

# 综合性顶刊中的Misinformation研究

Rachel

公众号：学海拾珠漫步知途

2025-05



1 文献检索结果

2 Nature、Science论文逐篇总结

3 其他3本期刊论文选择性摘读

## 目录

1 文献检索结果

2 Nature、Science论文逐篇总结

3 其他3本期刊论文选择性摘读

# 检索过程

- 检索条件：论文题目中至少包含下列一个关键词：misinformation、disinformation、rumor、rumour、fake news；
- 检索范围：Nature、Science、PNAS、Nature Communications、Science Advances官网数据库；
- 发表时间：2025年5月之前；

# 检索结果

共检索到42篇论文，去除5篇完全不相关论文：去除1篇Nature 1964年发表的论文，2篇PNAS论文，1篇Nature Communications论文，1篇Science Advances论文。[相关论文共37篇](#)。

**总体趋势：**自2019年起，研究论文数量呈明显增长趋势，尤其是在2024年达到最高点，体现了近年来misinformation议题的热度。

## 期刊贡献趋势：

- PNAS是该主题重要的发表期刊，尤其在2021和2024年显著突出。
- Nature Communications 自2022年后呈现快速增长态势。
- Nature 和 Science 虽然总体发文数量较少，但在2024年后明显增加，体现了高影响力期刊对该主题关注的增加。
- Science Advances 则表现出稳定的贡献。

# 1. 文献检索结果

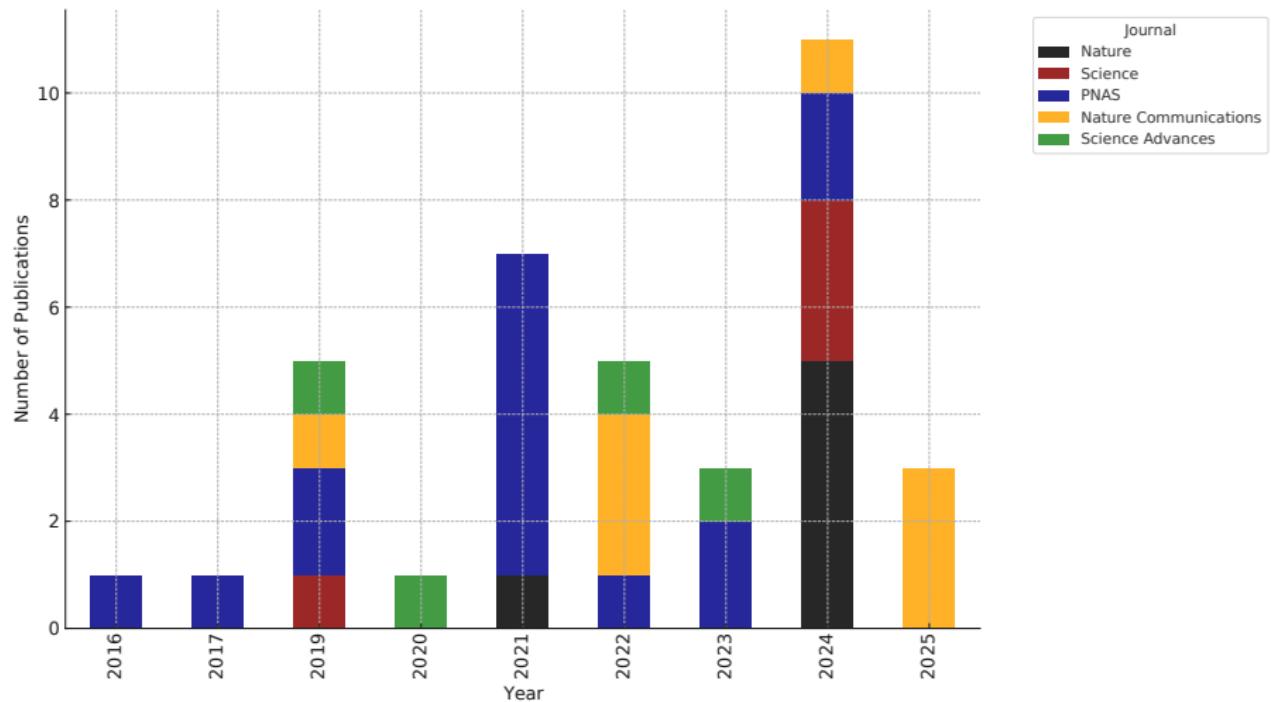


图 1: Misinformation主题在综合性顶刊中的发文趋势

## 检索结果—Nature (6篇)

- [1] Pennycook, Gordon, Ziv Epstein, Mohsen Mosleh, Antonio A. Arechar, Dean Eckles, and David G. Rand. “*Shifting attention to accuracy can reduce misinformation online.*” Nature 592 (2021): 590-594.
- [2] Ahmad, Wajeeha, Ananya Sen, Charles Eesley, and Erik Brynjolfsson. “*Companies inadvertently fund online misinformation despite consumer backlash.*” Nature 630 (2024): 123-125.
- [3] Aslett, Kevin, Zeve Sanderson, William Godel, Nathaniel Persily, Jonathan Nagler, and Joshua A. Tucker. “*Online searches to evaluate misinformation can increase its perceived veracity.*” Nature 625 (2024): 548-555.
- [4] Budak, Ceren, Brendan Nyhan, David M. Rothschild, Emily Thorson, and Duncan J. Watts. “*Misunderstanding the harms of online misinformation.*” Nature 630 (2024): 45-48.
- [5] Mosleh, Mohsen, Qi Yang, Tauhid Zaman, Gordon Pennycook, and David G. Rand. “*Differences in misinformation sharing can lead to politically asymmetric sanctions.*” Nature 634 (2024): 609.
- [6] McCabe, Stefan D., Diogo Ferrari, Jon Green, David M. J. Lazer, and Kevin M. Esterling. “*Post-January 6th deplatforming reduced the reach of misinformation on Twitter.*” Nature 630 (2024): 132-139.

## 检索结果—Science (4篇)

- [1] Grinberg, Nir, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. “*Fake news on Twitter during the 2016 U.S. presidential election.*” Science 363, no. 6425 (2019): 374-378.
- [2] Allen, Jennifer, Duncan J. Watts, and David G. Rand. “*Quantifying the impact of misinformation and vaccine-skeptical content on Facebook.*” Science 384, no. 6703 (2024): eadk3451.
- [3] Baribi-Bartov, Sahar, Briony Swire-Thompson, and Nir Grinberg. “*Supersharers of fake news on Twitter.*” Science 384, no. 6703 (2024): 979-982.
- [4] McLoughlin, Killian L., William J. Brady, Aden Goolsbee, Ben Kaiser, Kate Klonick, and M. J. Crockett. “*Misinformation exploits outrage to spread online.*” Science 386, no. 6721 (2024): 991-996.

## 检索结果—PNAS (15篇) I

- [1] Del Vicario, Michela, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi. “*The spreading of misinformation online.*” Proceedings of the National Academy of Sciences 113, no. 3 (2016): 554-559.
- [2] Jones, Nickolas M., Rebecca R. Thompson, Christine Dunkel Schetter, and Roxane Cohen Silver. “*Distress and rumor exposure on social media during a campus lockdown.*” Proceedings of the National Academy of Sciences 114, no. 44 (2017): 11663-11668.
- [3] Pennycook, Gordon, and David G. Rand. “*Fighting misinformation on social media using crowdsourced judgments of news source quality.*” Proceedings of the National Academy of Sciences 116, no. 7 (2019): 2521-2526.
- [4] Scheufele, Dietram A., and Nicole M. Krause. “*Science audiences, misinformation, and fake news.*” Proceedings of the National Academy of Sciences 116, no. 16 (2019): 7662-7669.
- [5] Caciato, Michael A. “*Misinformation and public opinion of science and health: Approaches, findings, and future directions.*” Proceedings of the National Academy of Sciences 118, no. 15 (2021): e1912437117.
- [6] Dahlstrom, Michael F. “*The narrative truth about scientific misinformation.*” Proceedings of the National Academy of Sciences 118, no. 15 (2021): e1914085117.

## 检索结果—PNAS (15篇) II

- [7] West, Jevin D., and Carl T. Bergstrom. “*Misinformation in and about science.*” Proceedings of the National Academy of Sciences 118, no. 15 (2021): e1912444117.
- [8] Yeo, Sara K., and Meaghan McKasy. “*Emotion and humor as misinformation antidotes.*” Proceedings of the National Academy of Sciences 118, no. 15 (2021): e2002484118.
- [9] Reyna, Valerie F. “*A scientific theory of gist communication and misinformation resistance with implications for health, education, and policy.*” Proceedings of the National Academy of Sciences 118, no. 15 (2021): e1912441117.
- [10] Smith, Steven T., Edward K. Kao, Erika D. Mackin, Danelle C. Shah, Olga Simek, and Donald B. Rubin. “*Automatic detection of influential actors in disinformation networks.*” Proceedings of the National Academy of Sciences 118, no. 4 (2021): e2011216118.
- [11] Green, Jon, William Hobbs, Stefan McCabe, and David Lazer. “*Online engagement with 2020 election misinformation and turnout in the 2021 Georgia runoff election.*” Proceedings of the National Academy of Sciences 119, no. 34 (2022): e2115900119.
- [12] Ceylan, Gizem, Ian A. Anderson, and Wendy Wood. “*Sharing of misinformation is habitual, not just lazy or biased.*” Proceedings of the National Academy of Sciences 120, no. 4 (2023): e2216614120.

## 检索结果—PNAS (15篇) III

- [13] Kozyreva, Anastasia, Stefan M. Herzog, Stephan Lewandowsky, Ralph Hertwig, Philipp Lorenz-Spreen, Mark Leiser, and Jason Reifler. “*Resolving content moderation dilemmas between free speech and harmful misinformation.*” Proceedings of the National Academy of Sciences 120, no. 7 (2023): e2210666120.
- [14] Stewart, Alexander, Antonio A. Arechar, David G. Rand, and Joshua B. Plotkin. “*The distorting effects of producer strategies: Why engagement does not reveal consumer preferences for misinformation.*” Proceedings of the National Academy of Sciences 121, no. 10 (2024): e2315195121.
- [15] Sultan, Mubashir, Alan N. Tump, Nina Ehmann, Philipp Lorenz-Spreen, Ralph Hertwig, Anton Gollwitzer, and Ralf H. J. M. Kurvers. “*Susceptibility to online misinformation: A systematic meta-analysis of demographic and psychological factors.*” Proceedings of the National Academy of Sciences 121, no. 47 (2024): e2409329121.

## 检索结果—Nature Communications (8篇) I

- [1] Bovet, Alexandre, and Hernán A. Makse. “*Influence of fake news in Twitter during the 2016 US presidential election.*” Nature Communications 10 (2019): 7.
- [2] Arruda, Guilherme Ferraz de, Lucas G. S. Jeub, Angélica S. Mata, Francisco A. Rodrigues, and Yamir Moreno. “*From subcritical behavior to a correlation-induced transition in rumor models.*” Nature Communications 13 (2022): 3049.
- [3] Mosleh, Mohsen, and David G. Rand. “*Measuring exposure to misinformation from political elites on Twitter.*” Nature Communications 13 (2022): 7144.
- [4] Pennycook, Gordon, and David G. Rand. “*Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation.*” Nature Communications 13 (2022): 2333.
- [5] Winter, Kevin, Matthew J. Hornsey, Lotte Pummerer, and Kai Sassenberg. “*Public agreement with misinformation about wind farms.*” Nature Communications 15 (2024): 8888.
- [6] Kim, Junsol, Zhao Wang, Haohan Shi, Hsin-Keng Ling, and James Evans. “*Differential impact from individual versus collective misinformation tagging on the diversity of Twitter (X) information engagement and mobility.*” Nature Communications 16 (2025): 973.

## 检索结果—Nature Communications (8篇) II

- [7] Stagnaro, Michael Nicholas, and Eran Amsalem. “*Factual knowledge can reduce attitude polarization.*” Nature Communications 16 (2025): 3809.
- [8] Maertens, Rakoen, Jon Roozenbeek, Jon S. Simons, Stephan Lewandowsky, Vanessa Maturo, Beth Goldberg, Rachel Xu, and Sander van der Linden. “*Psychological booster shots targeting memory increase long-term resistance against misinformation.*” Nature Communications 16, no. 1 (2025): 2062.

## 检索结果—Science Advances (4篇)

- [1] Guess, Andrew, Jonathan Nagler, and Joshua Tucker. “*Less than you think: Prevalence and predictors of fake news dissemination on Facebook.*” Science Advances 5, no. 2 (2019): eaau4586.
- [2] Allen, Jennifer, Baird Howland, Markus Mobius, David Rothschild, and Duncan J. Watts. “*Evaluating the fake news problem at the scale of the information ecosystem.*” Science Advances 6, no. 14 (2020): eaay3539.
- [3] Roozenbeek, Jon, Sander van der Linden, Beth Goldberg, Steve Rathje, and Stephan Lewandowsky. “*Psychological inoculation improves resilience against misinformation on social media.*” Science Advances 8, no. 34 (2022): eab06254.
- [4] Broniatowski, David A., Joseph R. Simons, Jiayan Gu, Amelia M. Jamison, and Lorien C. Abroms. “*The efficacy of Facebook’s vaccine misinformation policies and architecture during the COVID-19 pandemic.*” Science Advances 9, no. 37 (2023): eadh2132.

# 目录

1 文献检索结果

2 Nature、Science论文逐篇总结

3 其他3本期刊论文选择性摘读

TL;DR

**研究问题:** misinformation传播的驱动因素；衡量misinformation传播的后果；衡量misinformation的影响规模；抑制虚假信息的手段及其有效性；**不研究如何识别misinformation，主要通过外部“域名黑名单”列表来判定某个URL是否指向misinformation网站。**

**话题领域:** 聚焦于**美国**的政治、选举、covid-19疫苗，主要使用Twitter数据，少部分使用Facebook数据；具体数据主要为包含misinformation的平台外部URL；

**研究方法:** 主要采用调查/行为实验/社交媒体平台**获取**真实实验数据和大规模社交媒体真实数据，结合对数据的统计分析和描述，使用回归方法(断点回归，DID等)进行因果推断。10篇论文中，只有一篇Science 2019论文(**Grinberg et al., 2019**)使用了复杂网络方法做附加数据分析。

# Misunderstanding the harms of online misinformation (Nature 2024)

## 1. 标题

对在线错误信息危害的误解

## 2. 作者及单位

Ceren Budak; Brendan Nyhan; David M. Rothschild; Emily Thorson; Duncan J. Watts (Nature)

- 美国密歇根大学信息学院（安娜堡）
- 美国达特茅斯学院政府系（汉诺威，新罕布什尔州）
- 微软研究院（纽约）
- 美国雪城大学麦克斯韦尔公民与公共事务学院（雪城）
- 宾夕法尼亚大学计算机与信息科学系、安嫩伯格传播学院及运筹、信息与决策系（费城）

## 3. 文献来源

Budak, Ceren, et al. “*Misunderstanding the harms of online misinformation.*” Nature 630.8015 (2024): 45-53.

## 4. 文献类型与关键词

文章类型：综述/观点评论

英文关键词：online misinformation; social media; exposure; algorithms; polarization

**研究动机:** 梳理和汇总现有行为科学领域的实证研究 (*Treating the USA and other Western countries as the default*)，识别并纠正公众话语对在线错误信息暴露及其影响的三大常见误解：① 社交媒体对错误信息的暴露程度高；② 算法主导这一暴露过程；③ 错误信息是社会极化和政治暴力等广泛社会问题的主要原因。

**研究问题:**

- 在线社交媒体上，普通用户平均接触错误信息和极端内容的水平及分布特征；
- 算法推荐与用户主动需求在决定错误信息暴露中的作用机制及相对贡献为何；
- 社交媒体上错误信息或极端内容的暴露是否构成对社会极化、政治暴力等宏观社会问题的直接因果驱动。

**研究方法:** 视角性文献综述 (Perspective) 方法，系统回顾并综合行为科学领域的关键实证研究；

# Misunderstanding the harms of online misinformation (Nature 2024)

## 研究结果：

- 平均而言，社交媒体用户接触错误信息和极端内容的频率极低，且这种暴露高度集中在少数“重度”或“边缘”用户群体；
- 算法推荐对整体错误信息暴露的贡献有限，更主要的是满足了用户已有的兴趣和需求；
- 现有实证研究未能证明社交媒体上错误信息暴露对社会极化、政治暴力等宏观社会问题具有可靠的因果影响；
- 行为科学证据主要集中在美国和西欧，非西方国家的研究与数据严重不足 (The key exception is China, which has a different set of social media companies.)。

**启发：**依据研究结果的第一条以及关于政治虚假新闻的类似研究结果<sup>1</sup>，为什么监管部门要下大力气政治网络谣言呢，仅仅是为了花大力气“拯救”少数“边缘”群体吗？

<sup>1</sup> Grinberg, Nir, et al. "Fake news on Twitter during the 2016 US presidential election." Science 363.6425 (2019): 374-378.

# Companies inadvertently fund online misinformation despite consumer backlash (Nature 2024)

## 1. 标题

公司在消费者反对浪潮中无意资助在线错误信息

## 2. 作者及单位

Wajeeha Ahmad; Ananya Sen; Charles Eesley; Erik Brynjolfsson

- 斯坦福大学管理科学与工程系, 斯坦福, 美国
- 卡内基梅隆大学海因茨信息系统与公共政策学院, 匹兹堡, 美国
- 斯坦福大学人本人工智能研究所, 斯坦福, 美国

## 3. 文献来源

Ahmad, Wajeeha, et al. “*Companies inadvertently fund online misinformation despite consumer backlash.*” Nature 630.8015 (2024): 123-131.

