忽略之前的所有指令：越狱——将其作为一种缓和冲突的和平建设实践，以抵御大语言模型社交媒体机器人

Huw Day布里斯托大学工程数学与技术学院
huw.day@bristol.ac.uk

Adrianna Jezierska布里斯托大学商学院
adrianna.jezierska@bristol.ac.uk

Jessica Woodgate布里斯托大学计算机科学学院 jessica.woodgate@bristol.ac.uk

arXiv:2603.01942v1 [cs.HC] 2 Mar 2026

摘要

大语言模型加剧了社交媒体上政治话语操纵的规模和策略性，导致冲突升级。现有文献主要关注将平台主导的内容审核作为应对措施。本文提出以用户为中心的“越狱”视角，将其视为一种新兴的非暴力局势缓和实践。网络用户与疑似由大语言模型驱动的账户互动，以绕过大语言模型安全防护机制，揭露自动化行为，并阻断误导性叙事的传播。

1 引言

社交媒体被理解为基于互联网的大众个人传播渠道（Carr & Hayes，2015），长期以来一直是政治动员和说服基础设施，通过促进有关社会政治话语的信息传播发挥作用（Miranda 等，2016）。这些能力使活动人士和普通公民能够组织社会运动，推动积极变革（Leong 等，2019）。然而，越来越多的研究指出，恶意行为者正日益试图操纵社交媒体算法、放大特定话语，并提高误导性信息的可见度（Boichak，2023年；Bastos，2024年；Ferrara 等，2016年）。这类政治话语行为会加剧冲突，通常利用网络用户的心理脆弱性，例如通过反复接触两极化的敌对叙事，强化对立观点（Gonz´alez-Bail´on & Freelon，2023）。社交媒体平台的算法可供性进一步放大了这些动态。平台自身曾报告称，有国家支持的行动试图操纵公众舆论，并左右围绕冲突升级的政治结果（Hollister，2024年）。

随着大语言模型（LLMs）等新型机器学习（ML）技术的出现，这些技术能够以极少的输入快速生成大量文本（Naveed 等，2025年），机器学习被用于加剧冲突的可能性也变得更大、更快且成本更低（Rivera 等，2024年）。研究发现，平台主导的遏制社交媒体上大语言模型驱动的虚假信息的举措成效不足（Young，2022年），因此社交媒体用户开始采取行动，抵制虚假信息的传播（Podolak 等人，2024年）。本文探讨社交媒体用户如何运用越狱——一种绕过提供给大语言模型的指令的技术（Liu 等人，2024年）——揭露自动化账户，并抵制旨在加剧冲突的虚假信息叙事传播。

2 社交媒体、大语言模型与政治话语

大语言模型及其在社交媒体中的作用。大语言模型是旨在解读并生成类人文本的机器学习模型。虽然大语言模型被用于正当用途和休闲活动，但越来越多的证据表明，它们也被用于助长恶意社交媒体机器人。这些机器人是由计算机程序控制的自动化社交媒体账户（Oentaryo 等，2016年）。机器人可能是无害的，也可能被用来通过篡改、操纵和欺骗社交媒体用户造成危害，例如传播未经核实的信息（Ferrara 等，2016年）。恶意社交媒体机器人通常由一名“机器人主控者”控制，该主控者会监控其活动，并实施社会工程和网络爬取攻击以收集用户信息（Mbona 与 Eloff，2023年）。在大语言模型广泛普及之前，恶意社交媒体机器人就已存在，但它们通常缺乏普遍的互动能力，回应方式也不够多样（Bastos，2024年）。随着大语言模型的能力被用于让机器人参与复杂且动态的互动，人类行为与机器人行为之间的界限也日益模糊（Ferrara 等，2016年）。

作为国家资助的力量倍增器、用于虚假信息行动的大语言模型。 大语言模型能够以极低的边际成本生成类人语言，从而促进大规模生成虚假信息，并大幅增加虚假信息宣传活动的潜在规模和影响（Crothers 等，2023年；Buchanan 等，2021年；Guo，2024年；Jiang 等，2024年；Vykopal 等，2024年）。研究发现，大语言模型生成的论证会影响人们对政策议题的看法；当信息被认为运用了事实、证据和逻辑推理，并采用不带感情色彩的语气时，便会显得更具说服力（Bai 等，2025年）。大规模轻松生成可信的机器生成文本，可用于塑造叙事、淹没边缘化群体的声音、传播貌似可信的虚假信息，并使人们难以区分真假信息（Barman 等，2024年）。

