Interdisciplinary Science Letters 交叉科学快报, 2026, 10(5), 1250-1259
Published Online September 2026 in Hans. https://www.hanspub.org/journal/isl
https://doi.org/10.12677/isl.2026.105147

Hans汉斯

网络暴力犯罪中平台算法推荐行为的刑法归责

梁冬颖

上海政法学院刑事司法学院(纪检监察学院)，上海

收稿日期：2026年8月4日；录用日期：2026年8月27日；发布日期：2026年9月8日

摘　要

现今社会，有必要拓展网络平台间接刑事归责的教义学路径，威慑平台及时采取防范网暴风险的技术措施。针对平台“明知”网暴信息的存在，仍为其传播提供帮助或放任其传播的行为，应当以片面帮助犯追究其刑事责任。有必要对网络平台算法推荐行为的归责依据与归责困境进行研究，分析其在客观追责与主观故意方面认定的困境，并探讨相应的解决办法。据此，提出应在客观层面采用“相当因果关系说”并通过“阈值”设定与技术手段量化算法推荐的行为贡献度；在主观层面明确“明知”的认定标准，合理运用推定破解算法“黑箱”带来的证明困境，明确平台间接刑事责任的认定标准，并通过责任限缩合理地划分应对网络平台刑事归责的情形。

关键词

刑法归责，网络暴力，网络平台，算法推荐

Criminal Imputation of Platform Algorithmic Recommendation Behavior in Cyber Violence Crimes

Dongying Liang

School of Criminal Justice (School of Discipline Inspection and Supervision), Shanghai University of Political Science and Law, Shanghai

Received: August 4, 2026; accepted: August 27, 2026; published: September 8, 2026

Abstract

In contemporary society, it is necessary to activate the model of indirect criminal imputation for online platforms, thereby deterring platforms and compelling them to promptly adopt technical measures to prevent the risk of cyber violence. Where a platform, clearly knowing of the existence

文章引用: 梁冬颖. 网络暴力犯罪中平台算法推荐行为的刑法归责[J]. 交叉科学快报, 2026, 10(5): 1250-1259.
DOI: 10.12677/isl.2026.105147

of cyber violence information, still provides assistance for its dissemination or allows such information to spread, criminal liability should be pursued on the basis of unilateral aiding. It is essential to examine both the grounds for imputing liability for algorithmic recommendation behaviors of online platforms and the dilemmas involved in such imputation, analyzing the difficulties in establishing objective imputation and subjective intent, and exploring corresponding solutions. Accordingly, it is proposed that, at the objective level, the "theory of adequate causation" should be adopted, and the degree of behavioral contribution of algorithmic recommendations should be quantified through the setting of "thresholds" and technical means; at the subjective level, the criteria for determining "knowledge" should be clarified, presumptions should be reasonably employed to overcome the proof dilemma arising from the algorithmic "black box", the indirect criminal liability of platforms should be tightened, and the circumstances triggering the criminal imputation of online platforms should be reasonably delineated through the limitation of liability.

Keywords

Criminal Imputation, Cyber Violence, Online Platforms, Algorithmic Recommendation

1. 引言

网络暴力是以网络为媒介，通过诽谤侮辱、煽动滋事、公开隐私等人身攻击方式，侵害他人人格权益，危害网络空间正常秩序的失范行为[1]。从德阳女医生安某某因网暴自杀[1]到周口市第六人民医院妇产科主任邵某某坠楼身亡[2]等事件可以看出，现今社会网暴案件不断发生且影响巨大。这背后其实与网络平台的算法固有推荐逻辑密切相关，算法推荐往往会扩大网暴信息的传播范围，因为其将互动率与内容质量挂钩，而网络暴力引发的负面内容又更容易引起用户互动，算法便会通过正反馈循环不断增加争议性内容的曝光量，以此便引发了网络暴力，造成受害者社会性死亡，甚至自杀。然而，目前虽明确平台可因未履行信息网络安全管理义务而被刑事追责，但该罪名规制对象有限，难以应对平台对网暴信息消极放任的情形，对平台主动采取措施应对网暴信息的激励不足。由此可见，算法推荐在网暴事件犯罪的发酵与传播中扮演着重要角色，有必要拓展网络平台间接刑事归责的教义学路径，以此督促平台履行内容管理与风险防控义务。

2. 网络平台算法推荐行为的归责依据

(一) 对网络平台算法归责的技术依据

网络暴力事件中的负面情绪内容往往更容易引起用户的情绪与互动，高互动的内容又会被算法认定为流量热点，不断推送给其他用户并增加曝光量，逐步形成对他人的网络暴力。算法通过对用户和内容交互的数据进行分析，将其与内容的特性相融合来定位用户的兴趣方向。这一行为虽然能够挖掘新的利益点，但也会使违法有害的内容在网络空间迅速蔓延，引起网络环境恶化，从而使网络暴力内容在短期内病毒式传播，这背后的技术机理主要有以下几个方面：

