夺回掌控权：Instagram 上针对仇恨言论的屏幕式个人内容审核

Anna Ricarda Luther
aluther@ifib.de
不来梅信息管理研究所
有限责任公司
德国不来梅
不来梅大学
德国不来梅

Hendrik Heuer
hendrik.heuer@cais-research.de
高级互联网研究中心
(CAIS)
德国波鸿
伍珀塔尔大学
德国伍珀塔尔

Stephanie Geise
sgeise@uni-bremen.de
媒体与传播中心
与信息研究中心（ZeMKI）
德国不来梅
不来梅大学
德国不来梅

Sebastian Haunss
s
ebastian.haunss@uni-bremen.de
不平等与社会政策研究中心（
SOCIUM）德国不来梅 不来梅大学 德
国不来梅

Andreas Breiter
abreiter@ifib.de 不来梅信息管理
研究所有限公司 德国不来梅 不来
梅大学 德国不来梅

arXiv:2603.01187v1 [cs.HC] 1 Mar 2026

摘要

仇恨言论仍是社交媒体面临的紧迫挑战，平台审核往往无法保护被针对的用户。个人审核工具让用户能够决定如何筛选内容，可以弥补其中的一些不足。然而，用户希望在哪些屏幕上（例如评论区、Reels 标签页或主页信息流）使用个人审核，以及他们最看重哪些功能，仍是开放式问题。为弥补这些空白，我们与 40 名曾遭遇仇恨言论的活动人士开展了三轮德尔菲研究。我们将定量评分和排名与有关所需功能的开放式问题相结合。参与者优先选择在对话型和算法策划界面上使用个人审核。他们看重可在不同屏幕间实现可逆性和监督的功能；而基于输入、针对特定内容类型以及高度自动化的功能，则更具屏幕特异性。我们探讨了个人审核的重要性，并针对 Instagram 上的个人审核提出以用户为中心的设计建议。

CCS概念

• **以人为中心的计算 → 协作与社交计算实证研究；人机交互实证研究**。

关键词

社交媒体、内容审核、仇恨言论、个人审核、德尔菲研究

1 引言

仇恨言论是社交媒体上一个已被充分记录的问题 [12–14], , 广泛影响着社交媒体用户 [38]。在本文中，我们将仇恨言论操作化为任何基于种族、性别或宗教等特征贬低个人或群体的传播形式 [84]。接触仇恨言论已被证明会产生沉默效应，使用户不愿参与公共讨论或披露其身份的部分内容。但这些后果在线仇恨言论的影响会延伸到数字空间之外，对身心健康构成现实世界的风险 [71, 79]。这一影响的范围与严重性凸显了采取有效策略以减轻仇恨言论并营造更安全在线环境的迫切需要。

现有的内容审核系统主要采用平台审核（即审核决策适用于整个平台） [34], , 但研究表明，这种方式无法有效遏制仇恨言论 [13]。自动仇恨言论检测是平台审核的重要组成部分 [34],, 但通常缺乏判断某一陈述是否具有攻击性所需的上下文 [25] , 从而导致审核过度或审核不足 [37], , 尤其是对于边缘化用户发布的内容 [25]。用户自身往往了解这些上下文。个人审核工具能够让用户提供这些上下文，从而弥补自动化平台审核的不足。个人审核是指赋予用户影响其所看到内容的工具，例如通过词语过滤器或敏感度控制设置偏好，并且这些设置仅适用于其账户 [47]。

通过这种方式，个人审核有望创建更安全、更具包容性的在线空间，尊重多样化的文化与个人规范 [32, 39]。既有研究揭示了用户对个人审核工具的偏好及其设计考量 [39, 44, 46]。既有研究发现的一个问题是，个人审核工具需要在控制粒度上匹配用户的审核目标。更具体而言，研究结果表明，用户对于社交媒体应用的不同部分，可能有不同的个人审核需求 [46]。

基于这一视角，我们开展了实证研究，探讨个人审核需求是否以及如何因社交媒体的不同部分而异。为系统地考察这些差异，我们引入了“屏幕”这一理论概念（例如评论区、主页信息流或 Reels 标签页）。由于个人审核需求可以反映在用户所需的功能中 [46], , 我们进一步考察不同屏幕上的个人审核功能

。我们将个人审核功能操作化定义为用户可配置的独立用户界面元素，以便在各自的账户中避免有害内容。我们提出以下研究问题（RQs）：

- 哪些屏幕应为遭受仇恨言论攻击的活动人士提供个人审核？（研究问题1）
- 遭受仇恨言论攻击的活动人士希望各个屏幕具备哪些个人审核功能？（研究问题二）

为回答这些研究问题，我们对居住在德国且曾遭遇仇恨言论的40名活动人士开展了三轮德尔菲研究。我们招募活动人士，是因为他们在政治议题上的高度可见参与度使其面临更高的仇恨言论遭遇风险 [55], ，而且其中许多人属于遭受仇恨言论不成比例攻击的边缘化群体 [10]。他们广泛且往往无法避免地依赖社交媒体来扩大自身声音，这进一步增加了他们遭遇骚扰的风险，也提高了管理骚扰的利害关系 [59]。我们聚焦于Instagram，因为它是德国使用最广泛的社交媒体平台 [74], ，尤其受到年轻用户和行动主义用户的青睐 [49, 74]。既有研究也记录了 Instagram 上仇恨言论带来的挑战，因此，该平台是研究个人审核需求的相关且高暴露环境 [57]。

我们发现，个人审核选项对于代表对话空间和算法策划空间的屏幕尤为重要。尽管所有屏幕都需要具备监督和可逆性的功能，但各屏幕所要求的功能也存在差异。基于输入的功能更常被要求用于对话界面，因为在这类界面中伤害类型是已知的。对于算法策划空间，用户则要求针对不同内容类型的功能，以及高度依赖自动化、同时只需极少用户配置的功能。

我们首次基于实证阐述了遭受仇恨言论针对的活动人士希望进行个人审核的屏幕（研究问题1）。针对每个屏幕，我们给出了被认为最重要的功能（研究问题二）。在讨论中，我们将识别出的用户需求转化为针对个人审核工具的具体设计建议。这些贡献共同推进了社交媒体上个人审核的研究，并为设计既可扩展又能回应受仇恨言论影响者需求的内容审核实践提供了可操作的路径。

2 背景

2.1 社交媒体上的不同内容审核类型

我们将本研究置于区分不同内容审核方法的既有学术研究脉络中。Jhaver 等人（2023年）根据各项内容审核决策的适用范围，对社交媒体上的内容审核方法进行了分类，归纳出以下三类：平台审核（适用于整个平台的内容审核决策）、社区内容审核（适用于特定社群的内容审核决策）和个人审核（适用于单个用户视图的内容审核决策）。他们还进一步区分了基于内容的内容审核决策（例如，因内容包含特定关键词而将其删除）和基于账户的内容审核决策（例如，因用户屡次违反平台准则而封禁该用户） [46]。鉴于Instagram 上的社区内容审核仅限于内容创作者各自的频道 [42], ，下文将聚焦个人审核如何弥补平台审核的局限。

2.2 平台审核的局限性

下文将阐述平台内容审核的局限性，并重点说明个人审核如何在一定程度上应对这些局限。平台内容审核的一项关键局限在于其采用一刀切方法。尽管社交媒体上确实存在因其造成的明显伤害而理应一律删除的内容（例如儿童性虐待材料 [34]），但对于大量其他内容而言，这种界限并不那么明确 [50]。何为仇恨言论，取决于不同文化、社群和个人之间各异的规范、身份认同与具体情境 [45, 48, 66, 77]。主要社交媒体平台的用户群体遍布全球，规模达数十亿 [70], ，因此，全平台审核在满足人们对何为可接受言论的多元期待方面，能力有限。个人审核可以将个体用户的视角纳入审核流程，为这种一刀切方法增添更多细致考量。

此外，社交媒体平台主要采用自动化系统来实现基于内容的平台审核 [34]。既往研究一再表明，自动化仇恨言论检测系统对边缘化社群存在偏见 [61]。这体现了边缘化的层层叠加：那些最常成为仇恨言论目标的人，同时也因带有偏见的系统而处于不利地位，从而加剧了他们在线面临的伤害。自动仇恨言论检测中的偏见以两种方式显现：一方面对针对边缘化群体的仇恨言论审核不足 [4, 62] ，同时又对反仇恨言论以及对侮辱性用语的重新挪用过度审核 [23, 25]。重新挪用是指边缘化群体将指称该群体的贬损性用语用作自我描述的做法，构成一种赋权或抵抗形式 [15]。

提升自动仇恨言论检测的性能仍然困难，原因在于若干关键挑战。例如，如何有意义地纳入准确检测仇恨言论所需的语境，仍是一个悬而未决的问题 [25]。然而，用户自身往往清楚用户生成内容是否构成反仇恨言论，或某个词语是否已被重新挪用。个人审核将允许用户提供这一语境，从而能够有效弥补平台审核的这一不足。

用于平台内容审核的自动化仇恨言论检测工具还面临数据漂移问题，即底层数据分布随时间发生变化 [24], ； 概念漂移问题，即输入数据与目标标签之间的关系发生变化 [6], ； 以及模型漂移问题，即模型性能因模式或社会规范不断演变而下降 [52]。这些动态变化至关重要，

因为新的侮辱性用语或暗语不断出现，现有词语的含义也会随时间改变 [72]。此外，违规者可能会故意使用模糊表达或改变拼写，以逃避检测 [35]。因此，预训练模型必然会落后于这些变化。讽刺、反语或隐语的使用进一步加剧了这一问题，导致检测准确率下降 [37, 60]。由于在大多数平台准则中，反复检测有害内容是基于账号的平台审核的先决条件 [41, 54, 73]，这类审核也存在同样的局限性。因此，我们将重点放在个人审核上。通过赋予用户控制审核的能力，审核措施的有效性可以更轻松地适应不断演变的语言和社会规范。

总之，平台审核受到偏见和缺乏语境感知的困扰，对最常遭受仇恨言论攻击的群体造成了尤为严重的伤害。这些局限凸显了采用个人审核等补充方法的必要性，因为这种方法能够将个体用户的需求纳入内容审核流程。

2.3 个人审核

学者和平台设计者都已开始认识到，个人审核是改进社交媒体内容审核的重要方向。在一项针对 984 名美国成年人的全国代表性调查中，Jhaver 和 Zhang（2023年）发现，对于社交媒体上违反规范的内容，用户更倾向于使用个人审核工具，而非平台主导的审核。学者也越来越多地呼吁采取个人审核解决方案 [32, 46, 51]，以改进社交媒体上的内容审核实践。作为回应，社交媒体平台已纳入个人审核功能，多个第三方也开发了个人审核工具。示例包括 Gobo Social [58]、Bodyguard [9]、Block Party [8], 和 FilterBuddy [44]。例如，FilterBuddy 旨在解决基于规则的词语过滤器的不足：这类过滤器要求用户手动输入贬损性词语的每一种语言变体，并预先考虑范围往往令人难以招架的各种词语 [44]。FilterBuddy 通过自动生成变体、支持导入群组专属关键词列表，以及允许用户之间共享列表 [44]，减轻了这一负担。

2.3.1 将屏幕视为个人审核中控制粒度的观察视角。尽管个人审核对用户 [39, 47] 和学者 [32, 46, 51] 都至关重要，且个人审核功能也日益得到应用，但个人审核工具的设计仍存在一些开放式问题。其中一个关键的缺口是此类工具所需的控制粒度。以往研究表明，参与者在社交媒体的不同部分（如新闻动态、私信或用户个人资料）都遭受过伤害，并指出用户在这些部分的个人审核需求可能有所不同 [46]。例如，有些用户可能希望始终过滤掉新闻动态中的某些词语，但不希望在私信中这样做。这些发现促使我们通过实证研究，考察用户需求是否会因平台的不同部分而异。我们将这些部分界定为屏幕，并提出以下问题：遭受仇恨言论攻击的活动人士希望在哪些屏幕上进行个人审核？（研究问题1）

2.3.2 个人审核的关键功能。基于个人审核需求可能因屏幕而异这一假设，我们进一步探讨用户希望在相应屏幕上使用哪些功能。此前的研究已开始探索用户对基于账户和基于内容的个人审核分别有何功能需求 [39, 46]。例如，对现有内容管理工具进行评估的访谈，以及探索新个人审核选项的技术探针，揭示了用户所需功能类型方面的关键见解 [46]。同样，针对残障社交媒体用户开展的设计探针研究指出，需要能够配置对歧视残障人士内容的过滤方式 [39]。尽管这些既有研究奠定了坚实基础，却未将这些功能请求与屏幕关联起来，也未让用户对这些功能进行评分和排序，从而明确优先事项。对这些功能进行优先级排序至关重要，这有助于将个人审核工具的设计聚焦于影响最大且最受期待的解决方案。通过采用德尔菲法，我们得以实现这一目标 [21, 40]，并深入探究第二个研究问题：遭受仇恨言论攻击的活动人士希望在不同屏幕上使用哪些个人审核功能？（研究问题二）

3 方法

3.1 德尔菲法

德尔菲法是一种结构化、多轮次且依赖路径的调查方法，旨在促成一组专家之间达成共识 [80]。我们选择德尔菲法开展本研究，因为它允许参与者在多轮过程中反复完善自己的判断。德尔菲法依赖路径，意味着条目会根据此前匿名汇总的参与者意见进行增补 [21]。未达到预先设定的一致意见门槛的条目也可以删除 [40]。其匿名形式能够限制社会动态对结果的影响 [22, 40]。此外，德尔菲研究常用于技术预测，以应对新兴的设计空间；在此类情境下，达成共识能够为未来发展提供有意义的指导 [2, 80]。最后，德尔菲流程采用漏斗式结构，能够得出稳定的优先事项，为平台设计提供参考 [21]。这种结构化、迭代且以共识为导向的方法，使我们能够系统评估用户根据个人审核功能所应用的屏幕提出了哪些需求。