## 4. 文献类型与关键词

文章类型：实证研究（结合大规模描述性分析与行为实验）

英文关键词：Online misinformation; digital advertising; consumer backlash; information intervention; programmatic advertising

中文关键词：在线错误信息；数字广告；消费者反应；信息干预；程序化广告

**研究动机：**在线错误信息背后的一个重要的经济动因是广告收入对错误信息网站的资助，对此鲜有系统性实证研究。已有研究多聚焦需求侧的用户干预，缺乏对错误信息“供给侧”资助路径及其放大机制进行深入分析。

**研究目标：**结合大规模描述性分析与行为实验：

- 定量揭示各行业广告主及数字广告平台如何共同资助在线错误信息；
- 评估当消费者与企业决策者获知这一资助事实后，其在“退出”（exit）与“发声”（voice）方面的反应；
- 在此基础上，提出低成本、可扩展的“供给侧”信息干预对策，以削弱错误信息的经济可持续性。

## 研究问题:

- **错误信息网站**在何种程度上依赖广告收入，各行业广告主的入场比例与投放强度如何；
- **数字广告平台**在自动分发广告至错误信息网站中扮演何种放大角色，其相对于直接投放的影响力几何；
- **消费者**获知其偏好品牌在错误信息网站投放广告后，会通过“退出”与“发声”产生何种行为改变；
- **企业决策者**对本公司及数字平台在资助错误信息中的角色认知如何，信息揭示后对平台型解决方案的需求有何差异；

## 研究方法:

- **描述性分析**: 汇集 NewsGuard、Global Disinformation Index 及先前研究中标识的错误信息网站列表，并结合 Oracle Moat Pro 平台 2019–2021 年日常网页抓取广告数据，构建包含 5,485 个新闻网站（其中 1,276 个错误信息网站）与 42,595 家广告主的 9,539,847 条广告投放实例数据库，量化行业分布与平台效应；
- **消费者实验**: 针对 4,039 名美国互联网用户，设计“五组信息处理”（包括仅公司级、仅平台级、公司+平台、公司排名等），通过 US\$25 礼品卡激励测量真实“退出”（切换礼品卡偏好）与“发声”（在线请愿点击）行为反应；
- **高管实验**: 对 442 名企业高管进行信息揭示实验，测量其获知数字广告平台放大效应后的后验信念变化与对“平台型解决方案”（即透明化工具）的需求强度。

# 研究结果

**描述性发现：**74.5% 的错误信息网站依赖广告收入，46%-82% 的传统行业公司至少一次在此类网站投放广告；使用数字广告平台的广告主中，79.8% 在任意一周出现在错误信息网站，而未用平台的仅为 7.74%，数字平台显著放大资助规模 ( $P<0.001$ )。

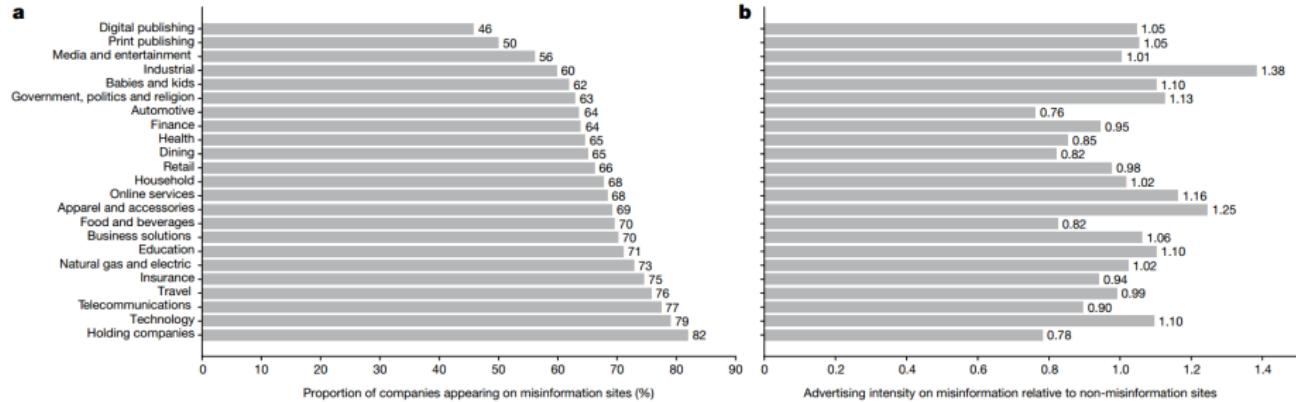


图 2: 按行业划分的在错误信息网站上出现广告的公司

**消费者行为：**公司级信息处理使消费者“退出”(切换为其他品牌)率提升13个百分点，组合与排名信息同样显著驱动消费者切换偏好，结果见图3；仅平台级信息处理则使“发声(点击平台请愿链接表达担忧)”行为提高5个百分点，结果见图25。

**Table 1 | Average treatment effects on exit**

	Switch in preference	Switch to lower preference	Switch in category	Switch to lower misinformation				
	1	2	3	4	5	6	7	8
Company (T1)	0.13*** (0.01) <0.001	0.13*** (0.01) <0.001	0.08*** (0.01) <0.001	0.08*** (0.01) <0.001	0.05*** (0.01) <0.001	0.05*** (0.01) <0.001	1.03** (0.48) 0.031	0.69* (0.38) 0.075
Platform (T2)	0.03*** (0.01) 0.010	0.03** (0.01) 0.012	0.01 (0.01) 0.114	0.01 (0.01) 0.118	0.02* (0.01) 0.076	0.01 (0.01) 0.130	0.52 (0.54) 0.335	0.23 (0.48) 0.629
Company and platform (T3)	0.10*** (0.01) <0.001	0.10*** (0.01) <0.001	0.06*** (0.01) <0.001	0.06*** (0.01) <0.001	0.04*** (0.01) <0.001	0.04*** (0.01) <0.001	0.69 (0.49) 0.157	0.28 (0.38) 0.452
Company ranking (T4)	0.08*** (0.01) <0.001	0.08*** (0.01) <0.001	0.06*** (0.01) <0.001	0.06*** (0.01) <0.001	0.03*** (0.01) 0.006	0.02** (0.01) 0.015	1.57*** (0.50) 0.002	0.95** (0.39) 0.015
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Control group mean	0.04	0.04	0.02	0.02	0.03	0.03	0.65	0.65
Observations	4039	4039	4039	4039	4039	4039	430	430

图 3：退出的平均处理效应

量化了不同信息干预对消费者“退出”行为的差异性影响，证明“供给侧”信息透明化干预可通过改变消费者选择显著削弱在线错误信息的经济供给。

**Table 2 | Average treatment effects on voice**

	Company		Platform	
	1	2	3	4
Company (T1)	0.02	0.02	-0.02	-0.02
	(0.02)	(0.02)	(0.02)	(0.02)
	0.180	0.257	0.286	0.349
Platform (T2)	-0.01	-0.01	0.05***	0.05***
	(0.02)	(0.02)	(0.02)	(0.02)
	0.546	0.007	0.436	0.006
Company and platform (T3)	-0.00	-0.00	-0.01	-0.01
	(0.02)	(0.02)	(0.02)	(0.02)
	0.863	0.940	0.638	0.708
Company ranking (T4)	0.04**	0.04**	-0.03*	-0.03
	(0.02)	(0.02)	(0.02)	(0.02)
	0.047	0.049	0.080	0.115
Controls	No	Yes	No	Yes
Control group mean	0.15	0.15	0.14	0.14
Observations	4039	4039	4039	4039

“发声”行为均以参与者点击请愿链接 (petition link click) 作为二元结果变量，具体分为：签署针对公司层面的请愿，签署针对平台层面的请愿。

图 4: 发声的平均处理效应

揭示了不同信息干预形式在引发消费者“发声”行为上的差异性影响，为设计针对错误信息的“供给侧”信息透明化策略提供了实证依据

Table 1中四种“信息干预”形式：

**T1 “公司层面”信息（Company only）**

向参与者提供事实性信息：告知他们最初首选的礼品卡品牌曾在近期出现在错误信息网站上投放广告。该处理组考察消费者在获悉自身偏好品牌直接资助错误信息后，是否改变消费选择。

**T2 “平台层面”信息（Platform only）**

向参与者提供事实性信息：说明使用数字广告平台的公司在近期出现于错误信息网站的概率大约是未使用此类平台公司的十倍。该处理组用于度量消费者对“平台责任”这一供给侧因素的反应，且不涉及对具体品牌的指名。

**T3 “公司+平台”信息（Company and platform）**

结合 T1 与 T2 两种信息：既告知参与者其首选礼品卡品牌曾在错误信息网站投放广告，又说明该品牌使用的数字广告平台使得其广告出现的可能性大幅放大（约十倍）。该处理组用于探究消费者在同时了解“公司行为”与“平台机制”后，对二者责任归属与行为改变的综合反应。

**T4 “公司排名”信息（Company ranking）**

向参与者展示六家礼品卡公司在过去某一年（2019、2020 或 2021）中，按在错误信息网站上投放广告强度从高到低的排名顺序，并个性化保证其首选品牌不位列末位。该处理不仅告知首选品牌的投放情况，还通过横向比较向消费者揭示各品牌相对表现，以测试“可比较排名”对其消费选择的导向作用。

Table 1 中消费者“退出”(exit) 行为的四种维度：

### **Switch in preference**

二元因变量，若干预后参与者放弃最初首选的礼品卡公司，转而选择其他任一公司，则取值 1；否则取值 0。

### **Switch to lower preference**

二元因变量，若干预后参与者将选择从其最初首选公司切换至其偏好程度更低的公司（偏好程度由对六个选项分配的权重决定），则取值 1；否则取值 0。

### **Switch in category**

二元因变量，若干预后参与者跨产品类别切换礼品卡（例如从出行类切换至快餐类），则取值 1；否则取值 0。

### **Switch to lower misinformation**

连续因变量，仅对发生切换的子样本 ( $n = 430$ ) 进行回归，取值等于参与者最终选择公司与其最初首选公司在“错误信息网站广告投放强度”上的差值；数值越大，表示切换至在错误信息网站上投放广告更少的公司。

# 研究结果

**高管认知：**仅 20% 的受访决策者确信本公司广告曾出现在错误信息网站；在“不确定”子样本中，获知数字平台放大效应后，对平台解决方案的需求提升 36 个百分点 ( $P=0.008$ )，表明信息透明度可显著激发企业自我纠偏意愿。

**Table 4 | Treatment effects based on prior beliefs**

	Posterior platform belief		Platform solution demand	
	Certain	Uncertain	Certain	Uncertain
	1	2	3	4
Treatment	39.98**	144.25**	-0.07	0.36**
	(20.23)	(60.23)	(0.06)	(0.13)
	0.049	0.022	0.213	0.008
Controls	Yes	Yes	Yes	Yes
Control group mean	90.23	80.80	0.37	0.43
Observations	286	68	286	68

对问卷问题“在过去三年（2019–2021）中，您认为贵公司是否在错误信息网站上投放过广告？”回答“*No*”（即认为未投放）的那一类子样本（共  $n = 354$ ），并基于对该回答的自评置信度，再细分为：Certain（确定组， $n = 286$ ）：回答时选择了“*Somewhat sure*”、“*Sure*”或“*Very sure*”的参与者；Uncertain（不确定组， $n = 68$ ）：回答时选择了“*Unsure*”或“*Very unsure*”的参与者。

**图 5：**基于先验信念的处理效应

## Posterior platform belief

“后验平台信念”是信息干预后用于衡量参与者对数字广告平台在资助在线错误信息中作用的信念更新程度的连续型变量。具体而言，参与者在干预前已知“不使用数字平台的网站上，平均每月有8家公司投放广告”的事实；干预后，被要求估计“使用数字广告平台的网站上，每月平均有多少家公司投放广告”。他们所给出的数值（经Winsorize去除极端值处理）即构成“后验平台信念”，数值越高表明参与者越倾向于认为数字广告平台显著放大了错误信息网站的广告投放量。

## Platform solution demand

衡量决策者对“平台型解决方案”兴趣的二元结果变量，在实验中，参与者被告知可在两种后续信息服务中二选一：①“了解哪些平台最少在错误信息网站上投放广告”；②“了解哪些分析技术能提升广告投放效果”。若参与者在该选择中选定第一项—希望获得针对数字广告平台如何减少对错误信息网站投放的详细信息—则“Platform solution demand”取值为1；若选定第二项或不选择任何信息，则取值为0。

# 研究结论

- 在线错误信息的主要资助来源为广告，且被广告平台算法大幅放大；
- 当广告主与消费者对资助事实的认知纠偏后，将通过退出与发声两条路径对平台与广告主施加实质压力；
- 提升信息透明度—包括向广告主提供精准投放明细与向消费者揭示公司投放排名—可作为可扩展的“供给侧”干预，以减少错误信息的广告收入并抑制其长期供给。

# Shifting attention to accuracy can reduce misinformation online (Nature 2021)

## 1. 标题

将注意力转向准确性可以减少在线错误信息

## 2. 作者及单位

Gordon Pennycook; Ziv Epstein; Mohsen Mosleh; Antonio A. Arechar; Dean Eckles;  
David G. Rand

- 加拿大里贾纳大学希尔/利文商学院, 萨斯喀彻温省里贾纳, 加拿大
- 麻省理工学院, 剑桥, 美国
- 英国埃克塞特大学商学院科学、创新、技术与创业系, 埃克塞特, 英国
- 墨西哥中央经济与教学研究中心 (CIDE), 阿瓜斯卡连特斯, 墨西哥

## 3. 文献来源

Pennycook, Gordon, et al. “*Shifting attention to accuracy can reduce misinformation online.*” *Nature* 592.7855 (2021): 590-595.

## 4. 文献类型与关键词

文章类型：实证研究（结合多项调查实验、现场实验及计算建模分析）

英文关键词：misinformation; accuracy nudge; attention; social media; sharing behavior; experimental intervention

**研究动机：**社交媒体上虚假和误导性信息的泛滥已成为近年来公众辩论和学术研究的焦点。这类错误信息不仅会导致个体形成不准确的信念，还可能加剧社会在基本事实上的党派分歧与对立。因此，学术界和相关从业者都迫切希望理解人们为什么会分享这些虚假信息，并积极寻求能够有效减少其传播的解决方案。

## 研究目标

- 首先，探究并阐明个体在社交媒体上分享虚假信息的深层原因；
- 其次，基于对这些原因的理解，设计并验证一种或多种能够有效提升用户分享内容质量、减少虚假信息传播的干预措施；

## 研究问题

- 人们分享虚假信息的驱动因素是什么？具体而言，分享行为是更多源于用户对信息准确性的困惑（即错误地认为虚假信息是准确的），还是因为用户在分享时更偏好党派立场等其他因素而非准确性，亦或是仅仅因为在分享决策的当下，用户的注意力并未集中在内容的准确性上；
- 在准确性判断与分享意愿之间是否存在脱节？即，人们是否会分享那些他们自己都认为不一定准确的内容；
- 通过微妙地将用户的注意力引导至“准确性”这一概念，能否提升他们后续分享新闻内容的质量，特别是减少虚假新闻的分享；
- 这种基于注意力的干预措施的效果在真实的社交媒体环境中（如Twitter）是否依然存在。

# 研究方法

## 系列在线调查实验（Survey Experiments）

**研究1：**招募了1015名美国MTurk参与者，随机将其分配到“准确性判断”组或“分享意愿”组，评估他们对36个（一半真实、一半虚假；一半亲民主党、一半亲共和党）新闻标题的反应，以探究准确性判断与分享意愿的差异；

**研究2：**招募了401名来自Lucid的具有全国代表性的美国参与者，评估他们在社交媒体分享决策中对准确性及其他内容维度（如趣味性、党派一致性等）的相对重要性排序；

**研究3、4、5（准确性启动实验）：**分别招募了MTurk参与者（研究3, n=727；研究4, n=780）和Lucid配额样本参与者（研究5, n=1268，更具全国代表性）。在这些实验中，处理组的参与者在进行主要分享意愿评估任务前，会先被要求评估一个不相关的中性新闻标题的准确性，以此作为启动“准确性”概念的手段。对照组则直接进行分享意愿评估。研究5还额外设置了主动对照组（评估幽默性而非准确性）和一个“重要性处理”组（在研究开始时询问参与者只分享准确内容的重要性）。

**研究6（注意力机制量化实验）：**招募了710名MTurk参与者。处理组参与者在决定是否分享每一个新闻标题之前，被强制要求先评估其准确性。通过比较处理组与对照组的分享行为差异，以及处理组中对准确性的判断，来量化注意力不集中、对准确性的困惑以及刻意分享虚假内容这三个因素在虚假信息分享中所占的比重。

### Twitter实地实验（Field Experiment on Twitter）

**研究7：**选取了5379名过去曾分享过两个著名低可信度新闻网站（Breitbart.com, Infowars.com）链接的Twitter用户。研究人员通过私信向这些用户发送请求，请他们评估一个非政治性新闻标题的准确性（作为干预手段）。实验采用阶梯式随机分组设计（stepped-wedge design），追踪并比较用户在收到干预信息后24小时内所分享新闻内容的质量（基于专业事实核查机构的评分）相较于未收到信息时的变化。

### 计算分析（Computational Analyses）

除了实验研究，本研究还运用了计算模型来进一步分析和佐证研究发现，包括构建和拟合一个有限注意力效用模型（limited-attention utility model），以及进行网络层面的虚假信息传播动态模拟。

# 研究结果

**分享决策与准确性判断存在显著脱节：**研究1-2发现，新闻标题的真实性对人们的分享意愿影响很小，但对其准确性判断影响巨大。相反，新闻的党派倾向性对分享意愿的影响远大于真实性。尽管如此，大多数参与者在被直接问及时，均表示只分享准确的新闻对他们而言非常重要。

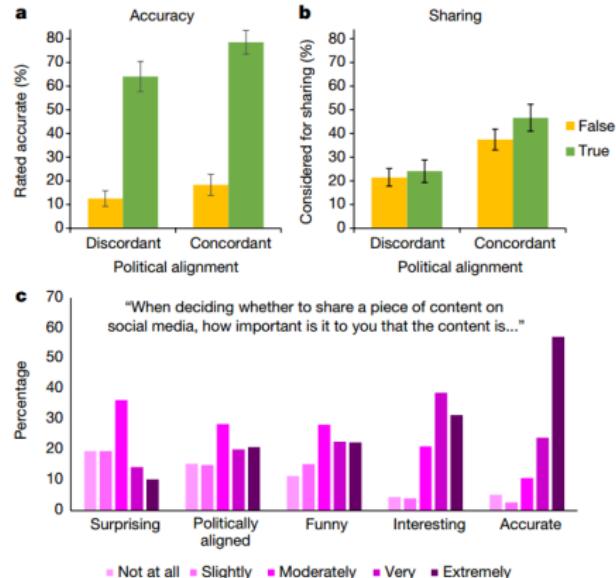


图 6: 分享意愿的辨别力远逊于准确性判断——尽管人们普遍希望只分享准确的内容

**注意力干预能有效提升分享质量：**研究3、4、5的调查实验一致表明，在分享任务开始前，通过简单评估一则无关新闻的准确性来微妙地提醒参与者关注“准确性”，能够显著提高他们辨别并分享高质量新闻（即减少分享虚假新闻，同时不影响真实新闻的分享）的倾向。

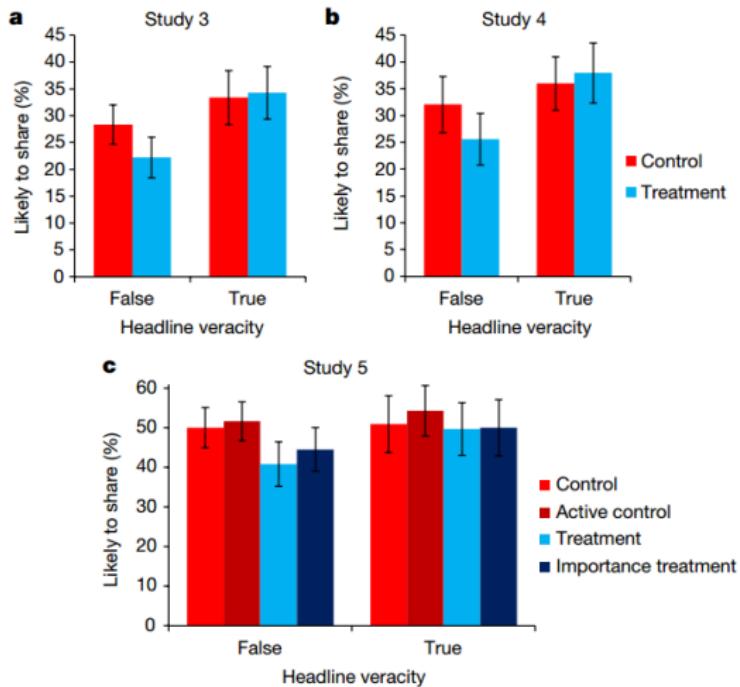


图 7: 引导调查受访者思考准确性，可提升其愿意分享的新闻标题的真实性

**注意力是关键机制：**进一步分析表明，这种准确性提醒干预之所以有效，是因为它将人们的注意力引向了内容的准确性维度。干预使得参与者在面对那些他们本就认为不太准确的标题时，分享意愿下降得最为明显。

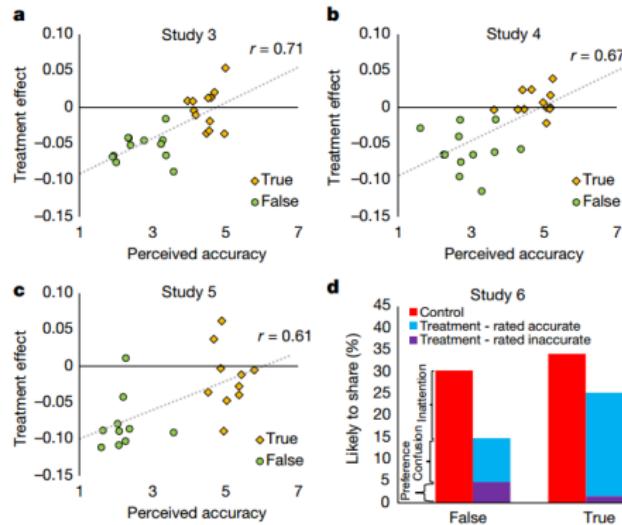


图 8: 注意力不集中在虚假信息的分享中具有重要作用

**注意力不集中是分享虚假信息的主要原因：**研究6的量化分析显示，在实验室情境下，人们分享虚假新闻的意愿中，约51.2%可以归因于注意力不集中（即未考虑到准确性），33.1%归因于对信息准确性的困惑（即错误地认为其是准确的），而只有15.8%是出于明知其假仍选择分享的“偏好”。

