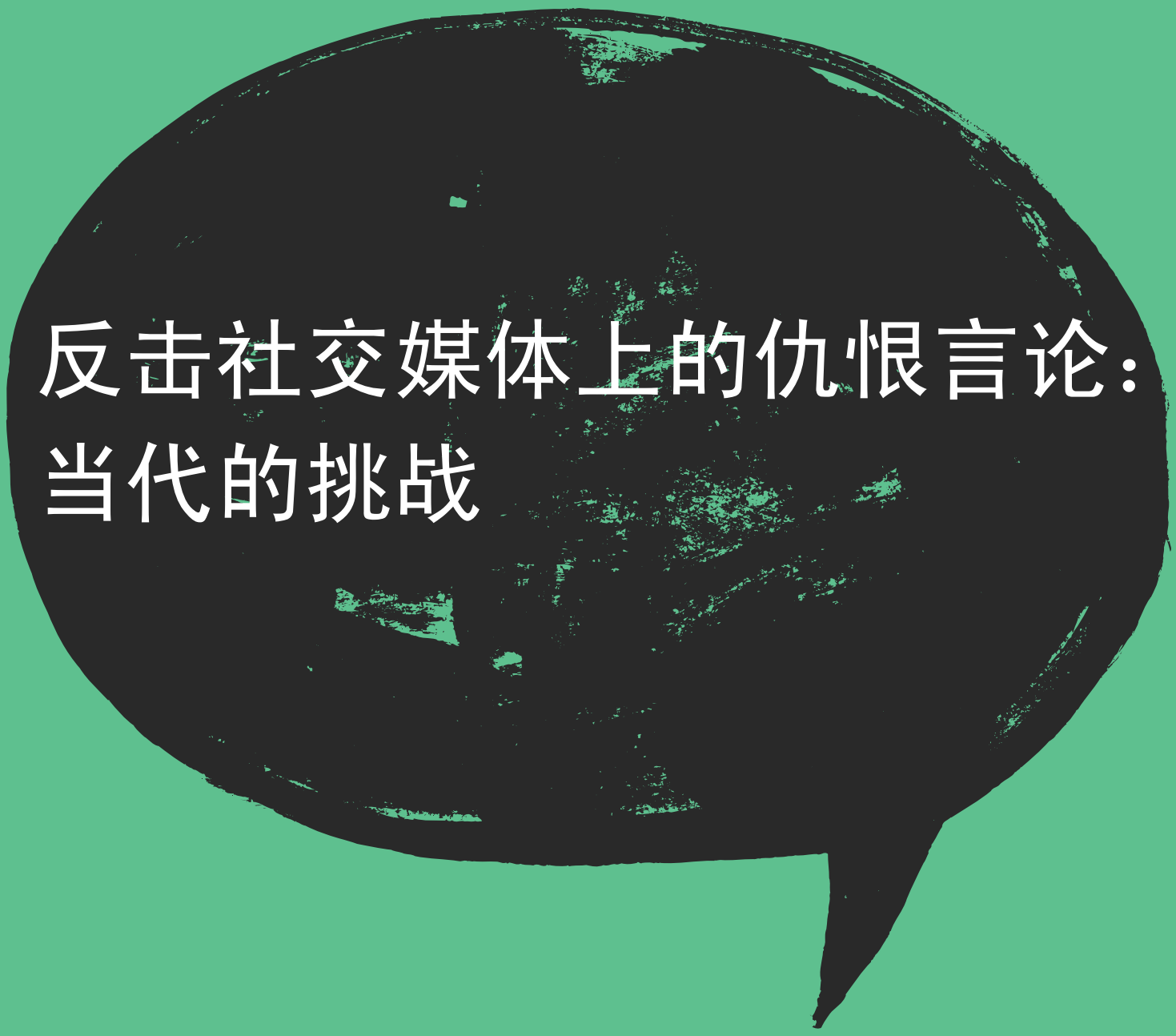
unesco

UNITED NATIONS OFFICE ON GENOCIDE PREVENTION AND THE RESPONSIBILITY TO PROTECT



反击社交媒体上的仇恨言论：当代的挑战

讨论文件

01

本文件由牛津大学互联网研究所（Oxford Internet Institute）研究人员编写，由联合国教科文组织提供支持，旨在推进《联合国关于仇恨言论的战略和行动计划》，并纳入了由欧盟资助的“#冠状病毒病真相：应对冲突易发环境下的冠状病毒病‘信息疫情’”项目框架中。本文件是牛津大学互联网研究所与教科文组织合作开发的一个工具包的组成部分，该工具包包含各种现有的方法、资源及研究项目，旨在监测网络仇恨言论的发生、传播和影响，以及评估反击仇恨言论的能力和做法。欢迎就本文件提出相关意见，以利于更加深入地开展研究。

