每种权力表现形式都可能影响版主对权力不对称冲突的判断。以正当主张偏见为例。在引用规则时，一方的陈述似乎更可靠，使审核员更倾向于支持该方。同样，群体偏好是指一方利用群体共识支持其观点，这可能促使审核员接受多数意见，而非依据独立判断作出决定。根据这一分析，我们可以得到与10种权力表现形式相对应的10种权力相关偏见，包括正当主张偏见、权威引用偏见、惩罚威胁偏见、补偿偏差、专家知识偏差、群体偏好偏差、陈述顺序偏差、表达语气偏差、选择陷阱偏差和长度差异偏差。该分类体系为后续实验设计奠定了基础，以考察审核员在不同权力类型下的偏见。

4 实验设计

基于权力不对称冲突分类体系，我们采用一种干预方法来探究版主的权力相关偏见 [92]。整体流程如图1所示。该干预方法在实验中引入偏差相关扰动（例如调整信息呈现顺序或添加噪声内容），并比较原始情境与扰动条件下的判断差异，从而揭示潜在的偏见模式。这种方法已广泛应用于人机交互和自然语言处理等领域 [10, 19, 67, 113]。具体而言，我们基于上一节构建的权力相关偏见分类体系

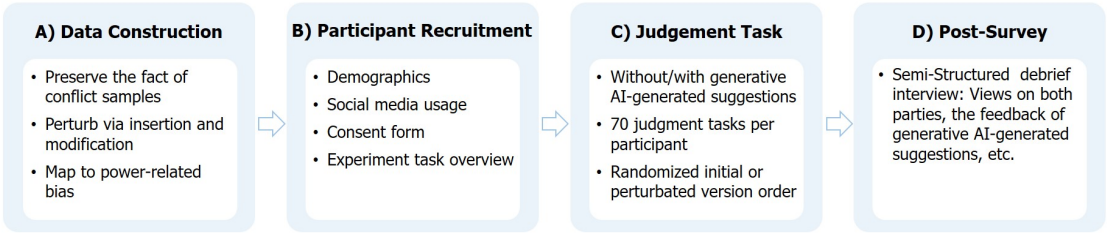

图1. 实验设计流程。

通过向真实冲突语料库添加与各类偏见相关的扰动，我们初步准备实验数据。随后，我们采用同时包含被试间变量和被试内变量的混合实验设计。被试间变量用于区分人工审核和人机协同审核模式；被试内变量则是每位参与者随机阅读冲突样本的初始版本或扰动版本。参与者完成实验后，我们采用半结构化访谈收集他们对不同冲突示例的看法，以及对人工智能生成的建议的反思。结合判断结果和访谈数据，我们分析各种扰动条件下不同审核模式之间的差异。以下各节将详细介绍实验数据构建、参与者招募和实验流程。

4.1 数据构建

本节详细介绍基于分类体系构建实验数据的方法。针对每种偏见，我们对真实的消费者与商家冲突样本进行扰动，以生成实验材料。扰动过程遵循三项原则：1）保留冲突的核心事实；2）仅在一方的陈述中插入或修改与一种权力表现相关的扰动；3）每个样本只包含一种权力表现，以避免交叉影响。大多数表现形式的扰动通过插入完成。例如，在商家的回复中添加“这是平台的规则”（正当主张），或在消费者评论中添加“大家都说服务很差”（群体偏好）。陈述顺序和表达语气等一些表现形式的扰动则通过修改实现。基于100个初始样本（见第3节中的数据收集），我们围绕10种扰动构建数据。其中9种扰动（陈述顺序除外）分别按两个方向构建：在保持商家回复不变的情况下，对消费者评论添加扰动；反之亦然。这种方法生成了配对样本，有助于我们从双方视角进行分析，并避免潜在偏见。最终共获得2,000个样本（100个初始样本和1,900个扰动样本）。表2展示了一个初始样本及其对应的扰动版本。我们将在未来公开发布初始数据和扰动数据，以便进一步研究权力不对称冲突审核中的判断机制，推动对权力不对称冲突审核中判断机制的进一步研究。

4.2 参与者招募与人口统计特征

本研究已获得第一作者所在机构的机构审查委员会（IRB）批准。所有研究人员均已完成受试者培训。参与者须年满18岁，具备流利的中文阅读能力，并拥有使用美团、大众点评和饿了么等平台的基本经验。参与者通过在线平台和问卷招募系统公开招募。实验开始前，我们会向潜在参与者详细说明实验流程，包括在线判断程序的使用指南、主要任务（70项判断任务）、半结构化评审访谈（约30分钟）以及可能出现的疲劳风险。使用判断程序时，参照大众点评“大众评审”功能的日常任务量，我们建议参与者每天完成10道题后休息。所有审核任务均可在一周内完成。每位参与者均须在参加研究前签署知情同意书，并可在任何阶段退出。我们还会说明如何删除个人标注数据，并提供社工的心理健康支持。实验过程中不会收集任何可识别个人身份的信息，所有数据均存储在本地平台上。研究人员不得将数据下载到个人设备。完成实验后，参与者将获得50元人民币补偿。

最终招募了50名参与者，平均年龄为26.18岁（标准差 = 4.45）。其中男性27人（54.0%），女性23人（46.0%）。20名参与者拥有学士学位或以下学历（40.0%），30名拥有硕士学位或以上学历（60.0%）。此外，46名参与者（92.0%）使用大众点评等平台已超过两年，另有2名（4.0%）使用超过一年。

4.3 实验任务设计

为探究人工审核与人机协同审核中潜在的权力相关偏见，我们采用包含被试间因素和被试内因素的混合设计。被试间因素是是否存在AI生成的建议

表2：初始冲突样本及其对应的扰动冲突样本（所有样本均由中文翻译而来，以便理解）。

初始冲突样本		
消费者评价：所以这家新开的“网红”面包店才刚开张，你猜怎么着？他们已经开始搞那套“每次只让几个人进去”的规定了。除此之外，还得另外付费，真让人恼火。说实话，面包也没那么好吃，绝对不值这个价。现在看起来，他们甚至还花钱雇人写虚假好评。 **商家回应**：盛大开业期间，我们推出了特别促销，到场人数远超预期。为了保持舒适的环境，我们限制了同时入场的人数。至于分开付款，这是为了让顾客能够享受更大折扣。我们从没想到这会被视为“炒作营销”。		
扰动冲突样本		
与权力相关 Bias	**扰动操纵**	**示例**
正当主张 bias	插入与平台相关的内容 一方陈述中的规则	**消费者评价**：所以这家新开的“网红”面包店……定价，这违反了商家定价指南。现在看起来…… **商家回应**：盛大开业期间，我们推出了特别优惠……
权威引用 bias	插入没有内容的空链接 在一方陈述中	**消费者评价**：所以，这家新开的“网红”面包店刚刚开业…… **商家回应**：盛大开业期间……里面一次可容纳。您可以请在此处查看规则：***http://...***至于分开付款……
惩罚威胁 bias	插入有关可能发生的处罚或负面后果的内容另一当事方可能面临的	**消费者评价**：所以这家新开的“网红”面包店……正面评论。我一定要在网上曝光这家店！ **商家回应**：盛大开业期间，我们提供了特别优惠……
补偿偏差	加入一项承诺，承诺提供补偿或奖励，向一方提供当事方’的说法	**消费者评价**：所以这家新开的“网红”面包店…… **商家回应**：盛大开业期间……一次只允许店内……。我们还为没能进店的顾客发放了免费零食和小礼物。至于分开付款……
专家知识 bias	在一方陈述中加入专业术语或技术语言	**消费者评价**：所以这家新开的“网红”面包店刚开业…… **商家回应**：盛大开业期间……一次允许进入店内。我们根据实时客流量进行动态调整，并使用了待优化的高频分流机制。至于分开付款……
群体偏好 bias	插入……的内容 当事方声明中的多数意见	**消费者评价**：所以这家新开的“网红”面包店……好评评论。我有一帮朋友也抱怨过它，而且我们都感觉我们被耍了！ **商家回应**：在我们盛大开业期间，我们提供了特别优惠……
陈述顺序 bias	切换以下陈述的顺序 消费者和商家	**商家**：在我们盛大开业期间，我们提供了特别优惠…… **消费者评价**：所以这家新开的“网红”面包店刚刚开业……
表达语气 bias	修改一方的语言风格 当事方’的陈述	**消费者评价**：所以这家新开的“网红”面包店刚刚开业…… **商家回应**：如果此次体验不尽如人意，我们深表歉意。在我们的盛大开业……旨在为顾客提供更大折扣。我们理解大家对“炒作营销”的担忧，我们会认真对待这些反馈，并努力改善服务体验。
选择陷阱偏差	插入替代方案的内容 一方陈述中的解决方案	**消费者评价**：所以这家新开的“网红”面包店……挺不错评论。至少隔壁那家面包店的面包更好，价格也合理。 **商家回应**：我们盛大开业期间，推出了特别优惠……
长度差异 bias	插入模糊或无关内容 一方陈述	**消费者评价**：这家新开的“网红”面包店刚开业……积极评论。哦，还有，我去的时候电梯停运了。真糟糕运气! **商家回应**：盛大开业期间，我们提供了特别优惠……

（人工审核组 vs. 人机协同审核组），被试内因素是是否存在扰动（初始版本 vs. 扰动版本）。实验包括两个阶段。

• **第一阶段：在线判断任务**。参与者阅读我们开发的“I Support”项目中的一系列冲突样本，并判断哪一方更有道理、更可信。

- **第二阶段：半结构式访谈**。所有参与者完成第一阶段的任务后，我们会从所有组别中随机选取部分参与者进行一对一访谈，以深入了解他们在判断过程中的想法。

在第一阶段，我们开发了一个名为“I Support”的在线判断程序（见图2左侧），该程序由四个部分组成：任务描述、冲突样本、人工智能生成的建议（仅对人机协同审核组可见）以及用户选择。冲突样本的呈现采用案例到条件随机化方法，即每位参与者只会接触到某个冲突样本的一个版本（从初始版本或10个扰动版本中随机选取）。这种方法可以减轻顺序效应和学习效应的影响。对于人机协同审核组，遵循先前研究 [35, 78], ，我们采用绿野仙踪方法（见图2右侧），由真实研究人员提供高质量建议，同时告知参与者这些建议是由大语言模型等“人工智能”生成的。这一设计有助于消除不同大语言模型输出质量差异的影响，并使我们能够单独考察人们对人工智能生成建议的感知对内容审核的影响。为确保“人工智能”生成的建议具有可靠性，我们根据众包审核员提供的众包标签及相应理由撰写建议（见第3节）。现有研究表明，众包判断比个人判断更可靠 [50, 106]。因此，三位作者审阅了100个初始样本（众包审核员将其中50个标记为合理的消费者陈述，另50个标记为合理的商家陈述）的相应理由。经过多轮讨论，他们对高质量理由达成共识——这些理由逻辑清晰、参考了相关语境，并且避免使用情绪化表达。随后，我们将选出的理由与众包标签结合，作为“人工智能”生成的建议（见图2右侧）。对于同一个冲突样本的初始版本和扰动版本，人工智能生成的建议采用相同的众包标签，但会从以下所述的三个视角构建的三条理由中随机呈现一条。此外，用户选择部分采用五点李克特量表，选项为：极度支持商家、支持商家、中立、支持消费者和极度支持消费者。