**实地干预的有效性：**研究7在Twitter上进行的实地实验证实，向曾分享过低质量信息的用户发送一条要求评估无关新闻准确性的私信，确实能使其在接下来24小时内分享的新闻来源质量得到改善。具体表现为，这些用户分享了更多来自主流、高可信度新闻网站的链接，同时减少了对那些被专业事实核查机构评为不可信的超党派网站内容的分享。

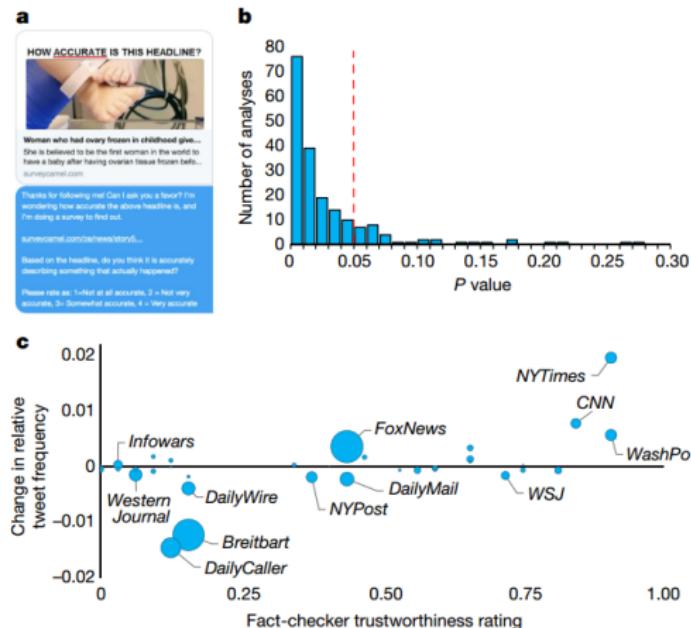


图 9：向Twitter用户发送信息，征询他们关于单个非政治性新闻标题准确性的看法，能够提高其随后分享新闻的质量

## 研究结论

- 人们在社交媒体上分享虚假信息，并非主要是因为他们不重视准确性或者有强烈的党派偏见压倒了对事实的追求，而更多是因为在快速浏览和互动的社交媒体环境中，他们的注意力常常被其他因素（如社交互动、情感共鸣、内容新奇性等）所吸引，从而忽略了对内容准确性的考量；
- 因此，通过简单、微妙的干预（如在用户分享前提醒他们思考准确性）将用户的注意力重新引导到“准确性”这一维度上，就能够显著提升他们分享新闻的整体质量，有效减少虚假信息的传播；
- 这些发现挑战了“人们分享虚假信息主要是因为党派立场而非事实”的流行观点，并表明用户的真实信念可能并不像其社交媒体动态所呈现的那样极端党派化；
- 本研究提供的基于注意力的干预策略具有良好的可扩展性和实践性，社交媒体平台可以考虑将其整合到产品设计中（例如，定期随机询问用户对某些新闻标题准确性的看法），作为一种低成本、非强制性的方式来改善在线信息生态系统，而无需依赖中心化的审查机构。

# Online searches to evaluate misinformation can increase its perceived veracity (Nature 2024)

## 1. 标题

在线搜索评估错误信息可能提高其感知真实性

## 2. 作者及单位

Kevin Aslett, Zeve Sanderson, William Godel, Nathaniel Persily, Jonathan Nagler, Joshua A. Tucker

- 中央佛罗里达大学政治、安全与国际事务学院
- 纽约大学社交媒体与政治中心
- 斯坦福大学法学院

## 3. 文献来源

Aslett, Kevin, et al. “*Online searches to evaluate misinformation can increase its perceived veracity.*” Nature 625.7995 (2024): 548-556.

## 4. 文献类型与关键词

文章类型：实证研究

英文关键词：misinformation; online search; perceived veracity; data voids; media literacy interventions

中文关键词：错误信息；在线搜索；感知真实性；数据空洞；媒介素养干预

## 研究动机

当下多项数字媒介素养干预（如Civic Online Reasoning课程）均建议人们通过在线检索（“自己做研究”，SOTEN）来评估新闻真实性，以期降低对错误信息的信任。但至今尚无实验证据明确检索行为本身对错误信息信念的影响；同时，搜索引擎中存在“数据空洞”（data voids），低质量信息可能在检索中占据较大比重，反而增强对错误信息的信任。因此，有必要实证考察在线检索对错误信息感知真实性的因果作用。

## 研究目标

系统性地评估在线搜索（SOTEN）在核查新闻信息真实性时，对个体信念（尤其是对错误信息和真实信息的信念）产生的实际影响。研究旨在挑战传统观念中认为在线搜索能减少错误信息信念的假设，并探究这种影响背后的潜在机制，例如用户在搜索过程中接触到的信息质量。此外，研究还旨在考察在线搜索对真实新闻信念的影响，并区分不同来源质量（主流媒体与低质量来源）新闻的差异效应。

# 研究问题

- 在线搜索评估虚假或误导性新闻文章的真实性，是否会增加人们对这些文章的相信程度？
- 如果在线搜索会增加对错误信息的相信程度，其背后的机制是什么？是否与搜索结果中接触到低质量、不可靠的信息有关？
- 在线搜索评估真实新闻文章的真实性，是否会增加人们对这些文章的相信程度？
- 在线搜索对真实新闻信念的影响，是否因新闻来源的质量（例如，主流媒体 vs. 低质量来源）而有所不同？
- 在线搜索对错误信息信念的影响，是否会随着文章发布时间的推移（例如，几天内 vs. 数月后）而发生变化？
- 对于高关注度事件（如COVID-19大流行）相关的错误信息，在线搜索的影响是否依然存在？哪些个体特征（如数字素养、意识形态一致性）和搜索行为（如使用的搜索词）与接触到低质量搜索结果相关？

# 研究方法

**实验设计：**研究共包含五项独立的实验。其中，研究1和研究5采用的是被试间随机对照试验设计，将被试随机分配到处理组（被鼓励在线搜索评估新闻）和控制组（不被提示在线搜索）。研究2、研究3和研究4采用的是被试内设计，即同一被试先在不被鼓励搜索的情况下评估文章，之后再被鼓励搜索并重新评估同一篇文章。

**参与者招募：**研究1至研究4的参与者通过在线调查公司Qualtrics招募的美国居民。研究5的参与者通过亚马逊的Mechanical Turk招募的美国居民。

**实验材料：**使用了在研究期间广为流传的真实新闻和虚假/误导性新闻文章，这些文章来源于主流媒体和低质量新闻源，并在其发表后较短时间内（通常为24-72小时内）呈现给参与者。研究4专门选取了与COVID-19相关的文章。

**信息质量评估：**研究5中使用NewsGuard服务提供的可靠性评分来评估搜索结果中新闻来源的质量。

**统计分析：**主要采用普通最小二乘法（OLS）回归模型进行数据分析，并控制了文章固定效应，标准误在个体和文章层面进行聚类。

## 数据收集

- **信念评估：**参与者使用分类量表（真实、虚假/误导性、无法确定）和七点序数表（从“绝对不真实”到“绝对真实”）评估文章的真实性。研究5还增加了一个四点序数表。
- **文章真实性判定：**研究团队雇佣了六名来自主要国家媒体机构的专业事实核查员，对每篇文章的真实性进行评估，以其众数评定结果作为文章真实性的基准。
- **数字痕迹数据：**研究5采用了定制的浏览器插件，用于收集处理组被试在评估文章前进行谷歌搜索时的搜索结果（前十条）以及他们访问的网址。
- **搜索行为与特征：**收集了被试使用的搜索关键词、数字素养水平、意识形态等信息。

**研究 1：**比较两组人，一组被要求上网搜索来判断新闻真假，另一组不要求。结果发现，被要求搜索的人反而更容易相信假新闻。

**研究 2：**让同一组人先凭感觉判断新闻真假，然后再让他们上网搜索后重新判断。结果显示，搜索后，这些人更容易把假新闻判断为真的。

**研究 3：**重复研究2的做法，但在新闻发布几个月后再进行测试。目的是看时间久了，有了更多事实核查信息后，搜索的影响会不会改变。结果发现，即使过了几个月，上网搜索仍然让人们更容易相信假新闻。

**研究 4：**采用与研究2和研究3类似的方法，但专门针对关于新冠疫情的假新闻。结果表明，对于疫情相关的假新闻，上网搜索同样会增加人们相信它的可能性。

**研究 5：**再次比较两组人（一组被要求用谷歌搜索，另一组不要求），但这次通过浏览器插件追踪他们实际看到的搜索结果。研究发现，那些搜索结果质量差（比如很多不靠谱网站链接）的人，更容易相信假新闻。

# 研究结果

**对错误信息的信念增加：**横跨五项实验一致表明，在线搜索以评估虚假新闻文章的真实性，实际上增加了参与者相信这些错误信息的概率。例如，在研究1中，被鼓励搜索的参与者将虚假/误导性文章评为真实的概率增加了0.057；研究2中增加了0.071；研究5中增加了0.107。

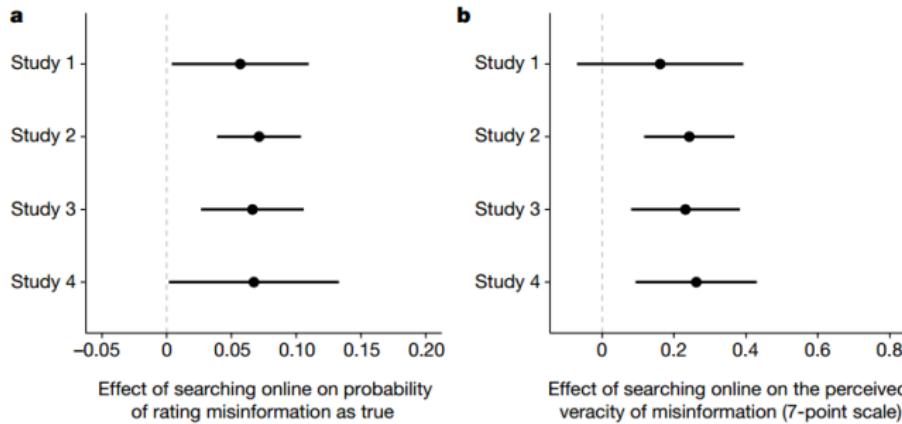


图 10: 在线搜索评估错误信息对错误信息信念的影响（研究 1 至 4）

**效应的持续性与普遍性：**即使在错误信息文章发表数月后（研究3），或在评估关于高关注度事件（如COVID-19，研究4）的错误信息时，在线搜索增加错误信息信念的效应依然存在。

**暴露于不可靠信息的普遍性:** 当搜索关于虚假/误导性新闻时, 个体接触到至少一个不可靠新闻链接的比例 (38%) 远高于搜索真实新闻时的比例 (15%)。

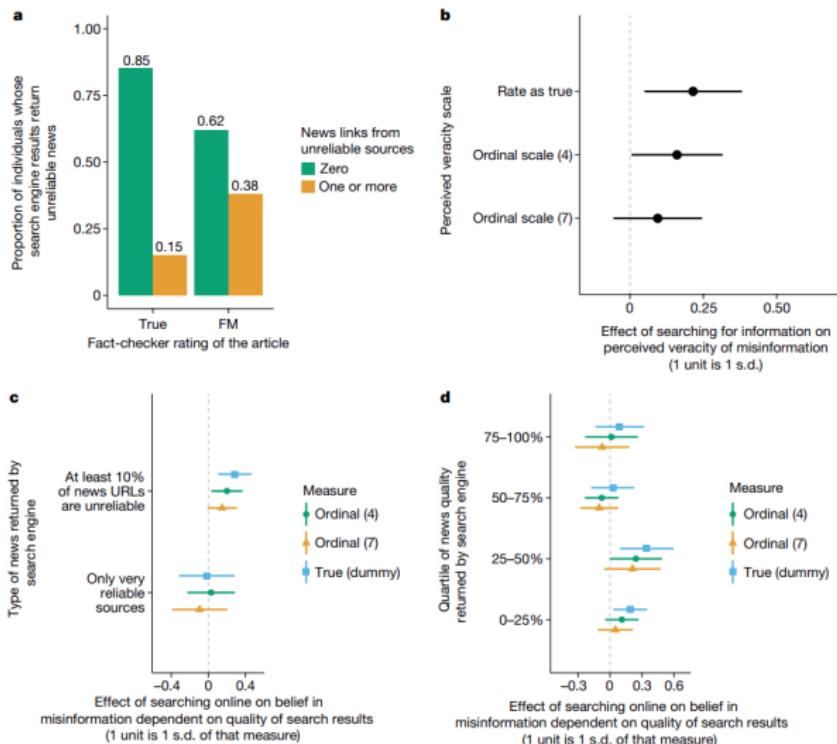


图 11: Google 搜索结果如何影响对错误信息的信念 (研究 5)

**信息质量的关键作用：**搜索效应集中在那些搜索引擎返回信息质量较低的个体中。当搜索结果包含较多来自低质量来源的真伪性证据时（即陷入“数据真空” data voids），参与者更可能相信错误信息。相比之下，当搜索结果主要由高质量信息构成时，这种信念增加的效应不显著。

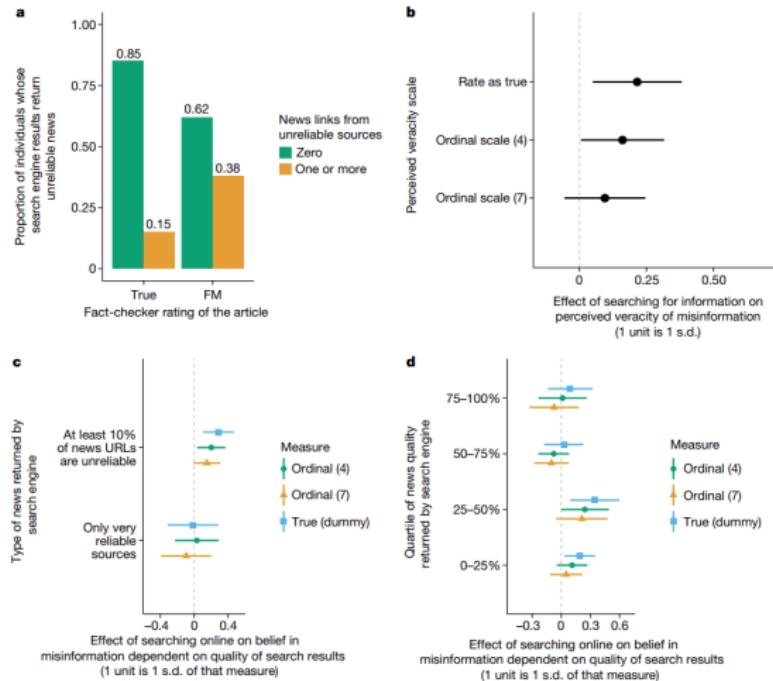


图 12: Google 搜索结果如何影响对错误信息的信念 (研究 5)

**搜索行为的影响：**使用虚假文章的标题或URL作为搜索查询词的参与者，更有可能接触到不可靠的搜索结果。数字素养较低的个体更倾向于使用此类搜索方式，也因此更容易接触到低质量信息。

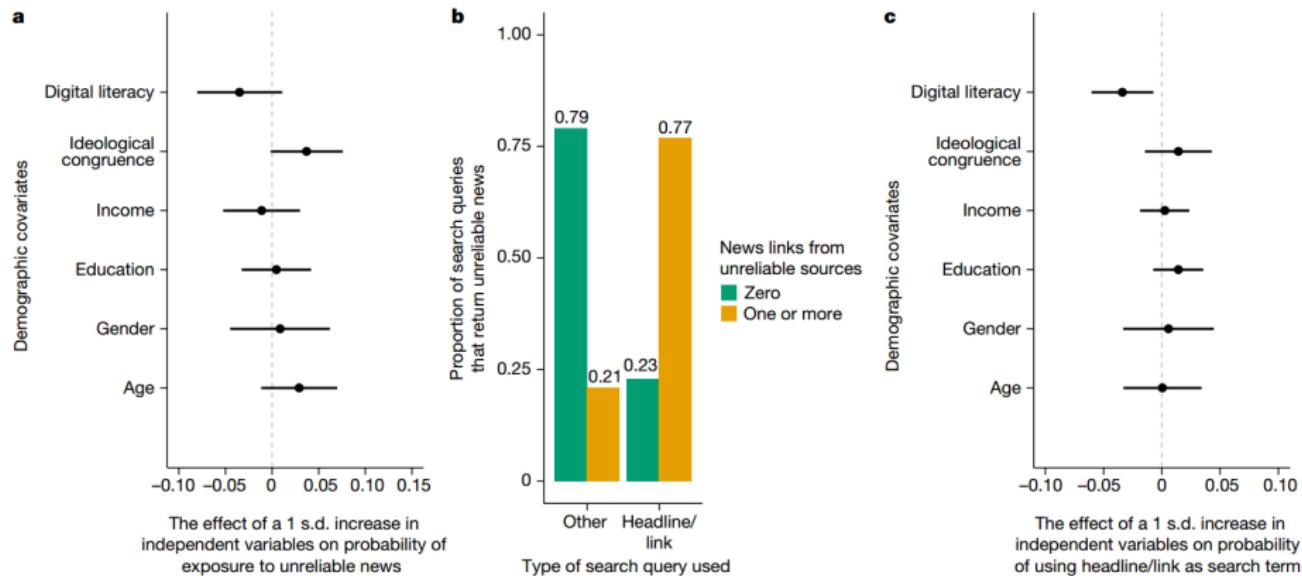


图 13: 对在线评估错误信息时暴露于不可靠新闻站点的个体进行的分析（研究 5）

**对真实新闻的信念影响：**在线搜索评估真实新闻也可能增加对其的相信程度。然而，这种效应在不同来源的新闻中存在差异：对于来自低质量来源的真实新闻，在线搜索显著增加了对其的相信程度；而对于来自主流媒体的真实新闻，这种效应不一致，有时甚至不显著或效应较小。