应对并反击仇恨言论需要从多层面开展工作，包括消除其根源和驱动因素、防止其转化为暴力行动，并应对其带来的社会后果。要制定应对仇恨言论的有效措施（包括通过教育手段），必须利用清晰可靠的数据更好地监测和分析这种现象。在数字时代，这也意味着更好地了解网络仇恨言论的发生、危害程度及影响范围。

从方法论角度看，研究如何识别网络仇恨言论面临着诸多挑战——包括用以框定问题的定义、社会和历史背景、语言的微妙之处、在线社群的多样性以及网络仇恨言论的形式（语言类型、图像等）。从技术角度看，由于检测系统可靠性不稳定、专有算法晦涩不明、各家公司持有的数据无法访问等原因，网络仇恨言论研究困难重重。只有明确了如何应对这些挑战，才能进一步了解网络仇恨言论如何出现和扩散，也才能制定有效的应对措施。

《联合国关于仇恨言论的战略和行动计划》确定了一系列监测和分析仇恨言论的优先领域，规定联合国相关实体应能够“识别、监测、收集数据和分析仇恨言论趋势”。在关注网络仇恨言论时，鼓励联合国实体就“滥用互联网和社交媒体传播仇恨言论与驱使个人走向暴力的因素之间的关系开展更多研究”；“了解新技术和数字平台带来的与仇恨言论传播相关的新风险和新机遇”；最后，鼓励“制定应对新形式的数字仇恨言论的行动协议”。

在过去的一年中，仇恨言论席卷全球，进一步加剧了对特定群体的不容忍和歧视现象，破坏了社会和政治制度的稳定，而冠状病毒病疫情进一步凸显了《联合国关于仇恨言论的战略和行动计划》的相关性。

本讨论文件旨在概述解决社交媒体仇恨言论时需要考虑的重要问题：或由社交媒体公司作出具体规定，或通过反击措施和立法手段，或采取预防性教育措施。本文件分为三个部分：第1部分侧重于仇恨言论的定义及相关的法律框架；第2部分介绍了用于监测网络仇恨言论的各种工具和技术，并探讨了网络仇恨言论泛滥程度的衡量标准；第3部分探讨了可能采取的对策和预防措施。

1 仇恨言论的定义

应对仇恨言论并制定相关立法，困难始于如何为其定义。对于仇恨言论，并没有一个国际公认的定义。相反，还引出了涉及意见和表达自由、歧视和鼓吹或煽动歧视、敌意或暴力等的一些法律问题。

苏珊·贝尼希的**危险言论（Dangerous Speech）项目**[1] 表明，“仇恨言论”这一术语有两个主要难点。首先，“仇恨”是一个模糊的词语，可以表示不同的仇恨程度，并可能导致不同的后果——“仇恨言论中的‘仇恨’是意味着说话者仇恨、试图说服他人仇恨、还是想要让他人感觉受到仇恨？” [2]其次，“仇恨言论”的核心是指某些个人或某个群体因其身份/群体成员身份而成为攻击目标。这就要求法律或定义具体说明是否认为所有身份和群体都属于该项法律管辖的范畴，如果不是，则应说明包括哪些群体。“危险言论项目”认为，过于宽泛的法律可能会被滥用，不利于弱势群体或政治和公民反对派，有时反而会伤害仇恨言论法律原本要保护的群体。 然而，也可以认为，一项定义如果过于狭隘地关注特定群体和身份，可能会导致法律排斥现象或缺乏解决该问题的法律工具的现象。

虽然本讨论文件受其范畴所限，无法详细研究这些挑战，但是，纵览一下世界各地的国际法和国家法律，也能说明该问题的复杂性以及对仇恨言论存在的不同解释。

在全球层面，除了不具约束力的《世界人权宣言》外，《公民及政治权利国际公约》（ICCPR）规定了表达自由的权利（第十九条），并禁止构成煽动歧视、敌视或强暴的任何鼓吹仇恨的主张（第二十条）。第十九条和第二十条还规定了对表达自由的限制应**“以经法律规定，且为下列各项所必要者为限：（a）尊重他人权利或名誉；（b）保障国家安全或公共秩序、或公共卫生或风化。”**

