联合国 A/74/486

大 会

Distr.: General
9 October 2019
Chinese
Original: English

第七十四届会议

议程项目 70(b)

促进和保护人权：人权问题，包括增进人权和基本自由切实享受的各种途径

促进和保护意见和表达自由权*

秘书长的说明

秘书长谨向大会转递促进和保护意见和表达自由权特别报告员大卫·凯伊根据人权理事会第 34/18 号决议提交的报告。在本报告中，特别报告员评估了适用于管制网络“仇恨言论”的人权法。

* 本报告逾期提交，以反映最新动态。

19-17368 (C) 151019 251019

促进和保护意见和表达自由权特别报告员的报告

摘要

在对限制“仇恨言论”呼声越来越高的世界里，国际人权法为规范国家和公司处理网络表达的方式提供了标准。本报告根据人权理事会第 34/18 号决议提交，促进和保护意见和表达自由权特别报告员在报告中说明了这些标准如何为考虑监管方案的政府和决定如何在网上尊重人权的公司提供了框架。报告员首先介绍了国际法律框架，重点说明了联合国条约和与俗称的“仇恨言论”相关的条款的重要解释。他随后强调了国家的主要义务，并讨论了公司对内容的审核可如何确保尊重用户和公众的人权。最后，他向国家和公司提出了建议。

目录

一. 导言

1. “仇恨言论”这一缩略语在协定国际法中没有定义，具有双重的模糊性。这种含糊不清以及对其含义缺乏共识可能被滥用，导致大量合法表述受到侵蚀。许多政府像使用“假新闻”一样使用“仇恨言论”，来攻击政敌、不信仰者、持不同政见者和批评者。然而，这个短语的弱点(“这只是言论”)似乎也阻碍了政府和企业应对真正的危害，比如煽动对弱势群体的暴力或歧视的言论，或使边缘化群体噤声所造成的危害。这种情况让常常感到网络侵犯猖獗的公众感到沮丧。

2. 在对限制“仇恨言论”呼声越来越高的世界里，国际人权法提供了标准，以规范国家和公司处理网络表达的方式(A/HRC/38/35，第 45 段)。[1] 在本报告中，促进和保护意见和表达自由权特别报告员说明了这些标准如何为考虑监管方案的政府和决定如何在网上尊重人权的公司提供了框架。特别报告员首先介绍了国际法律框架，重点说明了联合国条约和与俗称的“仇恨言论”相关的条款的重要解释。他随后强调了国家的主要义务，并讨论了公司对内容的审核可如何确保尊重用户和公众的人权。最后，他为国家和公司提供了建议。

3. 本报告是自 2015 年发布的一系列报告中的第六份，特别报告员在这些报告中探讨了适用于信息和通信技术部门中意见和言论自由的人权标准。[2] 应结合先前提出的标准和建议阅读本报告，这些标准和建议在本报告中不一定重复说明。与以往的报告一样，本报告广泛借鉴了现有国际标准和过去几年中民间社会的大量投入。

二. 国际人权法中有关“仇恨言论”的规范

4. 根据国际人权法，仇恨言论的局限似乎要求调和两套价值观，即民主社会对允许公开辩论和个人自主与发展的要求，与防止攻击脆弱群体、确保所有人平等无歧视地参与公共生活这一同样紧迫的义务。[3] 各国政府经常利用由此产生的不确定性来威胁合法的表达，如政治异议和批评或宗教分歧。[4] 然而，表达自由、平等和生命权与不歧视义务是相辅相成的；人权法允许国家和企业专注于保护和促进所有人的言论，特别是那些权利往往受到威胁的人，同时也处理损害享受所有权利的公开和私人歧视问题。

---

[1] 本报告用“仇恨言论”一词指称没有使用这一特定术语的人权法中的义务和限制。见 Susan Benesch，“Proposals for improved regulation of harmful online content”，为以色列民主学会编写的论文，2019。Benesch 创造了一个类似术语“危险言论”，指“一个群体针对另一个群体催化暴力的能力”。另见 Susan Benesch，“Dangerous speech: a proposal to prevent group violence”，2012。

[2] 见 A/HRC/29/32 (加密和匿名)、A/HRC/32/38 (信通技术部门对权利的影响的调查分析)、A/HRC/35/22 (数字接入行业)、A/HRC/38/35 (在线内容审核)和 A/73/348 (人工智能与人权)。

[3] 特别见前特别报告员弗兰克·拉鲁关于仇恨言论的报告(A/67/357)。

[4] 同上，第 51 至 54 段。

表达自由

5. 《公民权利和政治权利国际公约》第十九条㈠保护持有主张而不受干涉的权利，而第十九条㈡保障表达自由权，即通过任何媒介寻求、接受和传递各种消息和思想的权利，而不论国界。许多其他全球和区域性条约明确保护表达自由。[5] 《公民权利和政治权利国际公约》的专家监督机构人权事务委员会强调，这些自由是“个人充分发展不可或缺的条件……是每个自由民主社会的基石。”这些自由“构成了充分享受各种其他人权的基础”。[6]

6. 由于表达自由是享有所有人权的根本，对其限制必须是例外情况，应符合狭义条件并受到严格监督。人权事务委员会强调，限制即使是正当的，也“不得危及权利本身”。[7] 《公民权利和政治权利国际公约》第十九条㈢阐明了限制的例外性质，承认各国只有在法律规定和为尊重“他人的权利或名誉，或保护国家安全、公共秩序、公共卫生或道德”所必需的情况下，才可根据第十九条㈡限制表达。这些都是狭义的例外情况(特别见 A/67/357 第 41 段和 A/HRC/29/32 第 32 至 35 段)，限制言论的当局应负责证明限制是合理的，而不是由发言者来证明他们有权利发表这样的言论。[8] 任何限制都必须满足三个条件：

(a) **合法性**：限制必须由精确、公开、透明的法律做出，避免向当局提供不受限制的自由裁量权，并对言论受到管制者给予适当通知。规则应受到公众意见和常规立法或行政程序的约束。程序性保障，特别是由独立法院或法庭保证的程序性保障，应该为权利提供保护；

(b) **正当性**：应证明限制是为了保护第十九条㈢中规定的一项或多项利益，即尊重他人的权利或名誉，或保护国家安全、公共秩序、公共卫生或道德；

(c) **必要性和相称性**：国家必须表明限制是保护合法利益所需要的，而且是实现所称目标的限制性最低的手段。人权事务委员会将这些条件称为“严格检验”，根据这些检验，施加限制的“目的仅限于明文规定的目的，并且必须与所指特定需要直接相关”。[9]

7. 各国往往声称对表达施加限制有适当目的，但未能证明其限制能够通过合法性或必要性和相称性的检验(见 A/71/373)。因此，在采用规则时必须严格、一秉诚意，并提供强有力、透明的监督。《公民权利和政治权利国际公约》第二条㈢(乙)

[5] 例如，见《消除一切形式种族歧视国际公约》，第五条；《儿童权利公约》，第十三条；《残疾人权利公约》，第二十一条；《保障所有移民工人及其家属权利国际公约》，第十三条；《美洲人权公约》，第十三条；《非洲人权和民族权宪章》，第九条；《欧洲人权公约》，第十条。

[6] 人权委员会，关于言论和表达自由的第 34 号一般性意见(2011)，第 2 和 4 段；另见同上，第 5 至 6 段。

[7] 同上，第 21 段。人权事务委员会澄清说，“限制不得损害权利的本质”，并补充“授权实行限制的法律必须使用精确的标准，对于实施限制者不能给予无限的权限”(人权委员会，关于行动自由的第 27 号一般性意见(1999)，第 13 段)。

