新闻与传播评论 JCR
Journalism & Communication Review
第 79 卷 第 2 期 2026 年 3 月
Vol. 79 No. 2 Mar. 2026. 090-101

DOI:10. 14086/j. cnki. xwycbpl. 2026. 02. 008

集体把关:竞争传播环境和意见交互阈值的作用

李丹珉

摘要:在数字媒体快速发展的时代,虚假信息大规模传播已成为威胁人类社会安全的重大挑战。近年来,Meta 等科技巨头纷纷推出社区笔记功能,试图通过用户协作的集体把关模式抑制虚假信息传播。在此背景下,集体把关的有效性值得人们重新审视。利用 NetLogo 多主体仿真方法,可模拟集体把关对虚假信息支持者态度的影响。研究发现:在不同竞争传播环境和意见交互阈值下,集体把关效果存在显著差异。在虚假信息强势传播和高意见交互阈值情况下,集体把关能减少的虚假信息支持者以及极端支持者比例更高。中立多数更有助于压缩虚假信息支持者比例,中立少数更有助于缩小极端支持者规模。竞争传播环境对集体把关的影响受意见交互阈值调节。在低阈值和中阈值情况下,竞争传播环境变化对集体把关结果影响更大。总体来看,集体把关能否有效遏制虚假信息传播取决于具体情境。虚假信息治理应采取综合干预手段,将集体把关和其他辟谣手段结合起来,提升整体治理效果。

关键词:虚假信息;信息治理;模拟仿真
中图分类号:G206　　**文献标志码**:A　　**文章编号**:2096-5443(2026)02-0090-12
基金项目:上海市哲学社会科学规划课题(2023ZXW002)

一、导言

后真相时代,随着媒介技术不断发展,越来越快的信息传播速度与不断降低的内容生产门槛共同导致网络舆论场中虚假信息泛滥。虚假信息助长暴力与偏见,加剧分裂与冲突,破坏社会凝聚力。世界经济论坛(The World Economic Forum)发布的《2024 年全球风险报告》(*Global Risks Report 2024*)将虚假信息列为未来两年全球面临的最大风险之一。[1]部分研究者认为,虚假信息泛滥与把关失效密切相关。[2]在新媒体冲击之下,传统媒体把关能力日渐弱化,如何采取更有效的手段以应对网络中泛滥成灾的虚假信息,成为人们不得不思考的问题。

2023 年 5 月,一张显示五角大楼附近发生爆炸的虚假图片在社交媒体上疯传。随后,X(原 Twitter)新增一项功能,允许用户为具有误导性的内容添加澄清说明。[3]2024 年 6 月,YouTube 开始测试社区笔记功能,鼓励用户通过添加注释的方式,标注视频中的错误信息。[4]2025 年 1 月,Facebook 母公司 Meta 宣布,将用社区笔记系统替代其现行的第三方事实核查项目。[5]这些互联网巨头普遍希望利用用户集体把关的力量遏制虚假信息传播。

用户通过评论、举报、编辑、投票、批注等社交互动行为实现信息筛选、过滤和传播的过程被称为集体把关(collective gatekeeping)。[6]传统事实核查依赖新闻机构(如 BBC Verify)或第三方组织(如 PolitiFact)的专业团队对信息进行审核。集体把关则主要由去中心化的普通用户完成,信息传播由社群共识所决定。除了社区笔记系统,Wikipedia 内容编辑系统以及 Reddit 投票系统也符合集体把关特征。随着越来越多的社交媒体平台尝试利用集体把关应对虚假信息,人们开始围绕这一机制的有效性展开讨论。有观点认为,集体把关能让用户共同决定哪些内容具有误导性,这种做法有助于构

建更为开放的社交媒体环境;不过,也有业界人士表示,面对病毒式传播的虚假信息,集体把关的作用可能微乎其微。[7]

上述分歧反映出集体把关研究仍有待深入,集体把关对虚假信息传播的影响有待进一步检验。在新闻传播学和信息科学领域,围绕集体把关议题,学者们着重探讨影响用户参与虚假信息纠正的因素以及用户集体把关中考虑的信息标准等问题[8-9],少有研究从宏观角度验证集体把关的有效性。此外,尽管已有学者从动态视角出发,模拟社会互动中的舆论演化过程,但相关研究主要聚焦不同个体属性与群体结构条件下群体意见如何趋于一致或分化为多个对立意见簇[10-12]。从集体把关视角出发,探讨用户行为如何影响虚假信息传播的研究仍较为有限。有鉴于此,本研究采用仿真研究方法,模拟不同情境下集体把关网络的运行效果,测量集体把关对虚假信息传播的影响。在学理层面,通过总结社群驱动的集体把关特点,丰富把关理论现有研究成果;在实践层面,通过探索集体把关在打击虚假信息传播方面发挥的作用,为虚假信息治理提供可行建议。

二、文献回顾与问题提出

集体把关这一概念来源于把关理论。1947 年,勒温(Kurt Lewin)在思考为什么不同美国家庭的饮食习惯不同时,首次明确提出“把关人”的概念。[13]勒温发现食物通过不同渠道进入美国家庭餐桌,家庭主妇犹如把关人,会将不受欢迎的食物移出购买清单。20 世纪 50 年代,新闻传播学界围绕信息把关现象展开首波理论探讨,怀特(David Manning White)通过研究一位美国地方报纸编辑对电讯稿的筛选行为,提出新闻选择的基本把关模式。[14]传统把关理论认为,信息传播过程主要由专业媒体机构所控制。不过进入 21 世纪后,以舒梅克(Pamela Shoemaker)为代表的学者们发现,互联网兴起后,传统媒体的权威性逐渐减弱,信息流动开始由用户行为、社群共识和平台算法等多重因素共同决定。[15-17]随着社交媒体互动性的增强,个体在依据自身标准筛选信息并与他人持续互动的过程中,涌现式的集体把关效用会显现出来。[18]虽然布伦斯(Axel Bruns)、梅拉兹(Sharon Meraz)等学者普遍承认集体把关有纠正虚假信息的潜能,但也强调集体把关效果受信息传播环境与用户行为结构的限制[19-21],本研究遵循这一思路对相关文献进行回顾。