作为对这些原则的补充，《拉巴特行动计划》提出了一个**“六项门槛测试”**，以此证明言论自由限制的合理性，同时考虑**社会政治背景、言论发表者的地位、煽动对立的意图、言论内容、传播的范围及造成危害的可能性。**

用于应对仇恨言论的其他重要文件还有《消除一切形式种族歧视国际公约》（ICERD），该公约提出了比《公民及政治权利国际公约》第二十条更严格的条款，即对“鼓吹仇恨”者惩处而不问意图，并将传播仇恨的行为纳入了应受惩处的做法之列。这一领域的相关文件还有《防止及惩治灭绝种族罪公约》（CPPCG）、《消除对妇女一切形式歧视公约》（CEDAW）等。

在与联合国官员以及学术界和民间社会专家讨论的基础上，言论自由组织ARTICLE 19制定了《关于表达自由和平等的卡姆登原则》（Camden Principles）。这些原则为《公民及政治权利国际公约》条款提供了解释性指南，并力求通过详细说明与“煽动”相关的问题以及“歧视”、“敌视”和“强暴”的构成要素，阻止行为者滥用第二十条。

[1] 危险言论项目，2021年, https://dangerousspeech.org/
[2] 苏珊·贝尼希（2021年）《危险言论：实用指南》。危险言论项目，2021年，第7页。 https://dangerousspeech.org/

在将国际法和国际原则转化为国家法律时，各国对仇恨言论的定义方式略有不同，包括仇恨言论的表达方式、潜在的目标、以及产生何种危害才能将该言论本身视为仇恨言论。在抵御网络仇恨言论时，缺乏统一的定义是主要挑战之一，而这又不一定受国界的限制。

仇恨言论的定义对于研究和宣传工作也很重要，尤其是在确定其社会后果时。仇恨言论造成的危害可能是在个人层面（以心理伤害的形式）、群体和社区以及社会层面（以侵蚀权利和公共财物的形式）。由于仇恨言论所针对的是具有某种群体特征的人群，因此在社区危害层面进行分析尤为重要。仇恨言论所造成的危害在一般人群中分布不均，但边缘化群体首当其冲。对于身受其害的人们来说，危害是累积性的，之前所经历的仇恨言论是评估仇恨言论所造成危害的关键性可变因素。[3]

网络仇恨言论

网络仇恨言论与线下仇恨言论没有本质区别。但是不同之处在于，网络仇恨言论发生/出现时具有互动性质，并且在特定词语、指摘和阴谋论的使用和传播方面有所不同，这些内容可以非常迅速地发酵、达到顶峰、然后消失。仇恨信息可能会在数小时甚至数分钟内像病毒般迅疾传播开来。

教科文组织2015年《反击网络仇恨言论》报告指出，网络仇恨言论可以低成本制作和传播，不像其他书面作品那样经过编辑过程，其接受情况可以视帖子的受欢迎程度而大不相同，并且还可以跨国发布，因为平台服务器和总部并不需要与用户及其目标受众处在同一个国家。网络仇恨言论还可以持续更长时间，经历数轮发酵，连接到新网络或死灰复燃，以及匿名发布。因此，谁来管理网络空间、是否及何时应该删除内容，一直存在着广泛的争论。

这场争论以德国《**网络执法法案**》（**NetzDG**）等法律为代表。该法案于2017年发布，要求拥有超过200万用户的社交媒体平台实施透明程序，管控非法内容（包括仇恨言论），在24小时内删除被认定为非法的内容，并定期报告所采取的措施。该法案受到了严厉批评，因为它推动平台扮演“私有化审查”的角色，去做本应由法院做出的决定，同时还有人警告说法案规定的时间限制和罚款将导致平台“过度删除”内容以避免高额罚款。2020年，该法案进行了修订，要求社交媒体平台将识别出的非法内容转发给联邦刑事警察局。另一项几乎同时进行的修订要求平台更加方便用户举报非法内容，并允许就是否删除帖子的裁决提出上诉，从而加强了用户的权利。

在这种情况下，制定应对线上线下仇恨言论的法律，往往充满着与定义方面的挑战、完成在法律框架内尊重言论自由的任务相关的复杂审查过程。鉴于这些挑战，还必须采用法律措施以外的方法应对仇恨言论。