[8] 人权委员会，第 34 号一般性意见(2011)，第 27 段。

[9] 同上，第 22 段。

规定，各国有义务确保为违反《公民权利和政治权利国际公约》的情形寻求补救的个人享有“由主管司法、行政或立法当局，或由国家法律制度规定的任何其他主管当局所确定的”权利(另见 A/HRC/22/17/Add.4，第 31 段)。

构成煽动的鼓吹仇恨

8. 《公民权利和政治权利国际公约》第二十条㈡规定，缔约国有义务以法律禁止“鼓吹构成煽动歧视、敌意或暴力的任何民族、种族或宗教仇恨”。各国没有义务将这种表达定为犯罪。前任特别报告员解释说，第二十条㈡涉及：(a) 鼓吹仇恨；(b) 构成煽动的鼓吹；(c) 可能导致歧视、敌意或暴力的煽动(A/67/357，第 43 段)。

9. 与第二十条㈡通过对民族、种族或宗教仇恨的关注所提供的保护相比，联合国人权标准为防止歧视提供了更广泛的保护。《公民权利和政治权利国际公约》第二条㈠保障“不分任何区别”的权利，第二十六条则明文规定“法律应禁止任何歧视并保证所有的人得到平等的和有效的保护，以免受任何理由的歧视”。国际标准确保提供保护，以免受基于种族、肤色、性别、语言、宗教、政治或其他见解、民族或社会出身、财产、出生或其他身份，包括土著血统或身份、残疾、移民或难民身份、性取向、性别身份或间性者身份的负面行为的影响。[10] 保护的范围随时间扩大，例如，年龄和白化病等其他类别现在也受到明确保护。鉴于保护范围在全世界范围内扩大，禁止煽动应被理解为适用于目前国际人权法所涵盖的更为广泛的类别。

10. 至关重要的一点是，根据第二十条㈡，只有其主张构成煽动的人，其言论才可被禁止。对于所鼓吹的仇恨不构成煽动歧视、敌意或暴力，例如，支持少数群体，甚至是对宗教信条或历史事件做出冒犯性解释的人，或为了报告目的或提高对该问题的意识而例举仇恨和煽动情形的人，不得根据第二十条(或人权法的任何其他规定)使其沉默。这种表达应受国家保护，即使国家不同意该表达或因此受到冒犯。[11] 在国际人权法中没有“质问者否决权”。[12]

11. 在《公民权利和政治权利国际公约》前一年通过的《消除一切形式种族歧视国际公约》呼吁各国“消除一切种族歧视或煽动种族歧视的行为”，同时适当考虑到其他受人权法保护的权利，包括言论自由(见《消除一切形式种族歧视国际公约》第四条和第五条)。该公约第四条规定缔约国除其他外有义务：(a)“宣告凡传播以种族优越或仇恨为根据的思想，煽动种族歧视，对任何种族或属于另一肤色或人种的人群实施强暴行为或煽动此种行为，概为犯罪行为，依法惩处”；(b) “宣告凡提倡和煽动种族歧视的组织以及相关有组织活动和其他一切

[10] 另见 Article 19，*“Hate Speech” Explained: A Toolkit* (London, 2015)，p. 14。关于网上暴力侵害妇女行为，见 A/HRC/38/47。

[11] 人权委员会，第 34 号一般性意见(2011)，第 11 段。

[12] 见 Evelyn M. Aswad, “To ban or not to ban blasphemous videos”, *Georgetown Journal of International Law*, vol 44, No. 4 (2013)。

宣传活动概为非法，应予禁止，同时确认凡参与此类组织或活动均属犯罪行为，应依法惩处”。

12. 《公民权利和政治权利国际公约》第二十条㈡和《消除一切形式种族歧视国际公约》第四条涉及特定类别的表达，通常被称为“仇恨言论”。[13] 与第十九条㈡相比，这些规定的措辞模糊不清。[14] 《公民权利和政治权利国际公约》第十九条㈡中的表达自由涉及由积极动词(寻求、接受、传递)体现、范围在最大程度上广泛(任何思想、不论国界、通过任何媒体)的权利，而第二十条㈡和第四条所规定的禁止情形，虽然范围比一般性的“仇恨言论”要狭窄得多，但却使用了难以界定的情绪性语言(仇恨、敌意)和取决于具体情形的禁止(鼓吹煽动)。人权事务委员会决定，《公民权利和政治权利国际公约》第十九条和第二十条“互相兼容，互为补充”。[15] 即便如此，这些条款仍要求做出解释。

13. 人权事务委员会在第 34 号一般性意见(2011 年)中指出，在国家限制言论自由的任何情况下，包括第二十条㈡所界定的言论自由，仍然必须“证明限制及其规定完全符合第十九条。”[16] 2013 年，在联合国人权事务高级专员主持下召集的一个高级别人权问题专家组通过了第二十条㈡的解释。[17] 在《关于禁止构成煽动歧视、敌意或暴力的鼓吹民族、种族或宗教仇恨言论的拉巴特行动计划》中，主要术语定义如下：

> “‘仇恨’和‘敌意’是指对目标群体的谴责、敌意、憎恨等强烈的非理性情绪；‘鼓吹’一词应被理解为有意公开宣扬对目标群体的仇恨；‘煽动’一词是指关于民族、种族或宗教团体的言论，这些言论对属于这些团体的人造成迫在眉睫的歧视、敌意或暴力风险(A/HRC/22/17/Add.4，附录，脚注 5)。”[18]

14. 《拉巴特行动计划》共确定了六个因素，以决定将煽动行为定为刑事犯罪所需达到的严重程度(同上，第 29 段)：

(a) “发表和传播言论时普遍存在的社会和政治背景”；

(b) 发表言论者的地位，“特别是个人或组织相对于言论所针对者的地位”；

(c) 意图，即“过失或鲁莽不足以构成《公民权利和政治权利国际公约》第二十条下的犯罪”，规定仅仅分发或传播并不等同于鼓吹或煽动；

---

[13] 见 Jeremy Waldron, *The Harm in Hate Speech* (Harvard University Press, 2012)。

[14] 考虑到充满矛盾的历史，这种模糊不清并不令人惊讶。见 Jacob Mchangama, “The sordid origin of hate-speech laws”, *Policy Review* (December 2011 and January 2012)。

[15] 人权委员会，第 34 号一般性意见(2011)，第 50 段。

[16] 同上，第 52 段，特别是就《公民权利和政治权利国际公约》第二十条㈡而言，见第 50 段。

[17] 例如，见消除种族歧视委员会，关于打击种族仇恨言论的第 35 号一般性意见(2013)。

[18] 前特别报告员弗兰克·拉鲁认为评估煽动时的一个关键因素是“有关表达是否会导致真正和紧迫的暴力危险”(A/67/357，第 46 段)。另见 Article 19, *Prohibiting Incitement to Discrimination, Hostility or Violence* (London, 2012), pp. 24-25。

(d) 言论的内容和形式，特别是“讲话的挑衅和直接程度，以及所使用论述的形式、风格和性质”；

(e) 言语行为的范围或影响，例如“其受众的数量和规模”，包括是否“是一页传单，还是通过主流媒体或互联网进行的广播，言语的频率、数量和程度，受众是否有手段就煽动采取行动”；

(f) 造成以下情形的可能性(包括紧迫程度)，即使得“一定程度的伤害风险必须得到确定”，包括(按照《行动计划》的建议，通过法院)确定“言论将成功煽动针对目标群体采取实际行动的合理可能性”。

