联合国 A/HRC/38/35

大会

Distr.: General
6 April 2018
Chinese
Original: English

人权理事会
第三十八届会议
2018 年 6 月 18 日至 7 月 6 日
议程项目 3
促进和保护所有人权——公民权利、政治权利、
经济、社会及文化权利，包括发展权

促进和保护意见和表达自由权问题特别报告员

秘书处的说明

秘书处谨向人权理事会转交促进和保护意见和表达自由权问题特别报告员大卫·凯伊依照理事会第 34/18 号决议编写的报告。在本报告中，特别报告员介绍了网上用户原创内容的监管。他建议各国确保提供一个有利于在线自由表达的环境，同时建议各公司在所有运营阶段适用人权标准。人权法赋予公司工具，以尊重民主规范和抵抗专制命令的方式阐述自己的立场。随着个人越来越多地在网上行使基本权利，各公司和国家至少应该谋求从根本上提高规则制定和执行的透明度，确保用户有自主权。

GE.18-05436 (C) 300418 010518
1805436

促进和保护意见和表达自由权问题特别报告员的报告

目录

一. 导言

1. 在数字时代初期，约翰·佩里·巴洛宣布互联网将开创“一个新世界，任何人在任何地方均可表达信仰，无论这种信仰是多么奇特，而无需担心被胁迫而沉默或服从”。[1] 尽管互联网仍然是有史以来最伟大的全球信息获取工具，但是这种网络福音如今已经罕见。公众从用户自创内容中看到的是仇恨、滥用和虚假信息。政府看到的是恐怖分子的招募或令政府感到尴尬的异议和反对意见。民间社会组织看到的是公共职能，例如保护表达自由的职能被外包给不负责任的私营部门。尽管公司正在采取措施说明其规则和与政府的互动，但公司仍然是神秘的监管者，它们制定了一种“平台法”，从中难以看到明确性、一致性、问责制和补救办法。联合国、各区域组织和条约机构已经申明，线下权利同样适用于线上，然而公司是否保护用户权利或者国家是否从法律上激励公司这样做，这一点并不总是清晰的。

2. 在本报告中，特别报告员提议建立一个以人权为中心的用户自创内容审核框架。[2] 他试图回答以下一些基本问题，即：公司在确保它们的平台不妨碍国际法保障的权利方面有哪些责任？公司应在内容审核中适用哪些标准？国家是否应该监管商业内容的审核？如果应该，如何监管？法律希望国家实现透明度和问责制，以减轻对表达自由的威胁。我们是否应该对私营部门寄予同样的希望？数字时代中的保护和补救程序是什么样的？

3. 之前的报告已经讨论过其中一些问题。[3] 本报告的重点是用户自创内容的监管，主要由国家和社交媒体公司进行监管，但监管的方式适用于信息通信技术部门中的所有相关行为体。特别报告员概述了适用的人权法律框架并介绍了公司和国家的内容监管方法。他提出了公司应依照国际人权法采纳的内容监管标准和程序。

4. 对公司服务条款、透明度报告和次级来源的研究为本报告提供了初步依据。经过征求意见，收到了 21 份由国家提交的材料和 29 份由非国家行为体提交的材料(包括 1 份公司提交的材料)。特别报告员访问了硅谷的几家公司，并与其他公司进行了对话，力求了解它们的内容审核方法。[4] 他受益于 2017 年和 2018 年在曼谷和日内瓦与民间社会进行的协商和 2018 年与拉丁美洲、中东、北非和撒哈拉以南非洲的专家举行的在线讨论。[5]

[1] 约翰·佩里·巴洛,《网络空间独立宣言》，1996 年 2 月 8 日。

[2] “审核”是指互联网公司确定用户生成内容是否符合公司服务条款和其他规则所载标准的过程。

[3] A/HRC/35/22 和 A/HRC/32/38。

[4] 特别报告员访问了脸书、Github、谷歌、Reddit 和推特公司的总部，并与雅虎/Oath、Line 和微软公司的代表交谈。他还访问了非营利组织维基媒体基金会。他希望访问北京、莫斯科、首尔和东京的公司，以开展本报告相关工作。

[5] 特别报告员谨此感谢他的法律顾问阿莫斯·托阿和加州大学欧文分校法学院国际司法讲习班的学生。

二. 法律框架

5. 信通技术部门公司的活动牵涉隐私权、宗教自由和信仰权、见解和表达权、集会和结社自由权以及公众参与权等权利。本报告的重点是表达自由权，同时承认各项权利之间的相互依存性，例如隐私权是实现表达自由的一个重要途径。[6] 《公民权利和政治权利国际公约》第十九条的规则得到了全球的认可和 170 个国家的批准，同时与《世界人权宣言》相呼应，都规定保证“人人有权持有主张，不受干涉”以及“有权不论国界，通过任何媒介寻求、接收和传递各种信息和思想”。[7]

A. 国家义务

6. 人权法规定各国有义务确保为表达自由提供有利环境，并保障该权利的行使。保障表达自由的义务要求各国，除其他外，促进媒体多元化和独立性以及信息获取。[8] 此外，国际和区域性机构已敦促各国促进互联网普遍接入。[9] 各国还有义务确保私营实体不干涉见解和表达自由。[10] 人权理事会在 2011 年通过了《工商企业与人权指导原则》，其中原则 3 强调国家有责任确保为工商企业创造环境，使其能够尊重人权。[11]

7. 各国不得限制“持有主张、不受干涉的权利”。《公民权利和政治权利国际公约》第十九条第三款规定，国家对表达自由的限制必须符合以下既定条件：

- *合法性*。限制必须是由“法律规定”的。具体而言，必须通过正规法律程序采取限制，并以“足够准确”的方式区分合法和非法言论，以限制政府的酌处权。秘密采取的限制不符合这一基本要求。[12] 对合法性的保障与否，通常应受独立司法机关的监督。[13]

- *必要性和相称性*。各国必须证明，限制对权利行使造成最小负担，而且实际上保护或有可能保护有争议的国家正当利益。各国不能仅仅坚

[6] 见 A/HRC/29/32, 第 16-18 段。

[7] 另见《非洲人权和民族权宪章》，第 9 条；《美洲人权公约》，第 13 条；《保护人权与基本自由公约》，第 10 条。另见言论自由和信息获取自由研究中心 (Centro de Estudios en Libertad de Expresión y Acceso a la Información) 提交的材料。

[8] 《关于表达自由和“虚假新闻”、虚假信息和宣传行为的联合声明》，2017 年 3 月 3 日，第 3 节。另见人权事务委员会《关于见解自由和言论自由的第 34 号一般性意见》(2011 年)，第 18 和第 40 段；A/HRC/29/32, 第 61 段和 A/HRC/32/38, 第 86 段。

[9] 见人权理事会第 32/13 号决议，第 12 段；美洲人权委员会言论自由问题特别报告员《自由、开放和包容的互联网标准》(2016 年)，第 18 段。

[10] 见第 34 号一般性意见，第 7 段。

[11] A/HRC/17/31。

[12] 同上，第 25 段；A/HRC/29/32。

[13] 同上。

称有必要性，而是必须在通过限制性立法和限制具体表达内容时加以证明。[14]

- 正当性。任何限制要想合法，必须只保护《公约》第十九条第三款列举的权益：他人的权利或名誉，国家安全或公共秩序，或公共卫生或道德。例如，旨在保护他人权利的限制包括为保护“《公约》所承认的人权和国际人权法普遍承认的人权”[15] 而施加的限制。为保护隐私权、生命权、正当程序权、结社和参与公共事务权等权利而施加的限制，经证实通过合法性和必要性检验后，将是合法的。人权事务委员会警告说，为保护“公共道德”施加的限制不应“排他性地出自一种传统”，以求确保这种限制体现不歧视原则和权利的普遍性。[16]

