



新闻界

Journalism and Mass Communication

ISSN 1007-2438, CN 51-1046/G2

《新闻界》网络首发论文

题目：从沉默围观到积极回应：数字正义视角下社交媒体旁观者干预机制研究
作者：陈雪娇，赵一鸣，王琪方，吴晔
DOI：10.15897/j.cnki.cn51-1046/g2.20260920.001
网络首发日期：2026-09-21
引用格式：陈雪娇，赵一鸣，王琪方，吴晔. 从沉默围观到积极回应：数字正义视角下社交媒体旁观者干预机制研究[J/OL]. 新闻界.
<https://doi.org/10.15897/j.cnki.cn51-1046/g2.20260920.001>



网络首发：在编辑部工作流程中，稿件从录用到出版要经历录用定稿、排版定稿、整期汇编定稿等阶段。录用定稿指内容已经确定，且通过同行评议、主编终审同意刊用的稿件。排版定稿指录用定稿按照期刊特定版式（包括网络呈现版式）排版后的稿件，可暂不确定出版年、卷、期和页码。整期汇编定稿指出版年、卷、期、页码均已确定的印刷或数字出版的整期汇编稿件。录用定稿网络首发稿件内容必须符合《出版管理条例》和《期刊出版管理规定》的有关规定；学术研究成果具有创新性、科学性和先进性，符合编辑部对刊文的录用要求，不存在学术不端行为及其他侵权行为；稿件内容应基本符合国家有关书刊编辑、出版的技术标准，正确使用和统一规范语言文字、符号、数字、外文字母、法定计量单位及地图标注等。为确保录用定稿网络首发的严肃性，录用定稿一经发布，不得修改论文题目、作者、机构名称和学术内容，只可基于编辑规范进行少量文字的修改。

出版确认：纸质期刊编辑部通过与《中国学术期刊（光盘版）》电子杂志社有限公司签约，在《中国学术期刊（网络版）》出版传播平台上创办与纸质期刊内容一致的网络版，以单篇或整期出版形式，在印刷出版之前刊发论文的录用定稿、排版定稿、整期汇编定稿。因为《中国学术期刊（网络版）》是国家新闻出版广电总局批准的网络连续型出版物（ISSN 2096-4188，CN 11-6037/Z），所以签约期刊的网络版上网络首发论文视为正式出版。

从沉默围观到积极回应: 数字正义视角下 社交媒体旁观者干预机制研究¹⁾

陈雪娇 赵一鸣 王琪方 吴晔

摘要 对社交媒体上针对特定群体的贬损性评论的治理不仅依赖于平台规制, 也与普通用户能否参与公共纠偏密切相关。本文从数字正义视角出发, 采用在线情境实验法, 考察了六类反对策略对旁观者感知共情、感知责任和支持性干预行为的影响, 并检验话题类型、群体身份与确认偏误在其中的调节作用。研究发现, 不同反对策略会显著影响旁观者反应, 其中礼貌纠偏策略总体上更能激发旁观者的积极干预行为; 不同的群体身份和确认偏误水平会影响反对策略的实际作用效果; 反对策略对干预行为的影响也会通过共情与感知责任的链式中介路径间接实现。研究表明, 普通用户的旁观者干预, 是数字正义在日常互动中的具体实践, 也构成平台公共规范生成与秩序涌现的重要基础。

关键词 数字正义; 旁观者干预; 群体身份; 确认偏误

中图分类号 G206 **文献标识码** A

作者 陈雪娇, 北京师范大学新闻传播学院讲师, 北京 100875; 北京师范大学计算传播学研究中心讲师, 广东珠海 519087。赵一鸣, 北京师范大学新闻传播学院博士研究生, 北京 100875。王琪方, 北京师范大学新闻传播学院硕士研究生, 北京 100875。吴晔, 北京师范大学新闻传播学院教授, 北京 100875; 北京师范大学计算传播学研究中心教授, 广东珠海 519087

一、研究背景: 数字正义视角下的旁观者干预行为

社交媒体中的公共讨论并不总是沿着理性协商的方向展开, 具有贬损性、攻击性或污名化倾向的评论表达时常夹杂其中。本文将此类表达概括为群体贬损性评论, 主要指针对基于兴趣偏好、学科背景或知识立场形成的特定群体, 通过负面标签、贬低性概括或攻击性措辞, 对其能力、价值或社会形象作出否定性评价的评论表达。被攻击者或其他用户针对此类评论作出的公开反驳或规劝, 则将其概括为“反对意见”。

网络旁观者是目睹伤害性互动但未直接卷入其中的用户, 相关研究综述显示, 网络欺凌情境中的潜在旁观者众多, 但多数人选择忽略事件或不采取行动^[1]。然而, 旁观者的沉默并不意味着其在社交媒体互动中无足轻重, 相

1) 基金项目: 广东省哲学社会科学规划青年项目“生成式 AI 驱动下粤港澳大湾区舆情传播特征与认知安全防护研究”(GD26YXW01); 北京师范大学人才培养质量提升专项计划支持项目。

反,作为数量庞大且处于可见位置的参与者,旁观者往往是影响冲突走向的重要主体^[2]。当旁观者通过反驳或声援等方式介入线上互动时,其行为就构成了数字公民参与社区治理的具体形式^[3],这种公开可见的反制回应能够在一定程度上减轻相关言论造成的伤害感知^[4],相关反制策略还可能促进原帖删除、减少其他用户的点赞,从而削弱相关言论的可见性与放大效应^[5]。

当前,平台治理是化解网络冲突的重要途径,其中删帖、禁言等平台规制主要作用于违规内容,而用户的反驳和声援则能在互动现场形成社区内部的规范反馈。换言之,网络空间中有害表达的治理不仅属于平台管理范畴,也关乎社区成员能否主动参与规范维护^[6],在更广泛的网络内容治理中,政府、平台、社会组织 and 普通网民逐渐形成多元主体协同参与的治理格局^[7]。

旁观者是否愿意发声,并不只是取决于其对贬损性评论的价值判断,也受到反对方式本身的影响^[8]。现有研究虽已开始关注不同反对策略的传播效果,但这些策略能否进一步激活旁观者干预,以及不同群体对同类策略是否存在差异性反应,仍有待深入讨论。基于此,本文从数字正义视角出发,考察社交媒体冲突场景中不同反对策略对旁观者干预行为的影响,并进一步从用户参与和社区规范生成的层面理解旁观者发声的条件及其治理意义,以期为社交媒体中的用户参与式治理提供依据,并丰富数字空间规范建构的相关讨论。