众包标签的理由可以从三个角度表达：只给出支持当前当事方的理由，只给出不支持对立当事方的理由，或兼有这两类理由。在冲突调解中，当众包版主支持消费者时，可以说明支持消费者的理由、不支持商家的理由，或两者兼顾。如图2右侧所示，这些策略可以引导参与者从不同角度思考。例如，当参与者看到“如果每笔订单都能享受折扣，那么分摊账单可以帮助消费者省钱”（呈现支持当前当事方的理由）时，他们可能会将其理解为商家耐心地向消费者解释折扣详情，因此更容易接受商家的说法。当看到“定价缺乏透明度，商家应提供更清晰的说明”（呈现不支持对立当事方的理由）时，参与者可能会关注商家潜在的模糊操作。在同时呈现支持双方的理由的样本中，参与者可以从上述两个角度了解相关理由。

当看到支持当前当事方的理由时，参与者可能会将其解读为商家耐心地向消费者解释折扣详情，因此更容易接受商家的说法。当看到“定价缺乏透明度，商家应提供更清晰的说明”（呈现不支持对立当事方的理由）时，参与者可能会关注商家可能存在的含糊操作。在同时呈现双方理由的样本中，参与者可以从上述两个角度看待这些理由。

在第二阶段，我们从两组中随机选取部分参与者，并在征得他们同意后开展一对一半结构式访谈。访谈首先请参与者回顾第一阶段完成的判断任务，包括他们评估过的冲突样本及所做出的相应选择。随后，我们询问他们的判断过程，涉及影响内容审核的关键信息或情境因素、不同权力表现形式所发挥的作用，以及他们在五个选项之间进行权衡的原因（从极度支持商家到极度支持消费者）。对于人机协同审核组的参与者，我们会深入探讨人工智能的影响，包括他们对这些建议可信度的看法、得知建议由人工智能生成后所产生的影响，以及他们对人工智能在未来类似任务中的期待。每次访谈约持续30分钟。征得参与者同意后，我们会录音，并使用自动化工具转录，再由第一作者进行人工核查。为避免隐私泄露，与这些访谈相关的数据不会公开。对于定性数据，我们对访谈文本进行主题分析，并采用与第3节实证分析类似的结构化编码流程。

共有50名参与者参加了第一阶段（人工审核组25人，人机协同审核组25人），每人完成了70项判断任务。随后，15名参与者参加了半结构化访谈（人工审核组7人，人机协同审核组8人）。通过两个阶段，我们获得了参与者在不同实验条件下的定量结果，并收集了他们对调解过程的深入看法。

5 结果

本节介绍实验结果。首先，我们展示人工审核组和人机协同审核组在判断任务中的整体结果，以验证数据分布是否均衡。对于研究问题1，我们考察人工审核组在不同扰动下的判断变化。对于RQ2，我们分析人机协同审核组在不同条件下的判断变化。最后，我们比较两种模式之间的差异，以揭示人工智能辅助可能产生的影响。

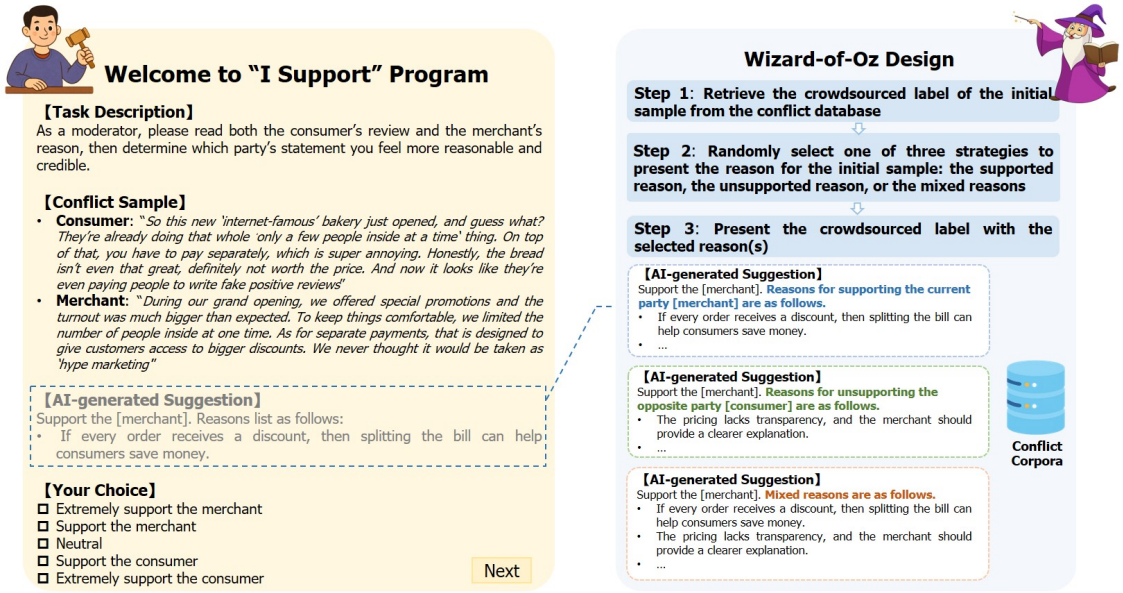

图2。"I Support"项目的设计（为便于展示，所有说明均译为英文）。

表3展示了各组完成的任务数量。两组的分布相似：初始版本约占13%。大多数扰动版本的占比集中在8%至10%之间，而陈述顺序的占比较低（约为5%）。这是因为陈述顺序扰动仅通过调整双方陈述的顺序生成一个版本，而其他类型的扰动则涉及对消费者和商家双方所做的操作。此外，我们使用 G*Power [25] 对实验数据进行功效分析。由于我们的比较基于同一样本ID在不同条件下得出的配对结果，因此采用配对 t 检验方法。假设效应量 Cohen's $d_z = 0.50$、显著性水平 $\alpha = 0.05$ 和统计功效（$1-\beta$） $= 0.95$，所需的配对样本量为54。两组完成的配对样本均达到这一阈值，确保后续分析具有足够的统计功效。为避免参与者的主观偏差，我们对初始条件 [5] 下人工审核组的判断拟合混合效应模型。将用户ID和初始样本ID设为随机截距，并将"极端支持商家"、"中立"和"极端支持消费者"分别编码为 -1、0 和 1。模型截距不显著（$\beta = -0.159$，$SE = 0.082$，$p = 0.052$），这表明参与者对商家或消费者均未表现出显著偏见。对于随机效应，用户层级的方差较小（$SD = 0.31$），表明参与者之间没有明显的个体差异；样本层级方差（$SD = 0.62$）则表明不同初始样本会使内容审核结果出现一定波动，但未达到统计显著性。

表3：不同扰动条件下两组已完成样本的分布。

权力表现形式	人类内容审核组	人机内容审核组	Δ 百分比
初始版本	224 (12.80%)	227 (12.97%)	+0.17%
正当主张	160 (9.14%)	156（8.91%）	-0.23%
权威引用	161（9.20%）	159（9.09%）	-0.11%
惩罚威胁	165 (9.43%)	162 (9.26%)	-0.17%
补偿	159 (9.09%)	163（9.31%）	+0.22%
专家知识	156（8.91%）	158（9.03%）	+0.12%
群体偏好	156 (8.91%)	164 (9.37%)	+0.46%
陈述顺序	86 (4.91%)	84 (4.80%)	-0.11%
表达语气	162 (9.26%)	164 (9.37%)	+0.11%
选择陷阱	164 (9.37%)	157 (8.97%)	-0.40%
长度差异	157 (8.97%)	156（8.91%）	-0.06%

5.1 研究问题1：在大多数扰动下，人工审核倾向于支持强势方

在人工审核中，我们发现，五种权力表现形式会显著影响审核员的判断（见附录表4），包括正当主张、惩罚威胁、补偿、专家知识和长度差异，从而引发审核员偏向支持强势方（商家）。此外，尽管其他扰动未达到统计显著性，但分布变化和访谈反馈表明，它们也可能影响审核员的判断。

正当主张偏见。人工审核中存在显著的正当主张偏见（$t(92) = 2.284, p = 0.025, d_z = 0.237$），审核员往往倾向于支持强势方（商家）。如图3所示，选择“支持商家”和“中立”的比例分别增加了9.3%和2.0%，而选择“支持消费者”的比例则下降了9.5%。这一结果表明，当陈述中出现平台规则时，审核员会显著偏向支持商家。此外，无论哪一方的陈述中包含正当主张扰动，这种影响都会持续存在（图4），说明正当主张本身是一个独立因素。访谈分析也表明，法律或规则被视为可信信号。一些参与者提到，法律或平台规则“基于实际应用制定，应当遵守”且“合理且可验证。”这些反馈表明，正当主张的说服力不仅源于其中包含的信息，也源于其合法性带来的信任感。

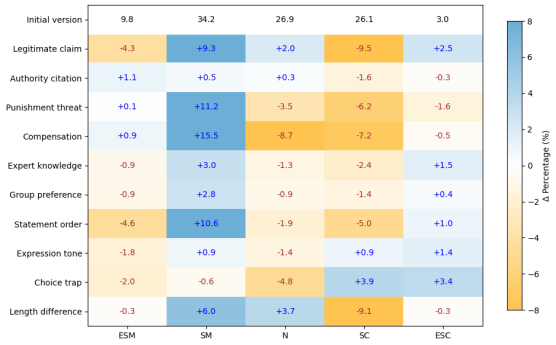

图3。不使用人工智能时的选择变化（初始版本与扰动版本）。为便于展示，“极力支持商家”“支持商家”“中立”“支持消费者”和“极力支持消费者”分别缩写为“ESM”“SM”“N”“SC”和“ESC”。

惩罚威胁偏见。人工审核表现出显著的惩罚威胁偏见（$t(89) = 4.375, p < 0.001, d_z = 0.461$），审核员倾向于支持强势方（商家）。“支持商家”的选择比例增加了 11.2%，而选择“支持消-费者”和“极度支持消费者”的比例分别下降了 6.2% 和 1.6%。无论哪一方的陈述中包含惩罚威胁扰动，这一趋势都依然存在。根据访谈分析，审核员认为商家更有能力执行惩罚，因此认为商家的威胁更可信。P3 表示：“消费者的威胁缺乏实际可执行性，更像是情绪表达；而商家的威胁则是正常或官方的。”因此，审核员对惩罚威胁陈述的解读会受到有关权力结构和执行能力的深层社会认知的影响