8. 根据《公约》第二十条第二款——要求各国禁止“鼓吹民族、种族或宗教仇恨，构成煽动歧视、敌视或强暴行动的主张”——施加的限制仍必须全部满足合法性、必要性和正当性这三项条件。[17]

B. 公司责任

9. 互联网公司已成为讨论和辩论、信息获取、商务和人文发展的中心平台。[18] 它们收集和保留数十亿人的个人数据，包括关于个人习惯、行踪和活动的数据，经常宣称自己的公司公民角色。2004 年，谷歌提出了“即使放弃一些短期收益，也要为世界谋福利”的志向。[19] “脸书”创始人宣布希望“发展社会基础设施，让人们有能力建设一个适合所有人的全球社区”。[20] 推特已承诺采取政策，“促进而非破坏自由的全球性对话”。[21] 俄罗斯社交媒体公司 VKontakte “将世界各地的人们相互连接”，而腾讯“为构建和谐社会出一份力，成为良好企业公民”的目标体现了中国政府的用语。[22]

10. 几乎没有几家公司在运营中适用了人权原则，而适用人权原则的公司则大多认为这些原则在如何应对政府威胁和要求方面作用有限。[23] 然而，《工商企业与人权指导原则》制定了“预期行为的全球标准”，应适用于公司的各项业务

[14] 见第 34 号一般性意见，第 27 段。

[15] 同上，第 28 段。

[16] 同上，第 32 段。

[17] 同上，第 50 段。另见 A/67/357。

[18] 例如，见 2017 年 6 月 19 日美国高等法院对 *Packingham* 诉北卡罗来纳州案的意见；2009 年 3 月 10 日欧洲人权法院对 *Times Newspapers Ltd. (Nos. 1 and 2)*诉联合王国案(申请号 3002/03 和 23676/03 号)的判决书，第 27 段。

[19] 根据 1933 年《证券法》作出的《证券登记声明》(S-1)，2004 年 8 月 18 日。

[20] 马克·扎克伯格，“建设全球社区”，脸书，2017 年 2 月 16 日

[21] 推特《S-1 登记声明》，2013 年 10 月 13 日，第 91-92 页。

[22] VKontakte，公司信息；腾讯，“公司信息”。

[23] 丹麦人权学会提交的材料。参见雅虎/Oath 提交的材料，2016 年。

活动，无论公司在何处运营。[24] 虽然《指导原则》不具有约束力，但是公司在全球公共生活中发挥着极其重要的作用，因此亟需采纳和执行该原则。

11. 根据《指导原则》确立的框架，公司至少应该：

(a) 避免造成或助长负面人权影响，努力预防或缓解经由其商业关系与其业务、产品或服务直接相关的负面人权影响，即使并非它们造成了此类影响(原则 13)；

(b) 作出高层政策承诺，承诺尊重用户人权(原则 16)；

(c) 开展尽职调查，查明和解决公司活动导致的实际和潜在人权影响，并对这种影响负责，包括开展定期风险和影响评估、与可能受影响的群体和其他利益攸关方进行有意义的协商、采取适当的后续行动来减轻或预防这些影响(原则 17-19)；

(d) 实施预防和减缓战略，在面临相互冲突的地方法律要求时，尽最大可能信守国际公认的人权原则(原则 23)；

(e) 持续审查在尊重权利方面的努力，包括定期与利益攸关方进行协商，并与受影响的群体和公众开展频繁、便利和有效的沟通(原则 20-21)；

(f) 提供适当的补救，包括通过用户能运用的业务层面申诉机制，同时不加剧他们的“丧失权利感”(原则 22、29 和 31)。

三. 内容监管方面的主要关切

12. 政府设法为公司审核内容创造环境，而公司则依据用户协议约束个人访问平台的行为，用户协议中的服务条款规定了个人可以表达哪些内容和如何表达内容。

A. 政府监管

13. 各国经常要求公司限制明显非法的内容，例如展现儿童性虐待的内容、直接和可信的伤害威胁以及煽动暴力的内容，假定这些要求也满足合法性和必要性的条件。[25] 一些国家更为严格，依靠审查和定罪来塑造网络监管环境。[26] 针对“极端主义”、亵渎、诽谤、“攻击性”言论、“虚假新闻”和“宣传”的限制性法律措辞宽泛，它们往往成为要求公司镇压合法言论的借口。[27] 各国越来越

[24] 《指导原则》，原则 11。

[25] 爱尔兰已建立了共同监管机制，与企业共同限制非法儿童性虐待资料：爱尔兰提交的材料。许多公司依靠图像识别算法检测和删除儿童色情制品：开放技术研究所提交的材料，第 2 页和 ARTICLE 19 提交的材料，第 8 页。

[26] 见 A/HRC/32/38, 第 46-47 段。关于互联网关闭问题，见 A/HRC/35/22, 第 8 至 16 段以及特别报告员的来文实例：第 UA TGO 1/2017、UA IND 7/2017 和 AL GMB 1/2017 号来文。

[27] 第 OL MYS 1/2018、UA RUS7/2017、UA ARE 7/2017、AL BHR 8/2016、AL SGP 5/2016 和 OL RUS 7/2016 号来文。阿塞拜疆禁止宣传恐怖主义、宗教极端主义和自杀行为：阿塞拜疆提交的材料。

多地将网络平台上的内容作为监管目标。[28] 其他法律可能会阻止人们行使见解和表达自由，从而干涉在线隐私。[29] 许多国家还部署了虚假信息和宣传工具，以限制独立媒体的可获取性和可信度。[30]

14. *责任保护*。从数字时代初期以来，许多国家采取了规则来保护媒介免于对第三方在其平台上发表的内容承担责任。例如，欧盟电子商务指令建立了一个法律制度，以保护媒介免于对内容承担责任，除非它们超越了为用户提供信息充当的“单纯渠道”、“缓存方”或“托管方”的角色。[31]《美国通信道德法》第 230 条为提供“交互式计算机服务”以托管或发布关于他人信息的一方普遍提供豁免，但这种豁免已经受到限制。[32] 巴西的媒介责任制度要求必须有法院命令才能限制具体内容，[33] 而印度的媒介责任制度建立了一个“通知和清理”程序，也需要有法院或类似裁决机构的命令。[34] 由民间社会专家联盟制定的 2014 年《马尼拉媒介责任原则》确定了应指导任何媒介责任框架的基本原则。

15. *规定公司义务*。有些国家根据模糊或复杂的法律标准，在没有事先进行司法审查的情况下，以严厉处罚作为威胁，将限制内容的义务强加给公司。例如，2016 年《中国网络安全法》加强了模糊的禁止措施，禁止传播会扰乱“社会或经济秩序”、民族团结或国家安全的“虚假”信息；它还要求各公司监测其网络并向当局报告违规行为。[35] 据报告称，该国几个最大的社交媒体平台已经因为不遵守要求而被处以高额罚款。[36]

16. 监测和迅速删除用户自创内容的义务在全球越来越普遍，甚至在民主社会中也建立了有可能破坏表达自由的惩罚性框架。德国的《网络执行法》(NetzDG)要求大型社交媒体公司删除不符合特定地方法律的内容，并对未能在极短时间内遵守要求的公司进行重罚。[37] 欧洲联盟委员会甚至建议成员国规定积极监测和过滤非法内容的法律义务。[38] 肯尼亚在 2017 年通过了选举期间社交媒体内容传

[28] 见第 OL PAK 8/2016 号和第 OL LAO 1/2014 号来文；进步通讯协会《放开言论自由：关于亚洲地区网络言论刑罪化法律的研究》，全球信息社会观察组织《2017 年特刊》。

[29] A/HRC/29/32。

[30] 例如，见 Gary King、Jennifer Pan 和 Margaret E. Roberts《中国政府如何编造社交媒体帖子以转移注意力而非参与论争》，《美国政治科学评论》第 111 期第 3 卷 (2017 年)，第 484-501 页。