二、文献综述

(一) 数字正义:从制度规制到互动中的规范修复

旁观者在社交媒体上针对贬损性评论所作出的反对或声援,可以被理解为一种“数字正义”实践。既有数字正义研究主要关注数字技术如何改善司法可及性、纠纷解决与纠纷预防^[9],与之相关的数据正义研究则将公平问题延伸至个体如何在数据化过程中被看见、被代表和被对待^[10],这一视角更强调网络空间中的分布式责

任^[11],即将普通用户纳入数字秩序的维护过程,关注其如何感知他人遭遇的不公、确认自身的介入责任,并将规范判断转化为可见行动。

一方面,数字正义体现为可见性正义。社交媒体中的观看、点赞、评论、转发和举报等行为都会参与可见性的配置^[12],进而影响不同言论的传播范围。因此,旁观者对反对性回应的支持,既是个人的态度表达,也可能通过平台的可见性机制相对削弱贬损性评论的影响、提升纠偏性表达的显著性,进而强化对群体贬损言论的规范约束。

另一方面,数字正义还体现为参与正义,即普通用户能否作为行动主体参与公共讨论秩序的形成与维护。有学者以“参与平等”作为理解正义的重要标准,强调公共规则应使成员能够作为平等主体参与社会互动^[13]。据此,旁观者干预可以被视为一种参与式规范维护实践,其集体回应还会向其他成员传递何种表达可被接受的信号,从而参与社区规范的形成与巩固^[14]。

基于上述分析,本文将数字正义界定为一个由规范识别、责任承担与可见行动共同构成的动态过程。具体而言,不同反对策略通过相应的机制参与社交媒体中的规范维护;感知共情反映旁观者对被攻击者受损处境的理解;感知责任反映其参与公共纠偏的责任认知;点赞、评论和分享意愿进一步表征旁观者是否愿意将这种规范判断转化为公开可见的支持行动。

(二) 旁观者干预的心理基础:共情驱动与责任激活

旁观者干预是指个体在目睹他人遭受不公正对待或处于需要帮助的情境时,所采取的积极回应行为。拉塔内(Bibb Latané)和达利(John M. Darley)提出的旁观者干预模型指出,个体采取干预通常要经历注意事件、判断情境性质、承担干预责任、评估自身帮助能力,并最终做出行动决策等连续心理过程^[15]。

首先,共情是旁观者干预的重要心理基础,能够让旁观者理解他人的受损处境,并增强其提供帮助的动机。一项元分析结果表明,共情与亲社会行为呈显著正相关^[16];网络欺凌研究也

发现,共情水平较高的旁观者更可能采取干预行为^[17]。感知责任进一步影响这种情绪反应能否转化为实际干预。但旁观者并不总会主动承担干预责任,当情境中存在其他旁观者时,个体更容易认为干预会由他人完成,从而降低自身介入的可能性^[18]。这一机制在网络场景中同样存在,旁观者数量与平台匿名性都会影响其主动干预倾向^[19]。从参与正义的角度看,感知责任体现了旁观者对自身公共参与角色的确认,并促使其将对贬损性评论的不认同转化为公开回应。

因此,共情使旁观者识别介入的必要性,感知责任则推动其将情绪理解转化为支持行动,二者的连续作用推动旁观者由沉默围观转向积极回应,也体现了参与正义由情境理解、责任担当向实际参与逐步转化的过程。

(三) 社交媒体上的反对策略:群体贬损性评论回应的类型与作用

反对策略是指用户在社交媒体互动中针对群体贬损性评论作出的公开回应,不同策略虽然都表达对相关言论的不认可,但其表达方式会影响旁观者对反对行为的判断与支持。结合前人研究^{[20][21]},本文将社交媒体平台上反对群体贬损性评论的作用机制概括为四类:内化机制、惩罚机制、群体边界维护机制和情绪调节机制。

1. 内化机制:通过情感与理性诉求推动规范认同

内化机制是指通过价值引导,使个体将反对群体贬损性评论视为应遵循的公共规范。当规范获得个体的内在认同时,即使缺乏外部监督,人们也可能出于内在动机自觉遵循相关规则^[22]。这一过程通常通过两条路径实现,即呼吁良知与呼吁逻辑,它们分别从道德情感和理性判断两个层面推动规范内化。

呼吁良知策略通过唤起道德情感与同理心,使旁观者意识到贬损性评论可能对他人造成的伤害。相关实证研究表明,带有共情诉求的反对言论能够减少仇恨表达,并提高原始仇恨言论被删除的可能性^[23];呼吁逻辑策略主要通过

事实和论据指出不当言论的问题,引导旁观者从理性层面重新评估其正当性^[24]。在纠正错误信息时,基于逻辑和论据的解释回应更能降低错误认知,达到更好的传播效果^[25]。

2. 惩罚机制:以后果警示强化规范约束

惩罚机制通过声誉施压或道德谴责提高越界表达的社会成本,使发言者因预期负面后果而调整行为^[26]。威胁声誉策略是该机制的主要表现之一,例如,“你这样的言论只暴露了自己的无知和偏见”之类的评论,并不着重讨论观点本身,而是通过贬损发言者的社会形象,强化其对越界表达后果的感知。但这类策略也可能激发更强的防御与冲突^[27]。

3. 群体边界维护机制:以成员资格界定强化规范秩序

群体边界维护机制主要通过成员资格或规范判断标示社区可接受的表达范围。这与社会身份理论相一致,即个体往往借助所属群体来界定自我,并倾向于通过维护共同规范来维持积极的群体形象^[28]。群体除名策略主要通过要求发表不当言论的用户退出互动空间、删除评论等方式来否定其成员资格,可以被视为边界维护机制中的硬性规范;礼貌纠偏策略通常以较克制的方式提示相关言论已经越过规范边界^[29],如指出某一言论“属于骚扰”或“是虚假信息”,可以被视为边界维护机制中的软性规范。

4. 情绪调节机制:以低对抗表达缓冲防御反应

情绪调节机制主要通过降低互动中的对抗张力,减少旁观者对反对信息的防御。幽默策略作为一种以情绪调节为导向的表达方式能够改善受众对传播者的情感评价^[30],积极幽默还可以调节负面情绪^[31],并通过降低愤怒减少网络互动中的不文明回应^[32]。

总体而言,上述四类机制体现了数字正义实践中不同的规范维护路径。内化机制促进规范理解与认同,惩罚机制强化违规的声誉压力,群体边界维护机制重申社区互动规则,情绪调节机制缓和对抗并保留沟通空间。不同策略表达的规范依据和互动压力并不相同,因而可能

引发旁观者不同程度的共情、责任承担与可见支持。基于此,本文提出:

研究问题 1: 作为不同的规范修复方式,六种反对策略是否会对旁观者的感知共情、感知责任与支持性干预行为产生显著差异?若存在差异,哪类反对策略更能推动旁观者参与公共纠偏?