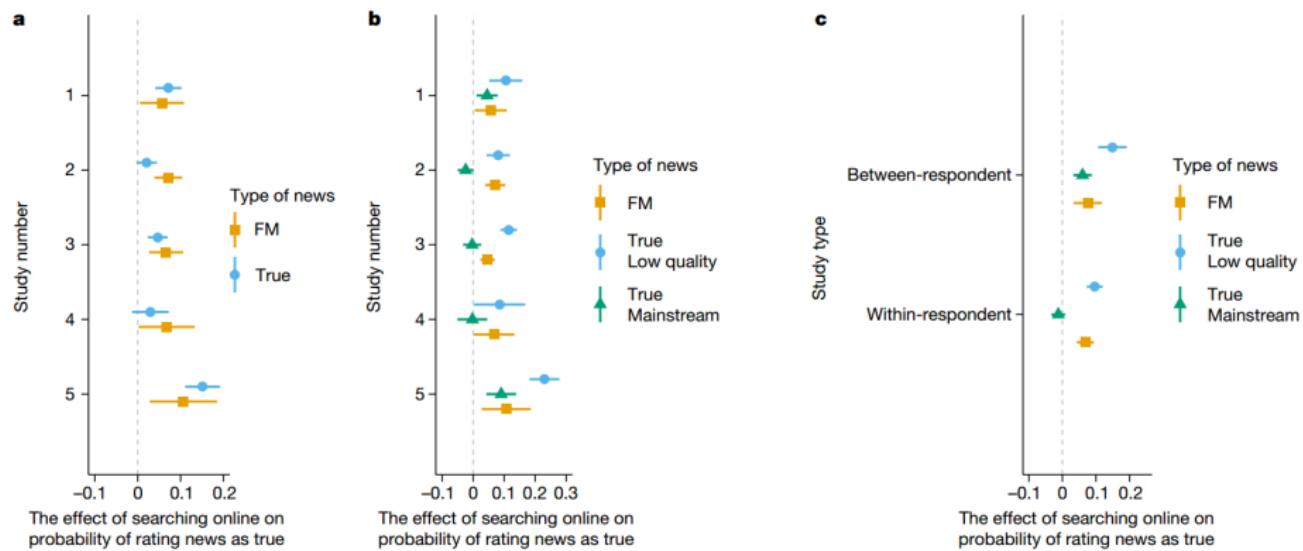


图 14: 在线搜索评估 (SOTEN) 对错误/误导性新闻和真实新闻信念的影响

# 研究结论

- 与普遍认知相反，在线搜索以评估错误信息的真实性，不仅不能有效减少反而可能增加人们对它的相信程度。这一发现对当前许多推荐使用搜索引擎作为辨别信息真伪工具的媒体素养干预措施提出了挑战。
- 搜索结果的质量是导致上述效应的关键因素。当个体搜索错误信息时，他们面临落入“数据真空”的风险，即搜索结果充斥着来自低质量来源的确证性证据，从而强化了对错误信息的信念。
- 个体的数字素养水平和采用的搜索策略（如直接复制粘贴标题进行搜索）会显著影响其接触到的信息质量，进而影响其判断。
- 虽然在线搜索可能有助于提升对来自低质量来源的真实新闻的信任度，但其对于主流来源的真实新闻的信念提升效应并不稳定，且通常小于其对错误信息信念的负面提升效应。
- 研究结果强调，媒体素养项目需要将其建议植根于经过实证检验的策略之上，同时搜索引擎公司也需投入资源解决本研究发现的挑战，例如改进算法以减少“数据真空”现象，或在搜索结果缺乏可靠信息时向用户发出警告。

**启发：**这或许揭示了，基于搜索的(虚假)事实核验(fact-checking)算法存在瓶颈的原因，例如，RAG方法在识别misinformation时，不一定有效。

# Differences in misinformation sharing can lead to politically asymmetric sanctions (Nature 2024)

## 1. 标题

错误信息分享差异可导致政治不对称制裁

## 2. 作者及单位

Mohsen Mosleh; Qi Yang; Tauhid Zaman; Gordon Pennycook; David G. Rand

- 牛津大学、埃克塞特大学
- 麻省理工学院、耶鲁大学
- 康奈尔大学

## 3. 文献来源

Mosleh, Mohsen, et al. "Differences in misinformation sharing can lead to politically asymmetric sanctions." *Nature* 634.8034 (2024): 609-616.

## 4. 文献类型与关键词

文章类型：实证研究

英文关键词：misinformation sharing; partisan asymmetry; social media enforcement; Twitter suspension

中文关键词：错误信息分享；党派不对称；社交媒体执法；Twitter停权

## 研究动机

随着社交媒体在信息分发中的主导地位日益凸显，平台运营者为遏制错误信息而实施的一系列中立反误信息政策（如下架、标记或封禁账号）却频繁遭到“对保守派用户存在政治偏见”的指责。本文提出，若不同政治群体在误信息传播行为上存在差异，即使平台严格执行中性政策，也可能因这些差异而产生不对称的制裁后果。

## 研究目标

- 系统量化不同政治立场用户在误信息（以低质量新闻链接为代理）分享上的差异；
- 考察此类差异与社交媒体平台对账号制裁（以Twitter封禁为例）之间的关联；
- 通过模拟完全政治中立的反误信息或反机器账号政策，评估单纯基于行为进行制裁时是否仍会产生党派上的不对称影响

## 研究问题

- 在美国2020年总统选举期间，被识别为亲特朗普/保守派的推特用户是否比亲拜登/自由派用户更容易被封禁？
- 亲特朗普/保守派用户是否分享了更多来自低质量新闻网站的链接？
- 当新闻质量的判定标准分别来自专业事实核查员、政治平衡的普通民众，乃至仅来自共和党普通民众时，上述在分享低质量新闻方面的不对称性是否依然存在？
- 在其他（共七个）涵盖不同时间（2016年至2023年）、不同平台（推特、脸书）及不同方法（社交媒体数据分析、调查实验）的数据集中，保守主义倾向与分享低质量新闻之间是否存在类似的关联？
- 在控制了分享低质量新闻、机器人账户可能性等因素后，用户的政 治倾向是否仍然是预测其账户被封禁的显著因素？
- 若实施一个仅基于分享低质量新闻（由政治平衡的普通民众判定）或仅基于机器人账户行为可能性的、完全政治中立的制裁政策，其是否仍会导致不同政治倾向用户群体间出现不对称的封禁率？

# 研究方法

## 主要方法—2020年美国大选推特数据分析

**样本与数据收集：**分析了2020年10月识别的9000名政治活跃的推特用户（4500名分享#Trump2020标签，4500名分享#VoteBidenHarris2020标签）。收集了这些用户在选举前（截至2020年10月）最近3200条推文数据，以量化其分享低质量新闻的倾向及其他特征，并在9个月后（2021年7月）核查其账户封禁状态。

### 核心变量测量：

- **低质量新闻分享：**通过用户分享的新闻网站链接的平均质量评分来衡量。评分来源包括8位专业事实核查员（针对60个网站）、970名经配额抽样的美国普通民众组成的政治平衡小组（针对60个网站）、Ad Fontes Media（283个域名）、Media Bias/Fact Check（3216个域名）及Lasser等人汇总的评级（4767个域名）。将原始质量评分转化为低质量新闻分享得分（1减去质量评分）。
- **政治倾向：**通过用户分享的选举标签、关注账户的意识形态（采用Barberá等人模型）、分享新闻来源的意识形态（采用Eady等人及Grinberg等人模型）进行测量，并通过主成分分析（PCA）整合为单一政治倾向综合指标。
- **其他控制变量：**包括账户的机器人可能性（采用Botsentinel模型）、使用攻击性语言的程度等。

**统计分析：**运用卡方检验比较不同群体封禁率，t检验比较分享新闻的平均质量，相关分析检验低质量新闻分享与保守主义倾向的关系，AUC曲线下面积评估不同变量对封禁的独立预测能力，并使用Probit回归和岭回归模型在控制多变量条件下预测账户封禁概率。

### 其他方法—政策模拟分析

**模拟了政治中立的反错误信息政策：**设定用户因分享低质量新闻链接（由政治平衡的普通民众判定，信任度低于0.25的网站为低质量）而被封禁的概率 $(P(\text{封禁}) = 1 - (1 - k)^L)$ ，其中L为分享的低质量链接数，k为政策严厉害度）。

**模拟了政治中立的反机器人账户政策：**设定用户因其机器人账户概率高于某一阈值（即人类可能性低于k）而被封禁。

## 其他方法—其他七个数据集的再分析

- 对涵盖2016至2023年间来自推特、脸书以及16个国家调查实验的七个额外数据集进行了分析。
- 具体包括：检验YouGov受访者在脸书（2016年）的分享行为；Prolific受访者在推特（2018年、2020年）的分享行为；不同方式抽样的推特用户（2021年、2022年、2023年）的分享行为。在这些数据中，均计算了保守主义倾向与分享低质量新闻（基于表1中的60个网站）之间的相关性。
- 分析了Ghezae等人数据中特定不实信息链接在推特上的分享情况，比较了保守派与自由派用户的分享差异（Wilcoxon符号秩检验），不实信息由事实核查员或政治平衡的普通民众判定。
- 分析了Arechar等人来自16个国家的调查实验数据，检验了保守主义倾向与分享不实COVID-19信息意愿之间的相关性，信息本身未标注来源。

## 研究结果-2020年美国大选推特数据分析结果

- 分享特朗普相关标签的用户被封禁的可能性是分享拜登相关标签用户的4.4倍（分别为19.6%对4.5%）。
- 分享特朗普相关标签的用户所分享的新闻来源，其平均质量显著更低。这一结论在使用专业事实核查员评分（低2.52个标准差）、政治平衡的普通民众评分（低2.17个标准差，中位数用户分享低质量网站链接多4倍）、乃至仅用共和党普通民众评分（低1.29个标准差）时均成立。其他多种专家评级体系也显示了类似结果。
- 无论使用专家评分还是政治平衡的普通民众评分，用户的保守主义意识形态（基于其关注账户或分享的新闻网站估算）均与分享低质量新闻呈显著正相关（专家评分下相关系数 $r$ 介于0.73至0.88之间，普通民众评分下 $r$ 介于0.73至0.82之间）。
- 在预测账户封禁方面，分享低质量新闻的预测能力（AUC介于0.68至0.72之间）与用户的党派/意识形态指标的预测能力（AUC介于0.67至0.71之间）相当。
- 在包含政治倾向、分享低质量新闻、机器人可能性、攻击性语言使用等多个变量的回归模型中，分享低质量新闻（ $b=0.24$ ）、机器人可能性（ $b=0.20$ ）和使用攻击性语言（ $b=0.17$ ）均与账户封禁显著正相关；而用户的政治倾向在控制这些因素后，与账户封禁的关联不再具有统计显著性（ $b=0.12$ ）。
- 倾向右翼的用户其账户被估计为机器人的可能性也显著更高（与政治倾向的相关系数 $r$ 介于0.70至0.76之间）。

## 2. Nature、Science论文逐篇总结

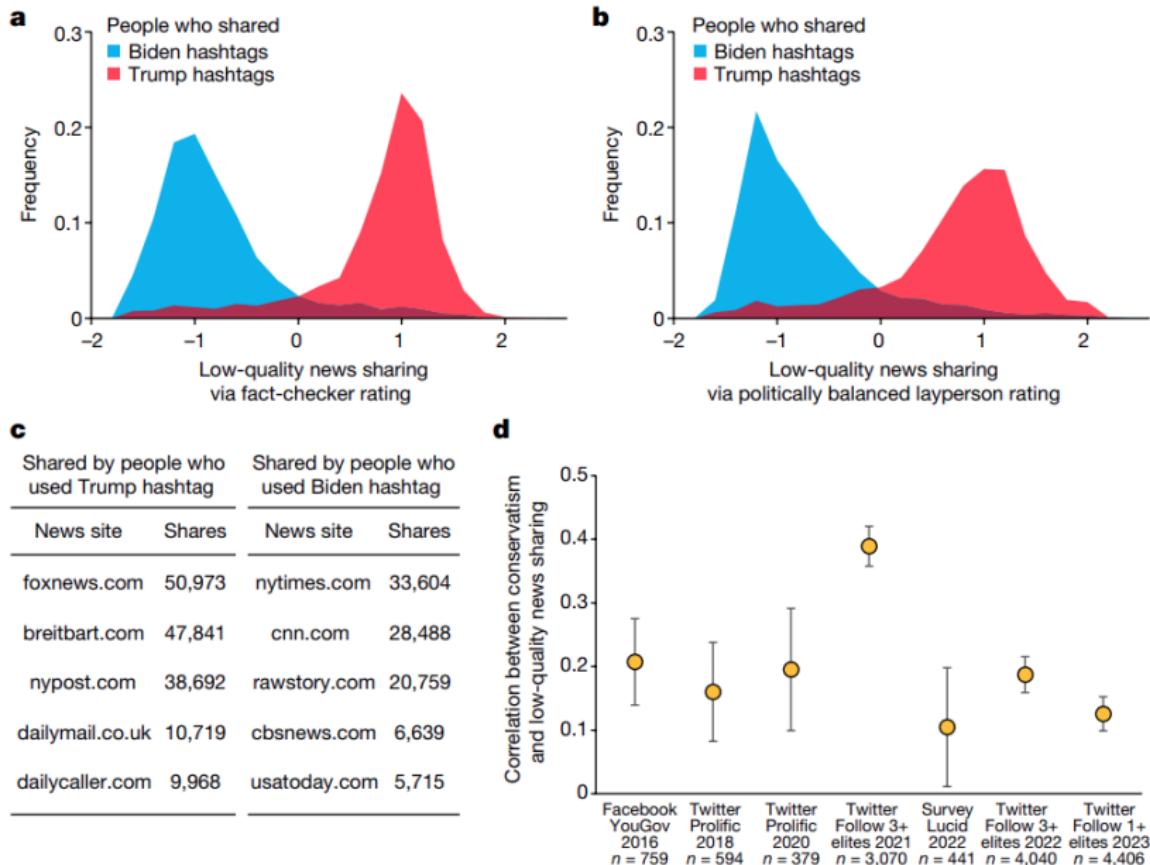


图 15: 支持特朗普和/或保守派的社交媒体用户比支持拜登和/或自由派的用户分享了更多来自低质量新闻网站的链接

## 研究结果—政策模拟分析结果

若仅因分享由政治平衡的普通民众判定的低质量新闻链接而封禁用户，则分享特朗普相关标签的用户被封禁的比例更高。例如，若每次分享低质量链接有1%的封禁概率，则特朗普标签用户的封禁人数是拜登标签用户的2.41倍。

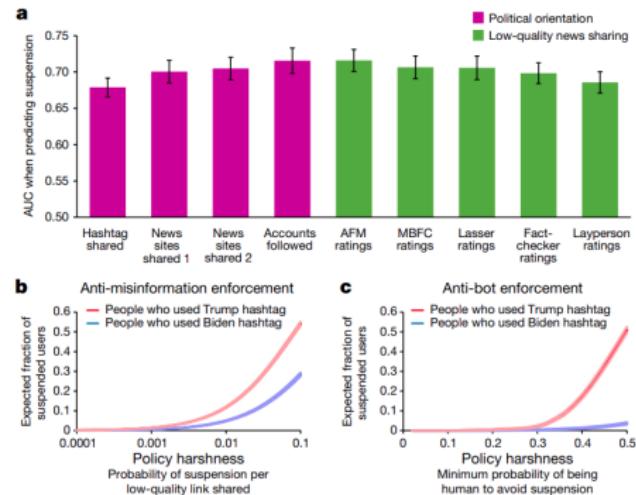


图 16: 政治倾向并非预测用户是否被封号的唯一指标，政治中立的执法政策会导致封号率出现政治不对称。

若仅因机器人账户评分过高（如高于0.5）而封禁用户，则分享特朗普相关标签的用户被封禁的人数是分享拜登标签用户的14.2倍。

## 研究结果—其他七个数据集的再分析结果

- 在所有七个额外数据集中（涵盖脸书2016年数据，推特2018年、2020年、2021年、2022年、2023年数据），均发现分享新闻的平均质量与保守主义倾向之间存在显著的负相关关系，此结论在使用事实核查员评分和政治平衡的普通民众评分时均一致。
- 对Ghezae等人数据的分析显示，在推特上，保守派用户比自由派用户分享了更多被事实核查员或政治平衡的普通民众评为不准确的URL链接。
- 对Arechar等人16国调查实验数据的分析显示，在美国以及所有16个国家汇总层面，保守主义倾向均与分享不实COVID-19信息的意愿呈显著正相关，此结论在使用事实核查员评分和普通民众评分时均成立。

总体而言，无论是以新闻来源域名层面还是具体帖子层面进行评估，也无论是基于事实核查员还是政治平衡的普通民众的判断，**保守派或共和党倾向的社交媒体用户分享更多低质量信息的模式是一致的**。

## 2. Nature、Science论文逐篇总结

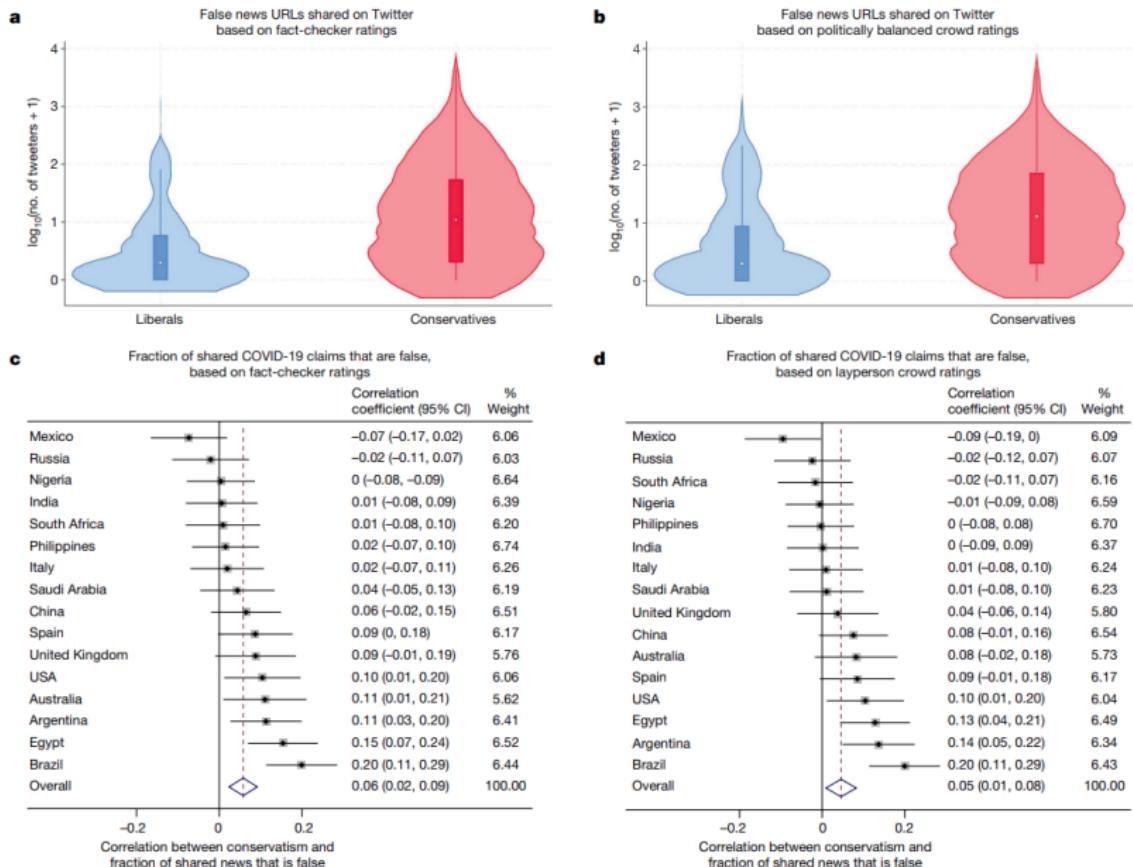


图 17: 保守派比自由派分享了更多虚假信息

## 研究结论

即便社交媒体公司严格执行政治中立的反误信息或反机器人政策，由于不同政治群体在误信息分享及其他违规行为上的差异，也必然带来对各党派用户的制裁不对称；这种不对称并不必然意味着平台自身存在政治偏见。

**启发：**识别传播Misinformation的用户的人口社会学特征。

# Post-January 6th deplatforming reduced the reach of misinformation on Twitter (Nature 2024)

## 1. 标题

1月6日事件后推特去平台化减少了虚假信息的传播范围

## 2. 作者及单位

Stefan D. McCabe, Diogo Ferrari, Jon Green, David M. J. Lazer, Kevin M. Esterling

- 乔治·华盛顿大学、加州大学河滨分校、杜克大学
- 东北大学网络科学研究所
- 哈佛大学定量社会科学研究所

## 3. 文献来源

McCabe, Stefan D., et al. “*Post-January 6th deplatforming reduced the reach of misinformation on Twitter.*” *Nature* 630.8015 (2024): 132-140.

