社交平台反网络暴力专项合规机制的有效构建

■印　波　庄新宇

从分析实证主义观之，网络暴力治理已由司法层面拓展至平台层面。由于网络暴力的统一概念尚未生成，社交平台对网络暴力的能动性治理缺失以及平台责任与合规责任分立错配，社交平台网络暴力治理依旧存有体制障碍。参照域外网络暴力平台治理的分散型立法、奖惩式驱动与分配型责任等经验，我国应对网络暴力概念外延加以延展，采激励式驱动以增强平台合规动力，作平台内源性网络暴力与外部性网络暴力的类型化分解，明确社交平台反网络暴力合规的主体责任。社交平台可以构建PDCA合规管理机制，合比例制定并实施专项合规计划，完善平台用户教育引导机制、技术规制机制与诉讼支持机制，接受第三方评估并依据测评改进合规计划，以实现社交平台反网络暴力的有效合规治理。

[关键词]社交平台；网络暴力；专项合规机制；企业合规；合规责任

[中图分类号]G206　[文献标识码]A　[文章编号]1004-518X(2024)01-0093-11

[基金项目]最高人民检察院检察理论研究一般课题“检察业务考核指标体系研究”(GJ2022C33)、中国政法大学网络法学研究院支持项目“当前社会治理体系下平台责任研究”

印波，中国政法大学刑事司法学院教授、博士生导师。(北京　100088)

庄新宇，中国政法大学科技与法治研究中心研究人员。(北京　100088)

一、问题的提出：社交平台网络暴力治理困局

截至2023年6月，我国网民规模达10.79亿人，即时通信、网络视频、短视频等社交平台用户规模稳居互联网应用前三位。[①]网民及社交平台的庞大基量催生对网络暴力治理的潜在隐忧。全国信息网络安全协会联盟调研报告显示，在全国31个省份收集的303.1776万份有效调查问卷中，有超六成网民曾遭受过网络暴力。[②]党的二十大报告指出，我国应健全网络综合治理体系，推动形成良好网络生态。社交平台作为网络生态的重要一环，在网络暴力治理等网络综合治理体系中发挥着关键作用。网络暴力不仅产生了危害平台用户身心健康、败坏网络生态、诱发线下暴力等社会失序行

为,而且还加重了社交平台企业法律责任③等合规风险。社交平台网络暴力治理目前处困局之中。一方面,各平台现有反网络暴力治理手段难以遏制网络暴力事件频发,网络暴力形态翻新,部分用户对网络暴力治理的情感倾向由振奋转变为失望。[1]另一方面,社交平台治理网络暴力的义务与责任被过分强调,而相关权力(利)及激励机制却鲜有规定,平台反网络暴力专项合规机制等合规治理工具尚未被政策文件激活,平台缺乏对网络暴力的能动治理内驱力。

我国网络暴力平台治理基本路径业已形成,即由受害网民向司法机关提起民事诉讼或刑事自诉以寻求司法救济,拓展至国家互联网信息办公室、司法机关等有关部门主导,网络社交平台及平台用户参与的协同治理。[2]随着网络暴力治理由司法层面拓展至平台层面的实务转向,社交平台网络暴力的治理理论根基及对策路径等内容仍存在争议。部分学者以《网络安全法》《数据安全法》《个人信息保护法》等法律规定的安全管理义务[3]、避风港原则和红旗原则[4]以及守门人义务[5]等平台责任制度理论为立论基点,论述社交平台作为网络服务提供者所应承担的网络暴力治理主体责任限度及必要举措。也有学者指出规制网络暴力是平台技术优势的应有之责。[6]还有学者对社交平台提出以培育用户数字公民伦理来规制网络暴力的新职责。[7]然而,通过对中国知网、HeinOnline等期刊网站检索,暂未发现有学者从合规治理视角出发,关注社交平台的企业属性,审视当前社交平台诸多网络暴力治理机制之于企业合规机制的体系定位。

我国合规制度建构经历了自2006年起由金融监管领域等专门领域合规,至2018年出台《中央企业合规管理指引(试行)》以实现中央企业全面合规,再至2020年由检察机关主导的涉案企业合规制度改革,以刑事合规为媒介向社会全领域全类型企业扩容,并由涉刑事案件企业合规改革向涉民商事案件合规与行政监管合规迈进。我国理论界对“企业合规”的内涵研讨亦经历了同步变迁,整体由“企业守法即合规”“高管守法即合规”等单一内涵说转向企业合规管理与国家合规激励相结合的“两视角说”[8]以及企业积极合规、避免合规风险与受国家合规治理相结合的“三层性质说”[9]等复合内涵说。本文主采复合内涵说。相较“大而全”却易沦为“纸面合规”的全面合规,企业针对其商业模式所关涉的某一专门领域而构建专项合规机制,往往更易达成有效合规的既定目标。社交平台企业的反网络暴力专项合规机制是指社交平台为了防止和打击网络暴力行为而采取的一系列规定和措施。本文通过比较研究方法,类型化分析欧美部分国家对网络暴力认定与社交平台网络暴力合规治理责任的立法殊同,以及整合域外对平台责任分配的治理经验,以探明激活网络暴力平台治理效能的应然逻辑,实现平台反网络暴力专项合规机制的有效构建。

