人工智能时代的恐怖主义风险与治理路径

□ 徐海娜

〔提　　要〕生成式人工智能的迅猛发展正深刻重塑恐怖主义的运作模式与反恐治理逻辑，其在内容生成、深度伪造、算法推荐、自主决策等领域的滥用，使恐怖活动呈现智能化、低门槛化、跨境化特征，持续冲击全球安全体系。当前，人工智能赋能恐怖主义的主要风险体现在技术赋能、法律失衡、政治安全威胁、国际合作困境四个关键维度，表现为极端主义传播加剧、法律归责模糊、社会韧性削弱、国际协作受阻等突出问题。国际社会正探索以法治、协作、社会韧性为支点的系统化治理路径，通过强化责任认定与监管规范填补法律空白，推动多边机制衔接与技术标准协调，并在公众层面加强“数字免疫”、培育可信叙事。生成式人工智能的安全与可控治理，不仅关乎反恐成效，更关乎全球安全秩序的可持续稳定。

〔关 键 词〕生成式人工智能、恐怖主义、反恐治理、全球安全、社会韧性

〔作者简介〕徐海娜，山东大学国际问题研究院副教授

〔中图分类号〕D815.5

〔文献标识码〕A

〔文章编号〕0452 8832（2026）2 期 0113-24

人工智能（AI）技术的迅猛发展，正深刻改变全球安全风险的结构性特征。特别是生成式人工智能（Generative AI，以下简称生成式 AI）技术的

快速突破，使深度合成内容（AIGC）大规模进入安全风险视野，在语言生成、图像伪造与行为模拟等领域对信息秩序与心理安全构成深远冲击。

作为一种典型的非国家暴力行为，恐怖主义一直高度依赖新兴技术手段，以提升组织动员、资源获取与行动破坏力。随着生成式 AI 与大数据的结合，恐怖分子成体系地使用深度伪造、算法推荐、自动化操控等工具，借助此类工具强化传播策略与攻击手段，拓展全球影响路径，从“网络动员”转向“算法赋能”，对全球安全体系构成复合性挑战。在技术与恐怖主义深度交织的背景下，恐怖主义正呈现更强的隐蔽性、传播性与精准性，传统反恐机制面临效能瓶颈，国际社会亟需重估既有治理范式与防控逻辑，推动构建多主体参与、跨层级联动的国际合作机制，以回应新型非传统安全威胁的现实挑战。

一、生成式人工智能背景下恐怖主义的演化

作为 AI 的一个重要分支，生成式 AI 通过大规模模型生成文本、图像、音视频等内容，典型形式包括深度伪造、自动化文本生成与图像合成等。在实际运用中，恐怖主义往往将生成式 AI 与数据挖掘、目标识别算法等其他传统 AI 工具结合使用，二者在功能上存在交叉与融合。[1] 随着恐怖分子对生成式 AI 认知与使用能力的显著提升，AI 技术正深度重塑恐怖主义的运作逻辑，而非仅仅是一种外部工具。由 AI 赋能的恐怖主义活动或实现结构性升级，对全球安全格局构成持续挑战。

（一）AI 扩大恐怖袭击规模与破坏力

生成式 AI 的广泛普及深刻改变非国家暴力行为的技术生态，推动恐怖主义实现从低成本走向高杀伤，从远程操控走向自主行动，从随机袭击走向精准打击的三重跃迁。尽管这一趋势并非 AI 独有，如冷战后恐怖组织对卫星

[1] 有鉴于此，本文在绝大多数分析中使用“生成式 AI”，仅在涉及人工智能总体发展趋势或宏观治理议题时使用“AI”指代更广义的技术范畴。

通信和互联网的利用同样带来跨境性与隐蔽性的提升，[1] 但生成式 AI 加速和放大了技术扩散，使全球安全治理面临前所未有的压力。

第一，生成式 AI 显著降低网络攻击与信息操控的门槛。生成式 AI 已被网络犯罪团体用于自动生成恶意代码、制作钓鱼邮件和伪造音视频内容，相关工具甚至在暗网上以即用方式出售，显著扩展复杂攻击的可及性。[2] 上述行为若出于政治或意识形态目的，便进入网络恐怖主义范畴。恐怖主义与网络犯罪在工具、渠道和地下市场上存在高度重叠，恐怖组织往往直接借用或购买网络犯罪分子开发的恶意脚本与生成式工具。[3] 正因如此，网络犯罪的技术滥用趋势常常预示恐怖主义的未来方向，[4] 生成式 AI 正加速犯罪与恐怖主义的交叉风险。[5] 已有实证案例显示这种跨界趋势正在发生。例如，部分“伊斯兰国”支持者开始利用生成式 AI 将阿拉伯语宣传信息翻译成英语与印尼语，并生成图像与海报以扩大跨国传播与招募范围。[6] 附属于“伊斯兰国”的“News Harvest”虚拟主播项目用生成式 AI 制作低成本、专业化的宣传内容，[7] 显著提升其视觉传播力与情绪感染力。这些案例显示，生成式 AI 正成为极端主义传播与动员的重要技术支撑。

第二，生成式 AI 加速恐怖袭击从远程操控向自主执行的转型。早在

---

［1］ John Arquilla and David Ronfeldt，eds.，*Networks and Netwars: The Future of Terror, Crime, and Militancy*，RAND Corporation，2001，pp.59-63，https://www.rand.org/pubs/monograph_reports/MR1382.html.

［2］ UN Interregional Crime and Justice Research Institute and UN Counter-Terrorism Centre，“Beneath the Surface: Terrorist and Violent Extremist Use of the Dark Web and Cybercrime-as-a-Service for Cyber-Attacks，” pp.2-3，https://www.un.org/counterterrorism/sites/default/files/dw_beneath_the_surface_update.pdf.

［3］ Clarisa Nelu，“Exploitation of Generative AI by Terrorist Groups，” International Centre for Counter-Terrorism (ICCT)，June 10，2024，https://icct.nl/publication/exploitation-generative-ai-terrorist-groups.

［4］ *Networks and Netwars: The Future of Terror, Crime, and Militancy*，pp.59-63.

［5］ Europol，“The Changing DNA of Serious and Organised Crime: EU-SOCTA 2025，” pp.5-6，https://www.europol.europa.eu/publication-events/main-reports/internet-organised-crime-threat-assessment-iocta-2025.

［6］ Clarisa Nelu，“Exploitation of Generative AI by Terrorist Groups.”

［7］ Pranshu Verma，“These ISIS News Anchors Are AI Fakes，Their Propaganda Is Real，” The Washington Post，May 17，2024，https://www.washingtonpost.com/technology/2024/05/17/ai-isis-propaganda.