[31] 欧洲议会和欧洲理事会 2000 年 6 月 8 日第 2000/31/EC 号指令。

[32] 《美国法典》第 47 卷第 230 节。另见《允许各国和受害者打击网络性贩运行为的法案》(H.R. 1865)。

[33] 《网络民法》(第 12.965 号联邦法)，第 18-19 条。

[34] 印度最高法院，*Shreya Singhal 诉印度联盟案*，2015 年 3 月 24 日的判决。

[35] 第 12 和第 47 条；中国人权提交的材料，2016 年，第 12 页。对《网络安全法》早期草案提出的意见，见第 OL CHN 7/2015 号来文。另见全球之声《网民报告：互联网审查法案在埃及盛行》，2018 年 3 月 16 日；南非共和国《电影与出版物修正法案》(B 61-2003)。

[36] 美国笔会《封锁消息流：中国政府对社交媒体的控制》(2018 年)，第 21 页。

[37] 《关于加强社交网络法律执行力度的法案》(《网络执行法》)，2017 年 7 月。见第 OL DEU 1/2017 号来文。

[38] 欧洲联盟委员会，关于有效处理非法在线内容的措施建议(最近一次更新：2018 年 3 月 5 日)。

播准则，要求各平台在 24 小时内“删除平台上用于传播不良政治内容的帐号”。[39]

17. 出于合理的国家关切，例如隐私和国家安全，诉诸监管是可以理解的。但是，这种规则会威胁表达自由，对公司造成重大压力，以致于它们在大力规避责任的同时可能会删除一些合法的内容。这种规则还会将监管职能委任给私营部门，而私营部门缺乏基本的问责工具。关于迅速和自动删除的要求，可能会导致事先限制的新形式，而这些限制已经威胁到版权领域的创新性努力。[40] 一般应由公共机构裁定事实和法律的复杂问题，而不是私营部门，后者的现行程序可能不符合正当程序标准，其动机主要是经济动机。[41]

18. *全球删除*。一些国家正在要求域外删除链接、网站和其他据称违反当地法律的内容。[42] 这种要求引发的严重关切是，国家可以“不分国界”地干涉表达自由权。按照这些要求的逻辑，审查将跨越国界，以便开展最严格的审查。应该要求各国在每个相关的辖区通过正规的法律和司法程序提出这种删除请求。

19. *没有国家法律依据的政府要求*。公司将通过正规法律渠道提出的关于删除据称非法内容的请求与基于公司服务条款提出的删除请求区别开来。[43] (依法删除通常仅在提出请求的司法管辖区适用；基于公司服务条款的删除通常适用于全球。)国家当局越来越多地寻求在法律程序之外删除内容，甚至是根据公司服务条款提出这种要求。[44] 一些国家已经建立了专门的政府单位，负责将需要删除的内容转交公司。例如，欧盟互联网转介小组“举报网上恐怖主义和暴力极端主义内容，并与网络服务提供商合作，以删除这些内容”。[45] 澳大利亚也有类似的转介机制。[46] 据报告称，东南亚各方与政府联合，试图利用基于服务条款的删除请求限制政治批评言论。[47]

---

[39] 见第 OL KEN 10/2017 号来文；Javier Pallero《洪都拉斯：新法案扬言要抑制网上言论》，Access Now, 2018 年 2 月 12 日。

[40] 见欧洲联盟委员会《欧洲议会和欧洲理事会关于数字单一市场中版权问题的指令提案》，COM (2016) 593 final, 第 13 条；Daphne Keller《欧盟委员会平台提案中的过滤问题》，斯坦福法学院互联网和社会中心，2017 年 10 月 5 日；Karisma 基金会提交的材料，2016 年，第 4-6 页。

[41] 欧盟法律规定，搜索引擎必须确定在“被遗忘权”框架下提出的申诉的有效性。欧洲法院 (大审判庭) 2014 年 5 月 13 日对*谷歌西班牙诉西班牙数据保护局和 Mario Costeja González* 案 (C-131/12 号案件)的判决；ARTICLE 19 提交的材料，第 2-3 页和 Access Now 提交的材料，第 6-7 页；谷歌《谷歌“被遗忘权”透明度更新报告》；Theo.Bertram 等人，《被遗忘权的三年》(谷歌，2018 年)。

[42] 例如，见美国笔会《禁封信息流》，第 36-37 页；加拿大最高法院对*谷歌公司诉* Equuestek Solutions *公司案*的 2017 年 6 月 28 日判决；欧洲法院，*谷歌公司诉法国国家信息与自由委员会*(第 C-507/17 号案件)；全球网络倡议提交的材料，第 6 页。

[43] 比较《推特透明度报告：删除请求》(2017 年 1-6 月)和《推特透明度报告：基于服务条款的政府删除请求报告》(2017 年 1-6 月)。另见脸书“政府请求：常见问题”。

[44] ARTICLE 19 提交的材料，第 2 页和全球网络倡议提交的材料，第 5 页。

[45] 欧盟互联网转介小组，第一年年度报告，第 4.11 节；欧洲数字权利组织提交的材料，第 1 页和 Access Now 提交的材料，第 2-3 页。

[46] 澳大利亚提交的材料。

[47] 东南亚媒体联盟，第 1 页。

20. 各国还对公司施加压力，通过不具约束力的要求，要求它们加快删除内容，这些删除大多不够透明。巴基斯坦对 YouTube 施行三年禁令，这迫使谷歌创建了当地版本，该版本容易顺从于政府删除“冒犯性”内容的要求。[48] 据报告称，“脸书”和以色列同意对努力和工作人员作出协调，以监测和删除“煽动性”网络内容。双方没有透露该协议的细节，但以色列司法部长声宣称，在 2016 年 6 月至 9 月期间，“脸书”同意了几乎所有删除“煽动性”内容的政府请求。[49] 根据国家意见协调内容方面的行动的安排加剧了人们的担忧，即公司在不受法院及其他问责机制监督的情况下履行公共职能。[50]

21. 2016 年《关于打击网上非法仇恨言论的欧盟行为守则》涉及欧盟与四家主要公司在删除内容方面的协议，要求这四家公司与“可信的举报者”合作并宣传“独立的反击言论”。[51] 虽然宣传反击言论可能对于打击“极端主义”或“恐怖主义”内容效果甚好，但是大力推行这种方法可能会把各平台变成宣传平台，大大超出既定的合理关切范畴。[52]

B 公司内容审核

公司遵守国内法

22. 每家公司原则上都承诺遵守业务所在地的地方法律。正如“脸书”所说：“如果我们经过仔细的法律审查，认定某内容违反了当地法律，那么我们将在相关国家或领土境内删除此内容。”[53] 移动聊天和社交媒体应用程序“微信”的所有者腾讯公司更加严格，要求在中国使用此平台的任何人和“在世界任何地方”使用此平台的中国公民遵守内容限制，这些限制体现了中国法律或政策。[54] 一些公司之间也相互协作，并与监管机构协作，删除儿童性虐待图片。[55]

23. 如果相关的国家法律含糊不清，可作不同的解释或不符合人权法，这会使遵守法律的承诺变得复杂。例如，反“极端主义”法律并未对这一关键术语下定义，这使政府当局有酌处权，可以基于值得怀疑的理由压迫公司删除内容。[56] 同样，公司往往受到压力，要遵守国家法律，按刑事罪论处据说亵渎神明、批评

[48] 数字权利基金会提交的材料。

[49] 阿拉伯社交媒体发展中心 7amleh 提交的材料。

[50] 进步通信协会，第 14 页和 7amleh。

[51] “可信的举报者是指给予某些组织的地位，使其能够通过一个不对普通用户开放的特殊报告制度或渠道举报非法内容。”欧洲联盟委员会，《关于打击网上非法仇恨言论的行为守则》：初步执行成果 (2016 年 12 月)