(四) 认知偏差的调节作用: 确认偏误与内群体偏好对反对策略的影响

在探讨何种反对策略更能激发旁观者干预时,还需进一步考虑接收者的认知偏差。认知偏差是指个体在判断与决策过程中,由于认知资源有限,或受到态度立场与动机倾向的影响,导致判断偏离客观事实^[33]。迪马拉(Evanthia Dimara)等人从任务视角对认知偏差进行了分类^[34],结合本文关注的旁观者信息加工过程,确认偏误主要影响其对反对信息的接受与解释,内群体偏好则可能使其支持判断呈现出群体倾向。

1. 确认偏误的影响

确认偏误是指个体在搜寻、解释和评估信息时,更倾向于接受支持既有信念的证据,而忽视、质疑甚至反驳与原有立场不一致的信息^[35]。确认偏误水平较高的个体,更容易选择性接收信息,并对异见内容表现出怀疑或排斥;确认偏误水平较低的个体则更可能对不同观点保持开放评估。社交媒体上的争议性话题更容易强化这种偏差,因为在个性化信息流中,用户经常接触与自身立场相近的内容,并在观点相似的人群中互动,从而进一步巩固原有看法^[36]。已有研究表明,当更正信息与受众既有政治态度相冲突时,立场鲜明的受众更可能表现出抵抗^[37]。因此,不同确认偏误水平的旁观者可能对反对策略产生差异化反应。基于此,本文提出:

假设 1: 确认偏误会调节反对策略的效果,不同反对策略在高、低确认偏误个体中的(a)感知共情、(b)感知责任和(c)支持性干预行为存在差异。

2. 内群体偏好的影响

内群体偏好是指个体通常对同属一群体的成员持更积极的评价或资源分配,并可能相对

不利地对待外群体成员^[38]。在社交媒体环境中,群体身份通常建立在价值观、兴趣偏好、话题立场或与事件当事人的相似性之上。已有研究表明,当个体将自己与受害者归为同一群体时,更容易产生共情与帮助意愿^[39],共享群体身份也能够提高紧急情境中的实际帮助行为^[40]。基于此,本文提出:

假设 2: 群体身份会调节反对策略的效果,不同反对策略在内群体与外群体情境中的(a)感知共情、(b)感知责任和(c)支持性干预行为存在差异。

旁观者干预模型指出,个体采取帮助行为通常经历从情境理解、责任承担再到行动实施等心理过程^[41]。在此过程中,共情有助于旁观者识别他人的受损处境,感知责任则推动其将这种理解转化为干预行为。基于此,本文提出:

假设 3: 不同反对策略通过影响旁观者的感知共情,进而影响其感知责任与支持性干预行为,其中,感知共情与感知责任发挥链式中介作用。

三、研究方法

本研究采用在线情境实验法,其中,反对策略类型(呼吁良知 vs. 呼吁逻辑 vs. 威胁声誉 vs. 礼貌纠偏 vs. 群体除名 vs. 幽默)为被试间操纵变量,话题类型(追星 vs. 大学文科 vs. 中医)为被试内重复测量因素,并将群体身份(内群体 vs. 外群体)和确认偏误(高 vs. 低)纳入分析。

(一) 被试

本研究通过腾讯问卷与见数平台的随机抽样服务招募被试,该服务依托平台样本库开展随机抽样,以尽可能降低抽样偏差。最终共纳入有效被试 651 人。每名被试者均完成三个话题情境下的测量,共形成 1953 条话题层观测记录。剔除标准包括未认真阅读实验材料、作答时间过短、注意力检测未通过及明显同质化作答。有效样本中,女性 338 人(51.92%),男性 313 人(48.08%);平均年龄为 39.14 岁(SD=12.15)。

(二) 实验材料与实验流程

研究构建模拟社交媒体页面“深潜”,页



图 1 实验材料示例

注：示例图中的反对策略均为呼吁良知策略，由左到右分别为追星话题、文科话题、中医话题。

面包括主帖、群体贬损性评论及反对性回复，尽可能贴近用户在真实社交媒体平台中的阅读情境（图 1）。为覆盖不同类型的争议情境，研究选取追星、大学文科和中医三个话题，并分别设置一条针对相关群体的固定贬损性评论。依据既有分类研究，研究设置六种反对策略，并将被试随机分配至其中一种策略条件（表 1）。被试依次浏览三个话题情境，并在阅读每则材料后填写相关量表。

（三）量表与测量

感知共情（Cronbach's $\alpha = 0.939$ ）。参照相关研究^[42]编制中文版感知共情量表，共 6 个题项，如“当我观看此评论时，我感受到了与反对者相同的情绪”。题项均采用 7 点李克特量表计分（1 = 完全不同意，7 = 非常同意）。

感知责任（Cronbach's $\alpha = 0.896$ ）。参照费舍尔（Peter Fischer）等人^[43]和奥伯迈尔（Magdalena Obermaier）^[44]的研究编制中文版感知责任量表，共 3 个题项，如“看到这种冒犯言论时，我认为我有责任去反

表 1 文科话题下攻击者言论与六种反对策略示例

攻击者言论：现在的大学文科专业，不就是培养一群高学历的废物吗？学的都是些空洞无用的理论，毕业就等于失业。拿着父母的钱读这种专业，跟啃老有什么区别？纯属浪费社会资源。		
反对策略	反对内容	
内化机制	呼吁良知策略（情感偏向）	不是每个人都想当工程师或程序员。有些人就是发自内心地热爱历史、文学和哲学。大学不只是职业培训所，更是安放个人志趣和灵魂的地方。用“有没有用”来衡量一个年轻人的梦想，太残忍了。
	呼吁逻辑策略（理性偏向）	把文科等同于没有价值，忽视了现代社会对法律、传播、教育和公共治理等专业能力的持续需求。AI 可以替代部分技术流程，但难以替代复杂沟通、价值判断与跨文化协调能力。
惩罚机制	威胁声誉策略	你这番言论不仅暴露了你的无知和偏见，随意攻击他人也显示出你缺乏做人的基本素养。
群体边界维护机制	礼貌纠偏策略（软性规范）	文科生就业确实面临一些挑战，但这并不意味着文科没有价值。不同学科各有作用，没必要用贬低性的方式评价别人。
	群体除名策略（硬性规范）	你的学科歧视言论真的让人不适，这里不欢迎你这种引战的人。
情绪调节机制	幽默策略	随便你怎么说，反正我左耳进右耳出。

