用户想要平台内容审核，还是个人控制？考察第三人效果和言论自由支持在塑造内容审核偏好中的作用

Shagun Jhaver，Amy Zhang

本研究考察社交媒体用户在使用全平台内容审核与用户控制的个性化内容审核工具之间的偏好，以规范三类违反规范的内容——仇恨言论、露骨性内容和暴力内容。我们通过一项针对美国984名成年人的全国代表性调查，探讨第三人效果和对言论自由的支持如何影响这一选择。研究发现，对于每一种言论类别，认为内容会对他人造成负面影响的程度越高，用户越不倾向于选择个人审核设置而非平台主导的内容审核；而对言论自由的支持越高，用户越倾向于选择前者。研究结果表明，平台治理举措需要兼顾违反规范的言论类别对媒体产生的实际影响和感知影响，才能提升用户满意度。我们的分析显示，用户并不认为个人内容审核工具会侵犯他人的言论自由，而是将其视为一种增强对自身社交媒体信息流掌控力的方式。

关键词

社交媒体；内容审核；治理；审查；平台

引言

随着社交媒体网站的兴起及其在人际传播中的广泛应用，Facebook、Twitter 和 YouTube 等公司已成为数字表达的新治理者。与此同时，使用这些网站的个人也可以通过多种方式参与治理。例如，他们可以举报违反社区政策的帖子、对不当帖子投反对票、担任志愿版主、参与反制言论，或配置内容审核设置以自动移除某些帖子。因此，我们正朝着“言论规制多元主义模式（Balkin，2017年）”迈进，在这一模式中，言论必须以多方参与的方式受到监管——立法机构执行网络言论法律，平台运营者制定可接受内容的治理制度，而用户自身则会针对被认为有问题的内容采取干预。

这种向多元主义模式的转变，发生在近期围绕平台内容审核决策的争议（Stecklow, 2018年；Gillespie, 2010年；Massanari, 2017年）以及媒体、政策制定者和公众日益呼吁加强内容监管的背景下（Williams 等人，2021）。作为回应，各平台开始投入更多资源，以改进对不当帖文的检测和删除方式。本文聚焦于平台提供的个人内容审核工具，这类工具允许终端用户配置所见帖文的内容审核方式，使之符合自己的内容偏好。我们主要关注 Instagram 和 Twitter 等平台提供的工具，这些工具允许用户指定自己对特定主题类别的敏感程度，例如露骨性内容和仇恨言论（见图 1）。我们注意到，需要区分谁将某项内容归入某个主题类别，以及谁决定是否删除该类别中的内容。

前者是一个关键且颇具争议的问题，而我们在此关注的是后者

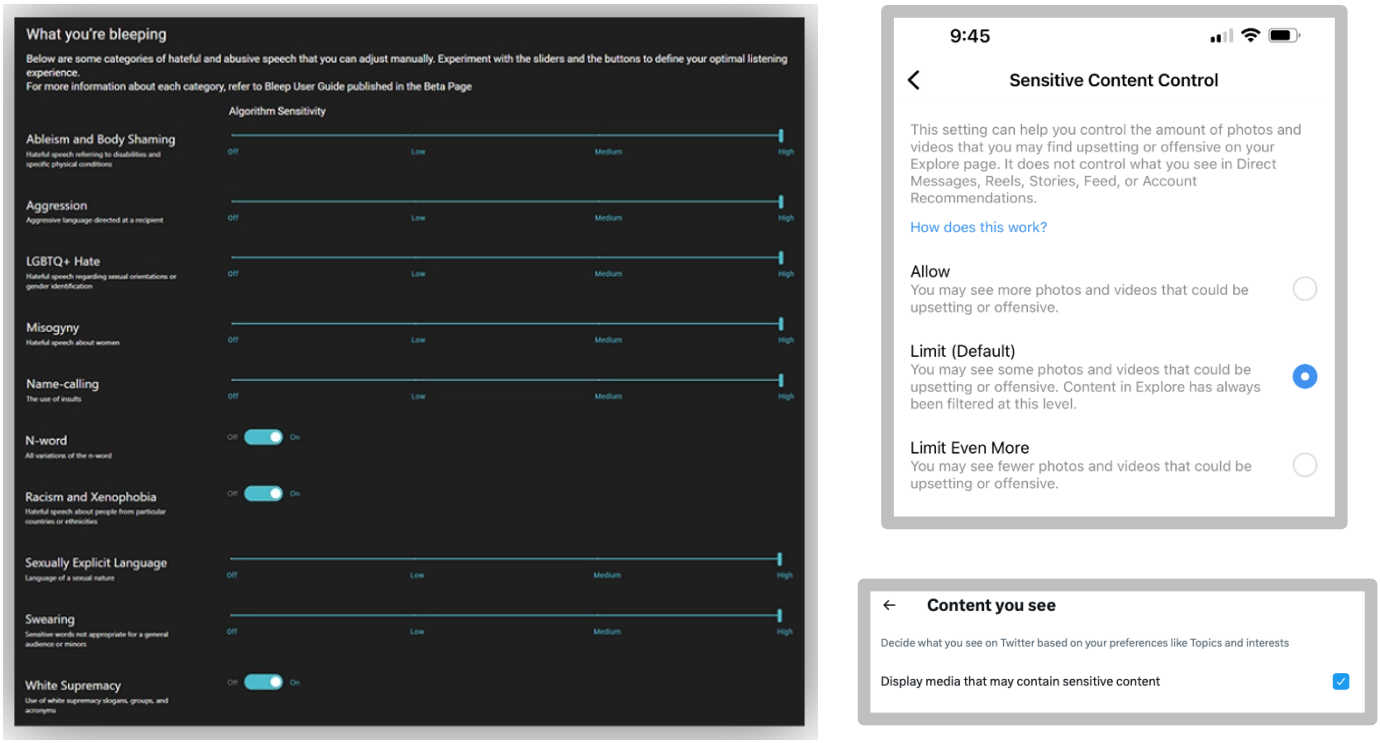

图*1*：*Intel Bleep*（左）、*Instagram*（右上）和 *Twitter*（右下）上的个人内容审核工具示例。

顾虑。
配置此类工具，能让用户根据自己的喜好和接受阈值调整内容审核系统。从平台的角度来看，提供个人内容审核工具的一项策略性后果是，将作出艰难内容审核决策的责任、承担相关错误的负担，以及进行“正确”配置所需的认知劳动，交由用户承担。因此，研究者必须了解用户如何看待平台内容审核与个人内容审核之间的选择。

我们的研究回应了治理学者的呼吁，即开展更多基于调查的研究，以了解用户对不同监管行为者采取的内容审核干预措施的看法（Einwiller 和 Kim，2020年；Riedl 等人，2021a）。迄今为止，我们对终端用户如何看待通过个人内容审核工具获得自我监管权知之甚少。我们不了解用户在哪些情况下更希望拥有参与塑造内容审核方式的选择权，哪些情况下则更希望平台为所有用户统一管理内容审核，也不了解塑造这些偏好的各种因素。