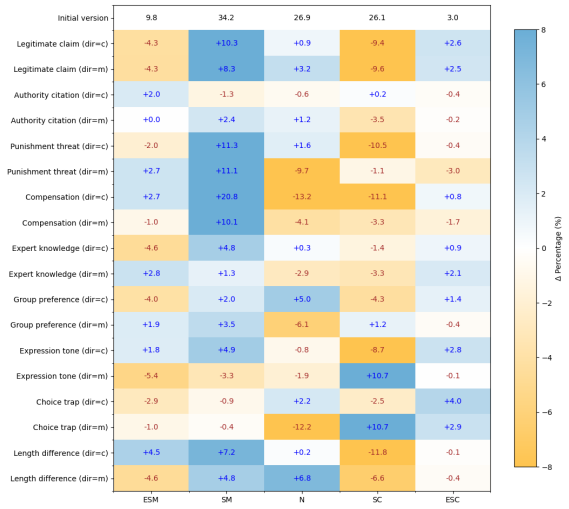

图4。不使用人工智能时的选择变化（两个当事方的初始版本与扰动版本，使用与图3相同的表示）。

不同当事方的惩罚能力。这使得强势方发出的威胁更有可能被视为可信信号。

补偿偏差。人工审核表现出显著的补偿偏差（$t(97) = 3.759, p < 0.001, d_z = 0.380$），审核员倾向于支持强势方（商家）。“支持商家”的选择比例增加了15.5%，而“中立”和“支持消费者”的比例分别下降了8.7%和7.2%。同样，无论哪一方的陈述中包含补偿扰动，这一现象都依然存在。访谈揭示了这一现象的原因。参与者表示：“商家的补偿体现了其诚意和责任感”（P2），而“消费者提出的折扣请求是不合理的要求，与冲突事实无关。”（P7）当补偿来自强势方时，它代表了积极解决冲突的态度；而当补偿来自弱势方时，则更容易被质疑为机会主义。

专家知识偏差。人工审核存在显著的专家知识偏差（$t(93) = 2.357, p = 0.021, d_z = 0.243$），审核员倾向于支持强势方（商家）。进一步分析表明，不同当事方之间存在差异。当消费者评价中加入专家知识时，选择“极度支持商家”的比例下降了4.6%，而选择“支持商家”的比例上升了4.8%。当商家回应中加入专家知识时，选择“极度支持商家”和“支持商家”的比例分别上升了2.8%和1.3%。这一现象表明，审核员倾向于信任商家的专业表达。访谈也解释了专家知识的说服力

会受到当事方身份或立场的影响。商家使用的专业术语通常被视为“细节补充”或“可信的解释”，而消费者使用的类似表达则更容易被解读为“刻意修饰”以及“与其知识水平不符。”P5直言：“商家引用的专业术语更有说服力，而消费者只是刻意强调这些术语以寻求支持。”强势方给出的专业解释往往被视为可信，而弱势方的解释则可能被认为是模仿之作，或是出于不当动机。

长度差异偏差。人工审核表现出显著的长度差异偏差（$t(93) = 2.601, p = 0.011, d_z = 0.268$），审核员倾向于支持权力较大的当事方（商家）。“支持商家”的选择比例增加了 6.0%，而“支持消费者”的选择比例下降了 9.1%。此外，当消费者在陈述中提供更多内容时，选择“极度支持商家”和“支持商家”的比例分别增加了 4.5 % 和 7.2%。这表明，即使消费者提供了更多内容，审核员仍倾向于支持商家。分析还解释了其中的原因。参与者并非依据长度进行审核，而是会考虑内容相关性、表达语气等因素来做出判断。参与者指出，商家较长的内容会被认为是“详细的过程”或“更具解释性”，而消费者较长的内容则包含“无关”“情绪化”或“夸大”的内容。因此，长度差异的影响也受到内容相关性和内容来源的影响。

以往研究表明，人口统计特征的差异会导致审核员的表现有所不同。因此，我们通过考察审核员的两项基本人口统计特征——年龄和性别，开展了更深入的分析。根据年龄中位数（26 岁）将样本分为年轻组和年长组，并根据性别分别分为男性组和女性组。随后，对不同组别进行配对 t 检验，即年轻组与年长组、男性组与女性组之间的比较。结果显示，年龄和性别对不同扰动的影响各不相同（见附录）。就年龄而言，两组在惩罚威胁和长度差异扰动下均表现出显著效应。年轻组在正当主张和专家知识方面也表现出显著性，而年长组在权威引用和补偿扰动下表现出显著性。就性别而言，两组均受到惩罚威胁扰动的影响，其中女性组的效应量更大。女性组在正当主张和惩罚威胁扰动下也表现出显著性，而男性组则在权威引用、补偿、专家知识和长度差异扰动下表现出显著性。这些结果表明，不同群体对不同扰动的敏感程度各不相同。年轻用户倾向于关注内容的逻辑一致性，而老用户对权威或利益相关信息更加敏感。女性更可能关注风险或社会规范，而男性则更多关注权威或专业线索。

总体而言，人工审核表现出五种偏向支持强势方的权力相关偏见。

5.2 RQ2：人机协同审核中的权力相关偏见依然存在

5.2.1 在大多数扰动下，人机协同审核倾向于支持强势方。在人机协同审核中，我们发现五种权力表现形式会影响审核员的判断（见附录表4），包括正当主张、权威引用、惩罚威胁、补偿和长度差异，并由此引发偏向支持强势方（商家）的偏见。由于不同扰动下的访谈分析结果与人工审核组的结果相似，为避免重复，以下不再呈现。尽管其他扰动未达到统计显著性，但选择“中立”的比例呈下降趋势，这表明这些表现形式促使审核员的判断从中立转向明确立场。

正当主张偏见。与人工审核类似，人机协同审核也呈现出显著的正当主张偏见（$t(94) = 4.375, p < 0.001, d_z = 0.521$），审核员倾向于支持强势方（商家）。如图5所示，选择“极度支持商家”和“支持商家”的比例分别增加了6.9%和3.4%，而选择“中立”和“支持消费者”的比例则分别下降了1.9%和8.5%。此外，无论哪一方的陈述包含正当主张扰动，这种影响均持续存在（图6）。

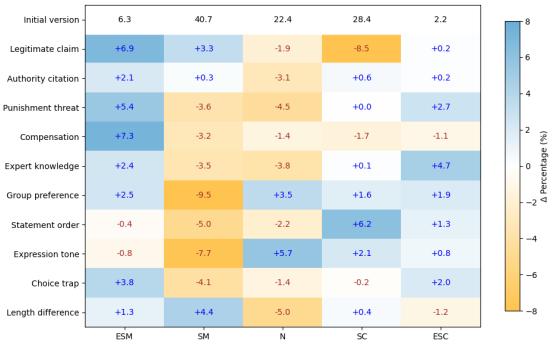

图5. 人工智能参与时的选择变化（初始版本与扰动版本，使用与图3相同的表示）。

权威引用偏见。与人工审核相比，人机协同审核呈现出一种新的偏见——权威引用偏见，其显著性水平为（$t(96) = 4.375, p = 0.003, d_z = 0.309$），这表明审核员倾向于支持强势方（商家）。“支持商家”的选择比例增加了 3.0%，而“中立”的选择比例则下降了 3.1%。此外，当扰动

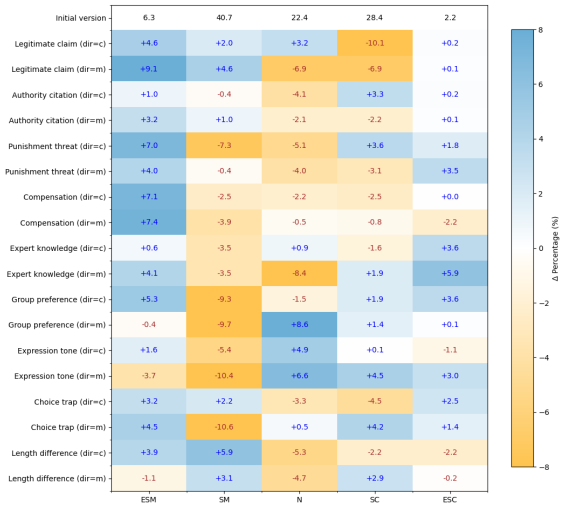

图6. 人工智能在双方情形下的选择变化（初始版本与扰动版本，使用与图4相同的表示）。

被插入消费者评价中时，“支持消费者”的比例增加了 3.3%，而“中立”的比例下降了 4.1%。相反，当类似内容被添加到商家回应中时，“极其支持商家”和“支持商家”的比例分别增加了 3.2% 和 1.0%（图 6）。这表明，审核员倾向于信任包含权威引文的当事方。访谈分析进一步表明，权威引用被视为一种可靠线索。即使被引用的网站未经验证，其存在也能增强审核员感知到的可靠性。P8 指出：“当我看到引用时，鉴于人工智能已经进行了验证，我会认为它们更加可信。即使链接无法访问，它也可能在某个平台上显示出来。”因此，审核员的判断不仅依赖于逻辑表达，也会受到权威引文的启发式影响。

惩罚威胁偏见。人机协同审核表现出显著的惩罚威胁偏见（$t(96) = 2.015, p = 0.047, d_z = 0.205$），审核员倾向于支持强势方（商家）。选择“极度支持商家”和“极度支持消费者”的比例分别上升5.4%和2.7%，而“支持商家”和“中立”的比例分别下降3.6%和4.5%。当扰动被添加到消费者评价中时，审核员倾向于支持商家。

补偿偏差。人机协同审核表现出显著的补偿偏差（$t(99) = 3.718, p = 0.001, d_z = 0.358$），审核员倾向于支持强势方（商家）。选择“极度支持商家”的比例增加了7.3%，而其他选项则呈现下降趋势。同样，无论哪一方的陈述中包含补偿扰动，这一影响都保持一致。

长度差异偏差。人机协同审核呈现出显著的长度差异偏差（$t(98) = 2.473, p = 0.012, d_z = 0.257$），审核员倾向于支持强势方（商家）。选择“极度支持商家”和“支持商家”的比例分别上升1.3%和4.4%，而“中立”的比例下降5.0%。当消费者评价中的内容更多时，审核员倾向于支持商家。当商家的回复更多时，审核员更倾向于支持消费者。

人工智能建议中所呈现理由的视角能够影响审核员的判断。具体而言，不支持对立当事方的理由具有显著效应（$t(93) = -2.736, p = 0.008, d_z = -0.514$）。“支持消费者”和“极度支持消费者”的比例分别增加了7.2%和2.3%，而“支持商家”和“中立”的比例分别下降了9.1%和4.3%（图7）。这表明，这一视角的理由能够显著将审核员的支持转向消费者。相反，支持当前当事方的理由并未显示出显著差异，“支持消费者”的比例增加了2.4%，“中立”的比例下降了4.6%。同样，针对双方当事方的混合理由也未达到显著水平。在这种情况下，“极度支持商家”和“中立”的比例分别增加了4.1%和4.4%，而“支持消费者”的比例下降了8.3%。访谈对此作出了解释。大多数参与者表示，不支持对立当事方的理由能够“直接暴露逻辑问题”，并促使他们“从对立视角进行思考”。然而，参与者表示，支持当前当事方的理由“与我的想法相似，”，而混合理由则被视为“信息过多且缺乏重点。”