(一)竞争传播环境与集体把关

在多元信息并存的社交媒体环境下,公众的认知判断与观点形成并非由单一信息所决定,而是持续受到立场对立、相互竞争的多种信息源的共同影响。[22]所谓竞争传播环境,是指立场不同的信息(如虚假信息与辟谣信息)围绕公众的认知资源展开动态博弈,进而形成的具有明显对抗性的传播场域。[23]在不同的竞争传播环境中,由于信息传播的结构性差异与用户意见生态的多样性,同一条虚假信息在不同平台上可能呈现出截然不同的传播路径。

竞争传播环境的特征首先表现为虚假信息是否在与辟谣信息的竞争中占据优势。有学者发现,与北欧和西欧国家相比,美国公民初始相信虚假信息的比例更高。[24]还有学者指出,虚假信息在八卦报纸等低合法性媒体上的影响力更大且更容易被用户所相信。[25]总体而言,当虚假信息在初始支持者比例、信息影响力、信息可信度等方面占据优势时,竞争传播环境呈现虚假信息强势特征;反之,若辟谣信息在这些方面更具优势,则出现虚假信息弱势情况;而当双方势均力敌时,则形成均势竞争情况。

除了虚假信息与辟谣信息之间的竞争关系外,中立者的结构性分布也是影响竞争传播环境的重要因素。相关研究显示,中立者比例提升在一定条件下有助于降低社群间的对抗强度。[26]一方面,中立者作为连接不同社群的桥梁,有助于打破回音室效应;另一方面,中立者的存在能够降低网络中极端意见的集中程度,促使各类意见在网络中均衡分布。

针对不同竞争传播环境下的集体把关效果,学者们提出不同观点。一类观点认为,社交媒体平台上大规模用户的参与并不一定提升信息把关效果,反而可能增强虚假信息的扩散能力。[27]与此相对,另一类观点强调集体把关的纠偏潜力,认为用户可以通过指出信息中的事实性错误或不当表述,

帮助其他用户更好地理解新闻事件,促使讨论空间的规范形成。[28]这些研究分歧表明,竞争传播环境对集体把关效果的影响具有复杂性。为进一步探索不同竞争传播环境下的集体把关效果,本研究提出如下研究问题:

研究问题 1:在不同的竞争传播环境下,集体把关效果是否存在差异?

(二)集体把关中的意见交互阈值

意见交互阈值体现个体的开放性。意见交互阈值越高,个体越愿意与持不同意见的人互动,对异质性意见的接受度越高。[11]根据社会判断理论,人们在接触到新信息后,会利用已有观念对新信息进行感知评价。人们在面对新信息时,通常会根据该信息与自身立场的接近程度,将其归入接受区、不确定区或拒绝区。只有当双方观点相差不大时,用户意见才可能发生改变。[29]

信息影响力、信息可信度、个体从众性都是影响意见交互阈值的关键因素。信息影响力能够影响用户的信息采纳意愿,社交媒体平台上高点赞和高转发的信息更容易被用户接受。[30]信息可信度是指人们对某一信息真实性和可靠性的评估。有研究发现,即便信息本身具有客观真实性,若未能获得用户的信任和认同,其在网络舆论场中的传播效能亦可能明显受限。[31]个体从众性是指个体在信息判断过程中是否容易受他人观点的影响,从众性较高的个体更容易接受群体观点并根据群体观点调整自身立场。[32]

既有研究普遍认为,意见交互阈值越高,社交网络中的信息流动性越强,不同观点之间越容易交叉传播。[33]不过,意见交互在虚假信息传播中的作用具有双重可能性。一方面,如果意见交互发生在同质性较强的社群内部,虚假信息在群体内部的影响力可能会被进一步放大[34];另一方面,当用户基于不确定性进行决策时,他们更容易接触到与虚假信息相悖的辟谣意见,这种接触会促使用户更新自己的观点,从而限制虚假信息传播[35]。这些看似矛盾的研究结果提醒我们,在不同竞争传播环境下,意见交互阈值可能会发挥不同作用,应对不同意见交互阈值下虚假信息的集体把关情况进行分层讨论。基于此,本研究提出以下两个研究问题:

研究问题 2:在不同的意见交互阈值下,集体把关效果是否存在差异?

研究问题 3:竞争传播环境对集体把关的影响是否会受意见交互阈值的调节?

三、集体把关的演化仿真

尽管各种集体把关方式在具体运作机制上存在差异,但它们确实拥有某些共同特征。首先,信息筛选依靠用户群体协作而非单一机构或个体决策来完成;其次,用户群体内部存在分化立场,人们对同一条信息的观点并非完全一致,支持、中立、反对虚假信息的网民共存;最后,用户对信息的认知会动态变化,人们对信息真伪的评估会受他人反馈和外界环境的影响。基于此,本研究试图建构能够映射这些特征的集体把关演化仿真模型。一方面,该模型创设去中心化的信息传播场域,场域中的虚假信息支持者、中立者、反对者共同参与把关活动,这些行动主体共同决定虚假信息传播结果;另一方面,在互动过程中,用户对信息的态度会随集体反馈情况发生改变。沿此逻辑延伸,本研究将进一步探索不同情境下的集体把关效果,以期为社交媒体虚假信息治理提供思路。

(一)集体把关的场域建构

为了更好地模拟现实中的集体把关场域,本研究对 BA 无标度网络模型进行改进。该模型可以通过增长机制和优先连接机制生成具有幂律分布特征和动态适应性特征的网络。本研究一方面利用三角构成法提高网络的聚类系数,另一方面引入边的权重,使生成的网络更接近现实世界中的复杂网络。

在网络生成阶段,模型首先创建 N_0 个初始节点,并在不允许重复连边的条件下随机添加连边,使网络达到预设的初始连边规模。随后进入增长阶段,模型每轮新增一个节点,并仅在度未达到上限 K_{imax} 的既有节点中选择连接对象;在此基础上采用偏好连接机制,使连接度更高的节点以更大概率获得新连边。在新节点与既有节点建立连边后,若该既有节点存在其他邻居,模型将从其邻居中