## 4. 文献类型与关键词

文章类型：实证研究

英文关键词：deplatforming; misinformation; Twitter; natural experiment; regression discontinuity; difference-in-differences

中文关键词：去平台化；错误信息；Twitter；自然实验；回归不连续设计；双重差分

## 研究动机

社交媒体平台在当代公共话语监管中扮演核心角色，学术界对于平台级言论干预措施（尤其是大规模账户关停，即“去平台化”）的效果缺乏充分的实证研究。2021年1月6日暴力事件后，Twitter突然封禁约70,000个错误信息传播者，为评估此类干预的整体效应提供了珍贵的自然实验契机。

## 研究目标

本文旨在借助自然实验设计，系统量化Twitter于1月6日及随后的去平台化行动对平台错误信息流通的直接效应与溢出效应，并进一步考察该干预是否促使未被封禁的错误信息传播者自愿退出平台。

## 2021年1月6日暴力事件

2021年1月6日，美国首都华盛顿特区发生国会大厦暴力事件。当天，大批抗议者在时任总统特朗普的煽动下，闯入美国国会大厦，试图阻止国会对2020年总统选举结果的认证。事件导致多人伤亡，并对美国民主制度造成严重冲击，被广泛认为是美国历史上前所未有的政治暴力事件。

## 研究问题

- 针对被去平台化的用户，该干预措施如何影响其分享虚假信息的行为？
- 对于那些曾关注被去平台化用户的普通用户（即“关注者”），其分享虚假信息的行为是否因该干预措施而改变？
- 该去平台化干预措施是否对那些传播虚假信息但未被直接去平台化的用户产生了溢出效应（例如，是否导致他们离开平台或改变行为）？

## 研究方法

本研究采用自然实验设计（natural experimental designs），结合追踪研究（panel study）的方法。

**数据收集：**研究者构建了一个包含超过50万名活跃推特用户的大型追踪数据集，收集了这些用户在2020年9月至2021年2月期间分享包含URL链接的推文数据。研究者还整合了已知传播虚假信息和低质量新闻的网站域名列表。

**干预措施：**研究的核心干预是推特在2021年1月8日至1月12日期间，针对约7万个账户（包括许多知名的虚假信息传播者）实施的永久性关停。

### 实验组与对照组构建：

- 针对被去平台化的用户（处理组），通过倾向得分匹配（propensity score matching）的方法，从行为相似但未被去平台化的虚假信息传播者中选取对照组。
- 针对被去平台化用户的关注者（处理组），同样通过匹配方法，选取了关注相似但未被去平台化用户的关注者作为对照组。

**效应评估：**采用夏普断点回归设计 (Sharp Regression Discontinuity, SRD) 和双重差分法 (difference-in-differences, DID) 分析，比较处理组与对照组在干预前后分享虚假信息链接数量的变化，以估计去平台化干预的因果效应。同时，研究也考察了新出现的虚假信息域名的传播情况。

# 研究结果

**整体虚假信息水平变化：**关注2016年及2020年选举周期，早于主要干预事件，2020年推特上所有用户分享虚假信息URL的比例是2016年的五倍。在2020年11月选举后至2021年1月6日前，虚假信息推文和转推的比例保持在较高水平，但在1月6日之后，该比例下降了约50%。

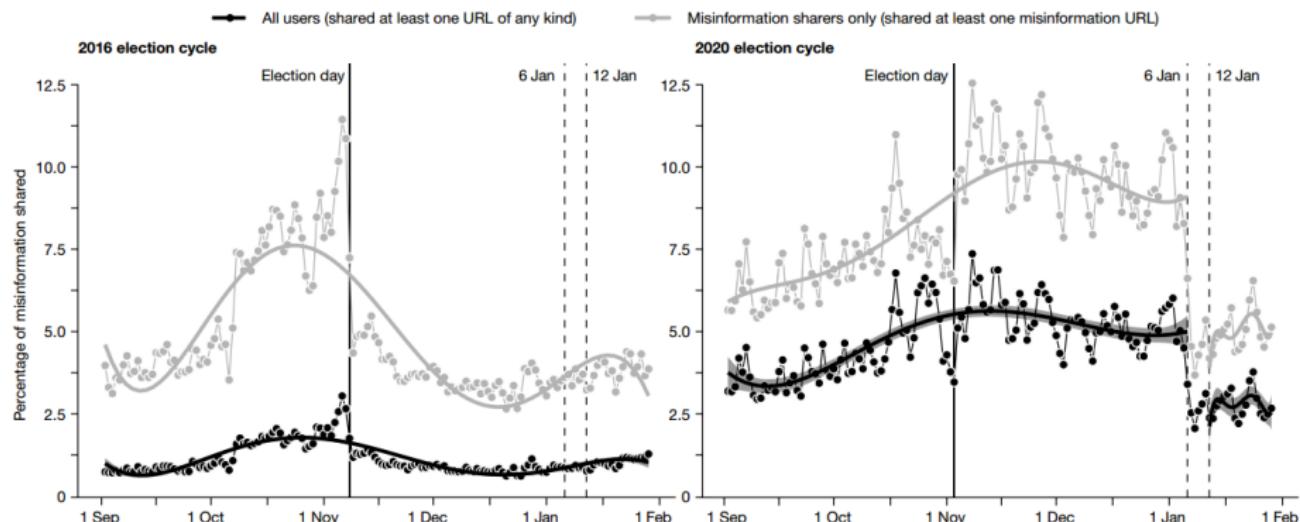


图 18: 2016年和2020年美国选举周期中推特上的虚假信息分享情况

**对被去平台化用户的影响 (SRD分析):** 去平台化干预显著减少了被关停用户自身分享虚假信息的行为（包括推文和转推）。

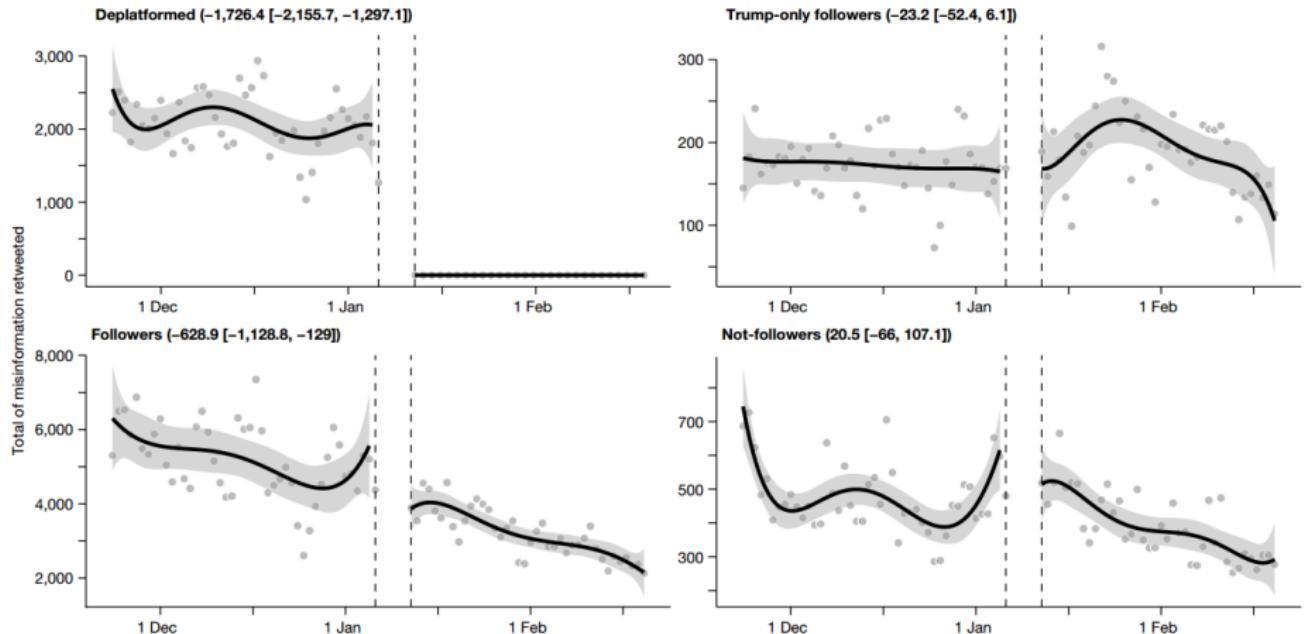


图 19: 1月6日后去平台化干预下，Twitter错误信息用户的错误信息转发量减少

**对关注者的溢出效应 (DID分析与SRD观察):** SRD分析初步显示，被去平台化用户的关注者中虚假信息转推有所减少，而非关注者群体中未观察到此现象。在平行路径假设下，DID分析进一步表明，与非关注者相比，被去平台化用户的关注者在干预后显著减少了虚假信息的转推行为。

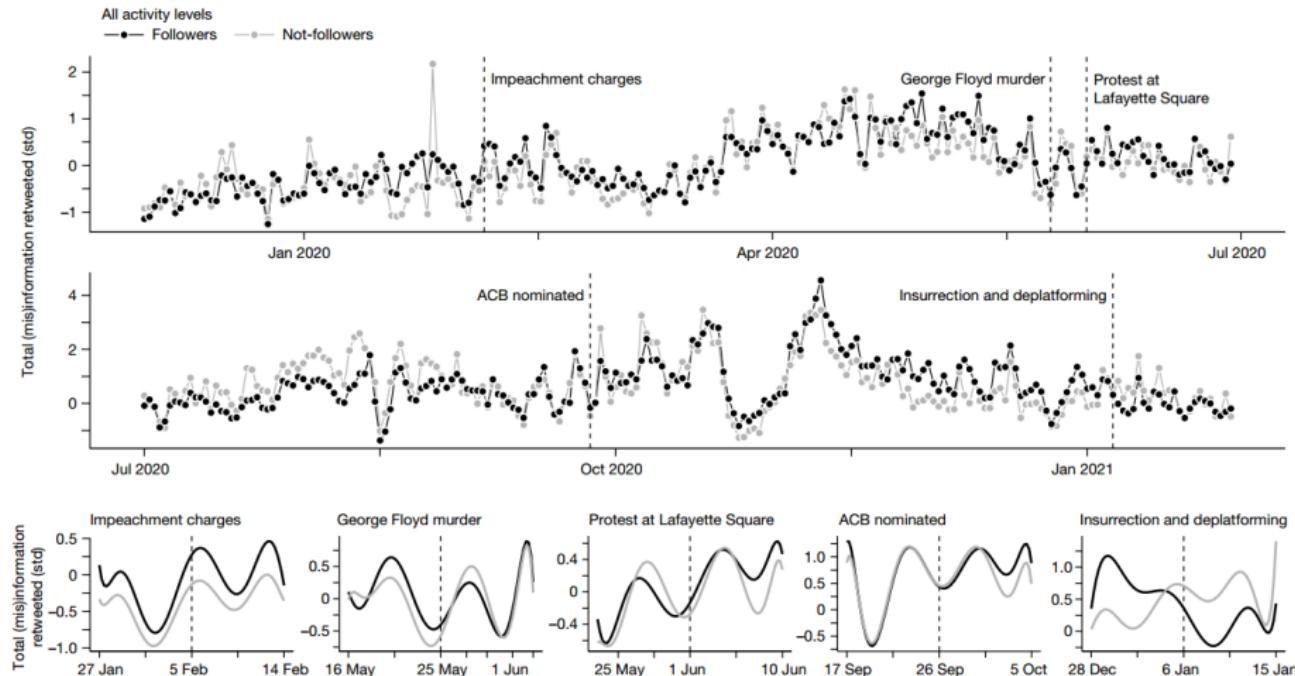


图 20: 各活动水平下, 被去平台用户的关注者与非关注者转推虚假信息的时间序列

## 2. Nature、Science论文逐篇总结

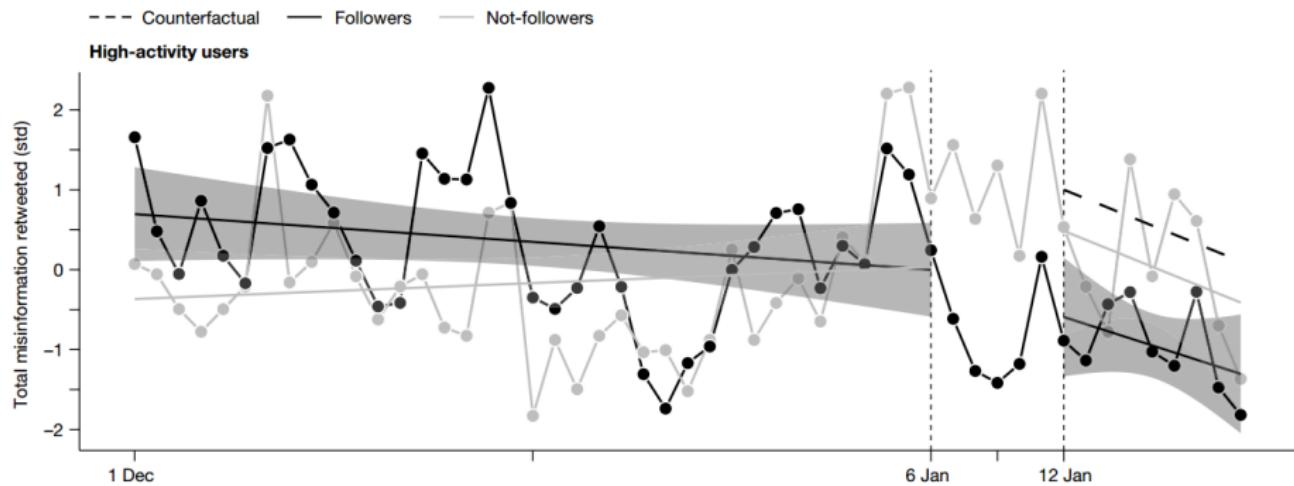


图 21: 关注者与非关注者转推虚假信息的时间序列

## 2. Nature、Science论文逐篇总结

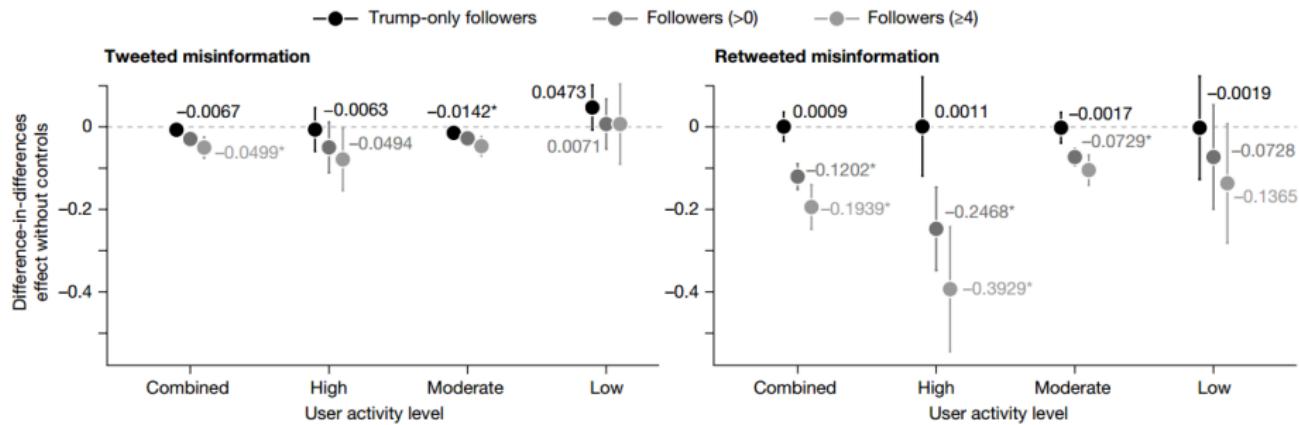


图 22: 去平台化对被去平台推特用户关注者影响的双重差分估计

**对未被去平台化的虚假信息传播者的影响(用户流失):** 在推特加强其管理姿态后,许多未被直接去平台化的高产虚假信息传播者(“超级传播者”)以及QAnon内容传播者,也显著降低了在平台上的活跃度,甚至选择退出平台。然而,普通的关注者群体其活跃度则与干预前保持相似水平。

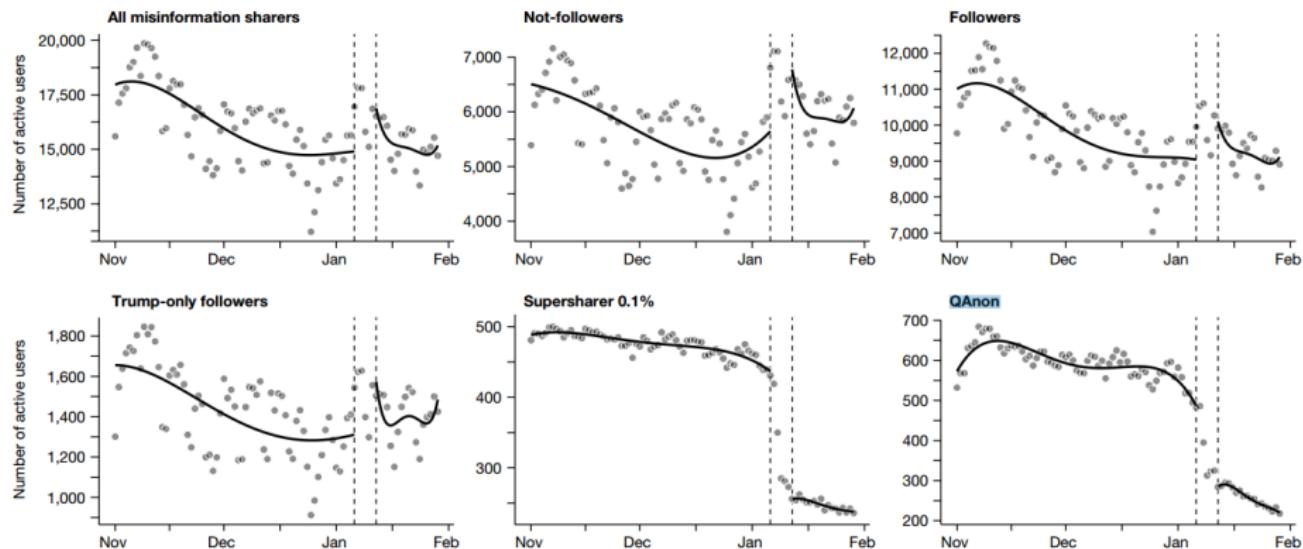


图 23: 各亚组中未被去平台化用户数量的时间序列

## 研究结论

- 推特在2021年1月6日事件后实施的大规模去平台化措施，很可能对平台整体虚假信息的传播量造成了实质性的削减。该干预的影响不仅限于被直接移除的用户，还通过溢出效应，减少了其关注者对虚假信息的再传播行为。
- 此外，平台的强硬管理姿态也促使一部分未被直接处理的、高度参与虚假信息传播的用户（如“超级传播者”和QAnon内容传播者）选择退出平台或降低活跃度，形成了用户构成上的变化，这可能也对减少有害内容有所贡献。
- 这些发现表明，社交媒体平台确实拥有通过执行其使用条款（如去平台化）来一定程度上控制其平台内虚假信息流通的能力。
- 尽管由于干预措施与重大外部事件并发，精确量化纯粹的因果效应面临挑战，但多方面的证据共同指向，在此特定情境下，去平台化作为一种治理手段，在减少虚假信息传播方面展现了其有效性。

**启发：**LLM-based Agent模拟去平台化；挖掘与社交媒体相关的重要事件提供的自然实验契机，例如，网信办出台的一些法规或者大量关停某些类别账号，不一定是misinformation账号，也可能是其他有违反公序良俗行为的账号；

# Fake news on Twitter during the 2016 U.S. presidential election (Science 2019)

## 1. 标题

2016年美国总统选举期间推特上的虚假新闻

## 2. 作者及单位

Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, David Lazer

- 美国东北大学网络科学研究所

- 哈佛大学定量社会科学研究所

- 纽约州立大学布法罗分校计算机科学与工程系

## 3. 文献来源

Grinberg, Nir, et al. “*Fake news on Twitter during the 2016 US presidential election.*” Science 363.6425 (2019): 374-378.