[52] 为努力开发全行业技术工具，删除平台上的恐怖主义内容，这些公司设立了全球网络反恐论坛。谷歌《全球网络反恐论坛情况更新》，2017 年 12 月 4 日。

[53] 脸书,“政府请求：常见问题”。另见谷歌依法删除请求；推特的规则和政策；Reddit 内容政策。

[54] 腾讯，《服务条款：导言》；腾讯，《腾讯微信软件许可及服务协议》。

[55] 联合国教育、科学及文化组织，《促进网络自由：互联网媒介的作用》(巴黎，2014 年)，第 56-57 页。

[56] 见 Maria Kravchenko《2016 年俄罗斯不当执行反极端主义立法》，SOVA 信息和分析中心，2017 年 4 月 21 日；Danielle Citron《极端主义言论、强制遵守和审查渗透》，《圣母大学法律评论》，第 93 期第 3 卷(2018 年)，第 1035-1071 页。

国家、诽谤公职人员或虚假的内容。如下文所述，《指导原则》提供了工具，以尽量减小这种法律对个人用户的影响。全球网络倡议是一个多利益攸关方倡议，旨在帮助信通技术公司应对人权挑战，并为这些工具的使用提供了额外指引。[57] 其中一个工具是透明度：许多公司每年按国别报告它们所收到和执行的政府请求数量。[58] 然而，公司并不总是披露足够的信息说明它们是如何应对政府请求的，也没有定期报告政府根据服务条款所提出的请求。[59]

公司内容审核标准

24. 互联网公司要求其用户遵守适用于在它们的平台表达言论的服务条款和“社区标准”。[60] 用户必须接受公司服务条款才能使用平台，服务条款明确了解决争端的管辖权并保留公司在内容和帐号的行动方面的自由裁量权。[61] 平台的内容政策是服务条款的一部分，阐明了对用户表达内容和表达方式的限制。大多数公司在制定内容标准时没有明确以任何可以管理言论的特定法律文本为依据，例如国内法或国际人权法。然而，中国搜索引擎巨头百度禁止“违反《中华人民共和国宪法》所载基本原则”的内容。[62]

25. 内容审核政策的制定通常有法律顾问、公共政策和产品经理以及高管的参与。公司可以建立“信任与安全”团队，负责处理垃圾邮件、欺诈和滥用行为，还可以建立反恐团队，负责处理恐怖主义内容。[63] 一些公司已经制定了机制，用于征求外部团体对于内容政策的专业意见。[64] 呈指数增长的用户自创内容已经促使公司制定详细和不断发展的规则。由于一系列因素，从公司规模、收入、业务模式到“平台的品牌和口碑、风险承受力和想要吸引的用户参与类型”等，各个公司的规则各异。[65]

[57] 全球网络倡议，《表达自由和隐私原则》，第 2 节。参与这一倡议的社交媒体公司包括脸书、谷歌、微软/领英和雅虎/Oath。

[58] 见下文第 39 段。此外，Automattic、谷歌、微软/必应和推特等公司定期在 Lumen 数据库发布政府删除内容和知识产权请求，尽管发布的内容不一定全面。

[59] 数字权利排名组织，《2017 年企业问责制指数》，第 28 页。

[60] Jamila Venturini 等人《服务条款与人权：网络平台协议分析》(里约热内卢，Revan 公司，2016 年)。

[61] 百度用户协议(“百度有权出于任何原因自行决定删除和删除本服务中的任何内容。”)；腾讯服务条款(“我们保留出于任何原因屏蔽或删除您的内容的权利，包括我们认为应当这样做或适用的法律和法规要求这样做。”)；推特服务条款(“我们可以随时以任何理由或无需理由暂停或终止您的帐号，或停止向您提供全部或部分服务。”)。

[62] 百度服务条款，第 3.1 节。

[63] Monika Bickert《难题：我们如何打击恐怖主义》，2017 年 6 月 15 日。

[64] 例如，见推特的信任与安全理事会和 YouTube 的可信举报者计划。

[65] Sarah Roberts《内容审核》(加州大学洛杉矶分校，2017 年)。另见 ARTICLE 19 提交的材料，第 2 页。

内容标准方面的关切领域

26. *含糊的规则*。公司规定禁止威胁使用或宣传恐怖主义、[66] 禁止支持或颂扬危险组织的领导人，[67] 禁止宣传恐怖主义行为或煽动暴力的内容[68]，这些规定和反恐立法一样过于含糊。[69] 公司关于仇恨、骚扰和滥用行为的政策也没有明确指出哪些行为会构成犯罪。推特禁止“煽动对某一受保护团体的恐惧”，而“脸书”则区分对受保护特性的“直接攻击”与单纯“令人反感或冒犯性的内容”，这些都是主观而不稳定的内容审核依据。[70]

27. *仇恨、骚扰、虐待*。含糊不清的仇恨言论和骚扰政策已经引发了对于政策执行不一致现象的投诉，即：处罚少数群体而增强主导或强势群体地位的现象。用户和民间社会举报：暴力侵害和虐待妇女行为，包括人身威胁、厌恶女性的评论、未经同意发布或发布伪造的亲密图片以及人肉搜索行为；[71] 威胁要伤害被剥夺政治权利的人、[72] 少数民族和种姓[73] 以及遭受暴力迫害的族裔群体；[74] 以及虐待难民、移民和寻求庇护者的行为。[75] 与此同时，据报告称各平台压制以下内容：男女同性恋、双性恋、跨性别者和性别奇异者激进主义，[76] 反对镇压性政府的主张，[77] 对种族清洗的报道[78] 以及对种族主义现象和权力结构的批评意见。[79]

28. 仇恨言论问题的解决涉及范围广、复杂程度高，是一项长期挑战，可能导致公司限制此类言论，即使这种言论与负面结果并无明确关联(根据《公民权利和政治权利国际公约》第二十条，鼓吹仇恨的主张与煽动行为有关联)。但是，公司应当阐明这种限制的理由，并证明任何内容方面的(例如删除内容或暂停帐号)的必要性和相称性。通过真实报告具体案件，在仇恨言论政策执行方面实现有意

[66] 推特规则和政策 (暴力极端主义团体)。

[67] 脸书社区标准 (危险组织)。

[68] Youtube 政策 (暴力或图片内容政策)。

[69] 见 A/HRC/31/65, 第 39 段。

[70] 脸书社区标准(仇恨言论)；推特规则和政策(仇恨行为政策)。

[71] 大赦国际《推特有毒：毒害妇女》；进步通信协会提交的材料，第 2 页。

[72] 7amleh 和进步通信协会提交的材料，第 15 页。

[73] Ijeoma Oluo 《脸书合谋压制黑人妇女》，Medium, 2017 年 8 月 2 日；通信治理中心提交的材料，第 5 页和进步通信协会提交的材料，第 11-12 页。

[74] 缅甸人权状况特别报告员李亮喜女士在人权理事会第三十七届会议上的发言，2018 年 3 月 12 日。

[75] 进步通信协会提交的材料，第 12 页。

[76] 电子前沿基金会提交的材料，第 5 页。

[77] 同上；进步通信协会和 7amleh 提交的材料。

[78] Betsy Woodruff 《脸书禁止罗辛亚人报道种族清洗》，The Daily Beast, 2017 年 9 月 18 日；ARTICLE 19 提交的材料，第 9 页。

[79] Julia Angwin 和 Hannes Grasseger《脸书的秘密审核保护白人男性而非黑人儿童免受仇恨言论》，ProPublica, 2017 年 6 月 28 日。

义和一贯的透明度，这也可以提供一定程度的见解，即使最详细的说明也无法提供同样的见解。[80]

29. *背景信息*。公司强调在评估一般限制的适用性时务必要考虑背景信息。[81] 然而，对背景信息的关注并未阻止公司删除以下内容：具有历史、文化或教育价值的裸体描绘；[82] 描述冲突的历史和文献记录；[83] 战争罪的证据；[84] 针对仇恨群体的反击言论；[85] 试图挑战或重申种族主义、恐同或仇外言论的内容。[86] 时间和人工审核资源的限制、对自动化的过度依赖或对语言和文化的细微差别所知甚少，可能会限制公司开展有意义的背景信息审查。[87] 公司已敦促用户为有争议的内容补充背景信息，但这一要求的可行性和有效性尚不明确。[88]