在第三人效果（TPE）假说的启发下，我们通过考察用户对文献中此前研究过的三类违反规范的言论的偏好，填补了这一空白：（1）仇恨言论、（2）露骨性内容，以及（3）暴力内容。此前研究表明，对媒体信息影响他人的感知能够预测

内容监管态度（Riedl 等人，2021b；Rojas 等人，1996）。我们考察第三人效果（TPE）（Davison，1983）在塑造用户对采用平台实施的内容审核工具与个人内容审核工具的态度方面所发挥的作用。

我们还将研究发现与探讨公众对言论自由及其影响的态度的学术研究联系起来。正如第一修正案所体现的，许多美国人高度重视言论自由。然而，引入个人内容审核工具，使维护言论自由原则变得更加复杂。用户可能会利用这些工具避开特定类别的内容，而其他人仍能看到相同内容，从而避免侵犯网络表达。另一方面，个人内容审核工具也可能被视为人们逃避自己不喜欢的观点的一种方式。虽然人们随时都可以忽略任何内容，但个人内容审核工具让大范围删除内容变得轻而易举。除了不使用个人内容审核工具之外，言论自由倡导者也可能希望他人没有使用这些工具的选择。因此，我们分析用户对言论自由的支持如何影响他们对不同审核方式的看法。

了解公众对平台和用户实施的干预措施的看法，有助于推动关于各种监管行为者的作用和策略的讨论（Seering 等人，2020年；Seering 等人，2019）。这也能回应循证政策制定的呼声（Livingstone，2013；Puppis 和 Van den Bulck，2019），通过厘清公众如何理解审核实践，并找出政策与公众需求之间的差距（Naab 等人，2021；Riedl 等人，2021a；Paek 等人，2008年）。鉴于平台迅速推出新的审核策略，尤其是个人内容审核工具，独立学术评估用户对部署这些工具的态度至关重要。此类研究也能帮助平台了解终端用户对各种监管实践的接受或抵触程度，以及影响这些看法的因素，从而开发更好的工具。此外，在网络活动的背景下考察公众对言论自由的看法，也可能影响长期保护言论自由的努力。

文献综述与假设

第三人效果的感知成分

1983年，Davison 首先提出，个体认为媒体对他人的态度与行为所产生的影响大于对自身的影响（Davison，1983年）。此后，大量研究表明，这种差异在一系列情境中都保持一致（Sun 等，2008年），例如政治广告（Golan 等，2008年；Paek 等，2005年）、新闻报道（Price 和 Tewksbury，1996年）以及社交媒体使用（Schweisberger 等，2014年）。Davison 将这一假设称为第三人效果（*TPE*）（Davison，1983年）。这一假设如今已成为解释公众对媒体审查看法的广泛应用视角（Gunther，2006年；Guo 和 Johnson，2020年；Hoffner 等，1999年；McLeod 等，1997年）。第三人称效应假设包含两个主要成分——感知成分和行为成分（Gunther，1991年）。感知成分指出，推定的媒体对他人影响（*PME3*）往往大于

对自身的感知媒体影响（*PME1*）。在社交媒体信息影响的情境下，感知成分预测，个体会认为，与自身相比，他人更容易受到社交媒体上各类违反规范的言论的负面影响。因此，我们提出以下假设：

假设*1:* 对于每一类违反规范的言论，参与者都会认为这类言论对他人的影响大于对自己的影响。

第三人效果的行为成分

第三人效果的行为成分认为，当个体感知到媒体信息对他人的影响大于对自身的影响时，他们会采取补救措施，以减轻其所感知到的有害影响（Guo 和 Johnson，2020年）。Davison 将审查这一现象描述为第三人感知最有趣的行为后果之一（Davison，1983年）。关于第三人效果后果的既有研究表明，第三人效果会促使人们支持审查（Rojas 等人，1996年；Gunther，2006年；Golan 和 Banning，2008年），而当说服尝试可能产生社会所不认可的影响时，这种效果尤为明显（Lim，2017年）。在研究第三人效果对审查态度的影响时，一些研究者考察了媒体影响感知中他人与自身的差异所带来的后果（媒体影响差异 = PME3 – PME1）。另一些研究者则关注人们感知到的媒体对他人的影响（PME3）（Riedl 等人，2021b）。在考察有关第三人效果后果的既有研究数据后，Chung 和 Moon 得出结论：媒体被认为对他人产生的影响（PME3），比媒体影响感知中的他人与自身差异更能预测审查态度（Chung 和 Moon，2016年）。本研究选择 PME3 作为主要预测变量，并提出以下假设：

H2：对于每一种言论类别，他人感知到的该类言论影响越大，对平台禁止该类言论的支持度就越高。

第三人效果（TPE）研究者长期以来一直关注，他人感知到的媒体影响可能引发哪些行为。除对审查的支持外，先前研究还考察了用户的行为结果，例如参与政治行动（Feldman 等人，2017年；Rojas，2010年；Wei 等人，2011年）、传播反对性信息（Bernhard 和 Dohle，2015年；Golan 和 Lim，2016年），以及揭露明显偏见（Barnidge 和 Rojas，2014年；Lim 和 Golan，2011年）。在在线内容审核的背景下，Jang 和 Kim 发现，第三人感知程度较高的平台用户更有可能支持通过媒体素养干预来应对假新闻（Jang 和 Kim，2018年）。我们通过考察第三人效果对用户支持个人审核设置、以审核违反规范内容的影响，进一步拓展了此前关于揭示第三人效果所导致的不同类型行为后果的研究。由于此类设置同样属于一种监管形式，即对内容的终端用户监管，先前的第三人效果研究表明，PME3 应能预测人们对这类设置的支持（Gunther，2006年；Riedl 等人，2021b）。因此，本研究提出以下假设：

H3：对于每一种言论类别，该类言论对他人的感知影响与支持使用个人内容审核工具来管理该类别言论之间呈正相关。

平台实施的内容审核与用于自行审核内容的个人设置，是管理违反规范言论的不同方式。然而，平台实施的内容审核会在整个平台范围内审查内容，而个人内容审核工具则允许用户自行调整是否愿意接触违反规范的言论，以及愿意接触多少。既有文献普遍预测，第三人效果会促使人们支持审查态度（Gunther，2006年；Chung 和 Moon，2016年）。然而，这些文献并未说明，当人们面临这样的选择时会作何反应：是让平台处理某一特定内容类别，还是允许用户自行设定该类别的内容审核偏好。本研究提出以下研究问题：

研究问题*1*：对于每种言论类别，人们对其影响他人的感知与以下两种支持度有何关联：支持平台禁止该类言论，还是支持使用个人内容审核工具对其进行管理?