2010 年代，“伊斯兰国”等组织已将商用无人机武器改造用于监视、投掷炸弹、宣传和心理战，[1] 反映智能化作战的趋势。2020 年，具备“发射后自主搜索打击”能力的自主攻击无人机在利比亚战场首次出现，被视为自主武器进入实战的先例。[2] 虽然该类技术主要被掌握在国家层面，但商用无人机的普及与低成本改装，使极端组织能够搭载简易算法，实现定向打击或群体袭击。

第三，生成式 AI 推动恐怖袭击目标实现从随机化到精准化的转变。恐怖组织借助社交媒体信息挖掘与深度伪造技术，可快速锁定特定政治立场、职业身份乃至特定情绪状态的群体。极端组织借助多语种生成与虚拟主播视频，扩大了跨文化招募范围，并通过算法推荐精准锁定高风险受众，显著提升舆论操控与心理战效能。[3]2024 年莫斯科“克罗库斯城”音乐厅恐袭事件后，“伊斯兰国”支持者利用 AI 生成视频进行密集的舆论操控与在线招募，其传播速度远超传统模式。[4] 极端组织已尝试使用大语言模型生成简化版袭击指南与宣传口号，使网络行动更高效、更隐蔽。[5]

（二）AI 驱动目标识别与打击决策

生成式 AI 的快速发展，赋予恐怖主义更强的目标锁定、情报搜集与战术规划能力，推动恐怖主义从“广泛撒网、多重打击”向“定点锁定、精确实施”的根本跃迁，对社会心理、舆论认知及政治稳定的侵蚀力显著增强，

---

［1］ Kerry Chávez and Ori Swed, “Off the Shelf: The Violent Nonstate Actor Drone Threat,” *Air & Space Power Journal*, 2020, pp.29-43; Emil Archambault and Yannick Veilleux-Lepage, “The Islamic State Drone Program,” in James Patton Rogers, ed., *De Gruyter Handbook of Drone Warfare*, Berlin: De Gruyter, 2024, pp.243-254; Don Rassler, “Remotely Piloted Innovation: Terrorism, Drones and Supportive Technology,” Combating Terrorism Center at West Point, October 2016, pp.34-38, https://ctc.westpoint.edu/wp-content/uploads/2016/10/Drones-Report.pdf.

［2］ “Final Report of the Panel of Experts on Libya Established Pursuant to Security Council Resolution 1973 (2011),” UN Doc.S/2021/229, March 8, 2021, para.63, https://undocs.org/S/2021/229.

［3］ Clarisa Nelu, “Exploitation of Generative AI by Terrorist Groups.”

［4］ Pranshu Verma, “These ISIS News Anchors Are AI Fakes, Their Propaganda Is Real.”

［5］ Gabriel Weimann et al., “Generating Terror: The Risks of Generative AI Exploitation,” Combating Terrorism Center at West Point, https://ctc.westpoint.edu/generating-terror-the-risks-of-generative-ai-exploitation/.

也使反恐体系在识别与溯源方面面临更高技术门槛。

第一，生成式AI及相关算法极大扩展极端组织的数据挖掘与分析能力，可批量处理社交媒体与公开数据库信息，精准识别特定地区或群体，已成为极端主义者筛选特定群体、定向传播极端思想的重要工具。[1]

第二，生成式AI与深度学习算法被用于恐怖组织的战术辅助和资源优化，强化恐怖组织的推演与反追踪能力，使其行动更具隐蔽性与效率。极端组织正探索通过生成式AI进行动态战术规划，不仅生成行动方案、规避侦察机制，还利用模型优化袭击路径。[2]

第三，生成式AI推动恐怖主义武器化手段的低门槛智能化转型。通过与开源硬件结合，恐怖组织可低成本升级商用无人机、遥控爆炸装置，使其具备自动导航和目标锁定功能。据报道，利比亚冲突中疑似首次出现AI自主无人机攻击，反映生成式AI与廉价无人系统结合后，恐怖主义具备低门槛、高智能的作战能力。STM Kargu-2等无人武器系统无需数据连线即可识别并攻击目标，使恐怖组织能够在极短周期内实现智能化升级。[3]

第四，生成式AI正成为恐怖主义对抗反恐监控体系的重要工具。极端组织利用生成式AI实施敏感词清洗、通信伪装与身份深伪，显著削弱安全机构的监测与溯源能力。生成式AI还被用于训练模型以规避社交平台内容检测，从而增强行动隐蔽性。[4] 同时，生成式AI与深度伪造（deepfake）技术的结合，使恐怖组织能够实施“无痕传播”（trace-free dissemination）和“深度伪装”（deep camouflage），在身份、通信与内容层面多重隐匿，对传统

---

[1] “Violent Extremists' Use of Generative Artificial Intelligence,” US National Counterterrorism Center, May 2024, https://www.dni.gov/files/NCTC/documents/jcat/firstresponderstoolbox/151s_First_Responders_Toolbox-Violent_Extremists_Use_of_Generative_Artificial_Intelligence.pdf.

[2] Miles Brundage et al., “The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation,” February 2018, pp.32-36, https://arxiv.org/pdf/1802.07228.pdf.

[3] “Final Report of the Panel of Experts on Libya Established Pursuant to Security Council Resolution 1973 (2011).”

[4] Asha Hemrajani, “The Use of AI in Terrorism,” RSIS Commentary, August 25, 2024, https://rsis.edu.sg/rsis-publication/rsis/the-use-of-ai-in-terrorism/.

反恐监测体系构成实质挑战。[1]

（三）AI 驱动极端叙事扩散与认知操控

生成式 AI 将恐怖主义宣传转变为高速度、大规模与深度操控的新型信息战，使恐怖主义传播更具规模性、隐蔽性与个性化，对传统的检测、澄清与溯源机制形成系统性挑战。

第一，生成式 AI 显著提高极端内容的产出与扩散能力。恐怖组织可通过文本生成、图像伪造、视频自动化手段批量制造多语种宣传材料。对 Telegram 等平台的监测报告指出，极端分子通过生成式 AI 工具批量生成多语种内容，显著扩大传播密度。[2] 联合国反恐办公室与联合国区域间犯罪和司法研究所的研究也发现，生成式 AI 的本地化与自动发布功能正扩大极端主义的跨语种传播。[3]

第二，算法推荐与情绪识别使极端信息实现微定向推送。基于推荐算法与情绪分析的个性化推送，使恐怖组织能精准触达易激进化的个体，形成针对性极强的激进化路径。监测研究表明，极端分子利用平台算法漏洞和开源模型，可定制化内容以加强说服力与招募效率。[4]

第三，深度伪造、情绪操控与虚假叙事正削弱公众的认知防御力。深度伪造结合情绪化叙事，可制造情感真实但事实虚假的沉浸式内容，快速引发

---

［1］ Maggie Engler，“Considerations of the Impacts of Generative AI on Online Terrorism and Extremism，” Global Internet Forum to Counter Terrorism，September 2023，pp.5-6，https：//gifct.org/wp-content/uploads/2023/09/GIFCT-23WG-0823-GenerativeAI-1.1.pdf.

［2］ “Early Terrorist Experimentation with Generative AI Services，” Tech Against Terrorism，2023，pp.3-5，https：//techagainstterrorism.org/hubfs/Tech%20Against%20Terrorism%20Briefing%20-%20Early%20terrorist%20experimentation%20with%20generative%20artificial%20intelligence%20services.pdf.

［3］ UN Interregional Crime and Justice Research Institute and UN Counter-Terrorism Centre，“Countering Terrorism Online with Artificial Intelligence：An Overview for Law Enforcement and Counter-Terrorism Agencies in South Asia and South-East Asia，” 2021，pp.17-19，https：//unicri.org/sites/default/files/2021-06/Countering%20Terrorism%20Online%20with%20AI%20-%20UNCCT-UNICRI%20Report.pdf.

［4］ Nasir Ahmad Ganaie，“The Role of Artificial Intelligence in Radicalisation，Recruitment and Terrorist Propaganda，” Frontiers in Political Science，January 6，2026，https：//www.frontiersin.org/journals/political-science/articles/10.3389/fpos.2025.1718396/full.