30. *实名制要求*。为了应对网上滥用行为，一些公司提出了“真实身份”要求；[89] 其他公司则更为灵活地处理身份问题。[90] 实名制要求作为防止网上滥用行为的保障措施是否有效，这一点值得怀疑。[91]事实上，严格坚持实名制已经暴露了那些使用假名保护自己的博主和活跃分子，使他们面临遭受严重人身伤害的危险。[92] 实名制还封锁了男女同性恋、双性恋、跨性别者和性别奇异者、活跃分子、男扮女装的演员和使用非英文或非常规名称的用户帐号。[93] 由于经常需要网络匿名来保护弱势用户的人身安全，因此人权原则默认保护匿名制，仅遵守能够保护其身份的限制。[94] 制定精细的冒名规则，限制用户以迷惑性或欺骗性手段假扮另一人，这可能是保护其他用户身份、权利和声誉的更适当的方法。[95]

---

[80] 见下文第 52 和第 62 段。

[81] 推特《我们的政策制定方法和执行理念》；Youtube 政策 (背景信息的重要性)；Richard Allan《难题：全球网络社区中应该由谁来决定什么是仇恨言论？》，脸书 Newsroom, 2017 年 6 月 27 日。

[82] OBSERVACOM 提交的材料，第 11 页和 ARTICLE 19 提交的材料，第 6 页。

[83] WITNESS 提交的材料，第 6-7 页。

[84] 同上。

[85] 电子前沿基金会提交的材料，第 5 页。

[86] 进步通信协会提交的材料，第 14 页。

[87] 见 Allan《难题》。

[88] Youtube 政策(背景信息的重要性)；脸书社区标准(仇恨言论)；

[89] 脸书社区标准(使用真实身份)。请注意，脸书现在根据具体情况允许实名政策有例外情况，但批评意见认为这是不够的：Access Now 提交的材料，第 12 页。百度甚至要求使用个人识别信息：百度用户协议。

[90] 推特帮助中心，“用户名注册帮助”；Instagram,“注册 Instagram”。

[91] J. Nathan Matias,“实名制的谬误”，Coral Project, 2017 年 1 月 3 日。

[92] Access Now 提交的材料，第 11 页。

[93] Dia Kayyali, “脸书的姓名政策再次来袭，这次瞄准印第安人”，电子前沿基金会，2015 年 2 月 13 日。

[94] 见 A/HRC/29/32, 第 9 段。

[95] 推特规则和政策(冒名政策)。

31. *虚假信息。*虚假信息和宣传对信息获取构成挑战，也对公众对媒体和政府机构的整体信任度构成挑战。公司面临越来越大的压力，要解决通过链接第三方虚假新闻文章或网站、虚假帐号、欺骗性广告和操纵搜索排名而散布虚假信息的问题。[96] 然而，由于封锁网站或删除特定内容等强硬的行动方式可能会严重干涉表达自由，公司应该仔细制定任何处理虚假信息的政策。[97] 公司已经采取了各种应对措施，包括与第三方事实查证机构达成协议，加大力度执行广告政策，加强监测可疑帐号，修改内容集展和搜索算法以及培训用户识别虚假信息。[98] 一些措施，尤其是加强新闻内容限制的措施，可能会威胁独立和另类新闻来源或讽刺性内容。[99] 从政府当局采取的立场中可以看出，政府似乎过度期待单凭技术力量解决此类问题。[100]

公司内容审核程序和工具

32. *自动举报、删除和发布前过滤。*用户自创内容规模之大，促使各大公司开发了自动审核工具。自动化已经主要被用于举报供人工审核的内容，有时也被用于删除内容。自动扫描工具在音乐和视频上传时进行扫描以发现版权侵权问题，这引发了人们对于过度封锁的关切。还有人呼吁将上传过滤扩大至恐怖主义相关内容和其他内容领域，这有可能导致公司建立全面而过度的发布前审查制度。[101]

33. 自动化可以提供关键词过滤、垃圾邮件检测、哈希匹配算法和自然语言处理等一系列工具，为评估海量用户自创内容的公司带来价值。[102] 哈希匹配被广泛用于儿童性虐待图片的识别，但是识别“极端主义”内容通常需要进行背景信息评估，因此难以在没有明确的“极端主义”规则或人工评估的情况下应用哈希匹配。[103] 自然语言处理也同样如此。[104]

34. *用户和可信的举报。*用户举报使个人能够针对不当内容向内容审核员投诉。举报通常不能对适当的界限进行差别细微的讨论(例如，为什么内容可能具有冒犯性，但总体而言最好予以保留)。[105] 举报还被“操控”，用来增加对各平台的压力，让它们删除支持性少数群体和穆斯林的内容。[106] 许多公司已制定了专门

[96] 同上；Allen Babajanian 和 Christine Wendel，“#虚假新闻：是无伤大雅还是不可容忍？”，威尔顿庄园论坛第 1542 号报告，2017 年 4 月。

[97] 2017 年《联合宣言》。

[98] 进步通信协会提交的材料，第 4-6 页和 ARTICLE 19 提交的材料，第 4 页。

[99] 进步通信协会提交的材料，第 5 页。

[100] 见第 OL ITA 1/2018 号来文。参见欧洲联盟委员会，《处理虚假信息的多角度方法：虚假新闻和虚假信息问题独立高级别小组的最终报告》(卢森堡，2018 年)。

[101] 据报告称，大不列颠及北爱尔兰联合王国开发了一个工具，在上传时自动检测和删除恐怖主义内容。内政部，《发布新技术帮助打击网络恐怖主义内容》，2018 年 2 月 13 日。

[102] 民主和技术中心《混合信息？自动化媒体内容分析的局限性》(2017 年 11 月)，第 9 页。

[103] 开放技术研究所提交的材料，第 2 页。

[104] 民主和技术中心《混合信息？》，第 4 页。

[105] 关于用户举报，见 Kate Crawford 和 Tarleton Gillespie《什么是举报？社交媒体的报告工具和投诉词汇表》，《新媒体与社会》第 18 期第 3 卷(2016 年 3 月)，第 410-428 页。

[106] 同上，第 421 页。

的“可信”举报者名册，这些人通常是专家和高影响力用户，据称有时还有来自政府的举报者。[107] 很少有或根本没有公开信息说明如何甄选专业举报者、他们如何解释法律或社会标准、或者他们对公司决策有何影响。

35. *人工评估*。自动化常常要辅以人工审查，大型社交媒体公司组建了大规模的内容审核员团队，负责审查被举报的内容。[108] 被举报的内容可能被转给内容审核员，他们一般有权决定——通常在几分钟内决定——内容是否适当，是要删除还是批准。如果难以决定特定内容是否适当，审核员可将其上交公司总部的内容团队审查。接下来，将由公司官员——通常是公共政策或“信任与安全”团队偕同总法律顾问——作出删除决定。对于总体或特定情况下的删除讨论，公司披露的信息有限。[109]

36. *帐号或内容方面的行动*。不当内容可能触发一系列的公司行动。公司可以按一个管辖区、一组管辖区、整个平台或一组平台来限制内容删除。它们可以施加年龄限制、予以警告或停用。[110] 违规行为可能导致帐号暂停，而重复违规可能导致帐号注销。在版权执法以外的极少数情况下，公司提供“反通知”程序，允许用户发布质疑删除的内容。

37. *通知*。一个常见的投诉是，发布被举报内容的用户，或者投诉滥用行为的人，都不会收到任何关于删除或其他行动的通知。[111] 即使公司发布通知，通常也只会提及所采取的行动和一般的行动理由。至少有一家公司已经尝试在通知中提供更多的背景信息，但是尚不清楚在批量通知中提供额外详情是否能在所有情况下构成充分的解释。[112] 透明度和通知齐头并进：稳健的业务层面透明度能够提高用户对平台内容删除方法的认识，从而缓解逐个通知的压力，如果总体透明度较低，在没有针对具体个案定制通知的情况下，用户更可能无法理解各项删除。