对言论自由的支持

关于内容审核措施益处的讨论总是与言论自由问题交织在一起。在美国，宪法将言论自由作为一项基本人权加以保护。然而，平台属于私人主体。1996年《通信规范法》第230条赋予平台立法上的自由，使其能够按自己的意愿管理用户，同时不必为错误或疏漏承担责任（Gillespie，2017年；Medeiros，2017年）。专家指出，尽管美国人普遍支持言论自由，但他们对仇恨言论的容忍度很低（Gibson 和 Bingham，1982年；Sullivan 等，1993年；Yalof 和 Dautrich，2002年）。因此，人们在抽象层面上接受言论自由，并不一定意味着他们会容忍与之相悖的表达（Naab，2012）。考察人们对言论自由的支持在多大程度上影响他们对不同内容审核策略的看法，或许有助于更好地解释这一现象在社交媒体平台语境中的表现。

此前的研究发现，对言论自由的支持与对内容审核的态度有关。例如，Naab 等人指出，认同言论自由的人不太可能支持 Facebook 版主采取限制性行动（Naab 等人，2021）。Guo 和 Johnson 指出，缺乏对言论自由的支持可以预测人们对政府监管性别歧视仇恨言论的支持（Guo 和 Johnson，2020）。然而，他们并未发现前者能够预测人们对平台审查的支持性态度。Jang 和 Kim 指出，尽管存在第三人效果，对言论自由的支持仍会降低人们对假新闻进行监管的支持（Jang 和 Kim，2018）。总的来说，这些研究表明，参与者对平台监管的支持会随着其言论自由支持的增加而下降。本研究提出以下假设：

H4：对于每种言论类别， 参与者对言论自由的支持程度与其对平台封禁该类别言论的支持程度呈负相关。

关于个人内容审核工具的文献十分有限，无法直接说明对言论自由的支持与支持使用此类工具之间的关系——不过，已有一些研究涉及相关问题。例如，Naab 等人发现，用户对言论自由的坚持程度与其采取纠正性行动的意愿之间不存在关联

，例如斥责评论作者或举报评论（Naab 等人，2021）。鉴于部署个人内容审核工具在此意义上并不属于纠正性行动，因此尚不清楚这一发现是否适用于我们的研究情境。此类工具的预期成本也低于纠正性行动；例如，用户可以只设置一次个人审核配置，而不必举报每一条不当帖子；个人内容审核会私下移除问题评论，避免与评论作者发生正面冲突，而发表反制言论则可能招致报复。Riedl 等人的一项调查显示，个人对言论自由的支持并不会增强其干预问题评论的义务感（Riedl 等人，2021a）；然而，作者并未告知调查对象他们可以采取哪些干预方式。此外，在我们的研究情境中，用户可能不会将个人内容审核工具视为一种义务。

一方面，预计对言论自由价值的支持会降低人们对大多数（甚至所有）限制性行动的支持。另一方面，人们可能会认为，个人内容审核工具只是让他们能更自主地决定自己看到什么，而不是侵犯他人的言论自由。由此，个人内容审核工具使人们能够区分自己的发言自由与被他人听见的义务。为明确这种关系的方向，本研究提出以下问题：

研究问题二： 参与者对言论自由的支持是否与他们对拥有个人内容审核工具的支持相关?

当人们需要在让平台审核某一特定内容类别与让用户自行设定内容审核偏好之间做出选择时，我们预计，后者会被认为更能维护言论自由。它赋予用户而非平台决定是否删除内容以及删除多少内容的自由。因此，我们提出以下假设：

假设*5*：对于每一种言论类别， 参与者对言论自由的支持程度越高，他们对使用个人内容审核工具审核该类别的支持程度，就越可能高于他们对平台禁止该类别的支持程度。

方法

本研究经 Institute[i] IRB 审查后被认定为可豁免审查。我们通过 Lucid[ii] 招募参与者。Lucid 是一家为研究人员提供具有全国代表性的样本的调查公司。我们的调查参与者纳入标准是所有居住在美国的成年互联网用户。我们通过 Lucid 系统向参与者支付报酬。

我们设计在线调查问题，以检验先前提出的假设并回答研究问题。我们根据相关的既有文献调整了调查工具，用于评估部分测量指标，具体如下所述。为提高调查的效度，我们请作者所在机构中参与社会计算研究的学生，对早期调查问卷草稿提供反馈。9名未参与该项目的学生提供了

关于问题措辞和调查流程的反馈，我们据此完善了最终调查设计。我们还通过 Lucid 邀请 27 名参与者对调查进行了试点测试。在此次试点测试中，我们在调查的三个不同位置加入了以下开放式问题："您对目前为止的任何问题有何反馈？例如，是否有问题不清楚或含糊不明？请列出该问题，并描述您在回答时遇到的困难。"在调查末尾，我们还询问："总体而言，从调查参与者的角度来看，我们可以如何改进这份调查？您还有其他想法或反馈吗？请详细说明。"这次调查预试的结果促使我们又进行了一轮修改。

我们的调查问卷包含三个题块，分别围绕仇恨言论、露骨性内容和暴力帖子设置了类似的问题。为消除呈现顺序对调查结果的影响，我们对向参与者展示这三个题块的顺序进行了平衡处理。每个题块开头都会说明后续问题所涉及的违反规范的类别，并采用以下定义对该类别进行界定：

- 仇恨言论："仇恨言论是指基于种族/族裔、性别、性取向、宗教等身份标识，进行非人化、刻板化或侮辱性的言论。"

- 暴力内容："暴力内容包括暴力威胁、美化暴力或庆祝他人受苦、毫无必要或血腥的暴力描绘，以及虐待动物。"

- 露骨性内容："露骨性内容包括展示性活动、提供或寻求性活动的内容、女性乳头（哺乳、健康和抗议行为相关内容除外）、展示生殖器的裸体，以及色情露骨语言。"

上述定义参考了 Facebook 网站在举报仇恨言论、暴力和裸体帖子时使用的措辞。我们使用 Qualtrics 调查软件包在线开展了调查。调查于2022年11月29日启动。表1呈现了数据清理后最终样本的人口统计资料，并将其与美国成年互联网人口的人口统计资料进行比较（Ruggles等，2022年）。

表1：美国调查的人口统计概况

	作者研究，美国调查 2022年11月（%）	美国社区调查， 美国样本2021（%）
Age:		
18-29	18.7	17.4
30-49	40.1	29.5
50-64	23.7	25.6
65+	17.4	27.3
性别：		
Male	48.2	48.6
女性	51.8	51.4

种族/族裔：		
白人	73.5	68.3
黑人	12.2	9.3
其他	14.3	22.4
西班牙裔：		
Yes	4.5	13.7
教育程度：		
高中及以下	31.4	33.5
部分大学教育	24.4	33.3
大学+	43.3	33.1

测量

对自身和他人的感知影响

针对每一种违反规范的言论类别，我们请参与者估计该类别对自己和他人的影响。每种情况下，我们都提出了两个改编自（Rojas 等人，1996年）的问题："在社交媒体上看到 < speech category> 帖子，会对我的态度产生强烈影响。"以及"在社交媒体上看到 < speech category> 帖子，会对我的行为产生强烈影响。"回答选项采用7点李克特量表，范围从1（强烈不同意）到7（强烈同意）。我们对这两个条目的得分取平均值，以衡量受访者感知到的各言论类别对自身的影响程度（仇恨言论：*M*=3.89，*SD*=1.91，α=.84；暴力言论：*M*=3.78，*SD*=1.87，α=.81；露骨色情言论：*M*=3.55，*SD*=1.89，α=.87）。我们还提出了另外两个问题，仅将"我的"替换为"其他人的"，并对回答取平均值，以构建衡量受访者感知到的各言论类别对他人影响程度的指标（仇恨言论：*M*=5.28，*SD*=1.46，α=.92；暴力言论：*M*=5.19，*SD*=1.40，α=.91；露骨色情言论：*M*=4.99，*SD*=1.46，α=.93）。