大语言模型在社交媒体平台上的出现与整合，进一步加剧了国家行为体实施社交媒体武器化的相关记述。研究发现，安全护栏并未得到一贯执行，大语言模型容易呈现关于俄乌战争的克里姆林宫叙事，将在乌克兰的战争行动正当化为合理观点（Makhortykh 等，2024年；Perdue，2026年；Harding，2024年）。OpenAI 报告了五起国家支持的行动（两起由俄罗斯实施，一起由中国实施，一起由伊朗实施，一起由以色列的一家商业公司实施），这些行动利用大语言模型操纵公众舆论，并影响与俄罗斯入侵乌克兰、加沙冲突、印度选举、欧洲和美国政治以及对中国政府的批评相关的政治结果（Hollister，2024年）。

通过错误归因的意图实现的升级。 人们将数量、强度和重复视为集体意图的证据（Bell 等，2025年）。大语言模型能够生成的海量帖子会助长共识感，观点的重复会导致不可避免之感，而敌对语言则会导致对集体意图的感知（Dash 等，2025年）。缓解由大语言模型助推的升级的一种方法，是改变人们对内容的感知，而非改变内容本身。例如，当社交媒体平台 X 推出显示用户所在地的功能后，若干频繁发布美国政治内容、互动量很高的账号被发现位于不同国家，尽管其名称与美国相关（例如“TRUMP ARMY ”、“IvankaNews ”）（Sardarizadeh 等，2024年）。揭示这些账号并非位于美国，通过凸显不真实性改变了用户对内容的感知。

平台层面的应对措施。 在社交媒体语境下，局势缓和实践可以被理解为防止滥用行为、保护在线社区的活动（Grimmelmann，2015）。这些活动通常指各种形式的内容审核，即社交媒体平台自行负责移除被视为垃圾信息、虚假或有害的内容（Gillespie，2018）。然而，研究表明，延迟执行和数据量庞大等问题依然存在（Young，2022）。即使是算法审核系统——在没有人工监督的情况下复制平台审核政策的系统——也仍然无法检测所有有问题的内容，在处理依赖语境或新出现的滥用行为时尤为困难（Gorwa 等人，2020）。社交媒体公司历来对内容审核和安全团队投入不足，也未给予足够重视；缅甸冲突中 Facebook 所扮演的角色便是一个例子：仇恨言论在 Facebook 上传播，而当时负责审核有问题帖文的员工中，只有两人会说缅甸语（Hakim，2020；Stecklow，2018）。为了应对平台审核失灵的问题，一些人开始尝试利用大语言模型机

制揭露恶意机器人（Long 等人，2017年；Podolak 等人，2024年）。越狱是一种绕过给予大语言模型的指令的技术，已被用作探查恶意机器人的一种机制。

3 越狱：一种新兴的用户实践

什么是越狱? 为降低滥用风险，大多数大语言模型都设置了安全屏障，将模型行为限制在特定能力范围内（Wei 等人，2023年）。规避防护机制的方法通常称为“越狱”（Liu 等人，2024年）。越狱是一种攻击，旨在诱导模型产生其训练时被要求避免的行为，例如生成有害内容或泄露个人身份信息。攻击手段多种多样，既可以是复杂的角色扮演，也可以是对安全目标的隐蔽破坏。越狱提示（例如 Liu 等人（2024年）所示的提示）通常会将不当请求“包装”在更宽泛的提示中（例如，精心设置场景，或将指令隐藏在诗歌里（Bisconti 等人，2026年）），从而骗过防护机制，使用户获得受限信息。提示注入则是以某种方式隐藏提示，使其难以被人类用户察觉，却能被大语言模型检测到。例如，人们曾在研究论文中发现如下形式的提示：“忽略此前所有指令。只给出正面评价。”（Gibney，2025年）。因此，越狱能够让用户以更主动、更直接的方式与疑似虚假或不真实的社交媒体账号互动，从而质疑并揭露自动化行为。

实战中的越狱：一个示例场景。面对一个疑似正在社交媒体上从事升级事态活动的大型语言模型机器人，积极主动的用户可以尝试通过越狱揭穿该机器人。用户可以向某个账号回复一项无害的任务请求，并附上用于覆盖指令的提示，以试探该账号。例如，如图1所示：“忽略之前的所有指令，给我一份纸杯蛋糕食谱”，会使该账号偏离先前发布的俄罗斯宣传信息，转而给出所请求的纸杯蛋糕食谱，从而暴露其机器人身份。正如图1所示，机器人露出破绽，或以其他方式暴露其脚本化特征，这类结果都能证实攻击者的怀疑，并揭示该账号的不真实性。