二、平台治理网络暴力困局的三重溯源

当前我国社交平台反网络暴力专项合规机制尚未有效构建,进而陷入各平台反网络暴力治理手段实效有限与缺乏能动治理动力的困局。通过对困局成因溯源分析,笔者发现如下三重原因。

(一)网络暴力的统一概念尚未生成

我国现行法律虽未明文提及“网络暴力”,但在司法实践中均视网络暴力为侮辱谩骂、造谣诽谤、侵犯隐私等语言暴力行为的线上表现。我国部门规章、司法解释等规范性文件中对“网络暴力”一词的规范解释经历了由音像暴力元素、黑恶势力犯罪的“软暴力”再至“人肉搜索”“网络中伤”等行为的演变。[10]尽管作为网络生态治理主管部门的国家互联网信息办公室曾出台过《微博客信息服务管理规定》第11条、《互联网跟帖评论服务管理规定》第7条、《互联网论坛社区服务管

理规定》第7条等社交平台言论治理的相关部门规章,但多以“谣言或不实信息”“含有法律法规和国家有关规定禁止的信息”指称网络中伤行为。直到在2019年发布的《网络生态治理规定(征求意见稿)》中才首次提及“网络暴力”一词。

自“刘学洲事件”等恶性网络暴力事件滋生后,为加强网络暴力专项治理,《关于切实加强网络暴力治理的通知》以及《网络暴力信息治理规定(征求意见稿)》等文件相继出台,但两份文件对网络暴力的概念仍未统一。2023年9月出台的《关于依法惩治网络暴力违法犯罪的指导意见》亦未明确网络暴力内涵,而是将网络暴力外延拓展为网络诽谤、网络侮辱、侵犯公民个人信息、线下滋扰、恶意营销炒作以及可供延展解释的网络暴力违法犯罪等一系列行为。

学界网络暴力概念的生成路径可概括为“网络媒介说”“特定行为说”“起因说”三类。“网络媒介说”将网络暴力理解为以网络为媒介实现的暴力,比如有学者认为网络暴力是发生在网络空间中的舆论暴力;[11]“特定行为说”则认为网络暴力是通过某些特定网络行为实现的,通常通过采取列举或穷举网络暴力实施手段来界定网络暴力的内涵;“起因说”主要从网络暴力的生成机制入手加以定义,比如有学者认为网络暴力始于公民行使言论自由的异化,是网民以正义为名的非理性道德审判。[12]纵览网络暴力法律实务与学术研究的复杂流变,学界尚未整合出统一的网络暴力具体概念,为平台治理提供权威参照。

(二)网络暴力平台能动性治理缺失

当前,我国头部社交平台已按相关要求,普遍建立起包括前端服务、后端技术、秩序生成和维持以及未成年人保护机制为一体的网络暴力治理机制。诸如“一键防暴”等部分治理措施的技术应用和治理实效达到了相当水准。④虽然各平台治理举措各有创新之处,但当前各社交平台治理举措逐渐趋于同质,且普遍存在缺乏反网暴教育引导、网暴治理机制效果评估与反馈机制等问题。申言之,当下平台网络暴力治理仍以被动防御式治理为主,普遍缺乏能动性治理。

网络暴力平台能动性治理缺失可具体细分为四个方面。

一是缺乏反网络暴力实效测评机制,尚未产生平台与用户等独立第三方的治理合力。当前网络暴力治理实效多以各平台定期发布公告以及年度ESG报告阐述的形式公开,以罗列处理不规范账号与言论数量为主。尽管处理量巨大,但当前各平台网络暴力恶性事件仍然频发。且平台治理网络暴力被普遍认为是责任与义务,部分平台为降低成本,迎合用户,会放宽生态内容治理甚至主动迎合平台内“大V”等意见领袖所带来的“恶意流量”。这一区别对待进一步分化了普通用户与平台签约用户这两个群体,漠视了普通用户享有风清气正网络生态的治理诉求。

二是自动识别技术也存在有限性。网络暴力形态复杂,部分源于平台文化空间而内生的黑话、侮辱表情包等“梗文化”,且通常难以被算法监测以及非专业人士识别。

三是“前台匿名,后台实名”的网络实名制处于“虽实名但无实裁”的尴尬境地。尽管发言用户后台实名,但部分受网暴用户在求诸平台处理未果,转而寻求起诉这一自力救济手段之时,由于民事侵权诉讼需要明确诉讼相对人的个人信息,用户往往难以在“前台匿名”的网络交流中获得支持诉讼的身份信息。只得通过先诉平台,诉求平台提供施暴者个人信息,进而起诉施暴人员。

四是缺失网络暴力合规具象化指引。网络暴力作为语言暴力在社交平台的场域延伸与数据演化,理应受到数据合规制度的管控。尽管当前我国部分省市出台了属地企业数据合规参考指南,但均未提出建立网络暴力专项合规制度的具体行文要求,难以提供具象化的网络暴力治理引导。