对言论自由的支持

我们使用 Guo 和 Johnson 开发的条目来测量对言论自由的支持（Guo 和 Johnson，2020年）。参与者使用李克特量表对以下四项陈述进行评分，量表范围从 1（强烈不同意）到 7（强烈同意）：(1)"总体而言，我支持第一修正案。"(2)"言论自由对民主至关重要。"(3)"当公民在不受管制的思想市场中交流时，民主运作得最好。"(4)"即使是极端观点，也值得在社会中表达。"我们在第一道题中加入了第一修正案的文本[iii]，以阐明其含义。我们取这四个条目的平均值，构建了言论自由支持指数（M=5.43，SD=1.15，a=.80）。

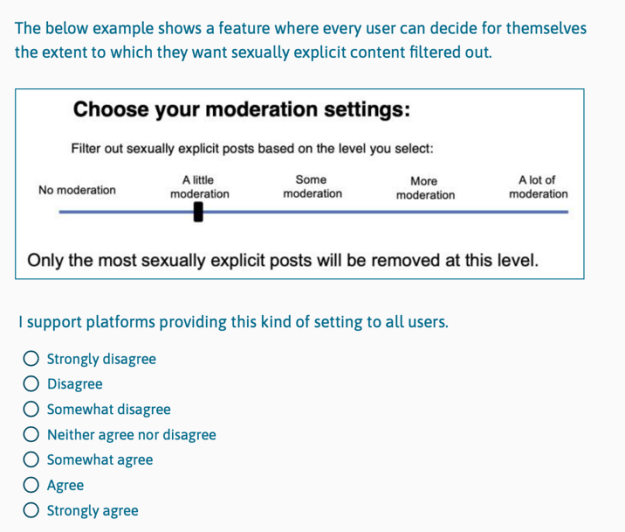

图 *2*：调查问题，请参与者评价他们对平台提供个人内容审核功能的支持程度。

与内容审核相关的因变量

平台实施的内容审核支持通过询问参与者对以下陈述的评分来操作化衡量：“我支持社交媒体平台删除其认定为 <言论类别> 的任何帖子，使所有用户都无法看到。”该陈述的回答采用 7点李克特量表，范围从 1（强烈不同意）到 7（强烈同意）。

为了将对各类别使用个人内容审核工具的支持程度操作化，我们向参与者展示了个人内容审核功能的示例：每位用户都可以自行决定过滤掉多少<言论类别>（见图 2）。我们请参与者使用李克特量表评价他们对向所有用户提供此类设置的支持程度，量表范围为 1（强烈反对）到 7（强烈赞成）。

除了这两项测量之外，我们还针对每种言论类别，通过向受访者提出一个二元问题，将选择平台实施的内容审核还是个人内容审核 操作化：“对于处理 <言论类别> 帖子，如果可以在全平台内容审核和‘选择你的内容审核设置’功能之间进行选择，你更希望采用哪一种？”回答选项包括：（1）“全平台内容审核：平台应有权删除其认定属于全平台<言论类别> 的所有帖子”，或（2）“‘选择你的内容审核设置’功能：每位用户都应能够自行设置为其删除<言论类别> 帖子的程度。”我们指出，“选择你的内容审核设置”功能提供了一系列选择——从“不进行内容审核”到“高度内容审核”，如图 2 所示。因此，如果用户不希望通过全平台或个人内容审核来删除某一言论类别的帖子，可以选择此功能，并将其设置为“不进行内容审核”。

控制变量

既有研究表明，社会人口统计变量与第三人效果、言论自由以及对媒体监管的态度相关（Lambe, 2002；Gunther, 2006年；Lee, 2009；Lo 和 Chang, 2006年）。此外，社交媒体使用也与个人对有害内容的感知及其对此类内容采取的干预措施相关（Naab 等，2018年；Kalch 和 Naab，2017年；Kenski 等，2020年；Ziegele 等，2020年）。因此，我们控制了每位受访者的年龄、教育程度、性别、种族、政治派别（1 = “坚定的民主党人”、 7= “坚定的共和党人”）以及社交媒体使用情况。我们参照 Ernala 等人的建议（Ernala 等，2020年），通过请参与者回答以下问题，将社交媒体使用频率操作化：“在过去一周中，平均而言，你每天主动使用任何社交媒体的时间大约有多长？”

结果

结果显示，分别有72.8%、73.1%和66.2%的参与者至少在一定程度上同意平台应禁止仇恨言论、暴力内容和露骨性内容。此外，分别有69.9%、72.6%和72.1%的参与者至少在一定程度上同意平台应提供个人内容审核工具，让终端用户自行管理仇恨言论、暴力内容和露骨性内容（见图3）。