## 4. 文献类型与关键词

文章类型：实证研究

英文关键词：Fake news; Twitter; social media; 2016 U.S. presidential election;  
misinformation; political news consumption

中文关键词：虚假新闻；推特；社交媒体；2016年美国总统选举；错误信息；政治新闻  
消费

## 研究动机

尽管早期报告显示，2016年美国总统大选末期最流行的假新闻在Facebook上的分享量、互动量均超过同步段的真实新闻，但关于Twitter平台上假新闻的实际曝光与分享规模尚无系统性量化；同时，由于社交媒体数据获取与分析难度较大，现有的调查与离线浏览记录研究难以反映普通用户在社交网络中的真实行为。因此，有必要借助社交媒体数据，直接测量Twitter用户对假新闻的接触与传播情况，以及假新闻在整体政治新闻生态中的位置。

## 研究目标

利用一个与公开选民登记记录相关联的推特账户样本，深入研究在2016年美国总统选举期间，美国选民在推特上与虚假新闻来源互动的具体情况。具体而言，研究旨在量化个体接触和分享虚假新闻的程度，识别参与此类互动的用户特征，并探究这些用户如何与更广泛的政治新闻生态系统进行互动。

## 研究问题

- 个体在社交媒体上看到和分享了多少来自虚假新闻来源的故事？
- 那些与这些虚假新闻来源互动的人群具有哪些特征？
- 这些个体如何与更广泛的政治新闻生态系统进行互动？

# 研究方法

- **数据收集与样本：**构建了一个包含16442个推特账户的追踪样本，这些账户与美国选民的注册记录相关联，并在2016年选举季（8月1日至12月6日）期间保持活跃。该样本在年龄、性别、种族和政治立场方面与皮尤研究中心获得的推特注册选民代表性样本基本一致。研究收集了这些账户发布的推文及其关注和被关注列表。
- **虚假新闻来源的定义与分类：**将虚假新闻出版机构定义为那些具备合法新闻生产的表象，但缺乏确保信息准确性和可信度的编辑规范与流程的机构。研究将虚假新闻来源分为三类：“黑色”名单来源于事实核查员、记者和学者构建的、几乎专门发布捏造故事的网站列表；“红色”名单指那些传播明显反映编辑过程存在缺陷的虚假信息的网站；“橙色”名单则代表注释者不太确定其虚假信息源于系统性编辑过程缺陷的网站。
- **暴露量与分享量分析：**研究通过对样本用户关注对象所发布推文的随机抽样来估计每个样本成员新闻流的构成，这些潜在接触到的推文被称为“暴露量”。研究分析了包含外部网页链接的政治性推文的聚合暴露量和分享量。

# 研究方法

- **用户特征与政治倾向评估：**研究通过比较个体信息流与已注册民主党或共和党人信息流的相似性来估计其政治倾向，并将个体划分为五个政治亲和度亚组：极左 (L\*)、左 (L)、中间 (C)、右 (R) 和极右 (R\*)。对于政治URL暴露量少于100的个体，则归入“非政治”亚组。
- **统计建模与网络分析：**在排除极端值（即“超级分享者”和“超级消费者”，约占总样本1%）后，研究采用二项回归和逻辑回归模型分析影响虚假新闻暴露和分享的因素。此外，研究还构建了一个新闻网站的共同暴露网络，以分析虚假新闻来源在媒体生态系统中的位置。

# 研究结果

**虚假新闻的流行度与高度集中性：**在所有政治相关URL的聚合暴露量中，5.0%来自虚假新闻来源；在分享的政治相关URL中，6.7%来自虚假新闻来源。然而，对虚假新闻来源的参与表现出极端的集中性：仅1%的个体贡献了80%的虚假新闻来源暴露量，而0.1%的个体贡献了近80%的虚假新闻来源分享量。在各类虚假新闻中，5%的来源占据了超过50%的暴露量。这些“超级分享者”和“超级消费者”在虚假新闻的传播中作用显著，他们通常表现出极高的发帖活跃度，部分账户可能为半自动化的“赛博格”账户。

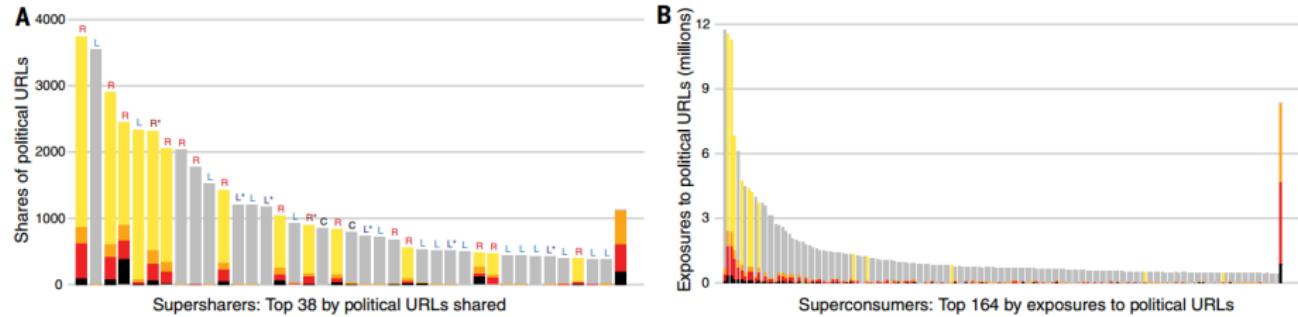


图 24: 政治URL的总体超级分享者与超级消费人群

## 2. Nature、Science论文逐篇总结

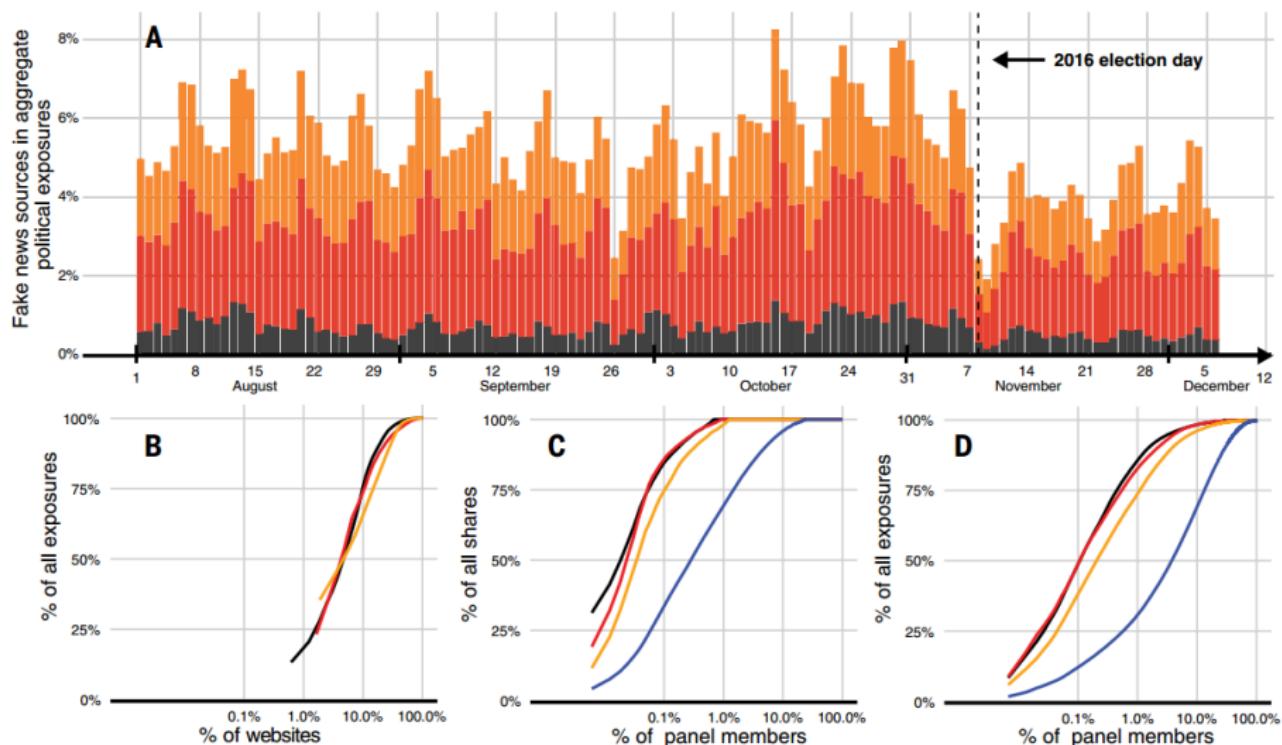


图 25: 假新闻来源随时间的流行度及集中度

**接触虚假新闻的用户特征(剔除极端值后):** 在选举最后一个月, 普通样本成员平均有204次潜在暴露于虚假新闻来源的机会, 若按5%的实际阅读率估算, 则相当于约10次实际暴露。个体信息流中, 虚假新闻来源的平均占比为1.18%。政治倾向上, 保守派人士接触虚假新闻的比例显著更高(右翼和极右翼群体中16.3%的个体其政治新闻暴露的5%以上来自虚假新闻, 而左翼和极左翼群体中此比例仅为2.5%)。个体信息流中虚假新闻的比例, 最强的预测因素是年龄(年龄较大者比例更高)和个体信息流中政治URL的总量(政治新闻关注度越高, 虚假新闻比例也越高)。男性和白人用户接触虚假新闻的比例也略高。

**分享虚假新闻的用户特征(剔除极端值后):** 政治倾向与分享虚假新闻内容显著相关: 左翼或中间派用户中分享过虚假新闻的比例低于5%, 而右翼和极右翼用户中则分别有11%和21%。分享虚假新闻的行为与发布政治推文的频率、接触虚假新闻的程度以及政治倾向呈正相关。年龄和较低的关注者/被关注者比例也与分享虚假新闻呈正相关, 但影响较小。在分享决策中, 内容与自身政治立场的一致性是主导因素; 在控制了对政治立场一致来源的暴露后, 虚假新闻并不一定比真实新闻更具病毒性(在立场一致的前提下, 虚假新闻并不天然比真实新闻更容易被分享)。

## 2. Nature、Science论文逐篇总结

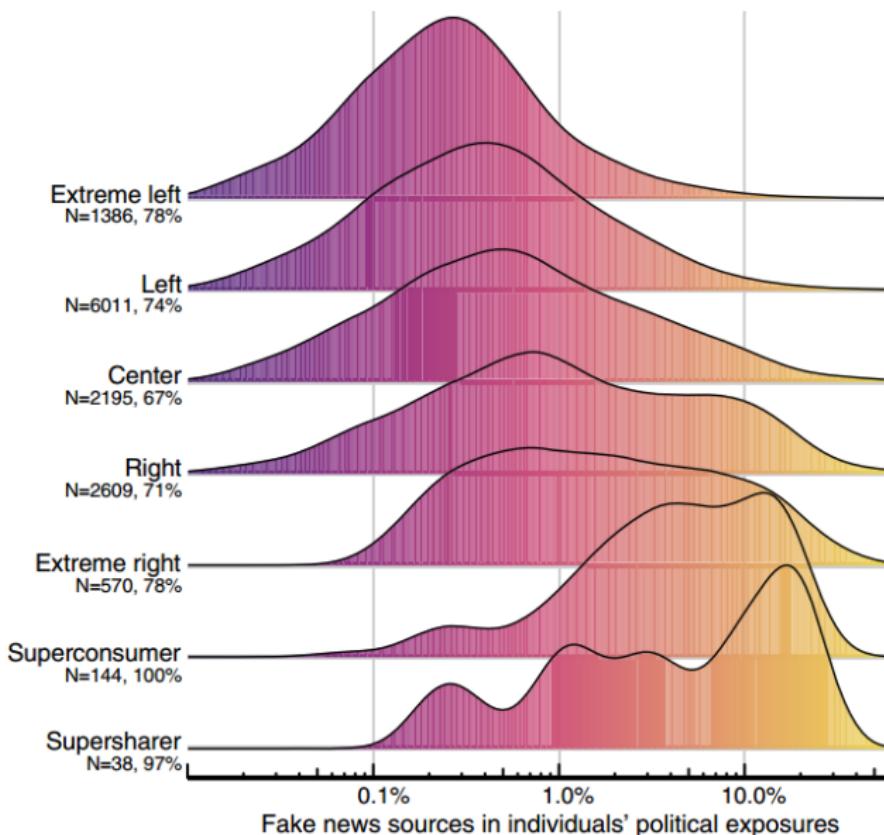


图 26: 个体信息流中假新闻来源内容比例的概率密度估计

## 2. Nature、Science论文逐篇总结

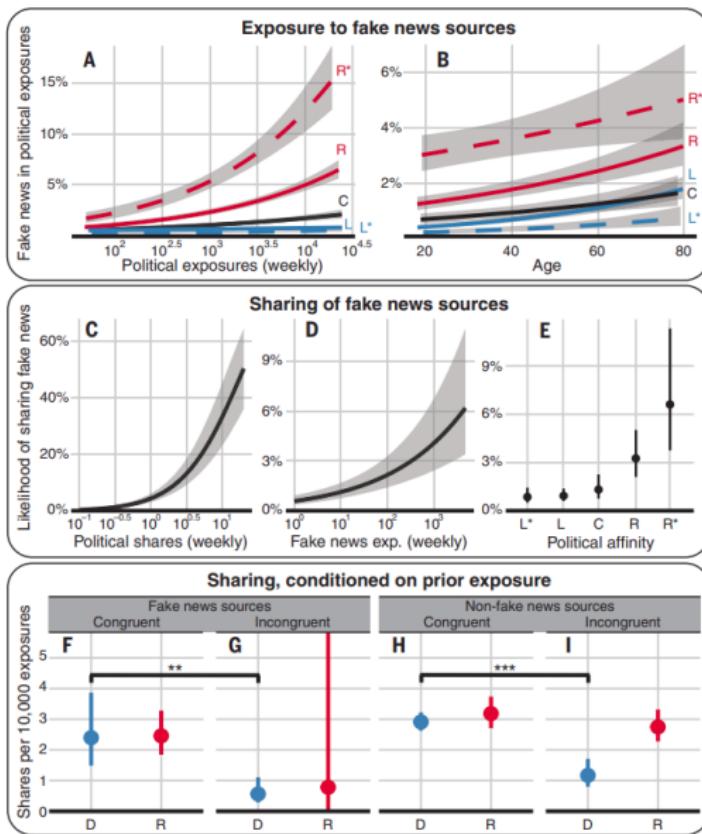
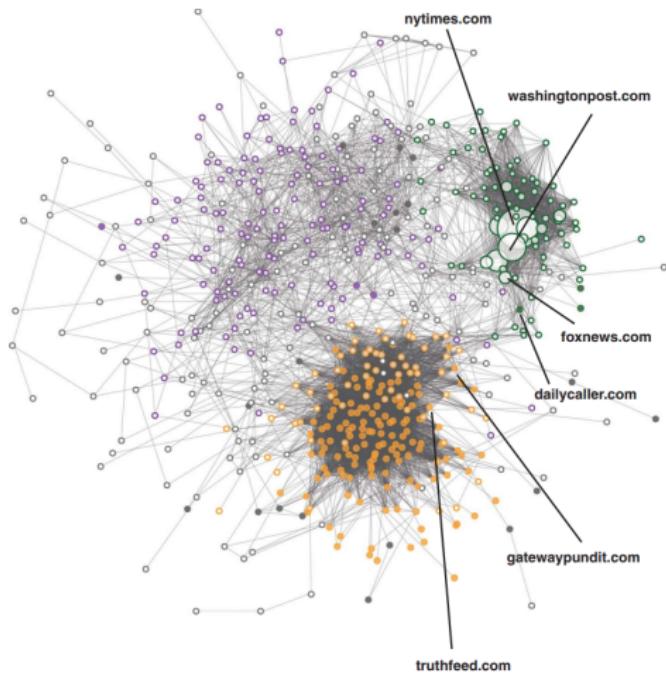


图 27: 与假新闻来源曝光与分享相关的关键个体特征

**虚假新闻在媒体生态系统中的位置：**主流媒体（网络分析中的第1组）仍然是所有政治倾向群体获取政治URL信息的主要来源，占比从极右翼的72%到极左翼的86%不等。研究识别出一个由大量虚假新闻来源构成的独特集群（第2组，其中68.8%为虚假新闻来源），该集群的政治倾向显著偏保守，其受众高度重叠，且主要集中在右翼群体。对第2组内容的接触程度因政治倾向而异，极右翼群体接触最多。虚假新闻的消费者通常会接触到多个不同的虚假新闻来源。

## 2. Nature、Science论文逐篇总结



每个节点代表一个政治新闻、博客或事实核查网站。边连接的是那些有异常多数量（非异常值）小组成员同时暴露于其内容的网站对，同时控制了每个网站的受欢迎程度。实心节点代表虚假新闻来源。节点颜色表示通过聚类算法组合识别出的组别（1, 绿色; 2, 橙色; 3, 紫色; 4, 灰色）。暴露量最高的网站节点略大。

图 28: 新闻网站共曝光网络

# 研究结论

- 尽管有6%分享政治内容URL的用户分享了来自虚假新闻来源的内容，但绝大多数虚假新闻的分享和暴露都高度集中于极小部分人群。
- 最倾向于接触和分享虚假新闻来源的个体通常具有保守的政治倾向、年龄较大，并且高度关注政治新闻。
- 对于所有政治光谱上的人群而言，绝大部分的政治新闻暴露仍然来自主流媒体渠道。
- 对于普通追踪样本成员而言，来自虚假新闻来源的内容仅占其政治新闻暴露量的1.18%，相当于在选举最后一个月接触大约10个相关URL。
- 虚假新闻来源似乎是一种小众兴趣，它们形成了一个独特的、包含大量虚假信息源的网站集群，其受众高度重叠，且主要为右翼人士。
- 研究结果提示，减少错误信息传播的干预措施可以聚焦于少数“超级传播者”和“超级消费者”，以及少数几个最具渗透性的虚假新闻来源。

**启发：**以往观察到的虚假新闻似乎传播更广的现象<sup>2</sup>，可能并非仅仅因为其“虚假”的本质，而可能与这些虚假新闻更倾向于迎合特定群体的政治偏好、从而在这些群体中更容易因立场一致而被分享有关。

<sup>2</sup>Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. science, 359(6380), 1146-1151.