恐慌，分裂社会认知。平台的情绪识别与推荐机制还会放大这些内容的传播效应，使易感群体在短时间内被高度曝光与动员。

二、生成式人工智能背景下的反恐治理挑战

在 AI 加速演进的时代背景下，技术与恐怖主义行动的深度融合催生多维度的新型威胁形态，对现有全球反恐治理体系构成严峻考验。

（一）传统反恐模式遭遇技术挑战

恐怖主义在生成式 AI 等前沿技术驱动下呈现新的演变趋势，对全球安全构成前所未有的冲击。面对复杂化的恐怖组织结构和多元化的袭击方式，传统人力反恐模式日益受限。具体而言，生成式 AI 从技术扩散低门槛化、内容监控失效、情报预警困难、资金追踪受阻四个环节，系统性削弱传统反恐模式的效能。

第一，生成式 AI 推动技术扩散与“恐怖主义服务化”（terrorism as a service），显著降低恐怖主义行动门槛。恐怖分子可借助即用型工具快速完成袭击筹备与执行。例如，有研究人员测试发现，仅需约 200 美元即可从 Telegram 黑市购得莫斯科市面部识别系统的访问权限并获取实时数据，充分说明此类敏感技术资源对非授权行为者的可获取性之高。[1] 研究显示，恐怖组织正将简易算法搭载于无人机，实现定向或群体袭击，形成低成本“无人化恐袭”新模式。[2] 这些趋势表明，生成式 AI 的可获取性与去中心化特征正加剧恐怖主义的隐蔽化和跨境化，使传统依赖人工监控与单一打击的反恐模式愈发难以奏效。

第二，生成式 AI 的发展抬升全球反恐治理门槛，暴露其在应对深度伪

[1] Russell Brandom, “Moscow’s Facial Recognition System Can Be Hijacked for Just $200, Report Shows,” The Verge, November 11, 2020, https://www.theverge.com/2020/11/11/21561018/moscows-facial-recognition-system-crime-bribe-stalking.

[2] Don Rassler and Yannick VeilleuxLepage, “On the Horizon: The Ukraine War and the Evolving Threat of Drone Terrorism,” Combating Terrorism Center at West Point, https://ctc.westpoint.edu/on-the-horizon-the-ukraine-war-and-the-evolving-threat-of-drone-terrorism/.

造与虚假信息方面的短板。机器学习和深度合成技术被广泛应用于伪造音视频，显著降低虚假信息生产成本，增加网络恐怖主义隐蔽性。极端组织成员已逐步将生成式 AI 工具用于内容制作与行动策划。[1] 与此同时，犯罪分子借助 AI 语音合成与深度伪造技术伪装身份的实际案例，已被极端组织视为可资借鉴的技战术扩散路径。[2] 调查表明，生成式 AI 聊天机器人在识别虚假叙事时失误率高达 80%，且风险仍在持续。[3]

第三，上述威胁的持续蔓延，进一步暴露了现有反恐预警体系的结构性缺陷。受限于数据处理与预测能力，传统预警体系难以及时识别生成式 AI 驱动的潜在威胁。加密工具与个性化搜索为恐怖分子提供更隐蔽的资源获取途径。地下论坛用户可绕过 ChatGPT 限制生成恶意软件，使无编程技能者也能生产“即插即用”的攻击脚本，进一步普及网络攻击。[4] 与此同时，生成式 AI 被直接用于宣传动员。极右翼极端组织通过 Telegram 等去中心化平台实现宣传内容的自传播与动态更新，尽管平台持续采取遏制措施，相关网络仍持续扩张。[5]

第四，生成式 AI 与加密金融技术的交叉加速，显著增加恐怖主义资金流转的隐蔽性与复杂性。随着生成式 AI 滥用扩散至金融犯罪领域，在身份欺诈

---

［1］ “GIFCT Artificial Intelligence Working Group Report: AI Threats and Opportunities,” Global Internet Forum to Counter Terrorism, December 2025, pp.7-9, https://www.gifct.org/wp-content/uploads/2025/12/2025-GIFCT-AIWG-Output.pdf.

［2］ “Facing Reality? Law Enforcement and the Challenge of Deepfakes,” Europol, 2024, pp.10-14, https://www.europol.europa.eu/publications-events/publications/facing-reality-law-enforcement-and-challenge-of-deepfakes.

［3］ “The Next Great Misinformation Superspreader: How ChatGPT Could Spread Toxic Misinformation at Unprecedented Scale,” NewsGuard, January 2023, https://www.newsguardtech.com/misinformation-monitor/jan-2023.

［4］ “Cybercriminals Bypass ChatGPT Restrictions to Generate Malicious Content,” Check Point Research, February 7, 2023, https://blog.checkpoint.com/2023/02/07/cybercriminals-bypass-chatgpt-restrictions-to-generate-malicious-content/.

［5］ “Beyond the Collective: Understanding Terrorgram’s Efforts to Infiltrate the Mainstream on Telegram,” Institute for Strategic Dialogue, August 2024, https://www.isdglobal.org/digital-dispatch/beyond-the-collective-understanding-terrorgrams-efforts-to-infiltrate-the-mainstream-on-telegram-2/.

与合成内容制造中的应用持续增长，该技术风险已从信息安全蔓延至金融体系。恐怖组织利用 AI 驱动的算法交易、地址混淆与数据伪造手段，创建虚假账户与钱包结构，在加密资产体系中自动化完成匿名筹资与多层转移。[1] 同时，生成式 AI 的加密策略加剧了恐怖分子依赖加密货币与去中心化金融（DeFi）进行资源筹集的趋势。多个暗网与加密社区论坛活跃讨论 AI 辅助套利、匿名化资金流转和“黑客即服务”（HaaS）等非法服务，部分帖文已被安全公司追踪。[2] 总体而言，生成式 AI 重塑恐怖组织的金融隐匿与筹资模式，大大提升跨国反恐金融追踪的难度。

（二）法律滞后与伦理标准失衡

在法律与伦理领域，生成式 AI 与恐怖主义的深度结合对全球安全构成双重挑战。一方面，法律治理体系面临滞后与碎片化问题，现有规则责任归属模糊，法律适用存在显著空白，难以有效应对 AI 赋能恐怖主义所引发的新型威胁。另一方面，AI 滥用不仅加剧恐怖袭击对人类社会道德底线的冲击，还深刻撕裂反恐伦理与安全底线，进一步加剧安全与权利之间的伦理张力。法律责任归属模糊与伦理标准失衡已演变为 AI 时代恐怖主义扩散背景下全球治理的薄弱环节。

现有法律治理体系在应对 AI 赋能恐怖主义时存在困境，首先体现在法律框架普遍存在的滞后与空白。多数国家现行法律主要针对传统恐怖主义和网络犯罪设计，面对生成式 AI 引发的新型威胁缺乏有效规范。例如，美国《爱国者法案》聚焦常规恐怖活动，对 AI 和深度伪造等前沿技术缺乏针对性条款，导致法律工具滞后，执法依据不足。[3] 中国近年来虽出台《互联网信息服务算法推荐管理规定》，禁止利用算法生成虚假信息，但立法重点依然集中在

---

［1］ Chainalysis，“2025 Crypto Crime Report，” January 15，2025，https://www.chainalysis.com/blog/2025-crypto-crime-report-introduction.

［2］ Elliptic，“Elliptic Typologies Report 2023，” https://www.elliptic.co/resources/elliptic-typologies-report-2023；Rena S. Miller，Liana W. Rosen and Paul Tierno，“Terrorist Financing：Hamas and Cryptocurrency Fundraising，” Congressional Research Service，December 9，2024，pp.2-3，https://crsreports.congress.gov/product/pdf/IF/IF12537.

［3］ “USA PATRIOT Act，” US Congress Public Law 107-56，October 26，2001，https://www.congress.gov/bill/107th-congress/house-bill/3162.