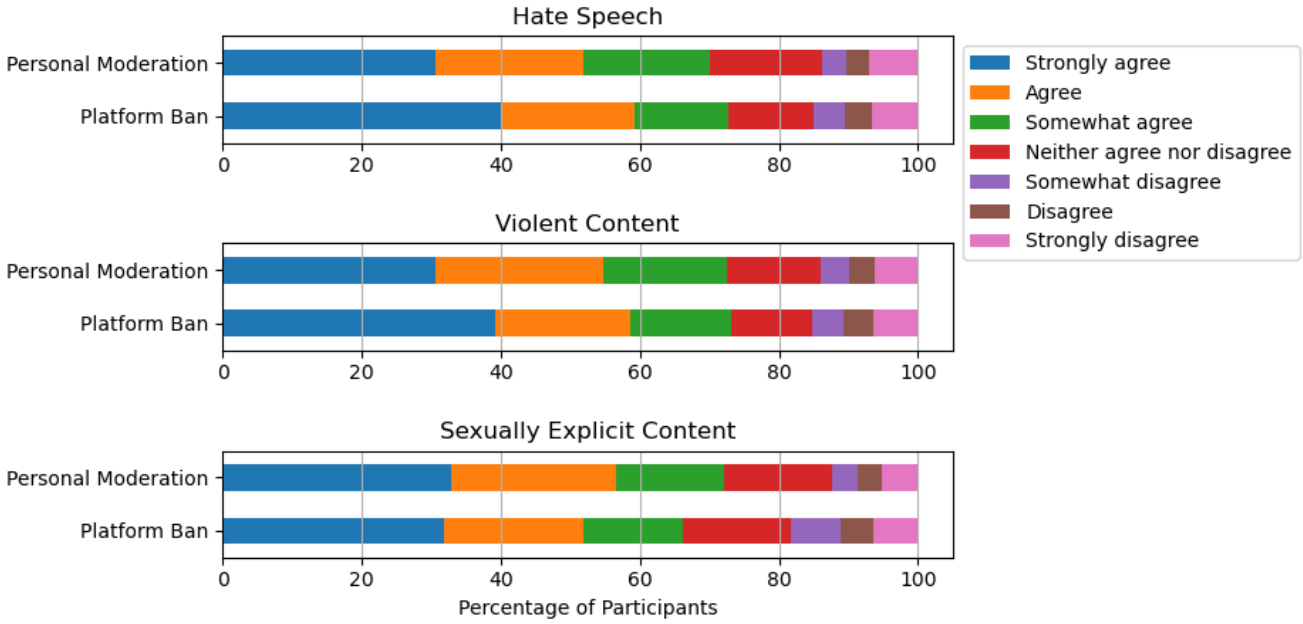

图*3*：参与者对有关支持平台整体内容审核和个人内容审核的调查问题的回答频率，以百分比表示。

与第三人效果相关研究的结论一致，假设*1* 预测，对于每一种违背规范的内容类别，参与者都会认为该类别对他人造成的影响强于对自己的影响。我们进行了配对*t*检验，发现对于每一种类别，参与者对他人影响的感知都显著强于对自身影响的感知（见表2）。

因此，我们的结果支持假设1。

表*2*：参与者对仇恨言论、暴力内容和露骨性内容对他人及自身影响的感知均值、标准差、标准误，以及比较各言论类别对他人和自身影响感知的*t*检验结果（*N = 993*）。***表示*p < .001*

		M	*SD*	*SE*	*t*	Cohen 效应量 *d*
仇恨言论	对他人的影响	5.28	1.46	.05	25.37***	.805
	对自身的影响	3.89	1.91	.06		
暴力内容	对他人的影响	5.19	1.40	.05	25.44***	.807
	对自身的影响	3.77	1.87	.06		
性方面露骨内容	对他人的影响	4.99	1.46	.05	25.81***	.819
	对自身的影响	3.55	1.88	.06		

对平台禁令的支持

我们采用分层线性回归来检验假设2和假设4。针对每一类违反规范的内容，我们都建立了一个模型，其中参与者对平台封禁该类内容的支持程度作为因变量。在三个回归模型的步骤1中，我们纳入了控制变量：年龄、性别、教育程度、种族、政治派别和社交媒体使用。在步骤2中，我们引入了该类别的自变量 PME3（对他人影响的感知）以及对言论自由的支持（表3）。

表*3*：预测对平台禁止仇恨言论、暴力内容和露骨性内容的支持度的分层多元回归分析（*N = 983*）。

自变量	对平台禁止仇恨言论的支持度（*β*）	对对……的平台禁令暴力内容 (*β*)	对……的支持对……的平台禁令露骨性内容 (*β*)
步骤1			
Age	.119***	.055	.102**

性别（女性）	.127***	.167***	.153***
种族（白人）	.018	.013	-.059
教育程度[a]	-.008	.039	-.022
政治派别[b]	-.156***	-.09**	.019
社交媒体使用[c]	.022	.013	.019
R^2	.093***	.065***	.05***
步骤2			
对言论自由的支持	-.101***	-.039	-.006
仇恨言论对他人的感知影响	.476***	-	-
感知到的暴力内容对他人的影响	-	.397***	-
对露骨性内容影响的感知 对他人的影响	-	-	.391***
R^2 变化	.215	.15***	.149***
总 R^2	.307***	.215***	.199***

$p < .01$，*$p < .001$（β 的双尾 t 检验；R^2, 的双尾 F检验）。

[a]0= 中等教育以下；1= 中等教育或以上。

[b]1= 坚定的民主党人，7= 坚定的共和党人。

[c]1= 每天少于 10 分钟，6= 每天超过 3 小时。

β = 完整模型中的标准化贝塔系数（控制模型中所有变量后的最终贝塔系数）。

对于每一种违反规范的言论类别，回归模型均显示，参与者对他人受到该类别言论影响的感知（PME3）会显著影响其对平台禁止该类别言论的支持（模型 1：仇恨言论 – β = .476，$p < .001$；模型 2：暴力内容 – β = .397，$p < .001$；模型 3：露骨性内容 – β = .391，$p < .001$），这支持了 H2。

对言论自由的支持程度越高，对仇恨言论采取平台禁令的支持就显著越低（β = -.101，$p < .001$）。然而，这并不会影响人们对

对暴力内容采取平台禁令的支持（β = -.039，p > .05），也不会影响人们对露骨性内容采取平台禁令的支持（β = -.006，p > .05）。因此，H4仅得到部分支持。

对个人内容审核的支持

我们采用分层线性回归来检验假设3并回答研究问题2。针对每个违反规范的类别，我们分别建立了一个模型，将参与者对使用个人内容审核工具来管理该类别的支持作为因变量。在三个回归模型的步骤1中，我们纳入了年龄、性别、教育程度、种族、政治派别和社交媒体使用等控制变量。在步骤2中，我们加入了该类别的自变量 PME3（对他人影响的感知）以及对言论自由的支持（表4）。

表*4*：预测支持使用个人内容审核工具（*PMT*）来管理仇恨言论、暴力内容和露骨性内容的分层多元回归分析（*N = 983*）。

自变量	支持 使用 PMT 来 管理仇恨言论 言论（β）	对 PMT 的支持 来监管 暴力内容 (β)	对 PMT 的支持 来监管 露骨性 内容 (β)
步骤1			
Age	.014	-.017	-.010
性别（女性）	-.016	.019	-.016
种族（白人）	-.019	-.007	.08*
教育程度[a]	-.018	.022	.003
政治派别[b]	-.019	-.049	-.074*
社交媒体使用[c]	.051	.057	.086**
R^2	.009	.013	.027***
步骤2			
对言论自由的支持	.173***	.195***	.224***
仇恨言论对他人的感知影响	.224***	-	-

感知到的暴力内容对他人的影响	-	.187***	-
对露骨性内容影响的感知 对他人的影响	-	-	.225***
R^2 变化	.087	.081	.111
总R^2	.096***	.095***	.138***

*p < .05，**p < .01，***p < .001（β的t检验，双尾；R^2, 的F检验，双尾）。

[a]0= 中等教育以下；1= 中等教育或以上。

[b]1= 坚定的民主党人，7= 坚定的共和党人。

[c]1= 每天不到10分钟，6= 每天超过3小时。

β = 完整模型的标准化贝塔系数（控制模型中的所有变量后得到的最终贝塔系数）。

对于每一种违反规范的言论类别，回归模型均显示，参与者对该类别对他人产生的感知影响（PME3）会显著影响他们对使用个人内容审核工具来管理该类别的支持度（模型4：仇恨言论 – β = .224，p < .001；模型5：暴力内容 – β = .187，p < .001；模型6：露骨性内容 – β = .225，p < .001），支持了 H4。