3.2 研究程序

我们与来自德国、遭受仇恨言论针对的40名社会媒体活动人士开展了一项三轮德尔菲研究。每一波次均通过SoSci Survey以在线调查问卷的形式实施。在正式启动研究之前，我们与本机构的八名同事和研究生进行了三轮非正式预测试。他们就问题表述、说明清晰度以及调查问卷流程提供了反馈，我们在每一轮中加以采纳，最终形成了最终设计。

当地法规并不要求对此类研究进行正式的机构伦理审查。尽管如此，第二作者所在机构的机构审查委员会（机构审查委员会）审查并批准了该研究方案。

每一轮都有其特定目的：第1轮对个人内容审核工具进行了初步评估，并根据用户自由输入扩展了回应选项；第2轮则提供了

对这一初步评估的验证；第3轮则对最重要的屏幕和功能进行了排名。研究流程概览见表1。

在第1轮中，参与者收到以下情境：

> “想象有一款类似 *Instagram* 的社交媒体应用。该应用经常不会移除仇恨言论。作为一名经验丰富的用户，你知道这会带来哪些挑战。现在，你有机会设计一种新工具，让你能更好地控制账户上的仇恨言论过滤方式。想象一下，你可以决定这个工具如何运作。我们想了解你希望这个工具是什么样的。请想想你对仇恨言论的经历，并考虑这个工具如何保护你免受社交媒体上的仇恨言论侵害。你无需考虑自己的想法和愿望在技术上是否可行。”（由德语翻译而来）。

基于这一情境，我们请参与者使用5点李克特量表（1 = 不重要， 2 =不太重要，3 = 既非重要也非不重要， 4 = 比较重要， 5 = 重要，-1 =不知道）对 Instagram 界面中不同屏幕的重要性进行评分，以了解用户认为哪些屏幕对于控制内容过滤最为重要（研究问题1）。我们还请参与者说明他们对每个屏幕的重要性评分。

每个屏幕都配有简短说明和屏幕截图，以确保参与者对各个屏幕的理解一致。所有屏幕、相应说明及其标签均列于表2中。为帮助读者理解，我们在结果部分的图2中展示了相应屏幕的图片。

仅当参与者将某一屏幕评为中性、较为重要或重要时，我们才请该参与者对该屏幕所对应的假设性个人内容审核工具的功能进行评分。随后，我们请参与者为该屏幕提出额外功能。我们向参与者呈现了三项源自文献的功能，以帮助他们将想法具体化。这些示例并非意在限制回答，而是为了激发反思，并确保所有参与者拥有共同的讨论起点，正如德尔菲研究中的常见做法 [22]。第一个选项是**基于词语的过滤**。该功能允许用户输入特定字符串。明确包含这些字符串的内容会被过滤（例如，侮辱性用语或关键词）。选择此功能作为初始输入，是因为它既已在一些社交媒体平台上实现 [30, 42], ，同时也因其局限性而受到批评 [44]。第二个选项是**基于主题的过滤**。该功能使用户能够输入特定主题，随后与这些主题相关的内容将被过滤。与**基于词语的过滤**不同，这并不仅仅是字符串匹配，还包括同义词、语义相似词，或内容涉及用户所输入的更广泛主题。选择此功能作为初始输入，是因为先前研究已探讨其减少接触有害内容的潜力 [65]。然而，主题过滤器在不同屏幕之间的重要性差异仍是一个开放性问题。第三个功能是**自定义对检测到的仇恨言论的响应**。用户可以决定如何处理检测到的仇恨言论：将其移除、模糊化、隐藏在警告之后，或移入“稍后审核”文件夹。选择将此功能作为初始输入的依据是先前的研究，这些研究指出，用户的错失恐惧既会对其个人审核评价产生负面影响 [39, 46]，甚至也会对基于账户的个人审核工具的采用产生负面影响 [43]。随后，我们鼓励参与者通过自由文本输入为每个屏幕提出额外功能。

作为德尔菲研究的核心环节，第2轮的流程与第1轮相同，但做了三项关键调整。首先，我们聚焦于获得多数参与者支持的屏幕和功能（≥ 50%的参与者对该屏幕或功能给出的李克特评分为 ≥ 4）。其次，我们向参与者展示了第1轮中各屏幕的重要性评分中位数，以反馈群体的集体观点，促进反思 [21]。第三，我们将参与者在自由文本回答中的建议纳入其中，扩充了个人审核功能列表。这一设计鼓励参与者进行反思、重新评估最初的意见，并结合来自同行反馈的新过滤思路展开交流。

为生成个人审核功能的排名，第3轮重点对个人审核功能进行排名。我们仅纳入第2轮中获得多数参与者支持的屏幕和功能（有 ≥ 50% 的参与者给出了李克特评分 ≥ 4）。我们通过一项点数分配任务编制排名 [19]。例如，对于 所有评论 屏幕，我们展示了 14 项功能，参与者可以分配 14 分。因此，他们可以将所有分数都分配给一项功能，也可以将分数平均分配给所有功能，或采用介于两者之间的任何分配方式。这种方法促使参与者权衡不同功能之间的取舍，确保最终排名反映的是相对优先级，而非孤立的重要性判断。我们将分数换算为百分比，以便进行比较，并根据百分比对功能进行排名。

3.3 招募与样本

我们招募了经常使用社交媒体且曾在网上遇到仇恨言论的活动人士。我们将“活动人士”界定为积极参与社会运动组织的个人。要参与本研究，参与者必须：(1) 参与某项社会运动；(2) 在过去一个月中，至少两次专门为参与相关运动的活动而使用社交媒体；以及 (3) 曾在网上遭遇仇恨言论。

我们采用了多步骤招募策略。首先，我们根据对德国三座中大型城市中所有社会运动的系统性收集结果，通过电子邮件联系了相关社会运动和组织。这些城市以公民参与活跃、政治格局多元以及基层网络成熟为特点。我们还通过此前与社会运动合作时建立的联系招募参与者。此外，我们还采用了滚雪球抽样。

我们观察到以下参与者流失情况：58 名参与者完成了第1轮，47 名完成了第2轮，40 名完成了第3轮。虽然每一轮所有参与者的自由文本输入都用于为后续波次的功能建议提供依据并扩充建议集合，但评分和排名的分析仅限于完成全部三轮的 40 名参与者。此决定旨在保持各波次面板的一致性，确保观察到的评分差异反映的是参与者随时间发生的内部变化，而非面板构成的变化。因此，以下样本描述仅涉及那些……的参与者

表1：三轮德尔菲研究程序概览。第一轮中呈现的所有屏幕均在表2中说明

第1轮（第一波）	第2轮（第二波）	第3轮（第三波）
参与者完成了初始调查问卷： (1) 参与者阅读了该情境。 (2) 参与者**评估了重要性** 对所有**屏幕** 进行评分，并解释了评分理由。 (3) 对所有被评为重要的屏幕： (a) 参与者**评估了重要性** **3项功能** （来自文献）。 （b）参与者 **新增了功能请求** 通过自由文本输入提出。	参与者遵循了与第一波相同的流程 但有以下区别： （1）所有与屏幕相关的问题 搜索结果 **已被删除**。（ 屏幕未获得多数 第一波中的投票）。 （2）参与者还**对更多功能请** **求进行了评分**（新增了37项功能 ，依据第一波的自由文本输入）。 （3）所有**来自第一波的中位数评分**均 显示在已有项目旁边 W1.	参与者遵循了相同的流程 与第二波相同，但： (1) 参与者并未对功能进行评分， 他们**排名**了这些功能。 (2) 参与者还**对更多功能进行了排名-** **功能请求** （新增了 6 项功能 基于第二波中的自由文本输入）。 （3）所有**第二波的中位数评分**均 显示在已存在的项目旁边 W2.

表2：本表概述了所有屏幕，包括其描述以及分析中使用的标签。表格按照这些屏幕呈现给参与者的顺序排列。在三轮中，参与者均使用 5点李克特量表，对每个屏幕上个人审核的重要性进行评分。

屏幕标签	说明
主页信息流	当用户滚动浏览主页信息流时
探索选项卡	探索选项卡中所有可见的内容
所有评论	所有类型内容下的所有评论
特定帖子的评论	所有单条帖子的评论
自己帖子的评论	所有针对其自身内容的评论
所有私信	所有发给相应用户的私信
来自特定账户的私信	个人用户向相应用户发送的所有私信
搜索结果	基于搜索条目显示的所有内容
Reels 标签页	Reels 标签页上可见的所有内容

完成了全部三个波次。参与者的平均年龄为32.3岁（标准差 = 9.5），年龄最小的参与者为20岁，年龄最大的为63岁。样本由16名女性、12名男性和11名非二元性别者组成。另有一名参与者未透露其性别。值得注意的是，与类似研究相比，非二元性别参与者的比例相对较高 [46],，这可能是因为我们在关注 LGBTQIA+行动主义的社会运动中招募了参与者。这一代表性为实证研究中常被忽视的群体提供了宝贵见解 [69]。与以往研究的做法一致 [26, 49]，该调查小组涵盖了多元的行动主义背景，使我们能够了解行动主义群体的不同观点。参与者可以选择多个词语来描述自己的行动主义，也可以通过自由文本输入描述自己的政治行动主义。样本包括从事左翼政治行动主义（14人）、LGBTQIA+ 行动主义（11人）、气候与环境正义行动（8人）、女权主义与妇女权利（5人）、族裔与宗教少数群体倡导（5人）、反法西斯主义（4人）、难民倡导（2人）、反能力歧视（1人）和数字安全（1人）的参与者。

参与者每完成一次研究波次，即可获得 20 欧元（22.60 美元）；每次平均耗时不到 30 分钟（SD = 5.3）。

3.4 分析

我们检查了完成全部三个波次的参与者的所有回答，确认回答完整，并检查是否存在直线作答模式。我们还考察了参与者是否提供了能表明他们理解问题要求的自由文本输入，例如提供与问题相关的说明或示例。所有回答均符合这些标准，因此无需剔除任何案例。

我们按照 Braun & Clarke 的反思性主题分析方法，对自由文本回答进行了分析 [11]。此举旨在为第2轮和第3轮提出新的功能建议（研究问题二）。第一作者反复阅读数据，并通过不断梳理内容来熟悉数据。第一作者采用开放式编码流程，从数据中提炼初始编码。随后进行了归纳式编码和轴心编码 [20]。第一作者与第二作者在六次修订会议中共同讨论编码的制定，直至解决所有分歧。通过反复聚类、完善和合并编码，我们确定了代表新颖筛选选项的类别（研究问题二）。例如，我们将代码 多媒体筛选器 拆分为子代码，例如基于图像内容的过滤 或基于音频内容的过滤，同时将

代码 过滤攻击性语气 和过滤居高临下的语气 合并为基于语气的过滤。

我们将定量数据表示为将屏幕（研究问题1）或功能（研究问题二）评为重要的参与者百分比（李克特量表评分 ≥ 4）。为了对第3轮的具体功能建议进行排序，我们分析了功能请求的排名，排名依据是各项功能获得的分数百分比。本文重点讨论获得全部分数中超过 5% 的功能。

4 结果

4.1 研究问题1：哪些屏幕应为遭受仇恨言论攻击的活动人士提供个人审核?

为了深入了解遭受仇恨言论攻击的行动主义用户希望在哪些屏幕上进行个人审核，我们请参与者评估在各种屏幕上进行个人审核的重要性。十名参与者中有七名认为除搜索结果之外的所有屏幕都很重要（李克特量表评分 ≥ 4）。该屏幕是唯一一个未被大多数参与者评为重要的屏幕，因此随后从第2轮和第3轮中移除。所有波次中各屏幕的全部评分见附录（表8）。下文将介绍第3轮的结果。

绝大多数（至少十分之八）参与者将所有评论 （100%）、自己帖子的评论 （92.5%）、主页信息流 （87.5%）、*Reels* 标签页 （87.5%）和所有私信 （85.0%）评为重要。在这五个屏幕中，许多参与者特别强调，描述对话空间的三个屏幕对个人审核尤为重要：至少六成参与者将所有评论 （77.5%）、自己帖子的评论 （75.0%）和所有私信 （67.5%）评为高度重要。不到四成的参与者将描述算法策划空间的两个屏幕，即*Reels* 标签页 （27.5%）和主页信息流 （37.5%）评为高度重要。

对于所有评论、自己帖子上的评论 和所有私信这些屏幕，参与者将评论中高度暴露于仇恨言论描述为在这些屏幕上进行个人审核的关键动机（例如 P2、P17、P21、P25 和 P39）。与此同时，每个屏幕也呈现出不同的动机：对于所有评论，P16 强调无法控制“我在评论中看到谁”。P16 将此与主页信息流直接对比，他们在主页信息流中使用“关注”标签页，以确保只显示来自已关注账号的内容。评论则缺乏这种控制。对于自己帖子上的评论，P31 指出该屏幕“直接影响 *[他们的]* 心理空间和安全感”，这一担忧也为若干参与者所共有（例如 P25、P32 和 P36）。

参与者还指出，在算法策划界面 *Reels* 标签页 和 主页信息流上进行个人审核，动机各不相同。参与者强调，主页信息流是社交媒体应用中所有互动的入口，因此是进行个人审核尤为重要的屏幕（例如，P31、P32 和 P39）。对于*Reels* 标签页，参与者强调个人审核的重要性，因为 Reels“会自动播放，且不会预览短视频的内容”（P27）。此外，参与者还报告称，在*Reels* 标签页 上花费大量时间，也是该屏幕获得高重要性评分的原因之一（例如，P27、P32，和 P38）。对*Reels* 标签页 上显示的内容类型缺乏控制，也是 *Reels* 标签页 需要进行个人审核的重要原因（P39）。

有三个屏幕被不到80%的参与者评为重要：特定帖子的评论（77.5%）、特定账户的私信（75.0%）以及探索选项卡（72.5%）。P11解释说，他们将特定帖子的评论和特定账户的私信的个人审核重要性评得较低，因为他们认为这对“进行筛选的人来说相当复杂。”其他人表示，这种控制程度过于精细，他们更希望对所有评论和所有私信采用相同的个人审核（例如P7、P11、P17和P37）。对于屏幕特定账户的私信，P31表示“让 *[被]* 仇恨言论或骚扰针对的风险更高，且这种风险更多来自特定个人或特定类型的账户 （例如陌生人、重复违规者、公开回复喷子）”，其他人则强调他们“无法提前选择特定账户”（P7和P15）。对于探索选项卡，参与者（P1、P3和P25）解释说，他们很少使用这个屏幕，尤其是“对于政治账户”（P25），这降低了该屏幕上的个人审核对他们的重要性。

4.2 研究问题二：遭受仇恨言论针对的活动人士按屏幕请求了哪些个人审核功能?