[3] 凯瑟琳·盖伯（Katharine Gelber）与卢克·麦克纳马拉（Luke McNamara）（2016年）“仇恨言论危害证明”（Evidencing the harms of hate speech），Social Identities，22:3

2
衡量和监测仇恨言论的工具和技术

有关网络仇恨言论的检测、监测和审核的政策和工具因环境、行为者和平台而异。

检测方法可以大致分为两类：最初依赖于关键词过滤器和群众外包的较全面检测方法，以及依赖于人类内容版主的检测方法（这些版主审查已被用户标记为仇恨言论的内容并决定其是否可以归类）。人工方法虽然具有捕捉上下文并快速应对新情况的独特优势，但是这项工作属于劳动密集型，耗时且昂贵，对可扩展性和快速解决方案构成了限制。由于出现了这些挑战，由于社交媒体上的内容量不断增加，也由于机器学习和自然语言处理技术方面的进展，各个平台和研究人员开发出了各种自动化检测解决方案，而且越来越依赖于这些新方案。许多新的举措合并使用了多种方法。与这些方法相关的一些关键术语包括：

- **机器学习：**利用计算机算法的技术，可以通过经验和数据的使用进行自动改进。
- **自然语言处理：**处理和分析大量自然语言数据的技术。
- **基于关键词的方法：**使用本体或字典识别包含潜在仇恨关键词的文本的方法。
- **分布语义：**根据词语、短语和句子在大样本数据中的分布情况，对其相似性进行量化和分类的方法。
- **情感分析：**对特定文本中就某一主题所表达的态度进行解释的方法。
- **源元数据：**一些方法通过数据的元信息为模型提供信息，如与消息相关联的用户数据，包括关注者数量等基于网络的特征。
- **深度学习：**采用多层处理方式从原始输入数据中逐步提取更高级别特征的一类机器学习算法。

社交媒体公司已经发生了很大的转变，过去是对用户标记为仇恨言论的帖子做出反应，而现在是在用户看到此类内容之前即通过公司的自动化系统就主动检测和处理此类内容。虽然这些方法对于大规模处理仇恨言论是必要的，但是也会产生一些复杂的影响：仇恨言论自动检测不可避免会出错，并且也可能会删除非仇恨内容。过度删除内容可能会引起寒蝉效应，破坏言论自由。

这些公司为了改进监测工作，正在不断开发仇恨言论检测工具。例如，Perspective**应用程序编程接口**[4] 是Jigsaw（谷歌内部孵化公司）和谷歌反滥用技术团队开发的开源工具，已用于各新闻机构及谷歌产品。该工具利用机器学习，根据对话中的潜在危害性对短语进行评分。已经有七种语言的版本（英文、法文、德文、意大利文、葡萄牙文、西班牙文及俄语）可供使用。**脸书**表示，其自身平台上仇恨言论监测和检测工具的最新版本改进了对语言的语义理解和对内容的理解，包括对图像、评论及其他元素的分析。[5]

研究人员和民间社会组织也努力开发出了各种仇恨言论检测工具。其中包括：

- 肯尼亚的**Umati**平台是肯尼亚最早用人工监测以相关语言撰写的在线帖子的平台之一。
- 戴维森（Davidson）等人（2017年）开发了**HateSonar**，使用了以网络论坛和推特的数据为基础进行训练的逻辑回归方法。
- 美国反诽谤联盟（ADL）和加利福尼亚大学伯克利分校D-Lab开发了“网络仇恨指数”（OHI），旨在通过机器学习将人类对仇恨言论的理解转变为可部署在互联网内容上的可扩展工具，以发现网络仇恨言论的范围和传播情况。
- 阿兰·图灵研究所“措施与对策”（Measures & Counter-measures）项目团队开发了一种使用深度学习方法的工具，用于检测社交媒体上针对东亚的偏见。
- 文（Moon）等人（2020年）开发了一种韩国仇恨言论检测工具，除了“仇恨”之外，他们还使用“偏见”标签来训练该模型。
- **Hatemeter**使用机器学习和自然语言处理技术检测反穆斯林仇恨言论。该平台提供英文、法文及意大利文版本。
- **COSMOS**使用情感分析和自然语言处理技术，通过关键词规范实时收集和分析来自推特的数据。
- **MANDOLA**把情感分析、自然语言处理、机器学习和深度学习方法结合起来，检测仇恨内容。