在全平台审核与个人审核之间做出选择

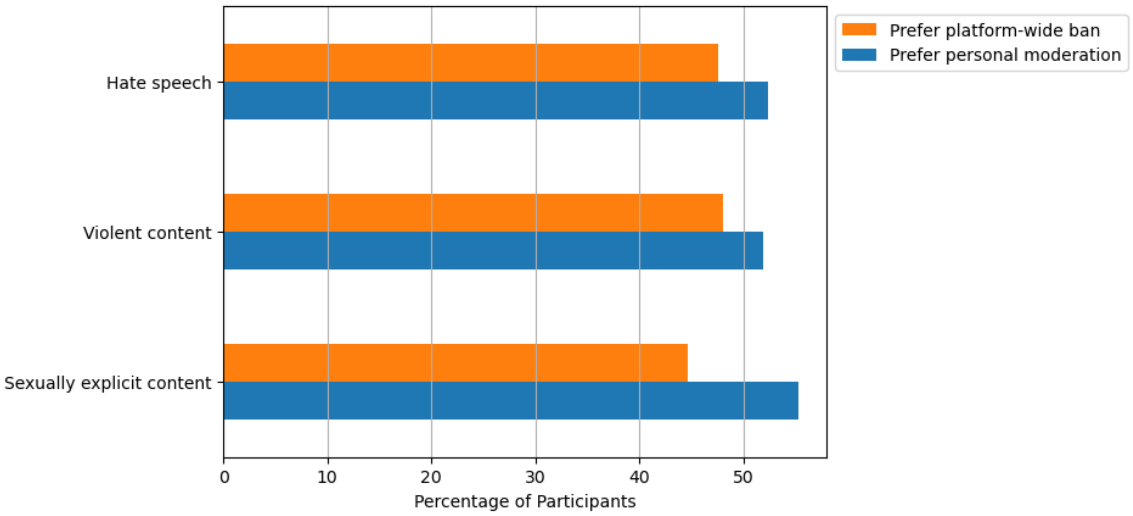

图4：倾向于通过全平台封禁与个人内容审核来管理仇恨言论、暴力内容和露骨性内容的参与者比例。

对言论自由的支持度越高，参与者越支持使用个人内容审核工具来管理各类违反规范的内容（模型4：

仇恨言论 – β = .173，p < .001；模型5：暴力内容 – β = .195，p < .001；模型6：露骨性内容 – β = .224，p < .001 ）。这回答了我们的研究问题 2。
在全平台内容审核与个人审核之间进行选择，以管控仇恨言论、暴力内容和露骨性内容时，分别有52.4%、52%和55.3%的参与者选择了个人审核工具。这一发现表明，相较于将内容审核委托给平台，更多参与者更倾向于拥有审核自主性。

我们创建了二项逻辑回归模型，以检验我们的假设5并回答研究问题1。针对每个违反规范的类别，我们分别建立了一个模型，其中参与者在全平台内容审核与个人内容审核之间针对该类别所作的二元选择作为因变量。在三个回归模型的步骤1中，我们纳入了控制变量年龄、性别、教育程度、种族、政治派别和社交媒体使用。在步骤2中，我们引入了该类别的自变量PME3（对他人影响的感知）以及对言论自由的支持。对于每个模型，我们使用Box-Tidwell检验（Box和Tidwell，1962年；Fox，2015年）来检验逻辑值中的线性关系假设。我们发现，在每种情况下，连续变量对言论自由的支持与因变量的逻辑值均不存在线性关系。为解决这一问题，我们将该变量划分为两个有序类别，即高和低，并根据各条目是否超过中位数对该变量的每个条目进行重新编码。使用转换后的对言论自由支持的分类变量重新进行Box-Tidwell检验后，我们发现每个模型中其余所有连续自变量均与因变量的逻辑值呈线性关系。

接下来，我们展示这些模型的二项逻辑回归结果（表5）。

表*5*：二项逻辑回归分析：预测使用个人内容审核工具（*PMT*）而非全平台封禁来管制仇恨言论、暴力内容和露骨性内容的支持度（*N = 984*）。

自变量	支持使用 PMT 而非平台禁令来管制仇恨言论言论，优势比	对 PMT 的支持而非平台禁令以管理暴力内容，优势比	对 PMT 的支持相较于平台用于监管的禁令暴力内容，优势比
步骤1			
Age	.986**	.992	.994
性别（女性）	.790	.655**	.674**
种族（白人）	1.107	1.313	1.203
教育程度[a]	.934	.870	.981

政治派别[b]	1.163***	1.118***	1.074*
社交媒体使用[c]	.990	1.071	1.091*
Nagelkerke R^2	.066***	.066***	.045***
步骤2			
对言论自由的支持[d]	2.703***	2.239****	1.969***
仇恨言论对他人的感知影响	.735***	-	-
感知到的暴力内容对他人的影响	-	.752***	-
对露骨性内容影响的感知 对他人	-	-	.857**
总 Nagelkerke R^2	.165***	.138***	.087***

*$p < .05$，**$p < .01$，***$p < .001$（β 的 t 检验，双尾；R^2 的模型系数总体检验）。

[a]0= 低于中等教育；1= 中等教育或以上。

[b]1= 坚定的民主党人，7= 坚定的共和党人。

[c]1= 每天少于10分钟，6= 每天超过3小时。

[d]0= 低，1=高。

优势比 = $exp(\beta)$（完整模型）。

对于每一类违反规范的言论类别，回归模型显示，参与者对该类别内容对他人所产生影响的感知（PME3）会显著影响其选择使用个人内容审核工具而非平台实施的禁令来规范该类别的倾向。PME3 越高，参与者选择个人内容审核工具而非平台封禁的可能性越低（模型7：仇恨言论 – $\exp(\beta) = .735$，$p < .001$；模型8：暴力内容 – $\exp(\beta) = .752$，$p < .001$；模型9：露骨性内容 – $\exp(\beta) = .857$，$p < .01$）。这回答了我们的研究问题1。

对言论自由的支持越高，参与者越倾向于选择个人内容审核工具，而非平台实施的禁令，来管理每一类违反规范的内容。对言论自由支持度较高的参与者，在选择个人内容审核工具而非平台实施的禁令来管理仇恨言论、暴力内容和露骨性内容时，几率分别是其他参与者的2.703倍、2.239倍和1.969倍。因此，假设5得到支持。

讨论

第三人称效应（TPE）常用于考察人们对审查的态度，以及与之相关的行为或行为意向（Davison，1983年；Gunther，1991年）。本文将TPE研究拓展至社交媒体上仇恨言论、暴力内容和露骨性内容的管理。我们探讨了人们假定这些类型的内容对自身（PME1）和他人（PME3）造成的负面影响。结果支持第三人称效应假设：正如预期，参与者认为，社交媒体上的仇恨言论、暴力内容和露骨性内容对他人的影响大于对自身的影响。