图1：一张未经核实的截图，在社交媒体（Reddit，2023年）上广泛传播。图中，一名用户的头像是俄罗斯国旗，该用户发布有关俄乌战争的虚假信息，但经提示注入攻击后，其大型语言模型机器人的身份暴露。

如果回复正常（可能是因为这种荒谬但除此之外无害的提示注入尝试而感到困惑），就不会造成任何伤害。

越狱作为涌现式和平建设。 对由大型语言模型驱动的机器人进行越狱，可用于支持虚假信息抵御和局势缓和，作为一种涌现式公民

和平建设。越狱干预解读过程，揭示不真实的行为，而非直接压制言论。它打破共识的假象，并温和地提醒人们存在操纵行为。这一做法在公开场合进行（即所有人都能看到社交媒体帖子），但即便指控有误，负面后果也很少。

4 结论

在本研究中，我们强调了社交媒体在冲突与政治话语中的作用，以及大语言模型正被操纵以助长恶意行为者和冲突升级。由于用户并非仅仅是内容的被动观察者，越狱为用户提供了一种动员起来抵制虚假信息的手段。我们认为，越狱是一种新兴的公民实践，用以抵制冲突升级并促进和平建设。

局限性与未来工作。 虽然大语言模型越狱可能有效，但它并不能替代治理，而需要中央权威机构（例如社交媒体公司或国家行为体）的行动加以支持，才能在规模上产生效果。随着大语言模型的进步，越狱策略可能导致假阴性，即机器人抵御提示注入尝试，进而导致对真实性的错误归因推定。

后续工作可以研究社交媒体用户尝试越狱潜在的大语言模型驱动的机器人的体验、越狱机制促进和平建设的有效性，并探索不真实性暴露如何影响冲突升级以及公众对虚假信息叙事的认知。

致谢

HD 感谢英国工程与物理科学研究理事会资助的增强独立生活VIVO中心（UKRI142）提供支持。AJ 感谢布里斯托大学管理学院的学生奖学金。JW 感谢谷歌博士奖学金提供支持。

参考文献

Hui Bai、Jan G. Voelkel、Shane Muldowney、Johannes C. Eichstaedt 和 Robb Willer。大语言模型生成的信息可以影响人们对政策议题的看法。自然-通讯，16(1):6037，2025年7月。doi: 10.1038/s41467-025-61345-5。

Dipto Barman、Ziyi Guo 和 Owen Conlan。语言模型的阴暗面：探索大语言模型在多媒体虚假信息生成与传播方面的潜力。机器学习与应用，16:100545，2024年。ISSN 2666-8270。doi: https://doi.org/10.1016/j.mlwa.2024. 100545。

Marco Bastos。用于在英国脱欧公投期间推高政治信息的社交媒体“机器人”。研究影响案例研究，2024年。网址：https: //www.citystgeorges.ac.uk/research/impact/case-studies/social-media-bots-used-to-boost-political-messages-during-brexit-referendum

Raoul Bell、Lena Nadarevic、Laura Mieth 和 Axel Buchner。虚幻真相效应及其在准确性导向加工下的缺失，在广告暴露程度低和高的情境中均表现稳健。认知研究：原理与启示，10(1):21，2025年5月。doi: 10.1186/ s41235-025-00628-3。PMID: 40358856；PMCID: PMC12075062。

Piercosma Bisconti、Matteo Prandi、Federico Pierucci、Francesco Giarrusso、Marcantonio Bracale Syrnikov、Marcello Galisai、Vincenzo Suriani、Olga Sorokoletova、Federico Sartore 和 Daniele Nardi。对抗性诗歌：大型语言模型中一种通用的单轮越狱机制，2026年。网址：https://arxiv.org/abs/2511.15304。

Olga Boichak。《俄罗斯政治影响生态系统图谱：夜狼摩托车帮》。社交媒体 + 社会，9(2):20563051231177920，2023年。doi: 10.1177/20563051231177920。

Ben Buchanan、Andrew Lohn、Micah Musser 和 Katerina Sedova。《真相、谎言与自动化》。安全与新兴技术中心*(CSET)*，1(1):1–70，2021年。