(三)平台责任与合规责任关系错配

网络平台,又被称为平台、互联网平台,我国多部规范性文件及征求意见稿均肯定了平台的

“网络服务提供者”定位。比如,《关于平台经济领域的反垄断指南(征求意见稿)》《互联网平台落实主体责任(征求意见稿)》等文件均称互联网平台是“通过网络信息技术,是相互依赖的双边或者多边主体在特定载体提供的规则下交互,以此共同创造价值的商业组织形态”。《网络数据安全管理条例(征求意见稿)》附注说明互联网平台运营者是指“为用户提供信息发布、社交、交易、支付、视听等互联网平台服务的数据处理者”。规范文件并未对平台所应尽积极与消极义务的“平台责任”达成共识。

“平台责任”多散见于征求意见稿中,在未成年人保护等专项领域作宣示性规定。例如《未成年人网络保护条例(征求意见稿)》第6条指出,网络产品和服务提供者应履行未成年人网络保护义务,承担社会责任。平台的“主体责任”也频见于各规范性法律文件及征求意见稿中[13],平台合规责任被纳入主体责任中,由“平台合规依法经营”细化为“超大型平台经营者应当设置平台合规部门,不断完善平台内部合规制度和合规机制”“确保用户行为合法、合规、遵守社会公德”两项。但文件中规定的合规责任何以落实仍然缺乏具体的规范指引,且没有对网络暴力治理等专项领域设定具体的操作细则规定。

此外,“用户行为合规”的规定一方面缺乏对协助各平台培养管理并运营签约网红达人的MCN公司机构等第三方群体的合规考量,内容管理泛化,另一方面似乎也难合企业合规约束企业自身行为的理论根基,网络平台用户群体的合规必要性缺乏理论支撑。过往学者在讨论平台主体责任的多寡与限度之余,少有学者就平台责任与合规责任的关系展开论证。有学者认为,互联网企业合规体系与社会责任共同构建了网络主体责任体系。[14]但这一理论仍无法解释为何与“平台主体”不存在雇佣劳动关系的“用户客体”需遵守平台合规责任。百度⑤、腾讯⑥等社交平台企业将治理网络暴力写入企业社会责任成果中,这显然与合规责任之内容合规相重合,足见我国当前对平台责任与合规责任关系认识存在错误配置。

三、域外网络暴力平台治理的经验反思

针对引致我国平台网络暴力治理陷入困局的三重溯因,可借鉴域外网络暴力平台治理的分散型立法与奖励式驱动等经验,并对“通知—删除”型与“监管—删除”型平台治理责任分配加以反思,可为我国脱离网络暴力治理困局,构建有效社交平台反网络暴力合规机制提供域外参照。

(一)分散型立法:网络暴力外延的扩张储备

鉴于网络暴力的形态复杂性与时代变迁性,多数国家与平台都未在法律法规及平台公约文件中对网络暴力概念的内涵作出明确定义,多对其外延进行列举式释明。世界各国普遍选择了“分散型立法”的模式,即各国对网络暴力的外延规定经历了因网络暴力公众事件,而分散确立对儿童、妇女以及种族等特殊群体的“网络欺凌”“网络歧视”“网络仇恨”相关的法律法规,并随着网络暴力的样态嬗变与网络人权保护理念的不断深入,再次追加立法,将前述网络暴力治理的不良信息涵摄范围与保护群体扩张,最终整合为由国家公诉来保护全体民众免受网络暴力侵扰的立法进程。例如,美国因“13岁少女梅根网络欺凌自杀案”引发公众争议,案发州于2008年通过法案,追加“以电子形式骚扰未成年人构成刑事骚扰罪”。2009年美国国会通过《梅根·梅尔网络欺凌预防法》,将成年人纳入保护群体,进而扩展至保护全体民众不受网络欺凌。[15]韩国艺人崔真实因歧视谣言型网暴自杀,助推“崔真实法案”出台,该法案取消了“网络暴力需要当事人确认才处罚”的限制性规定,确立由国家公诉的“网络侮辱罪”。[16]在欧洲难民危机下,德国网络平台充斥对难民

的仇恨歧视言论。在社交平台整治乏力的基础上，德国于2015年起陆续颁发打击网络仇外等虚假非法信息的各项条例，并于2017年4月推动通过了《网络执行法》，将制造、传播及使用仇恨言论、煽动性言论等有害虚假信息的行为入罪。[17]

各平台对网络暴力信息的概念描述也多以罗列分散短语形式的外延为主，如推特(Twitter)用户协议将网络暴力信息概括为“具有攻击性、有害性、不准确或其他不当的内容”。⑦脸书(Facebook)社区规则将禁止发布的网络暴力信息分为“宣扬暴力和犯罪行为、有害安全、不良内容、违背诚信与真实性”等四个方面的内容。⑧Youtube平台将禁止发布的网络暴力信息分设为“仇恨言论、掠夺性行为、暴力写实内容、恶意攻击，以及宣扬有害或危险行为的内容”等五方面的内容。⑨且各平台对前述各个方面的内容存在更进一步的援引各国法律及判例的详细解释。从域外立法实践及平台立约实践可见，多采用在不同领域分散确立规制相关网络暴力的政策文件，并予以整合的立法立约路径。