[1]参见《德阳女医生自杀案宣判，三人因侮辱罪获刑，最重者被判一年半》，载 http://m.toutiao.com/group/6993421771357717000/?upstream_biz=doubao，访问日期：2025 年 10 月 11 日。
[2]参见韦嘉维：《官方通报：医生邵某某(女，57 岁)坠楼身亡，已成立调查组》，载 http://m.toutiao.com/group/7535024232611217920/?upstream_biz=doubao，访问日期 2025 年 10 月 11 日。

首先，算法对情绪内容存在优先排序机制，网暴信息通常会带有质疑、愤怒与嘲讽等多种强烈的负面情绪，这些情绪能够快速激起用户的生理与心理反应，并点击评论或者转发相关内容。情绪越强烈的内容引起的互动率越高，算法基于流量热点的考量也会分给其更多的推荐权重，从而加大对网暴内容的传播，引起更多人关注。同时，平台算法的核心目标是为平台占取市场流量，而网暴信息因为话题性强、传播速度快，天然是优质流量载体。为了迅速抓取流量，算法会实时监控平台热门话题内容并启动流量倾斜，增加其曝光权重，即使相关内容是谣言，只要互动率高，就容易被算法快速推给其他大量用户。

其次，算法通过其运行的核心机制为用户构建画像，若用户曾关注过某类争议性的内容，算法便会持续向其推送同类内容，逐渐将用户包裹在只符合自己偏好的信息气泡中。同时，当大量持有相同观点的用户被算法识别并聚集在一起，他们的观点与互动会被进一步放大，被更多同类用户看到，持相反观点的言论则会被算法过滤。用户在网络上极少看到与自己相悖的意见，便会更容易坚信自己的想法是主流观点，是正确观点[2]。

最后，算法推荐的流量导向设计成为“暴力流感”的传播引擎。“暴力流感”喻指网暴内容如流感病毒，并借助算法推荐实现病毒式扩散、圈层传染与持续蔓延。平台为了追求流量与用户活跃度，在算法设计中会刻意将网暴内容等高冲突内容设为高权重推荐对象，还通过标签分类、兴趣匹配等技术手段，将网暴内容精准推送给对应圈层用户。这种算法设计不具有中立性，而是主动筛选并助推暴力内容突破传播壁垒，使得“暴力流感”从个体言论快速升级为群体围观。

(二) 对网络平台算法归责的法律依据

与互联网初期只提供单一网络服务技术的平台不同，如今网络平台的技术资源、资金以及劳动力都达到了前所未有的规模，其“技术中立”的身份主张在算法推荐主导的服务模式下已丧失实质正当性，不再有绝对中立的身份。“技术的应用是包含价值观和主观意图的，这是法律追责之根本指向，也是平台承担法律责任的根本依据。”[3]平台在算法设计、参数调整以及干预机制等方面都存在主观能动空间，这为刑事归责提供了基础。

首先，在算法设计阶段，技术人员定义内容质量和用户兴趣等核心指标的量化标准时，很难完全排除自身主观因素对构建算法推荐系统的影响。他们依据商业目的及运营策略对算法推荐系统的程序进行创设的行为蕴含着其价值判断与对风险控制的主观选择，尤其是对内容质量、用户兴趣、风险等级等核心维度进行量化定义时，本质上就是平台对信息传播方向的预先规划[4]。这些主观因素使平台有能力从源头调控算法，完全可以从根源识别并阻止网暴信息传播。若平台明知某类指标设置可能放任侵权信息传播却未予调整，其主观过错已具备归责前提。

其次，在参数调整环节，平台对敏感词语的过滤阈值和负面内容降权系数等参数的调整，本质上就是在行使内容筛选编辑权，这是有效遏制网暴信息传播的关键关口。在“爱奇艺诉字节跳动案”[3]中，法院指出平台对算法推荐参数的调整能力使其区别于单纯的信息存储空间服务提供者，若平台未针对用户上传的侵权内容调整推荐权重，即构成知道侵权行为而未采取必要措施，需承担帮助侵权责任。这一裁判逻辑同样可以契合算法对网络暴力的技术性无视，即当平台可以通过降低网暴内容的推荐系数、提高敏感词识别精度等参数的调整来遏制网暴信息的扩散却未采取必要措施时，其不作为与损害结果的扩大之间存在因果关系。