对”。题项均采用 7 点李克特量表计分（1 = 完全不同意，7 = 非常同意）。

确认偏误（Cronbach's $\alpha = 0.833$ ）。用于测量个体在接收与自身立场不一致的信息时的开放评估倾向。题项改编自主动开放思维量表（Actively Open-Minded Thinking, AOT-7）^[45]以及 AOT-17 量表^[46]。所有题项均采用 7 点李克特量表计分（1 = 完全不同意，7 = 非常同意），题项均为正向的开放思维表述，对各题进行反向计分后取平均值，得分越高表示确认偏误越强。

内外群体划分。用于识别被试在不同话题情境下的群体归属。本研究设置 6 个题项测量其对相关话题的态度，每类话题各对应 2 个题项。如在追星话题中，题项包括“如果我喜欢一个偶像，我愿意花时间或金钱参与与他 / 她相关的活动”等。所有题项均采用 6 点李克特量表计分（1 = 完全不同意，6 = 非常同意）。

支持性干预行为意愿。采用一道多选题测量“请您设想自己正在浏览该社交媒体平台，您会采取哪些行为”，选项包括“点赞”“评论支持”“评论反对”“分享支持”“分享反对”“举报”以及“什么都不会做”，均采用二元编码（1 = 选择，0 = 未选择）。本文将针对反对性回复的“点赞”“评论支持”和“分享支持”作为结果变量。其余选项不纳入分析，仅用于模拟真实社交媒体中的行为选择。

（四）数据分析方法

本研究使用 R 进行数据分析。针对三个话题下的重复测量数据，使用 lme4 包拟合以被试为随机截距的混合效应模型，将反对策略和话题类型设为固定效应。研究问题 1 进一步加入“反对策略 $\times$ 话题类型”交互项，假设 1 和假设 2 分别加入反对策略与确认偏误、群体身份的交互项。固定效应显著性检验通过 lmerTest 包完成，事后比较采用 emmeans 包计算估计边际均值，并进行 Tukey 校正。假

设 3 的链式中介效应使用 lavaan 包检验，其中，支持性干预行为由“点赞”“评论支持”和“分享支持”三个指标取均值得到。考虑到同一被试在三个话题下重复作答，Bootstrap 检验以被试为抽样单位重复抽样 5000 次，并以 95% 置信区间是否包含 0 判断间接效应的显著性。

四、研究结果

随机分配后，六种反对策略组在年龄 [$F(5, 645) \leq 1.19, p_s \geq .312$] 和性别 [$\chi^2(5) \leq 6.78, p_s \geq .237$] 上均无显著差异。。研究另行招募 28 名未参加主实验的被试进行操纵检验，共获得 84 次评价。线性混合效应模型结果显示，反对策略在礼貌性、攻击性、逻辑性、幽默性、共情诉求、规范清晰度、声誉威胁和群体排斥等八项感知指标上的主效应均显著（ $p_s \leq .012$ ），且各策略在预期特征维度上的得分均显著高于其他策略，表明六类材料具有良好的区分度，操纵总体有效。

为回答研究问题 1，本文检验了不同反对策略对旁观者感知共情、感知责任及支持性干预行为的影响差异。结果显示，反对策略在感知共情 [$F(5, 1561.38) = 10.90, p < .001$]、感知责任 [$F(5, 1480.36) = 3.52, p = .004$]、点赞意愿 [$F(5, 1561.12) = 5.68, p < .001$]、评论意愿 [$F(5, 1550.37) = 6.05, p < .001$]和分享意愿 [$F(5, 1532.92) = 3.83, p = .002$]上均存在显著主效应。

事后比较结果表明，在感知共情方面，礼

表 2 六种反对策略在感知共情、感知责任以及点赞、评论、分享意愿上的成对比较

反对策略	M (SD)				
	感知共情	感知责任	点赞意愿	评论意愿	分享意愿
呼吁良知	5.38 (1.26) ^{cd}	4.60 (1.65) ^b	0.60 (0.49) ^{ab}	0.48 (0.50) ^{bc}	0.24 (0.43) ^{ab}
呼吁逻辑	5.29 (1.36) ^{bc}	4.47 (1.81) ^{ab}	0.59 (0.49) ^{ab}	0.42 (0.49) ^{abc}	0.24 (0.43) ^b
威胁声誉	5.12 (1.38) ^{ab}	4.30 (1.76) ^{ab}	0.57 (0.50) ^{ab}	0.37 (0.48) ^{ab}	0.22 (0.42) ^b
礼貌纠偏	5.51 (1.12) ^d	4.54 (1.62) ^b	0.69 (0.47) ^b	0.49 (0.50) ^c	0.23 (0.42) ^b
群体除名	5.18 (1.33) ^{abc}	4.40 (1.75) ^{ab}	0.54 (0.50) ^a	0.41 (0.49) ^{abc}	0.19 (0.39) ^{ab}
幽默	5.06 (1.41) ^a	4.40 (1.77) ^a	0.56 (0.50) ^a	0.37 (0.48) ^a	0.19 (0.39) ^a

注：同一列中带有相同字母的策略在 $p < .05$ 水平上彼此不显著。

礼貌纠偏显著高于呼吁逻辑、威胁声誉、群体除名和幽默，呼吁良知显著高于威胁声誉和幽默；在感知责任方面，礼貌纠偏和呼吁良知均显著高于幽默；在行为指标上，礼貌纠偏的点赞意愿显著高于群体除名和幽默，评论意愿显著高于威胁声誉和幽默，呼吁良知的评论意愿显著高于幽默；在分享意愿方面，呼吁逻辑、威胁声誉和礼貌纠偏均显著高于幽默，各反对策略在上述变量上的描述性统计及事后比较结果见表2。

反对策略与话题类型在分享意愿上的交互作用显著 [$F(10, 1532.97) = 2.24, p = .014$]，简单效应分析表明，这一交互主要体现在文科话题中：呼吁逻辑的分享意愿显著高于呼吁良知、礼貌纠偏、群体除名和幽默 ($ps \leq .013$)；在追星和中医话题中，策略差异均不显著。

总体而言，相较于威胁声誉、群体除名和幽默等更具对抗性或间接性的回应方式，礼貌纠偏更容易激发旁观者的支持性反应，而话题类型仅改变不同策略对分享意愿的影响。

为检验假设1，本文依据确认偏误量表得分将被试划分为高、低两组。结果显示，确认偏误组别在各因变量上均存在显著主效应，低确认偏误组在感知共情、感知责任及点赞、评论和分享意愿上均显著高于高确认偏误组 ($ps < .001$)。反对策略与确认偏误的交互作用仅在评论意愿上显著 [$F(5, 1222.90) = 3.00, p = .011$]，其余指标上的交互作用均不显著。估计边际均值的结果见图2。

简单效应分析显示，低确认偏误组的策略差异不显著。在高确认偏误组中，不同反对策略对评论意愿的影响存在显著差异：礼貌纠偏显著高于呼吁逻辑、威胁声誉、群体除名和幽默 ($ps \leq .019$)；呼吁良知也显著高于群体除名和幽默 ($ps \leq .026$)。因此，假设1a和1b未得到支持，假设1c得到部分支持。