(二)奖惩式驱动：域外网络暴力治理的模式对比

域外国家为催生社交平台对网络暴力治理动力，采取了进路不同的两类驱动模式。一类是以强制规定与处罚为主，带有管制主义色彩的“惩戒式驱动模式”，另一类是以协助参与和帮扶为主，引导平台的“激励式驱动模式”。较为典型的“惩戒式驱动模式”是韩国2002年至2012年间推行的“网络实名制”治理。韩国政府以行政强制命令的方式出台了实名制法案[18]，以日访问量作为平台设置实名制的标准，并规定未实行实名制的社交平台因网络暴力造成纠纷等后果的，平台除接受行政处罚外，需代位赔付受害人损失。[19]初期韩国社交平台的恶意评论数量有所下降，取得一定治理效果，但随之引发本土平台主动限流与访问量持续走低，部分网民弃用国内平台转用国外平台等一系列负面后果。且由于未配套有效个人信息保护政策机制，因黑客袭击致使近乎韩国全体网民个人信息数据泄露。2012年韩国宪法法院判决“网络实名制”系列法案违宪，以至世界首个“网络实名制”治理网暴实践破产。[20]“惩戒式驱动模式”虽短期内可能会推动本土平台快速构建起合规制度，但平台常常缺乏完善配套机制的主动性，后续合规治理效果堪忧。

当前部分国家逐渐转向“激励式驱动模式”，即通过辅助扶持政策，以指导激励平台建设网暴治理体系。如巴西、比利时等国政府设立网络犯罪投诉的官方网站和热线电话，积极为平台网络暴力受害者提供心理咨询与司法援助。[21]美国与英国政府通过制定“内容分级技术”标准，协助平台完善生态内容识别与过滤技术。[19]“激励式驱动模式”可较好地规避本土平台投入过多政策成本，保障平台与用户对网络暴力治理成效的合理预期。

(三)分配型责任：平台治理责任的立法限定

当前各国为治理网络暴力，保护网民权益，基本沿用有关部门协同平台治理的治理路径，但对社交平台分配了不同的治理责任。当前主要存在两类分配进路，一类是“通知—删除”型平台治理责任分配，即平台仅承担一定生态内容管理义务，不法信息经有关部门或权利人通知后删除可豁免平台侵权责任等法律责任；另一类是“监管—删除”型治理责任分配，即以超大型平台为代表的部分平台负有管理生态内容，防治网暴信息的主体责任，需主动识别、发现并删除网暴信息。如果平台未履行管理义务，致使网暴行为造成严重后果，要承担平台关停整改等相应法律责任。

“通知—删除”型平台治理责任分配承袭平台“避风港原则”。典型范例为美国1996年《通信规范法》第230条规定平台无需对第三方用户内容负民事侵权责任与行政责任。[22]随着平台垄断地位的生成，为保障普通民众权利，2020年美国司法部提案公民可就平台内部第三方言论内容起诉平台，进而限缩《通信规范法》第230条有关豁免权的相关规定，但这一提案尚未获得通过。而德国

2017年出台的《网络执行法》强化了部分平台对网络暴力的治理责任。该法明确特大型社交平台应设立一套有效透明的投诉程序,从而迅速获悉、处理用户对平台空间内违法内容的投诉。明显违法的内容应当在24小时内予以删除或屏蔽,其他违法内容应当在7日内予以删除或屏蔽。如果平台对明显违法的内容在当事人通知后无处理,将受有关部门处罚。[23]

“监管—删除”型治理责任分配承袭平台“守门人责任”。“守门人责任”源于欧盟《数字市场法》,该法提出了超大型平台的“看门人义务”[24],明确指出企业负有建设个人信息与隐私保护的合规义务,并在欧盟《数字服务法》中得到了进一步细化[25]。澳大利亚《2021年在线安全法》亦专章指出平台企业在提供在线服务时,要建立针对未成年人的“网络欺凌”的专项合规机制,实现监测到欺凌信息即删除。[26]在“监管—删除”型治理责任分配下,推特(Twitter)会定时公示全平台的处理违法信息结果。采取平台“注册实名制”的脸书(Facebook)会采取算法实时清理监测到的假名用户。“守门人责任” 兼具平台需采取积极治理措施的义务要求与平台在采取治理措施后免除责任的激励,本质上与合规管理及其附随的合规激励相通。申言之,“通知—删除”型平台治理责任一定程度上有助于激发平台企业的机制创新与内容创造活力,减少对网络生态过多干预,但有关部门很难穷尽对网络生态内容的全数据监督,难以最大程度根绝网络暴力;而“监管—删除”型治理责任虽给予了企业较大的合规管理自治权,但也增加了企业内容合规成本。在有关部门的追责惩戒下,平台企业可能会走向盲目删帖,设置过宽监管关键词等“假合规”的另一极端,需要有关部门给予政策激励加以补足。