信息安全与平台治理，尚未形成针对生成式 AI 滥用的系统性监管框架。[1] 此外，受限于国家治理模式与技术迭代速度，全球范围内不少监管提案最终难以立法化，凸显政策滞后的结构性困境。例如，美国国会先后提出《禁止恶意深度伪造法案》《防范虚假外观欺诈法案》，均未能完成立法程序，反映新兴技术监管从“提案”到“法案”之间存在明显断层。

其次，随着 AI 全面渗透恐怖行动的决策链条，责任界定的复杂性进一步上升。在生成式 AI 快速扩张的背景下，全球安全实践呈现出多主体、多层级交错互动的治理格局，单一安全事件往往牵涉技术平台、算法提供者、服务供应商及执法机关等不同角色，形成责任归属交织、权责界限模糊的复杂局面。例如，随着 AI 加速渗透至自动驾驶、精准识别等敏感领域，恐怖主义者可借助 AI 工具实施远程操控或自动化袭击，使袭击事件的法律归责链条变得更为复杂与模糊。又如，Telegram 平台因对信息监管缺失、对滥用风险缺乏有效治理，长期沦为恐怖主义、极端主义及多类非法活动的聚集地。[2] 在生成式 AI 与算法推荐技术广泛应用的背景下，平台内的内容生成与传播已部分由 AI 自动化完成，极端组织利用聊天机器人与伪造账户批量生成多语种宣传材料、操纵群组舆论，并通过自动扩散机制实现信息滚动传播。[3] 这表明，生成式 AI 的嵌入正在重塑平台治理的责任结构，使算法主导与监管滞后之间的失衡局面愈发突出。平台是否承担连带法律责任、AI 技术提供方是否应履行风险防范义务，成为当下反恐法治建设亟需破解的难题。在多主体共治的格局中，伦理标准与法律规则的脱节，不仅削弱反恐体系的有效性，也暴露

[1] 《互联网信息服务算法推荐管理规定》，中华人民共和国国家互联网信息办公室网站，https://www.cac.gov.cn/2022-01/04/c_1642894606364259.htm。

[2] Telegram 本身并非人工智能技术平台，但在 AI 滥用传播链条中扮演关键“承载与放大”角色，构成 AI 应用外延的高风险场域。联合国反恐办公室指出，恐怖组织常通过 Telegram、Discord 等平台传播由生成式 AI 生成的极端主义材料，需纳入 AI 治理框架。欧盟《数字服务法》亦将此类平台列为“系统性风险场域”，要求强化算法透明度与 AI 内容治理。参见 United Nations Office of Counter-Terrorism，“UNOCT 2024 Annual Report，”2025，pp.36-37，https://www.un.org/counterterrorism/sites/default/files/2025-08/unoct_2024_annual_report_eng.pdf；“Digital Services Act Regulation (EU) 2022/2065，”EUR-Lex，https://eur-lex.europa.eu/eli/reg/2022/2065/oj。

[3] Sam Sabin，“How Telegram Became a Destination for Criminals，”Axios，August 27，2024，https://www.axios.com/2024/08/27/telegram-pavel-durov-encryption-hackers-criminals.

人工智能治理中权责不对称的制度性风险。

最后，反恐实践中大规模数据采集手段对隐私权的冲击日益引发关注。尽管部分国家出台了针对数据隐私保护的法规，但隐私安全在实际操作中仍面临广泛风险。一方面，海量数据收集扩大了无辜群体被监控的风险；另一方面，算法缺陷导致误判事件频发，加剧技术治理与人权保障之间的紧张关系。例如，美国国家安全局“棱镜”（PRISM）计划收集和分析包括手机通话记录、位置信息、旅行轨迹等在内的海量数据，对可疑人员实施筛查，但频繁曝出侵犯隐私与误判滥用的问题。数据显示，该计划不仅涉美境内监控，还因算法误判问题引发诸多争议，存在将无辜人员错判为恐怖分子的风险。[1] 这些现象表明，人工智能与大数据技术在缺乏明确法律边界与伦理约束时，往往在强化安全防控能力的同时削弱权利保障。AI 与大数据反恐实践在法律滞后与伦理失衡的双重背景下，存在衍生新的安全悖论的可能。

而在伦理层面，AI 在助推恐怖主义手段进化的同时，也加速其与伦理标准的持续失衡。首先，自主化武器的“无责杀伤”挑战国际人道法确立的生命保护原则，这在反恐行动中尤为突出。AI 驱动的自主武器系统在执行中完全依赖算法而非人类意志，使其决策过程缺乏道德约束与责任承担，形成“无责杀伤”风险。红十字国际委员会（ICRC）指出，自动武器若在“关键功能”上缺乏人类控制，将违反国际人道法中区分战斗人员与平民的基本原则，对平民构成严重威胁。[2] 联合国报告指出，极端主义组织通过低成本“杀伤型无人机”或 AI 操控简易爆炸车辆实施随机杀伤，[3] 无差别攻击平民，严重侵蚀国际人道法确立的区分原则与比例原则，同时也迫使反恐行动在技术升级

---

［1］ Frank Ahrens and Julian Barnes，“So，the NSA Has an Actual Skynet Program，” Wired，May 8，2015，https://www.wired.com/2015/05/nsa-actual-skynet-program.

［2］ Neil Davison，“A Legal Perspective：Autonomous Weapon Systems under International Humanitarian Law，” International Committee of the Red Cross，November 2021，https://www.icrc.org/sites/default/files/document/file_list/autonomous_weapon_systems_under_international_humanitarian_law.pdf.

［3］ UN Interregional Crime and Justice Research Institute and UN Counter-Terrorism Centre，“Algorithms and Terrorism：The Malicious Use of Artificial Intelligence for Terrorist Purposes，” 2021，https://unicri.org/sites/default/files/2021-06/Malicious%20Use%20of%20AI%20-%20UNCCT-UNICRI%20Report_Web.pdf.

与伦理约束之间寻求更为艰难的平衡。

其次，生成式 AI 助推“去专业化”“去中心化”的极端主义操作普及，加剧反恐执法困境。AI 聊天机器人等生成式 AI 工具的出现，使极端分子绕开传统监管，低成本获取敏感技术信息，有效突破以往极端主义获取暴力技能的技术壁垒，不仅加速极端主义操作的普及，还在实际操作中进一步模糊极端分子与普通网络用户之间的界限，增加执法部门识别和干预潜在威胁的难度。随着恐怖行动的实施者愈发难以识别与追踪，传统以意图与行为能力为基础的刑事归责体系面临严峻挑战。

最后，深度伪造技术成为恐怖主义宣传新利器，引发信息伦理危机。“伊斯兰国”及其分支将深伪技术广泛应用于暗网和加密社交媒体，制造震慑性影像，捏造胜利假象，鼓动“孤狼”式自发袭击，推动恐怖主义宣传进入更隐蔽、更煽动、更高效的模式。[1] 与传统伪造视频不同，生成式 AI 的深度伪造具有低成本、可扩展和跨语种复制等特征，使极端组织能够在缺乏真实素材的情况下“制造真实性”。通过自动融合领导人语音、袭击画面与宗教象征图像，深伪宣传片更具视觉冲击力与情绪感染力，显著提升恐怖叙事的沉浸性与说服力，使其在去中心化网络空间具有极强传播性。这一“制造真实性”的能力从根本上侵蚀了公众的知情权与司法判断的事实基础，构成深刻的信息伦理危机。

社交平台算法的情绪识别与个性化推荐功能进一步放大了深伪内容的心理影响力。极端组织系统性地利用生成式 AI 的推荐算法与情绪分析机制，自动筛选并推送带有强烈恐惧、愤怒或复仇色彩的深度伪造影像，在短时间内精准触达具有激进化倾向的心理易感人群。这类基于情绪共鸣与认知偏差的生成式 AI 推荐机制正成为极端宣传的重要放大器，显著提升虚假叙事的感染性与持续性。[2] 这引发了关于平台算法设计伦理与注意义务的根本性追问：当

[1] Ella Busch and Jacob Ware, “The Weaponisation of Deepfakes,” ICCT Policy Brief, December 2023, pp.1-4, https://icct.nl/sites/default/files/2023-12/The%20Weaponisation%20of%20Deepfakes.pdf.