15. 2013 年，消除种族歧视委员会，即《消除一切形式种族歧视国际公约》的专家监督机构，仿效了人权委员会和《拉巴特行动计划》的做法。委员会澄清了《消除一切形式种族歧视国际公约》第四条的“适当注意”措辞，即要求严格遵守言论自由保障。[19] 消除种族歧视委员会强调(这表明有关解释趋同)，第四条所规定的刑事定罪在某些情况下应予保留，具体如下：

> “对一些种族歧视表达形式的刑事定罪应保留给证明确凿无疑的严重情况，不太严重的情况，应通过刑法以外的手段处理，除其他外，应考虑到对所针对的个人和群体所产生影响的性质和程度。刑罚的采用应符合合法性、相称性和必要性原则。”[20]

16. 消除种族歧视委员会解释说，《公民权利和政治权利国际公约》第十九条界定的条件也适用于《消除一切形式种族歧视国际公约》第四条的限制。[21] 消除种族歧视委员会指出，在将传播和煽动界定为应受法律惩罚的犯罪方面，各国必须在确定特定言论是否属于应禁止类别时考虑到一系列因素，包括言论的“内容和形式”，发表言论期间的“经济、社会和政治环境”，“发表言论者的地位或身份”，“言论的影响范围”及其目标。委员会建议《消除一切形式种族歧视国际公约》的缔约国考虑“有关言论可能造成发表言论者所希望或意图发生的行为的紧迫风险”。[22]

17. 消除种族歧视委员会还指出，《消除一切形式种族歧视国际公约》要求禁止“侮辱、嘲笑或诽谤个人或群体，或为仇恨、蔑视或歧视行为辩护”，强调这种表达只有在“明显等同于煽动仇恨或歧视”的情况下才可被禁止。[23] “嘲笑”和“辩护”是极其宽泛的用语，一般不受国际人权法的限制，国际人权法保护冒犯和嘲弄的权利。因此，与煽动和《公民权利和政治权利国际公约》第十九条㈢所建立框架的联系有助于将这种禁止限制在最严重的类别之内。

---

[19] 消除种族歧视委员会第 35 号一般性意见(2013)，第 19 段。消除种族歧视委员会认为，适当注意条款在言论自由方面具有尤其重要的意义，认为这是“在衡量对言论自由限制的合法性时最具相关性的参考原则”。

[20] 消除种族歧视委员会第 35 号一般性意见(2013)，第 12 段。

[21] 同上，第 4 段和第 19 至 20 段。

[22] 同上，第 15 至 16 段。

[23] 同上，第 13 段。

18. 《拉巴特行动计划》还澄清，刑事定罪应仅适用于《公民权利和政治权利国际公约》第二十条㈡规定的最严重的煽动行为，一般情况下，应首先考虑其他方法(A/HRC/22/17/Add.4，附录，第 34 段)。这些方法包括社会领导人反对仇恨言论、促进容忍和社区间尊重的公开声明；教育和文化间对话；使人们更容易获得反击仇恨讯息的信息和思想；促进人权原则和标准并开展培训。承认除法律禁止以外的措施突出表明，对于应对仇恨言论问题的国家来说，禁止往往不是限制性最低的措施。

可能不构成鼓吹或煽动的仇恨言论

19. 其他类型的言论可能不符合第二十条㈡或第四条的定义或门槛，但涉及到鼓吹仇恨等问题。这就引起了国家是否可以限制不构成煽动歧视、敌意或暴力的“鼓吹仇恨”的问题。换言之，他们是否能在“仇恨言论”被定义为“基于个人或群体的身份，换言之，基于他们的宗教、族裔、国籍、种族、肤色、血统、性别或其他身份因素，攻击或使用贬低或歧视性语言”的言论时(如联合国消除仇恨言论战略和行动计划小组最近所定义的那样)，对其加以限制？[24] 显然，这种语言缺乏第二十条㈡和第四条中的煽动含义，尽管国家和企业应该通过教育、谴责和其他手段来打击这种态度，但施加法律限制需要满足国际人权法中的严格标准。

20. 对于涉及联合国仇恨言论战略所界定的言论，即具有仇恨性但不构成煽动的言论的内容，《公民权利和政治权利国际公约》第十九条㈢提供了适当指导。其条件必须加以严格适用，以使任何限制——以及针对言论采取的任何行动——都符合合法性、必要性、相称性以及正当性的条件。鉴于其模糊性，联合国仇恨言论战略使用的措辞如果意在指导根据法律做出的禁令，那么在合法性方面将存在问题，尽管可以将其作为打击歧视和仇恨的政治和社会行动的依据。任何采用这种定义的国家也需要为限制提供合法的限制理由。在大多数情况下，第十九条㈢所界定的他人的权利可能是适当的依据，着重关注与歧视或干涉隐私，或保护公共秩序有关的权利。但在每一种情况下，国家仍必须证明采取行动的必要性和相称性，惩罚越严厉，就越需要证明在严格意义上的必要性。[25]

21. 根据国际人权标准，某些限制是特别不受欢迎的。人权事务委员会首先举例指出，“禁止不尊重宗教或其他信仰体系的表现，包括亵渎宗教法不符合《公约》”，除非在亵渎神灵也可被定义为鼓吹宗教仇恨，并构成满足条件的煽动的情况下。[26] 明确地说，鉴于第十九条保护个人及其表达和意见自由的权利，反亵渎神灵法律

[24] 在《拉巴特行动计划》中，提到了低于《公民权利和政治权利国际公约》第二十条㈡所设定门槛的言论，但这种言论或者“可能有理由对其提起民事诉讼或做出行政处罚”，或者不需要做出处罚，但“仍然引起对容忍、礼貌和尊重他人权利问题的关切”(A/HRC/22/17/Add.4，第 20 段)。

[25] 《公民权利和政治权利国际公约》第十九条㈢的公共道德例外不太可能成为依据，但值得注意的是，人权事务委员会已经澄清，“为了保护道德的限制必需基于不光是来自单一传统的原则”(人权委员会，第 34 号一般性意见(2011)，第 32 段，引用人权委员会关于思想、良心和宗教自由权的第 22 号一般性意见(1993)，第 8 段)。

[26] 人权委员会，第 34 号一般性意见(2011)，第 48 段。在这种情况下，亵渎并不重要；只有构成煽动的鼓吹才有相关性。

不符合《公民权利和政治权利国际公约》第十九条㈢的合法性条件；该条和《公民权利和政治权利国际公约》第十八条都没有保护思想或信仰免受嘲笑、谩骂、批评或其他被视为冒犯性的“攻击”。几个人权机制确认了废除亵渎神灵法的呼吁，因为这些法律可能造成对宗教思想的辩论，并且会发挥作用，使政府能倾向于一种宗教思想，而不是其他宗教、信仰或非信仰系统(特别见 A/HRC/31/18，第 59 至 61 段)。

22. 第二，“惩罚对历史事实发表见解”的法律与《公民权利和政治权利国际公约》第十九条不符，将否认大屠杀和其他暴行定为刑事犯罪的法律及类似法律受到质疑，这些法律常常援引“仇恨言论”作为依据。人权事务委员会指出，“错误的”和“对过去事件进行不正确解释”的意见不可受到一般性禁止，对表达这种意见的任何限制“不应超出《公民权利和政治权利国际公约》第十九条㈢所允许的范围”或“第二十条所要求的”。[27] 根据这些解释和其他解释，在没有根据上述定义和背景进行进一步评估的情况下，否认暴行的历史准确性不应受到刑事处罚或其他限制。根据国际人权法实施任何此类限制都应评估《拉巴特行动计划》中提到的六个因素。