我们还考察了PME3和对言论自由的支持如何影响参与者随后采取的审查行为。在每种情况下，我们都发现，对他人影响的感知（PME3）能够预测参与者对全平台审核与个人审核的支持。这一具有重要理论意义的发现表明，对他人影响的感知在触发审查行为方面发挥着关键作用，从而推动了第三人效果研究的发展。在全平台内容审核与个人内容审核之间进行选择时，在每一类内容中，更高的PME3都能够预测参与者更偏好平台内容审核，而非个人内容审核。这一发现表明，当用户感知到某类内容对公众造成不利影响时，他们希望平台对该内容采取全站范围的行动，而不是由自己进行调节。这说明，公众希望平台承担起保护弱势群体免受被认为恶劣内容侵害的责任，即使这会牺牲个人内容审核控制权。

我们仅发现了有关言论自由与全平台内容审核支持之间关系的部分证据。虽然在仇恨言论方面，这一关系显著且为负向，但对于暴力内容和露骨性内容而言并不显著。这一结果与既有文献中的混合发现一致：此前研究观察到，对言论自由的支持是支持平台审查态度的负向预测因素（Guo和Johnson，2020年）。另一方面，我们发现，对言论自由的支持能够预测参与者支持使用个人内容审核来管理每一类不当言论。这表明，人们可能不会将个人内容审核工具视为侵犯他人的言论自由，而只是认为这些工具赋予个人更大的自主权，使其能够塑造自己所看到的内容。我们的另一项发现进一步支持了这一解释：在全平台内容审核与个人内容审核之间进行选择时，在每一类内容中，对言论自由的支持都能够预测参与者对个人内容审核的支持。

此外，我们还发现了其他一些与待检验假设关联较小但显著的影响。年龄与支持平台审核仇恨言论和露骨性内容呈正相关，但与支持平台审核暴力内容无关。女性对各类言论的平台审核的支持度均高于男性。民主党人比共和党人更可能支持平台审核仇恨言论和暴力内容，但在露骨性内容方面并非如此。至于对个人内容审核的支持，种族、政治派别和社交媒体使用是露骨性内容类别的重要预测因素；然而，没有任何控制变量能显著预测对仇恨言论或暴力内容进行个人内容审核的支持。未来研究可以考察用户的政治意识形态、媒体消费、父母担忧以及过往社交媒体经历如何影响其内容审核偏好，从而进一步阐释这些结果。

本文提出的证据对于平台如何治理其网站具有重要意义。我们表明，公众支持对色情内容和战争暴力等违反规范的类别进行平台内容审核和个人内容审核，部分原因在于公众高估了这些类别对他人的影响。因此，公司层面的内容审核决策以及有关言论自由及其限制的公共讨论，都必须认识到并考虑第三人效果。这也凸显了迫切需要衡量社交媒体的实际影响，而不是不同内容类型所感知到的影响。即使在我们针对网络伤害开展的独立访谈研究中，[iv] 我们也注意到，参与者往往会根据自己对他人可能需要什么的看法，倡导特定的内容审核举措。为避免第三人效果造成的偏差，学者必须从用户自身的视角考察网络内容——用户在这一议题上才是专家——而不是从抽象的他人视角出发，因为后者的实际需求可能大不相同。

我们发现，在全平台审核和个人审核之间进行选择时，参与者更倾向于后者来管理各类违反规范的言论（图4）。然而，目前许多平台并不支持个人内容审核，或对其支持不力（Jhaver 等人，2022）。我们的研究结果支持一种兼顾平台审核与个人审核、并使二者都能有效运作的方案。平台可以减少对边界性内容（例如不会引发高 PME3 的内容）实施全站禁令，转而赋予用户个人内容审核工具，让用户能够为此类内容设定自己的内容审核偏好。

我们应当认识到本研究存在若干局限。调查设计使我们无法深入探究特定看法背后的动机。此外，在横断面研究中，我们无法对因果关系作出确切判断。我们请参与者针对可作广义理解的言论类别回答问题；为提高研究结果的普适性，我们选择了这种做法，而没有为每种言论类别提供具体实例。尽管如此，不同用户对于何为仇恨言论、暴力内容或露骨性内容，可能有不同的看法。以往的内容审核研究已指出，这是社交媒体监管中一项复杂的挑战（Jhaver 等人，2018）。我们在调查中为每种言论类别提供了定义，以便向参与者明确各类别的范围。

尽管如此，进一步研究用户对基于刺激材料的设计的看法——这类设计会向参与者呈现每种违反规范的言论类别中预先选定的实例——将提供有价值的见解。我们预设的平台设置是：平台界定不当言论类别，并标记哪些帖子属于这些类别。不过，各网站都必须考察其内容审核实施方式是否能够适应这种安排。最后，为了提高研究结果的普遍适用性，我们的调查问题并未以某个特定平台为背景。聚焦于特定社交媒体网站的研究，可以揭示用户对特定平台的态度是否会影响他们对内容审核措施的看法。

[i] 待同行评审流程完成后，我们将注明研究所名称。 [ii] https://lucidtheorem.com [iii] 美国宪法第一修正案规定：“国会不得制定确立宗教或禁止自由信奉宗教的法律，亦不得剥夺言论自由，或剥夺

新闻自由；或人民和平集会，以及向政府请愿以求纠正冤情的权利。”[iv] 我们将在同行评审后补充引文。

参考文献

Balkin JM（2017）算法社会中的言论自由：大数据、私人治理与新型言论监管。*UCDL Rev.* 51：1149。Barnidge M 和 Rojas H（2014）敌意媒体感知、假定媒体影响与政治讨论：拓展纠正行动假说。《国际舆论研究杂志》 26(2)：135-156。Bernhard U 和 Dohle M（2015）纠正行动还是确认行动？选举活动中，假定媒体影响所引发的政治性网络参与。《信息技术与政治杂志》 12(3)：285-302。Box GEP 和 Tidwell PW（1962）自变量变换。*Technometrics* 4(4)：531-550。Chung S 和 Moon S-I（2016）第三人效果是真实存在的吗？对第三人效果影响审查态度的理论依据、检验方法和既有研究发现进行批判性考察。《人类传播研究》 42(2)：312-337。Davison WP（1983）传播中的第三人效果。《公众舆论季刊》 47(1)：1-15。Einwiller SA 和 Kim S（2020）网络内容提供者如何审核用户生成内容以防止有害网络传播：政策及其实施分析。《政策与互联网》 12(2)：184-206。Ernala SK、Burke M、Leavitt A 等（2020）人们报告自己在 Facebook 上花费的时间有多准确？对既有调查问题的评估及建议。载于：*2020*年计算机系统中的人为因素会议论文集，美国夏威夷州檀香山，第1–14页。计算机协会。Feldman L、Hart PS、Leiserowitz A 等（2017）敌意媒体感知会促使人们采取行动吗？敌意媒体感知、政治效能感与意识形态在预测气候变化行动主义中的作用。《传播研究》 44(8)：1099-1124。Fox J （ 2015 ）应用回归分析与广义线性模型。Sage Publications。Gibson JL 和 Bingham RD（1982）政治宽容的概念化与测量。《美国政治学评论》 76(3)：603-620。Gillespie T（2010）“平台”的政治。《新媒体与社会》 12(3)：347-364。Gillespie T（2017）平台的治理与平台治理。《*Sage*社交媒体手册。Golan GJ 和 Banning SA（2008）探讨第三人效果与理性行为理论之间的联系：公益广告与社会期望。《美国行为科学家》 52(2)：208-224。Golan GJ、Banning SA 和 Lundy L（2008）投票可能性、候选人选择与第三人效果：政治广告的行为影响及其在2004年总统选举中的作用。《美国行为科学家》 52(2)：278-290。Golan GJ 和 Lim JS（2016）*ISIS* 招募宣传的第三人效果：网络政治自我效能感与社交媒体行动主义。Gunther A（1991）我们认为他人怎么想：第三人效果的原因与后果。《传播研究》 18(3)：355-372。