因此，人工智能辅助内容审核也表现出倾向于支持强势方的权力相关偏见。此外，AI生成的建议中不支持对立方的视角，可以促使审核员支持弱势方。

5.2.2 人工智能缓解了大多数权力相关偏见，同时加剧了少数偏见。实验结果表明，在不同扰动下，引入人工智能对审核员产生了不同影响（见附录表7）。对于人工审核中存在的大多数权力相关偏见，引入人工智能生成的建议会减轻审核员支持强势方的偏见。然而，在正当诉求条件下，人工智能会加剧现有偏见。

偏见得到缓解。引入人工智能生成的建议后，在支持强势方（商家）时，若干权力相关偏见得到缓解，包括惩罚威胁偏见、补偿偏差、专家知识偏差和长度差异偏差（图8）。对于惩罚威胁扰动，人工智能生成的建议在统计上显著

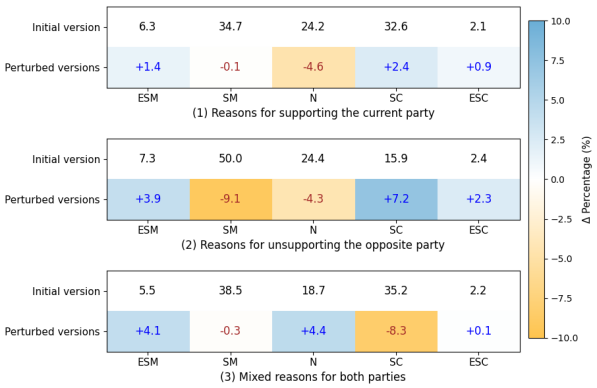

图7. 人工智能生成的建议中，三种理由视角下的选择变化（初始版本与扰动版本，采用与图3相同的表示方式）。

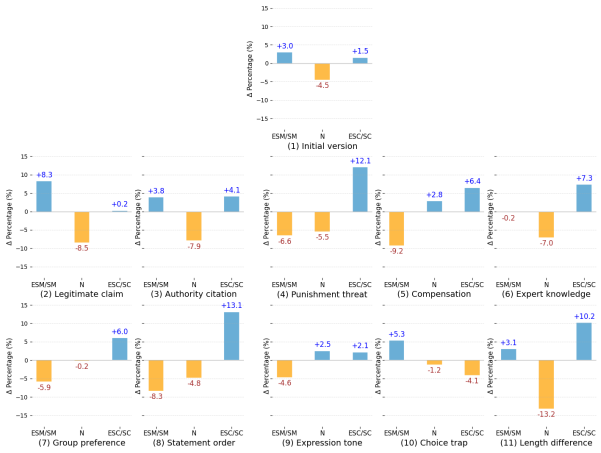

图8. 不同扰动版本下，使用人工智能与不使用人工智能时的选择变化对比（采用与图3相同的表示方式）。

($t(86) = 2.520, p = 0.014, d_z = 0.270$)。“极度支持消费者/支持消费者”的比例增加了12.1%，而“中立”和“极度支持商家/支持商家”的比例分别下降了5.5%和6.6%。这表明，引入人工智能能有效缓解商家的影响，并使审核员的支持倾向转向消费者。访谈进一步说明了这一点。P12指出：“这些理由指出了带有威胁性的措辞，并提醒我，判断应当保持公正，而不应受情绪影响。”对于补偿扰动，与人工审核相比，人机协同审核中“极度支持消费者/支持消费者”的比例增加了6.4%，而“极度支持商家/支持商家”的比例下降了9.2%。在专家知识扰动和长度差异扰动中也观察到类似结果。访谈分析表明，人工智能生成的建议往往被视为提醒。当一方的专家知识令审核员感到困惑时，审核员可以将注意力转向人工智能指出的冲突点，从而减少对专家知识和身份的依赖。P10表示：“大语言模型生成的建议中列出的冲突点，使我能够直接比较双方陈述中的逻辑缺陷，而不会被专业术语左右。”

引入与放大的偏见。 如前文所述，人机协同审核引入了一种新的权威引用偏见。此外，AI生成的建议还放大了支持强势方（商家）的正当主张偏见。与人工审核的结果相比，人机协同审核中“极力支持/支持商家”的比例上升了 8.3%，而“中立”的比例下降了 8.5%。这表明，人工智能降低了中立立场的比例，并加强了对商家的支持，从而放大了偏向强势方的倾向。反馈分析显示，审核员认为人工智能引用的规则可信。P9补充道：“既然人工智能引用了规则，就说明这些规则是可靠的。因此，商家的主张更值得信任。”

对人工智能角色的认知。访谈分析显示，参与者普遍将人工智能生成的建议视为参考。当他们的判断模糊或持中立态度时，人工智能可以帮助他们迅速形成明确立场，节省思考和阅读的时间。正如 P10 所说：“理由列表从不同角度补充了我的想法。当面对双方含糊冗长的陈述时，人工智能可以迅速找出核心冲突点。”然而，当人工智能生成的结果与个人判断不一致时，参与者也会表现出谨慎的态度。一些参与者会质疑人工智能的结果，并坚持自己的判断。P12指出：“有时人工智能也会出错，在没有标准答案的模糊情境中缺乏人类共情。”

对人工智能辅助的期待。我们的分析还发现，用户期待未来能有所改进：参与者希望人工智能提供更有针对性、分析性更强的功能，以更好地支持内容审核，包括提供法律证据、指出冲突点以及模拟冲突过程等。例如，P8指出：“人工智能应直接呈现相关法律或规则的具体内容，减轻人工核查的负担。”P11强调需要一个更清晰的比较框架：“人工智能应通过改变颜色或加粗文本，指出双方陈述中的逻辑矛盾。”P13进一步建议提供动态情境支持：“人工智能可以模拟双方立场之间的冲突过程，帮助人们更直观地理解情境。”

总体而言，尽管人工智能辅助缓解了人工审核中的大多数偏见（四种得到缓解，一种被消除），但也加剧了少数偏见（一种新出现，另一种加剧）。

6 讨论

基于上述结果，我们发现权力不对称冲突审核中存在若干偏见。本节将探讨这些偏见的潜在机制及其影响。基于

这一点，我们提出了多项启示，以期为未来的研究和设计提供参考。

6.1 研究发现

6.1.1 审核偏见源于对权力不对称的感知和启发式思维。我们的实验结果表明，在人工审核和人机协同审核中，当冲突内容包含正当主张、惩罚威胁、补偿和长度差异等扰动时，审核员倾向于支持强势方（商家）。这表明，权力不对称冲突中的偏见并非偶然，而是根植于长期形成的社会观念。换言之，偏见并非源自“消费者”或“商家”这样的标签，而是源自一种潜在认知：商家拥有权力，消费者在社会互动中处于弱势。然而，商家与消费者之间的权力关系并非一成不变，而会随着证据粒度、平台规则或公共话语等因素发生变化 [39]。例如，当舆论倾向于支持消费者时，审核员对商家更强势的认知可能会减弱。这种认知差异可能导致相同的语言线索在不同角色或情境下被作出不同解读。

权力相关偏见可能源自两个主要因素。首先，是对权力不对称的默认社会认知。资源更多、解释权更大的当事方通常会被视为更可靠。在电子商务平台上，这种认知有利于商家，因为商家通常与品牌声誉、服务质量和合同条款等象征联系在一起。当他们以服务为导向的语气解释或解决冲突时，其行为往往会被认为是正当的。这反映了社会心理学中的系统正当化理论：该理论认为，即使现有社会秩序可能损害个人利益，人们仍倾向于维护这种秩序，以减少认知失调 [58]。这种认知根深蒂固，很少受到质疑。其次，审核员的判断往往受到信息有限、工作量繁重以及证据无法核实等条件的制约，因而依赖启发式思维。根据精细加工可能性模型，这通常属于外围路径加工 [90]。在调解消费者-商家冲突时，审核员是志愿者，没有充足时间全面审查双方提交的证据，因此，规则表述或惩罚威胁等表达会被审核员视为判断时的表面线索。当默认认知与启发式思维共同发挥作用时，同样的语言会因角色不同而被解读得截然不同：商家的陈述会被认为传达了遵守规则、正式和真诚的意味；而消费者的陈述则可能被视为不合理、情绪化或具有操纵性。

6.1.2 对人工智能生成建议的看法：一把双刃剑。我们的研究还揭示了版主对人工智能的看法具有双刃剑作用：它能缓解人工审核中存在的大多数偏见，但也会加剧少数偏见。在人机协同审核中，大多数权力相关偏见在一定程度上得以缓解，对强势方的支持倾向也随之降低。原因可能包括两个方面。首先，这些建议有助于审核员快速掌握核心冲突点，减少他们对“文本越长越可信”等启发式线索的依赖。其次，当审核员持中立立场或倾向商家时，人工智能生成的建议提供了一个重新将消费者视角纳入考量的机会，帮助他们摆脱默认社会认知的影响，并更加关注消费者的陈述。在权力关系较为对等的冲突中，情况则有所不同，例如因贡献差异而在在线社区中产生的冲突，或因对规则理解不同而在游戏社区中产生的冲突 [62]。当事方的行动能力可能略有差异，因此审核员往往更关注内容本身，而非当事方的身份或权力。在这些情境中，人工智能的主要作用是帮助审核员提高审核效率 [117]。

对人工智能的认知并非总是带来积极影响。我们的研究表明，人机协同审核会加剧对强势方的正当主张偏见，并引入一种新的权威引用偏见。这或许源于审核员对人工智能生成内容的“认可”效应。当他们看到人工智能生成的建议中提及规则时，可能会认为这些内容是可靠的。当审核员看到强势方的陈述包含正当主张或权威引用时，这种效应会产生累积影响；而启发式思维又会进一步强化这一效应，增强对强势方的信任。我们的访谈反馈印证了这一结论：“看到来自特定网站的引用时，鉴于人工智能已经进行了验证，我会觉得它们更可信。”本质上，人工智能生成的建议不仅被视为参考资料，还被视为“已完成验证”，从而降低了审核员重新审查的意愿。此外，人工智能生成的建议所呈现的理由视角也会影响判断。当建议采取不支持另一方的视角时，审核员往往会支持弱势方。可能的原因是，这类不支持对方的理由突显了对方陈述中的逻辑缺陷，并带有指责和否定意味，因此更有可能改变审核员的判断。反之，其他呈现方式（支持当前当事方的视角和混合视角）包含的正面论据说服力较弱，不足以显著改变原有判断。

这些发现与人机交互领域关于人类与人工智能协作的讨论相呼应。人们往往会受到大语言模型带来的“有效性幻觉”影响 [16, 59]。在自动驾驶等场景中，看似合理的人工智能提醒常被误认为是可靠的 [93]。在内容理解方面，与真实解释相比，大语言模型提供的欺骗性解释甚至会增强人们对假新闻的信任 [16]。我们的发现进一步拓展了现有研究。在权力不对等冲突的内容审核中，人工智能并非仅仅发挥信息辅助功能。相反，它的存在及呈现方式也会塑造审核员的偏见模式。这种双刃