23. 第三种并非煽动的言论可能涉及这样一种情况，即发表言论者“单独针对某个可识别的受害者”，但不寻求“煽动他人基于受保护的特征对他人采取行动”。[28] 同样，参照《公民权利和政治权利国际公约》第十九条㈢，这种言论可能会受到限制，以保护他人的权利或保护公共秩序。通常，国家在“仇恨犯罪”的一般类别下限制这种表达——对人身攻击或财产攻击的惩罚因其背后的仇恨动机而加重。

24. 第四，需要重点强调的是，可能具有攻击性或带有偏见、引起对不容忍严重关切的表达，往往不会达到足以受到任何限制的严重程度。存在各种仇恨性的表达，尽管是丑恶的，但并不涉及煽动或直接威胁，例如表明对受保护群体的偏见。这种情绪不会受到《公民权利和政治权利国际公约》或《消除一切形式种族歧视国际公约》的禁止，而采取其他限制或不利行动则要求分析第十九条㈢中的条件。《拉巴特行动计划》确定的将煽动定为犯罪的六个因素也将为考虑如何评估公共机关对这种言论的反应提供有价值的基准。的确，没有限制并不意味着没有行动；各国可以(并且应该按照人权理事会第 16/18 号决议)采取强有力的措施，如由政府谴责偏见、教育、培训、公共服务公告、社区项目，打击这种不容忍，确保公共当局保护个人不受基于此类仇恨主张的歧视。

25. 最后，《防止及惩治灭绝种族罪公约》要求各国将煽动种族灭绝定为犯罪。在某些情况下，例如在缅甸，国家对煽动种族灭绝不采取行动可能会对脆弱社区造

[27] 人权委员会，第 34 号一般性意见(2011)，第 49 段。见 Sarah Cleveland, *Hate Speech at Home and Abroad*, in Lee C. Bollinger and Geoffrey R. Stone, eds., *The Free Speech Century* (New York, Oxford University Press, 2019)。另见 A/67/357，第 55 段。

[28] Article 19，“Hate Speech” Explained，p. 22。

成非常严重的后果。这种不作为本身应受到谴责，正如煽动本身必须受到反对和惩罚一样。[29]

区域一级的人权规范

26. 欧洲、美洲和非洲人权体系也阐明了与“仇恨言论”有关的标准。欧洲人权法院强调，言论自由保护可能会“冒犯、震惊或扰乱”的各种言论。[30] 然而，法院对那些以禁止“仇恨言论”为由，继续通过法律禁止亵渎神灵或继续将否认种族灭绝定为刑事犯罪的国家采取了相对尊重的态度，这与在全球层面观察到的趋势相反。[31] 法院通常会完全回避“仇恨言论”问题，不是以表达自由为依据，而是以“践踏权利”为依据，认为对有关侵犯行为的指控不可受理。[32] 在要求中间方对其平台上发表的仇恨言论承担责任方面，欧洲的规范可能处于变化之中。[33] 相比之下，美洲人权委员会的标准往往更接近上述国际标准，而非洲系统的标准则处于相对早期的阶段。[34] 在任何情况下，都不能援引区域人权规范来为背离国际人权保护标准提供理由。

27. 人权事务委员会特别否定了欧洲法院的判断余地原则，指出“缔约国在任何特定情况下均须具体表明导致其限制言论自由的对第3款所列任何理由的确切威胁性”。[35] 人权事务委员会不会仅仅因为国家当局声称他们通常更了解当地情况而给予国家酌处权。

联合国仇恨言论文书总结

28. 近年来，国际人权框架不断发展，使表面上看起来相互矛盾的规范合理化。简言之，表达自由是民主社会最有价值的一项法律权利，与整体人权法中其他权利相互依存并支持其他权利。同时，反歧视、平等以及平等和有效的公众参与奠定了整体人权法的基础。《公民权利和政治权利国际公约》第二十条和《消除一切形式种族歧视国际公约》第四条所规定的表达方式对这两套规范提出了挑战，这

[29] 特别见 A/HRC/39/64，第 73 段。《防止及惩治灭绝种族罪公约》第三条(c)要求将“直接和公开煽动实施种族灭绝”定为刑事犯罪。

[30] 欧洲人权法院，*Handyside* 诉联合王国，申请编号 5493/72，1976 年 12 月 7 日的判决，第 49 段。见 Sejal Parmer，“The legal framework for addressing ‘hate speech’ in Europe”，在处理媒体中仇恨言论国际会议上的发言，萨格勒布，2018 年 11 月。

[31] 见欧洲委员会，“Hate speech”, fact sheet, October 2019；和 Evelyn M. Aswad, “The future of freedom of expression online”，*Duke Law and Technology Review*, vol. 17 (August 2018)。

[32] 关于实践的概述，见欧洲委员会，“Guide on article 17 of the European Convention on Human Rights: prohibition of abuse of rights”, 2019 年 8 月 31 日更新。

[33] 比较欧洲人权法院大法庭，*Delfi AS* 诉爱沙尼亚，申请编号 64569/09，2015 年 6 月 16 日的判决，与欧洲人权法院，第四分庭，*Magyar Tartalomszolgáltatók Egyesülete* 和 *Index.hu Zrt* 诉匈牙利，申请编号 22947/13，2016 年 2 月 2 日的判决。另见 Article 19，“Responding to ‘hate speech’: comparative overview of six EU countries”，2018。

[34] 见美洲人权委员会，“Hate speech and incitement to violence”，美洲人权委员会，*Violence against Lesbian, Gay, Bisexual, Trans and Intersex Persons in the Americas* (2015)。

[35] 人权委员会，第 34 号一般性意见(2011)，第 36 段。

是公共生活的所有参与者都必须承认的。因此，对表达自由权的限制必须是例外情况，国家有责任证明这种限制符合国际法；根据《公民权利和政治权利国际公约》第二十条和《消除一切形式种族歧视国际公约》第四条作出的禁止必须符合《公民权利和政治权利国际公约》第十九条㈢的严格和狭义条件；各国一般上应使用除定罪和禁止之外可利用的手段，如教育、驳斥言论、促进多元化，来处理各种“仇恨言论”。

三. 治理网络仇恨言论

A. 国家义务与网络仇恨言论的治理

29. 严格遵守国际人权法标准可以防止政府的过度行为。作为首要原则，各国不应将互联网公司用作限制言论自由的工具，根据国际人权法，国家本身将无法做出这些限制。他们对公司的要求，无论是通过监管还是通过威胁使用监管，都必须在国际法下是正当的并与国际法相一致。针对内容采取的某些类型的行动显然不符合《公民权利和政治权利国际公约》第十九条㈢，例如关闭互联网，将网络上的政治异议或对政府的批评定为犯罪(见 A/HRC/35/22)。对发表非法仇恨言论者的惩罚不应仅仅因为言论是在网络上发表而加重。

30. 可以假设有这样一个国家，正在考虑颁布立法，要求网络中间方对未能对“仇恨言论”采取特定行动承担责任。这种“中间方责任”法的一般目的是限制表达，无论是特定平台的用户还是平台本身的表达，有时旨在履行《公民权利和政治权利国际公约》第二十条㈡所规定的义务。对这种提案的任何法律评估都必须考虑第十九条㈢规定的累积条件，以确保符合自由表达方面的国际标准。[36]

合法性

31. 《公民权利和政治权利国际公约》第十九条㈢要求在对发布“仇恨言论”追究责任时，这一词语本身和确定发生了仇恨言论所涉及的因素都必须得到界定。在提出对未能移除“煽动”内容追究责任时，必须根据《公民权利和政治权利国际公约》第二十条㈡和《消除一切形式种族歧视国际公约》第四条界定这种煽动的内容，包括界定上述《拉巴特行动计划》中提到的关键术语。如果一国希望以《公民权利和政治权利国际公约》第二十条和《消除一切形式种族歧视国际公约》第四条规定以外的理由管制仇恨言论，则必须界定有关内容实际上为非法；[37]《公民权利和政治权利国际公约》第十九条㈢要求的精确性和明确性意味着国家法律应限制政府行为体执行规则时或私人行为体利用规则压制合法表达的过度自由