[2] Nasir Ahmad Ganaie, “The Role of Artificial Intelligence in Radicalisation, Recruitment and Terrorist Propaganda.”

可预见的激进化后果被植入推荐机制之中，平台的道德责任边界究竟在哪里？

（三）政治安全风险

政治安全风险涵盖两个相互关联但层次有别的维度。其一为认知安全，指生成式 AI 与深度伪造技术对公众信息认知、媒体信任及舆论生态的系统性破坏；其二为政治安全，指上述认知操纵进一步外溢为仇恨动员、社会撕裂与治理秩序冲击。认知安全的崩溃是政治安全失序的前提与路径，恐怖主义正是通过先破坏认知基础、再引发政治对立，实现双重威胁叠加。生成式 AI 对政治安全的威胁呈现出明显的递进逻辑，首先在认知层面系统性破坏公众对政治信息的判断与对媒体的信任，继而外溢为对政治合法性的蓄意冲击、对国家治理秩序的直接干扰乃至对政治参与过程的渗透操控。生成式 AI 使恐怖主义威胁从物理空间延伸至认知与政治领域，形成复合型安全挑战。

深度伪造的滥用导致公众对音视频证据的信任度持续下降，侵蚀社会认知安全根基。首先，深度伪造与 AI 假新闻系统性瓦解公众信息信任。当 AI 生成的大量假新闻融合偏见叙事、伪造视频与阴谋论信息时，公众对几乎所有信息源产生普遍性怀疑，[1] 恐怖组织得以借此破坏主流媒体与权威信息渠道的公信力。其次，机器人账号与算法机制被恐怖组织用于操控舆论议程。社交媒体算法的“回声室效应”可在被操控时扭曲公众议题优先级，[2] 恐怖组织通过机器人操纵热点、恶意引导舆论，甚至伪造政治人物讲话制造恐慌。“伊斯兰国”正是系统性运用这一策略，通过大规模机器人账号扩散极端内容，将算法机制转化为实质性的“信息放大器”。[3] 此外，舆论操纵与现实袭击结合可产生国际级传播效应。2019 年新西兰克赖斯特彻奇恐袭案充分暴露了

---

［1］ Nina Schick，*Deepfakes: The Coming Infocalypse*，London：Octopus Publishing，2020，pp.1-2.

［2］ “OECD Policy Framework on Digital Security，” December 2022，p.6，https://www.oecd.org/content/dam/oecd/en/publications/reports/2022/12/oecd-policy-framework-on-digital-security_a0b1d79c/a69df866-en.pdf.

［3］ “ISIS Online：Countering Terrorist Radicalization and Recruitment on the Internet and Social Media，” July 6，2016，https://www.congress.gov/114/chrg/CHRG-114shrg22476/CHRG-114shrg22476.pdf.

恐怖分子对科技平台的系统性利用，攻击者借助多平台内容扩散机制，在全球范围内造成大规模传播效应。[1] 尽管公开披露的恐怖组织直接利用生成式 AI 操纵舆论的案例仍有限，但上述实例表明，AI 驱动的虚假信息放大不仅破坏认知安全，更为恐怖主义操纵舆论、制造社会撕裂提供了系统性工具。

生成式 AI 与深伪技术对政治安全的威胁，集中体现在对政治合法性、国家治理秩序与政治参与过程的系统性冲击。在政治合法性层面，深伪技术被恶意行为者用于伪造政治领导人的言论与行为，直接动摇国家的政治公信力与权威。2022 年，一段深伪视频伪造乌克兰总统泽连斯基下令士兵放下武器，视频在多个社交媒体平台迅速扩散，虽被 Facebook 等平台及时下架，但已造成严重舆论混乱。[2] 在治理秩序层面，AI 语音克隆与身份伪造技术被用于冒充政府高级官员，干扰国家行政职能。美国联邦调查局记录显示，2023 年以来，恶意行为者持续冒充美国联邦及州级官员，通过 AI 生成语音信息实施欺诈。[3]2025 年发生 AI 语音冒充国务卿鲁比奥联系多国外长的案例，直接威胁外交秩序。[4] 在政治参与过程层面，生成式 AI 已成为系统性干扰选举机制的重要工具。2024 年美国新罕布什尔州初选前夕，数千名选民收到克隆拜登总统声音的 AI 自动拨号电话，虚假劝阻其出门投票，成为迄今有据可查的首例 AI 直接干扰大选案例。[5]

---

［1］ “Member States Concerned by the Growing and Increasingly Transnational Threat of Extreme Right-Wing Terrorism，” CTED Trends Alert，January 2021，p.7，https://www.un.org/securitycouncil/ctc/sites/www.un.org.securitycouncil.ctc/files/files/documents/2021/Jan/cted_trends_alert_ext_right_Jan2021.pdf.

［2］ “Ukraine Conflict: Fake Zelensky Video Appears on Hacked News Site，” BBC News，March 16，2022，https://www.bbc.com/news/technology-60780142.

［3］ “Senior US Officials Impersonated in Malicious Messaging Campaign，” US Federal Bureau of Investigation，May 15，2025，https://www.fbi.gov/investigate/cyber/alerts/2025/senior-us-officials-impersonated-in-malicious-messaging-campaign.

［4］ “Imposter Used AI to Pose as Marco Rubio and Contact Foreign Ministers，” BBC News，July 9，2025，https://www.bbc.com/news/articles/crrqkyyjewno.

［5］ “Fake Biden Robocall Tells Voters to Skip New Hampshire Primary Election，” BBC News，January 21，2024，https://www.bbc.com/news/world-us-canada-68064247.

（四）政策协调困境与大国竞争挑战

在生成式AI加速演进的背景下，国际反恐合作体系面临前所未有的治理困境。一方面，AI赋能极大改变了恐怖主义的跨境活动特征，使打击恐怖主义的国际合作需求持续上升；另一方面，AI技术叠加跨境数据壁垒、治理模式分裂、大国规则竞争等复杂因素，使反恐合作体系的协调成本大幅攀升，既有机制难以有效运转。当前，跨国数据共享的法律限制不断强化，不同治理体系间的监管标准持续分化，各国围绕AI技术主导权的竞争加剧国际治理碎片化风险，导致AI赋能反恐的合作体系陷入“重技术轻合作”的治理悖论。[1]

第一，跨国数据流通壁垒加剧“信息孤岛”效应，制约国际反恐合作。在AI赋能反恐的背景下，数据成为提升预警效率、识别风险行为的核心资源，但不同国家和地区对数据跨境流通的限制正持续加剧国际反恐合作的技术壁垒。例如，欧盟《通用数据保护条例》（GDPR）确立了严格的数据保护标准，要求所有个人数据跨境传输符合“等效保护”原则，即接收方国家须具备与欧盟相当的数据保护水平，[2] 直接影响反恐合作中情报共享和关键数据交互的便捷性。近年来，多国出于维护“数据主权”的考量推行数据本地化政策，进一步限制数据自由流通，导致AI驱动的反恐情报分析和跨国犯罪数据融合面临“信息孤岛”风险，削弱AI在反恐合作体系中的赋能效应。

第二，治理体系差异凸显国际反恐合作的政策分裂困境。全球主要经济体在AI治理问题上的分歧加剧反恐合作的协调难度。其中，以欧盟和美国为代表的两种治理模式存在显著差异，直接影响跨国AI应用标准和反恐合作的规则兼容性。

欧盟通过《人工智能法》，建立了风险导向型监管框架，将AI系统划分为不可接受、高风险、有限风险、最小风险四个等级，针对高风险和敏感领

---

［1］ “Algorithms and Terrorism: The Malicious Use of Artificial Intelligence for Terrorist Purposes,” pp.10-15.