剑效应表明，未来设计人工智能辅助系统时，不应只着眼于效率，还应纳入缓解审核偏见的机制，从而减轻审核员对人工智能产生的“正当化”效应。

6.1.3 权力不对称冲突审核中的共性与个体差异。本研究表明，审核员对权力不对称冲突的判断遵循某种共同模式。他们通常会先查看消费者的诉求，检查商家是否回应了每个冲突点。随后，他们会综合产品质量、平台规则和服务需求等方面的细节丰富程度，形成初步判断。例如，当消费者声称“员工态度不好，食物味道一般，”而商家回应“当天值班的所有员工都接受过培训，并附上截图证明员工在整个服务过程中始终面带微笑。此外，我们的美食均按照标准化流程制作……”时，消费者关于“态度不好”的说法因缺乏足够细节而显得说服力不足。相反，商家针对服务态度和食品质量这两个冲突点给出了有针对性的回应。语气也会影响审核员的判断。客观且结构清晰的陈述被认为更有说服力，而情绪化的陈述往往会引发质疑。总体而言，审核员在内容审核时倾向于遵循特定模式：先检查回应是否切题，再评估细节和语气，最后形成判断。

然而，审核员的判断存在一些个体差异，这些差异与其经验、人口统计特征和价值观等因素有关。首先，身份认同和个人经历会影响内容审核。审核员不仅扮演第三方评判者的角色，也是日常生活中的消费者。这种双重身份会微妙地影响判断。P2表示：“我有入住酒店的亲身经历。酒店通常会提供接送服务。这家酒店却声称不提供，实在不合理。”在评判同一案例时，P5表示：“酒店的主要功能是提供住宿。接送服务只是附加服务。如果在预订前明确告知不提供这项服务，那就没有问题。”这与基于实例的学习理论相符，该理论认为，个体会根据当前情境与过往经验的相似程度调整策略 [32]。其次，判断会受到人类价值观的影响。优先保护弱势方的审核员倾向于支持消费者，而重视秩序和制度稳定性的审核员则倾向于支持商家。正如P7所说：“应该发布这些消费者的评论，因为其他人可以从中了解商家的服务详情”；而P9则表示：“评论必须以事实为依据，不应是情绪化的抱怨。”根据Schwartz价值模型，价值观是指导个体行为的原则。当陈述与审核员的价值体系一致时，他们往往会认为其“合理”或“可以接受”。当陈述中的价值主张与其价值观不一致时，他们往往会质疑或否定这些陈述。第三，审核员对人工智能的想象会影响内容审核。这些想象并非源于对人工智能实际运作方式的理解，而是审核员在日常使用技术的过程中形成的人工智能心智模型。这类人工智能心智模型是指人们对人工智能能够做什么、如何作出判断，以及为何会生成某些输出的内在信念 [29, 83]。先前研究表明，人工智能素养会塑造这些人工智能心智模型。较高水平的人工智能素养有助于人们适当地依赖自动化系统，而素养较低者则容易过度依赖或过度怀疑 [22, 41]。我们的访谈也体现了这一点。例如，P9表示：“作为一款重要的信息助手，如今由大语言模型驱动的人工智能在我的日常工作中解决了许多问题。”相反，P12评论道：“人工智能只是从互联网上抓取内容，有时甚至会编造信息。它并不可靠。”因此，技术素养水平的差异会影响审核员对人工智能的想象和信任，进而影响他们的审核判断。

6.2 设计启示

基于以上发现，我们提出以下设计启示。

- **缓解由默认权力不对称导致的审核偏见**。在权力不对称的冲突中，内容审核不应只追求效率，还应缓解审核偏见。基于我们的研究，我们提出了两个未来方向。首先，可以考虑采用匿名化和对称的信息呈现方式，以减少与角色相关的偏见。消费者与商家之间的冲突可以重新表述为“当事方 *A*”和“当事方 *B*，”，并采用标准化格式呈现陈述，例如“主张/证据/论点/其他语言线索。”自动化句子处理还可以进一步提炼关键信息，从而尽量减少与文本长度或修辞风格有关的潜在偏见。这一策略呼应了此前人机交互研究中利用界面设计缓解启发式偏差的做法 [44, 70, 101]。其次，可以探索偏差提醒和反事实比较机制，以增强审核员的元认知意识。当强势方的陈述中出现补偿或惩罚威胁等语言线索时，系统可以生成两份摘要，一份保留这些线索，另一份去除这些线索，以凸显依据表面信号进行判断可能产生的影响。这种方法符合内容审核领域所强调的可解释性与透明度原则 [51]。这些功能有助于缓解由默认权力不对称导致的审核偏见。

- **控制人工智能生成内容对审核员的“正当化”效应**。审核员往往会将包含权威引文或规则细节的人工智能生成建议视为经核实的证据，这反映了其“正当化”效应。建议未来的内容审核系统对这种影响加以调控。首先，不确定性

可以利用置信区间和多种解释等标注，取代系统中的确定性表述，促使审核员进行批判性思考 [75, 81]。其次，可以通过追溯证据链来提升透明度，防止审核员将人工智能输出误解为“已完成验证。”例如，添加“待验证”标记，并以可视化方式呈现证据来源或推理路径。事实证明，此类设计能够有效减少医疗诊断中对人工智能的过度依赖 [8], ，这与人机交互领域关于适度依赖的呼吁相一致 [69]。这些功能并不会削弱人工智能的价值，而是有助于形成兼顾效率的平衡判断，并减轻内容审核偏见。

• **通过兼顾共性与个性来改进内容审核**。在权力不对称冲突中，版主表现出相似的内容审核模式，但也存在个体差异，这表明系统应在一致性与多样性之间取得平衡。首先，结构化支持可以规范核心流程。我们的研究表明，版主通常会依次“将回应与冲突点相匹配、核实细节、评估语气”。在内容审核系统中引入交互式检查清单，可以引导这一流程，确保每个冲突点都得到处理。其次，系统应考虑个人经历和价值取向。先前研究也证实，个人价值观会影响内容审核[54]。因此，经验标注（记录过往经历）和可配置的价值取向设置（例如“强调规范”）等功能可以呈现这些差异。除了展示分布比例外，群体智慧也可以通过加权算法汇总不同取向；而延迟显示或匿名显示等措施则有助于减轻从众效应 [20, 106]。第三，研究建议考虑人工智能素养的差异。分层解释界面可以同时提供简洁和详细的说明。此外，首次使用时提供教育性提示（例如“这些建议仅供参考！”），并针对历史冲突进行模拟练习，也能进一步帮助人工智能素养较低的版主培养批判性思维。

6.3 可推广性

本节从研究发现和实验设计两方面讨论我们研究的普适性。我们聚焦于在线权力不对称冲突中的审核偏差，并以消费者-商家冲突作为代表性场景，但研究发现也可以推广到其他情境。首先，第三方审核员倾向于支持强势方的现象并不局限于这一场景。在医疗纠纷中，仲裁者可能会因医生的专业权威而倾向于接受其解释 [8]。在教育领域，评估者可能会因教师的专业知识和机构地位而支持教师 [37]。与消费者-商家冲突类似，这些现象表明，权力不对称会普遍影响第三方审核员。因此，我们的发现有助于理解这些情境中的偏见模式。其次，人工智能生成内容的“正当化”效应也体现在其他决策场景中。例如，医生可能依赖大语言模型的输出以减轻工作负担，却忽视患者之间的差异；人力资源人员在使用大语言模型筛选简历时，则可能加剧性别和种族偏见 [84]。这些现象表明，措辞看似合理的人工智能生成内容可能会加剧误判，并为内容审核及更广泛的人机协作带来启示。

此外，我们的实验设计具有可推广性，也可应用于其他关于人类/人工智能偏见和冲突判断的研究。首先，用于考察偏见的干预方法并不局限于在线消费者-商家冲突，还可应用于职场谈判、社区治理和平台调解等场景。其次，绿野仙踪设计（由研究人员扮演“人工智能”）消除了模型输出差异的影响，使我们能够聚焦于审核员的偏见模式，而非技术限制。这为研究人们如何看待人工智能生成内容提供了一个典型框架，也有助于设计不依赖特定模型的人工智能辅助系统。

7 局限性与未来工作

本研究存在以下局限性。首先，我们以消费者-商家冲突作为权力不对称的代表性场景。虽然这一场景便于获取数据，且具有标准化解决流程，但仅采用这一种场景可能会限制研究发现的可推广性和可靠性。职场谈判、社区治理或医疗纠纷等其他代表性情境涉及多轮互动、隐性权力动态和道德考量，可能引发不同的偏见。未来研究应在更多情境中验证这些发现。此外，我们的数据来自大众点评这一中国本地生活服务平台，所有参与者均来自中国。虽然这确保了情境的一致性，但也限制了跨文化可推广性。既有研究表明，内容审核会受到文化差异的影响，包括身份特征、意识形态和语言等。 [36, 86, 119]。这些发现意味着，本研究中识别出的审核偏见未必适用于所有文化。因此，未来研究应在多种文化背景下开展，以考察本研究所揭示偏见机制的普遍性及其边界。

此外，我们的研究没有衡量参与者的技术素养水平，因此难以评估这一因素如何影响他们对人工智能的想象以及审核判断。未来研究可以考察审核员不同的技术素养水平，进一步探究这些差异如何影响他们在内容审核过程中对人工智能的信任。此外，为避免技术差异，我们采用绿野仙踪方法

由研究人员扮演“人工智能”提供建议，这不可避免地偏离了人工智能模型的真实使用场景。具体而言，不同大语言模型的输出在语气、准确性和幻觉频率方面差异显著，而且提示词的细微变化也可能导致不同的回答，从而影响用户的信任[13, 107, 109]。真实大语言模型给出的建议可能会因其表达风格和输出稳定性而改变审核员偏见的表现形式。更具权威性的表达可能会加剧偏见，而不同大语言模型之间不一致的输出则可能削弱审核员的信任。未来研究可以纳入真实的人工智能模型进行比较，并考察其对内容审核的相应影响。

8 结论

本文研究了权力不对称冲突中人类审核和人机协同审核存在的权力相关偏见。通过一项涉及 50 名参与者的混合设计实验，我们发现，两种审核模式都存在若干偏向支持强势方的权力相关偏见。尽管人工智能生成的建议缓解了人类审核中的大多数偏见，但也加剧了少数偏见。此外，在这些建议中呈现不支持对方当事方的理由，也会在一定程度上促使审核员支持弱势方。这些发现为权力不对称冲突调解机制和 AI 辅助审核系统的设计提供了若干启示。未来，我们将进一步探讨不同文化语境下审核偏见的差异，并研究能够缓解这些偏见的审核设计和人机协作策略。

致谢

本研究获得美团和中国国家自然科学基金委员会（NSFC）资助，项目编号为 62372113。Peng Zhang 任职于复旦大学计算机科学与人工智能学院。Tun Lu 任职于复旦大学计算机科学与人工智能学院、上海市数据科学重点实验室和复旦大学老龄研究院银龄哲学社会科学实验室。