[36] 关于在中间方责任方面可适用原则的说明，见 Electronic Frontier Foundation，“Manila principles on intermediary liability”，2015。

[37] 各国已在很大程度上将恐怖主义和“极端主义”内容与“仇恨言论”区分开来，但同样的合法性原则也必须适用于这些主题。例如，见 A/HRC/40/52，第 75(e)段。“极端主义”一词经常被用来作为“仇恨言论”的代名词，尽管这不具备法律依据。“暴力极端主义”对增加清晰度没有任何助益。在网络环境中秉持诚意使用“极端主义”一语的政府似乎关注的是“恐怖主义和暴力极端主义意识形态”疯狂传播的问题，目标似乎是打击“极端主义”言论和“防止滥用互联网”（消除网络上恐怖主义和暴力极端主义内容的克赖斯特彻奇呼吁)。

裁量权，并必须规定应向个人发出适当通知，以规范其活动。[38] 如果定义缺乏明确性和精确性，就存在滥用、限制合法内容和无法解决相关问题的重大风险。处理“仇恨言论”的国家应将其定义与国际人权法的标准紧密联系起来，例如《公民权利和政治权利国际公约》第二十条㈡所规定的标准。

32. 若干国家已经通过或正在考虑采用规则，要求互联网公司在特定时间内删除“明显非法”的言论，通常是 24 小时，甚至短至 1 小时，或者在较长时间内删除其他非法内容。这些法律中最著名的是德国的《网络执行法》，要求公司从其平台上删除根据《德国刑法》的一些具体规定属于非法的言论。[39] 例如，《刑法》第 130 条规定，除其他外，应处罚“能够以某种方式扰乱公共和平，煽动对民族、种族、宗教团体或由族裔血统界定的团体的仇恨，由于部分人口或个人属于上述群体之一而煽动对他们的仇恨，或呼吁对他们采取暴力或任意措施”的人。[40] 法律显然没有定义其关键术语(特别是“煽动”和“仇恨”)，[41] 但通过《网络执行法》，对未能遵守这些条款的公司处以高额罚款。所依据的法律含糊不清，存在问题。虽然《网络执行法》应被理解为秉持诚意的努力，以应对网络仇恨及其线下后果造成的广泛关切，但未能定义这些关键术语对该法不利，无法证明其要求符合国际人权法。

33. 很少有国家让本国法院参与评估不符合当地法律的平台仇恨言论的过程，但应允许仅根据独立法院的命令要求承担责任，并提供应中间方或受行动影响的其他方面(如主体用户)的请求提出上诉的可能性。[42] 政府在增加对公司的压力，要求公司充当仇恨言论的裁判者。通过过程还应遵循严格的法治标准，提供充分机会听取公众意见，并评估备选方案和对人权的影响。[43]

必要性和相称性

34. 为鼓励消除网络仇恨言论并对未能这样做的互联网公司追究责任而开展的立法努力必须满足上述必要性和相称性标准。近年来各国一直在催促企业，让其几乎立即删除内容，要求它们开发过滤器，禁止上传被认为有害的内容。这种压力旨在推出自动化工具，作为一种出版前审查的方式。问题是，上传过滤器的要求“将使内容在发布之前就能够在没有任何形式正当程序的情况下被屏蔽，背离

[38] 这并不排除一个人可能就网上而不是线下发生的传统侵权行为对另一人提出民事权利主张的可能性。但是，根据《公民权利和政治权利国际公约》第十九条，必须界定可能导致可通过法律纠正的损害的表述。

[39] 德国，《加强执行社交网络法律的法案》(《网络执行法》)(2017)，第 1(3)节。

[40] 关于网上仇恨言论的法国法案中也提到了类似问题。见 FRA 6/2019 和法国政府的答复，可查阅：https://spcommreports.ohchr.org/Tmsearch/TMDocuments。

[41] 但是，参见德国联邦法院 2008 年 4 月 3 日的判决，案件编号 3 StR 394/07。

[42] 前任特别报告员指出，“执行限制措施的机构必须以既不武断也不歧视的方式行事，并有充分的防止滥用的保障措施，不受任何政治、商业或其他不当影响”(A/67/357，第 42 段)。

[43] 见 AUS 5/2019 号通讯和澳大利亚常驻联合国办事处代表团及日内瓦其他国际组织的答复，可查阅：https://spcommreports.ohchr.org/Tmsearch/TMDocuments。

了公认的假设，即国家而不是个人应承担对表达自由的限制之责任”。[44] 由于这类过滤器是出了名的无法处理通常构成仇恨内容的自然语言，因此它们可能导致严重不当的结果。[45] 此外，有研究表明，这类过滤器对历史代表性不足的社区造成了尤为严重的伤害。[46]

35. 推动针对仇恨言论(和其他类型内容)的上传过滤器是不明智的，因为它逼迫平台对合法内容进行监管和删除。这类过滤器加强了公司的权力，而监督或补救的机会却很少，即便有的话。相反，各国应推行法律和政策，推动公司保护言论自由，并通过以下组合特征，打击法律限制的仇恨言论形式：允许公众监督的透明度要求；由独立的司法当局执行国内法；按照《拉巴特行动计划》和人权理事会第 16/18 号决议提出的方针进行的其他社会和教育努力。

36. 一些国家已采取步骤，通过其他创造性和看似相称的手段解决非法仇恨言论问题。印度在某些情况下将关闭互联网作为一种处理内容问题的工具，严重地干扰了民众获得通信的机会，[47] 而印度的一些邦则采用了其他方法。其中一个方法涉及建立热线，方便个人向执法当局报告 WhatsApp 应用软件的内容，而另一个方法涉及建立“社交媒体实验室”来监控网络仇恨言论。虽然这类方法需要仔细开发以符合人权规范，但它们揭示了采用“创造性”和“跳出框框”的方法来解决仇恨言论，而无需将内容警察的角色外包给遥远的公司。[48]

37. 2019 年，法国的一个官方委员会提出了一种监管网络内容的方法，似乎可以在保护表达权的同时也为解决非法仇恨言论留出空间。虽然在撰写本报告时，该委员会的工作状况尚不清楚，但其提案涉及司法当局解决仇恨言论问题，以及多利益相关方关于提供对公司政策监督的举措。该委员会的结论是：

> 因此，为迫使最大的参与者对我们的社会凝聚力采取更负责任和更具保护性的态度而进行的公众干预似乎是合法的。鉴于公民自由问题岌岌可危，这种干预应该特别审慎。它必须(1) 尊重范围广泛的社交网络模式，这些模式特别多

[44] OTH 71/2018 号通讯，可查阅：https://spcommreports.ohchr.org/Tmsearch/TMDocuments。另见欧洲联盟委员会 2018 年 3 月 1 日关于有效处理网络非法内容的(EU) 2018/334 号措施建议，其中呼吁“采取主动措施，包括使用自动化手段，以便发现、识别和迅速删除或禁止获取恐怖主义内容”。

[45] 见民主和技术中心，《混合信息？自动化社交媒体内容分析的局限性》，2017 年 11 月 28 日。

[46] 关于上传过滤器引起对表达自由的严重关切，见 Daphne Keller，《落网的海豚：互联网内容过滤器和检察长关于 *Glawischnig-Pieczek* 诉脸书爱尔兰分公司案的意见》，斯坦福互联网与社会中心，2019 年 9 月 4 日。