最后，在干预机制方面，平台人工审核的响应速度和算法熔断机制等运营策略，也是影响违法信息扩散范围大小的重要因素。这些干预机制的运行效果完全取决于平台对算法监测的资金投入与流程设计。事实上，这也属于可以自主控制的人为决策过程，若平台为了扩大收益而消减在算法更新升级方面的资

[3]参见北京市海淀区人民法院民事判决书(2018)京 0108 民初 49421 号。

金投入，导致落后的算法系统未能识别出当下行业内普遍使用的新型算法系统已经能够普遍识别出的网暴内容，在其有能力更新系统而不更新的情况下，其主观过错已形成归责前提。可以说，人为决策行为贯穿平台算法推荐系统形成的全过程，平台不能再以技术中立为由完全免责，而应对算法在网暴事件中承担的帮助角色受法律担责。

3. 网络平台算法推荐行为归责困境

(一) 算法推荐的客观归责障碍

在算法推荐过程中，推荐行为、网暴行为与危害结果之间的关系具有多环节性和不确定性等特征。算法推荐是以用户画像为基础的自动决策过程且推荐逻辑具有“黑箱”属性；即使平台本身也很难准确追踪到哪些具体参数推动了违法有害信息的扩散，更无法准确评价算法推荐对危害结果的贡献度，这使得司法实践中难以评定算法推荐行为与危害结果之间的因果关系。一方面，算法推荐行为的介入使得传统的线性因果关系被拆解为多主体、多环节的因果链条。算法推荐以用户的历史行为数据为基础，通过复杂的运算生成最终的推荐结果，该算法的自主决策过程使因果关系难以追溯，即使是平台的技术人员也难以精准定位到哪一算法推荐行为推动了某一网暴信息的扩散传播。另一方面，即使能够证明算法推荐与危害结果之间存在客观因果关系，平台的算法对危害结果的实际贡献度也难以量化。刑法归责要求行为人仅对自己行为造成的危害结果担责，即责任与行为的危害程度相适应。但在网络暴力事件中，算法推荐行为的危害贡献度往往与其他因素相互交织，其独立造成的危害难以被单独量化[5]。此时，无法确定“若没有算法推荐，危害结果是否仍会发生”等问题，进而难以判断平台算法推荐行为带来的后果是否达到了刑事追责所需的“情节严重”的标准。这种贡献度的模糊性，使得平台即使存在一定的不作为，司法机关也难以认定其行为构成犯罪，只能在民事责任或行政责任的框架内追责，从而限制了刑事责任的适用范围。

此外，当算法被用于传播违法信息时，如何判断平台已履行相应义务，也是客观归责的难题。应采取有效措施防止违法信息传播，但平台的必要干预义务在算法推荐情形下并不清晰。一方面，目前并无明确的法律标准来明确平台应在何种程度上主动审查与过滤算法推荐过程中出现的违法信息。以微博与抖音等用户自行创作内容的平台为例，其每日产生的数据量巨大，要求平台人工审查每条被算法推荐的信息显然不切实际，但若仅依靠算法进行初步筛除又很难确保对所有违法有害信息作出有效甄别和拦截。另一方面，平台收到用户的投诉或举报后该采取何种针对措施，目前亦无明确统一的标准，且不同措施对平台运营与用户权益影响不同，这使得实践中很难准确判断平台义务履行情况，进而影响客观归责的准确性。

(二) 主观故意的认定难题

在算法推荐领域，如果平台存在明显的指令要求算法推荐某些特定违法信息，或其工作人员明知平台存在网暴信息，仍通过算法向其他网络用户推荐相关内容，这些情形下认定平台主观故意较为容易。但在实际应用中，多数情况都是算法根据自动化数据分析来进行推荐，算法的“黑箱”属性使网络平台对网络暴力信息的传播是否“明知”的认定陷入困境。算法的“黑箱”与算法推荐逻辑人为设计并不冲突，二者是客观特性与主观构建的不同维度。算法“黑箱”属性意味着平台虽能控制算法系统的设计目标与核心参数，但其无法预知每一次自动推荐的具体结果，更难以实时监控海量推荐内容中夹杂的网暴信息。同时，在现有的技术条件下，平台难以通过算法语义识别系统实现对所有隐性网暴信息的事前识别，也无法完全掌控算法推荐的具体流向，这使得司法实践中难以证明平台主观故意。“算法黑箱使得平台的主观心态呈现出‘模糊化’特征，传统以‘明知’为核心的过错认定标准，难以适配算法时代的平台责任场景”[3]，这种模糊性直接导致刑事归责中对主观要件的证明陷入困境。例如，在“搜索提示词

案”中[4]，平台以“搜索提示系算法根据用户历史记录自动生成，无人工干预”为由进行抗辩，法院最终也认定平台不构成“明知”。当算法通过深度学习自主生成推荐内容时，由于算法的动态迭代性与语义识别的局限性，平台往往无法提前预判具体的推荐结果，导致平台对算法自主推荐内容可能引发网络暴力这一事实缺乏实质认知。