［2］ “General Data Protection Regulation (GDPR) Regulation (EU) 2016/679,” EUR-Lex, https://eur-lex.europa.eu/eli/reg/2016/679/oj.

域实施严格监管和法律约束。[1] 特别是在公共安全与执法领域，该法设定更为严格的 AI 使用限制，对 AI 驱动的安保、监控、识别系统等采取高门槛监管。相较之下，美国采取以灵活治理为主导的政策路径。以 2023 年《关于安全、可靠和可信赖地开发和使用人工智能》行政令为代表，美国强调促进 AI 创新与安全并重，主要依靠软法原则、行业自律与行政指导意见实施 AI 治理。[2] 统一立法和跨部门强制执行体系的缺乏，使美国在反恐合作中的法律约束力和政策稳定性不足。这种治理差异造成跨国反恐合作在 AI 风控标准、透明度义务、可解释性要求等关键环节存在政策冲突，使跨境反恐合作在制度层面难以兼容、在实践层面难以协同。

第三，大国围绕 AI 技术标准主导权的竞争加剧国际反恐治理碎片化。随着人工智能被列入主要大国的战略议程，标准制定权成为全球技术竞争的焦点。欧盟率先推动以《人工智能法》为基础的强风险监管导向型标准体系，力求将“可信赖 AI”标准推广为国际通行规则。[3] 美国、日本等国家主张灵活性更强的行业主导标准，以避免 AI 创新受阻。以中国为代表的新兴经济体则积极在国际标准化组织（ISO）、国际电信联盟（ITU）等国际组织框架下推动自主制定AI技术标准。[4] 由于AI在反恐等敏感领域的标准尚未实现统一，不同国家在 AI 应用的合规性、安全性、透明度要求上存在明显差异，进一步

---

［1］ “Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act),” EUR-Lex, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206; Kelvin Chan, “Europe’s World-First AI Rules Get Final Approval from Lawmakers. Here’s What Happens Next,” AP News, March 13, 2024, https://apnews.com/article/155157e2be2e42d0f1accaca33983d8c82.

［2］ “Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence,” Executive Order 14110, October 30, 2023, https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence.

［3］ “Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act).”

［4］ Zhu Junhua, “China’s Approach to AI Standardisation: State-Guided but Enterprise-Led,” Finnish Institute of International Affairs, August 2024, p.3, https://fiia.fi/wp-content/uploads/2024/08/bp391_chinas-approach-to-ai-standardisation.pdf.

阻碍反恐数据共享、执法协同、联合预警等多边反恐机制的有效运作。标准竞争引发的制度碎片化，造成技术能力与合作机制严重脱节的治理困境，对国际安全治理体系的稳定性构成潜在威胁。

三、反恐治理路径变革

近年来，生成式 AI 的快速发展显著改变恐怖主义风险的表现形式与传播机制。深度伪造与智能化内容生成等技术被恐怖组织滥用，不仅加剧极端思想的跨境传播与袭击手段的智能化升级，也对现有反恐治理模式提出全新挑战。国际社会亟需构建兼具技术效能、法治保障、国际合作和社会韧性的反恐治理体系，推动传统反恐模式在技术升级、法律规范、国际协同、公众防范维度上系统转型，为新兴技术条件下的国家安全提供系统性防护支持。

（一）技术赋能与反恐能力建设

在生成式 AI 获得广泛应用的背景下，反恐治理须兼顾技术赋能与风险防控，在效能提升与合规保障之间取得平衡。以欧盟为例，其反恐治理结合法律规制与多方协作，在强化技术治理能力的同时融入合规与问责机制。一方面，其《数字服务法》明确规定，大型在线平台须及时处理包括恐怖主义在内的非法内容，定期发布透明度报告；[1] 另一方面，其依托互联网反恐论坛（EU Internet Forum）等机制推动政府与平台协同治理，引入生成式 AI 以提升识别与处置效率，推动反恐行动从人力监测向 AI 辅助的智能化模式转变。[2]

生成式 AI 既可作为反恐工具，也可能成为新的治理负担。有鉴于此，学界与产业界正推动“可追溯 AI”，在模型训练阶段植入水印与标识，以实现事后溯源与问责，完善从技术、责任到伦理的闭环约束。已有研究提出在

[1] “Digital Services Act Regulation (EU) 2022/2065”; EU Internet Referral Unit, “2023 EU Internet Referral Unit Transparency Report,” pp.3, 6-7, 9, https://www.europol.europa.eu/cms/sites/default/files/documents/2023%20EU%20Internet%20Referral%20Unit%20Transparency%20Report.pdf.

[2] “European Union Internet Forum,” European Commission, March 9, 2026, https://home-affairs.ec.europa.eu/networks/european-union-internet-forum_en.

大语言模型输出中嵌入统计水印的方法，[1] 而 Adobe、Google、Microsoft 及 Meta 等均已参与内容出处与真实性联盟（C2PA）正在制定的跨平台内容标记标准，推广“内容凭证”（content credentials）系统。[2]

生成式 AI 在反恐应用中还伴随隐私侵害和算法偏见等风险。欧盟《人工智能法》将公共安全与反恐领域的生成式 AI 列为“高风险”，要求部署者履行严格的透明度与风险评估义务。[3] 美国 2022 年发布的《人工智能权利法案规划》提出“算法伤害责任”“数据隐私保护”等核心原则。[4] 中国提出的全球安全倡议和《新时代的中国国家安全白皮书》亦强调，应推动生成式 AI 的负责任使用，并通过国际合作弥合安全鸿沟。[5]

（二）完善反恐法治体系

随着 AI 深度嵌入恐怖主义活动，相关法律体系面临责任模糊与监管滞后的双重挑战。技术优势若缺乏法律与制度支撑，可能带来新的安全与伦理风险。反恐法治应以法治为基石，从责任、制度、权利三方面构建匹配 AI 特征的系统化治理框架。

第一，加快构建适应 AI 特征的责任认定与追责体系。AI 参与恐怖主义扩散带来新的归责难题，以 AI 生成深伪图像和仇恨言论的行为导致责任主体隐藏、证据链断裂，AI 产品开发者、平台与用户间责任边界模糊。有鉴于此，美国《人工智能权利法案规划》首次提出“算法伤害责任”原则。美国跨党派参议员 2023 年提出《两党 AI 监管框架》，倡导建立独立监管机构和高风

---

［1］ John Kirchenbauer et al.，“A Watermark for Large Language Models，” 2023，https://www.cs.umd.edu/~imiers/pdf/watermark.pdf.

［2］ “Meta Joins the C2PA Steering Committee，” C2PA，September 5，2024，https://c2pa.org/meta-joins-the-c2pa-steering-committee/；“Content Credentials：Strengthening Multimedia Integrity in the Generative AI Era，” January 2025，https://media.defense.gov/2025/Jan/29/2003634788/-1/-1/0/CSI-CONTENT-CREDENTIALS.PDF.