38. *申诉和补救办法*。各平台允许对一系列行动提出申诉，从个人概况或页面删除到具体文章、照片或视频的删除。[113] 然而，即使可以申诉，用户可用的补救办法似乎是有限的或不及时的，甚至是完全没有补救办法可用，而且在任何情况下，补救办法对于大多数用户甚至是民间社会专家而言都是晦涩的。例如，如果删除对内容发布者造成了具体损害——声誉、身体、道德或财务等损害——那么恢复内容是不足以补救的。同样，在公众抗议或辩论期间暂停帐号或删除内容，可能会对政治权利产生重大影响，公司对此也没有任何补救措施。

---

[107] YouTube 帮助，YouTube 可信赖举报者计划；YouTube 帮助,“参与 YouTube 贡献者计划”；

[108] 见 Sarah Roberts, 《商业内容审核：数字劳动者费力不讨好的工作》，《媒体研究出版物》，论文 12 (2016 年)。

[109] 参见维基百科：大胆更新、修改和讨论的循环。Reddit 鼓励审核员为新用户和有疑惑的用户提供“有用的规则解释、温馨提示和链接”(Reddit Moddiquette 审核指南)。

[110] Youtube 政策(与裸露和色情内容有关的政策)；YouTube 帮助，“创作者在 YouTube 上的影响力”。

[111] ARTICLE 19 提交的材料，第 7 页和进步通信协会提交的材料，第 16 页。

[112] 见 https://twitter.com/TwitterSafety/status/971882517698510848/。

[113] 电子前沿基金会和可视化影响《如何申诉》，onlinecensorship.org。脸书和 Instagram 只允许对暂停帐号提出申诉。参见 Gitbub 提交的材料，第 6 页。

透明度

39. 公司制作了透明度报告，公布政府对内容删除的要求和用户数据方面的汇总数据。这种报告显示了公司面临的那种压力。透明度报告按国别说明依法删除要求的数量、[114] 已采取某种行动或限制内容的要求的数量，[115] 以及越来越多的关于特定法律依据的描述和示例。[116]

40. 然而，正如主要的互联网透明度审查结论所言，公司披露“尽量少的信息来说明如何制定和执行用于自我监管和共同监管的*私营*规则和机制”。[117] 具体而言，公司披露了“极少的”信息说明针对根据服务条款提出的私下删除请求所采取的行动。[118] 内容标准措辞宽泛，为平台酌处权留有余地，而公司没有充分说明这种酌处权。媒体和公众监督促使公司用解释性的博客文章[119] 和有限的假设示例[120] 来补充通用政策，但是缺乏细节说明公司是如何制定和适用内部规则的。[121] 虽然服务条款通常有当地语言版本，但是透明度报告、公司博客和相关内容却没有，对于非英语用户而言更不清晰。因此，用户、公共当局和民间社会常常对根据服务条款所采取行动的不可预测性表示不满。[122] 缺乏充分参与，加上公众批评意见日益增多，迫使公司不断评估、修改和捍卫规则。

四. 适用于公司内容审核的人权原则

41. “脸书”创始人最近表示，希望建立一个能让公司“在不同地方更准确反映本社群价值观”的程序。[123] 该程序和相关标准可参见人权法。私营规范依每家公司的不同商业模式而不同，加上含糊的社区利益主张，导致了不稳定、不可预测和不安全的用户环境和更严格的政府监督。对于那些力求为地理和文化多样的用户群制定共同规范的公司而言，国内法是不合适的。但是，如果人权标准得到透明和一贯的执行，同时参考有意义的用户和民间社会意见，便可提供一个框架，要求国家和公司对跨越国界的用户负责。

---

[114] 《推特透明度报告：删除请求》(2017 年 1-6 月)；《谷歌透明度报告：政府删除内容请求》；2016 年 Reddit 公司《透明度报告》。脸书未按国别提供所收到的请求总数。

[115] 例如，见脸书《透明度报告(法国)》(2017 年 1-6 月)；谷歌《透明度报告：政府删除内容请求(印度)》；推特《透明度报告(土耳其)》。

[116] 同上。

[117] 数字权利排名组织提交的材料，第 4 页。原文斜体。

[118] 同上，第 10 页。

[119] 见 Elliot Schrage《难题概述》，脸书 Newsroom, 2017 年 6 月 15 日；Twitter Safety,“强制执行新规则以减少仇恨行为和滥用行为”，2017 年 12 月 18 日。

[120] 例如，见 Youtube 政策(暴力或图片内容政策)。

[121] Angwin 和 Grasseger,“脸书的秘密审查规则”。

[122] 数字权利排名组织提交的材料，第 10 页；OBSERVACOM, 第 10 页；进步通讯协会，第 17 页；国际图书馆协会联合会，第 4-5 页；Access Now, 第 17 页；欧洲数字权利组织，第 5 页；

[123] Kara Swisher 和 Kurt Wagner, “Recode 网站就剑桥分析咨询公司争议等问题采访脸书首席执行官马克·扎克伯格的采访记录”，Recode, 2018 年 3 月 22 日。

42. 有了人权框架，能够对不当的国家限制予以强有力的规范性回应，前提是公司遵守类似的规则。《指导原则》和配套的“软性法律”文本为公司提供了指导，说明应如何预防或缓解政府对删除内容的过度要求。与此同时还规定了尽职调查、透明度、问责制和补救原则，限制平台通过产品开发和政策制定干涉人权。致力于在各项业务中——而不仅仅在符合公司利益时——落实人权标准的公司在要求各国对相同标准负责时，将有更牢固的依据。此外，如果公司调整服务条款以使其更符合人权法，各国会发现更加难以利用这些公司审查内容。

43. 人权原则还使各公司能够创造一种包容性环境，顾及用户的不同需要和利益，同时建立可预测的和一致的基本行为标准。随着围绕公司是否同时行使媒介和编辑职能的争论日益激烈，人权法向用户作出承诺，他们可以依靠基本规范，超越国内法可能施加的限制，来保护自己的言论。[124] 但是，人权法并不是刻板或教条式的，不会要求公司允许有言论损害他人权利或损害国家保护正当国家安全或公共秩序利益的能力。面对一系列在数字空间中的影响大于线下影响的恶行——例如旨在压制妇女和性少数群体的厌恶女性或恐同骚扰行为，或煽动各种暴力的行为——人权法不会剥夺公司使用工具的权利。相反，它将提供一个全球公认的框架，以制定工具和通用的词汇表，解释这些工具的性质、用途以及对用户和国家的适用。

A. 内容审核的实质性标准

44. 数字时代使快速传播和庞大覆盖面成为可能，但是它不具备人文环境属性。根据《指导原则》，公司在评估内容限制措施的必要性和相称性时，可以考虑各自平台的规模、结构和独特功能。

45. *默认人权*。服务条款应避免一般性和自助服务“社区”需求中根深蒂固的自主决定方式。公司应作出高层政策承诺，以符合人权法的方式，为用户维护平台，让他们在平台上提出意见、自由表达观点和获取各种信息。[125] 这些承诺应支配公司进行内容审核和处理复杂问题(例如计算宣传[126] 和用户数据收集和处理)的方法。公司应将人权法的相关原则直接纳入其服务条款和“社区标准”之中，确保内容相关的行动将遵循合法性、必要性和正当性标准，这些标准同样约束着国家对言论的监管。[127]

46. *“合法性”*。公司规则通常不够清楚和具体，用户无法以合理的确定性预测哪些内容会让他们越线。这在“极端主义”和仇恨言论方面尤其明显，在缺乏严格人工背景评估的情况下，这两个受限领域容易受到过度删除的影响。“具有新闻价值”的一般例外兴起，使公众更难理解特定背景下的规则。[128] 尽管对公众