随机抽取一个节点与新节点再连边,以形成三角闭合结构并增强局部聚类。为刻画加权网络中连接强度差异,本研究在网络生成阶段为连边赋予权重属性 w,将其操作化为行动者之间的连接强度。为在模型中以最小参数成本呈现强弱连接的结构异质性,参考加权网络研究中以边权重表征连接强度的常见处理[36-37],每次生成连边均以 0.2 的概率设为强连接(w =0.8),否则设为弱连接(w =0.2)。

竞争传播环境下,行动主体可以分为支持、中立、反对虚假信息的三类人群。支持者倾向于主动传播虚假信息,反对者会通过提出反驳或质疑来阻断虚假信息传播。中立者起初对虚假信息无明显态度,他们的传播行为受社交压力或外界舆论的影响。在模拟仿真研究中,行动者数量的设定可能会影响输出结果的稳定性,因此需要对比不同行动者数量的模拟结果以有效识别模型的结构敏感性程度。[38]在保证计算效率与实验可重复性的前提下,参考过往研究[39],将初始行动者数量分别设置为 600 人、900 人、1200 人。

(二)主体行为设置

为了更好地模拟集体把关中个体意见的演化过程,本研究在 Hegselmann-Krause(HK)模型的基础上,针对行动主体属性和意见改变规则进行扩展设定。该模型构建一个封闭的社交网络,其中包含若干行动主体,每个行为主体都拥有一个初始意见值。在模型的每一步仿真中,行动主体都会与其邻近的其他主体发生互动,其意见值根据与邻居意见的关系按特定规则进行更新。当主体的意见值在一定时间内趋于稳定或模型运算达到预设的最大迭代次数时,程序停止运行。HK 模型以其简单有效的结构被广泛用于模拟群体意见的演化过程,该模型在研究网络环境下共识的出现与意见的分化时表现出重要价值。

核心行动者都拥有四个基本属性:意见值和意见倾向、信息影响力、信息可信度、个体从众性。①意见值和意见倾向:意见值和意见倾向代表行动者对虚假信息的态度。本研究将初始状态下个体 i 的意见值设定为 $O_i(t)$,表示个体 i 在 t 时刻的意见值。意见值分布于连续区间[0,1],意见值所处区间可反映不同的意见倾向。其中,[0,0.33]表示反对虚假信息的倾向,(0.33,0.67)表示中立倾向,[0.67,1]表示支持虚假信息的倾向。此外,如果意见值小于 0.1 或大于 0.9,则视为极端反对或极端支持虚假信息的倾向。②信息影响力:信息影响力 I_i 反映行动者 i 在网络中的结构性影响力。本研究中,I_i 的取值范围为[0,1],数值越大说明行动者 i 对他人意见变化的推动作用越强。③信息可信度:信息可信度 T_i 的取值范围为[0,1],表示行动者 i 在互动中对信息的信任倾向。④个体从众性:个体从众性 C_i 意味着行动者 i 受外部意见影响的容易程度。C_i 在[0,1]的范围内随机取值,数值越大说明行动者 i 越容易改变自己的态度。

个体行动与意见更新规则:模型初始状态下,所有行动者围绕同一条虚假信息形成意见值 $O_i \in [0,1]$,并据此被划分为反对者、中立者、支持者三类。三种意见倾向的人数比例可以在 0~100%之间变化,其总和为 100%。程序运行过程中,每一轮(tick)都会先根据当前行动者的意见值 O_i 重新标记行动者立场。在互动阶段,模型将支持者与反对者视为积极表达立场的行动者,对这两类主体执行意见更新;中立者暂时不主动更新自身意见,但其中立立场仍作为邻近信息环境的一部分参与和他人的互动。具体来说,非中立主体 i 会与网络中的相邻个体发生潜在互动,并进一步受到意见交互阈值的约束:只有当相邻个体 j 与 i 的意见差异满足 $|O_i(t) - O_j(t)| < \epsilon$(程序中对应阈值参数 z)时,双方才发生意见交互。若存在可互动邻居,则主体 i 的态度将向该邻居的态度方向发生一次调整,其调整幅度由邻居的信息影响力 I_j、行动者自身的个体从众性 C_i、双方的信息可信度 T_i 和 T_j 共同调节。更新完成后,意见值被限制在[0,1]区间内,以保证意见范围的可解释性。具体计算公式为:

$$O_i(t+1) = O_i(t) + (O_j(t) - O_i(t)) \cdot I_j \cdot T_i \cdot T_j \cdot C_i$$

(三)工具模拟与情境设置

本研究使用由美国西北大学计算机科学专家威伦斯基(Uri Wilensky)开发的多主体建模语言 Netlogo 6.3.0 进行仿真模拟。由于 Netlogo 可以通过参数变化组合生成多种情境[40],通过调整变量

取值,模拟不同虚假信息传播情境详见表 1。

竞争传播环境:本研究将竞争传播环境设定为二阶变量,该变量由信息竞争情况和中立者比例两个一阶变量共同构成。信息竞争情况反映虚假信息与辟谣信息的传播竞争关系。参考前人关于谣言辟谣过程中情绪传播的多主体仿真研究,辟谣相关的关键客观因素(如可信度、辟谣速度等)可以被归一化至[0,1]区间,并可以采用 0.2、0.5、0.8 的分层参数设定构建多组情景。[41]基于这一情景化设定思路,根据虚假信息是否在与辟谣信息的竞争中占据优势,将信息竞争情况划分为三类。①弱势传播:初始支持/反对虚假信息的人数比例为 1∶3,虚假信息的影响力与可信度均为 0.2,辟谣信息的影响力与可信度均为 0.8;②均势传播:人数比例为 1∶1,虚假信息和辟谣信息的影响力与可信度均为 0.5;③强势传播:人数比例为 3∶1,虚假信息的影响力与可信度均为 0.8,辟谣信息的影响力与可信度均为 0.2。