[4] 有关Perspective应用程序编程接口的更多信息，请访问 https://www.perspectiveapi.com/
[5] 来源 https://ai.facebook.com/blog/ai-advances-to-better-detect-hate-speech/

需要注意的是，网络仇恨言论的监测取决于对数据的访问，尤其是来自社交媒体平台的数据。此外，当今很多现有工具都采用单语形式且通常仅限于英文，因此需要对多语言检测方法的性能进行更多研究。此外，对社交媒体平台上仇恨言论的绝大多数研究和监测都集中在美国和欧洲，导致不仅在工具和数据方面存在差距，而且在对仇恨言论在其他地区传播的程度和动态的理解方面也存在差距。鉴于仇恨言论的背景性质，弥合这一差距更为重要。

社交媒体平台上仇恨言论的流行程度

推特、脸书、Instagram及YouTube采用基于当今可用的各种方法的自动检测工具，因此报告的标记内容和/或删除内容日益增多。2021年1—3月期间，YouTube删除了85 247个违反其仇恨言论政策的视频。YouTube之前的两份报告也发布了类似数据。脸书同一季度报告经处理的内容总数为2 520万条，而Instagram报告的内容为630万条。推特最新的透明度报告称，该公司在2020年7—12月期间删除了1 628 281条被视为违反其仇恨言论政策的内容。

在社交媒体平台上，仇恨言论的流行程度是通过用户查看的内容样本来确定的。换句话说，它仅捕捉平台上（估计）剩余的仇恨言论，不包括该公司已主动检测和删除的仇恨言论。迄今为止，脸书是报告流行度指标的唯一平台。该公司报告称，在2021年1—3月期间，仇恨言论的流行度为0.05%—0.06%，与之前的两份报告相比略有下降。部分研究表明，主流平台（如推特和维基百科）上仇恨言论的流行度不到总内容的1%，而在较小众的替代性平台（如Gab和4chan）上，托管内容的5%—8%可能具有辱骂性质。 [6]关于社交媒体平台上仇恨言论流行度的证据仍然不太完整，部分原因是平台缺乏透明度和数据访问权限。

[6] 赞内图（Zannettou）等，2018年；马修（Mathew）等，2018年；海因（Hine）等，2017年，维德根（Vidgen）等，2019年。

3
反击网络仇恨言论

首先值得强调的是，反击仇恨言论——并进而选择适当的工具和策略及预防措施——因多种因素而变得复杂。在一些背景下，不同行为者对主要问题给出的答案几乎没有共识。仇恨言论如何造成危害以及危害何时严重到需要对言论进行监管？更详细地说，与哪些仇恨言论行为相关的哪些类型的危害需要根据国际人权法和言论自由准则进行监管？

互联网架构也使反击仇恨言论明显更具挑战性。其中包括网络内容的永久性、流动性、匿名性和跨辖区特征、平台架构的多样性以及涉及多利益攸关方的多元化互联网治理系统。

尽管面临这些挑战，许多团体和个人仍以各种方式反击网络仇恨言论，并采取预防性措施，加强网络用户对仇恨言论的抵御能力。

国家法律追索权

反击仇恨言论的一个重要途径是法律追索权。尽管对仇恨言论和网络仇恨言论的立场因地区而异，并且这种立场随着对这一问题加深理解而不断演变，但是如本文件开头所述，依然有许多国际原则、地区协定、国家层面的法律以及与国际人权准则一致的法理学示例包含了与线上线下仇恨言论相关的条款。

然而，很快出现了认为网络仇恨言论法律过于严苛的质疑。其中包括对权利均衡的疑虑、当权者对权利滥施限制的可能性、依赖门槛制止煽动暴力，以及对仇恨言论与线下暴力之间的关系知之甚少，等等。更重要的是，对于反击网络仇恨言论，法律追索权的一个关键问题是各国政府对网络数字空间的权力有限。有效应对网络仇恨言论不能仅仅依靠国家法律追索权。[7]

2016年，多家大型科技公司就欧盟委员会的《反击网络非法仇恨言论行为守则》（*Code of Conduct on Countering Illegal Hate Speech Online*）达成一致，该准则要求这些公司在收到报告后的一天内审查仇恨言论。由于服务条件和仇恨言论的可操作定义均存在很大差异，这种方法具有挑战性，但这是在仇恨言论领域加强合作以及将法律方法与法外方法结合起来的一项意义重大的工作。