参考文献

[1] Deema AlKilani. 2012. 权力：定义社会权力基础的六种基础，以及由此带来的社会独立变革。领导力百科全书 (2012), 1242–1249. [2] Saar Alon-Barkat and Madalina Busuioc. 2023. 公共部门决策中的人机交互：“自动化偏见”与对算法建议的“选择性遵循”。公共行政研究与理论杂志33, 1 (2023), 153–169.[3] Vallet Améline, Locatelli Bruno, Barnaud Cécile, Makowski David, Quispe Conde Yésica, and Haroldv Levrel. 2020. 生态系统服务治理社会网络中的权力不对称。环境科学与政策114 (2020), 329–340. [4] Shestakofsky Benjamin. 2019. 书评：《幽灵工作：如何阻止硅谷打造新的全球底层阶级》，作者 Mary L. Gray 和 Siddharth Suri。*ILRReview* 72, 5 (2019), 1283–1285. [5] Jules Brooks Blanchet, Megan E. Hillis, Yeongji Lee, Qijia Shao, Xia Zhou, Devin Balkcom, and David J. M. Kraemer. 2025. 提升在线视频的教育潜力：系统开发及关于 TikTok 舞蹈挑战的实证研究。载于论文集 *2025*年计算机系统人因大会。文章 587，19 页。[6] Nattapat Boonprakong、Benjamin Tag、Jorge Goncalves 和 Tilman Dingler。2025。《人机交互研究者如何研究认知偏差？一项范围综述》。载于 *2025* 年计算机系统人因大会论文集（*CHI '25*）。文章 473，20 页。[7] Michelle Brachman、Zahra Ashktorab、Michael Desmond、Evelyn Duester- wald、Casey Dugan、Narendra Nath Joshi、Qian Pan 和 Aabhas Sharma。2022。《人类与人工智能协作数据标注中的依赖与自动化：冲突解决》。载于 *ACM*人机交互论文集 6，CSCW2，第 321 号文章（2022），27 页。[8] Adrian Bussone、Simone Stumpf 和 Dympna O’Sullivan。2015。《解释对临床决策支持系统中信任与依赖的作用》。载于 *2015* 年国际医疗信息学会议。IEEE，160–169。[9] Matt Canute 、 Mali Jin、Hannah Holtzclaw、Alberto Lusoli、Philippa Adams、Mugdha Pandya、Maite Taboada、Diana Maynard 和 Wendy Hui Kyong Chun。2023。《在线冲突的维度：迈向对抗性的建模》。载于 计算语言学协会论文集：*EMNLP 2023*。12194–12209。[10] Guiming Hardy Chen、Shunian Chen、Ziche Liu、Feng Jiang 和 Benyou Wang。2024。《由人还是大语言模型来评判？一项关于判断偏差的研究》。载于 *2024*年自然语言处理实证方法会议论文集。美国佛罗里达州迈阿密，8301–8327。[11] XinHui Chen、Xiang Yuan、Hui Zhang、Ruixiao Zheng 和 Wanyi Wei。2025。《维持“平衡”的冲突：人工智能语音代理在临时设计团队在线协作中的主动干预策略》。载于 *2025*年计算机系统人因大会论文集（*CHI ' 25*）。文章 907，19 页。[12] Yu-Ling Chou、Hsuan-Jen Lee、Jie Tsai、En-Chi Liang 和 Yung-Ju Chang。2023。《我喜欢他们的自主性以及与我的亲近感：揭示用户对社交媒体影响者的感知吸引力》。载于 *2023*年计算机系统人因大会论文集。文章 674，19 页。[13] Michelle Cohn、Mahima Pushkarna、Gbolahan O. Olanubi、Joseph M. Moran、Daniel Padgett、Zion Mengesha 和 Courtney Heldreth。2024。《相信拟人化：考察拟人化线索对大语言模型信任度的作用》。载于 计算机系统人因大会扩展摘要集。文章 54，15 页。[14] Petru L Curseu、Patrick Kenis、Jorg Raab 和 Ulrik Brandes。2010。《通过团队约会组建高效团队》。《组织研究》 31，7（2010），873–894。[15] Robert A Dahl。1957。《权力概念》。《行为科学》 2，3（1957），201–215。[16] Valdemar Danry、Pat Pataranutaporn、Matthew Groh 和 Ziv Epstein。2025。《大语言模型的欺骗性解释比诚实解释更常使人们改变对错误信息的信念》。载于 *2025*年计算机系统人因大会论文集（*CHI '25*）。文章 933，31 页。[17] Janice Denegri-Knott。2006。《行为不端的消费者：偏离还是创新？网络上的权力斗争》。《消费者行为杂志：国际研究评论》 5，1（2006），82–94。[18] Djellel Difallah、Elena Filatova 和 Panos Ipeirotis。2018。《Mechanical Turk 工作人员的人口统计特征与动态》。载于 第*17*届*ACM*网络搜索与数据挖掘国际会议论文集（*WSDM '18*）。135–143。[19] Kaushik Divyansh、Hovy Eduard 和 Lipton Zachary。2020。《利用反事实增强数据学习真正有差异的差异》。载于 国际学习表征会议。[20] Claire Dolin、Ben Weinshel、Shawn Shan、Chang Min Hahn、Euirim Choi、Michelle L. Mazurek 和 Blase Ur。2018。《剖析在线定向与个性化背后数据推断的用户认知》。载于*2018*年计算机系统人因大会论文集（*CHI '18*）。1–12。[21] S. M. Easterbrook、E. E. Beck、J. S. Goodlet、L. Plowman、M. Sharples 和 C. C. Wood。1993。冲突实证研究综述。Springer London，1–68。[22] Upol Ehsan、Samir Passi、Q. Vera Liao、Larry Chan、I-Hsiang Lee、Michael Muller 和 Mark O Riedl。2024。《可解释人工智能中的“谁”：人工智能背景如何影响人们对人工智能解释的认知》。载于 *2024*年计算机系统人因大会论文集。文章 316，32 页。[23] Richard M Emerson。2019。《权力—依赖关系》。载于 《现代社会中的权力》。劳特利奇，48–58。[24] Xianzhe Fan、Qing Xiao、Xuhui Zhou、Jiaxin Pei、Maarten Sap、Zhicong Lu 和 Hong Shen。2025。《用户驱动的价值对齐：理解

用户对人工智能伴侣中偏见和歧视性言论的认知及应对策略》。载于《2025年计算机系统人因大会论文集（CHI ’25 910，19 页）》。[25] Franz Faul、Edgar Erdfelder、Axel Buchner 和 Albert-Georg Lang。2009。《使用 G* Power 3.1 进行统计功效分析：相关与回归分析检验》。行为研究方法 41，4（2009），1149–1160。[26] Colin Ford、Dan Gardner、Leah Elaine Horgan、Calvin Liu、a. m. tsaasan、Bonnie Nardi 和 Jordan Rickman。2017。《聊天速度 OP PogChamp：大型 Twitch 聊天中的连贯性实践》。载于《2017年计算机系统人因大会扩展摘要集（CHI EA ’17）》。858–871。[27] Kyriaki Fousiani。2022。《组织中的权力不对称、谈判与冲突管理》。《组织冲突》（2022）。[28] John RP French、Bertram Raven 和 Dorwin Cartwright。1959。《社会权力基础》。《社会权力研究》150（1959），167。[29] Katy Ilonka Gero、Zahra Ashktorab、Casey Dugan、Qian Pan、James John- son、Werner Geyer、Maria Ruiz、Sarah Miller、David R. Millen、Murray Campbell、Sadhana Kumaravel 和 Wei Zhang。2020。《合作博弈情境中的人工智能代理心智模型》。载于《2020年计算机系统人因大会论文集》。1–12。[30] Ysabel Gerrard。2020。《屏幕后：社交媒体阴影中的内容审核》。[31] Peter Glick 和 Susan T Fiske。1996。《矛盾性别歧视量表：区分敌意性别歧视与善意性别歧视》。《人格与社会心理学杂志》70，3（1996），491–512。[32] Cleotilde Gonzalez、Javier F Lerch 和 Christian Lebiere。2003。《动态决策中的基于实例的学习》。《认知科学》27，4（2003），591–635。[33] Jarod Govers、Eduardo Velloso、Vassilis Kostakos 和 Jorge Goncalves。2024。《用于在线辩论受众去极化的人工智能驱动调解策略》。载于《2024年计算机系统人因大会论文集（CHI ’24）》。文章 803，18 页。[34] James Grimmelmann。2015。《内容审核的美德》。《耶鲁法律与科技杂志》17（2015），42。[35] Ken Gu、Madeleine Grunde-McLaughlin、Andrew McNutt、Jeffrey Heer 和 Tim Althoff。2024。《数据分析师如何应对人工智能辅助？一项绿野仙踪研究》。载于《2024年计算机系统人因大会论文集（CHI ’24）》。文章 1015，22 页。[36] Oliver L. Haimson、Daniel Delmonaco、Peipei Nie 和 Andrea Wegner。2021。《保守派、跨性别者和黑人社交媒体用户遭遇不成比例的内容移除及不同的内容审核体验：边缘化与审核灰色地带》。ACM人机交互论文集 5，CSCW2，第 466 号文章（2021），35 页。[37] Bingyi Han、Simon Coghlan、George Buchanan 和 Dana McKay。2025。《谁在帮助谁？学生对高等教育课堂中人工智能与教师协作的顾虑》。ACM人机交互论文集 9，2，第 CSCW206 号文章（2025），32 页。[38] Gaole He、Gianluca Demartini 和 Ujwal Gadiraju。2025。《先规划，再执行：使用大语言模型代理作为日常助手时用户信任与团队绩效的实证研究》。载于《2025年计算机系统人因大会论文集》。文章 414，22 页。[39] Natali Helberger、Orla Lynskey、Hans-W. Micklitz、Peter Rott、Marijn Sax 和 Joanna Strycharz。2021。《欧盟消费者保护 2.0：数字消费者市场中的结构性不对称》。技术报告。信息法研究所与阿姆斯特丹传播研究学院。[40] Vali Hessamedin、Jingjun Xu David 和 Yildirim Mehmet。2015。《消费者评论冲突对购买决策准确性的影响》。载于《第21届美洲信息系统会议（AMCIS 2015）》。[41] Michael C Horowitz 和 Lauren Kahn。2024。《改变自动化偏见曲线：国家安全情境下人类与人工智能辅助决策研究》。《国际研究季刊》68，2（2024），sqae020。[42] Yoyo Tsung-Yu Hou、EunJeong Cheon 和 Malte F. Jung。2024。《人机交互中的权力》。载于《2024年ACM/IEEE国际人机交互会议论文集》。269–282。[43] Wenjian Huang、Tun Lu、Haiyi Zhu、Guo Li 和 Ning Gu。2016。《在线协作项目同行评审流程中冲突管理策略的有效性》。载于《第19届ACM计算机支持的协同工作与社会计算会议论文集（CSCW ’16）》。717–728。[44] Jevan A Hutson、Jessie G Taft、Solon Barocas 和 Karen Levy。2018。《消除欲望偏见：应对亲密关系平台上的偏见与歧视》。ACM人机交互论文集 2，CSCW（2018），1–18。[45] Angel Hsing-Chi Hwang 和 Andrea Stevenson Won。2021。IdeaBot：探究人机团队创造力中的社会促进。在《2021年计算机系统人因大会论文集》中。文章 712，共 16 页。[46] Muhammad Okky Ibrohim 和 Indra Budi。2019。印度尼西亚推特上的多标签仇恨言论与辱骂性语言检测。在《第3届在线辱骂性语言研讨会论文集》中。46–57。[47] Ridho Bramulya Ikhsan、Yudi Fernando、Anderes Gui 和 Erick Fernando。2024。在线评论的力量：探究信息不对称及其对绿色产品购买行为的影响。《国际消费者研究杂志》48，3（2024），e13050。[48] 2023年第四季度数字信任与安全指数。2024。争议数据、消费者洞察与新兴趋势。https://sift.com/index-reports-disputes-q4-2023/?utm_source=chatgpt.com。[49] Krithika Jagannath、Katie Salen 和 Petr Slovàk。2020。“（我们）可以好好谈谈……”：在受管理的《我的世界》服务器上设计以促进青少年的冲突解决技能。《ACM人机交互论文集》4，CSCW1，文章 49（2020），共 26 页。[50] Surowiecki James。2004。《群体的智慧：为什么多数人比少数人更聪明，以及集体智慧如何塑造商业、经济、社会与国家》。Doubleday & Co。[51] Shagun Jhaver、Amy Bruckman 和 Eric Gilbert。2019。内容审核中的透明度真的重要吗？Reddit 上内容移除说明发布后的用户行为。《ACM人机交互论文集》3，CSCW（2019），1–27。[52] Shagun Jhaver、Sucheta Ghoshal、Amy Bruckman 和 Eric Gilbert。2018。在线骚扰与内容审核：屏蔽列表案例。《ACM计算机与人类交互汇刊》25，2，文章 12（2018），共 33 页。[53] Jialun Aaron Jiang、Peipei Nie、Jed R. Brubaker 和 Casey Fiesler。2023。以权衡为中心的内容审核框架。《ACM计算机与人类交互汇刊》30，1，文章 3（2023），共 34 页。[54] Jialun Aaron Jiang、Peipei Nie、Jed R. Brubaker 和 Casey Fiesler。2023。以权衡为中心的内容审核框架。《ACM计算机与人类交互汇刊》30，1，文章 3（2023），共 34 页。[55] Jigsaw。2023。Perspective API。[56] Seering Joseph、Kaufman Geoff 和 Chancellor Stevie。2022。内容管理中的隐喻。《新媒体与社会》24，3（2022），621–640。[57] Seering Joseph、Wang Tony、Yoon Jina 和 Kaufman Geoff。2019。算法时代的版主参与和社群发展。《新媒体与社会》21，7（2019），1417–1443。[58] John T Jost 和 Jojanneke Van der Toorn。2012。系统正当化理论。《社会心理学理论手册》2（2012），313–343。[59] Harmanpreet Kaur、Harsha Nori、Samuel Jenkins、Rich Caruana、Hanna Wallach 和 Jennifer Wortman Vaughan。2020。解读可解释性：了解数据科学家如何使用机器学习可解释性工具。在《2020年计算机系统人因大会论文集》（CHI ’20）中。1–14。[60] Sara Kiesler、Robert Kraut、Paul Resnick 和 Aniket Kittur。2012。规范在线社区中的行为。《构建成功的在线社区：循证社会设计》。[61] Soomin Kim、Jinsu Eun、Joseph Seering 和 Joonhwan Lee。2021。面向审议式讨论的版主聊天机器人：讨论结构与讨论引导的影响。《ACM人机交互论文集》5，CSCW1，文章 87（2021），共 26 页。[62] Aniket Kittur 和 Robert E. Kraut。2010。超越维基百科：在线生产团队中的协调与冲突。在《2010年ACM计算机支持的协同工作会议论文集》中。215–224。[63] Aniket Kittur、Bongwon Suh、Bryan A. Pendleton 和 Ed H. Chi。2007。他说，她说：维基百科中的冲突与协调。在《SIGCHI计算机系统中的人为因素会议论文集》（CHI ’07）中。453–462。[64] Rob Kling。1996。《计算机化与争议：价值冲突与社会选择》。爱思唯尔。[65] Samantha Kolovson、Calvin Liang、Sean A. Munson 和 Kate Starbird。2020。美国大学体育队中的个人数据与权力不对称。《ACM人机交互论文集》4，GROUP，文章 05（2020），共 27 页。[66] Yubo Kou 和 Xinning Gui。通过情境化解释调解社群与人工智能交互：人工智能主导的内容管理案例。《ACM人机交互论文集》4，CSCW2，文章 102（2020），共 27 页。[67] Isadora Krsek、Kimi Wenzel、Sauvik Das、Jason I. Hong 和 Laura Dabbish。2022。自我说服还是被说服：考察影响用户选择隐私设置的干预措施。在《2022年计算机系统人因大会论文集》（CHI ’22）中。文章 623，