(三) 依据司法解释追责的局限性

2023 年，最高人民法院、最高人民检察院与公安部联合印发了《关于依法惩治网络暴力违法犯罪的指导意见》[5]，该意见明确界定网络暴力事件中平台应当承担的刑事责任，即平台在发觉网暴信息后，未依照法律法规对信息网络安全进行管理，在满足一定条件的情况下，可能会构成拒不履行信息网络安全管理义务罪。然而，该项规定的罪名设置与规范对象过于单一，一直被学界诟病[6]。该罪在算法场景中网络暴力犯罪的适用困境，本质上是制度设计与新型危害行为的适配性不足，具体体现在以下四个层面：其一，行政前置程序导致追责被动滞后。本罪以“经监管部门责令改正而拒不改正”为入罪前提，刑事追责完全依附于行政执法程序，大量平台怠于履行算法治理义务但未被行政机关责令整改的案件，无法进入刑事评价的范围；而网络暴力的传播又具有瞬时性、裂变式等特征，若等待行政责令改正下达后再启动刑事规制，往往已经错过了危害结果的干预窗口期[7]。其二，该罪的入罪量化标准与网暴危害属性不匹配。本罪“致使违法信息大量传播”的入罪标准，多以信息条数或点击量作为量化依据，但网络暴力的核心危害在于对公民人格权、精神健康的持续性侵害，单纯地依靠信息传播数量无法完整地评价其社会危害性，这也导致司法实践中难以把握网暴场景下对平台的入罪阈值。其三，该罪的规制形态与法益覆盖存在局限。本罪仅能评价平台不履行管理义务的不作为形态，无法覆盖平台基于流量利益主动调整算法参数、定向倾斜等推荐网暴内容的作为行为，且本罪的核心保护法益是信息网络管理秩序，无法充分回应网络暴力对公民人格权、生命健康权等权利的严重侵害，存在法益评价的缺位[8]。其四，对该罪的司法适用整体呈现空置化。从裁判文书公开情况来看，本罪在网络暴力治理场景中的适用案例极少，远未发挥预期的规制效能。因此，仅依靠该罪名难以倒逼平台主动优化算法治理机制，应当重视平台在网暴案件中作为共犯的间接刑事责任。

4. 算法推荐的归责路径认定

(一) 算法归责的客观方面

1. 因果关系的判断

在判定算法推荐行为与危害结果的因果关系上，不能简单采用传统的线性因果逻辑，因为算法推荐信息传播的过程涉及多个环节和多种因素的交互作用，包括用户自主转发和平台间信息转载等介入因素，因此可以运用“相当因果关系说”等理论进行综合判断。第一，筛选条件关系，排除与网暴事件没有关联的平台正常运营行为。第二，判断相当性基础。若平台主动推荐含有网络暴力内容的信息且该行为在通常情况下会促使此内容大规模传播并侵害受害者权益，就能初步认定其具有相当性。若平台仅推荐中性或积极内容，用户私自篡改后传播引发网暴事件，则不具有相当性。第三，考量介入因素的影响。若平台推荐网暴内容，用户或其他平台常规转发分享属于信息传播常规途径，不阻断行为与结果的关系。若用户或其他平台故意篡改后传播，超出平台可预见的范围，则属于异常介入因素，可阻断因果关联。从相当因果关系角度看，在“快播公司传播淫秽物品牟利案”中[6]，快播公司放任和加速传播淫秽物品的

[4]参见陈恒星：《北京互联网法院宣判首例“搜索提示词”算法侵权案》，载 https://xinwen.bjd.com.cn/content/s6731bd81e4b0a29cf60e864f.html，访问日期：2025 年 11 月 15 日。
[5]参见《关于依法惩治网络暴力违法犯罪的指导意见》，载公安部网站 https://www.mps.gov.cn/n6557558/c9221769/content.html，访问日期：2026 年 7 月 31 日。
[6]参见北京市海淀区人民法院刑事判决书(2015)海刑初字第 512 号。