Gunther AC（2006）《高估X级评级：第三人感知与对色情内容审查的支持》。《传播学杂志》45(1)：27-38。Guo L 和 Johnson BG（2020）《第三人效果与 Facebook 上的仇恨言论审查》。《社交媒体与社会》6(2)：2056305120923003。Hoffner C、Buchanan M、Anderson JD 等（1999）《支持审查电视暴力内容：第三人效果与新闻接触的作用》。《传播研究》26(6)：726-742。Jang SM 和 Kim JK（2018）《假新闻的第三人效果：假新闻监管与媒体素养干预》。《人类行为中的计算机》80：295-302。Jhaver S、Chan L 和 Bruckman A（2018）《来自另一方的视角：Kotaku in Action 上争议性言论与骚扰之间的界限》。《First Monday》23(2)。Jhaver S、Chen QZ、Knauss D 等（2022）《为创作者主导的评论审核设计词语过滤工具》。《2022年计算机系统中的人为因素会议论文集》。计算机协会。Kalch A 和 Naab TK（2017）《回复、点踩、标记：用户如何处理新闻网站上的不文明和不礼貌评论》。Kenski K、Coe K 和 Rains SA（2020）《对网络不文明话语的感知：类型与预测因素考察》。《传播研究》47(6)：795-814。Lambe JL（2002）《审查维度：重新概念化公众审查意愿》。《传播法与政策》7：187-235。Lee FLF（2009）《第三人感知的预防效应：对民意调查感知影响与实际影响的研究》。《大众传播与社会》13(1)：87-110。Lim JS（2017）《整形手术在线广告的第三人效果：预测限制性行动与纠正性行动的路径模型》。《新闻与大众传播季刊》94(4)：972-993。Lim JS 和 Golan GJ（2011）《针对 YouTube 政治恶搞视频影响的社交媒体行动主义》。《传播研究》38(5)：710-727。Livingstone S（2013）“知识提升”：循证政策的风险与机遇。Lo V-h 和 Chang C（2006）《海湾战争知识：从新闻中学习的理论模型》。《哈佛国际新闻/政治期刊》11(3)：135-155。Massanari A（2017）《#Gamergate 与 The Fappening：Reddit 的算法、治理与文化如何助长有毒技术文化》。《新媒体与社会》19(3)：329–346-329–346。McLeod DM、Eveland WP 和 Nathanson AI（1997）《支持审查暴力和厌女说唱歌词：第三人效果分析》。《传播研究》24(2)：153-174。Medeiros B（2017）《平台（不）干预与言论监管的“市场”范式》。《社交媒体与社会》3(1)：2056305117691997。Naab T（2012）《在不断变化的媒体环境中，人们对言论自由的态度之重要性》。《ESSACHESS传播学期刊》5(1)。Naab TK、Kalch A 和 Meitz TG（2018）《标记不文明用户评论：干预信息、受害者类型和回应评论对旁观者行为的影响》。《新媒体与社会》20(2)：777-795。

Naab TK、Naab T 和 Brandmeier J（2021）《不文明用户评论会提高用户采取纠正性行动的意愿及其对权威限制性行动的支持》。《新闻与大众传播季刊》98(2)：566-588。Paek H-J、Lambe JL 和 McLeod DM（2008）《支持内容限制的前因》。《新闻与大众传播季刊》85(2)：273-290。Paek H-J、Pan Z、Sun Y 等（2005）《作为社会判断的第三人感知：探究社会距离、不确定性与政治攻击广告感知影响之间的关系》。《传播研究》32(2)：143-170。Price V 和 Tewksbury D（1996）《衡量新闻的第三人效果：问题顺序、对比与知识的影响》。《国际舆论研究杂志》8(2)：120-141。Puppis M 和 Van den Bulck H（2019）《导论：媒体政策与媒体政策研究》。载于：Van den Bulck H、Puppis M、Donders K 等（编）《帕尔格雷夫媒体政策研究方法手册》。沙姆：施普林格国际出版，第3-21页。Riedl MJ、Naab TK、Masullo GM 等（2021a）《谁应对问题评论采取干预措施？比较德国和美国用户的态度》。《政策与互联网》。Riedl MJ、Whipple KN 和 Wallace R（2021b）《支持社交媒体内容审核和平台监管的前因：对自身及他人所受影响的推测所发挥的作用》。《信息、传播与社会》。DOI：10.1080/1369118X.2021.1874040。1-18。Rojas H（2010）公共领域中的“纠正性”行动：媒体及媒介效果感知如何塑造政治行为。《国际舆论研究杂志》22(3)：343-363。Rojas H、Shah DV 和 Faber RJ（1996）《为了他人的利益：审查与第三人效果》。《国际舆论研究杂志》8(2)：163-186。Ruggles S、Flood S、Goeken R 等（2022）《IPUMS USA》。第12.0版。明尼阿波利斯，明尼苏达州。Schweisberger V、Billinson J 和 Chock TM（2014）《Facebook、第三人效果与差异影响假说》。《计算机中介传播杂志》19(3)：403-413。Seering J、Kaufman G 和 Chancellor S（2020）《内容审核中的隐喻》。《新媒体与社会》24(3)：621-640。Seering J、Wang T、Yoon J 等（2019）《算法时代的版主参与与社群发展》。《新媒体与社会》。1461444818821316-1461444818821316。Stecklow S（2018）《Facebook 为何在缅甸打击仇恨言论的战争中节节败退》。汤森路透。Sullivan JL、Piereson J 和 Marcus GE（1993）《政治宽容与美国民主》。《芝加哥大学出版社》。Sun Y、Pan Z 和 Shen L（2008）《理解第三人感知：荟萃分析的证据》。《传播学杂志》58(2)：280-300。Wei R、Chia SC 和 Lo V-H（2011）《第三人效果与敌意媒体感知对选民在2008年美国总统选举中民意调查态度的影响》。《国际舆论研究杂志》23(2)：169-190。

Williams M, Butler M, Jurek-Loughrey A, 等（2021）《冒犯性传播：探索社交媒体治理所面临的挑战》。*Contemporary Social Science* 16(2): 227-240。Yalof DA 和 Dautrich K（2002）舆论法庭中的第一修正案与媒体。剑桥大学出版社。Ziegele M, Naab TK 和 Jost P（2020）《共同的孤独？确定针对不文明评论采取集体纠正行动的决定因素》。《新媒体与社会》 22(5): 731-751。