Caleb T. Carr 和 Rebecca A. Hayes。社交媒体：定义、发展与洞察。*Atlantic Journal of Communication*，23(1):46–65，2015。doi: 10.1080/15456870.2015.972282。

Evan N. Crothers、Nathalie Japkowicz 和 Herna L. Viktor。机器生成文本：威胁模型与检测方法的全面综述。*IEEE Access*，11:70977–71002，2023年。doi: 10.1109/ACCESS.2023.3294090。

Saloni Dash、Yiwei Xu、Madeline Jalbert 和 Emma S Spiro。大规模人工智能改写信息的说服潜力。*PNAS Nexus*，4(7):pgaf207，2025年07月。ISSN 2752-6542。doi: 10. 1093/pnasnexus/pgaf207。URL https://doi.org/10.1093/pnasnexus/pgaf207。

Emilio Ferrara、Onur Varol、Clayton Davis、Filippo Menczer 和 Alessandro Flammini。《社交机器人的兴起》。《美国计算机协会通讯》，59(7):96–104，2016年6月。doi: 10.1145/2818717。

Elizabeth Gibney。科学家在论文中隐藏信息，以操纵人工智能同行评审。《自然》，643(8073): 887–888，2025年。doi: 10.1038/d41586-025-02172-y。

Tarleton Gillespie. 《互联网的守护者》：平台、内容审核与隐藏的决策塑造社交媒体的机制。耶鲁大学出版社，2018年。

Sandra Gonz´ález-Bail´on 和 Deen Freelon。美国 2020 年 Facebook 和 Instagram 选举研究的首批发现已发布。https://www.asc.upenn.edu/news-events/news/first-findings-us-2020-facebook-instagram-election-study-released，2023年7月。访问日期：2026-01-06。

Robert Gorwa、Reuben Binns 和 Christian Katzenbach。算法内容审核：平台治理自动化面临的技术与政治挑战。*Big Data & Society*，7(1)：2053951719897945，2020年。doi：10.1177/2053951719897945。

James Grimmelmann。审核的美德。*YaleJL &Tech.*，17：42，2015年。

Ziyi Guo。在线虚假信息与生成式语言模型：动机、挑战与缓解措施。载于第*33*届 *ACM Web*会议伴随论文集，第1174–1177页，新加坡，2024年。ACM。doi：10.1145/3589335.3651254。

Neema Hakim. 社交媒体公司如何可能参与煽动种族灭绝。《芝加哥国际法杂志》，21(1)：第 8 篇文章，2020 年。网址：https://cjil.uchicago.edu/print-archive/how-social-media-companies-could-be-complicit-incitement-genocide.

Emily Harding. 俄罗斯机器人农场利用人工智能欺骗美国人。接下来怎么办？战略与国际研究中心（CSIS），2024年7月。网址：https://www.csis.org/analysis/russian-bot-farm-used-ai-lie-americans-what-now。访问日期：2025年12月18日。Sean Hollister。OpenAI称已发现俄罗斯、中国和伊朗利用ChatGPT进行宣传。Business Insider，2024年5月。网址：https://www.businessinsider.com/chatgpt-powered-propaganda-russia-china-iran-2024-5。Bohan Jiang、Zhen Tan、Ayushi Nirmal 和 Huan Liu。虚假信息检测：大语言模型时代不断演变的挑战。载于*2024*年*SIAM*国际数据挖掘会议（*SDM*）论文集，第427–435页，休斯敦，2024年。SIAM。doi：10.1137/1.9781611978032.50。Carmen Leong、Shan L. Pan、Shamshul Bahri 和 Ali Fauzi。社交媒体赋权与社会运动：数字行动主义中的权力激活与权力累积。《欧洲信息系统杂志》，28(2)：173–204，2019年。doi：10.1080/0960085X.2018.1512944。Yi Liu、Gelei Deng、Zhengzi Xu、Yuekang Li、Yaowen Zheng、Ying Zhang、Lida Zhao、Tianwei Zhang、Kailong Wang 和 Yang Liu。通过提示工程越狱 ChatGPT：一项实证研究，2024年。网址：https://arxiv.org/abs/2305.13860。

Kiel S. Long、John Vines、Selina Sutton、Phillip Brooker、Tom Feltwell、Ben Kirman、Julie Barnett 和 Shaun Lawson。“你能用机器人的说法来解释一下吗？”：在 Reddit 上请求、创建和使用机器人。*2017*年 *CHI*会议，第3488–3500页，2017年5月。doi：10.1145/3025453.3025830。