中立者比例是指中立者在场域中的结构性占比。曾有学者在对社会网络中初始中立者比例进行遍历(范围覆盖 0.01~0.99)之后发现,初始中立者比例影响系统最终走向,尤其是在低敏感度情况下,初始中立者比例较低则行动者走向两极分化,初始中立者比例较高则行动者走向全部中立。[42]据此,本研究设置中立少数(20%)与中立多数(60%)两种情况。值得一提的是,个体从众性属于相对个体化的变量,会因人格特质不同而产生差异。[43]此外,从建模技术的角度来看,如果将个体从众性设为随机变量,可以避免 HK 模型的群体演化路径过于线性、缺乏分化。因此,将个体从众性设为随机变量,用以增强模型的复杂演化能力。

意见交互阈值:意见交互阈值的取值范围为[0,1],数值越接近 1,意味着意见交互阈值越高,个体之间越容易互相产生影响。受过往研究启发[44],将意见交互阈值分为低阈值(0.2)、中阈值(0.5)、高阈值(0.8)三种。低阈值意味着个体只愿意与意见非常接近的人互动;中阈值说明个体既不完全开放,也不完全封闭,他们愿意与意见有一定差异的人沟通;高阈值代表个体愿意与意见差异较大的人交流。

在情境设置的过程中,将信息竞争情况与中立者比例两项因素结合起来构建 6 种竞争传播环境类型,并以编号 1~6 加以区分。例如,竞争传播环境 1 即代表“弱势传播+中立少数”情况。竞争传播环境类型分别搭配不同的意见交互阈值(低:0.2,中:0.5,高:0.8),构成 6(竞争传播环境)×3(意见交互阈值)的实验设计,共模拟 18 组情境。

表 1 模拟情境

竞争传播环境	弱势传播	竞争传播环境	均势传播	竞争传播环境	强势传播	意见交互阈值
1	中立少数(20%-20%-60%)	3	中立少数(40%-20%-40%)	5	中立少数(60%-20%-20%)	低(0.2)
						中(0.5)
						高(0.8)
2	中立多数(10%-60%-30%)	4	中立多数(20%-60%-20%)	6	中立多数(30%-60%-10%)	低(0.2)
						中(0.5)
						高(0.8)

注:表中括号内的百分比表示支持者、中立者、反对者在模拟群体中的人数占比。

本研究使用 2 个指标测量集体把关效果:①支持虚假信息的人数变化率;②极端支持虚假信息的人数变化率。测量以每轮开始时的人数为初始值,以迭代 1500 次后的人数为结束值。通常情况下,变化率为(结束值-初始值)/初始值×100%。由于在本研究模拟的多数情境中,结束值小于初始值,这就导致多数变化率为负值。为了方便阅读,给变化率取相反数,以使图表更加直观。此时,支持虚假信息的人数变化率如果为正,意味着支持虚假信息的人数减少,反之意味着支持虚假信息的

人数增多;极端支持虚假信息的人数变化率如果为正,意味着极端支持虚假信息的人数减少,反之意味着极端支持虚假信息的人数增多。

集体把关网络中的意见值分布具有随机性,模型每次运行结果存在一定差别。为了减少随机误差并提高数据的可靠性,参考前人研究设计[45],数据分析使用模型20轮运行的结果。模型具体运行机制如图1所示。

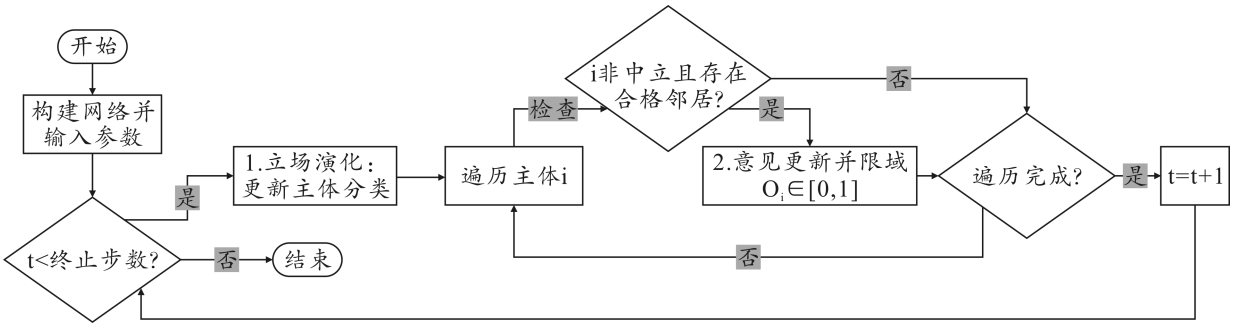

图1 模型运行机制图

四、研究发现

为检验不同行动者数量(600、900、1200)对模型输出结果的影响,利用单因素方差分析对其进行敏感性测试。结果显示,行动者数量变化不会引起"支持虚假信息人数的变化率"($p=0.764>0.05$)和"极端支持虚假信息人数的变化率"($p=0.978>0.05$)的显著变化,这表明模型输出结果对行动者数量变动不敏感。因此,将三个行动者数量组的数据合并起来用于后续统计分析,以扩大样本规模并增强分析稳健性。在敏感性测试的基础上,利用双因素方差分析,比较不同竞争传播环境和意见交互阈值对集体把关效果的影响。表2为模拟仿真结果的描述统计表。

表2 描述统计表

竞争传播环境	支持虚假信息的人数变化率			极端支持虚假信息的人数变化率		
	阈值低	阈值中	阈值高	阈值低	阈值中	阈值高
1(弱势传播+中立少数)	0.179 (0.040)	0.814 (0.067)	0.985 (0.013)	−0.141 (0.245)	0.589 (0.190)	0.998 (0.007)
2(弱势传播+中立多数)	0.298 (0.060)	0.907 (0.040)	0.967 (0.019)	−0.359 (0.266)	0.849 (0.109)	0.994 (0.014)
3(均势传播+中立少数)	0.195 (0.090)	0.914 (0.027)	0.992 (0.006)	0.471 (0.165)	0.845 (0.058)	1.000 (0.002)
4(均势传播+中立多数)	0.396 (0.090)	0.956 (0.018)	0.982 (0.013)	0.185 (0.273)	0.928 (0.052)	0.997 (0.007)
5(强势传播+中立少数)	0.180 (0.102)	0.949 (0.020)	0.970 (0.018)	0.776 (0.103)	0.971 (0.020)	0.997 (0.006)
6(强势传播+中立多数)	0.518 (0.126)	0.965 (0.021)	0.974 (0.017)	0.637 (0.211)	0.985 (0.021)	0.997 (0.006)