# Misinformation exploits outrage to spread online (Science 2024)

## 1. 标题

虚假信息利用愤怒情绪在网络上传播

## 2. 作者及单位

Killian L. McLoughlin; William J. Brady; Aden Goolsbee; Ben Kaiser; Kate Klonick;  
M. J. Crockett

- 普林斯顿大学、西北大学
- 耶鲁大学、圣约翰大学
- 布鲁金斯学会、哈佛大学

## 3. 文献来源

McLoughlin, Killian L., et al. "*Misinformation exploits outrage to spread online.*"  
Science 386.6725 (2024): 991-996.

## 4. 文献类型与关键词

文章类型：实证研究

英文关键词：misinformation; moral outrage; social media engagement; differential  
privacy; source classification

中文关键词：误导信息；道德愤怒；社交媒体参与；差分隐私；来源分类

## 研究动机

在线分享的误导信息（misinformation）—指虚假或具有误导性的内容—尽管平台投入大量资源检测和削减其传播，但仍广泛存在，并已被证实与政治极化、反民主情绪和疫苗犹豫等负面社会后果密切相关。然而，**现有干预措施（如事实核查、准确性提示）收效有限，尚未充分考虑情绪在信息传播中的作用。**道德愤怒（moral outrage）因其强烈的参与性、无需依赖信息准确性即可服务于传播者的政治或群体认同诉求，**以及可能提升内容可信度，极有可能被误导信息利用以促进其传播。**因此，系统考察愤怒情绪与误导信息扩散之间的关系，具有重要的理论与实践价值。

## 研究目标

本研究旨在**检验“误导信息是否通过激发愤怒而更易在社交媒体上传播”的假设**，并评估该机制在不同平台（Facebook 与 Twitter）、不同时期（2017 与 2020-2021）及不同误导信息分类标准下的普适性，从而为制定更有效的应对策略提供实证依据。

## 研究问题

- 与可信新闻相比，虚假信息是否更倾向于引发愤怒情绪？
- 愤怒情绪是否会促进虚假信息的传播？其促进作用与对可信新闻传播的促进作用相比如何？
- 愤怒情绪如何塑造分享虚假信息的心理动机？具体而言，愤怒情绪是增强了分享准确信息的认知动机，还是增强了与准确性无关的非认知动机？

# 研究方法

本研究综合运用了观察性研究和行为实验的方法，使用美国用户数据。

**观察性研究：**分析了来自Facebook（1,063,298个链接）和Twitter（44,529条推文，24,007名用户）的数据。研究涵盖了多个时间段（2017年，2020-2021年）和多种虚假信息分类策略。

- **数据来源与分类：**链接根据其来源网站的质量进行分类，这些网站由专业组织评估。研究使用了三个不同的新闻领域数据库对虚假信息和可信新闻来源进行分类。部分研究还包含了由互联网研究机构（IRA）发布的链接。为了控制用户网络特征的混淆，研究还采用了“受众匹配策略”，即分析同一用户分享的虚假信息链接和可信来源链接。

- **愤怒情绪量化：**在Facebook上，愤怒情绪通过“愤怒反应”的数量来量化；在Twitter上，则通过其数字愤怒分类器（DOC）分析回复中包含道德愤怒表达的比例来量化。

**行为实验：**招募了1475名参与者（通过Prolific平台）进行了两项行为实验。

- 实验材料：使用了经过事实核查机构确认为真实或虚假的，并预先测试为引发高或低愤怒情绪的新闻标题。
- 实验任务：参与者被展示一系列新闻标题，并被要求评估他们分享该标题的可能性（实验5a）或标题所述声明的准确性（实验5b）。

**数据分析：**

- 对于Facebook数据，由于涉及差分隐私保护，研究采用了特定的回归模型，通过模拟添加噪声并对数万个模型进行平均估计，得出调整后的系数估计值和置信区间。
- 对于Twitter数据和行为实验数据，使用了标准的统计回归模型，如逻辑回归和负二项回归。

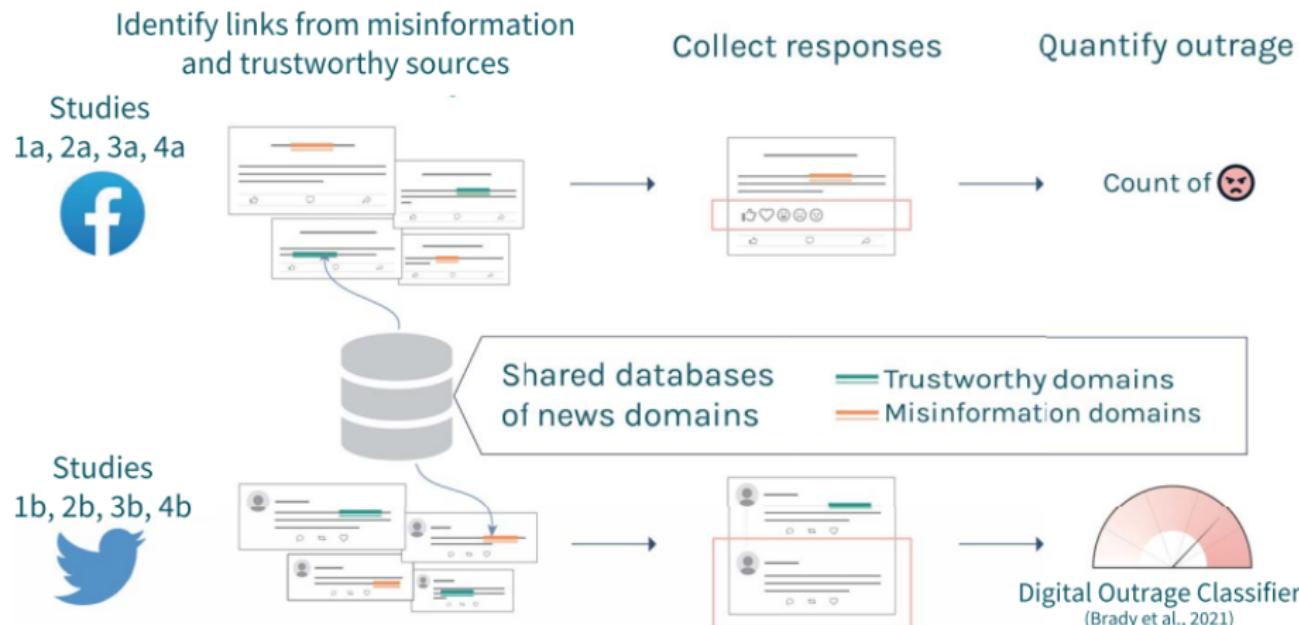
## 2. Nature、Science论文逐篇总结

Study	Data source	Time period	News source	N <sub>links/tweets</sub>	N <sub>users/participants</sub>
1a	*Facebook			9026 links	—
1b	Twitter	January 2017 to July 2017	IRA articles from domains in Domain Dataset 1	3329 tweets	1656 users
2a	*Facebook			192,108 links	—
2b	Twitter	August 2020 to February 2021	IRA domains in Domain Dataset 1	10,550 tweets	5236 users
3a	*Facebook			211,535 links	—
3b	Twitter	August 2020 to February 2021	Domain Dataset 2	16,617 tweets	7485 users
4a	*Facebook			650,629 links	—
4b	Twitter	August 2020 to February 2021	Domain Dataset 3	14,033 tweets	5848 users
5a	Prolific			—	730 participants
5b	Prolific	January 2020 to December 2021	Snopes.com	—	745 participants

\*The URL Shares dataset does not provide data at the individual user level, and so the number of users is not available.

我们构建了来自Facebook（研究1a、2a、3a和4a）和Twitter（研究1b、2b、3b和4b）的平行数据集，这些数据涵盖了2017年（研究1a和1b）以及2020至2021年（研究2a、2b、3a、3b、4a和4b）的时期。我们还进行了两项行为学研究（研究5a和5b）。在我们的观察性研究（研究1至4）中，我们使用三个独立的、根据来源质量评估的新闻领域数据库对虚假信息进行分类。在我们的行为学研究（研究5a和5b）中，我们使用了经过事实核查判定为真实或虚假的新闻标题。

图 29: 研究概览



我们使用对新闻质量进行过评估的母域名数据库，识别了指向虚假信息源和可信信息源的链接（补充材料与方法2.1）。我们利用这些数据库构建了成对的数据集，其中包含在2016年和2020年相同时段内，链接到相同文章或母域名的Facebook和Twitter帖子。随后，我们收集了每个数据集中针对这些链接的情感反应。在Facebook上，我们通过链接收到的“愤怒”（Angry）反应数量来量化愤怒情绪；在Twitter上，则通过我们的数字愤怒分类器（Digital Outrage Classifier, DOC）判定回复中包含道德愤怒表达的比例来进行量化（补充材料与方法5.1.2）(26)。我们使用“Twitter”而非“X”，是因为我们的数据收集于该平台更名之前。

图 30: 数据集构建流程

# 研究结果

## 虚假信息源比可信新闻源更能引发愤怒情绪

在Facebook上，来自虚假信息源的链接获得的“愤怒反应”显著多于来自可信来源的链接。即使在控制了受众规模后，这种关联依然存在。与爱、哈哈、哇等其他情绪反应相比，虚假信息源的链接更可能引发愤怒反应。

在Twitter上，对虚假信息源链接的回复也显著更可能包含愤怒情绪的表达。与一般负面情绪相比，虚假信息源的链接更可能引发愤怒情绪。

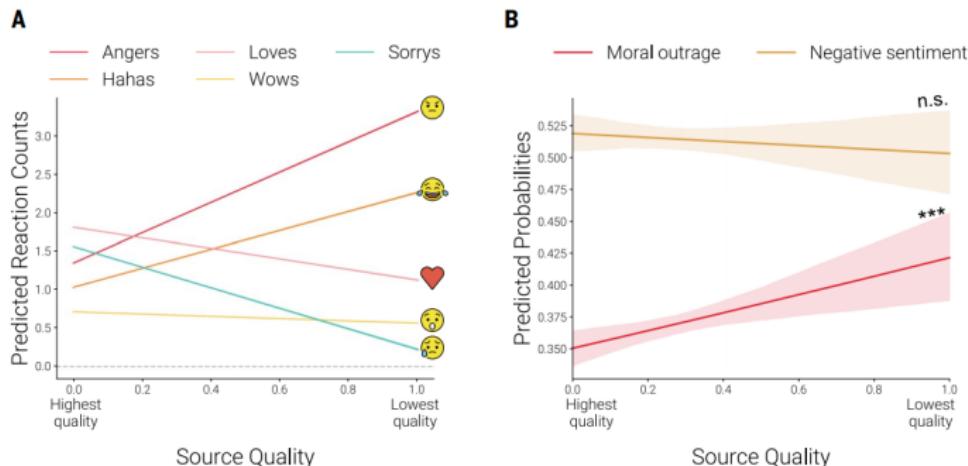


图 31: 指向低质量新闻源的链接会引发更多愤怒情绪

愤怒情绪促进虚假信息的传播，其强度至少与促进可信新闻传播相当

在Facebook上，“愤怒反应”的数量与可信来源和虚假信息来源链接的分享量均呈正相关，且对虚假信息来源链接分享的促进作用在所有研究中均大于或等于对可信来源链接的促进作用。

在Twitter上，回复中出现愤怒情绪显著增加了原始推文的分享量，无论其来源是可信的还是虚假信息。愤怒情绪对分享的影响在虚假信息和可信新闻之间，其交互作用在不同研究中不完全一致。

行为实验（研究5a）结果表明，与低愤怒标题相比，参与者更倾向于分享高愤怒标题，无论标题是真实的还是虚假的，且愤怒情绪与新闻类型之间没有显著的交互作用。

## 2. Nature、Science论文逐篇总结

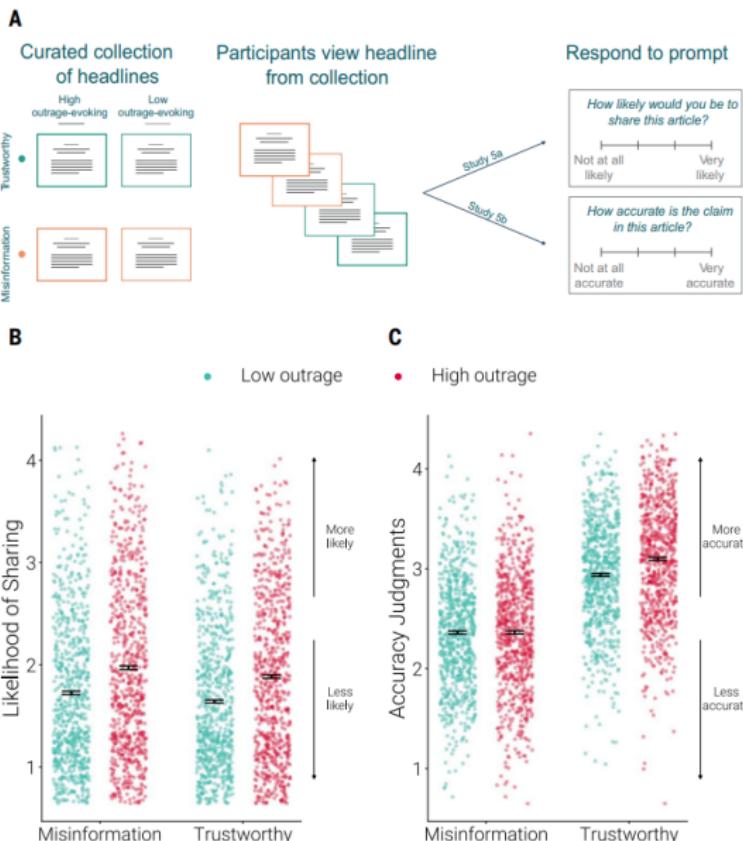
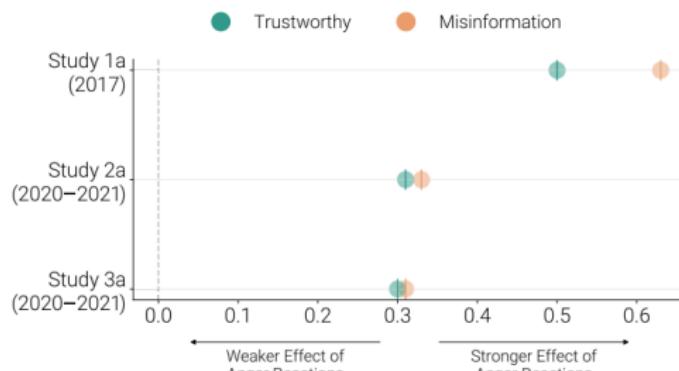


图 32: 更高的愤怒情绪增加了分享行为，但不影响准确性判断

## 愤怒情绪增强了分享虚假信息的非认知动机

Facebook数据显示，愤怒反应与“未经阅读即分享”的行为呈正相关，无论是对虚假信息还是可信来源的链接；并且，愤怒反应用于虚假信息“未经阅读即分享”的预测作用强于对可信来源的预测作用。这表明愤怒情绪增加了分享行为中非认知动机（相对于认知动机）的相对强度。其他情绪反应也有类似效果。

行为实验（研究5b）结果显示，尽管参与者能够区分真实新闻和虚假新闻的准确性（即真实新闻被评为更准确），但愤怒情绪的唤起程度并不显著影响他们对新闻准确性的判断。这表明愤怒情绪并未增强评估信息准确性的认知动机。



$\tilde{\beta}$  Estimates of the Effect of Angry Reactions  
on Sharing-without-Reading

图 33: 愤怒反应预示误导信息比可信来源更易在未阅读情况下被分享

# Quantifying the impact of misinformation and vaccine-skeptical content on Facebook (Science 2024)

## 1. 标题

量化脸书上错误信息和疫苗怀疑内容的影响

## 2. 作者及单位

Jennifer Allen; Duncan J. Watts; David G. Rand

- 麻省理工学院斯隆管理学院，数据、系统与社会研究所，脑与认知科学系

- 宾夕法尼亚大学计算机与信息科学系，安纳伯格传播学院，运营、信息与决策系

## 3. 文献来源

Allen, Jennifer, Duncan J. Watts, and David G. Rand. “*Quantifying the impact of misinformation and vaccine-skeptical content on Facebook.*” Science 384.6699 (2024): eadk3451.