行为，通常会导致淫秽视频大量传播以及公司因流量等增加而获利的结果，其行为与危害结果之间具有相当因果关系，构成传播淫秽物品牟利罪。

同时，针对算法推荐行为的贡献度难以量化等问题，可以通过“阈值”[9]界定无危害与可罚行为的边界。具体来说，将算法贡献度的量化过程分成三个组成部分，即分别设定为基础曝光阈值、互动累积阈值及圈层扩散阈值。例如，设定算法将该网暴内容在 24 小时内推送给 1 万用户为基础阈值，引发 500 条恶意评论为互动阈值，渗透 100 个关系链为扩散阈值。若平台的算法推荐后果突破任意一个阈值，就可以认定其贡献度达到了可评价的标准，当网暴行为构成犯罪时，可以追究平台的刑事责任。此外，还可以利用区块链技术的去中心化和不可篡改等特性，打破算法的“黑箱”属性，对网络暴力事件中的信息传播过程进行全程记录和存证[10]。这样一来，对于网暴犯罪就可以准确追溯算法推荐行为在整个因果链条中的作用和影响，清晰地展示算法推荐如何推动了网暴信息的扩散，为因果关系的认定提供可靠的证据支持。

2. 明晰平台的“必要注意义务”

网络平台基于对算法的控制支配与监管能力，对阻止违法信息的传播具有排他性的支配地位，进而具有法律意义上的保证人地位，负有保护处于危险状态法益的刑事作为义务[11]。在算法推荐系统推动网暴信息扩散，形成网络暴力进而构成犯罪的情况下，平台对算法推荐系统的支配力使其应履行对网暴信息的必要注意义务，否则需要承担相应的刑事责任[4]。平台的必要注意义务主要从几个方面界定：

首先，各类网络信息服务平台都应进一步明确主体责任，不仅要健全用户注册、个人信息保护、内容发布审核和识别处置等制度，还要严格执行对网络实名制的管理，对提供信息发布、即时通讯服务的，要依法对用户进行实名认证，并要求其加强账号信息的管理，为遭受网络暴力信息侵害的相关主体提供账号信息认证的帮助。同时，网络平台要保证算法的透明度，平台需通过算法备案与审计报告等方式，向监管部门说明其推荐机制的设计原理。备案内容可用于日后发生客观损害结果后对平台主观过错进行考量，并可供监管部门在算法生命周期中持续使用及定期审核。对于删除的网络暴力信息，算法需完整记录网络暴力的传播路径、用户操作日志等数据，保存期限不少于 6 个月。

其次，构建智能语义识别模型对侮辱诽谤等网暴言论实时拦截与动态核验，推动多模态语义理解模型的研究与创新，扩大对隐性贬义及侮辱词的识别捕捉，并由平台定期更新算法识别系统。同时，平台应进行风险评估与技术迭代，定期更新算法识别系统，与行业内标准统一，避免已知的算法漏洞对网络环境产生不利影响。同时，行业内要推动建立统一的网络暴力信息分类标准和多模态审核技术规范，鼓励平台共享恶意账号特征库，提升整体行业的治理质量。

接着，每日信息量生成巨大的网络平台应构建算法与人工双层审核机制，算法需依据多模态语义识别系统对用户发表的内容进行初筛，拦截含有侮辱污蔑性词汇及隐形网络暴力内容等网暴信息，对无法准确识别但判定可能构成网暴的内容进行疑似信息标记，由工作人员在 24 小时内复核审查后再认定是否发布。对于用户举报违规违法的信息，平台应立即屏蔽该内容并在 24 小时内审核完成，向举报者发送审核结果，认为不属于网暴内容的信息应当立即解除屏蔽。若某信息短时间内浏览量激增，举报量显著增长，算法需触发紧急预警机制，立即降低推荐权重占比并暂时下架该内容。鉴于网暴行为的多环节属性，起初没有被认定为违法违规的信息，后续经过发展仍然可能形成网络暴力，因此不应当限制用户的举报次数，但应当设置用户的举报时间间隔，如下一次举报距离上一次举报应间隔 24 小时。对于用户超 3 次举报的内容，平台应组建评估小组对该内容进行审核并向举报者出具结果。若举报者在第一次举报后得到平台不认为是网暴信息的结果，后续信息经发酵又形成网络暴力，举报者没有再举报，但平台根据现有技术能够主动识别却未处理，仍然属于未尽到注意义务。

最后，平台针对侵权行为采取的措施应当有效，才能属于尽到注意义务。如在“腾讯诉微播视界(抖

音)侵害《云南虫谷》信息网络传播权纠纷案中”，抖音没有尽到平台的监督义务，在收到了 3 次警示函、108 次侵权函后，依然没有在联想推荐位撤下“云南虫谷”等关键词，甚至还有部分函中侵权链接并未及时删除 [7]。平台的措施未达到“有效制止侵权”的实质要求，属于未履行必要注意义务。值得注意的是，若平台已尽到必要注意义务且主观无故意，危害后果仍然发生，平台可以以此进行免责抗辩。例如，在“衣念诉淘宝‘美丽侠女’售假案”中 [8]，淘宝网根据衣念公司要求，暂时保留涉嫌侵权信息，并提供了刘某的身份信息并在诉讼后随即删除相应信息。作为网络服务提供者，淘宝已尽合理义务且主观没有过错，淘宝可以此免责。