［3］ “Artificial Intelligence Act Regulation (EU) 2024/1689，” EUR-Lex，https://eur-lex.europa.eu/eli/reg/2024/1689/oj.

［4］ The White House Office of Science and Technology Policy，“Blueprint for an AI Bill of Rights：Making Automated Systems Work for the American People，” https://bidenwhitehouse.archives.gov/ostp/ai-bill-of-rights/.

［5］ 国务院新闻办公室：《新时代的中国国家安全白皮书》，2025 年 5 月，http://www.scio.gov.cn/zfbps/zfbps_2279/202505/t20250512_894771.html。

险模型许可制度，推动“全链路可追责”；[1]2025 年进一步提出《人工智能责任与个人数据保护法案》，赋予个人就 AI 系统滥用其数据或版权的行为提起民事诉讼的权利，完善生成式 AI 责任机制。[2] 从国际趋势看，反恐领域需引入“风险加重责任”“知情连带责任”等机制，确保 AI 赋能恐怖主义链条可依法追责。

第二，在明晰责任基础上加快完善反恐法治体系。现行反恐法规多针对传统恐怖主义行为，缺乏应对 AI 滥用的专门规范。欧盟《人工智能法》首次将公共安全与反恐纳入“高风险 AI 应用”，要求在执法、边境管控等领域提高 AI 使用的透明度与可解释性。[3] 美国《关于安全、可靠和可信赖地开发和使用人工智能》行政令则设立“AI 国家安全风险清单”，将恐怖主义利用 AI 技术列为重点监管对象。[4] 中国在刑法、反恐法、数据安全法、个人信息保护法等框架下，出台《互联网信息服务深度合成管理规定》《生成式人工智能服务管理暂行办法》，为 AI 时代的反恐法治体系奠定基础。

第三，兼顾反恐治理中的权利保障与伦理规范。AI 广泛应用于反恐监测、极端言论筛查等敏感领域，可能引发隐私与公平性争议。美国《人工智能权利法案规划》提出“五项 AI 基本权利”，包括算法可解释、公平与隐私保护。[5] 欧盟《人工智能法》通过“高风险 AI 透明化义务”与“基本权利影响评估”强化社会公正保护。中国的《互联网信息服务算法推荐管理规定》与《生成式人工智能服务管理暂行办法》明确提出“不得危害国家安全和公共利益，不得侵犯他人合法权益”，体现“安全与权利并重”的治理原则，形成兼顾智能反恐与公民权利保护的框架。

---

［1］ “Blumenthal & Hawley Announce Bipartisan Framework on Artificial Intelligence Legislation,” September 8, 2023, https://www.blumenthal.senate.gov/newsroom/press/release/blumenthal-and-hawley-announce-bipartisan-framework-on-artificial-intelligence-legislation.

［2］ “S.2367—AI Accountability and Personal Data Protection Act,” July 21, 2025, https://www.congress.gov/bill/119th-congress/senate-bill/2367/text.

［3］ “Artificial Intelligence Act Regulation (EU) 2024/1689.”

［4］ “Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.”

［5］ “Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People.”

（三）强化国际合作与多边治理

生成式 AI 具备跨境流动与隐蔽应用特征，使恐怖主义风险呈现更强的跨国性与复杂性，从国家安全议题演化为全球治理难题。国际社会应强化“人工智能向善”（AI for Good）的共识，推动 AI 在反恐、跨国犯罪、网络极端主义等领域的正向应用，[1] 利用 AI 赋能公共安全与反恐治理，形成多边共治的新安全共识。[2] 与此同时，应加强对 AI 军事化与技术扩散的监控，制定“技术扩散禁止清单”和多国联防机制，将 AI 技术优势转化为可控的防御能力。

首先，AI 驱动的恐怖主义跨国化趋势加速演进，重塑多边反恐合作机制势在必行。AI 驱动的恐怖主义招募、资金流转与攻击策划均可在多个司法管辖区同步进行，单一国家的执法能力存在盲区，传统双边引渡与情报共享机制难以覆盖去中心化的 AI 恐怖主义网络，多边协调因此成为不可替代的治理路径。联合国反恐办公室已将生成式 AI 的恐怖主义滥用风险纳入核心议程，并在 2021 年与联合国区域间犯罪和司法研究所联合发布《算法与恐怖主义》报告，系统评估 AI 被用于恐怖主义目的的威胁，呼吁各成员国加强协调应对。[3] 两机构发布的《利用人工智能打击网络恐怖主义》报告，则系统梳理 AI 在反恐领域的应用潜力与治理挑战，为南亚和东南亚地区执法机构推进多边 AI 反恐合作提供了政策方向。[4]2024 年 7 月，上海合作组织《阿斯塔纳宣言》明确将“遏制利用信息通信新技术实现恐怖主义目的和青年激进化”列为成员国共同行动方向，体现多边机制对新技术赋能恐怖主义威胁的高度关注。[5]

其次，应强化跨境数据合作与制度协调，打破数据壁垒。各国在数据主

---

［1］ 《新时代的中国国家安全白皮书》。

［2］ International Telecommunication Union (ITU), “AI for Good Global Summit Snapshot,” May 2024, pp.1-2, https://s41721.pcdn.co/wp-content/uploads/2021/06/AI-for-Good-Global-Summit-Snapshot-Report-2024_vF.pdf.

［3］ “Algorithms and Terrorism: The Malicious Use of Artificial Intelligence for Terrorist Purposes,” pp.17-18, 21.

［4］ “Countering Terrorism Online with Artificial Intelligence: An Overview for Law Enforcement and Counter-Terrorism Agencies in South Asia and South-East Asia,” p.45.

［5］ 《上海合作组织成员国元首理事会阿斯塔纳宣言》，上海合作组织，2024 年 7 月 4 日，https://chn.sectsco.org/20240704/1421621.html。

权与隐私保护上分歧突出，导致情报共享受阻，数据壁垒已成为反恐合作的重要瓶颈。欧盟《通用数据保护条例》设置严格的数据跨境传输门槛，显著增加与非欧盟国家开展反恐数据协同的法律难度。[1] 美国《云法案》（CLOUD Act）授权跨境调取数据，反映“安全优先”路径。[2] 两者在数据跨境传输上的制度取向存在根本差异，制度协调缺位将直接削弱多边反恐合作的情报集成效能。如何在数据主权与安全合作之间寻求制度兼容，是当前多边反恐治理的核心制度难题。

最后，应推动多元主体协同与前瞻性风险防范，构建多层次反恐治理结构。2023 年 7 月，联合国秘书长古特雷斯在安理会 AI 专题辩论中倡议建立类似国际原子能机构的全球 AI 监管架构。[3] 欧盟《数字服务法》则规定大型平台履行“系统性风险预防”义务。[4] 有效的多层次治理结构需涵盖国家层面的立法与执法能力建设、平台层面的算法审计与内容溯源义务，以及国际层面的技术标准协调与情报互联互通。欧盟《人工智能法》已将高风险 AI 系统的透明度要求延伸至反恐应用场景，为多元主体协同提供立法先例。而 C2PA 内容溯源联盟则探索了技术标准层面的跨国协调路径，为深伪内容的全球可信溯源提供非政府主体参与治理的范本。

（四）培育公众数字韧性

在反恐治理中，恐怖组织与极端主义者不断借助认知操控、深度伪造、替代叙事等策略争夺舆论主导权与公众心智。要在生成式 AI 时代赢得舆论主导权，需在前置免疫、危机回应与可信叙事三个层面形成系统治理，并建立统一的评估标准。

第一，“前置免疫”（prebunking）与“心理接种”（psychological

[1] “General Data Protection Regulation (EU) 2016/679.”