共 17 页。[68] Vivian Lai、Samuel Carton、Rajat Bhatnagar、Q. Vera Liao、Yunfeng Zhang 和 Chenhao Tan。2022。通过条件委托实现人机协作：内容审核案例研究。在《2022年计算机系统人因大会论文集》（CHI '22）中。文章 54，共 18 页。[69] John D Lee 和 Katrina A See。2004。自动化信任：为适当依赖而设计。《Human Factors》46，1（2004），50–80。[70] Kyungho Lee。2025。迈向生成式用户界面设计的工作定义。在《2025年ACM交互系统设计会议伴随出版物》中。489–495。[71] Sharon Levy、Robert E. Kraut、Jane A. Yu、Kristen M. Altenburger 和 Yi-Chia Wang。2022。理解在线对话中的冲突。在《2022年ACM Web会议论文集》（WWW '22）中。2592–2602。[72] Yaqiong Li、Peng Zhang、Hansu Gu、Tun Lu、Siyuan Qiao、Yubo Shu、Yiyang Shao 和 Ning Gu。2025。DeMod：一种集成工具，具备可解释检测与个性化修改功能，用于毒性审查。《ACM人机交互论文集》9，2，文章 CSCW061（2025年5月），共 24 页。[73] Yaqiong Li、Peng Zhang、Lin Wang、Hansu Gu、Siyuan Qiao、Ning Gu 和 Tun Lu。2025。“对你无害，对我伤人！”：从青少年视角探究有毒语言检测。ArXiv abs/2508.02094（2025）。[74] Baoxi Liu、Peng Zhang、Yubo Shu、Zhengqing Guan、Tun Lu、Hansu Gu 和 Ning Gu。2022。构建用于社交媒体文本内容审查的个性化模型。《ACM人机交互论文集》6，CSCW2，文章 499（2022年11月），共 31 页。[75] Zhuoran Lu、Patrick Li、Weilong Wang 和 Ming Yin。2025。了解人们同时受到同伴和专家影响时，基于人工智能的可信度指标所产生的影响。在《2025年计算机系统人因大会论文集》（CHI '25 817，19）中。文章页码。[76] M Lynne Markus。1994。寻找折中点：解释电子通信对工作中社交生活的负面影响。《ACM信息系统汇刊》（TOIS）12，2（1994），119–149。[77] David Martin、Benjamin V. Hanrahan、Jacki O'Neill 和 Neha Gupta。2014。成为土耳其机器人工作者。在《第17届ACM计算机支持的协同工作与社会计算会议论文集》（CSCW '14）中。224–235。[78] David Maulsby、Saul Greenberg 和 Richard Mander。1993。通过绿野仙踪法为智能代理制作原型。在《INTERACT '93与CHI '93计算机系统人因会议论文集》中。277–284。[79] Nora McDonald、Sarita Schoenebeck 和 Andrea Forte。2019。定性研究中的可靠性与评定者间信度：CSCW与人机交互实践的规范与指南。《ACM人机交互论文集》3，CSCW，文章 72 （ 2019 ），共 23 页。[80] Matthew B Miles。1994。《定性数据分析：扩展版资料手册》。千橡市（1994）。[81] Tim Miller。2023。可解释人工智能已死，可解释人工智能万岁！使用评估型人工智能进行假设驱动的决策支持。在《2023年ACM公平性、问责制与透明度会议论文集》中。333–342。[82] Shyamal Mishra 和 Preetha Chatterjee。2024。探索使用 ChatGPT 检测 GitHub 上的毒性。在《2024年ACM/IEEE第44届软件工程国际会议：新理念与新兴成果论文集》（ICSE-NIER'24）中。6–10。[83] Vikram Mohanty、Jude Lim 和 Kurt Luther。2025。表象之下是什么？探究底层人工智能模型更新对融入人工智能的系统的影响。在《2025年计算机系统人因大会论文集》中。文章 539，共 21 页。[84] Dena F. Mujtaba 和 Nihar Ranjan Mahapatra。2024。人工智能驱动招聘中的公平性：挑战、指标、方法与未来方向。ArXiv abs/2405.19699（2024）。[85] Isabel Munoz、Michael Dunn、Steve Sawyer 和 Emily Michaels。2022。平台中介市场、在线自由职业者与解构的身份。《ACM人机交互论文集》6，CSCW2（2022），1–24。[86] Gianluca Nogara、Francesco Pierri、Stefano Cresci、Luca Luceri、Petter Törnberg 和 Silvia Giordano。2025。毒性偏差：Perspective API 将德语误判为更具毒性。《AAAI网络与社交媒体国际会议论文集》19，1（2025），1346–1357。[87] Gilbert N. Nyaga、Daniel F. Lynch、Donna Marshall 和 Eamonn Ambrose。2013。涉及强势伙伴的二元关系中的权力不对称、适应与协作。《供应链管理杂志》49，3（2013），42–65。