注:表中数据以"均值(标准差)"的形式呈现。

（一）竞争传播环境对集体把关的影响

针对研究问题 1，研究发现，在不同的竞争传播环境下，集体把关网络运行后支持虚假信息的人数变化率（$F=192.915$，$p<0.001$，$\eta_p2=0.476$）和极端支持虚假信息的人数变化率（$F=322.902$，$p<0.001$，$\eta_p2=0.603$）均存在显著差异，见表 3。

表 3　双因素方差分析结果

	支持虚假信息的人数变化率		极端支持虚假信息的人数变化率	
	F	η_p2	F	η_p2
校正模型	2034.480***	0.970	546.328***	0.897
竞争传播环境	192.915***	0.476	322.902***	0.603
意见交互阈值	16228.344***	0.968	2896.826***	0.845
竞争传播环境×意见交互阈值	116.490***	0.523	187.941***	0.639

注：* $p<0.05$，** $p<0.01$，*** $p<0.001$。

就支持虚假信息的人数变化率而言，通过 LSD 多重比较后发现，除了“3（均势传播+中立少数）和 5（强势传播+中立少数）”（$M_{\_diff}=0.0005$，$p=0.939>0.05$），其他竞争传播环境之间的均值差异均显著（$p<0.001$）。由表 4 可知，6（强势传播+中立多数）的集体把关效果最为突出，这说明中立者的调和作用可能在虚假信息强势传播情况下尤为重要。在相同的信息竞争情况下，中立多数的抑制作用强于中立少数（例如，竞争传播环境 6>5，竞争传播环境 4>3）。此外，1（弱势传播+中立少数）作为基准参考组，发挥的集体把关效用最弱。

表 4　相对均值排序

支持虚假信息人数的变化率		极端支持虚假信息人数的变化率	
竞争传播环境的相对均值排序	均值差值（$M_{\_diff}$）	竞争传播环境的相对均值排序	均值差值（$M_{\_diff}$）
6（强势传播+中立多数）	+0.1597	5（强势传播+中立少数）	+0.4332
4（均势传播+中立多数）	+0.1185	6（强势传播+中立多数）	+0.3911
2（弱势传播+中立多数）	+0.0647	3（均势传播+中立少数）	+0.2901
3（均势传播+中立少数）	+0.0410	4（均势传播+中立多数）	+0.2216
5（强势传播+中立少数）	+0.0405	2（弱势传播+中立多数）	+0.0129
1（弱势传播+中立少数）	0（基准参考组）	1（弱势传播+中立少数）	0（基准参考组）

就极端支持虚假信息的人数变化率而言，通过 LSD 多重比较后发现，除了“1（弱势传播+中立少数）和 2（弱势传播+中立多数）”（$M_{\_diff}=-0.0129$，$p=0.376>0.05$），其他竞争传播环境之间的均值差异均显著（$p<0.01$）。由表 4 可知，在虚假信息强势传播情况下（竞争传播环境 5 与 6），极端支持者人数比例下降最为明显。此外，在极端支持者层面上，许多时候中立少数比中立多数的影响效果明显（例如，竞争传播环境 5>6，竞争传播环境 3>4）。

（二）意见交互阈值对集体把关的影响

针对研究问题 2，研究发现，在不同的意见交互阈值下，集体把关网络运行后支持虚假信息的人数变化率（$F=16228.344$，$p<0.001$，$\eta_p2=0.968$）和极端支持虚假信息的人数变化率（$F=2896.826$，$p<0.001$，$\eta_p2=0.845$）存在显著差异。此外，通过比较 η_p2 可以看出，与竞争传播环境相比，意见交互阈值对集体把关结果的影响更显著，且这种优势在支持虚假信息的人数变化率方面体现得更加明

显,见表3。

在支持虚假信息的人数变化率方面,通过LSD多重比较可以发现,低阈值-中阈值、低阈值-高阈值、中阈值-高阈值之间的均值差值分别为-0.623、-0.684、-0.061,两两之间差异显著($p<0.001$)。在极端支持虚假信息的人数变化率方面,低阈值-中阈值、低阈值-高阈值、中阈值-高阈值之间的均值差值分别为-0.600、-0.736、-0.136,两两之间差异显著($p<0.001$)。综合来看,无论是支持虚假信息的人数变化率,还是极端支持虚假信息的人数变化率,低阈值-高阈值引发的变化量都高于中阈值-高阈值引发的变化量。

(三)交互调节效应分析

为探讨研究问题3,笔者利用双因素方差分析检验"竞争传播环境×意见交互阈值"的交互作用。结果显示,支持虚假信息的人数变化率交互项($F=116.490$,$p<0.001$,$\eta_p2=0.523$)和极端支持虚假信息的人数变化率交互项($F=187.941$,$p<0.001$,$\eta_p2=0.639$)存在显著差异。为进一步揭示交互调节模式,笔者绘制不同意见交互阈值下的竞争传播环境组间成对比较森林图,见图2。图中横轴表示组间支持虚假信息的人数变化率或极端支持虚假信息的人数变化率的均值差值,纵轴表示6类竞争传播环境的编号成对组合,以95%置信区间判定显著性。