我们考察与不同屏幕相关的功能请求.

参与者分别针对每个屏幕指出了他们对个人审核的功能请求。我们报告这些功能在第1轮（第一波）和第2轮（第二波）的重要性评分，以及在第3轮（第三波）的最终排名。各屏幕按其感知重要性排序呈现（参见第4.1节，以及图2中这五个屏幕的可视化呈现）。下文未讨论的屏幕的功能请求、评分和排名见附录（表14–16）。

4.2.1 屏幕：所有评论。 本节呈现参与者请求明确应用于屏幕所有评论 的个人审核功能（该屏幕的可视化呈现见图2a，功能说明见表 3）。屏幕所有评论的所有功能请求、其在第1轮和第2轮中的相应评分，以及在第3轮中的最终排名，见附录（表9）。

对于界面所有评论，允许监督和可逆性的基于内容的功能在评分和排名上均较高。例如，功能自定义对检测到的仇恨言论的响应排名第一（第一波 93.9%，第二波 92.3%）（功能说明见表 3）。另一项允许可逆性的功能请求是功能更改过滤决策，排名第三。超过十分之九的参与者认为该功能重要（于第二波新增，第二波 92.3%）。

依赖用户输入的基于内容的个人审核功能也经常被要求用于<接近标签>所有评论</接近标签>界面。例如，<接近标签>所有评论 </接近标签>界面中第二重要的功能是<接近标签>基于词语的过滤</接近标签>。超过九成的参与者认为该功能很重要（第一波为 90.9%，第二波为 94.9%）。另一项基于用户输入的功能请求是<接近标签>基于主题的过滤</接近标签>，其排名第六（第一波为 66.7%，第二波为 92.3%）。

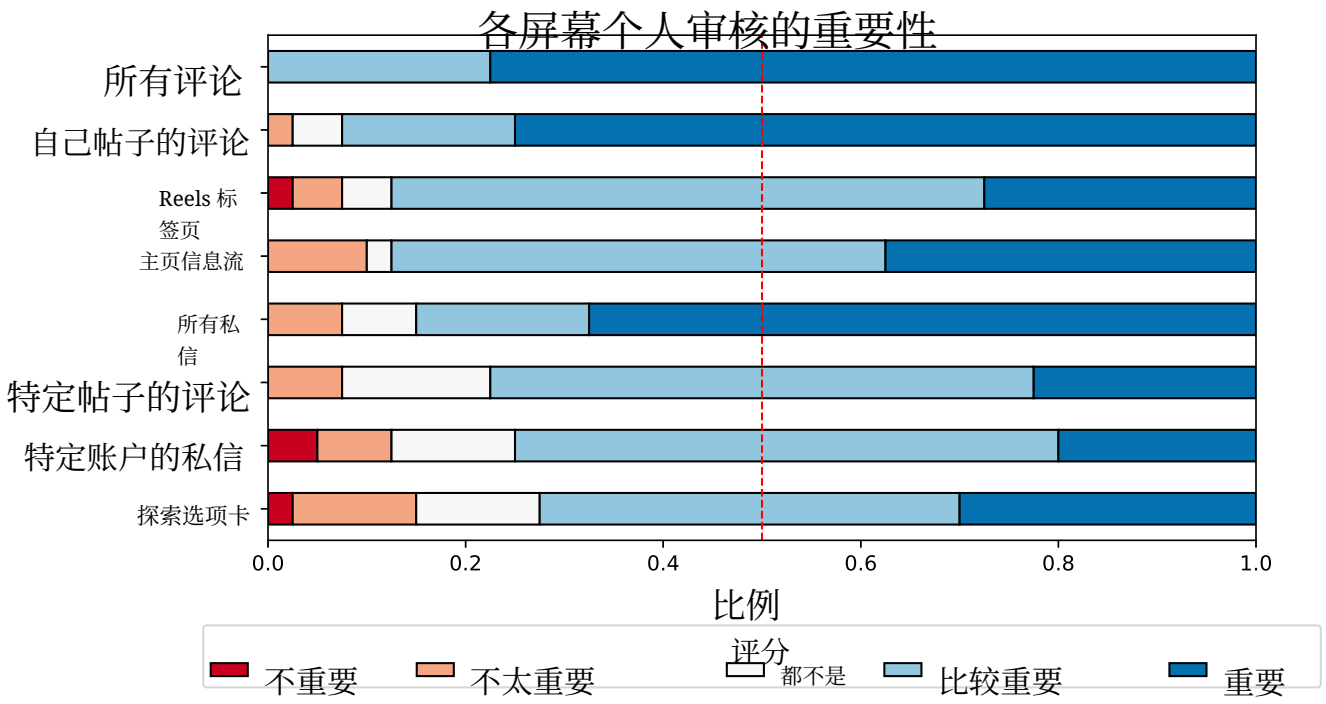


图1：在第3轮中，40名曾遭受针对自身的仇恨言论的社会媒体活动人士对所有屏幕的重要性评分分布。垂直红色虚线表示50%分界线，用以显示哪些评分得到了大多数参与者的认同。各屏幕按将其评为重要的参与者比例排序。

表 3：针对<接近标签>所有评论</接近标签>界面的功能请求，基于 40 名曾遭受针对自身的仇恨言论的活动人士社交媒体用户的输入。该表列出了第 3 轮的最终排名、功能名称，以及对功能运作方式的说明。请求按其第 3 轮排名排序。

Rank	**功能请求** 所有评论	**说明**
1.	自定义对检测到的仇恨言论的响应	选择删除、模糊处理、通过内容警告隐藏，或移至“稍后审核”文件夹。
2.	基于词语的过滤	移除所有评论中包含用户指定字符串的评论。
3.	更改过滤决策	可以根据需要查看并修改过滤决策。
4.	隐藏评论区	可以切换所有评论区的显示或隐藏。
5.	基于语气的过滤	移除所有评论中带有攻击性、侮辱性或居高临下的语气的内容。
6.	基于主题的筛选	移除所有评论中与用户指定主题相关的内容。
7.	基于用户的过滤	移除所有评论中来自用户指定的用户或用户群组的内容。
8.	基于沟通模式的过滤	移除所有评论中符合特定沟通模式的内容。

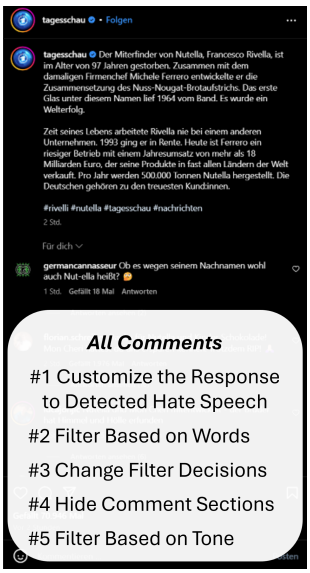
此外，参与者也提出了依赖先进仇恨言论检测机制的基于内容的功能，但这些功能的排名较低，例如基于语气的过滤排名第五。在第2轮中，超过七成的参与者认为这项功能很重要（71.8%）。类似地，功能请求基于沟通模式的过滤在第3轮中排名最后。参与者认为这项建议不太重要，只有略多于五成的参与者认为它很重要（第2轮新增，56.4%）。

最后，受访者还要求提供基于账户的个人内容管理功能：第3轮中排名第七的重要功能是基于用户的过滤（于第二波新增，第二波为 82.1%）。

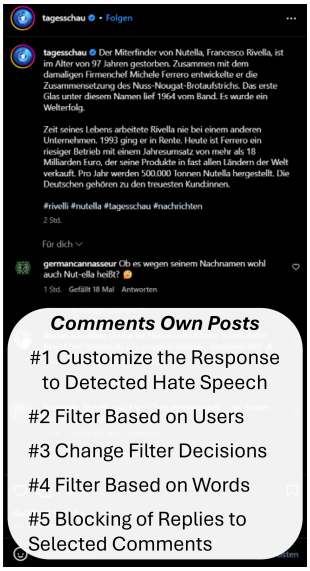
总体而言，这些发现凸显了监督、可逆性以及输入驱动功能对于所有评论个人审核的重要性。

4.2.2 屏幕：自己帖子上的评论。本节概述参与者认为适用于自己帖子上的评论屏幕的个人内容审核功能（该屏幕的示意图见图 2b，功能说明见表 4）。

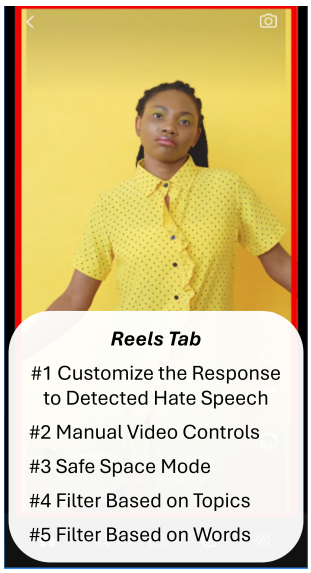
能够实现监督和可逆性的基于内容的功能在屏幕自己帖子的评论上排名最高。例如，自定义对检测到的仇恨言论的响应 被排名



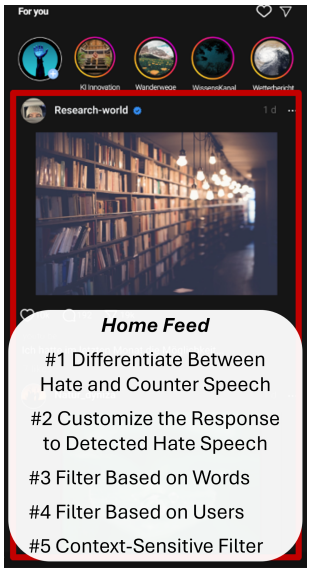
(a) 所有评论



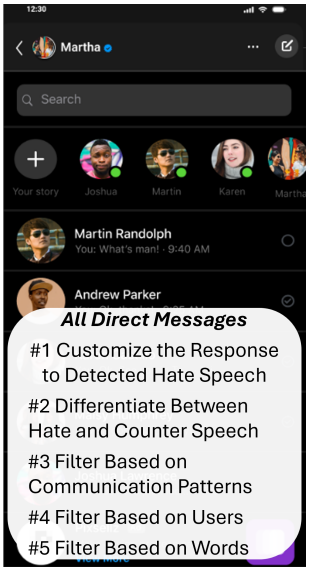
(b)自己帖子上的评论



(c) *Reels* 标签页



(d) 主页信息流



(e)所有私信

图2：此图展示了向参与者展示的屏幕示例图片：所有评论（**2a**）、自己帖子上的评论（**2b**）、***Reels*** 标签页（**2c**）、主页信息流（**2d**）和所有私信（**2e**）。**此外，还展示了各屏幕对应的排名最高的五项个人审核功能，并按照参与者排名排列。**

在第3轮中被参与者排在首位（94.6% 第一波，94.7% 第二波）。此外，参与者在第2轮通过自由文本输入要求能够更改过滤决策。该功能在第3轮排名第三，超过十分之九的参与者（92.1%）将其评为重要。

此外，参与者对账户特定功能的请求及其排名各不相同：基于用户的过滤（在第二波中新增，816 2. % W ）排名第二，而列出受信任的评论者（在第三波中新增）排名第九。

对于屏幕自己的帖子中的评论，参与者提出了多项基于输入的功能请求。排名第四的重要功能请求是基于词语的过滤。在第1轮，超过十分之八的参与者认为这项功能很重要（89.2%）；在第2轮，这一比例超过十分之九（93.3%）。参与者还建议增加自动屏蔽对选定评论的回复，该功能排名第五（第2轮新增，第二波为76.3%）；或者手动批准评论，该功能排名第六（第2轮新增，第二波为71.1%）。此外，参与者还建议增加基于帖子的筛选设置，该功能排名第七（第二波为76.3%）。P31

表 4：针对屏幕自己帖子上的评论的所有功能请求，依据来自40名曾遭受针对性仇恨言论的社会媒体活动人士的意见。该表列出了第3轮的最终排名、功能名称，以及对功能运作方式的说明。请求按照其第3轮排名排序。