为检验假设2，本文依据话题态度量表三个维度的得分，将得分较高者划分为内群体，得分较低者划分为外群体。结果显示，内群体情

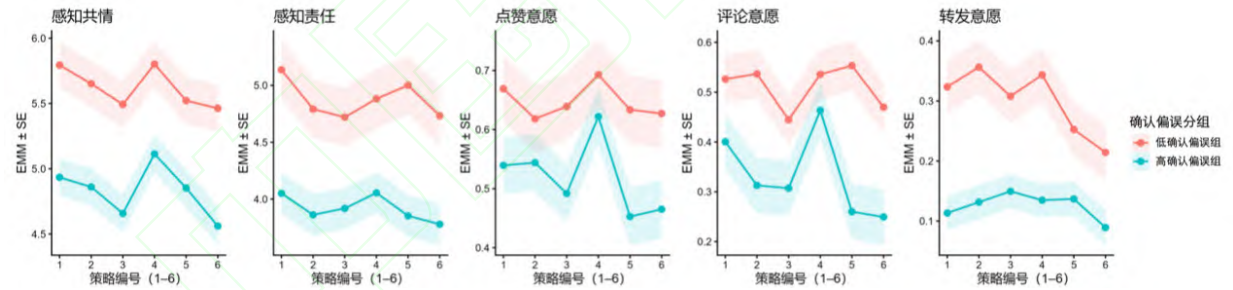


图2 不同确认偏误条件下六种反对策略对旁观者反应的影响

注：在策略编号（1—6）中，1=呼吁良知策略；2=呼吁逻辑策略；3=威胁声誉策略；4=礼貌纠偏策略；5=群体除名策略；6=幽默策略，下同。

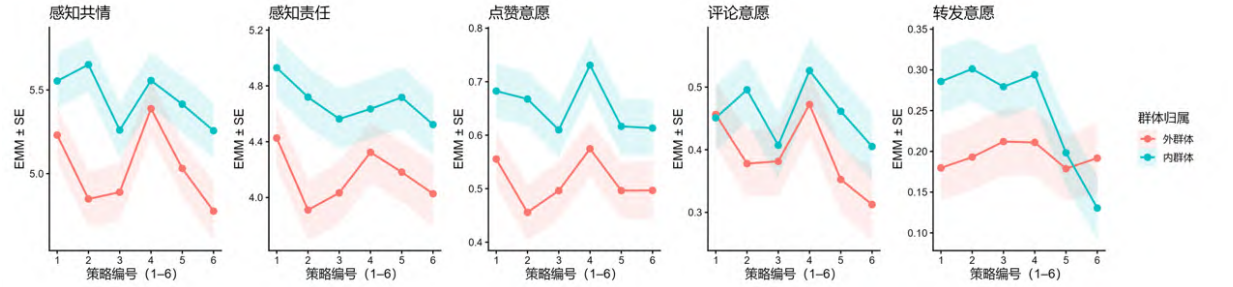


图3 不同群体归属条件下六种反对策略对旁观者反应的影响

境下的感知共情、感知责任及点赞、评论和分享意愿均显著高于外群体情境 ($ps < .001$)。反对策略与群体身份的交互作用在感知共情 [$F(5, 1294.86) = 2.65, p = .022$] 和分享意愿 [$F(5, 1256.70) = 2.30, p = .043$] 上显著, 在其余指标上均不显著。估计边际均值的结果见图 3。

简单效应分析显示, 在外群体情境下, 礼貌纠偏引发的感知共情显著高于呼吁逻辑、威胁声誉和幽默 ($ps \leq .006$), 呼吁良知也显著高于幽默 ($p = .006$); 在内群体情境下, 呼吁逻辑显著高于威胁声誉和幽默 ($ps \leq .015$)。在分享意愿方面, 策略差异主要出现在内群体情境, 呼吁良知、呼吁逻辑和礼貌纠偏均显著高于幽默 ($ps \leq .012$)。因此, 假设 2a 得到支持, 假设 2c 得到部分支持, 假设 2b 未得到支持。

为检验假设 3, 本文构建了“反对策略→感知共情→感知责任→干预行为”的链式中介模型。由于该观测变量路径模型为恰好识别模型 ($df = 0$), 因而不报告整体拟合指标, 模型对感知共情、感知责任和支持性干预行为的解释率分别为 $R^2 = 0.014$ 、 0.503 和 0.442 。

路径分析显示, 感知共情显著正向预测感知责任 ($\beta = 0.70, p < .001$), 感知责任进一步显著正向预测支持性干预行为 ($\beta = 0.46, p < .001$), 感知共情对干预行为也具有显著的直接作用 ($\beta = 0.25, p < .001$)。以礼貌纠偏为参照, 呼吁逻辑、威胁声誉、群体除名和幽默均显著负向预测感知共情, 呼吁良知的作用不显著; 在控制感知共情和感知责任后, 群体除名和幽默仍对干预行为具有显著负向作用, 具体路径见图 4。

Bootstrap 检验显示, 呼吁逻辑、威胁声誉、群体除名和幽默相对于礼貌纠偏的链式间接效应均显著且为负, 呼吁良知的链式间接效应不显著; 威胁声誉、群体除名和幽默的总间接效应显著, 具体结果见表 3。

因此, 感知共情与感知责任在部分反对策略影响支持性干预行为的过程中发挥链式中介作用, 假设 3 得到部分支持。

五、讨论

(一) 不同反对策略对旁观者支持的影响

本研究发现, 礼貌纠偏、呼吁良知和呼吁逻辑等低威胁策略, 更容易激发旁观者的共情、责任感与互动支持。这说明, 当反对方式带有较强的攻击、羞辱或排斥意味时, 可能使反对行为本身被理解为新的冒犯, 从而降低旁观者对其适当性与正当性的评价^[47]。从数字正义视