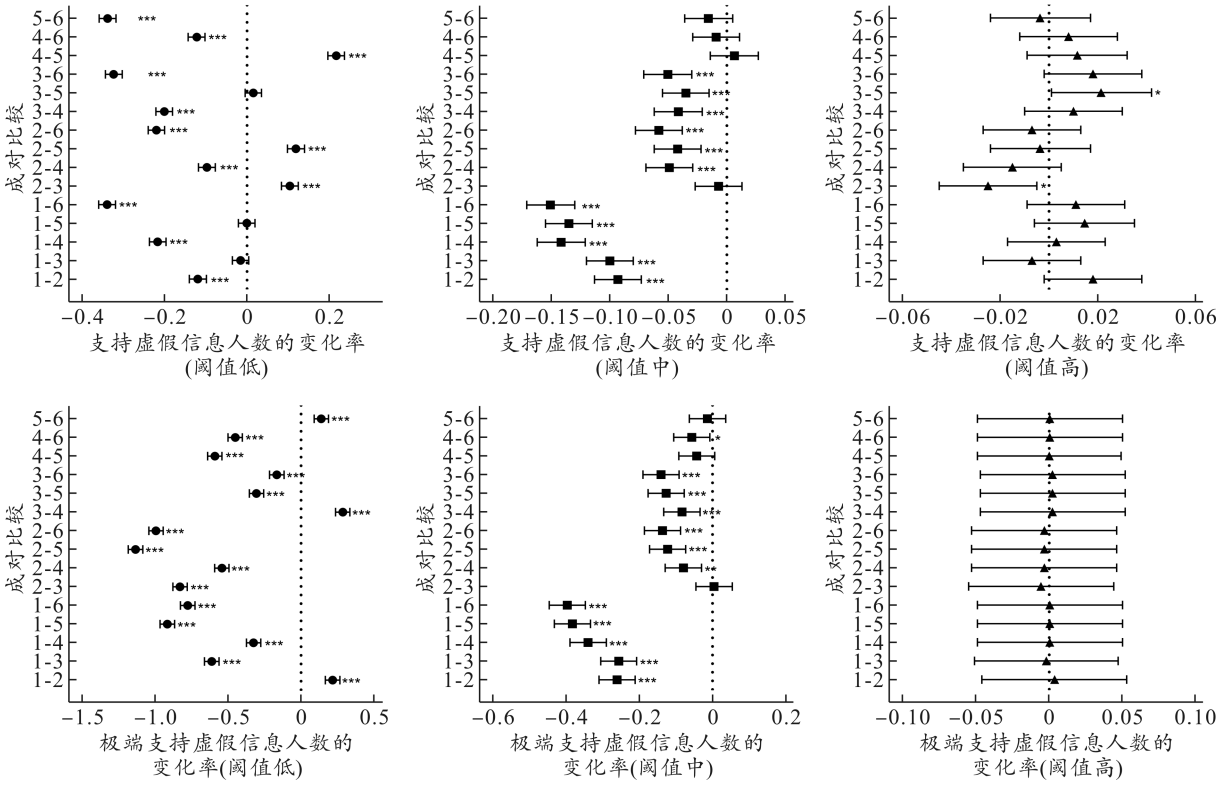

图2 成对比较森林图

结果显示,就支持虚假信息的人数变化率而言(图2上方三图),低阈值条件下不同竞争传播环境之间的差异最为显著,大部分成对比较的置信区间未跨越零点。中阈值下大多数组间比较仍显著,但效应量整体减小。在高阈值条件下,只有两组组间比较有显著差异(2-3,3-5),置信区间大部分穿过0,这说明不同竞争传播环境之间的差异已不再形成有效的集体把关差异。

极端支持虚假信息的人数变化率(图2下方三图)亦呈现出和支持虚假信息的人数变化率一致的调节趋势。不过,极端支持倾向对意见交互阈值的调节作用更敏感。具体来说,低阈值条件下,不

仅所有组别之间均出现显著差异,而且部分组别比较的变化量超过-1.0。中阈值条件下,多数组间差异保持显著;而在高阈值条件下,所有置信区间均跨越 0,集体把关效能整体趋于一致。

五、结论与讨论

(一)研究结论及理论对话

在社交媒体平台上虚假信息传播愈演愈烈的背景下,本研究利用 NetLogo 多主体仿真的方法,考察不同情境下集体把关对虚假信息传播的影响。

首先,从理论层面看,尽管有学者已经指出,集体把关受信息传播环境与用户行为结构的限制,但其代表性实证研究多聚焦于集体把关效果显著的案例,如 2011 年埃及“1·25 革命”和 2015 年天津爆炸事故等,提及的平台是 Twitter、微博等主流社交媒体。[20-21]可正如文献回顾中所提到的,在低合法性媒体上,用户对虚假信息的集体把关可能失效。[25]本研究明确,应将集体把关视为一种情境性影响机制,集体把关网络实际运行效果高度依赖竞争传播环境和意见交互阈值等变量的具体情况,这种定位有助于更加全面地分析集体把关和虚假信息传播的复杂关系。

其次,部分学界和业界人士担忧,在虚假信息强势传播情况下,集体把关不仅难以纠偏,反而可能加剧虚假信息扩散。[7]然而,本研究发现,即使是在虚假信息强势传播情况下,集体把关运行也能有效降低支持虚假信息和极端支持虚假信息的人数比例,反而是在虚假信息弱势传播情况下,集体把关运行效果不佳。这一发现丰富了信息传播风险治理的理解维度,说明集体把关的运行效果主要取决于舆论场中用户的集体反应能否被激活。在传播低频、关注度有限的场景中,由于缺乏足够的用户互动,集体把关容易陷入结构性沉默状态。

再次,在虚假信息治理研究中,人们普遍将中立者视为信息生态系统中的缓冲力量,认为中立者的存在有助于降低用户观点的极化程度[26]。笔者在此基础上进一步发现,中立多数有助于降低支持虚假信息的人数比例,但在遏制极端虚假信息支持者方面,中立少数可能更有效。在虚假信息强势传播和均势传播情况下,较低比例的中立者可能更容易渗透到虚假信息支持者的群体内部,进而发挥观点扰动作用。这提醒研究者,在去极端化的集体把关过程中,相较于中立者数量,更应关注中立者在网络中的分布情况。

最后,无论在哪种竞争传播环境下,较高的意见交互阈值都能显著提升集体把关效果。特别是在改变极端支持者态度方面,提高意见交互阈值尤为重要。此外,随着意见交互阈值从低到高提升,竞争传播环境对虚假信息支持者和极端支持者的影响会逐渐减弱甚至消失。这表明在开放的网络互动条件下,信息流动可能更有利于信息纠偏。未来研究应思考,如何让封闭圈群中的用户更加开放地接受异质信息。