科技公司的回应

2021年，YouTube和脸书均报告称本公司发现和标记的内容有所增加，与用户标记的内容相比，公司标记的内容比例更高。这是因为自动检测系统的使用越来越多。但是，与之前的报告期相比，这一趋势伴随着恢复内容的增加。2021年1—3月期间，脸书恢复了408 700条内容，Instagram恢复了43 700条。虽然报告表明平台处理的仇恨内容日益增加，但是我们并不知道这是随之增加的滥用情况、更加严格的平台政策、还是误报增加而导致的。

社交媒体公司位于国家司法管辖范围，受国家法律的直接影响，因此通常更能响应遏制仇恨言论的要求。但是，社交媒体平台不受地域限制，因此仅能依赖其恪守自身的服务条款，而这些条款与上一节中概述的国际协定规定的准则相比，或许更严格，或许不那么严格。社交媒体平台采取的行动包括删除被判定为仇恨言论的材料，以及向发布仇恨言论的用户发送警告，限制或禁止他们在平台上的活动。这些社区标准在不断变化，尤其是在它们对自动化审核方法与人工审核方法的依赖程度方面。

鉴于这些挑战，一场多利益攸关方运动呼吁提高互联网公司透明度，作为加强其问责制的一种手段，而近几年，这一呼吁势头日益强劲。至少30个国家和地区拟

[7] 自2013年以来，教科文组织的“法官倡议”在言论自由、信息获取和世界各地记者安全方面的国际和地区标准等方面提高了司法行为者的能力，特别是受到以下情况的推动：如何最好地处理仇恨言论案件这一问题是许多司法人员的主要利益点之一。23 000多名司法人员接受了有关这些问题的培训，特别是通过一系列慕课（MOOC）、实地培训和研讨会以及大量工具包和指南的出版。

议采取法律和监管措施，包括通过目前正在制定的《欧洲数字服务法案》（European Digital Services Act）。各家公司也采取了各种措施提高透明度。2021年，Access Now为70多家定期发布透明度报告的公司编制了索引，[8]但是所做的工作还远远不够。

教科文组织发布的《让阳光照进来：数字时代的透明度和问责制》（*Letting the Sun Shine In: Transparency and Accountability in the Digital Age*）简报，提出了将提高透明度作为国家对内容的过度监管（导致对人权限制过度）与因采取放任方式而不能有效解决有问题的内容（如仇恨言论和虚假信息）以外的第三种方式。该简报提供了全套26项高层次原则，涵盖了与内容和程序、尽职调查和补救、增强权能、商业层面、个人数据收集和使用以及数据获取相关的问题。

法外应对措施和预防性干预措施

其他法外应对措施来自民间社会的研究和倡导工作，或者可能侧重于加强网络用户对仇恨言论的抵御能力的预防性措施。这包括直接针对网络仇恨言论的起因和后果的举措（包括教育手段在内），以及呼吁实施更有效的法律措施和技术措施的倡议。

基于教育的举措是这些工作的核心，通常侧重于长期预防。教育干预措施有助于提高对仇恨言论有害后果的认识，从根源上解决问题，并有效警示用于线上线下传播仇恨的操纵技术和言辞。特别是，已在世界各地制定并实施了媒体与信息素养计划，旨在为网络用户提供批判性地检查网络内容信息并识别令人不安的仇恨内容和错误信息的技能。同样，还开展了旨在通过积极的反宣传应对仇恨言论的“反言论”（counterspeech）工作，如2017年脸书在德国、英国和法国主导开展的“线上公民勇气倡议”（Online Civil Courage Initiative）。其他民间社会的倡议侧重于倡导平台方面的变革。2020年7月，“停止以仇恨牟利”（Stop Hate for Profit）活动汇集了世界各地1 200多家公司加盟，呼吁抵制在主要平台上投放广告，要求对仇恨言论进行审核，并呼吁暂停在宣扬歧视某些群体的账户上投放广告。这项活动使呼吁应对网络仇恨言论的声势更加浩大（尤其是从冠状病毒病疫情期间越发针对边缘化群体的情况来看），促使一些社交媒体公司改变他们的社区指南。