Rank	功能请求 自己帖子上的评论	说明
1.	自定义对检测到的仇恨言论的响应	选择删除、模糊处理、隐藏在内容警告后，或移至“稍后审核”文件夹。
2.	基于用户的过滤	从用户指定的用户或用户群组中移除自己帖子上的评论用户。
3.	更改过滤决策	必要时可以审核和修改过滤决策。
4.	基于词语的过滤	移除包含用户指定字符串的自己帖子上的评论。
5.	屏蔽对选定评论的回复	自动屏蔽对自己帖子上的特定评论的回复，这些回复由用户手动选定。
6.	手动允许评论	允许用户审核并批准评论，然后再让它们显示在他们的帖子下方。
7.	基于帖子的筛选设置	允许用户在发布单条自己的帖子之前设置特定的筛选预设在发布之前。
8.	基于主题的过滤	移除其自有帖子中所有与指定主题相关的评论由用户指定。
9.	列出受信任的评论者	仅允许获批用户或关注者在其帖子下发表评论，不受限制地，而其他人的评论则会被过滤。

解释说，这项功能请求是指“发布帖子时选择一个筛选预设”，之后该预设将应用于该帖子的所有评论。最后，基于主题的过滤排名第八（第1轮为64.9%，第2轮为81.6%）。

允许自定义响应、按用户屏蔽以及修订过滤决策的功能，在评论自己的帖子屏幕的个人审核功能中排名最高。这表明，参与者优先考虑灵活的控制机制，以管理自己帖子下的评论。

4.2.3 屏幕：*Reels* 标签页。本节概述了参与者认为适用于 *Reels* 标签页屏幕的个人审核功能（该屏幕的示意图见图 2c，功能说明见表 5）。

对于*Reels* 标签页屏幕，可逆性和监督相关功能排名最高。自定义对检测到的仇恨言论的响应被评为最重要的功能（81 1, 91 8 9 2 . % W . % W），排名第二的功能是手动视频控制（第二波新增，88.9%）。P31进一步说明了对该功能的需求，强调除了视频控制之外，还应在“整个屏幕上显示主题、标题和触发警告 *[...]* ”。这样一来，用户便可以在充分了解情况后，决定是否观看该 Reel（P25）。

参与者还提出了一些需要高度自动化的功能。例如，参与者请求添加安全空间模式，该功能排名第三（在第三波中添加）。该功能设想了一种预先配置的环境，具有“最大化的过滤设置”（P8）。参与者还建议在*Reels* 标签页 中添加基于 *Reel* 评论的筛选，该功能排名第八（在第三波中添加）。这项个人审核功能会“显示覆盖整个屏幕的内容警告 *[ ]* 如果 *Reel* 下方的评论包含攻击性或仇恨情绪——即使尚未播放视频也是如此”（P31）。

有两项基于输入的功能被提出：基于主题的过滤 和基于词语的过滤。前者被评为第四重要的功能。近八成参与者（78.8%）在第1轮将此功能评为重要，几乎所有参与者在第2轮也如此（94.6%）。后者在第3轮排名第五。近七成参与者在第1轮将其评为重要（69.7%），支持在第2轮升至九成以上（91.9%）。

其他基于内容的功能请求则将过滤扩展到了不同的内容类型。基于音频内容的过滤 排名第六（在第二波中添加；第二波占比为85.7%）。基于 *Reel* 格式的过滤 排名第九（在第二波中添加；第二波占比为74.3%），指按视频风格或制作类型进行过滤（例如直播或拼接视频）。最后，参与者还请求了一项针对特定账户的功能，但排名较低：基于用户的过滤 排名第七（在第二波中添加；第二波占比为75.7%）。

总体而言，支持自定义系统响应、手动视频控制以及“安全空间”环境的个人审核功能，在屏幕*Reels* 标签页上排名最高。

4.2.4 屏幕：主页信息流。 本节介绍参与者针对屏幕主页信息流 所请求的个人审核功能（该屏幕的可视化呈现见图 2d，这些功能的说明见表 6）。

对于屏幕主页信息流，被请求的是高度依赖自动化、且只需极少用户配置的基于内容的功能。例如，区分仇恨言论与反仇恨言论 排名第一（在第二波中新增，第二波为 91.2%）。P13 举例说明：“一条帖子 *[其]* 明显是反犹太的，或仅仅是在讨论反犹太主义。”P14 说：“为什么允许那个人仇恨，却不允许我为人们辩护？”。此外，上下文敏感过滤 也被提出，并排名第五（在第二波中新增，第二波为 67.7%）。

唯一一项兼顾可逆性与监督功能、适用于主页信息流 屏幕的建议功能，是基于内容的功能自定义对检测到的仇恨言论的响应，排名第二（第一波为 96.3%，第二波为 94.1%）。

表 5：针对屏幕*Reels* 标签页的所有功能请求，基于遭受仇恨言论攻击的40名社会媒体活动人士的输入。该表提供第3轮的最终排名、功能名称以及该功能如何运作的说明。请求按其第3轮排名排序。

Rank	功能请求*Reels* 标签页	说明
1.	自定义对检测到的仇恨言论的响应	在移除、模糊处理和隐藏在内容警告之后之间进行选择，或移至“稍后审核”文件夹。
2.	手动视频控制	允许用户控制短视频 何时播放，并在播放前显示主题、标题和播放前的触发警告。
3.	安全空间模式	提供预先配置的“最大过滤”模式（P8），用于短视频。
4.	基于主题的过滤	移除与用户指定主题相关的所有短视频 。
5.	基于词语的过滤	移除包含用户指定字符串的所有短视频 。
6.	基于音频内容的过滤	将包含用户指定音频元素的所有短视频 静音或移除。
7.	基于用户的过滤	移除来自用户指定用户或用户群组的所有短视频 。
8.	基于 Reel 评论的筛选	如果短视频 的评论包含攻击性或仇恨情绪，则将其移除。
9.	基于*Reel* 的格式进行筛选	根据短视频 的格式（例如直播或拼接视频）将其移除。

表 6：针对主页信息流屏幕的所有功能请求，依据 40名社会媒体活动人士（曾遭受针对他们的仇恨言论）的反馈整理而成。该表列出了第3轮的最终排名、功能名称，以及该功能运作方式的说明。各项请求按照其第3轮排名排序。

Rank	功能请求 主页信息流	说明
1.	区分仇恨言论与反仇恨言论	可靠地区分有害的仇恨言论与正当的主页信息流中的反仇恨言论。
2.	自定义对检测到的仇恨言论的响应	选择移除、模糊处理、隐藏在内容警告后，或移至“稍后审核”文件夹。
3.	基于词语的过滤	移除包含用户指定字符串的主页信息流中的所有内容。
4.	基于用户的过滤	移除主页信息流中的所有内容，这些内容由特定用户生成或用户群组。
5.	上下文敏感过滤	根据时间或位置等上下文因素自动调整过滤决策。等因素。
6.	基于主题的过滤	移除与指定主题相关的主页信息流中的所有内容用户。
7.	基于图像内容的过滤	移除包含用户指定图像的主页信息流中的所有内容元素。
8.	基于视频内容的过滤	移除包含用户指定视频的主页信息流中的所有内容元素。
9.	基于音频内容的过滤	移除主页信息流中的所有内容 中包含用户指定音频的内容元素。
10.	基于示例的过滤	允许用户通过提供样本内容来训练系统，以优化过滤行为。

参与者提出了多项输入驱动功能，例如基于词语的过滤，排名第三（于第二波新增，第二波中占 97.1%）。此外，参与者还提出了基于主题的过滤 （第二波中占 81.5%，第二波中占 91.2%，第三波排名第六）和基于示例的过滤 （于第二波新增，第二波中占 67%，第三波排名第十）。

有一项针对账户的功能被提出，即基于用户的过滤。它在第3轮中排名第四（在第二波中加入，比例为88.2%）。参与者还建议根据特定内容类型进行过滤，不过这些功能的排名较低。基于图像内容的过滤排名第七（在第二波中加入，第二波中比例为76.5%）。基于视频的过滤内容排名第八（在第二波中加入，第二波中比例为73.5%）。类似地，基于音频内容的过滤排名第九（在第二波中加入，第二波中比例为67.7%）。

总体而言，参与者强调了精细过滤的重要性（例如，区分仇恨言论与反仇恨言论）、针对仇恨言论处理方式的自定义选项，以及可靠的关键词过滤或基于用户的屏蔽，同时对特定模态功能的优先级较低。

4.2.5 屏幕：所有私信。本节概述参与者认为与所有私信屏幕相关的个人内容审核功能（该屏幕的视觉呈现见图 2e，功能说明见表 7）。

表 7：所有私信屏幕的所有功能请求，基于对 40 名遭受仇恨言论攻击的活动人士社交媒体用户的调查。该表列出了第3轮的最终排名、功能名称，以及对功能工作方式的说明。请求按其第3轮排名排序。

Rank	功能请求 所有私信	说明
1.	自定义对检测到的仇恨言论的响应	选择移除、模糊处理，或隐藏在内容警告之后或移至“稍后审核”文件夹。
2.	区分仇恨言论与反仇恨言论	可靠地区分仇恨私信与正当的反仇恨言论或教育性讨论。
3.	基于沟通模式的过滤	移除所有私信中呈现特定沟通模式的内容。
4.	基于用户的过滤	移除来自用户指定的用户或用户群组的所有私信。
5.	基于词语的过滤	移除包含用户指定字符串的所有私信。
6.	基于图像内容的过滤	移除包含用户指定图像元素的所有私信。
7.	链接过滤	可以屏蔽所有包含链接的私信。
8.	基于语言的过滤	根据相应私信的语言自动调整过滤。

与其他对话空间一样，参与者希望个人内容审核工具具备基于内容的功能，以实现监督和可逆性。对于所有私信屏幕，参与者将自定义对检测到的仇恨言论的响应排在首位，而且该功能在两个波次中都获得了较高的重要性评分（第一波为 96.9%，第二波为 97.1%）。

参与者还要求提供基于内容的功能，这类功能高度依赖自动化，且只需用户进行最低限度的配置。例如，参与者希望个人内容审核工具能够可靠地区分仇恨言论与反仇恨言论。该功能排名第二（在第二波新增，第二波中占91.2%）。这一请求反映了一项关键关切：参与者希望有效过滤有害内容，同时也希望确保有价值的反仇恨言论——例如质疑、重新解读或揭露仇恨言论的回应——仍能被看到。另一个例子是要求基于沟通模式的过滤，该功能在第3轮中排名第三（在第二波新增，第二波中占76.5%）。参与者认为，识别重复出现的骚扰模式很有价值，例如协同针对或持续不断的不受欢迎联系。

参与者还要求一项针对特定账户的功能：基于用户的过滤。该功能被评为重要性第四的功能（在第二波新增，第二波中占85.3%）。由用户输入驱动的功能基于词语的过滤 排名第五，且其重要性在各波次中有所上升（第一波为78.1%，第二波为94.1%）。

内容类型和格式特定选项，即基于图像内容的过滤 （在 W3 中新增，在 W3 中排名第 6）、链接过滤 （在第二波中新增，占 73.5%，在 W3 中排名第 7）和基于语言的过滤 （在 W3 中新增，在 W3 中排名第 8），排名最低。参与者认为，这些选项对于所有私信屏幕的重要性较低。

总体而言，参与者强调，灵活的回应选项、准确区分仇恨言论与反仇恨言论，以及基于通信模式的过滤器，是对私信进行个人审核时最重要的工具。

5 讨论

我们展示了用户希望在哪些方面进行个人审核（研究问题1），以及他们最看重哪些功能（研究问题二）。在本讨论中，我们将这些发现置于现有文献的背景下，并阐明我们的结果如何为个人内容审核中的控制粒度提供一种概念化框架。此外，我们展示了如何对用户提出的功能进行分类，并将其与不同屏幕联系起来。最后，我们为社交媒体上的个人审核工具提出设计建议。

5.1 个人审核的缘由

5.1.1 优先处理屏幕，以实现有效的个人审核。 我们的研究结果表明，不同屏幕对个人审核的需求各不相同。通过确定用户认为最关键的屏幕优先级，我们为探索个人审核的研究人员、开发用户界面的设计师以及分配资源的平台提供了可行的指导。这种优先排序降低了实施复杂度，并确保相关工作聚焦于个人审核影响最大的情境。

另一项贡献是将屏幕确定为个人审核的一个有意义的控制单元，这一问题此前已有研究提出过 [46]。不同屏幕上的重要性分布各不相同，且其中一部分屏幕的评分较高，这表明这种概念化方式具有价值。这种细化至关重要，因为有意义的设计指导不仅需要了解用户重视哪些功能，还要了解他们希望在哪里应用这些功能。未来研究应深入探讨“屏幕”这一与个人审核相关的维度。因此，平台应调整个人审核功能，使其能够分别应用于不同屏幕。提供屏幕特定设置，既能扩展现有功能，也能契合本研究所揭示的用户优先事项。不过，本研究确定的屏幕及其相关可供性集合仅适用于 Instagram [5, 64]。未来还需开展更多研究，以确定用户对其他平台上基于屏幕的个人审核的偏好。

5.1.2 作为个人审核工具设计约束的劳动与可用性。设计个人审核工具时，若将我们研究中发现的屏幕特定差异纳入考量，就会在可用性和劳动方面引发新的问题。个人审核工具的设计必须确保其易于访问，并且确实会被需要它们的用户使用。这种张力在 Instagram 现有的功能中十分常见，例如主页信息流包含“为你推荐”标签页（对内容进行算法策展）和“关注中”标签页（已关注账号的按时间顺序的信息流）。尽管