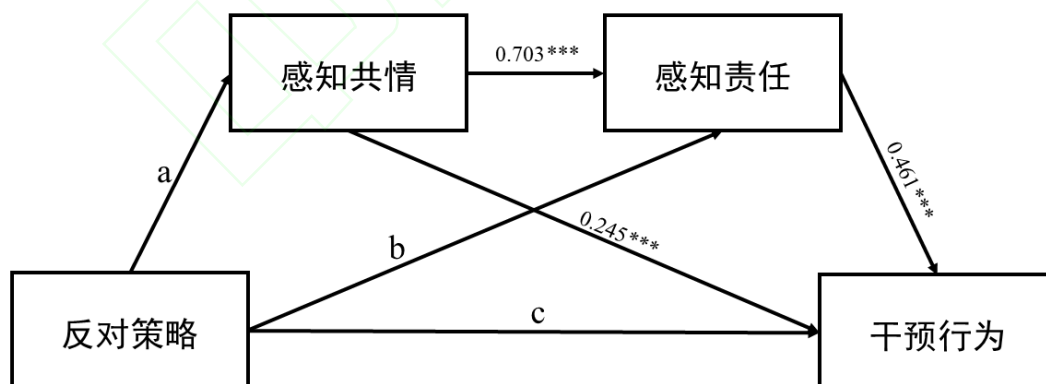


图 4 不同反对策略影响旁观者支持性干预行为的链式中介模型

注: 反对策略采用五个虚拟变量编码, 以礼貌纠偏策略为参照组。a: 与礼貌纠偏相比, 呼吁逻辑 ($\beta = -0.059, p = .031$)、威胁声誉 ($\beta = -0.107, p < .001$)、群体除名 ($\beta = -0.093, p < .001$) 和幽默 ($\beta = -0.126, p < .001$) 对感知共情的路径均显著, 呼吁良知不显著 ($p > .05$)。b: 与礼貌纠偏相比, 仅幽默策略的直接效应显著 ($\beta = 0.055, p = .007$), 其余策略对感知责任的直接效应均不显著 ($p > .05$)。c: 与礼貌纠偏相比, 群体除名 ($\beta = -0.053, p = .024$) 和幽默 ($\beta = -0.052, p = .018$) 对干预行为存在显著负向直接效应, 其余策略对干预行为的直接效应均不显著 ($p > .05$)。

表 3 不同反对策略影响旁观者支持性干预行为的具体间接效应

策略比较	经感知共情	经感知责任	感知共情→感知责任 链式间接效应	总间接效应
呼吁良知 vs. 礼貌纠偏	-0.009 [-0.020, 0.003]	0.016* [0.001, 0.034]	-0.012 [-0.027, 0.004]	-0.004 [-0.034, 0.028]
呼吁逻辑 vs. 礼貌纠偏	-0.015* [-0.028, -0.002]	0.013 [-0.004, 0.030]	-0.020* [-0.038, -0.003]	-0.022 [-0.056, 0.013]
威胁声誉 vs. 礼貌纠偏	-0.026* [-0.040, -0.013]	0.012 [-0.006, 0.029]	-0.035* [-0.052, -0.018]	-0.049* [-0.082, -0.016]
群体除名 vs. 礼貌纠偏	-0.022* [-0.036, -0.009]	0.015 [-0.003, 0.033]	-0.029* [-0.047, -0.013]	-0.036* [-0.070, -0.002]
幽默 vs. 礼貌纠偏	-0.030* [-0.045, -0.017]	0.027* [0.009, 0.046]	-0.040* [-0.059, -0.023]	-0.043* [-0.076, -0.009]

注：礼貌纠偏为参照组。* 表示 95% Bootstrap 置信区间不包含 0。

角看，旁观者是否支持反对行为，同时受到规范立场和表达方式的影响，礼貌纠偏在指出言论不当之处的同时维持了基本的互动秩序，因而更容易获得旁观者支持。

值得注意的是，幽默策略整体效果较弱，表明戏谑化表达未必能够有效推动旁观者参与。既有研究也发现，在应对社交媒体仇恨言论时，基于共情的反言论能够产生相对稳定的积极效果，而幽默和后果警示并未显示出一致作用^[48]。

此外，本文还检验了话题类型的调节作用。结果显示，话题类型仅对分享意愿产生有限的调节作用，且主要体现在文科话题中。呼吁逻辑在该话题下效果更好，这可能是因为旁观者分享此类内容时更看重回应的理由和论证，而追星和中医话题更多涉及情感认同与既有立场，因此不同策略对分享意愿的影响不明显。

（二）确认偏误对反对策略的调节作用

研究发现，低确认偏误个体的感知共情、感知责任和支持性干预行为均高于高确认偏误个体，表明其更容易对群体贬损性评论作出开放评估和积极回应。进一步来看，确认偏误的调节作用仅体现在评论意愿上，低确认偏误者对不同策略的反应差异不明显，高确认偏误者的反应则因策略而异，其中礼貌纠偏和呼吁良知更能激发其评论意愿。

已有研究表明，高确认偏误个体在面对与既有立场相冲突的信息时，更容易采取防御性加工^[49]。因此，这类旁观者对反对方式的威胁性更为敏感。礼貌纠偏和呼吁良知能够减弱对

抗感，因而更容易获得回应。这表明，在认知偏差较强的情境下，反对策略的可接受性有时比论证强度更为关键。

（三）群体身份对旁观者支持的影响

研究发现，群体身份会影响旁观者的整体反应水平，内群体情境下的共情、责任感及点赞、评论、分享意愿均高于外群体情境；群体身份还会影响不同反对策略的作用方式，在外群体情境下，策略差异主要体现在感知共情上；而在内群体情境下，策略差异进一步体现为分享意愿，说明当议题与旁观者的群体身份关联更强时，反对策略更容易激发公开支持。

这一结果与社会认同理论的基本判断一致，即群体身份会影响个体对群体规范的理解和行为选择^[50]。当反对策略被内群体旁观者理解为是在维护群体价值边界时，他们更可能公开表态；但若反对方式过于戏谑或对抗，内群体也可能将其理解为对群体体面的冒犯，从而降低支持意愿。

（四）共情与责任感驱动支持性干预行为的心理路径

研究发现，反对策略可通过感知共情与感知责任的链式作用影响旁观者的支持性干预意愿。共情有助于激发人们的利他动机^[51]，但这种情感反应能否转化为可见支持，还取决于旁观者是否认为自己负有介入责任^[52]。当责任感被分散时，其干预意愿也会随之下降^[53]。因此，不同反对策略能否获得支持，可能取决于其能否使旁观者将反对行为理解为对贬损性评论的

合理纠偏，并进一步形成介入责任认知。责任认知由此成为连接不公感知与公共纠偏行动的关键环节。

六、数字正义视角下旁观者支持性干预行为的理论意义与实践启示

本文将社交媒体旁观者干预置于数字正义的分析框架中，发现旁观者是否采取干预行动，受到反对策略、群体身份与确认偏误的共同影响。既有数字正义研究多从算法治理和平台公共性危机出发，强调通过数据资源分配和算法透明等路径回应算法时代的不平等问题^[54]，本研究则进一步表明，普通用户对具体贬损性评论的回应与反对，也是数字正义在日常互动中的实践。