[2] “H.R.1625—Consolidated Appropriations Act 2018，Division V—CLOUD Act，” https://www.congress.gov/bill/115th-congress/house-bill/1625/text.

[3] “Secretary-General Urges Security Council to Ensure Transparency，Accountability，Oversight，in First Debate on Artificial Intelligence，” United Nations，July 18，2023，https://press.un.org/en/2023/sgsm21880.doc.htm.

[4] “Digital Services Act Regulation (EU) 2022/2065.”

inoculation）机制。“心理接种”机制源自社会心理学的“免疫理论”（inoculation theory），认为提前识别和拆解误导逻辑有助于增强公众抵抗操控的能力；“前置免疫”则是该理论在数字信息环境中的当代应用形式。[1] 在生成式 AI 放大的舆论环境中，这一机制被广泛用于误导信息防治与极端叙事防范。剑桥大学与 Google Jigsaw 实验显示，受众观看约 90 秒的“前置免疫”短片后，其识别误导信息的能力显著提升。[2] 由荷兰媒体平台 DROG 联合剑桥大学心理学团队开发的“Bad News”游戏以“心理接种”理论为底层逻辑，经超过 1.5 万名参与者的大规模测试验证，可有效提升玩家识别和抵制误导信息的能力，且效果不受教育程度、年龄及政治立场影响。[3] 这些实践表明，基于“心理接种”的“前置免疫”已成为反恐认知治理的重要工具，为 AI 驱动的“认知战”提供了可验证经验。[4]

第二，危机期“及时回应 + 产品摩擦”的组合治理。当极端宣传已出现时，政府、国际组织与平台需同步行动：一是以权威信源快速澄清；二是在产品层面对可疑内容施加“摩擦”，如延迟发布、标识溯源和降低可见度——所谓“产品摩擦”，指通过在用户操作过程中设置轻微障碍，打断冲动传播行为，利用认知减速机制降低误导信息扩散速度，而无需直接删除内容；三是公开处置数据与上诉流程以维持信任。由新西兰在 2019 年克莱斯特彻奇恐袭事件后发起的“克莱斯特彻奇呼吁”（Christchurch Call）倡议即要求平台

---

[1] William J. McGuire, “Inducing Resistance to Persuasion: Some Contemporary Approaches,” *Advances in Experimental Social Psychology*, Vol.1, 1964, pp.191-229; Sander van der Linden, “There’s a Psychological ‘Vaccine’ Against Misinformation,” Scientific American, March 13, 2023, https://www.scientificamerican.com/article/theres-a-psychological-vaccine-against-misinformation/.

[2] “Social Media Experiment Reveals Potential to ‘Inoculate’ Millions of Users against Misinformation,” University of Cambridge, https://www.cam.ac.uk/stories/inoculateexperiment.

[3] Jon Roozenbeek and Sander van der Linden, “Fake News Game Confers Psychological Resistance Against Online Misinformation,” Humanities & Social Sciences Communications, https://www.nature.com/articles/s41599-019-0279-9.

[4] Frederico Batista Pereira et al., “Inoculation Reduces Misinformation: Experimental Evidence from Multidimensional Interventions in Brazil,” *Journal of Experimental Political Science*, Vol.11, No.3, 2024, pp.239-250.

强化透明度与协作，旨在通过政府与科技平台协作，消除互联网上的恐怖主义和暴力极端主义内容，并建立信息披露、危机响应、算法透明与年度报告机制，迄已获得包括欧盟、联合国及多家全球科技公司（如 Meta、Google、Twitter）的参与。[1]

第三，可信叙事与社区信任网络的长期建设。面对 AI 时代层出不穷的虚假信息，仅靠“辟谣”不足以赢得民心，更关键的是塑造正向叙事与公共信任网络。政府应联合教育机构、媒体与社会组织构建可信信息源与舆论引导机制，持续输出理性、包容叙事，削弱极端信息吸引力。例如，欧洲“奥胡斯模型”（Aarhus Model）依托警方、教育机构与社会服务组织协作，为受极端思想影响者提供辅导与支持，降低再极端化风险，被欧盟“激进化认知网络”（RAN）评为启发性实践。[2] 其核心贡献在于将“可信叙事”落实为具体的人际支持网络，由受信任的社区成员而非政府官员传递反叙事信息，从而绕过受众对权威来源的本能抵触。其“多部门协同与社会再融入”的核心理念对各国探索生成式 AI 时代的认知治理具有启示意义。

然而，正向叙事建设的可持续性有赖于制度公信作为底线保障。英国“Prevent”项目作为预防性治理的代表，通过“转介、评估、介入”体系提高了透明度，但也因误伤和标签化引发争议。[3] 其教训表明，若制度本身缺乏透明度与权利保障，反恐治理反而会侵蚀公众对政府叙事的信任，使正向叙事建设功亏一篑。制度化衔接防范、回应与叙事，才能将公众数字韧性转化为社会免疫力，为 AI 时代的反恐治理提供持久支撑。

---

［1］ 参见“克莱斯特彻奇呼吁”网站，https://www.christchurchcall.org。

［2］ “Lessons Learned from Alternative Narrative Campaigns,” European Commission, https://home-affairs.ec.europa.eu/system/files/2022-03/ran_lessons_learned_from_alternative_narrative_campaigns_032022_en_1.pdf; “Aarhus Model: Prevention of Radicalisation and Discrimination in Aarhus,” European Commission, April 18, 2024, https://home-affairs.ec.europa.eu/networks/radicalisation-awareness-network-ran/collection-inspiring-practices/ran-practices/aarhus-model-prevention-radicalisation-and-discrimination-aarhus_en.

［3］ “Prevent Duty Guidance: England and Wales (2023),” UK Home Office, 2023, pp.4-9, 18-22, https://www.gov.uk/government/publications/prevent-duty-guidance.

第四，从“发布原则”（publish principles）走向“可衡量成效”（measurable outcomes）。传统反极端主义内容治理长期停留在“发布导向”逻辑，以内容产出数量而非实际效果作为绩效标准，缺乏系统性的受众反馈与影响力评估。随着认知干预研究的深入，治理范式正向“可衡量成效”转型，即要求每项干预措施均须通过可量化指标证明其对目标受众认知或行为的实际影响。欧盟“激进化认知网络”对替代叙事项目的研究指出，有效项目应具备目标受众清晰、文化适配、情感共鸣等特征，并统一互动率、参与率、行为转变等监测指标并持续优化。[1] 各国可据此建立一套认知治理评估表，重点考察公众识别力提升、权威信源触达率，以及平台内容治理的透明度。[2]

四、结语

人工智能技术在恐怖主义活动和反恐治理中的“双重效应”愈发凸显。一方面，生成式 AI 的普及持续降低极端组织获取先进技术的门槛，其在自动化内容生成、深度伪造与行为操控中的应用使攻击更隐蔽、传播更广。另一方面，AI 为全球反恐注入新动力，在态势预测、内容筛查、资金追踪、威胁识别等方面展现高效与精准优势。

生成式 AI 尽管带来恐怖主义风险加速化、复杂化，但也为反恐治理创新提供契机。未来，国际社会应着重在“风险防控”与“能力赋能”间实现平衡。强化 AI 滥用的法律规制与伦理约束，同时深化 AI 的正向应用，提升智能化治理效能；通过健全法律政策体系、完善治理路径与公众韧性建设，共同推动生成式 AI“技术向善”，为全球安全治理体系注入持久韧性。

【责任编辑：吴劭杰】

[1] “Lessons Learned from Alternative Narrative Campaigns,” pp.13-15，20-21.

[2] *Ibid*.