先前研究记录了用户明确希望使用按时间顺序的信息流 [49, 63], , 但用户通常并不了解这一可能性 [49]。当设计一款涵盖本文所讨论的所有屏幕及全部功能的个人内容审核工具时，这一问题可能会进一步加剧，因为这要求实现方案应对更为复杂的情形。为应对这一挑战，我们强调本研究所提供的排名的价值。这有助于聚焦排名最高的屏幕和功能，从而帮助开发者和研究人员确定工作优先级。这种优先级排序还有望在配置这些工具时减少用户负担，而这一问题已被先前关于个人审核的研究频繁指出 [39, 46]。

当我们询问参与者对评分较低的屏幕的看法时，他们强调了配置个人内容审核设置所需付出的劳动；而对于评分较高的屏幕，他们并未提及这一点。这或许表明，尽管参与者意识到其中需要付出劳动，但对于他们认为个人内容审核影响很大的屏幕，他们仍愿意投入精力。尽管如此，即使只针对部分屏幕配置个人内容审核，也仍然需要用户了解并理解个人内容审核的可能性，并投入（至少是初期的）时间。先前的研究已指出，缺乏认知和必要的时间资源都会阻碍功能的采用，例如隐私相关功能的采用 [17]。综合来看，这些因素凸显了劳动是个人审核工具的一项核心设计约束。因此，未来的研究应探讨在配置个人内容审核时，如何平衡可控性与可用性。较高的可控性使用户能够按需调整审核方式，但过多的配置工作可能会带来认知负担 [46]。我们建议设计出易于发现且清晰可见、同时不会对较少受到仇恨言论影响的用户造成干扰的界面和个人审核工具。研究应探讨渐进式披露策略，作为解决这一问题的潜在途径。此外，提供有意义的默认设置也至关重要；同时，还应以确保用户对其功能形成共同理解的方式设计个人审核工具。先前研究也指出了这些问题 [46]。

5.1.3 对话界面中的个人审核。 我们的研究结果凸显了在评论区和私信等对话空间中进行个人审核的重要性。既有研究表明，对话场景给内容审核带来了挑战，因为含义取决于语境，社群间的规范各不相同，而人际交往历史也会影响解读 [25, 66]。我们的研究显示，对用户而言，这些空间恰恰是最需要个人审核功能的地方，因为用户往往清楚哪些词语、主题或个人可能造成伤害。这凸显了个人审核的独特作用：它并非要与自动化系统或平台级系统竞争，而是能让用户在自动化仇恨言论检测存在局限的人际交往情境中掌握主动权 [37]。由于这些发现基于 Instagram 的对话空间，因此能否推广至其他平台尚属假设。不过，我们认为，结构相似的平台或许同样能从中受益。即时消息、评论串和内容审核实践在不同平台间各有差异，社区规范也不尽相同；这些因素都会影响用户接触有害内容的情况，以及他们对个人审核实用性的感知。未来研究应探讨个人审核需求如何从对话情境延伸至其他平台。

5.1.4 算法策划界面上的个人审核。我们的结果表明，在主页信息流和 Reels 标签页等算法策划界面上，个人审核同样被认为十分重要。这一发现与对算法不透明性和平台治理中权力不对称的批评相一致 [27, 34]。然而，关于算法策划空间中用户控制的研究，大多从改进推荐系统的角度探讨这一问题 [36, 81] ，而非从内容审核的角度加以审视。

然而，内容审核与内容策展密不可分：当用户不想看到某些内容时，将其过滤掉与仅仅不予推荐之间的界限便会变得模糊。未被推荐的内容实际上可能会让人觉得是“被审核掉了” [46]。鉴于推荐并非中立过程，而是受商业逻辑影响、优先降低广告商风险而非解决用户伤害，这一点尤为重要 [56]。我们的研究结果表明，个人审核或许能够赋予用户更大的自主权，使其不仅能决定向自己突出显示哪些内容，也能决定应隐藏哪些内容，从而重新平衡这种不对称，并以个人审核补充平台主导的内容策展。然而，我们的研究结果也受到 Instagram 界面架构及其相关可供性的影响 [5, 64]。不同平台的信息流设计、排序算法和透明度各不相同，这可能会影响个人审核在其他平台算法策划界面上的实用性。未来还需开展更多研究，以确定哪些研究洞见具有普遍性，哪些仅适用于特定平台。

5.2 个人审核的内容

接下来，我们将对所有讨论过的屏幕中排名最高的五项功能请求进行分类。然后，我们会说明这些分类与提出相应请求的屏幕之间的关联。

5.2.1 监督与可逆性：个人审核的核心支柱。这一类别监督与可逆性功能基于先前的研究，描述了能够让用户监督内容审核结果并根据自身需求进行调整的基于内容的功能 [39]。我们研究中提出的监督与可逆性功能包括自定义对检测到的仇恨言论的响应、更改过滤决策以及手动允许评论。

在这些功能中，自定义对检测到的仇恨言论的响应在我们的研究中尤为重要。这一功能的重要性表明，它或许能够解决此前发现的个人审核相关问题。其中一个常见问题是担心错过内容 [39, 43]。允许用户查看个人内容审核工具的过滤决策后，用户便不必担心会错过内容，因为他们随时都能在“待审核”文件夹中查看所有内容。此外，这类功能还可以解决内容警示和触发警告因无法根据个人需求、具体情境或脆弱性进行充分定制而效果有限的问题 [16, 78] 。

大多数监督和可逆性功能都是针对所有评论 和自己帖子上的评论屏幕提出的。参与者在这些屏幕上遭遇内容审核系统故障的经历，或许可以解释他们为何更需要能够控制个人内容审核工具的决定，并在必要时撤销这些决定。

屏幕，或许可以解释这种需求为何有所增加：用户需要能够控制个人内容审核工具的决定，并在必要时撤销这些决定。

因此，我们建议在个人内容审核工具中实现基于内容的监督和可逆性功能。设计人员应探索响应自定义选项，例如选择移除、模糊处理、添加内容警告，或将内容移至单独的文件夹以供日后查看。若能结合修改个人内容审核决策的功能，便可进一步减少挫败感并增强信任 [39]。

总体而言，我们的研究结果凸显了监督和可逆性是跨屏幕有效开展个人审核的核心支柱，在审核评论区时尤为重要。

5.2.2 输入驱动功能在已知仇恨言论细节的情境中最为重要。我们将输入驱动功能定义为依赖用户输入、由用户明确指定希望审核哪些类型内容的个人审核功能。因此，根据 Jhaver 等人（2023）的分类，这也可归为基于内容的个人审核。在我们的研究中，基于词语的过滤、基于主题的过滤和基于示例的过滤均属于输入驱动功能。

词语过滤器尤其已集成到多个社交媒体平台中。例如，Instagram 允许用户过滤评论和私信请求中的词语 [42] ，Facebook 则支持过滤评论中的词语，并提供语言变体和批量上传等选项 [30]。然而，我们的研究结果表明，用户有必要区分不同屏幕上过滤的词语。对平台上的所有内容都过滤相同的词语，无法满足我们所发现的用户需求。这或许能解决此前指出的词语过滤器的缺陷，例如上下文不敏感 [46]。此前的研究也探讨过主题过滤器 [7, 65], , 但平台实现往往着重于呈现用户想看的内容，而非避开有害或不想看到的主题 [7, 36, 81]。我们的研究进一步强调了此前研究所指出的、从内容审核角度看待输入驱动功能的重要性 [46]。

针对评论区（包括所有评论 和自己帖子上的评论）的输入驱动功能，在参与者的排名中尤为靠前。评论相关输入驱动功能的重要性较高，这表明用户认为，这些功能在仇恨言论既常见又可预测的屏幕上尤其有效。相比之下，参与者认为，输入驱动功能对于算法策划界面（如主页信息流 或*Reels* 标签页）不那么重要。一个可能的原因是，在这些屏幕上，用户难以预判会出现什么内容，因此更难根据词语、主题或示例来设定有效的过滤器。这种对话式屏幕与算法型屏幕之间的区分，为输入驱动功能的设计增添了新颖的、针对特定屏幕的视角。

总的来说，输入驱动功能仍是个人审核的基础要素。不过，我们的结果表明，这些功能在对话界面中最为实用。

5.2.3 特定内容类型功能弥补算法信息流和私信中的关键缺口。特定内容类型功能是指根据内容格式（如文本、图像或视频）调整过滤方式的功能。我们研究中的特定内容类型功能包括基于音频、图像或视频内容的过滤、基于 *Reel* 格式的筛选，以及链接过滤。

在我们的研究中，特定内容类型功能在 *Reels* 标签页 和主页信息流等算法策划界面上尤其受到重视。参与者指出，这些信息流常常会向他们展示意料之外的多媒体内容，因此仅过滤文本并不足够。

在所有私信这一没有算法策展的界面中，参与者提出了图像和链接过滤器的需求。这表明，在私密互动中，参与者认为除了文本之外，针对其他内容类型的过滤器也很重要。以往研究表明，人们讨论个人审核时通常会着眼于文本 [65, 75], , 但我们的发现强调，保护用户免受通过各种内容类型传播的仇恨言论侵害，正变得日益重要。

总而言之，特定内容类型功能弥补了现有工具的不足，将个人审核扩展至非文本内容，并根据各类内容分别进行过滤。我们的研究表明，用户希望在算法策划的信息流中使用这些功能来应对不可预测性，并在私信中使用它们来处理针对性的基于图像和链接的骚扰。

5.2.4 账户特定功能至关重要，但仍有待完善。我们将账户特定功能定义为：允许用户围绕社会关系塑造其在线环境，并限制或放宽与个别用户或用户群体互动的功能 [45]。在我们的研究中，账户特定功能包括基于用户的过滤和列出受信任的评论者。

我们发现，账户特定功能在三个屏幕上的排名较高：自己帖子上的评论、主页信息流和所有私信。对于自己帖子上的评论，Instagram 已支持基于账户的个人内容审核。用户可以决定是否允许关注者、互相关注的人或所有人发表评论。用户也可以启用或禁用他人在其内容的 GIF 上发表评论。对于主页信息流，用户可以在“为你推荐”标签页（默认选项）和“关注”标签页之间切换，从而控制自己能看到哪些人的内容。对于所有私信，Instagram 已提供多种控制选项 [42]。鉴于这些选项已在多个平台上实现，这就引出了一个问题：哪些因素会促进或阻碍人们采用这些工具？近期研究显示，心理社会因素会影响基于账户的个人内容管理的采用，具体研究了 Facebook 上的静音功能 [43]。这项静音功能允许用户从动态消息中移除特定账户的内容。一项具有代表性的调查问卷显示，错失恐惧症（FoMO）会显著降低用户使用静音功能的意愿，而主观规范和对 Facebook 内容审核的信任则会正向影响用户对这些工具的使用 [43]。

这表明了平台设计方面的两个重要洞见。首先，现有的账户特定功能深受仇恨言论攻击用户的重视，但需要进一步扩展，以实现更精细的控制。我们的排名为应优先实施哪些功能提供了指导。其次，正如 Jhaver（2025年）所展示的，账户特定功能在减少用户遭受伤害的过程中，其有效性不仅取决于技术实现。这些功能的采用还取决于能否解决包括错失恐惧症（FoMO）在内的心理社会障碍，并培养用户对平台内容审核实践的信任。

5.2.5 在仇恨言论的具体细节未知时，高度自动化功能最为重要。高度自动化功能是指依赖先进仇恨言论检测机制、最大限度减少用户手动输入或配置的基于内容的功能。在我们的研究中，这些功能包括基于语气的过滤、基于交流模式的过滤、安全空间模式、区分仇恨言论与反仇恨言论，以及上下文敏感过滤。

用户在算法策划界面（*Reels* 标签页、主页信息流）上提出了对高度自动化功能的需求，因为这些功能可以在无需详细手动配置的情况下减少意外接触仇恨言论的机会。例如，针对 *Reels* 标签页，参与者要求提供安全空间模式，以移除任何令人痛苦的内容。何种内容令人痛苦因参与者而异。因此，针对不想要的内容训练个性化分类器是一条有前景的路径。这项工作可以借鉴Wang等（2025）的研究，该研究调查了用户在训练个性化内容分类器时的偏好。结果显示，用户在基于规则的训练、基于示例的训练以及向大语言模型发出提示之间存在细致入微的偏好 [75]。要为*Reels* 标签页设计个性化的安全空间模式，还需要对图像、音频和视频数据开展类似研究。

参与者很少要求在评论区提供此类功能。这与此前对自动化审核的批评相呼应：模糊性和不断演变的规范会削弱信任 [44, 46]。我们的研究结果进一步延伸了这些批评，表明尽管参与者欢迎在大多数屏幕上使用自动化，但在对对话界面进行内容审核时，他们对高度自动化仍持谨慎态度。

5.3 情境中的个人审核

尽管我们的研究考察了个人审核的设计空间，但将这些见解置于更广泛的文献背景中可以看出，个人审核只能构成应对仇恨言论这一更大规模社会技术响应的一部分。

一个关键问题是平台责任 [46]。增加个人审核的可能性，不应成为平台逃避责任的手段；对于显然应当在全平台范围内移除的内容，平台有责任将其移除。既有研究一再指出全平台审核的不足 [4, 47, 61], ，而这些不足也不应被期待由个人审核功能来解决。个人审核的目的应是补充全平台审核，以应对审核一刀切方法不可避免带来的局限。然而，要履行这些平台层面的责任，取决于审核流程中蕴含的专业知识。先前研究强调了内容审核流程所需的专业知识 [1]。这与当前行业缩减人工审核劳动的趋势形成了鲜明对比 [28, 29, 53]。然而，只要不会危及投资者利益或违反监管要求，平台的优先事项就是尽可能限制对审核工作的额外资金投入 [33]。