[47] 见 IND 7/2017 和 IND 5/2016 号通讯，可查阅：https://spcommreports.ohchr.org/Tmsearch/TMDocuments；另见联合国人权事务高级专员办事处新闻稿，《联合国人权专家敦促印度结束对克什米尔的通讯中断》，2019 年 8 月 22 日。

[48] Chinmayi Arun 和 Nakul Nayak，《关于网络仇恨言论与印度法律界定仇恨言论的初步调查结果》，2016 年 12 月 8 日，第 11 页。

样化，(2) 实行透明度原则，并将公民社会系统纳入其中，(3) 根据必要性和相称性原则，争取最低程度的干预，(4) 请法院确定个人内容的合法性。[49]

38. 这种做法值得进一步发展和考虑，因它以似乎能够尊重国际人权法的方式处理言论自由和社会凝聚力问题。

正当性

39. 政府对网络中介机构的监管应遵循与适用于所有政府言论限制的人权法所载的正当性准则相同的准则。如上所述，国家可能称之为“仇恨言论”的某些类型的言论不应受到《公约》第十九或二十条㈡的禁止。此外，限制煽动，例如煽动“对政权的仇恨”或“颠覆国家权力”的法律术语，是根据《公约》第十九条㈢(A/67/357，第 51-55)进行限制的非法依据。对仇恨言论的过于宽泛的定义，例如禁止煽动“宗教不和”或可能使国家面对暴力行为的言论，[50] 通常会使出于非法目的之言论限制成为可能，或者在政府管理网络中介公司时，对这些中介公司提出不符合人权法的的要求。

B. 公司的内容调节和仇恨言论

40. 网络仇恨内容是在互联网公司的平台上传播的，而网络公司的业务模式似乎只重视关注度和毒害性。[51] 目前最大的公司部署了“分类器”，使用人工智能软件根据特定的词语和分析识别违禁的内容，但或许只是取得了间歇性的成功。这些公司跨越司法管辖区进行运作，某一地方的内容可能会在其他地方产生不同的影响。网络仇恨言论通常涉及不知名的发表人，带有经协调的机器人威胁、虚假信息和所谓的深度虚假的信息，以及群体暴徒攻击。[52]

41. 互联网公司制订其平台的规则和公众展示(或品牌)。[53] 它们对人权产生了巨大影响，特别但不仅限于当它们成为公共和私人表达的主要形式的时候。在这些地方，限制言论可能导致公共沉默，或者无法处理煽动行为，而这会助长线下暴力和歧视(A/HRC/42/50，第 70-75 页)。不受管制的网络仇恨的后果可能是悲剧性的，脸书未能解决针对缅甸罗辛亚穆斯林社区的煽动行为就说明了这一点。公司

[49] 法国，《创建法国框架，使社交媒体平台更具责任感：以欧洲的眼光在法国行动》，向法国数字事务国务秘书提交的临时任务报告，2019 年 5 月。

[50] 见 JOR 3/2018 号通讯，可查阅：https://spcommreports.ohchr.org/Tmsearch/TMDocuments。

[51] 见 Tim Wu，*The Attention Merchants: The Epic Scramble to Get Inside Our Heads* (纽约，Vintage Books，2016 年)。

[52] 见 Gayathry Venkiteswaran，《“让暴民做这项工作”：仇恨的支持者如何威胁亚洲网络言论和宗教自由》(进步通信协会，2017 年 10 月)。

[53] 见 Kate Klonick，《新治理者：治理网络言论的人、规则和过程》，《哈佛法律评论》，第 131 卷，第 6 号，(2018 年 4 月)；David Kaye，《言论警察：治理互联网的全球斗争》(纽约，哥伦比亚大学全球报告，2019 年)。

没有政府的义务，但它们造成的影响则要求它们评估保护其用户表达自由权利的类似问题。[54]

42. 前几份报告认为，信息和通信技术行业的所有公司都应该适用联合国《工商企业与人权指导原则》，并通过设计和默认将人权纳入其产品。然而，公司在其平台上管理仇恨言论时，几乎完全没有联系到其产品对人权的影响。[55] 这是一个错误，它剥夺了公司做出遵守权利的决定并向政府和个人阐明其执行情况的框架，同时阻碍了公众使用全球理解的词汇进行索赔的能力。特别报告员重申呼吁公司执行人权政策，这些政策涉及以下机制：

(a) 定期审查其对人权的影响；

(b) 避免不利的人权影响，防止或减轻已确实产生的影响；

(c) 实施人权尽责程序，以“确定、防止和缓解人权影响，并对如何处理人权影响负责”，并制定补救伤害的程序。[56]

43. 对于如何将联合国人权标准应用于广泛的内容，总是会遇到难题，就像关于国家法律和区域人权法的难题一样。[57] 然而，上述指导可以帮助塑造公司在调节内容的每个阶段的权利保护：产品开发、定义、识别、行动和补救。全球规范为拥有跨界通信的全球用户的公司提供了一个坚实的基础，这些规则是《工商企业与人权指导原则》(原则 12)所要求的。[58]

人权尽职及审查

44. 处理仇恨言论应该从产品开发阶段的尽职要求开始。不幸的是，似乎很少有大型互联网公司进行与仇恨言论相关的以权利为导向的产品审查；即使有的话也没有公开。然而，信息和通信技术部门的产品正在不断更新和修订，因此，公司必须定期进行影响评估和重新评估，以确定其产品如何侵犯到享受人权。根据《工商企业与人权指导原则》，企业除其他事项外应有一个持续的程序，以确定仇恨言论如何影响其平台上的人权(原则 17)，包括通过平台自己的计算方法(A/73/348)。它们应借助内部和独立的人权专业知识，包括“与可能受影响的群体和其他利益攸关方进行切实磋商”(原则 18)。它们应该定期评估其对人权伤害所采取对策的有效性(原则 20)。

---

[54] 见 A/HRC/32/38，第 87 和 88 段；另见企业实现社会责任和世界经济论坛，《负责任的技术使用》，白皮书，2019 年 8 月。

[55] 在撰写本报告时，脸书刚刚发布了一份修订后的价值观声明，表明它将“寻求国际人权标准”以便做出涉及社标准的某些判断。见 Monika Bickert，《更新有关我们社区标准的价值观》，脸书，2019 年 9 月 12 日。

[56] 《工商企业与人权指导原则：实施联合国“保护、尊重和补救”框架》(A/HRC/17/31，附件)，原则 12(及评论)，原则 13 和 15。

[57] 见 Benesch，《关于改善监管的建议》。

[58] 见企业社会责任，《人权影响评估：脸书在缅甸》(2018 年 10 月)。

45. 缺乏透明度是所有公司内容调节过程中的一个主要缺陷。根据原则 21 的要求，对仇恨言论政策的外部审查(学术、法律和其他类型)存在重大障碍：虽然规则是公开的，但其在总体和粒子层面上的实施细节却几乎不存在。最后，公司还必须培训它们的内容政策团队、总法律顾问，特别是该领域的内容调节员，即那些进行实际限制工作的人(原则 16，评论)。公司的内容调节的目标应是保护和促进人权法的规范，而作为培训的一部分，则应确定这些规范。特别是，公司应通过评估上述合法性、必要性和正当性原则来评估其仇恨言论规则是否侵犯了表达自由。