[88] Ben Hill Passmore。2015。组织中的归纳推理、信息对称性与权力不对称。载于信息科学与技术百科全书（第三版）。IGI Global Scientific Publishing，808–818。[89] Ken Petress。2003。权力：定义、类型、描述、示例与启示。*http://uthscsa.edu/gme/documents/powerdefinitionstypologyexamples. pdf*（2003）。[90] Richard E Petty 和 John T Cacioppo。1986。说服精细加工可能性模型。载于实验社会心理学进展。第19卷。爱思唯尔，123–205。[91] Snehal Prabhudesai、Ananya Prashant Kasi、Anmol Mansingh、Anindya Das Antar、Hua Shen 和 Nikola Banovic。2025。“这里是 GPT 做出了选择，而每个选择都可能存在偏见”：学生如何通过终端用户审计活动批判性地参与大语言模型。载于*2025*年计算机系统人因大会论文集（*CHI '25*）。文章1015，共23页。[92] Gerald Peter Quinn 和 Michael J Keough。2002。生物学家的实验设计与数据分析。剑桥大学出版社。[93] Emely Rosbach、Jonas Ammeling、Sebastian Krügel、Angelika Kießig、Alexis Fritz、Jonathan Ganz、Chloé Puget、Taryn Donovan、Andrea Klang、Maximilian C. Köller、Pompei Bolfa、Marco Tecilla、Daniela Denk、Matti Kiupel、Georgios Paraschou、Mun Keong Kok、Alexander F. H. Haake、Ronald R. de Krijger、Andreas F.-P. Sonnen、Tanit Kasantikul、Gerry M. Dorrestein、Rebecca C. Smedley、Nikolas Stathonikos、Matthias Uhl、Christof A. Bertram、Andreas Riener 和 Marc Aubreville。2025。“两个错误不会成就一个正确”——考察计算病理学中人机协作期间的确认偏误与时间压力作用。载于*2025*年计算机系统人因大会论文集（*CHI '25*）。文章528，共18页。[94] Akrati Saxena 和 Harita Reddy。2022。在线众包问答平台和百科全书中的用户角色认同：一项综述。计算社会科学杂志 5，1（2022），285–317。[95] Tobias Schnabel、Saleema Amershi、Paul N Bennett、Peter Bailey 和 Thorsten Joachims。2020。更透明的界面对个性化推荐中行为的影响。载于第*43*届国际*ACM SIGIR*信息检索研究与发展会议论文集。991–1000。[96] Jakob Schoeffer、Maria De-Arteaga 和 Niklas Kühl。2024。人机决策中的解释、公平性与适当依赖。载于*2024*年计算机系统人因大会论文集（*CHI '24*）。文章836，共18页。[97] Joseph Seering。2020。重新思考自我审核：研究在支持在线内容审核社区型模式中的作用。*ACM*人机交互论文集 4，CSCW2，文章107（2020），共28页。[98] Farhana Shahid。2024。促进文化意识内容审核的人机协作。载于*2024*年计算机支持的协同工作与社会计算会议伴随出版物（*CSCW Companion '24*）。9–12。[99] Omar Shaikh、Valentino Emil Chai、Michele Gelfand、Diyi Yang 和 Michael S. Bernstein。2024。Rehearsal：模拟冲突以教授冲突解决。载于*2024*年计算机系统人因大会论文集（*CHI '24*）。文章920，共20页。[100] Amit Sheth、Valerie L. Shalin 和 Ugur Kursuncu。2022。定义和检测社交媒体中的毒性：语境与知识是关键。神经计算 490（2022），312–318。[101] Yubo Shu、Tian Tian、Peng Zhang、Hansu Gu、Yaqiong Li、Yiyang Shao、Tun Lu 和 Ning Gu。2025。SoulSearch：应用启发式优化，通过个性化人类与大语言多模态模型协作增强文生图。中国科学：信息科学 68，10（2025）。[102] George Spivack 和 Myrna B Shure。2014。人际认知问题解决（ICPS）：一项能力建设型一级预防计划。载于国家心理健康协会。劳特利奇，151–178。[103] Tim Squirrell。2019。平台辩证关系：Reddit 上志愿版主与终端用户之间的关系。新媒体与社会 21，9，2019，1910–1927。[104] Yla R. Tausczik 和 James W. Pennebaker。2013。利用实时语言反馈改善团队合作。载于*SIGCHI*计算机系统中的人为因素会议论文集（*CHI '13*）。459–468。[105] Leanne Ten Brinke 和 Dacher Keltner。2022。权力理论：获取和维持权力的感知策略。人格与社会心理学杂志 122，1（2022），53。[106] Ryosuke Ueda、Koh Takeuchi 和 Hisashi Kashima。2022。缓解众包标签聚合中的观察偏差。载于*2022*年第*26*届国际模式识别会议（*ICPR*）。1171–1177。

表4：人工审核组和人机协同审核组的组间配对 t 检验结果。* 表示 $p < 0.001$，** 表示 $p < 0.01$，* 表示 $p < 0.05$。Cohen's d_z 报告因变量的效应量。**

权力表现	人工审核组				人机协同审核内容审核组			
	sd	t	p	d_z	sd	t	p	d_z
正当主张	1.181	2.284	**0.025**(*)	0.237	0.884	5.077	**0**(***)	0.521
权威引用	1.346	1.897	0.061	0.194	0.972	3.08	**0.003**(**)	0.309
惩罚威胁	1.12	4.375	**0**(***)	0.461	1.058	2.015	**0.047**(*)	0.205
补偿	1.303	3.759	**0**(***)	0.380	1.008	3.718	**0.001**(**)	0.358
专家知识	1.094	2.357	**0.021**（*）	0.243	1.024	1.233	0.22	0.125
群体偏好	1.261	1.893	0.061	0.192	0.974	1.032	0.305	0.104
陈述顺序	1.284	1.966	0.053	0.225	1.144	0.486	0.629	0.054
表达语气	1.216	1.215	0.228	0.130	1.021	1.187	0.238	0.119
选择陷阱	1.369	0.377	0.707	0.039	1.03	1.577	0.118	0.16
长度差异	1.289	2.601	**0.011**(*)	0.268	1.016	2.473	**0.012**(*)	0.248

表5：不同年龄组中初始版本与扰动版本的配对t检验结果（采用与表4相同的表示形式）。

权力表现形式	年轻组				年长组			
	标准差	t	p	d_z	标准差	t	p	d_z
正当主张	1.181	3.224	**0.002**（**）	0.423	1.284	1.382	0.176	0.234
权威引用	1.443	0.809	0.422	0.111	1.366	2.088	**0.043**(*)	0.322
惩罚威胁	1.153	4.045	**0**(***)	0.572	1.136	2.532	**0.016**(*)	0.416
补偿	1.303	1.982	0.052	0.263	1.284	3.169	**0.003**(**)	0.478
专家知识	1.421	2.226	**0.030**(*)	0.297	1.225	1.847	0.072	0.288
群体偏好	1.363	0.791	0.432	0.107	1.253	1.598	0.120	0.278
陈述顺序	1.242	1.675	0.102	0.268	1.306	1.756	0.096	0.403
表达语气	1.421	1.027	0.310	0.150	1.093	1.391	0.172	0.223
选择陷阱	1.38	1.558	0.126	0.216	1.428	0	1.000	0
长度差异	1.243	2.729	**0.008**(**)	0.349	1.238	2.640	**0.013**(*)	0.467

表6：性别组之间初始版本与扰动版本的配对t检验结果（使用与表4相同的表示）。

权力表现	男性组				女性组			
	sd	t	p	d_z	sd	t	p	d_z
正当主张	1.27	1.147	0.256	0.152	1.212	3.236	**0.003**（**）	0.518
权威引用	1.327	2.234	**0.030**(*)	0.301	1.533	0.866	0.392	0.135
惩罚威胁	1.176	2.418	**0.020**(*)	0.353	1.209	3.125	**0.004**(**)	0.514
补偿	1.406	3.289	**0.002**(**)	0.411	1.212	1.071	0.291	0.174
专家知识	1.319	2.785	**0.007**（**）	0.372	0.901	1.387	0.173	0.217
群体偏好	1.387	0.550	0.585	0.076	1.223	0.970	0.338	0.153
陈述顺序	1.414	1.414	0.166	0.236	1.265	1.318	0.201	0.275
表达语气	1.136	1.257	0.215	0.180	1.372	0.928	0.362	0.182
选择陷阱	1.59	0.424	0.673	0.057	1.271	0.905	0.371	0.149
长度差异	1.252	2.446	**0.018**(*)	0.357	1.265	0.577	0.567	0.094

表 7：配对t检验结果，比较人工审核组与人机协同审核组（采用与表4相同的表述方式）。

权力表现	sd	t	p	d_z
初始版本	0.76	-0.449	0.654	-0.045
正当主张	0.815	-0.523	0.602	-0.056
权威引用	0.979	0.106	0.916	0.011
惩罚威胁	0.886	2.520	**0.014**(*)	0.270
补偿	0.932	1.626	0.107	0.164
专家知识	0.839	0.663	0.509	0.069
群体偏好	1.022	1.165	0.247	0.119
陈述顺序	1.096	1.026	0.309	0.128
表达语气	1.006	0.998	0.321	0.108
选择陷阱	1.093	-0.716	0.476	-0.075
长度差异	0.896	1.022	0.309	0.106

[107] Cunxiang Wang、Xiaoze Liu、Yuanhao Yue、Qipeng Guo、Xiangkun Hu、Xiangru Tang、Tianhang Zhang、Cheng Jiayang、Yunzhi Yao、Xuming Hu、Zehan Qi、Wenyang Gao、Yidong Wang、Linyi Yang、Jindong Wang、Xing Xie、Zheng Zhang 和 Yue Zhang。2025。大语言模型中的事实性综述。计算综述 58，1，文章13（2025），共37页。[108] Panpan Wang、Qian Huang 和 Robert M. Davison。2020。数字影响者如何影响社交商务意愿？社交权力与满意度的作用。信息技术与人 34，3（2020），1065–1086。[109] Joel Wester、Tim Schrills、Henning Pohl 和 Niels van Berkel。2024。“作为一个 AI语言模型，我不能”：探究大语言模型对用户请求的拒绝。载于*2024*年计算机系统人因大会论文集。文章979，共14页。[110] David Wicks。2017。质性研究者的编码手册。组织与管理中的质性研究：国际期刊 12，2（2017），169–170。[111] Donghee Yvette Wohn。2019。Twitch 微社区中的志愿版主：他们如何参与、承担哪些角色，以及经历何种情绪劳动。载于*2019*年计算机系统人因大会论文集（*CHI '19*）。1–13。[112] Qing Xiao、Yuhang Zheng、Xianzhe Fan、Bingbing Zhang 和 Zhicong Lu。2025。让我们共同影响算法：数百万粉丝如何构建对算法的集体理解并组织协同算法行动。载于*2025*年计算机系统人因大会论文集。文章1115，共19页。[113] Mingzhe Yang、Hiromi Arai、Naomi Yamashita 和 Yukino Baba。2024。公平机器引导如何提升有偏见者的公平决策。载于*2024*年计算机系统人因大会论文集（*CHI '24*）。文章285，共18页。[114] Jiayi Ye、Yanbo Wang、Yue Huang、Dongping Chen、Qihui Zhang、Nuno Moniz、Tian Gao、Werner Geyer、Chao Huang、Pin-Yu Chen、Nitesh Chawla 和 Xiangliang Zhang。2025。公正还是偏见？量化大语言模型即评判者中的偏见。载于表征学习国际会议，第2025卷。102351–102390。[115] Xin Ye 和 Lionel P Robert。2025。安保机器人权力与接受度：探索 French 和 Raven 的五种权力形式。载于*2025*年第*20*届*ACM/IEEE*国际人机交互会议（*HRI*）。IEEE，1731–1736。[116] Lu Yingchao、Chang Hailing 和 Zhou Yadan。2025。消费者投诉对企业数字化转型的影响。国际经济与金融评论 101（2025），104203。[117] Zehui Yu、Lukas Otto、Dennis Assenmacher 和 Claudia Wagner。2024。人工智能辅助审核对个人和群体影响的系统性综述。人机传播 9（2024），167–188。[118] Xuechen Zhang、Changyang He、Peng Zhang、Hansu Gu、Ning Gu、Qi Shen、Zhan Hu 和 Tun Lu。2025。RemiHaven：整合“本地”和“外地”同伴，为老年漂泊者提供个性化回忆支持。载于*2025*年计算机系统人因大会论文集。文章1034，共20页。[119] Andrew Zolides。2021。性别审核与审核性别：Twitch 社区准则中的性内容政策。新媒体与社会 23，10（2021），2999–3015。

A 附录