第二个核心问题是，有效的内容审核无法脱离更广泛的问责问题，即如何追究传播仇恨言论者的责任。既有研究表明，对在线传播仇恨言论者的惩处并不一致，而且往往力度有限 [68]。这一缺口涉及多个层面：媒介素养、平台治理和法律结构。暂停账户、发出警告以及社区准则等措施可以起到威慑作用 [31, 83]。此外，民主式审核方法可能有所助益，但也会引发有关正当程序、透明度以及多数主导的惩罚行动风险的争议 [18]。恢复性司法方法等互补方案则提供了潜在途径，通过着眼于修复而非惩罚来应对伤害 [82]。然而，这类模式有赖于对信任与安全基础设施的持续投入，而许多平台如今正在削减这方面的资源 [28, 29]。

此外，还需要更有力的法律框架，以确保违规者承担后果。仅凭个人或平台审核，无法应对反复且严重的仇恨言论违规行为。监管机构可以通过制定升级处理标准，并确保严重案件获得充足的法律资源进行追究，从而提供明确指导。

5.4 局限性与未来工作

尽管本研究提供了宝贵的见解，但研究也受到一些情境因素的影响，因此存在需要承认的局限性。我们的样本由德国的活动人士组成，这意味着他们在其他文化或政治背景下经历仇恨言论的情况可能有所不同。尽管如此，聚焦这一群体使我们得以了解那些特别容易遭遇仇恨言论的个体的观点，因此也能掌握他们对个人审核的充分了解的需求。此外，我们的样本来自意识形态光谱中倾向进步的一侧的运动。这可能影响了受访者表达的观点和优先事项，因为研究表明，与保守派相比，认同自由派或倾向左翼的个体更可能对内容审核持积极态度 [3, 67, 76]。三轮研究中参与度的变化可能造成偏差，使样本中的观点偏向已经支持个人审核方法的用户。还应注意，我们在各轮调查和各位参与者中均按相同顺序展示屏幕。这可能引发顺序效应，例如疲劳、习惯化，或随着时间推移对任务日益熟悉。由于各屏幕上的自由文本输入量相对稳定，因此没有明显迹象表明随着调查进行，参与度有所下降；不过，感知重要性或注意力焦点受到的细微影响仍无法排除。同样，固定顺序也可能影响参与者的相对评分：较早展示的屏幕可能成为锚点，影响对后续屏幕的判断；而较晚展示的屏幕也可能因参与者逐渐了解内容审核功能而受益于学习效应。最后，我们考察的屏幕反映了 Instagram 当前的架构。尽管许多主流平台都将用户互动组织在类似的屏幕中（例如信息流、消息空间、评论串），我们的研究结果能在多大程度上迁移到其他平台，取决于诸多因素，例如这些屏幕在功能、交互流程和社群规范方面的运作方式有多相似。因此，我们的结果应主要结合 Instagram 的背景来解读；不过，以屏幕为基础的抽象方式为评估其他平台上的个人审核需求提供了一个结构化方法，尤其适用于架构包含这些核心组成部分的平台。

基于这些局限性，本研究为未来研究开辟了有前景的方向。我们的结果表明，屏幕是分析个人审核的一个有意义但尚未得到充分探索的单位，未来研究应在不同平台上进一步确立这一分析维度。尤其是，探讨对话式空间与算法策展空间之间个人审核需求差异的研究，为未来研究提供了有前景的方向。特别是，对监督和可逆性功能进行研究与设计，有望提升用户自主权以及对个人审核工具的信任。此外，未来研究应采取更全面的视角，考察平台、社群与个人审核如何协调一致，以避免冗余或缺漏。这对于设计能够平衡用户自主权与安全的内容审核生态系统至关重要。总之，这些研究方向共同指向一项研究议程：将个人审核视为在线内容审核这一社会技术系统的组成部分，而非孤立的功能。

6 结论

本文考察了遭受仇恨言论攻击的行动主义用户如何设想个人审核工具，以及不同平台屏幕上的个人审核需求有何差异。通过一项三轮德尔菲研究，我们逐步提炼出参与者的优先事项，结果表明，个人审核需求高度依赖屏幕类型，其中对话空间和算法策划空间最需要强有力的用户控制。我们还确定了用户看重哪些类型的功能，以及这些偏好如何随屏幕类型而变化。

本研究将屏幕引入为个人审核需求控制粒度的一种概念化方式，并依据实证结果对各屏幕上的审核功能进行优先级排序，为研究和设计作出了贡献。这种基于屏幕的视角为构建契合日常平台使用实际情况的工具，提供了兼具实践意义与概念价值的框架。

展望未来，有效的个人审核必须置于应对仇恨言论的更广泛社会技术响应之中，其中包括平台层级治理和问责结构。这也要求研究如何将个人审核与平台审核及社区内容审核相结合并进行校准，从而形成协调一致的多层系统。未来研究应探讨依照屏幕类型调整的个人审核工具在现实部署及不同平台上的表现。通过凸显平台情境如何塑造用户需求，本研究为开发更贴合当代在线环境复杂性的个人审核系统提供了一种方法。

致谢

我们感谢所有参与者分享他们的见解，也感谢匿名评审者提出有益且具有建设性的反馈。本研究由德国联邦研究、技术与太空部以及欧盟“下一代欧盟”计划资助。

参考文献

[1] Nuredin Ali Abdelkadir、Tianling Yang、Shivani Kapania、Meron Estefanos、Fasica Berhane Gebrekidan、Zecharias Zelalem、Messai Ali、Rishan Berhe、Dylan Baker、Zeerak Talat、MilagrosMiceli、Alex Hanna 和 Timnit Gebru。2025年。《专业知识在有效审核有害社交媒体内容中的作用。见《2025年CHI计算机系统中的人为因素会议论文集》（CHI ’25）。计算机协会，美国纽约州纽约，文章 227, 21 10 1145 3706598 3714010 页。doi: . / . Ayman Alarabiat和Isabel [2] Ramos。. 信息系统研究中的德尔菲法（2004-2017）。《商业研究方法电子期刊》 2019 17, 2 2019 86 99 10 34190 17 2 04（6月），–。doi: . /JBRM. . . Meysam [3] Alizadeh、Fabrizio Gilardi、Emma Hoes、K. Jonathan Klüser、Maël Kubli和 Nahema Marchal。2022年。作为政治议题的内容审核：围绕特朗普封禁的Twitter话语。《定量描述期刊：数字媒体》 2 2022 1 44 10 51685 2022 023（），–。 doi: . /jqd. . Adeela Arshad-Ayaz、Ayaz Naseem和Hedia Hizaoui。2022年。穆斯林及少 [4] 数族裔加拿大青年对仇恨言论和社交媒体的看法。《当代教育问题期刊》 17, 1 2022 21 39 10 20355 29489（），–。doi: . / jcie Kader Arslan和Matthias Trier。. 迈向社交媒体平台可供性与使用文化的差异化视角—— [5] 2021 一项组织性文献综述。见ACIS 2021论文集。AIS电子图书馆（AISeL），美国佐治亚州亚特兰大， 1 11 2021 80 –。https://aisel.aisnet.org/acis / Firas Bayram、Bestoun S. Ahmed和Andreas Kassler。2022年。从概念漂移到模型退化： [6] 性能感知漂移检测器综述。《基于知识的系统》 245 2022 108632 10 1016 2022 108632（），。doi: . / j.knosys. . Rahul Bhargava、Anna Chung、Neil S. Gaikwad、Alexis Hope、Dennis J [7] en、Jasmin Rubinovitz、Belén Saldías-Fuentes和Ethan Zuckerman。2019。Gobo：探索社交媒体中隐形算法用户控制的系统。见《2019计算机支持的协作工作与社会计算会议配套出版物》（美国得克萨斯州奥斯汀）（CSCW ’19 Companion）。计算机协会，美国纽约州纽约，151–155。doi:10.1145/3311957. 3359452[8] Block Party。2025年。Block Party 应用。https://www.blockpartyapp.com/。检索日期：2025年9月10日。[9] Bodyguard.ai。2025年。Bodyguard.ai。https://www.bodyguard.ai/en。检索日期：2025年9月1日。[10] Catherine Borshuk。2004。对外群体行动主义动机的解释性研究。《质性报告》9（2004），300–319。第2期。doi:10. 46743/2160-3715/2004.1930[11] Virginia Braun和Victoria Clarke。2019。对反思性主题分析的反思。《体育、锻炼与健康质性研究》11，4（2019），589–597。doi:10.1080/2159676X.2019.1628806[12] Sergio A. Castaño-Pulgarín、Natalia Suárez-Betancur、Luz M. T. Vega和Harvey Mauricio Herrera López。2021年。互联网、社交媒体和在线仇恨言论。系统综述。《攻击与暴力行为》58（2021年），101608。doi:10. 1016 2021 101608 /J.AVB. . Melisa Castellanos、Alexander Wettstein、Sebastian Wachs、Julia Kansok-Dusche、Cindy Ballaschk、Norman [13] Krause和Ludwig Bilz。2023年。青少年中的仇恨言论：一项关于患病率和人口统计学差异的双国家研究。见《教育前沿》，第8卷。Frontiers Media SA，瑞士洛桑，1076249。doi:10.3389/feduc.2023.1076249[14] Antonio Castillo-Esparcia、Lucia Caro-Castaño和 Ana Almansa-Martínez。2023年。社交媒体上数字行动主义的演变：机遇与挑战。《信息专业人士》 32, 3 2023 1 16 10 3145 2023 03（），–。 doi: . /epi. .may. Carmen Cervone、Martha Augoustinos和Anne Maass。. 贬损与仇恨的 [15] 2020 语言：功能、后果与重新挪用。《语言与社会心理学期刊》40（2020年），80 – 101。 doi:10.1177/ 0261927X20967394[16] Ashleigh Charles、Laurie Hare-Duke、Hannah Nudds、Donna Franklin、Joy Llewellyn-Beardsley、Stefan Rennick-Egglestone、Onni Gust、Fiona Ng、Elizabeth Evans、Emily Knox等。2022年。内容警告与触发警告的类型学：系统综述。PLOS ONE 17，5（2022年），e0266722。doi:10.1371/journal. 0266722 pone.[17] Pithayuth Charnsethikul、Almajd Zunquti、Gale Lucas和 Jelena Mirkovic。2025年。应对社交媒体隐私：认知、偏好与可发现性。《隐私增强技术论文集》2025，4（2025年），620–638。doi:10. 56553 2025 0148 /popets- - Qijia Chen、Qunfang Wu和Giulio Jacucci。2025年。民主式内容审核：探索社交虚拟现实 [18] 中投票踢出的使用与感知。见《2025年CHI计算机系统中的人为因素会议论文集》（CHI ’25）。计算机协会，美国纽约州纽约，文章 1238, 18 10 1145 3706598 3713577 页。doi: . / . Marion Collewet 和 Paul Koster。. 基于点数分配实验的偏好估计。《选择建模期刊》 [19] 2023 48（2023年），100430。doi:10.1016/j. jocm. 2023 100430 . Juliet Corbin和 Anselm Strauss。2014。质性研究基础：发展扎根理论的技术与程序。Sage Publications, [20] 美国加利福尼亚州千橡市。[21] Kerstin Cuhls。2023年。德尔菲法：导论。Springer Fachmedien Wiesbaden，威斯巴登， 3 27 10 1007 978 3 658 38862 1_1 –。doi: . / - - - - Kerstin E. Cuhls。. 论证式德尔菲调查：对社会学研究的启示。《美国社会学家》55，2（2024年6月），[22] 2024 120–141。doi:10.1007/s12108-