合法性标准

46. 公司对仇恨言论的定义通常让人很难理解，尽管公司在这件事上的做法各不相同。有些公司没有这样的定义，而另一些公司的定义则含糊不清。例如，俄罗斯的社交网络 VK 禁止“宣传和/或助长种族、宗教、民族仇恨或敌意，宣传法西斯主义或种族优越感”或“包含极端主义材料”的内容。[59] 中国的短信应用软件微信则禁止“……事实上或在我们合理的观点中……那些(公开或不公开的)仇恨、骚扰、辱骂、种族或民族冒犯、诽谤、侮辱他人，威胁、亵渎或令人反感的内容。”[60] 另一些公司的定义则是多而且详细的，做了认真的工作准确地列述构成受限制仇恨言论的内容，但矛盾的是，多而详细的界定反而可能会造成困惑并缺乏明确性。三家占主导地位的美国公司——YouTube、脸书和推特——的政策经过了多年的演变和改进，[61] 每一家公司的政策分层方式都趋同于一套明显相似的规则。然而，虽然它们使用不同的术语来表明对“宣传”针对特定受保护群体的暴力或仇恨内容的限制，但它们并没有澄清它们如何定义宣传、煽动、针对群体等。除了其他问题，诸如意图和结果等主题很难在政策中判别(A/HRC/38/35，第 26 段)。

47. 公司应该重新审视它们的政策，或者采用新的政策，同时考虑到合法性测试。关于网络仇恨言论的符合人权的框架将借鉴上述定义指导，并提供以下问题的答案：

(a) 什么是受保护的个人或群体？人权法确定了需要明确保护的特定群体。信息和通信技术部门的公司应致力于根据不断发展的法律和规范性谅解，实施尽可能广泛的保护。公司应该清楚，它们不会限制“宣传一种积极的群体认同意识”，特别是对历史上处于不利地位的群体时，同时确认某些群体认同的表现形式，如白人至上主义，实际上可能构成仇恨内容；[62]

(b) 什么样的仇恨言论构成违反公司规则？公司应该通过考虑用户在平台上可能面临的干扰来制定仇恨言论政策。人权法提供指导，特别是指出为保护他人权利而设置限制的正当性。例如，公司可以考虑网络表达仇恨是如何能够煽动

[59] 见 https://vk.com/terms。

[60] 见 www.wechat.com/en/acceptable_use_policy.html。

[61] 见https://support.google.com/youtube/answer/2801939?hl=en，www.facebook.com/communitystandards/hate_speech 和 https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy。

[62] 第 19 条，《关于表达自由和平等的卡姆登原则》，原则 12。

威胁生命的暴力，侵犯他人的表达自由和获取他人信息，并干扰隐私或选举权，等等。公司不能代替政府来评估对国家安全和公共秩序的威胁，而出于这些理由的仇恨言论限制不应基于公司评估，而应基于国家的法令，而国家法令本身须遵守《公约》第十九条㈢的严格条件；

(c) 有没有公司限制的具体的仇恨言论内容？公司应说明如何禁止《公民权利和政治权利国际公约》第二十条㈡和《消除一切形式种族歧视国际公约》第四条所涵盖的表达方式。公司在确定这些禁止言论的定义时应从上述文书中引用。但煽动只是可能构成仇恨言论的问题内容的一部分。公司应该确定该类别处煽动之外还包括哪些内容，就像一些公司通过不断改变其政策那样。它们该做的不仅仅是识别；它们还应该通过制订一种判例法来展示它们的类别在规则的实际执行中究竟如何发挥作用(A/38/35，第 71 段)；

(d) 是否存在仇恨言论规则不适用的用户类别?国际标准明确规定，报道仇恨言论的记者和其他人应受到保护，免受内容限制或对其账户的不利行动的影响。此外，应用《拉巴特行动计划》的情境标准将导致对此类内容的保护。政治家、政府和军方官员以及其他公众人物则是另一回事。鉴于他们在煽动行为方面的突出地位和可能的领导作用，他们应该受到适用于国际标准的同样的仇恨言论规则的约束。在仇恨言论政策的背景下，默认公众人物应该遵守与所有用户相同的规则。在某些情况下，对情景的评价可能导致作出例外的决定，其中的内容必须作为例如政治言论来保护。但是几乎可以肯定的是，领导人口中说出的煽动言论要比其他用户说的更有害，这一因素应该成为平台内容评估的一部分。

48. 如果公司规则与国际标准不同，公司应提前对政策差异作出合理的解释，并明确说明差异。例如，如果一家公司决定禁止使用指代一个民族、种族或宗教团体的贬义词，而该用语本身不受人权法的限制，那么该公司根据人权法澄清其决定。此外，公司应特别警惕通过构成仇恨言论的虚假信息对其平台的滥用；特别是在紧张局势不断升级的环境中，公司应该明确陈述自己的政策，通过社区和专家的参与发展全面的理解，并坚决反对这种煽动。国际人权标准可以指导此类政策，而对这种情况下仇恨内容的毒害性可能需要作出快速反应和早期预警，以保护基本权利。

49. 公司应该定义它们如何判定用户违反仇恨言论规则。目前，很难知道在什么情况下可能会违反规则。在执行规则方面似乎存在严重的不一致性。执法的不透明性是问题的一部分。《拉巴特行动计划》确定了适用于根据《公约》第二十条㈡将煽动行为定为犯罪的一套因素，但这些因素在公司针对言论的行动方面也应具有分量。它们不需要像在刑事案件中那样被适用。然而，它们提供了一个有价值的框架，用于检查具体定义的内容——帖子、构成帖子的文字或图像——何时应予限制。

50. 公司可能会发现详细的情境分析是困难的，并且是资源密集型的。最大的公司严重依赖自动化，以便至少完成识别仇恨言论的第一层工作，需要的规则能够将内容分为一个类别(忽略)或另一个类别(删除)。它们使用人工智能来驱动这些系统，但这些系统在评价情境方面是出了名的差(A/73/348)。然而，如果公司真

的想在它们的平台上保护人权，它们必须确保它们清楚地定义规则，并需要人工评价。此外，人工评价必须不仅仅是评估特定词语是否归入特定的类别，而是必须基于在可能发现仇恨言论的社区的真实学习，由能够理解语言有时被用于隐藏煽动暴力行为的“代码”的人来评估讲话者的意图，考虑这个人和受众的性质，并评价仇恨言论可能导致暴力行为的环境。这些事情都不可能单靠使用人工智能完成，而定义和策略应该反映问题的细微差别。最大的公司应该承担这些资源的负担，并作为公开资源广泛分享其知识和工具，确保较小的公司和较小的市场能够获得此类技术。

必要性和相称性

51. 公司拥有以符合人权的方式处理内容的工具，在某些方面比国家享有的工具范围更广。这样的选择范围使它们能够根据问题的严重性和其他因素针对特定的问题内容制定相应的对策。它们可以删除内容，限制其毒害，标记其来源，暂停相关用户，暂停赞助该内容的组织，制定评级以突出个人对违禁内容的使用，在团队开展审核时临时限制内容，阻止用户将其内容货币化，阻碍分享其内容，并对内容贴出警告和标签，为个人提供更多的能力以拦截其他用户，最小化其内容放大功能，干扰机器人程序和协调的网络暴民行为，采用地理位置限制，甚至推动反宣传运动。并非所有这些工具都适用于每种情况，而且它们本身可能需要收到限制，但这些工具显示了公司在特定情况下可能拥有的删除之外的选择范围。换句话说，就像国家应该评估言论限制是否是限制最少的方法一样，公司也应该进行这种评价。在进行评价时，公司应承担在受影响用户提出要求时公开证明必要性和相称性的责任，无论用户是讲话者、声称的受害者、看到内容的另一个人还是公众成员。