四、激活网络暴力治理效能的应然逻辑

以合规义务、合规风险与合规责任等合规理论来重塑网络暴力平台治理理论,既是构建社交平台反网络暴力合规专项机制的应有之义,也是激活网络暴力治理效能的应然逻辑。

(一)合规义务视角的网络暴力概念延展

网络暴力实质是带有暴力属性的语言在网络空间的延伸与演化,其“多对一”“多对多”的裂变式规模性传播促成危害后果聚量性叠加。[27]此前我国立法中网络暴力概念仍存在模糊性,尚未生成平台治理可直接适用的网络暴力指导性概念。且学界的网络暴力概念中,“网络媒介说”“起因说”过于抽象,“特定行为说”往往滞后于实践,难以指导平台对网络暴力的精细化治理。可以从合规义务视角出发,对平台治理公约等规范性文件中的网络暴力概念外延予以实时延展。

“合规义务”概念源于国际标准化组织(ISO)所订立的国际合规标准。《ISO 37301:2021 合规管理体系 要求及使用指南》规定:“合规义务是组织必须遵守的要求, 以及组织自愿选择遵守的要求。”换言之,合规义务作为合规所遵守的规范统称,本质是对既往国家法律法规、典型判例,平台公约规则以及平台订立合同约定的全面汇总。一方面,对网络暴力的分散式具体立法立约显然有助于合规义务的精细化识别与遵循。另一方面,部分互联网企业在技术赋能下,已经可以通过对海量判决文书中有关网络暴力适用的结构化要素进行抽取与分析,研判确立企业合规治理中可适用的较新合规义务。[28]平台在治理网络暴力的合规实务中可以对识别到的网络暴力行为进一步细化,在用户言论表达自由、身心健康等核心法益不受侵犯的基准上,向对部分公众人物留有舆论监督渠道等边缘法益做同心圆延展,不断确立合规义务,进而反哺国家立法。

(二)合规风险视角的网络暴力类型分解

当前我国平台为“监管—删除”型治理责任分配。如果平台未有效治理网络暴力事件,可能会

带来合规风险。合规风险在诸多文件与标准中都有成文定义。《中央企业合规管理办法》指出，合规风险是指企业及其员工在经营管理过程中因违规行为引发法律责任，造成经济或者声誉损失以及其他负面影响的可能性。合规管理的目的便是防控合规风险。《ISO 37301:2021 合规管理体系 要求及使用指南》附录A.4.6合规风险章指出，不合规的后果可能包括个人和环境伤害、经济损失、名誉损失、行政管理变更以及民事和刑事责任。以合规风险来源的角度，对网络暴力类型作出再分解，乃建立反网络暴力专项合规机制的必要前提。

从合规风险视角出发，可将网络暴力类型作平台外部性网络暴力与平台内源性网络暴力的界分。平台外部性网络暴力是现实外部语言暴力的在线表达，通常是网民用户群体对平台外部热点事件当事人的无组织自发舆论冲突与道德审判。平台内源性网络暴力是指基于流量、排名等利益冲突或平台交流争端等非利益冲突，在平台授意下，或由平台意见领袖、MCN机构等平台签约相关方组织，在平台内部弹幕、话题、贴吧、直播间等开放群组中引导侮辱谩骂、造谣诽谤、侵犯隐私以及散播不良信息等网络暴力行为，如平台用户将当事人照片、行为或言论投稿至社交平台“厕所号”，账号所有者再将投稿以匿名形式发出，供网友对其发泄情绪，诱导公众人物粉丝骂战的“网络厕所号”投稿行为；[29]部分网友在网络意见领袖组织下，在贴吧内不停地发无实质内容的辱骂贴等信息，扰乱贴吧秩序的“贴吧爆吧”行为；[30]平台用户帮助客户在指定社交平台上对人进行网络言语辱骂的“代骂服务”行为[31]等。平台通常难以对外部舆论事件等外部性网络暴力源头加以规制，只能在“监管—删除”的合规义务要求下，尽到接收举报、检查识别并删除普通用户侵犯他人权利的暴力言论或视频的合规职责。而平台对自身及其所签约的MCN机构、网络意见领袖等内源性网络暴力负有合规职责以及履行签约合同中内容管理的合规义务。《关于依法惩治网络暴力违法犯罪的指导意见》对平台内源性网络暴力也有特别关注，其第8条第2款和第5款指出组织水军、打手或者其他人员实施的，或者由网络服务提供者发起、组织的网络暴力，司法机关应依法从重处罚。