023-09596-x 出版商：Springer Science and Business Media LLC。[23] Lucas Dixon、John Li、Jeffrey Sorensen、Nithum Thain和Lucy Vasserman。2018年。测量并缓解文本分类中的非预期偏见。见*2018*年*AAAI/ACM*人工智能、伦理与社会会议论文集 （美国路易斯安那州新奥尔良） （*AIES '18*）。计算机协会，美国纽约州纽约，67–73。doi:10.1145/3278721.3278729[24] 董思杰、王启同、Sahri Soror、Themis Palpanas和Divesh Srivastava。2024年。高效缓解数据漂移对机器学习流水线的影响。《*VLDB Endow.*论文集》 17 2024 3072 3081 10 14778 3681954 3681984 （），–。[25] doi: . / . Rebecca Dorn、Lee Kezar、Fred Morstatter和Kristina Lerman。. 2024. 语言模型的有害言论检测表现出性别酷儿方言偏见。见第*4*届*ACM*算法、机制与优化中的公平与可及性会议论文集 （墨西哥圣路易斯波托西） （*EAAMO '24*）。计算机协会，美国纽约州纽约，文章编号6，12页。doi:10.1145/3689904.3694704[26] Houda El mimouni、Sarah Rüller、Konstantin Aal、Yarden Skop、Norah Abokhodair、Volker Wulf和Peter Tolmie。2025年。探索算法抵抗：行动主义对社交媒体审查的回应。《*ACM*人机交互论文集》 9, 2, 072 2025 24 2024 *2* 10 1145 3710970 文章CSCW（5月），页。doi: . / Motahhare Eslami和Hendrik Heuer。.。重访算法体制中的透明度努力。阿姆斯特丹大学出版社，阿姆斯特丹，37–56。doi:10.1515/9789048556908-002[28] 欧盟委员会。2024年。欧盟委员会根据《数字服务法案》要求X提供有关减少内容审核资源的信息。https: // digital-strategy.ec.europa.eu/en/news/commission-requests-information-x-decreasing-content-moderation-resources-under-digital-services。检索日期：2025年9月4日。[29] 欧洲议会。2025年。执行副主席就《数字服务法案》下TikTok内容审核所作的答复（E-002454/2024）。https://www.europarl.europa.eu/doceo/document/E-10-2024-002454_EN.html。检索日期：2025年9月4日。[30] Facebook帮助中心。2025年。Facebook将什么视为仇恨言论? https://www.facebook.com/help/131671940241729。检索日期：2025年9月1日。[31] Diane Felmlee、Daniel DellaPosta、Paulina d C Inara Rodis和Stephen A Matthews。2020年。社交媒体反滥用政策能否奏效？一项关于网络性别歧视和种族主义侮辱语的准实验研究。*Socius* 6（2020年）。doi:10. 1177/2378023120948711[32] Francis Fukuyama、Barak Richman、Ashish Goel、Roberta R Katz、A Douglas Melamed和Marietje Schaake。2020年。面向主导型数字平台的中间件：应对民主威胁的技术解决方案。网络政策中心，弗里曼·斯波格利研究所 （2020年），24–37。[33] Daniel Gayo-Avello。2015。社交媒体、民主与民主化。*IEEE*多媒体 22, 2 2015 10 16 10 1109 2015 47 （），–。doi: . /MMUL. . 2018. [34] Tarleton Gillespie。.互联网的守护者：平台、内容审核以及塑造社交媒体的隐性决策。耶鲁大学出版社，美国纽黑文。[35] Tommi Gröndahl、Luca Pajola、Mika Juuti、Mauro Conti和N. Asokan。2018年。你所需要的只是"爱"：规避仇恨言论检测。见第*11*届*ACM*人工智能与安全研讨会论文集 （加拿大多伦多） （*AISec '18*）。计算机协会，美国纽约州纽约，2–12。doi:10.1145/3270101.3270103[36] Jaron Harambam、Dimitrios Bountouridis、Mykola Makhortykh和Joris van Hoboken。2019。将用户纳入考量以设计得更好：（新闻）推荐系统中用户控制机制的质性评估。见第*13*届*ACM*推荐系统会议论文集 （丹麦哥本哈根） （*RecSys '19*）。计算机协会，美国纽约州纽约， 69 77 10 1145 3298689 3347014 –。doi: . / . David Hartmann、Amin Oueslati、Dimitri Staufer、Lena Pohlmann、Simon Munzert和Hendrik Heuer。2025年。迷失于审核之中：商业内容审核API如何对针对群体的仇恨言论及语言变体过度审核与审核不足。见*2025*年*CHI*计算机系统中的人为因素会议论文集。计算机协会，美国纽约州纽约，1–26。doi:10.1145/3706598.3713998[38] HateAid和兰德克数字正义运动。2021年。《网络上的无边界仇恨——欧洲各地形势严峻》。https://hateaid.org/wp- content/uploads/2022/04/HateAid-Report-2021-DE.pdf [39] Sharon Heung、Lucy Jiang、Shiri Azenkot和Aditya Vashistha。2025年。"无知并非福"：设计个性化审核以应对社交媒体上歧视残障人士的仇恨言论。见*2025*年*CHI*计算机系统中的人为因素会议论文集（*CHI '25*）。计算机协会，美国纽约州纽约，文章 173, 18 10 1145 3706598 3713997 页。doi: . / . Chia-Chien Hsu和Brian A Sandford。. 2007. 德尔菲法：理解共识。实践评估、研究与评价 12，1（2007），10。doi: 10 7275 9 90 . /pdz -th Reddit公司。*.Reddit*规则（内容政策）。Reddit公司。[41] 2025. https://redditinc.com/ policies/content-policy 检索日期：2025年11月20日。[42] Instagram 帮助中心。2025年。Instagram 隐私设置与相关信息。网址：https://help.instagram.com/196883487377501/。检索日期：2025-09- 04。[43] Shagun Jhaver。2025。Facebook 上的个人审核配置：探讨错失恐惧、社交媒体成瘾、规范和平台信任的作用。《First Monday》 30, 9 2025 10 5210 30 9 14272 ( )。doi: . /fm.v i 。Shagun Jhaver、[44] Quan Ze Chen、Detlef Knauss 和 Amy X. Zhang。2022。为创作者主导的评论审核设计词语过滤工具。载于《2022年 CHI计算机系统中的人为因素会议论文集》（美国路易斯安那州新奥尔良）（CHI '22）。计算机协会，美国纽约州纽约市，文章 205, 21 10 1145 3491102 3517505 页。doi: . / 。Shagun Jhaver、Sucheta Ghoshal、[45] Amy Bruckman 和 Eric Gilbert。2018。在线骚扰与内容审核：屏蔽列表案例。《ACM计算机与人机交互汇刊》（TOCHI）25卷2期（2018），1–33。doi:10.1145/3185593[46] Shagun Jhaver、Alice Qian Zhang、Quan Ze Chen、Nikhila Natarajan、Ruotong Wang 和 Amy X. Zhang。2023。个性化社交媒体内容审核：用户对审核选择、界面设计与劳动的看法。《ACM 人机交互汇刊论文集》7卷，CSCW2（2023年9月），1–33。doi:10.1145/3610080[47] Shagun Jhaver 和 Amy X Zhang。2023。用户想要平台审核还是个人控制？考察第三人效应和言论自由支持在塑造审核偏好中的作用。《新媒体与社会》27卷5期（2023），2930–2950。doi:10.1177/14614448231217993[48] Jialun Aaron Jiang、Morgan Klaus Scheuerman、Casey Fiesler 和 Jed R Brubaker。2021。了解国际社会对在线有害内容严重程度的认知。《PLOS ONE》16卷8期（2021），e0256762。doi:10.1371/journal.pone.0256762[49] Anna Ricarda Luther、Hendrik Heuer、Stephanie Geise、Sebastian Haunss 和 Andreas Breiter。2025年。面向活动人士的社交媒体：重新构想安全、内容呈现与工作流程。载于《2025年 CHI计算机系统中的人为因素会议论文集》（日本横滨）（CHI '25）。计算机协会，美国纽约州纽约市，文章956，18页。doi:10.1145/3706598.3713351[50] Walid Magdy、Hamdy Mubarak 和 Joni Salminen。2025年。谁应制定标准？分析巴勒斯坦—以色列冲突期间 Facebook 上遭审查的阿拉伯语内容。载于《2025年 CHI计算机系统中的人为因素会议论文集》（CHI '25）。计算机协会，美国纽约州纽约市，文章 178, 16 10 1145 3706598 3713150 页。doi: . / 。[51] 2022. Maggie Engler。。中间件与内容审核的定制化——Maggie Engler。 h t t p s:// i ntegrityinstitute.org/blog/ middleware-and-the-customization-of-content-moderation[52] Dimitrios M. Manias、Ali Chouman 和 Abdallah Shami。2023。动态网络中的模型漂移。《IEEE Communications Magazine》61（2023），78–84。doi:10. 1109/MCOM.003.2200306 [53] Meta。2025年。更多发言，更少错误。https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/。检索日期：2025年9月1日。[54] Meta。2025年。限制账户——执法透明度。Meta。https://transparency.meta.com/enforcement/taking-action/restricting-accounts/ 检索日期：2025年11月20日。[55] Radu M. Meza、Hanna-Orsolya Vincze 和 Andreea Mogos。2018。语境中的在线仇恨言论目标。《Intersections》4卷4期（2018），26–50。doi:10.17356/ieejsp.v 4 4 503 i 。Silvia Milano、Mariarosaria Taddeo 和 Luciano Floridi。2020。[56] 推荐系统及其伦理挑战。《AI & Society》35卷4期（2020），957–967。doi:10. 1007/s00146-020-00950-y [57] Sandra Lopes Miranda。2023。分析 Instagram 上针对女性的仇恨言论。《Open Information Science》7卷1期（2023），20220161。[58] MIT 媒体实验室。检索日期：2025年9月1日。Project Gobo：你的社交媒体，由你做主。https://www.media.mit.edu/projects/gobo/overview/。[59] Mina Momeni。2017。社交媒体与政治参与。《新媒体与社会》19（2017），2094–2100。doi:10.1177/1461444817728054[60] Juan Manuel Pérez、Franco M Luque、Demian Zayat、Martín Kondratzky、Agustín Moro、Pablo Santiago Serrati、Joaquín Zajac、Paula Miguel、Natalia Debandi、Agustín Gravano 等。2023。评估语境信息对仇恨言论检测的影响。《IEEE Access》11（2023），30575–30590。doi:10.1109/ACCESS. 2023.3258973[61] Fabio Poletto、Valerio Basile、Manuela Sanguinetti、Cristina Bosco 和 Viviana Patti。2020。仇恨言论检测资源与基准语料库：系统综述。《La nguage Resources and Evaluation》55（2020），477–523。doi:10.1007/s10579-020-09502-8[62] Mattia Retta。2023。社交媒体仇恨词语的语用与话语分析。《Internet Pragmatics》6卷2期（2023），197–218。doi:10.1075/ip.00096.ret [63] Kristine M. Rogers。2022。用户对社交媒体信息流编排的偏好。载于《社交计算与社交媒体：设计、用户体验与影响》，Gabriele Meiselwitz（编）。Springer International Publishing，瑞士 Cham，185–204。[64] Alexander Ronzhyn、Ana Sofia Cardenal 和 Albert Batlle Rubio。2023。定义社交媒体研究中的可供性：文献综述。《新媒体与社会》25卷11期（2023），3165–3188。doi:10.1177/14614448221135187

[65] Joni Salminen、Juan Corporan、Soon-Gyo Jung 和 Bernard J. Jansen。2021。借助 Take Back Control 重新掌控社交媒体信息流。载于《2021年智能系统与应用创新国际会议》（INISTA）。IEEE，土耳其科贾埃利，1 7 10 1109 52262 2021 9548614 – 。doi: . / inista 。。[66] Joseph Seering、Tonya Wang、Jina Yoon 和 Geoff F. Kaufman。2019。算法时代的版主参与与社区发展。《新媒体与社会》21 2019 1417 1443 10 1177 1461444818821316 ( ), – 。doi: . / [67] Qinlan Shen 和 Carolyn Rose。2019。在线内容审核话语：探究 Reddit 隔离政策变更引发的两极化用户反应。载于《第三届在线辱骂性语言研讨会论文集》，Sarah T. Roberts、Joel Tetreault、Vinodkumar Prabhakaran 和 Zeerak Waseem（编）。计算语言学协会，意大利佛罗伦萨，58–69。doi:10.18653/v1/W19-3507[68] Alexandra A. Siegel。2020年。《在线仇恨言论》。剑桥大学出版社，剑桥，56–88。[69] Katta Spiel、Os Keyes 和 Pınar Barlas。2019。修补性别：人机交互中的非二元性别乌托邦。载于《2019年 CHI计算机系统中的人为因素会议扩展摘要集》（英国苏格兰格拉斯哥）（CHI EA '19）。计算机协会，美国纽约州纽约市，1 11 10 1145 3290607 3310425 – 。doi: . / 。Statista。 。[70] 2024 全球社交媒体用户每月活跃人数。Statista GmbH。
h
t
t
p
s://
d
e.
s
t
atist
a.com/statistik/daten/studie/739881/umfrage/monatlich-aktive-social-media-nutzer-weltweit/ 检索日期：2025年12月4日。[71] Francesca Stevens、Jason RC Nurse 和 Budi Arief。2021。网络跟踪、网络骚扰与成人心理健康：系统综述。《网络心理学、行为与社交网络》24卷6期（2021），367–376。doi:10.1089/cyber.2020.0253[72] Fatemeh Tahmasbi、Leonard Schild、Chen Ling、Jeremy Blackburn、Gianluca Stringhini、Yang Zhang 和 Savvas Zannettou。2021。"去吃蝙蝠吧，Chang！"：新冠肺炎疫情下网络社区仇华行为的兴起。载于《2021年万维网大会论文集》（斯洛文尼亚卢布尔雅那）（WWW '21）。计算机协会，美国纽约州纽约市，1122–1133。doi:10.1145/3442381.3450024[73] TikTok 安全中心。2025年。内容违规与封禁——账户与用户安全。TikTok。https://support.tiktok.com/en/safety-hc/account-and-user-safety/content-violations-and-bans 检索日期：2025年11月20日。[74] Dennis von Oehsen、André Rhody、Claudia Hess 和 Natalie Beisch。2025年。ARD/ZDF 媒体研究。（版本1.0.0） [数据集]。GESIS，科隆。https://doi.org/10.7802/2764。[75] Leijie Wang、Kathryn Yurechko、Pranati Dani、Quan Ze Chen 和 Amy X. Zhang。2025年。终端用户创作个性化内容分类器：比较示例标注、规则编写和大语言模型提示。载于《2025年 CHI计算机系统中的人为因素会议论文集》（CHI '25）。计算机协会，美国纽约州纽约市，文章1214，21页。doi:10. 1145/3706598.3713691[76] Sai Wang。2023。影响用户对社交媒体中基于人工智能（AI）的内容审核认知的因素。《计算机与人类行为》149 2023 107971 10 1016 2023 107971 ( ), 。doi: . /j.chb. 。[77] Galen Weld、Amy X. Zhang 和 Tim Althoff。2022。是什么让在线社群"更好"？衡量数千个子版块中的价值观、共识与冲突。《AAAI 网络与社交媒体国际会议论文集》16, 1 2022 1121 1132 10 1609 16 1 19363 （5月），– 。doi: . /icwsm.v i 。[78] Hannah Willems、Julia A Glombiewski、Richard J McNally 和 Philipp Herzog。2025。触发警告≠触发警告：比较措辞不同的触发警告对负面与正面结果的影响。《心理创伤：理论、研究、实践与政策》17卷，增刊1（2025），139–149。doi:10.1037/tra0001839[79] Matthew L Williams、Pete Burnap、Amir Javed、Han Liu 和 Sefa Ozalp。2020。机器中的仇恨：反黑人和反穆斯林社交媒体帖文能否预测线下种族与宗教仇恨加重犯罪。《英国犯罪学杂志》60, 1 2020 93 117 10 1093 049 ( ), – 。doi: . /bjc/azz Marianne Wilson、David Brazier、Dimitra Gkatzia 和 Peter Robertson。2024。与领域专家开展参与式设计：用于职业支持聊天机器人的一项德尔菲研究。载于《第六届 ACM 对话式用户界面会议论文集》（卢森堡卢森堡）（CUI '24）。计算机协会，美国纽约州纽约市，文章8, 12 10 1145 3640794 3665534 页。doi: . / 。[80] Stanisław Woźniak、Jacek Duszenko、Jan Kocoń 和 Przemysaw Kazienko。2025。通过用户可控档案改进基于大语言模型的推荐系统。载于《ACM 万维网大会2025年伴随论文集》（澳大利亚新南威尔士州悉尼）（WWW '25）。计算机协会，美国纽约州纽约市，2102 2111 10 1145 3701716 3717734 – 。doi: . / 。[81] Sijia Xiao、Coye Cheshire 和 Niloufar Salehi。2022。意义建构、支持、安全、惩罚、转变：以恢复性司法方法理解青少年应对网络伤害的需求。载于《2022年 CHI计算机系统中的人为因素会议论文集》（美国路易斯安那州新奥尔良）（CHI '22）。计算机协会，美国纽约州纽约市，文章146, 15 10 1145 3491102 3517614 页。doi: . / 。[82] Mustafa Mikdat Yildirim、Jonathan Nagler、Richard Bonneau 和 Joshua A. Tucker。2023。无需封禁：停用警告如何减少 Twitter 上的仇恨言论。《政治视角》21卷2期（2023），651–663。doi:10.1017/S1537592721002589[84] Ziqi Zhang 和 Lei Luo。2019。仇恨言论检测：一个已解决的问题？Twitter 上长尾问题这一棘手案例。语义网10, 5 (2019), 925–945。doi:10.3233/SW-180338