(二)现实启发

在 Meta 宣布将 Facebook 的事实查核机制改为类似 X 的社区笔记模式后,围绕这种新模式的争议就不断出现,许多人质疑集体把关能否真正有效地遏制虚假信息传播。从研究的结果看,许多情况下,集体把关的确可以遏制虚假信息传播。然而,集体把关不是万能的,其有效性受竞争传播环境和用户参与度的影响。某些时候,集体把关可能因心理反弹效应失效,反而强化部分用户对虚假信息的支持程度。因此,治理虚假信息需采取多层次、多策略的综合干预手段。

第一,集体把关可能对大型谣言的辟谣效果更好,对小型谣言的辟谣效果较差。大型谣言能引发广泛社会关注,即使在虚假信息强势传播情况下,由于参与讨论的人数众多,随着不断有用户对相关内容进行审核与补充,虚假信息支持者和极端支持者的数量也会显著减少。小型谣言由于内容贴近日常且传播范围有限,往往容易被大众忽视,参与辟谣的人数也相对较少。与大型谣言相比,它们对用户的影响更为隐蔽持久,可能会在不知不觉中改变用户认知、影响用户生活。为提升集体把关在治理不同类型谣言上的效能,或可根据谣言类型采取分层治理策略:针对大型谣言,治理主体可借

助社交平台的高曝光机制,鼓励公众通过评论、转发等方式参与信息传播,推动开展集中纠偏工作;针对小型谣言,要发挥社群管理员、群聊组织者、垂类博主等的辟谣作用,通过他们激活局部网络的信息纠错能力。

第二,针对中立者比例对集体把关效果的差异化影响,虚假信息治理应根据目标对象精准配置干预方式。具体来说,面对普通的虚假信息支持者,可通过提升中立者比例、营造理性多元的舆论氛围,缩小虚假信息扩散范围。面对立场极端的虚假信息支持者,平台可以通过算法推荐有意识地将中立观点嵌入信息网络结构中,以弱连接的方式将异质观点引入信息场,逐步打破极端虚假信息支持者封闭的认知系统。

参考文献:

[1] The World Economic Forum. Global risks report 2024. (2024-01-10)[2024-12-23]. https://cn. weforum. org/agenda/2024/02/https-www-weforum-org-agenda-2024-01-global-risk-report-2024-risks-are-growing-but-theres-hope-cn/.

[2] BENNETT W L, LIVINGSTON S. The disinformation order: disruptive communication and the decline of democratic institutions. European journal of communication, 2018, 33(2): 122-139.

[3] SATO M. Twitter is adding crowdsourced fact checks to images. The Verge. (2023-05-31)[2025-01-30]. https://www.theverge. com/2023/5/30/23742851/twitter-notes-images-crowdsourced-fact-checks-misinformation-moderation.

[4] The YouTube Team. Testing new ways to offer viewers more context and information on videos. YouTube Official Blog. (2024-06-17)[2025-01-30]. https://blog. youtube/news-and-events/new-ways-to-offer-viewers-more-context/.

[5] KAPLAN J. More Speech and Fewer Mistakes. Meta. (2023-05-31)[2025-01-30]. https://about. fb. com/news/2025/01/meta-more-speech-fewer-mistakes/.

[6] SHAW A. Centralized and decentralized gatekeeping in an open online collective. Politics & society, 2012, 40(3): 349-388.

[7] The New York Times. Meta says it will end its fact-checking program on social media posts. (2025-01-07)[2025-01-27]. https://www. nytimes. com/live/2025/01/07/business/meta-fact-checking.

[8] TANDOC JR E C, LIM D, LING R. Diffusion of disinformation: how social media users respond to fake news and why. Journalism, 2020, 21(3): 381-398.

[9] SAVOLAINEN R. Assessing the credibility of COVID-19 vaccine mis/disinformation in online discussion. Journal of information science, 2023, 49(4): 1096-1110.

[10] FENG Z, DU Y, HUANG J, et al. An evolutionary approach to extreme individual impact opinions based on time sunk costs. Intelligent data analysis, 2024, 28(6): 1627-1646.

[11] PERRIER R, SCHAWE H, HERNÁNDEZ L. Phase coexistence in the fully heterogeneous Hegselmann-Krause opinion dynamics model. Scientific reports, 2024, 14: 241. [2025-01-30]. https://www. nature. com/articles/s41598-023-50463-z. DOI: 10. 1038/s41598-023-50463-z.

[12] CAHILL P, GOTTWALD G A. A modified Hegselmann-Krause model for interacting voters and political parties. Physica a: statistical mechanics and its applications, 2025, 665: 130490.

[13] LEWIN K. Frontiers in group dynamics: II. Channels of group life; social planning and action research. Human relations, 1947, 1(2): 143-153.

[14] WHITE D M. The "gate keeper": a case study in the selection of news. Journalism quarterly, 1950, 27(4): 383-390.

[15] SHOEMAKER P J, VOS T P, REESE S D. Journalists as gatekeepers//WAHL-JORGENSEN K, HANITZSCH T. The handbook of journalism studies. Routledge, NewYork, 2009: 73-87.

[16] 白红义. 媒介社会学中的"把关":一个经典理论的形成、演化与再造. 南京社会科学, 2020(1): 106-115.

[17] 胡泳, 周凌宇. 把关理论与现代社会的重构. 新闻与写作, 2021(8): 41-51.

[18] 周勇, 黄雅兰. 从"受众"到"使用者":网络环境下视听信息接收者的变迁. 国际新闻界, 2013(2): 29-37.

[19] BRUNS A. Gatewatching, not gatekeeping: collaborative online news. Media international Australia, 2003, 107(1): 31-44.

[20] MERAZ S, PAPACHARISSI Z. Networked gatekeeping and networked framing on #Egypt. The international journal of press/politics, 2013, 18(2): 138-166.

[21] ZENG J, BURGESS J, BRUNS A. Is citizen journalism better than professional journalism for fact-checking rumours in China? How Weibo users verified information following the 2015 Tianjin blasts. Global media and China, 2019, 4(1): 13-35.

[22] 申彦, 许严妍. 多元异构耦合网络中竞争性舆情信息传播研究. 计算机应用研究, 2025(7): 2123-2131.

[23] KIM J N, DE ZÚÑIGA. Pseudo-Information, media, publics, and the failing marketplace of ideas: theory. American behavioral scientist, 2021, 65(2): 163-179.