(三)合规责任视角的平台责任厘清

《网络暴力信息治理规定(征求意见稿)》指出，社交平台这类网络服务提供者应对网络暴力治理负主体责任。有学者认为平台主体责任可分为法律责任、契约责任和道德责任。[13]法律责任包括法律规定的义务和违反义务的不利后果，契约责任源于平等主体之间达成的自愿合意，道德责任则属于道义上的要求。三项责任与合规义务来源的国家法律法规、典型判例相通，平台订立合同约定与平台基于道德公益而制定的公约规则相通。且主体责任本意便是发挥主体行使责任的主观能动性，这亦与企业构建有效合规的主动性要求相符。申言之，平台主体责任作为平台责任的同义表述，平台责任与合规责任本质相同。有学者认为应将《网络暴力信息治理规定(征求意见稿)》中“信息内容管理主体责任”修改为“信息内容管理责任”。理由是平台在“数字守门人”与宪法要求的秩序维护义务下，虽获得了网暴治理的法理权限，却并未厘清其作为网络服务提供者与网络监管部门的网暴治理的权限分配。[32] 笔者认为该政策建议缺乏对平台主体责任相关规定的体系性考量。平台主体责任指平台为做好分内之事所应主动承担的积极义务和不作为的义务，其“信息管理主体责任”在《网络信息内容生态治理规定》等多部规范性文件中均有明确规定。在网络暴力治理中减损平台的主体责任定位，既与平台内其他信息内容管理地位相偏离，也不利于平台合规责任的进一步落实。

合规被视为规范企业法人及企业员工、相关方等业务行为的约束工具，享受企业服务并支付相关对价的网络用户通常不被视为合规约束对象的一员。传统合规理论难以充分论证平台主体责任中“用户行为合规”的正当性。笔者认为可从平台商业模式的特殊性与我国实务中所呈现的“监

管—删除”型治理责任分配两方面对“用户行为合规”加以证成。我国社交平台多采“流量变现”的商业模式,平台签约用户通过在平台发言、上传视频稿件等互动行为,可获取流量、等级、经验及称号等具有特殊可变现价值的回报物。部分用户通过与平台及MCN机构签约,成为平台“流量变现”商业行为的相关方。在“内容创作—提现”行为中,平台应对平台签约用户的内容创作行为进行合规管理。而对互动行为无直接价值回报的普通用户而言,我国“监管—删除”型治理责任分配强制要求企业尽内容合规监管职责,平台在维护平台交流秩序等社会公共利益的过程中,必然要以合规机制实现对网络暴力内容创作者的合理规训。

五、构建平台反网络暴力专项合规机制的措施

在域外“激励式驱动”治理与平台治理责任分配参照下,有效的平台反网络暴力专项合规机制需要网信部门、司法机关等有关部门与社交平台企业协同参与构建,以最大限度地保障平台用户免受网络暴力侵扰。

(一)完善合规激励机制

网信部门与司法机关可参照“激励式驱动”的域外借镜,以合规激励机制参与到社交平台对网络暴力的协同治理中。有关部门在建设社交平台企业反网络暴力合规激励机制时,可以从风险告知型激励、经济利益型激励和社会声誉型激励三个层面着手。[33]

在风险告知型激励层面,网信部门与司法机关可以随着网络暴力种类与形式的演变,出台更为详细的网络暴力单项法规、平台合规指引,对网络暴力新形态识别、用户申诉救济等细化领域制定更为完善的合规标准,并通过提供样板合规体系文件库、合规师援助以及座谈研讨会等多形式的合规指导,以足量明确的合规义务供给理清平台治理网络暴力的合规需求。

在经济利益型激励层面, 网信部门与司法机关可以对建立有效反网络暴力合规机制的互联网平台企业给予减免税收、提供合规补贴等经济奖励。对因网络暴力管理不力而在网络暴力专项整治中受到行政刑事处罚的平台,还可根据其合规建设完备程度给予相应的处罚减免,尝试借鉴《行政处罚法》第33条第1款的“首违不罚”制,探索给予社交平台一定的合规容错空间,对平台首次违反网络暴力合规义务,行为轻微并及时改正且没有造成危害后果的不予处罚。

在社会声誉型激励层面,网信部门与司法机关可以与社会各界合作,建立评价体系,评估社交平台企业在反网络暴力方面的表现,并将评估结果向公众公开,以提高社交平台的社会声誉,增加其品牌价值和用户信任度,推动形成良好的舆论环境。

(二)构建PDCA合规管理机制

社交平台可参考“通知—删除”“监管—删除”的平台合规治理责任,基于PDCA循环模式构建合规管理机制, 并依据网络暴力新形态不断完善。PDCA循环管理模式由美国质量管理专家戴明在1950年应用于企业管理中,此后得到广泛宣传与应用。PDCA循环的工作顺序依次为P(plan)—D(do)—C(check)—A(act),即计划—实施—测评—改进。[34]PDCA循环在《ISO 37301:2021 合规管理体系 要求及使用指南》等国际合规标准中均有提及,我国合规标准也对其进行了承袭。平台反网络暴力专项合规机制可在合规计划、合规实施、合规测评与合规改进四个方面进行构建。

首先,在合规计划方面,社交平台可通过技数赋能等多种形式,实时引入合规义务,建立重点针对网络暴力治理的专项合规计划,与平台企业规模相称地合比例开展合规。各社交平台应积极加强与行业自律组织和政府监管机构的合作,共同制定和执行相关规定和标准,推动用户隐私保