[8] Access Now 网址： https://www.accessnow.org/transparency-reporting-index/

建议

为了给循证决策提供信息以遏制网络仇恨言论并防止仇恨言论转化为暴力，同时保护表达自由，必须对仇恨言论的趋势进行识别、监测、收集数据和分析，从而确定适当的应对策略。以下建议旨在确定关键行动，以应对新出现的病毒式仇恨言论带来的新挑战，尤其是解决这些仇恨言论给和平、稳定及人人享有人权造成的线下后果。

1. 促进为仇恨言论制定包容性定义，尊重言论自由

- 确保定义符合国际准则，尤其是《公民及政治权利国际公约》和《拉巴特行动计划》的规定。

2. 建立多利益攸关方联盟

- 鼓励在人权组织、互联网中介机构和公众之间共享数据和专门知识。
- 赋权利益攸关方，尤其是赋权当地社区，在专门根据其背景和语言定制的社交媒体上监测和检测仇恨言论。
- 围绕仇恨言论的趋势、发生情况以及反击方式，召集多利益攸关方对话。
- 倡导平台与专家群体和公众合作制定定义和操作程序，范围应扩展到北美和西欧以外的世界更多国家。

3. 收集数据并鼓励对已收集数据实行开放数据做法，同时尊重个人数据保护

- 收集仇恨言论针对个人的定性数据，以更好地了解危害的范围和性质。
- 倡导互联网平台公司改进其透明度做法，包括公开发布有关仇恨言论投诉和解决方案的数据以及有关其内容审核系统的准确性和运行的数据，尤其是以研究为目的的数据。
- 支持开发平价、无障碍、用户友好型工具和方法，用以在允许采取反制行动的时间框架内监测和检测多语言多文化背景下的仇恨言论。

4. 鼓励平台为内容已被删除的人员提供可行的补救措施

- 推动社交媒体公司与关注数字权利的民间社会团体之间的合作，以确保内容审核和删除流程符合社区需求。

5. 通过教育计划培养媒体与信息素养和数字技能

- 为制定抵御仇恨言论的能力培养教育计划提供资金和资源，了解当前仇恨言论的趋势，应对相关挑战。这需要社交媒体公司、研究机构与教育利益攸关方之间开展密切合作。
- 优先考虑警示网络仇恨言论有害影响的预防性教育方法，并在开展缓解和抵御工作的同时培养媒体与信息素养。
- 建立并支持教育机构与社交媒体公司之间的伙伴关系，通过有针对性的传播活动或将用户重新定向到外部资源，增加获取信息和资源的机会，以应对社交媒体平台上的仇恨言论。

6. 为在网络仇恨言论领域内工作的组织提供支持

- 支持并确保向致力于监测和反击仇恨言论的专业组织提供足够的资源，特别是那些最有能力考虑当地情况的组织。

本文是由教科文组织和联合国秘书长防止灭绝种族罪行特别顾问办公室（OSAPG）委托编写的讨论文件集的一部分。这些文件对《联合国战略和行动计划》做出了直接贡献，是在2021年9月和10月举行的通过教育应对仇恨言论多利益攸关方论坛和部长级会议的背景下发布的。

冠状病毒病疫情的爆发突出了《联合国战略和行动计划》的相关性，疫情在世界各地引发了仇恨言论浪潮——进一步加剧了对特定群体的不容忍和歧视，破坏了社会和政治制度的稳定。这些讨论文件试图解读与这一全球挑战相关的重要问题，并提出可能的应对措施和建议。

本文件由教科文组织表达自由与记者安全科委托撰写，是欧洲联盟资助的“#冠状病毒病真相：应对冲突易发环境下的冠状病毒病‘信息疫情’”（#CoronavirusFacts: Addressing the ‘Disinfodemic’ on COVID-19 in conflict-prone environments）项目的组成部分。文件由牛津大学互联网研究所的乔纳森·布莱特（Jonathan Bright）、安东内拉·佩里尼（Antonella Perini）、安妮·普洛因（Anne Ploin）和雷贾·维斯（Reja Wyss）起草。

联合国教育、科学及文化组织，丰特努瓦广场7号，75352 巴黎 07 SP，法国，2022年出版

原版书籍或期刊名称：*Addressing hate speech on social media: contemporary challenges*
联合国教育、科学及文化组织2021年出版

本文件的编写承蒙欧洲联盟提供财务支持。其内容完全由作者负责，不一定反映欧洲联盟的观点。

图案设计：Dean Dorat

CI/FEJ/2021/DP/01

由欧洲联盟资助