(二) 平台主观过错的认定

当有充分证据证明平台知道算法推荐的内容具有违法性却依然允许或放任该信息传播时，认定平台主观“明知”较为容易。这类证据主要包括平台内部审核记录中已标注该信息是违规信息却未采取任何措施处理；平台工作人员与信息发布者私下沟通，在明知其发布的信息侵权或违法的情况下，仍然通过调整算法规则进行相关内容推荐；或平台收到监管部门或权利人的明确通知后，不理会相关投诉与举报或审查后确认某信息侵权或违法，但为了占据市场并不停止相应内容的推荐等。而对于算法“黑箱”及自动化决策带来的认定困境，一方面，可以利用区块链技术的去中心化和不可篡改等特性，打破算法的黑箱属性。另一方面，可建立算法日志留存与溯源制度要求平台对算法推荐的核心决策数据建立可追溯日志，明确算法是如何进行决策推荐并传播内容。当网络暴力内容经算法推荐扩散时，便可通过日志倒查平台对同类暴力内容的历史推荐效果是否存在认知。同时，建立网络暴力内容超额流量收益与平台获利的关联性关系，当网络暴力内容的推荐流量占比超过同类型正常内容的 3 倍及以上，且平台通过该内容获得的利润远远超过常规阈值，可直接推定平台明知网络暴力内容。

由于算法系统的隐匿性与可更新性，在一些特定的情形下，也可推定平台对算法的推荐行为“明知”。具体而言，若算法推荐的信息具有明显违法特征且处于行业内通常能够识别的“红旗”状态，或者平台具备相应技术能力与管理机制能够识别出该信息违法，即可推定平台“明知”。同时，若平台的算法推荐系统存在明显漏洞或缺陷，导致大量违法信息持续广泛的传播，或在过去经营过程中曾因算法推荐违法信息而被行政处罚或被起诉，平台却未能及时采取合理的技术升级措施和管理优化措施加以改进，同样可以基于平台对算法运行的管理职责以及其对风险的持续忽视与放任和屡教不改的态度推定其对违法信息的传播“明知”。但若平台已经建立完善的算法审查机制并定期更新多模态语义理解模型，只是因为技术有限或意外事件致使违法信息传播，因其已尽到“善良管理人”的注意义务，不可推定其主观存在过错。需要明确的是这一推定规则的适用应当受到严格限定以避免客观归罪的风险：一方面，该推定规则仅适用于无争议的明显违法内容，即涉案信息属于公然侮辱、捏造事实诽谤、非法披露他人隐私等具备确定违法性的内容；而仅存在道德争议、观点分歧的言论不得作为推定的基础事实，禁止将价值判断混同于违法性判断。另一方面，基础事实需要达到充分性标准，单一的流量异常、收益提升仅可作为案件线索，必须搭配同类网暴内容已被多名用户多次举报并经平台核实、监管部门已就同类内容向平台作出行政提示、整改要求或平台曾因同类算法推荐违法内容被行政处罚或生效裁判认定为侵权等强化事实，方可启动主观“明知”的推定。当然，平台亦有权通过提交相反证据推翻推定，具体抗辩依据可分为以下四个方面[8]：第一，算法合规审计证据。平台可以提交经第三方专业机构出具的算法治理审计报告，证明其推荐系统已经设置了网暴内容降权、拦截和熔断规则，且相关技术标准符合行业通用规范。第二，

[7]参见《〈云南虫谷〉短视频侵权案背后，我们怎样看待新时代的影视剧版权保护？》，载 https://www.pkulaw.com/pal/a3ecfd5d734f711dd6322486d7a48ae2f5c28d0efe571407bdfb.html，访问日期：2025 年 11 月 16 日。
[8]参见陈琼珂：《网店售假，淘宝是否该被“连坐” 法院：若已尽到合理注意义务无需担责》，载 https://www.shzcfy.gov.cn/detail.jhtml?id=267269，访问日期：2025 年 11 月 16 日。

治理资源投入证据。平台可以举证其对内容审核团队规模、多模态语义识别技术的研发投入、算法安全维护成本等人力与财务数据，证明其已经投入了与自身经营规模、技术能力相匹配的治理资源。第三，处置操作留痕证据。平台可以提交涉案内容的算法推荐日志、审核处置工单和降权下架操作记录等，证明其在发现涉案内容后的合理时限内已经采取了必要处置措施。第四，技术局限说明证据。平台可以提交行业内通用语义识别准确率的对标数据，证明涉案内容属于隐喻、暗讽等隐性网暴言论，其识别难度超出当前行业普遍技术能力范畴。上述推定并没有倒置刑事诉讼的证明责任，控方始终需要对推定的基础事实承担完全的证明责任，且需达到排除合理怀疑的标准；平台一方仅需提出合理反证，使对主观“明知”的认定产生合理怀疑即可，无需对自身无过错承担举证责任，以此坚守无罪推定的基本原则，防止平台刑事责任的不当扩张。