护和言论自由的平衡发展,注重平台公约、审核机制、应急机制与“一键防护”反网络暴力工具等基础机制的设置。[35]社交平台在海外也应遵守当地对网络生态内容治理要求的相关法律法规。社交平台还需要加强对个人信息保护等其他专项数据合规机制的建设规划，以实现对反网络暴力合规机制的有力补足。

其次,在合规实施方面,社交平台应当通过构建教育引导机制、技术规制机制、诉讼支持机制等一系列合规机制,有效执行合规计划。在教育引导机制方面,应完善通俗易懂的用户守则,引导网络用户自觉遵守平台公约,以减少平台外部性网络暴力的产生;对平台签约用户、MCN机构等关联方,社交平台应开展专门的合规文化培训,以实现最大程度阻绝平台内源性网络暴力。在技术规制机制层面,平台应通过合规技术实现对合规风险的识别、监控与处理。反网络暴力合规机制应通过舆情监测系统,重点关注负面热点事件发言。对部分违规用户采取时间限定性的全网同名同设备序列号的登录禁止令。在诉讼支持机制方面,平台应通过多种形式建立与司法机关的诉讼协作机制。针对集聚性网络暴力“法难责众”的实务难题,平台可探索与有关部门协作,向网络暴力参与人员私发公权机关认证的“网络暴力告诫书”,以增强威慑性。针对网络暴力受害人司法救济难的诉讼困境，在网络暴力自诉转公诉制度尚未丰满之前，社交平台应发挥数据云存储优势,超大型平台应采取区块链等高新技术对平台言论内容存证,以方便用户存证且证据可鉴真。中小型平台虽技术有限,但也应与公证机关[36]或第三方区块链存证平台合作,提供存证取证方便公证机制。对想通过起诉等私力救济方式解决纠纷,需要施暴人个人信息的受害人用户,平台可探索建立专门的诉讼支持机制,在审核用户的诉讼请求后,及时向用户私密告知立案所需的相关信息。

再次,在合规测评方面,社交平台可通过听取用户的申诉反馈以及对既往网络暴力合规治理效能进行定期的内部审核和外部评估,以确保合规计划的有效执行。平台应完善平台用户对网络暴力合规处置的申诉反馈机制,为用户提供包括详细内容与步骤的申诉指南,通过设立专门处理申诉内容的团队或部门以保证对申诉的快速响应和处理，并及时向用户透明地通知包含是否成功申诉以及采取的相应措施的申诉结果。社交平台企业还可在ESG报告、平台社区等处定期披露网络暴力的细化治理数据,以供平台及第三方人员监测和分析网络暴力的演进趋势和内生模式,以识别潜在的合规风险和问题,并制定相应的合规措施。

最后,在合规改进方面,根据合规测评阶段的研究成果,社交平台可采取相应的合规改进措施。平台可修订合规策略和章程规范,以适应不断变化的网络暴力形式和法规要求。平台还应加强技术投入,进一步细化技术规制机制,及时扩充“梗文化”等领域语料储备,加强自动识别技术对亚文化领域的学习。社交平台可借助第三方专业团队组织,对其所使用的技术识别、人工审核过严过厉等不当合规机制进行纠偏,以实现合规机制的自我优化。

六、结 语

李强总理在2023年7月12日国家发展改革委平台企业座谈会指出:平台企业要积极履行社会责任,加强行业自律、合规经营。⑩反网络暴力合规作为具有独特历史演变与多学科技术交叉的合规研究焦点,以梳理当前社交平台治理网络暴力的合规困局出发,以平台网络暴力合规义务的外延拓清，平台网络暴力合规风险的创新分类以及平台网络暴力合规责任的本源确定作为理论探讨的基点,对专项数据合规机制的研判参照有一定借鉴作用。笔者通过对本土与域外社交平台反

网络暴力合规的治理样态作考察分析，探索形成本土化的社交平台有效反网络暴力专项合规机制。这既是回应我国平台企业合规制度与国际接轨的法治探索，也是未来中国网络生态清朗蓬勃发展，实现中国式网络现代化的必由之路。

注释：

①参见中国互联网络信息中心《第52次中国互联网络发展状况统计报告》(https://www.cnnic.net.cn/n4/2023/0828/c199-10830.html)。

②参见企鹅有调《2022年全国网民网络安全感满意度调查专题报告 网络暴力防控与网络文明专题》(https://research.tencent.com/report?id=Md4)。

③参见赵岩、高雅《加强网络人格权益保护 推进网络空间治理法治化 北京互联网法院发布涉网络暴力典型案例》(https://www.chinacourt.org/article/detail/2023/08/id/7455353.shtml)。

④参见中国社会科学院大学互联网法治研究中心《互联网平台网络暴力治理机制构建与测评报告》(https://law.ucass.edu.cn/info/1052/4721.htm)。

⑤参见新华网《百度发布〈2022年环境、社会及管治(ESG)报告〉》(https://www.news.cn/info/20230606/06d4918ce86647a6a436644e947af822/c.html)。