52. Evelyn Aswad 确定了公司在必要性框架中应该采取的三个步骤：评估它可以在不干扰言论本身的情况下保护合法目标的工具；识别对言论干扰最小的工具；评估并证明它选择的措施是否的确实现了目标。[63] 这种评估符合《工商企业与人权指导原则》要求企业确保防止或减轻伤害的呼吁，特别是因为这样的方法使企业能够评估两组可能的危害：由实施其规则引起的对言论的限制，以及因用户对其他用户或公众发表仇恨言论而造成的对言论的限制。从这一框架中汲取的方法使公司能够确定如何不仅回应真正的煽动，而且还对网络常见的各种表达——擦边仇恨言论和非煽动——作出回应。

补救

53. 国际人权法机制为网络仇恨言论的补救提供了丰富的思路。《公民权利和政治权利国际公约》和《消除一切形式种族歧视国际公约》要求对违反其规定的行为提供补救，《工商企业与人权指导原则》也要求提供补救。特别报告员在其 2018 年内容调节报告中，强调了公司根据《指导原则》补救不利人权影响的责任（A/HRC/38/35，第 59 段），因此无需在本报告详细重述。简而言之，补救过程必须从个人报告潜在违反仇恨言论政策的有效方式开始，并确保保护报告制度不被

[63] Aswad，《自由的未来》，第 47-52 页。

滥用为仇恨言论的一种形式。它应该包括一个透明和可查阅的过程，以上诉平台的决定，而公司则提供一个合理的回应，也应是公众可查阅的。

54. 至少，公司应该公开确定它们将对违反其仇恨言论政策的行为采取何种补救措施。暂停用户可能是不够的。公司应该根据违规的严重程度或用户的累犯程度做出分级的回应。它们应该开发强大的产品，保护用户的自主权、安全性和自由表达，以补救侵权行为。它们的方法可能包括对无论出于何种原因它们都不想禁止的问题言论去放大并去货币化，但公司应再次根据现有的定义，提前向所有用户明确和宣传其政策，并向所有人发出警告，使用户有机会撤回这些言论，并在必要时补救冒犯性评论的后果。它们可以开发程序，要求希望返回平台的被暂停的用户参与各种赔偿，例如道歉或其他形式的直接与他们所伤害的人接触。它们应该有教育、反言论、报道和培训的补救政策。对于最严重的过失，补救措施还应包括违规后影响评估和制定政策以结束违规行为。

55. 《拉巴特行动计划》和人权理事会第 16/18 号决议也提供了一些思路，供公司在对仇恨内容提供补救时借鉴。根据《拉巴特行动计划》，“各国应确保因煽动仇恨而遭受实际伤害的人有权获得有效补救，包括民事或非司法损害赔偿”。这些补救措施可以包括金钱赔偿、“改正权利”和“答辩权”(A/HRC/22/17/Add.4，附录，第 33–34 段)。人权理事会在第 16/18 号决议中确定了培训政府官员和促进少数群体表达其信仰的权利等工具。前任特别报告员敦促程序性补救措施，例如“诉诸司法和确保国内机构的有效性”，和实质性补救措施，例如“充分、迅速和与表达的重要性相称的赔偿，其中可能包括恢复名誉、防止再次发生和提供经济赔偿”(A/67/357，第 48 段)。但他也敦促采取一系列非法律补救措施，考虑到公司作为仇恨内容产生平台的创造者的责任，它们应该评估并实施这些补救措施。这种补救行动可以包括关于仇恨言论危害的教育努力，以及仇恨言论往往旨在将弱势群体推离平台的方式(即，让他们噤声)；促进对仇恨言论作出回应的机制，并给予更大的能见度；公开谴责仇恨言论，如宣传公共服务公告和公众人物声明；加强与社会科学研究人员的合作，评估问题的范围和最有效防止仇恨内容扩散的工具(同上，第 56-74 段)。

四. 结论和建议

56. 国际人权法应被理解为在打击仇恨、冒犯、危险或不受欢迎的言论时保护和尊重人权的重要框架。本报告中描述的网络仇恨言论是一种广泛的表达方式，可导致有害的结果。当这一短语被滥用时，它可以为恶意国家提供一个惩罚和限制言论的工具，而这种言论在尊重权利的社会中是完全合法的，甚至是必要的。但有些言论可以造成真正的伤害。它可以恐吓脆弱的社区使其保持沉默，特别是当它涉及鼓吹构成煽动敌意、歧视或暴力的仇恨时。如果放任不管，使其传播毒害，它会造成一个破坏公共言论的环境，甚至会伤害那些不是主题平台用户的人。因此，国家和公司必须下决心保护面临被迫沉默风险的人，并促进就即使是最敏感的公共利益问题进行公开和严肃的辩论，以解决仇恨言论问题。

给各国的建议

57. 国家对网络仇恨言论的做法应该从两个前提开始。首先，离线情景下的人权保护也必须适用于在线言论。不应该有特殊类别的在线仇恨言论，其惩罚高于离线仇恨言论。其次，各国政府不应通过法律或法外威胁要求中介公司采取国际人权法将禁止各国直接采取的行动。在这两个基础上，并参照上文概述的规则，各国在处理网络仇恨言论方面至少应做到以下几点：

(a) 在其法律中严格界定构成《公民权利和政治权利国际公约》第二十条㈡和《消除一切形式种族歧视国际公约》第四条所禁止内容的术语，并抵制将此类言论定为刑事犯罪，除非是在最严重的情况下，例如主张构成煽动歧视、敌对或暴力的民族、种族或宗教仇恨，并通过《拉巴特行动计划》所载的人权法的解释；

(b) 审查现有法律或制定符合合法性、必要性和相称性以及正当性要求的关于仇恨言论的立法，并将此类规则制定置于强有力的公众参与之下；

(c) 积极考虑和部署善政措施，包括人权理事会第 16/18 号决议和《拉巴特行动计划》中建议的措施，以处理仇恨言论，减少人们认为的对禁止表达的需要；

(d) 通过或审查中介责任规则，严格遵守人权标准，但不要求公司限制表述，去完成国家自己通过立法无法直接做到的事请；

(e) 建立或加强独立的司法机制，确保个人在遭受可认定的《公民权利和政治权利国际公约》第二十条㈡或《消除一切形式种族歧视国际公约》第 4 条所述损害时可获得司法和补救；

(f) 通过法律，要求公司详细公开地描述它们如何界定仇恨言论并执行其反对仇恨言论的规则，建立公司所采取的反对仇恨言论行动的数据库，并在其他方面鼓励公司在自己的规则中尊重人权标准；

(g) 积极参与旨在解决仇恨言论问题的学习论坛的国际进程。

给公司的建议

58. 公司长期以来一直规避将人权法作为其规则和规则制定的指南，尽管它们对其用户和公众的人权产生了广泛的影响。除了前几份报告中采用的原则外，为符合《工商企业与人权指导原则》，信息和通信技术部门的所有公司都应：

(a) 通过定期和公开的人权影响评估，评价其产品和服务如何影响其用户和公众的人权；

(b) 采取内容政策，将其仇恨言论规则与国际人权法直接挂钩，表明将根据国际人权法的标准执行这些规则，包括联合国相关条约和条约机构和特别程序任务负责人的解释以及其他专家意见，包括《拉巴特行动计划》；

(c) 定义它们认为仇恨言论的内容类别，并对用户和公众进行合理的解释，以及跨司法管辖区一致的方法；

(d) 确保仇恨言论规则的执行都涉及对情境以及内容对用户和公众造成的伤害的评价，包括确保人参与到任何自动化或人工智能工具的使用回路中；

(e) 确保情境分析包括被确定为仇恨言论的内容影响最大的社区，并确保社区参与确定解决平台上造成的伤害的最有效的工具；

(f) 作为解决仇恨言论整体努力的一部分，开发促进个人自主、安全和自由表达的工具，并酌情采用去放大、去货币化、教育、反言论、报告和培训作为禁用账户和删除内容的替代办法。