[124] Global Partners Digital 提交的材料，第 3 页；《指导原则》，原则 11。

[125] 《指导原则》，原则 16。

[126] 见 Samuel Wooley 和 Philip Howard,《全球计算宣传：执行摘要》(计算宣传研究项目第 2017.11 号工作文件) (牛津，2017 年)。

[127] Global Partners Digital 提交的材料，第 10-13 页。

[128] 见 Joel Kaplan,“社区和合作伙伴对于我们的社区标准的意见”，脸书 Newsroom, 2016 年 10 月 21 日；推特规则和政策。

利益的承认受到欢迎，但是公司还应说明在确定公众利益时会评估哪些因素，在评估新闻价值时除了公众利益以外还会考虑哪些因素。公司应付出更多努力，使用展现规则执行趋势的汇总数据、真实案例或展现具体规则解释及适用细节的广泛而详细的假设，更加详细地解释公司规则。

47. *必要性和相称性*。公司不仅应该更加详细地说明有争议的和具体背景下的规则，还应该披露资料和示例，深入说明在确定暴力行为及其严重程度和所采取的应对行动时会评估哪些因素。在仇恨言论方面，说明公司是如何解决具体案件的，这可能有助于用户更好地了解公司如何区分难以区分的冒犯性内容和煽动仇恨的内容，或者如何评估网络环境中说话者的意图或暴力行为的可能性。关于所采取行动的精确数据也将成为一个依据，以评估公司在制定限制措施方面的严谨程度。公司应说明在哪些情况下适用侵扰程度较低的限制(例如警告、年龄限制或废止)。

48. *不歧视*。有意义的不歧视保障措施要求公司超越形式主义办法。形式主义办法将所有受保护对象视作同样易受虐待、骚扰和其他形式的审查。[129] 事实上，这种办法似乎有违公司对背景信息的重视。其实，公司在开发或修改政策或产品时，应积极征求并考虑一贯易受审查和歧视的社区的关切。

B. 公司内容审核和相关活动的程序

回应政府请求

49. 正如公司透明度报告显示，各国政府施加压力，要求它们删除内容、暂停帐号并识别和披露帐号信息。在当地法律的要求下，公司似乎别无选择，唯有遵守要求。但是，公司可以开发工具，用于预防或缓解由于不符合国际标准的国家法律或要求造成的人权风险。

50. *预防和缓解*。公司往往声称自己认真对待人权。然而，公司在内部作出这种承诺，在出现争议时向公众作出专门的保证，这是不够的。公司还应在最高领导层通过并公布具体政策，“要求所有业务单位，包括各地子公司以有利于尊重表达自由、隐私和其他人权的方式处理法律的模糊之处”。这些承诺应产生政策和程序，用于解释和执行政府要求，缩小并“确保施加最少的内容限制”。[130] 公司应确保限制请求是书面的、援引了具体和有效的法律依据、由合法的政府机关以适当形式发布的。[131]

51. 如果遇到有问题的请求，公司应力求予以澄清或修改；寻求民间社会、同行公司、有关政府机关、国际和区域性机构以及其他利益攸关方的协助；探索可以质疑请求的所有法律手段。[132] 当公司收到国家根据公司服务条款或通过其他法外手段提出的请求时，应把这些请求转交法律合规程序，根据相关的地方法律和人权标准评估其有效性。

[129] 例如，见《消除一切形式种族歧视国际公约》第一条第四款和第二条(丑)款。

[130] 见 A/HRC/35/22, 第 66-67 段。

[131] 全球网络倡议提交的材料，第 3-4 页和 GitHub 提交的材料，第 3-5 页。

[132] 见 A/HRC/35/22, 第 68 段。

52. *透明度*。只有公司和国家之间的互动真正透明，用户才能在面对审查及相关人权风险时作出知情决定，决定是否以及如何参与社交媒体。应制定关于如何实现这种透明度的最佳做法。公司在报告国家请求时应辅以精确数据，说明所收到的请求类型(例如，诽谤、仇恨言论、与恐怖主义有关的内容)和采取的行动(例如，部分或全部删除、在特定国家或全球删除、暂停帐号、根据服务条款进行删除)。公司还应该尽可能经常提供具体例子。[133] 透明度报告应扩大到政府根据公司服务条款提出的要求[134]，还必须说明公私合作限制内容的举措，例如《关于打击网上非法仇恨言论的欧盟行为守则》、互联网转介小组和双边谅解等政府举措，例如据称的 YouTube 和巴基斯坦之间的谅解以及“脸书”和以色列之间的谅解。公司应记录根据这些举措提出的请求以及公司和请求方之间的沟通内容，还应尝试安排把这些请求的副本交由第三方保管。

规则制定和产品开发

53. *尽职*。虽然一些公司在评估它们如何应对国家提出的限制时承诺在人权方面做到尽职，但是尚不清楚它们是否实施了类似的保障措施以预防或缓解因自己制定和执行的政策而对表达自由构成的风险。[135] 公司应制定明确而具体的标准，用于确定哪些活动会触发这种评估。除了修订内容审核政策和程序以外，还应评估用户信息流和其他内容传送形式的策展、新功能或服务的发布、对现有功能或服务的修改、自动化技术的开发和入市决定(如为不同国家提供特定版本平台的安排)等情况。[136] 以往的报告还具体说明了评估应该检查的问题，以及为了将评估及其结论纳入相关业务所需的内部程序和培训。此外，这些评估应该是持续的，能够适应情况或运营环境的变化。[137] 多利益攸关方倡议，例如全球网络倡议提供了一个途径，使公司能够制定并完善评估程序和其他尽职程序。

54. *公众意见和参与*。参与协商的人一直提出的关切是，公司未能充分与用户和民间社会互动，在全球南部国家尤其如此。受影响的权利持有人(或其代表)和相关的地方或主题专家等的意见，以及切实采纳反馈意见的内部决策程序，都是尽职的组成部分。[138] 协商——特别是广泛的协商，例如征求公众意见——，使公司能够从多个角度考虑其活动对人权的影响，同时还能鼓励公司密切关注看似无害或表面上“社区友好”的规则是如何对社区造成重大和“超地方化”影响的。[139] 例如，与来自不同地理区域的各种土著群体互动，可以帮助公司制定更完善的指标，在评估裸露内容时考虑到文化和艺术背景。

---

[133] 例如，见《推特透明度报告：删除请求》(2017 年 1-6 月)；

[134] 推特已经开始公布数据，说明“由已知政府代表提交的、关于可能违反推特规则的内容的非法律请求”，禁止滥用行为、宣传恐怖主义和知识产权侵权。出处同上。另见微软《内容删除请求报告》(2017 年 1-6 月)。

[135] 数字权利排名组织提交的材料，第 12 页；《指导原则》，原则 17。

[136] 见 A/HRC/35/22, 第 53 段。

[137] 同上，第 54-58 段。

[138] 见《指导原则》，原则 18 和 A/HRC/35/22, 第 57 段。

[139] Chinmayi Arun《重新平衡言论监管：全球网络平台中的超地方内容》，哈佛大学伯克曼互联网与社会研究中心 Medium 文章汇编，2018 年；*Pretoria News*,《抗议谷歌和脸书“欺凌”裸露的女仆》，2017 年 12 月 14 日。

55. *规则制定的透明度*。公司似乎总是在不开展人权尽职调查或不评估实际影响的情况下修改产品和规则。它们至少应该就这些修改的影响征求感兴趣的用户和专家的意见，如有必要，应在保证对这种评估保密的情况下征求意见。它们还应明确向公众说明这些修改背后的规则和程序。