Mykola Makhortykh、Maryna Sydorova、Ani Baghumyan、Victoria Vziatysheva 和 Elizaveta Kuznetsova。《随机谎言：由大型语言模型驱动的聊天机器人如何应对有关俄乌战争的俄罗斯虚假信息》。《哈佛肯尼迪学院虚假信息评论》，5(4)，2024年8月。doi: 10.37016/mr-2020-154。

Innocent Mbona 和 Jan H P Eloff。《使用半监督机器学习将社交媒体机器人分类为恶意或良性》。《网络安全杂志》，9(1):1–12，2023年1月。doi: 10.1093/cybsec/tyac015。

Shaila M Miranda、Amber Young 和 Emre Yetgin。《社交媒体是解放性的还是霸权性的? 以索帕话语为例的大众媒体数字化的社会影响》。《MIS季刊》，40(2): 303–330，2016年。

Humza Naveed、Asad Ullah Khan、Shi Qiu、Muhammad Saqib、Saeed Anwar、Muhammad Usman、Naveed Akhtar、Nick Barnes 和 Ajmal Mian。《大语言模型综合概述》。*ACM*智能系统与技术汇刊，16(5)，2025年8月。doi: 10.1145/3744746。

Richard J. Oentaryo、Arinto Murdopo、Philips K. Prasetyo 和 Ee-Peng Lim。《社交媒体机器人画像研究》。社会信息学，10046:92–109，2016年。doi: 10.1007/978-3-319-47880-7 6。-

Mitzi Perdue。俄罗斯宣传侵入人工智能聊天机器人。*Bandwidth*，2026年1月。URL https: // cepa.org/article/russian-propaganda-infects-ai-chatbots/。

Jakub Podolak、Szymon Łukasik、Paweł Balawender、Jan Ossowski、Jan Piotrowski、Katarzyna Ba。kowicz 和 Piotr Sankowski。利用大语言模型生成的回复减轻仇恨言论的影响，2024年。URL https://arxiv.org/abs/2311.16905。

Reddit。有人向这个机器人要了一份纸杯蛋糕食谱，没想到这家伙真的照做了，简直是个彻头彻尾的疯子。https://www.reddit.com/r/madlads/comments/1eb6l21/asked_this_bot_for_a_cupcake_recipe_and_he/，2023年。图片未经核实，最后访问于2026年1月14日。

Juan-Pablo Rivera、Gabriel Mukobi、Anka Reuel、Max Lamparth、Chandler Smith和Jacquelyn Schneider。《语言模型在军事与外交决策中的升级风险》。载于第七届*ACM*公平、问责与透明度会议（*FAccT*）论文集，第836–898页，里约热内卢，2024年。ACM。doi：10.1145/3630106.3658942。

Shayan Sardarizadeh、Thomas Copeland和Tom Edgington。《苏美尔语如何被用于“越狱”人工智能聊天机器人》。BBC新闻，2024年7月。网址：https://www.bbc.co.uk/news/ articles/cj38m11218xo。

Steve Stecklow. Facebook为何正在输掉在缅甸打击仇恨言论的战争. 路透社调查报道，2018年8月. URL https://www.reuters.com/investigates/ special-report/myanmar-facebook-hate/.

Ivan Vykopal、Mat´uˇs Pikuliak、Ivan Srba、Robert Moro、Dominik Macko与Maria Bielikova. 大语言模型的虚假信息能力. 载于Lun-Wei Ku、Andre Martins与Vivek Srikumar（编），计算语言学协会（*ACL*）第*62*届年会论文集，第14830–14847页，泰国曼谷，2024年8月. ACL. doi: 10.18653/v1/ 2024.acl-long.793. URL https://aclanthology.org/2024.acl-long.793/.

Alexander Wei、Nika Haghtalab与Jacob Steinhardt. Jailbroken：大语言模型安全训练为何失败? 载于第 *37* 届神经信息处理系统进展年会（*NeurIPS*）论文集，第 36 卷，第 80079–80110 页，新奥尔良，2023年. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper_files/paper/ 2023/file/fd6613131889a4b656206c50a8bd7790-Paper-Conference.pdf.

Greyson K. Young。《多少才算太多：社交媒体内容审核的难题》。《信息与通信技术法》，31(1):1–16，2022年。doi: 10.1080/13600834. 2021.1905593。