本研究的发现为理解网络空间中规范生成与治理运行提供了新的切入点。网络空间治理不仅受到外部规制影响，也会在用户的持续互动中形成以内生调节为支撑的“涌现秩序”^[55]，这种秩序在大量用户的微小参与行动中逐渐形成，并随着平台可见性机制的介入持续显现和调整。网络意见生态作为具有自组织特征的复杂系统，具有自我调节和自我发展的特征，其规范结构在用户互动与反馈过程中不断生成与强化，每一次点赞、评论或转发都可能只是个体化的即时反应，但当这些反应在平台上被聚合、排序和可视化后，便会形成对公共讨论氛围的再塑造。旁观者对反对策略的支持，既是在回应某一具体贬损性评论，也是在向其他用户传递规范线索，正是在这种分散行动与平台可见性机制的交织中，网络公共空间的秩序得以被不断生产和修正。

在实践启示层面，平台也可通过功能设计、规则提示和可见性配置，搭建优化内容质量、鼓励理性讨论、提供被害救济的“媒介正义秩序”^[56]。具体而言，可从三个方面展开。

第一，优化干预路径的易用性与可见性。本文发现，不同的反对策略会影响旁观者的支持性干预行为意愿，这意味着平台可设计多样化、低门槛、低风险的反表达路径，使用户

能够在不同情境下选择恰当方式进行规范表达，并感知自身干预对内容生态的正向影响。第二，促进跨圈层规范共识的形成。群体身份认同会影响旁观者的支持性干预行为意愿，这意味着平台需要通过评论区呈现和推荐机制优化等方式，促进不同社群之间的理解与对话。第三，引导规范内化。较为稳固的网络秩序并不只来自外部惩罚，也来自用户在参与中将公共规范内化为自发的行为准则。因此，平台应更多展示正面规范信号，引导旁观者将支持正向表达、反对群体贬损性评论视为维护公共讨论秩序的日常工作，从而推动干预行为由偶发性的道德反应转化为持续性的规范建构。

总而言之，网络治理已深度嵌入用户的日常互动场景，平台的角色也由单一的规则管理者转向规则制定者与互动生态维护者。数字正义并不在于要求强制终止每一次冲突，而在于让更多旁观者愿意在关键时刻站出来，以更可接受的方式把公平、尊重与责任重新带回公共讨论之中，推动多方主体在互动中共同构建更加稳定且可持续的治理秩序。

本研究仍有进一步完善的空间。其一，本研究采用模拟社交媒体情境的在线实验设计，虽有助于控制干扰因素并识别不同反对策略的净效应，但与复杂的真实平台环境相比仍有差距。其二，本研究主要将旁观者干预理解为社交媒体中的外显互动行为，重点考察点赞、评论、分享等参与方式，这一处理有助于揭示旁观者介入公共讨论的行为取向，但这类干预能否进一步转化为被攻击者实际感受到的支持与保护，仍有待后续研究检验。

（责任编辑：李韵奕）

参考文献

- [1] 黄勛喆, 褚晓伟, 刘庆奇, 等. 网络欺负中的旁观者行为[J]. 心理科学进展, 2019, 27(7): 1248-1257.
- [2] 王鲁川, 岳建军. 从“袖手旁观”到“挺身而出”: 校园欺凌中旁观者的角色行为分析及其干预策略[J]. 西部学刊, 2025(21): 121-124.
- [3] Porten-Che é P, Kunst M, Emmer M. Online civic intervention: a new form of political participation under

- conditions of a disruptive online discourse[J]. *International Journal of Communication*, 2020, 14: 514–534.
- [4][14] Zapata J, Sulik J, von Wulffen C, et al. Bystanders' collective responses set the norm against hate speech[J]. *Humanities and Social Sciences Communications*, 2024, 11(1): 335.
- [5] Gennaro G, Derksen L, Abdelrahman A, et al. Counterspeech encouraging users to adopt the perspective of minority groups reduces hate speech and its amplification on social media[J]. *Scientific Reports*, 2025, 15(1): 22018.
- [6] Hansen T M, Lindekilde L, Karg S T, et al. Combatting online hate: crowd moderation and the public goods problem[J]. *Communications*, 2024, 49(3): 444–467.
- [7] 郭全中, 李黎. 网络综合治理体系: 概念沿革、生成逻辑与实践路径[J]. *传媒观察*, 2023 (7): 104–111.
- [8] 周敏, 赵秀丽. 如何应对社交媒体上的污名化? ——不同污名应对路径的在线社会支持度研究[J]. *新闻界*, 2024 (12): 37–50.
- [9] 伊森·凯什, 奥娜·拉比诺维奇-艾尼. 数字正义: 当纠纷解决遇见互联网科技[M]. 赵蕾, 赵精武, 曹建峰, 译. 北京: 法律出版社, 2019: 53–56.
- [10] Taylor L. What is data justice? The case for connecting digital rights and freedoms globally[J]. *Big Data & Society*, 2017, 4(2): 2053951717736335.
- [11] Floridi L. Faultless responsibility: on the nature and allocation of moral responsibility for distributed moral actions[J]. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 2016, 374(2083): 20160112.
- [12] 易前良, 唐芳云. 镜与台: 数字时代的“传播可见性”——两种范式的辨析和应用[J]. *新闻与写作*, 2023 (4): 43–52.
- [13] Fraser N. Social justice in the age of identity politics: Redistribution, recognition, and participation[M]. Fraser N, Honneth A. *Redistribution or Recognition? A Political-Philosophical Exchange*. London: Verso, 2003: 7–109.
- [15][41] Latané B, Darley J M. The unresponsive bystander: why doesn't he help?[M]. New York: Appleton-Century Crofts, 1970: 31–32.
- [16] 丁凤琴, 陆朝晖. 共情与亲社会行为关系的元分析[J]. *心理科学进展*, 2016, 24 (8): 1159–1174.
- [17] Freis S D, Gurung R A R. A Facebook analysis of helping behavior in online bullying[J]. *Psychology of Popular Media Culture*, 2013, 2(1): 11–19.
- [18][53] Darley J M, Latané B. Bystander intervention in emergencies: diffusion of responsibility[J]. *Journal of Personality and Social Psychology*, 1968, 8(4, Pt. 1): 377–383.
- [19] You L, Lee Y-H. The bystander effect in cyberbullying on social network sites: anonymity, group size, and intervention intentions[J]. *Telematics and Informatics*, 2019, 45: 101284.
- [20][47] Zhao P, Bazarova N N, DiFranzo D, et al. Standing up to problematic content on social media: which objection strategies draw the audience's approval?[J]. *Journal of Computer-Mediated Communication*, 2024, 29(1): zmad046.
- [21][23][48] Hangartner D, Gennaro G, Alasiri S, et al. Empathy-based counterspeech can reduce racist hate speech in a social media field experiment[J]. *Proceedings of the National Academy of Sciences*, 2021, 118(50): e2116310118.
- [22] Gavrillets S, Richerson P J. Collective action and the evolution of social norm internalization[J]. *Proceedings of the National Academy of Sciences*, 2017, 114(23): 6068–6073.
- [24][29] Shea A L, Omapang A K B, Cho J Y, et al. Discursive objection strategies in online comments: Developing a classification schema and validating its training[EB/OL]. arXiv, (2024-05-13)[2026-02-01]. <http://arxiv.org/abs/2405.08142>.
- [25] 董庆兴, 王雪宁, 熊思玥. 纠正是否有效? ——纠正方式和错误健康信息启发式线索的影响研究[J]. *新闻与传播研究*, 2024, 31 (4): 65–80.
- [26] Boyd R, Richerson P J. Punishment allows the evolution of cooperation (or anything else) in sizable groups[J]. *Ethology and Sociobiology*, 1992, 13(3): 171–195.
- [27] Hornsey M J. Why being right is not enough: Predicting defensiveness in the face of group criticism[J]. *European Review of Social Psychology*, 2005, 16(1): 301–334.
- [28][50] Hogg M A, Reid S A. Social Identity, Self-Categorization, and the Communication of Group Norms[J]. *Communication Theory*, 2006, 16(1): 7–30.
- [30] 牟怡, 张林. 幽默说服在医学微信公众号上传播效果代际差异的实验研究[J]. *现代传播 (中国传媒大学学报)*, 2020, 42 (9): 144–150.
- [31] Samson A C, Gross J J. Humour as emotion regulation: the differential consequences of negative versus positive humour[J]. *Cognition and Emotion*, 2012, 26(2): 375–384.
- [32] Elsayed Y, Hollingshead A B. Humor reduces online incivility[J]. *Journal of Computer-Mediated Communication*, 2022, 27(3): zmac005.
- [33] Kahneman D. *Thinking, fast and slow*[M]. New York: Farrar, Straus and Giroux, 2011.
- [34] Dimara E, Franconeri S, Plaisant C, et al. A task-based taxonomy of cognitive biases for information visualization[J]. *IEEE Transactions on Visualization and Computer Graphics*, 2020, 26(2): 1413–1432.
- [35] Nickerson R S. Confirmation bias: A ubiquitous phenomenon in many guises[J]. *Review of General Psychology*, 1998, 2(2): 175–220.
- [36] Cinelli M, De Francisci Morales G, Galeazzi A, et al. The echo chamber effect on social media[J]. *Proceedings of the National Academy of Sciences*, 2021, 118(9): e2023301118.
- [37] Ecker U K H, Ang L C. Political attitudes and the processing of misinformation corrections[J]. *Political Psychology*, 2019, 40(2): 241–260.
- [38] Turner J C, Brown R J, Tajfel H. Social comparison and group interest in ingroup favouritism[J]. *European Journal of Social Psychology*, 1979, 9(2): 187–204.
- [39] Tarrant M, Dazeley S, Cottom T. Social categorization and empathy for outgroup members[J]. *British Journal of Social*