[24] HUMPRECHT E, ESSER F, VAN AELST P. Resilience to online disinformation: a framework for cross-national comparative research. The international journal of press/politics, 2020, 25(3): 493-516.

[25] STEMPEL C, HARGROVE T, STEMPEL III G H. Media use, social structure, and belief in 9/11 conspiracy theories. Journalism & mass communication quarterly, 2007, 84(2): 353-372.

[26] MATAKOS A, TERZI E, TSAPARAS P. Measuring and moderating opinion polarization in social networks. Data mining and knowledge discovery, 2017, 31: 1480-1505.

[27] VOSOUGHI S, ROY D, ARAL S. The spread of true and false news online. Science, 2018, 359(6380): 1146-1151.

[28] NYHAN B, REIFLER J. When corrections fail: the persistence of political misperceptions. Political behavior, 2010, 32(2): 303-330.

[29] AGHBOLAGH H D, ZAMANI M, PAOLINI S, et al. Balance seeking opinion dynamics model based on social judgment theory. Physica d: nonlinear phenomena, 2020, 403: 132336. [2025-05-26]. https://doi. org/10. 1016/j. physd. 2020. 132336.

[30] 祝琳琳, 李贺, 刘嘉宇, 等. 社交媒体信息影响力形成因素与关联路径研究. 图书情报工作, 2024(13): 110-121.

[31] WATTS S A, ZHANG W. Capitalizing on content: information adoption in two online communities. Journal of the association for information systems, 2008, 9(2): 73-94.

[32] CIALDINI R B, GOLDSTEIN N J. Social influence: compliance and conformity. Annual review of psychology, 2004, 55(1): 591-621.

[33] XIAO R, YU T, HOU J. Modeling and simulation of opinion natural reversal dynamics with opinion leader based on HK bounded confidence model. Complexity, 2020, 2020(1): 7360302. [2025-05-26]. https://doi. org/10. 1155/2020.

[34] DEL VICARIO M, VIVALDO G, BESSI A, et al. Echo chambers: emotional contagion and group polarization on Facebook. Scientific reports, 2016, 6: 37825. [2025-05-28]. https://www. nature. com/articles/srep37825. DOI: 10. 1038/srep37825.

[35] GUO Z, CHO J H, LU C T. Mitigating influence of disinformation propagation using uncertainty-based opinion interactions. IEEE transactions on computational social systems, 2022, 10(2): 435-447.

[36] KITSAK M, GALLOS L K, HAVLIN S, et al. Identification of influential spreaders in complex networks. Nature physics, 2010, 6(11): 888-893.

[37] NEWMAN M E J. Analysis of weighted networks. Physical review E, 2004, 70(5): 056131. [2025-05-26]. https://doi. org/10. 1103/PhysRevE. 70. 056131.

[38] BORGONOVO E, PANGALLO M, RIVKIN J, et al. Sensitivity analysis of agent-based models: a new protocol. Computational and mathematical organization theory, 2022, 28(1): 52-94.

[39] HEGSELMANN R, KRAUSE U. Opinion dynamics driven by various ways of averaging. Computational economics, 2005, 25(4): 381-405.

[40] 葛岩, 秦裕林, 赵汗青. 社交媒体必然带来舆论极化吗: 莫尔国的故事. 国际新闻界, 2020(2): 67-99.

[41] ZENG R, ZHU D. A model and simulation of the emotional contagion of netizens in the process of rumor refutation. Scientific reports, 2019, 9: 14164. [2025-11-16]. https://www. nature. com/articles/s41598-019-50770-4. DOI: 10. 1038/s41598-019-50770-4.

[42] BALENZUELA P, PINASCO J P, SEMESHENKO V. The undecided have the key: interaction-driven opinion dynamics in a three state model. Plos one, 2015, 10(10): e0139572. [2025-12-20]. https://journals. plos. org/plosone/article? id

=10. 1371/journal. pone. 0139572. DOI:10. 1371/journal. pone. 0139572.
[43] WIJENAYAKE S, van BERKEL N, KOSTAKOS V, et al. Impact of contextual and personal determinants on online social conformity. Computers in human behavior, 2020, 108: 106302. [2025-11-16]. https://doi. org/10. 1016/j. chb. 2020. 106302.
[44] FORTUNATO S. On the consensus threshold for the opinion dynamics of Krause-Hegselmann. International journal of modern physics c, 2005, 16(2): 259-270.
[45] ADAM C, ARDUIN H. Agent-based epidemics simulation to compare and explain screening and vaccination prioritization strategies. Simulation, 2024, 100(4): 335-355.

Gatekeeping Emerges: The Role of Competitive Communication Environments and Opinion Interaction Thresholds

Li Danmin (East China University of Political Science and Law)

Abstract: In the context of rapidly developing digital media, the large-scale spread of misinformation has become a major challenge threatening societal security. In recent years, technology giants such as Meta have introduced features like Community Notes, aiming to curb misinformation through a model of collective gatekeeping based on user collaboration. Against this backdrop, it is necessary to revisit the effectiveness of collective gatekeeping. This study employs a NetLogo-based multi-agent simulation to examine how collective gatekeeping influences the attitudes of misinformation supporters. The results reveal that its effectiveness varies significantly under different conditions of competitive communication and thresholds for opinion interaction. When misinformation is dominant and the opinion threshold is high, collective gatekeeping is more effective in reducing both the proportion of misinformation supporters and extremist supporters. A neutral majority contributes to decreasing the overall number of supporters, while a neutral minority is more effective in reducing the share of extremists. Moreover, the influence of the competitive communication environment on collective gatekeeping is moderated by the opinion interaction threshold. At low and medium thresholds, variations in the communication environment produce greater effects on gatekeeping outcomes. Overall, the capacity of collective gatekeeping to curb misinformation is highly dependent on specific contexts. Therefore, misinformation governance should adopt a multi-pronged strategy that integrates collective gatekeeping with other debunking mechanisms to improve overall efficacy.

Key words: misinformation; information governance; simulation

■收稿日期:2025-03-08
■作者单位:李丹珉,华东政法大学韬奋新闻传播学院;上海 201620
■责任编辑:刘金波