A 附录

A.1 第1轮、第2轮和第3轮的屏幕重要性评分

表8：第1波、第2波和第3波的屏幕重要性评分概览。屏幕按第3轮中认为其重要的参与者比例排序。

屏幕	重要性 第1波	重要性第2波	重要性 第3波
所有评论	✔ 81.6% ✘ 13.2%	✔ 92.1% ✘ 2.6%	✔ 100.0% ✘ 0.0%
自己帖子的评论	✔ 84.6% ✘ 5.1%	✔ 89.5% ✘ 5.3%	✔ 92.5% ✘ 2.5%
Reels 标签页	✔ 61.5% ✘ 15.4%	✔ 78.9% ✘ 7.9%	✔ 87.5% ✘ 7.5%
主页信息流	✔ 65.8% ✘ 28.9%	✔ 76.3% ✘ 15.8%	✔ 87.5% ✘ 10.0%
所有私信	✔ 71.8% ✘ 17.9%	✔ 81.6% ✘ 15.8%	✔ 85.0% ✘ 7.5%
特定帖子的评论	✔ 72.5% ✘ 17.5%	✔ 76.3% ✘ 13.2%	✔ 77.5% ✘ 7.5%
特定账户的私信	✔ 57.5% ✘ 35.0%	✔ 73.7% ✘ 10.5%	✔ 75.0% ✘ 12.5%
探索选项卡	✔ 71.4% ✘ 22.9%	✔ 73.7% ✘ 15.8%	✔ 72.5% ✘ 15.0%
搜索结果	✔ 41.1% ✘ 33.3%	-	-

A.2 功能的重要性评分（第1轮、第2轮）及排名（第3轮）

表9：针对“所有评论”屏幕的所有功能请求，依据40名曾遭受针对性仇恨言论的社交媒体活动人士的反馈整理而成。该表显示了第1轮和第2轮的重要性评分，以及第3轮的最终排名和得分百分比。各项请求按照第3轮的排名排序。

功能请求 所有评论	重要性 第一波	重要性 第二波	排名 第三波（% 分数）
自定义对检测到的仇恨言论的响应	✔ 93.9% ✘ 3.0%	✔ 92.3% ✘ 0.0%	1. (13.8%)
基于词语的过滤	✔ 90.9% ✘ 0.0%	✔ 94.9% ✘ 2.6%	2. (12.2%)
更改过滤决策	-	✔ 92.3% ✘ 2.6%	3. (11.0%)
隐藏评论区	-	✔ 76.9% ✘ 10.3%	4. (10.0%)
基于语气的过滤	-	✔ 71.8% ✘ 7.7%	5. (10.0%)
基于主题的筛选	✔ 66.7% ✘ 12.1%	✔ 92.3% ✘ 0.0%	6. (9.0%)
基于用户的过滤	-	✔ 82.1% ✘ 5.1%	7. (9.0%)
基于交流模式的过滤	-	✔ 56.4% ✘ 25.6%	8. (8.8%)

表10：针对屏幕功能自己帖子上的评论的所有功能请求，依据40名曾遭受针对性仇恨言论的社会媒体活动人士提供的意见。表中展示了第1轮和第2轮的重要性评分，以及最终排名和第3轮的得分百分比。请求按第3轮排名排序。

功能请求 自己帖子上的评论	重要性第一波	重要性第二波	排名 第三波（% 分数）
自定义对检测到的仇恨言论的响应	✔ 94.6% ✘ 2.7%	✔ 94.7% ✘ 5.3%	1. (15.6%)
基于用户的过滤	-	✔ 81.6% ✘ 10.5%	2. (12.4%)
更改过滤决策	-	✔ 92.1% ✘ 7.9%	3.（11.9%）
基于词语的过滤	✔ 89.2% ✘ 2.7%	✔ 93.3% ✘ 6.7%	4.（11.7%）
自动屏蔽对所选评论的回复	-	✔ 76.3% ✘ 10.5%	5. (9.3%)
手动允许评论	-	✔ 71.1% ✘ 15.8%	6. (9.1%)
基于帖子的筛选设置	-	✔ 76.3% ✘ 15.8%	7. (7.9%)
基于主题的过滤	✔ 64.9% ✘ 21.6%	✔ 81.6% ✘ 7.9%	8. (7.7%)
列出受信任的评论者	-	-	9.（6.8%）

表11：根据40名遭受仇恨言论攻击的社会媒体活动人士提供的意见，列出屏幕*Reels* 标签页的所有功能请求。该表展示了第1轮和第2轮的重要性评分，以及第3轮的最终排名和得分百分比。请求按其第3轮排名排序。

功能请求*Reels* 标签页	重要性第一波	重要性第二波	排名 第3轮（% 分数）
自定义对检测到的仇恨言论的响应	✔ 81.8% ✘ 3.0%	✔ 91.9% ✘ 2.7%	1. (11.2%)
手动视频控制	-	✔ 88.9% ✘ 2.8%	2. (9.8%)
安全空间模式	-	-	3. (9.3%)
基于主题的过滤	✔ 78.8% ✘ 0.0%	✔ 94.6% ✘ 0.0%	4. (8.9%)
基于词语的过滤	✔ 69.7% ✘ 9.1%	✔ 91.9% ✘ 2.7%	5. (8.1%)
基于音频内容的过滤	-	✔ 85.7% ✘ 11.4%	6. (7.3%)
基于用户的过滤	-	✔ 75.7% ✘ 2.7%	7. (7.3%)
基于 Reel 评论的筛选	-	-	8. (6.6%)
基于 Reel 格式的筛选	-	✔ 74.3% ✘ 5.7%	9. (5.6%)

表12：针对屏幕主页信息流的所有功能请求，依据40名曾遭受针对性仇恨言论的社会媒体活动人士提供的意见整理而成。该表列出了第1轮和第2轮的重要性评分，以及第3轮的最终排名和得分百分比。各项请求按其第3轮排名排序。

功能请求主页信息流	重要性 第1波	重要性 第2波	排名 W3（% 分数）
区分仇恨言论与反仇恨言论	-	✔ 91.2% ✘ 2.9%	1. (16.3%)
自定义对检测到的仇恨言论的响应	✔ 96.3% ✘ 3.7%	✔ 94.1% ✘ 0.0%	2.（14.0%）
基于词语的过滤	✔ 77.8% ✘ 11.1%	✔ 97.1% ✘ 2.9%	3.（11.5%）
基于用户的过滤	-	✔ 88.2% ✘ 2.9%	4. (9.7%)
上下文敏感过滤	-	✔ 67.6% ✘ 17.6%	5. (9.4%)
基于主题的过滤	✔ 81.5% ✘ 14.8%	✔ 91.2%✘ 0.0%	6. (8.9%)
基于图像内容的过滤	-	✔ 76.5% ✘ 14.7%	7. (8.7%)
基于视频内容的过滤	-	✔ 73.5% ✘ 14.7%	8. (8.2%)
基于音频内容的过滤	-	✔ 67.6% ✘ 14.7%	9. (6.9%)
基于示例的过滤	-	✔ 67.6% ✘ 5.9%	10. (6.4%)

表13：基于40位遭受仇恨言论攻击的社交媒体活动人士用户的意见，列出屏幕所有私信的所有功能请求。表中展示了第1轮和第2轮的重要性评分，以及最终排名和第3轮的得分百分比。请求按第3轮排名排序。

功能请求 所有私信	重要性第一波	重要性第二波	排名 W3（% 分数）
自定义对检测到的仇恨言论的响应	✔ 96.9% ✘ 0.0%	✔ 97.1% ✘ 0.0%	1. (14.9%)
区分仇恨言论与反仇恨言论	-	✔ 91.2% ✘ 2.9%	2. (14.6%)
基于交流模式的过滤	-	✔ 76.5% ✘ 8.8%	3. (12.2%)
基于用户的过滤	-	✔ 85.3% ✘ 2.9%	4. (11.9%)
基于词语的过滤	✔ 78.1% ✘ 9.4%	✔ 94.1% ✘ 2.9%	5. (10.0%)
基于图像内容的过滤	-	-	6. (7.8%)
链接过滤	-	✔ 73.5% ✘ 23.5%	7. (6.5%)
基于语言的过滤	-	-	8.（5.4%）

表14：基于40位遭受仇恨言论攻击的社交媒体活动人士用户的意见，列出屏幕*CommentsSpecificPosts*的所有功能请求。表中展示了第1轮和第2轮的重要性评分，以及最终排名和第3轮的得分百分比。请求按第3轮排名排序。

功能请求 特定帖子的评论	**重要性第一波**	**重要性第二波**	**排名 第三波（% 分数）**
自定义对检测到的仇恨言论的响应	✔ 97.9% ✘ 0.0%	✔ 94.3% ✘ 0.0%	1. (20.8%)
基于词语的过滤	✔ 93.6% ✘ 0.0%	✔ 94.3% ✘ 2.9%	2. (18.0%)
基于主题的过滤	✔ 72.3% ✘ 10.6%	✔ 82.9% ✘ 2.9%	3. (14.7%)
基于用户的过滤	-	✔ 85.7% ✘ 5.7%	4.（14.7%）
自动调整筛选强度	-	✔ 77.1% ✘ 2.9%	5.（12.7%）
自动参照社区准则	-	-	6.（12.7%）
基于标签筛选	-	✔ 65.7% ✘ 17.1%	7. (7.3%)

表15：针对仇恨言论的40名社会媒体活动人士提出的、适用于“特定账户的私信”屏幕的所有功能请求。该表显示了第1轮和第二波的重要性评分，以及最终排名和第3轮的得分百分比。请求按其第3轮排名排序。

功能请求 特定账户的私信	**重要性第一波**	**重要性第二波**	**第3波排名（% 分数）**
自定义对检测到的仇恨言论的响应	✔ 94.7% ✘ 0.0%	✔ 94.7% ✘ 0.0%	1.（23.3%）
基于词语的过滤	✔ 89.5% ✘ 5.3%	✔ 91.7% ✘ 2.8%	2. (16.0%)
基于关注者的私信屏蔽	-	✔ 68.6% ✘ 8.6%	3. (15.1%)
基于信任评分的过滤	-	✔ 76.5% ✘ 14.7%	4. (14.3%)
基于用户的过滤	-	✔ 88.9% ✘ 2.8%	5. (14.3%)
基于主题的过滤	✔ 65.8% ✘ 15.8%	✔ 83.3% ✘ 2.8%	6.（11.8%）
上下文敏感过滤	-	-	7.（5.3%）

表16：针对仇恨言论的40名社会媒体活动人士提出的关于探索选项卡屏幕的所有功能请求。表中列出了第1轮和第2轮的重要性评分，以及最终排名和第3轮的得分百分比。各项请求按其第3轮排名排序。

功能请求 探索选项卡	**第1轮重要性**	**第2轮重要性**	**第3轮排名（得分百分比）**
自定义对检测到的仇恨言论的响应	✔ 86.0% ✘ 4.7%	✔ 91.2% ✘ 2.9%	1. (20.2%)
基于词语的过滤	✔ 72.1% ✘ 14.0%	✔ 91.2% ✘ 2.9%	2. (14.7%)
基于主题的过滤	✔ 74.4% ✘ 11.6%	✔ 94.1% ✘ 0.0%	3. (14.3%)
基于图像内容的过滤	-	✔ 79.4% ✘ 11.8%	4. (14.3%)
基于语气的过滤	-	✔ 64.7% ✘ 14.7%	5. (12.1%)
基于社区评分的过滤	-	✔ 67.6% ✘ 11.8%	6. (9.2%)
针对未知账户内容的警告	-	-	7. (9.2%)
基于内容传播度的过滤	-	-	8. (5.9%)