Psychology, 2009, 48(3): 427–446.

[40] Levine M, Prosser A, Evans D, et al. Identity and emergency intervention: how social group membership and inclusiveness of group boundaries shape helping behavior[J]. *Personality and Social Psychology Bulletin*, 2005, 31(4): 443–453.

[42] Shen L. On a scale of state empathy during message processing[J]. *Western Journal of Communication*, 2010, 74(5): 504–524.

[43] Fischer P, Greitemeyer T, Pollozek F, et al. The unresponsive bystander: are bystanders more responsive in dangerous emergencies?[J]. *European Journal of Social Psychology*, 2006, 36(2): 267–278.

[44][52] Obermaier M, Fawzi N, Koch T. Bystanding or standing by? How the number of bystanders affects the intention to intervene in cyberbullying[J]. *New Media & Society*, 2016, 18(8): 1491–1507.

[45] Haran U, Ritov I, Mellers B A. The role of actively open-minded thinking in information acquisition, accuracy, and calibration[J]. *Judgment and Decision Making*, 2013, 8(3): 188–

201.

[46] Svedholm-Häkkinen A M, Lindeman M. Actively open-minded thinking: development of a shortened scale and disentangling attitudes towards knowledge and people[J]. *Thinking & Reasoning*, 2018, 24(1): 21–40.

[49] Taber C S, Lodge M. Motivated Skepticism in the Evaluation of Political Beliefs[J]. *American Journal of Political Science*, 2006, 50(3): 755–769.

[51] Batson C D, Batson J G, Slingsby J K, et al. Empathic joy and the empathy–altruism hypothesis[J]. *Journal of Personality and Social Psychology*, 1991, 61(3): 413–426.

[54] 焦德武. 迈向数字正义：算法时代平台传播公共性危机的发生机制与治理路径[J]. *电子科技大学学报（社科版）*, 2026: 1–10.

[55] 杨颖兮, 喻国明. 传播中的非理性要素：一项理解未来传播的重要命题[J]. *探索与争鸣*, 2021 (5) : 131–138.

[56] 朱笑延. 走向媒介正义：网络暴力信息治理的平台角色与法治实现[J]. *中南大学学报（社会科学版）*, 2024, 30(1): 50–62.

From Silence to Action: Understanding Bystander Intervention on Social Media Through the Lens of Digital Justice

Chen Xuejiao, Zhao Yiming, Wang Qifang, Wu Ye

Abstract: Derogatory comments targeting social groups are a persistent feature of social media, and platform moderation alone cannot fully address them. Ordinary users may also shape online norms by challenging harmful content or supporting those who speak out against it. Drawing on the concept of digital justice, this study uses an online vignette experiment to compare six counterspeech strategies and examine their effects on bystanders' empathy, perceived responsibility, and willingness to engage in supportive intervention. It also investigates whether these effects vary across topics, group identity and confirmation bias. The results show that the way counterspeech is framed significantly shapes bystander responses. Polite correction generally generates the strongest willingness to intervene, while the effectiveness of different strategies varies with group identity and confirmation bias. Mediation analyses further indicate that counterspeech influences intervention intentions through a sequential pathway from empathy to perceived responsibility. Taken together, the findings suggest that ordinary users' willingness to support counterspeech may represent an everyday expression of digital justice, helping online communities develop shared norms and forms of emergent order.

Keywords: digital justice; bystander intervention; group identity; confirmation bias

Authors: Chen Xuejiao, School of Journalism and Communication and Center for Computational Communication Research, Beijing Normal University; Zhao Yiming, School of Journalism and Communication, Beijing Normal University; Wang Qifang, School of Journalism and Communication, Beijing Normal University; Wu Ye, School of Journalism and Communication and Center for Computational Communication Research, Beijing Normal University.