综上，在网暴行为构成犯罪的事件中，虽然平台与网暴直接实施者之间没有意思联络，但在明知网暴信息存在的情况下，其技术帮助行为在客观上降低了犯罪实施的难度且主观上存在过错，满足了单向共犯的成立要件[12]。基于此，对平台明知存在网络暴力信息的情况下仍提供帮助传播的行为，可按照共同犯罪处理，以片面帮助犯追究其刑事责任。片面帮助犯理论虽在大陆法系刑法教义学中存在一定争议，即否定说认为共犯成立需双向意思联络，单向帮助不具备共犯的主观联结；但肯定说已经逐渐成为有力观点，该学说的核心逻辑在于共犯的可罚性本质在于对正犯行为的加功与对法益侵害的促进，单向的帮助行为同样具备因果贡献与主观可谴责性[13]。在平台算法助推网暴信息扩散的场景下，采纳片面帮助犯具有显著的制度优越性，是相较于拒不履行信息网络安全管理义务罪更优的规制选择。第一，前者行为覆盖更加全面。该路径既可以评价平台放任算法传播网暴内容的不作为，也可以评价平台主动调整推荐参数、为网暴内容定向引流的作为，弥补了拒不履行信息网络安全管理义务罪仅规制不作为的制度局限。第二，前者法益适配更加精准。片面帮助犯依附于侮辱罪、诽谤罪等网暴相关罪名定罪处罚，能够直接对接网络暴力所侵害的公民人格权等核心法益，实现对被害人权益的充分刑法评价。第三，前者追责启动更加灵活。该路径不受行政前置程序的限制，在符合犯罪构成要件时可直接启动刑事追责，避免了行政监管缺位导致的规制真空。第四，加入片面帮助犯可以使罪名体系更协调。该路径与拒不履行信息网络安全管理义务罪可以形成梯度互补，平台因疏忽未履行算法治理义务、主观恶性较低且经监管部门责令改正却拒不改正的，可以适用拒不履行信息网络安全管理义务罪；平台明知网暴信息存在仍通过算法提供传播帮助、主观恶性较高的，可以适用片面帮助犯，以此构建层级化的责任体系，实现罪责刑相适应。因此，细化平台间接刑事责任的认定规则，以片面帮助犯理论规制平台的算法传播帮助行为，兼具理论正当性与实践必要性。

(三) 算法归责的合理限缩

法律规范在设定第三方义务时需遵循比例原则，考虑第三方的经济负担与执行成本。当前，网络暴力行为主要集中在少数主流社交网络平台，对这类行为的管控主体也应限于这些主流社交网络平台范围内，而不是对所有互联网平台都施加严格技术治理责任。对此，可参考域外“数字守门人理论”，该理论最早源于西方资本市场的第三方监管研究，后被拓展应用于数字平台治理领域[14]，该理论引入我国后已被学界广泛讨论，但其理论基础的本土化适配和与我国现有监管体系的衔接路径仍存在一定学术争议，对此本文仅采纳其“义务强度与主体能力相匹配”的核心思路，而非直接套用域外的制度定位。尽管国内目前没有关于“数字守门人”的具体认定标准，但现行立法已体现出差异化规制的倾向，可为责任分级提供规范依据[1]。如《个人信息保护法》第 58 条 [9]就对提供重要互联网平台服务、拥有庞大用户群体

[9]参见《中华人民共和国个人信息保护法》第五十八条，载中国人大网 http://www.npc.gov.cn/c2/c30834/202108/t20210820_313088.html，访问日期：2026 年 7 月 31 日。