规则执行

56. *自动化和人工评估*。自动化内容审核是用于大规模和大范围用户自创内容的一项功能，会导致内容行动与人权法不一致的明显风险。在公司预防和缓解人权影响的责任中，应当考虑到自动化的显著局限性，例如难以处理背景信息、广泛的语言线索和含义变化以及语言和文化的特殊性。自动化的实行，如基于在公司所在国形成的理解，可能会导致对全球所有用户群的严重歧视。最起码，为处理规模问题而开发的技术应该经过严格检查，开发过程中应该广泛参考用户和民间社会的意见。

57. 公司有责任促进准确、对背景信息敏感和尊重表达自由的内容审核做法，这也要求公司加强和保证对被举报内容的专业化人工评估，其中应包括加强对人工审核员的保护措施，使其符合适用于劳动权利的人权规范，同时郑重承诺在公司开展业务的每一个市场中都具备文化、语言及其他形式的专业知识。公司还应该有多样化的领导层和政策团队，以便能够将地方或主题事项专业知识应用于内容问题。

58. *通知和申诉*。用户和民间社会专家纷纷表示关切的是，被删除内容或暂停帐号或注销帐号的人获得的信息有限，举报滥用行为，例如举报厌恶女性的骚扰行为和人肉搜索行为的人获得的信息也有限。信息的缺乏为秘密规范创造了环境，这不符合清晰性、具体性和可预测性标准。这干涉了个人质疑内容行动或跟进内容相关投诉的能力；然而，在实践中，没有稳健的内容删除申诉机制，这种情况更加有利于举报者，而非内容发布者。有人可能争辩说，如果允许对每一个内容行动提出申诉，将是耗时和昂贵的。但是，各公司可以相互合作并与民间社会合作，共同探讨可扩展的解决办法，例如特定公司或全行业的监察员方案。在有关这类方案的构想之中，最好的一个是仿照新闻理事会建立独立的“社交媒体理事会”，实现全行业的投诉机制，促进对违规行为的补救措施。[140] 这一机制可以听取符合标准的个人用户申诉，针对反复出现的内容审核问题收集公众反馈意见，例如对某一特定主题领域过度审查的问题。各国应支持遵照人权标准运作的、可扩展的申诉机制。

59. *补救措施*。《指导原则》强调了补救“负面影响”的责任。然而，几乎没有任何公司规定补救。公司应当制定有力的补救方案，其中可包括恢复内容和确认关于名誉或其他损害的和解办法。有几家公司在内容规则方面已经进行了一些融合，从而让公司之间有可能合作，通过社交媒体理事会、其他的监察员方案或第三方裁决提供补救措施。如果公司继续不提供补救，可能将需要立法和司法干预。

[140] 见 ARTICLE 19,《社交媒体平台中的自我监管和“仇恨言论”》(伦敦，2018 年)，第 20-22 页。

60. *用户自治*。公司已经开发了工具，使用户能够塑造属于自己的网络环境。这包括对其他用户或特定类型的内容禁声和屏蔽。同样，各平台通常允许用户创建封闭或私人群组，由用户自行主持。虽然封闭群组的内容规则应符合基本人权标准，但是考虑到这种基于亲近关系的群组在保护见解自由、为弱势社区扩大空间和验证有争议或不受欢迎的想法等方面的价值，各平台应予以鼓励。考虑到对弱势个人的隐私和安全的影响，不应推崇实名制要求。[141]

61. 对于网络信息可核查性、相关性和有效性的关切越来越多，引发了公司应如何尊重信息获取权这一复杂问题。公司至少应披露策展方法的细节。如果公司基于用户之间的互动对社交媒体消息流中的内容进行排序，那么它们应该解释所收集的互动数据及其对排名标准的作用。公司应向所有用户提供可用和有意义的机会，让他们可以拒绝接受平台主导的策展。[142]

决策透明度

62. 尽管政府删除请求的整体透明度有所提升，但是公司大多未报告根据服务条款采取的行动。公司没有公布数据说明基于服务条款的私下请求数量和类型，更不用说已执行请求的占比。公司应制定透明度举措，以说明自动化、人工审核和用户或可信举报对于依照服务条款采取的行动有何影响。尽管有少数公司正在开始提供关于这些行动的一些信息，但是整个行业应该着手提供更多细节，说明有代表性的具体案件以及在公司政策解释和执行方面的重大进展。

63. 公司正在执行“平台法”，对内容问题采取行动，而不大量透露行动详情。理想情况是，公司应制定一种判例法，让用户、民间社会和各国能够了解公司是如何解释和执行其标准的。虽然这种“判例法”制度不会涉及公众所期待的那类法院和行政机构报告，但是详细的案件和实例资料库同案例报告一样，都能够澄清规则。[143] 一个有权评估整个信通技术部门投诉的社交媒体理事会有可能成为可信和独立的机制，以实现这种透明度。

五. 建议

64. 模糊的因素正在影响着全球个人行使表达自由的能力。此时此刻需要有彻底的透明度、有意义的问责制和补救承诺，以保护个人的能力，让他们能够把在线平台作为自由表达、获取信息和参与公共生活的论坛。本报告确定了一系列措施，具体如下。

对国家的建议

65. 各国应废除任何将线上或线下言论刑罪化或加以不当限制的法律。

66. 明智的监管，而非基于观点的严厉监管，应该成为常态，前者着重于确保公司的透明度和补救措施，使公众能够选择如何以及是否参与在线论坛。各国只应

[141] 见上文第 30 段。

[142] 例如，脸书允许用户按时间顺序倒序浏览信息流中的故事，但同时警告称“最终”将恢复到默认的内容策展设置。脸书帮助中心，“信息流中热门故事和最新故事之间有何差异？”

[143] 例如，见 Madeleine Varner 等人，“脸书如何看待仇恨言论？”，ProPublica, 2017 年 12 月 28 日。

设法根据独立公正的司法机关的命令限制内容，同时遵守正当程序和合法性、必要性和正当性标准。各国应避免对互联网媒介施加过度的处罚，不论是高额罚款还是监禁，因为过度的处罚会对表达自由产生严重的寒蝉效应。

67. 各国和各政府间组织应避免制定要求“主动”监测或过滤内容的法律或协议，这种要求既违反了隐私权，又有可能构成发表前审查。

68. 各国应避免采取这种监管模式，即政府机构而非司法机关成为合法言论问题的仲裁机构。各国应避免将内容裁决的责任交给公司，否则公司将有权逾越人权价值观，作出损害用户的判决。

69. 各国应公布详细的透明度报告，说明对媒介提出的所有内容相关请求，并在所有的监管考虑中纳入真实的公众意见。

对信通技术公司的建议

70. 公司应当认识到，确保平台表达自由的全球权威标准是人权法，而不是各异的国家法律或公司的私利，公司应当相应重新评估其内容标准。人权法赋予公司工具，以尊重民主规范和抵抗独裁命令的方式阐述和制定政策和程序。此方法从以权利为基础的规则入手，接着严格评估产品开发和政策制定的人权影响，并在各项业务中持续开展评估、再评估以及有意义的公众和民间社会协商。《工商企业与人权指导原则》以及由民间社会、政府间机构、全球网络倡议和其他各方针对具体行业制定的准则，提供了各互联网公司均应采纳的基准方法。

71. 互联网公司必须在所有运营阶段，从规则制定到私营规则解释所依据的“判例法”的实施和制订，着手就透明度采取截然不同的方法。要想实现透明度，必须有数字权利组织和民间社会其他相关部门的更多参与，还要避免公司与国家就内容标准及其执行达成秘密安排。

72. 鉴于公司对公共领域的影响，它们必须接受公共问责制。世界各地有效和尊重权利的新闻理事会提供了一种模式，用于在商业内容审核中实现最起码的一致性、透明度和问责制。第三方非政府手段，如果以人权标准为根基，可以提供申诉和补救机制，无需过高成本，不会让较小规模的实体或新的市场参与者望而却步。信通技术部门中负责审核内容或作为把关者行事的各方应把制定全行业的问责机制(例如社交媒体理事会)作为当务之急。