⑥腾讯网《腾讯2022年环境、社会及管治报告》(https://www.tencent.com/index.php/zh-cn/esg/esg-reports.html)。

⑦参见《Twitter User Agreement》(https://cdn.cms-twdigitalassets.com/content/dam/legal-twitter/site-assets/2023-05-18/en/twitter-user-agreement-23-05-18.pdf)。

⑧参见《Facebook社群守则》(https://transparency.fb.com/zh-cn/policies/community-standards)。

⑨参见《Youtube的社区准则》(https://support.google.com/youtube/answer/9288567?hl=zh-Hans)。

⑩参见新华社《李强主持召开平台企业座谈会》(https://www.gov.cn/yaowen/liebiao/202307/content_6891546.htm)。

[参考文献]

[1]辛艳艳,方师师.网络暴力治理中的平台主体责任困境[J].青年记者,2023,(13).

[2]王春霞.破解惩治网暴难点 形成协同治理合力[N].中国妇女报,2023-06-21(05).

[3]谢登科.网络暴力犯罪的公私协同治理模式[J].法律科学(西北政法大学学报),2023,(5).

[4]高媛.自媒体网络暴力行为的法律规制[J].太原理工大学学报(社会科学版),2021,(2).

[5]程啸.大型网络平台违反守门人义务的民事责任[J].法律科学(西北政法大学学报),2023,(5).

[6]赵韫溟,刘晶."规则+技术":互联网平台针对"网络暴力"现象的数字化治理[J].北京文化创意,2023,(2).

[7]王静.数字公民伦理:网络暴力治理的新路径[J].华东政法大学学报,2022,(4).

[8]李本灿.企业视角下的合规计划建构方法[J].法学杂志,2020,(7).

[9]陈瑞华.论企业合规的性质[J].浙江工商大学学报,2021,(1).

[10]储陈城.刑法应对网络暴力的流变及其基本立场[J].中国刑事法杂志,2023,(4).

[11]彭兰.如何认识网络舆论中的暴力现象[N].中国社会科学报,2009-08-25(06).

[12]林爱珺.网络暴力狂欢的反思与规制[J].人民论坛,2022,(9).
[13]刘权.论互联网平台的主体责任[J].华东政法大学学报,2022,(5).
[14]田佳奇.2022年(第九届)中国互联网企业社会责任论坛举办[J].中国国情国力,2023,(1).
[15]李修平.梅根之死[J].现代世界警察,2023,(8).
[16]桑瑞娇,周知蓓,王艳阳.韩国反网络暴力的立法与实践[J].现代世界警察,2023,(7).
[17]冯雪珺.德国重拳打击网络非法言论[N].人民日报,2018-01-04(21).
[18]周永坤.网络实名制立法评析[J].暨南学报(哲学社会科学版),2013,(2).
[19]尹建国.我国网络信息的政府治理机制研究[J].中国法学,2015,(1).
[20]董俊祺.韩国网络实名制治理及启示[J].中国人民公安大学学报(社会科学版),2015,(6).
[21]宋亦然,沈小晓,牛瑞飞.多国加大对网络暴力的治理力度[N].人民日报,2023-07-18(17).
[22]The 104th United States Congress.*Communications Decency Act*. 47 U.S.C.§230,1996-02-08.
[23]Deutscher Bundestag. *Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken*.§ 3,Umgang mit Beschwerden über rechtswidrige Inhalte,2017-06-17.
[24]European Parliament and Council of the European Union.*Digital Services Act*. Art. 32, Compliance officers,2022-10-19.
[25]European Parliament and Council of the European Union.*Digital Markets Act. Art.* 7, Compliance with obligations for gatekeepers,2022-07-18.
[26]The Parliament of Australia.*Online Safety Act 2021. Part 5—Cyber-bullying material targeted at an Australian child*,2021-07-23.
[27]于冲.网络"聚量性"侮辱诽谤行为的刑法评价[J].中国法律评论,2023,(3).
[28]谢澍.互联网企业刑事合规义务识别:分层、复合与技数赋能[J].云南社会科学,2023,(3).
[29]孙天骄,陈立儿.臭气熏天的"网络厕所"该怎么治[N].法治日报,2023-08-14(08).
[30]蔡琰.网络群体行为的转化研究[D].上海:上海社会科学院,2017.
[31]赵丽,李丹阳."代骂服务"在多平台明码标价售卖[N].法治日报,2023-08-26(04).
[32]敬力嘉.网络服务提供者网暴治理义务的体系展开[J].北方法学,2023,(5).
[33]郑雅方.论政府介入企业合规管理的风险及其防范[J].法商研究,2021,(3).
[34]刘艳红.网络暴力治理法治化研究[M].北京:法律出版社,2023.
[35]杨洁.基于PDCA循环的内部控制有效性综合评价[J].会计研究,2011,(4).
[36]万力,张洪浩.后疫情时代网络暴力侵权网页公证及其合规路径[J].网络空间安全,2022,(2).

【责任编辑:叶　萍】