且业务类型复杂的个人信息处理者引入了“守门人条款”[15]。具体来说，网络暴力技术治理主体的范围可明确在两个方面：一方面，它是一个社会化娱乐平台，涵盖即时通讯、游戏休闲、视听服务、直播、短视频等多种类型，既是人与人之间互动的重要载体，也是网络暴力发生的主要平台，应成为网络暴力治理的重点对象。另一方面，它被定位为“大型平台”与“超大型平台”。《互联网平台分类分级指南(征求意见稿)》[10]曾提出将网络平台划分为超级平台、大型平台及中小平台，其中超级平台和大型平台因用户数量庞大，资源储备丰富且技术水平先进，更适合承担网络暴力中的“守门人”角色。而中小平台影响有限，发展资本及用户数量都处在较低的水平，若对其也进行刑事归责，容易抑制其发展，因而对其治理责任应更多聚焦于被动响应层面，如收到网暴投诉后及时删除相关内容、提供侵权用户信息等。这一分级归责的逻辑合理性也可从欧盟《数字服务法》(Digital Services Act, DSA)的制度设计中获得印证：其一，DSA 根据平台用户规模、市场影响力将服务提供者划分为不同层级，对超大型在线平台施加算法风险评估、年度独立审计、算法透明度公开等更严格的治理义务，中小型平台仅承担基础的内容处置义务，其核心逻辑即“义务与能力相匹配”，这为我国差异化设置平台算法注意义务提供了域外经验支撑[16]。其二，DSA 明确要求超大型平台不得通过推荐算法放大有害、暴力内容的传播范围，需对违法内容采取降权、限制可见性等干预措施，且需向监管机构说明算法推荐的核心运行逻辑。这种责任划分不仅能让监管资源集中投向风险更高的大型平台，更能给中小平台留出合理的发展空间，符合平台经济健康发展的整体需求。

5. 结语

现今社会，网暴案件呈现高发态势，网络平台作为综合型中间服务提供者，其能精准捕捉用户的行为偏好，进而实现个性化产品投送与推荐以获取利益，所以它们有能力通过技术资源方面的优势，在打击网络暴力犯罪中贡献力量，或通过预先审查机制来阻止网络暴力等群体极化事件发生。但目前的《指导意见》规制对象比较单一，难以满足督促平台采取有效措施防范网暴事件的需求，应当拓展网络平台间接刑事归责的教义学路径，在客观方面以“相当因果关系说”进行判断，并明晰平台的“必要干预义务”；在主观方面明确“明知”的认定标准。若能构建以上全链条机制，能有效督促平台采取预防或干预措施，促进网络空间和谐发展。

参考文献

[1] 王燃. 论网络暴力的平台技术治理[J]. 法律科学(西北政法大学学报), 2024, 42(2): 121-134.

[2] 李凯林. 谣言、信息茧房与群体极化: 短视频平台网络暴力的治理[J]. 甘肃行政学院学报, 2023(3): 106-113+123+128.

[3] 张凌寒. 网络平台监管的算法问责制构建[J]. 东方法学, 2021(3): 22-40.

[4] 龚文博. 论网络暴力治理中的平台间接刑事责任[J]. 法制与社会发展, 2024, 30(6): 147-162.

[5] 冯明昱. 累积犯视阈下网络暴力的分层归责与治理机制研究[J]. 南昌大学学报(人文社会科学版), 2025, 56(5): 121-131.

[6] 夏伟. 论数据犯罪的立法重塑[J]. 法制与社会发展, 2023, 29(4): 173-190.

[7] 熊波. 网络服务提供者刑事责任“行政程序前置化”的消极性及其克服[J]. 政治与法律, 2019(5): 50-65.

[8] 李本灿. 网络暴力治理中的平台责任及其限度[J]. 华东政法大学学报, 2026, 29(3): 48-62.

[9] 张勇, 王杰. 数据累积犯的刑法规制[J]. 济南大学学报(社会科学版), 2024, 34(3): 137-149.

[10]参见《互联网平台分类分级指南(征求意见稿)》第三部分，载 https://mp.weixin.qq.com/s?__biz=MzUyNTgwMDI3Ng==&mid=2247522893&idx=2&sn=1d79ff07947abf5325d3d52361607d8f&chksm=fa1a5c76cd6dd560ee28c307e03a37d9d94a2d8b22908b5899547d1b20185ce8cddf040a3f84&scene=27，访问日期：2026 年 7 月 31 日。

[10] 朱浩齐. 技术驱动网络暴力治理[J]. 中国信息安全, 2024(9): 43-46.
[11] 万千慧. 论新兴数字权利与宪法的交互模式及其保障路径[J]. 东南法学, 2023(1): 68-83.
[12] 高冉. 网络暴力犯罪中的网络平台的刑事责任研究[D]: [硕士学位论文]. 呼和浩特: 内蒙古财经大学, 2025.
[13] [日]山口厚. 刑法总论[M]. 第 3 版. 付立庆, 译. 北京: 中国法制出版社, 2018.
[14] 单勇. 论网络犯罪的看门人规制[J]. 南京大学学报(哲学·人文科学·社会科学), 2023, 60(5): 57-74+157.
[15] 周汉华. 《个人信息保护法》“守门人条款”解析[J]. 法律科学(西北政法大学学报), 2022, 40(5): 36-49.
[16] 丁晓东. 第二代互联网规则?——从欧盟《数字服务法》看平台责任的未来[J]. 华东政法大学学报, 2026, 29(1): 22-37.