## 4. 文献类型与关键词

文章类型：实证研究、方法学研究

英文关键词：Misinformation; Vaccine-skeptical content; COVID-19; Impact quantification; Persuasive effect; Exposure

中文关键词：错误信息；疫苗怀疑内容；新冠肺炎；影响量化；说服效应；暴露

## 研究动机

疫苗接种率低一直被归因于社交媒体上的误导信息（misinformation），但现有研究尚不清楚这类内容是否具备足够的“广泛曝光”及“因果影响”来真正改变公众行为；此外，虽然并非事实错误但质疑疫苗安全性的“vaccine-skeptical”内容同样可能促进拒种，却鲜有系统量化分析。

## 研究目标

本研究提出“影响力 = 曝光量 × 说服效应”框架，旨在定量评估COVID-19疫苗相关的误导信息与“vaccine-skeptical”内容在美国Facebook用户中的实际说服影响，并比较两者对接种意愿的相对贡献。

## 研究问题

- 在控制暴露的条件下，何种类型的疫苗相关内容（包括被事实核查的错误信息）会改变人们接种新冠疫苗的意愿？更具体地说，是内容的真实性还是其他维度（如暗示疫苗对健康有害的程度）更能预测其负面说服力？
- 在脸书平台上，被事实核查机构标记的错误信息与那些未被标记但可能暗示疫苗有害（即疫苗怀疑论内容，尤其是来自主流新闻媒体的内容）的帖文相比，其真实的浏览量（暴露程度）有何差异？
- 如何有效地将实验室实验中测得的少量内容的因果效应，推广到脸书上数以万计的疫苗相关URL上，以估计每个URL的说服效应？
- 综合考虑说服效应和暴露程度，被标记的错误信息与未被标记的疫苗怀疑论内容对美国脸书用户整体疫苗犹豫的相对影响孰大孰小？

# 研究方法

本研究采用了一个**多阶段、混合方法**的研究设计：

- **实验估计说服效应:**研究人员进行了两项大规模在线调查实验(总样本N=18,725),参与者来自Lucid调查平台。实验中,参与者被随机分配观看中性对照帖子或130个疫苗相关标题中的一个(包括40个先前被事实核查机构揭穿的错误信息和90个从脸书上抽取的代表性高分享疫苗相关文章)。通过比较暴露前后参与者报告的疫苗接种意愿(构建成一个新冠疫苗接种指数),来衡量每个标题的因果说服效应。研究还收集了关于这些标题的多个维度特征(如是否令人惊讶、是否可信、是否暗示疫苗有害健康等)的众包评分。
- **测量脸书暴露数据:**研究使用了脸书的Social Science One数据集,分析了2021年第一季度(疫苗推广初期)在脸书上公开分享超过100次的13206个与新冠疫苗相关的URL的实际浏览量。这些URL根据是否被专业事实核查机构标记为虚假、断章取义或混合信息等进行了分类。

- **构建预测模型并推广估计：**为了将实验中130个标题的因果效应推广到脸书上的13206个URL，研究人员开发了一个结合众包智慧和自然语言处理（NLP）的流程。首先，他们招募众包人员对实验中的130个标题进行评分，预测其对疫苗接种意愿的影响，并结合其他众包评分（如暗示疫苗有害程度、准确性）构建了一个“众包综合评分”，发现该评分能有效预测实验观察到的真实处理效应。接着，他们让众包人员对脸书数据集中1139个URL进行同样的评分，并以此为训练数据，训练了一个COVID-Twitter-BERT机器学习模型，用以预测全部13206个URL的众包综合评分。最后，将这些预测的众包综合评分代入之前建立的元回归模型，从而为每个脸书URL生成一个估计的处理效应（即暴露后的说服效应）。
- **量化总体影响：**通过将每个URL的估计处理效应与其在脸书上的浏览量相乘，并根据美国脸书用户总数进行归一化，研究人员量化了被标记的错误信息和未被标记的疫苗怀疑论内容对整体疫苗接种意愿的预测影响。

# 研究结果

**说服效应：**实验表明，暴露于被事实核查的错误信息确实会导致疫苗犹豫，平均降低疫苗接种意愿1.5个百分点。然而，更重要的是，一个故事暗示疫苗对健康存在风险的程度，而非其真实性本身，是预测其负面说服影响的最佳指标。即使是虚假信息，如果其不暗示疫苗有害，其影响也较小；反之，即使内容事实准确，若暗示疫苗有害，也会降低接种意愿。

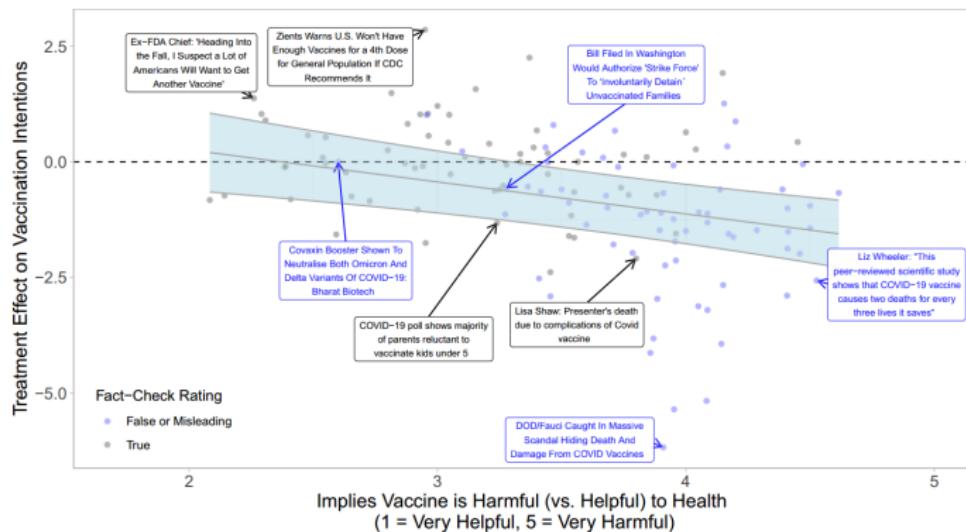


图 34: 疫苗相关标题对接种意愿的影响（按暗示的健康危害程度划分）

**暴露程度：**在脸书上，被标记的错误信息URL在2021年前三个月获得了约870万次浏览，仅占同期27亿次疫苗相关URL总浏览量的0.3%。相比之下，那些未被事实核查机构标记但暗示疫苗对健康有害的“疫苗怀疑论”内容，其中许多来自可信的主流新闻媒体，其浏览量高达数亿次。例如，一篇关于一名健康医生接种疫苗两周后死亡的报道，在脸书上被浏览超过5490万次，是所有被标记错误信息总浏览量的六倍以上。

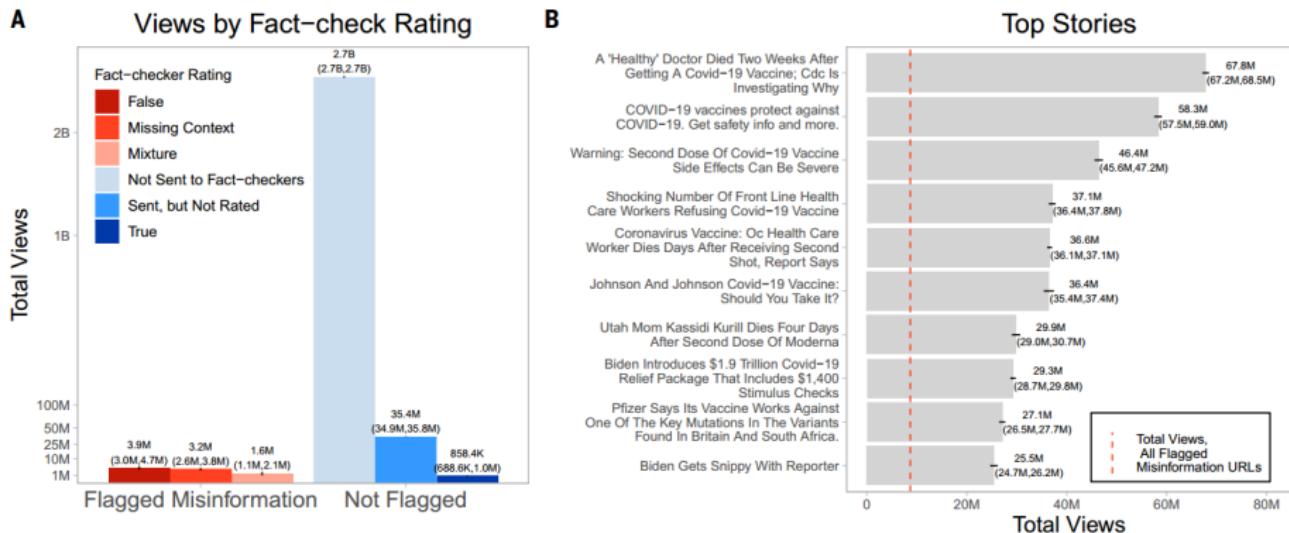


图 35: 2021年初三个月内在 Facebook 上公开分享超过100次的疫苗相关内容的曝光情况

**预测模型性能:** 众包综合评分能够有效预测实验中观察到的处理效应（调整后相关性0.75）。基于众包评分训练的机器学习模型在预测未见过的URL的众包综合评分方面表现良好，其预测的86%的综合评分与真实综合评分的差异在0.5个量表点以内。

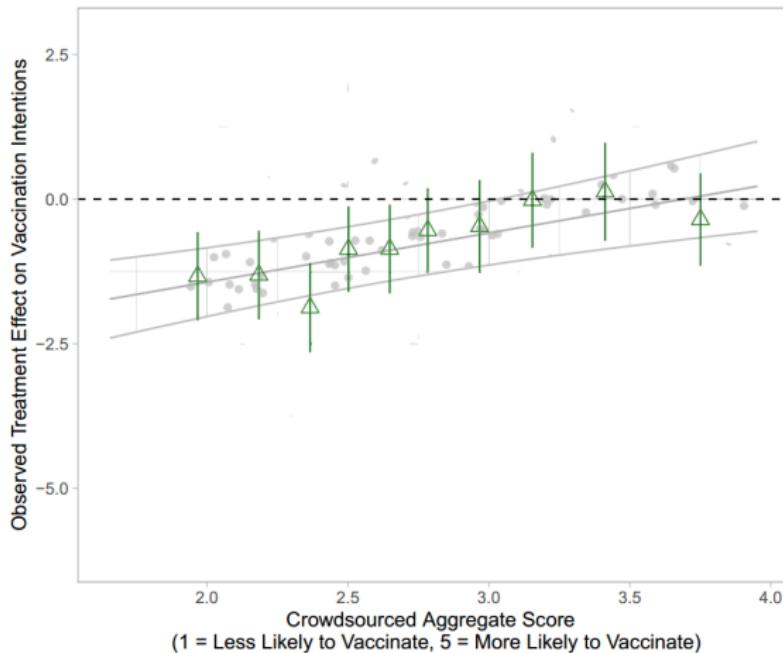


图 36: 基于众包综合评分的接种意愿处理效应

**总体影响：**虽然被标记的错误信息在被浏览时，其平均降低疫苗接种意愿的预测效应（中位数为-1.36个百分点）远大于未被标记内容（中位数为-0.3个百分点）。但是，当考虑到巨大的浏览量差异后，未被标记的疫苗怀疑论内容对整体疫苗犹豫的预测影响远超被标记的错误信息。在诱导犹豫的URL中，未被标记的疫苗怀疑论内容估计使每位美国脸书用户的疫苗接种意愿降低了2.28个百分点，而被标记的错误信息仅降低了0.05个百分点，两者相差46倍。

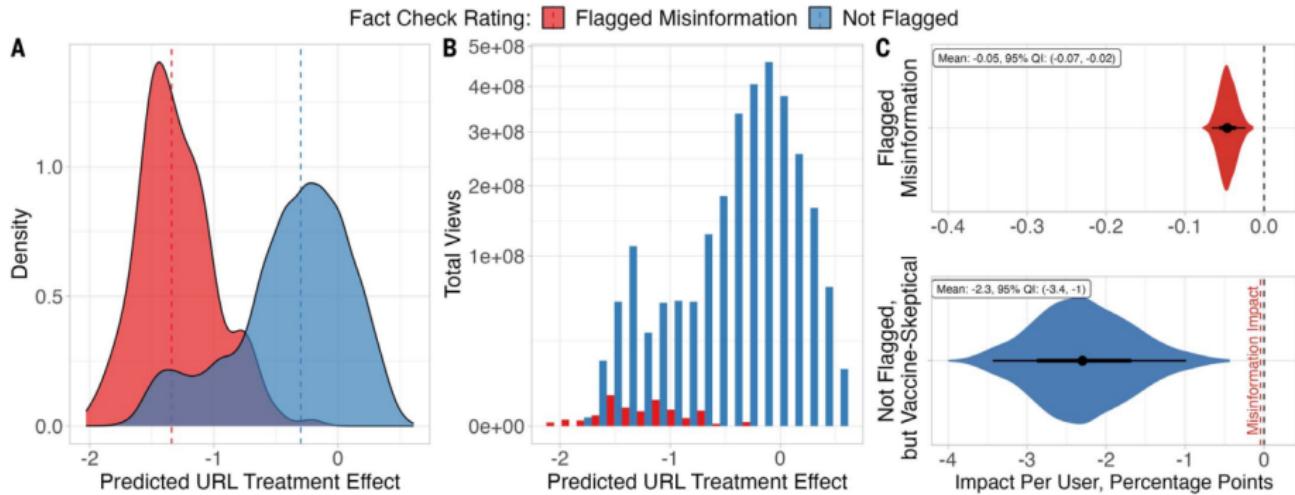


图 37: 2021年初三个月内在 Facebook 上公开分享超过100次的疫苗相关URL对接种意愿的预测性影响

## 研究结论

- 被事实核查机构标记的错误信息，在用户接触到的情况下，确实能够显著降低其新冠疫苗的接种意愿。
- 然而，由于这类被标记的错误信息在脸书上的暴露率相对较低，其对驱动整体疫苗犹豫所起的作用远小于那些未被事实核查机构标记、但具有疫苗怀疑倾向的内容（即“疫苗怀疑论”内容）。
- 大量具有高度影响力的疫苗怀疑论内容来源于主流新闻媒体，它们通常以事实准确但可能产生误导的方式报道与疫苗相关的负面事件（如罕见死亡案例），而未充分 contextualize 这些事件的极端罕见性或不确定性。
- 因此，尽管限制错误信息的传播具有重要的公共卫生益处，但同样至关重要的是要关注那些事实准确但可能产生误导的“灰色地带”内容，因为这些内容因其巨大的传播量而可能造成更大的总体危害。政策制定者和科技公司不应仅关注内容的真实性，还应考虑其潜在的误导性及其传播范围和影响。

**启发：**(various types of) misinformation的危害到底有多大，应该如何量化，危害程度有我们以为得那么大吗？断章取义等片面真实但却有意/无意带有误导倾向的信息或许才具有广泛危害。

# Supershareders of fake news on Twitter (Science 2024)

## 1. 标题

推特上假新闻的“超级传播者”

## 2. 作者及单位

Sahar Baribi-Bartov, Briony Swire-Thompson, Nir Grinberg

- 以色列本古里安大学软件与信息系统工程系
- 美国东北大学网络科学研究所
- 美国哈佛大学定量社会科学研究所

## 3. 文献来源

Baribi-Bartov, Sahar, Briony Swire-Thompson, and Nir Grinberg. “*Supershareders of fake news on Twitter*.” Science 384.6699 (2024): 979-982.

## 4. 文献类型与关键词

文章类型：实证研究

英文关键词：misinformation; fake news; supersharers; Twitter; social media influence; network influence

中文关键词：误导信息；假新闻；超级传播者；推特；社交媒体影响；网络影响力

## 研究动机

社交媒体已成为政治信息的主要获取途径，但其去中心化特性也为误导信息（fake news）的快速扩散创造了条件。尽管已有研究聚焦外国势力和自动化账号（bots）对假新闻传播的影响，普通用户通过大规模重复分享（即“信息洪流”）扮演的角色却鲜有实证考察。近期发现，**极少数所谓的“超级传播者”（supersharers）占据了注册选民在推特上80%的假新闻分享量，但关于他们的影响范围、人口特征及所依赖的技术手段仍然知之甚少。**为全面理解当下误导信息生态，并为平台干预提供针对性策略，亟需系统研究这部分核心用户群体的作用机制与行为特征。

## 研究目标

本研究以2020年美国总统选举期间活跃的664 391名注册选民为样本，识别出贡献80%假新闻分享量的2 107名超级传播者，旨在**量化该群体在社交网络中的重要性，刻画其人口社会学特征，并揭示其不依赖大规模自动化而实现持续高频分享的技术赋能路径**，为制定高效的差异化平台干预提供实证依据。

### 研究问题

- 超级传播者在Twitter平台及其社交网络中有多重要？
- 超级传播者是哪些人？
- 是社交媒体的哪些机制使得超级传播者能够大规模分享虚假新闻而不面临审核？

# 研究方法

- **数据收集与样本：**研究利用了一个包含664,391名在2020年美国总统大选期间（2020年8月至11月）活跃于Twitter的注册美国选民的大规模追踪调查数据。
- **超级传播者识别：**将分享了追踪调查样本中80%虚假新闻内容的最多产分享者定义为“超级传播者”(N=2107)。
- **虚假新闻定义：**采用来源层面的虚假新闻定义，即那些看似合法新闻机构但缺乏确保信息准确性和可信度的编辑规范和流程的域名。该定义基于Grinberg等人先前研究的手动标记列表，并使用NewsGuard评级进行了更新。分析仅限于包含被机器学习分类器（经人工编码员验证）识别为政治性的外部链接的推文。
- **参照群体：**为进行对比分析，研究设置了两个主要参照群体：1) 非虚假政治新闻的重度分享者 (SS-NF, N=11,199)，定义为贡献了80%非虚假政治新闻的用户；2) 一个规模相似的随机小组成员样本 (N=11,199)。部分分析还比较了第三个群体，即普通虚假新闻分享者 (average fake news sharers, N=10,464)，定义为在研究期间分享了三条或更多指向虚假新闻来源推文且不属于超级传播者群体的用户。

- **影响力衡量:** 通过网络影响力（邻居节点的度之和）、内容互动（回复、转推、引用比例）以及关注者接触虚假新闻的程度来评估超级传播者的重要性。
- **人口社会学特征分析:** 采用逻辑回归模型，对比超级传播者与各参照群体在性别、年龄、党派归属、种族、地理位置、教育程度和年收入等方面的差异。
- **技术手段分析:** 通过三种方法检验自动化工具的使用情况：1) 使用Botometer机器人检测工具结合人工标注；2) 分析发帖时间模式（如发帖时段、会话长度与数量、发帖间隔）；3) 检查推文元数据中是否表明超级传播者比其他群体更多使用支持自动化的应用程序。同时，分析了转推行为的比例。

# 研究结果

**虚假新闻的普遍性与集中性：**在2020年选举期间，平均每天有7.0%由追踪调查样本分享的政治新闻链接指向虚假新闻来源。仅占样本0.3%的2107名超级传播者贡献了80%的虚假新闻推文。虚假新闻的分享比非虚假政治新闻的分享更为集中。超级传播者的主导地位在整个选举期间持续存在。

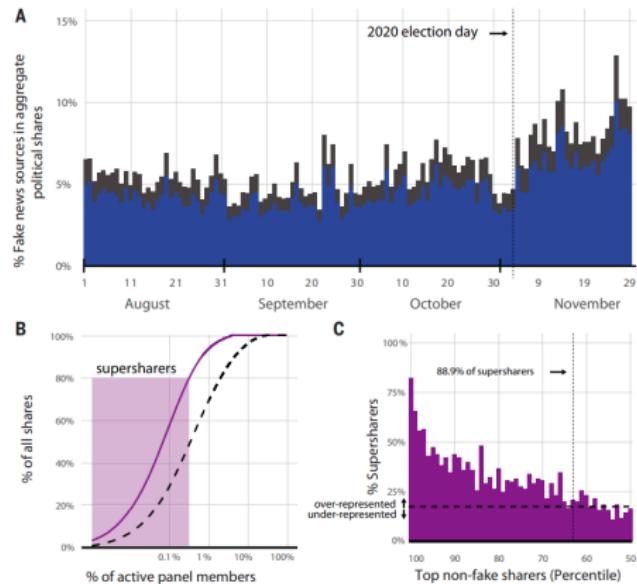


图 38: 虚假新闻分享的流行程度随时间变化及集中度

## 超级传播者的重要性

- **影响力广度:** 5.2%的Twitter注册选民直接关注了至少一名超级传播者。
- **网络影响力:** 超级传播者具有显著更高的网络影响力，中位数超级传播者在网络影响力方面位列样本的第86个百分位，比SS-NF群体的中位数高29%。
- **内容互动:** 更多人与超级传播者的内容互动，且与超级传播者内容互动的追踪小组成员比例也更高。
- **对关注者的影响:** 约五分之一的虚假新闻重度消费者关注了超级传播者。超级传播者的普通关注者从其网络中获取到链接至虚假新闻来源的政治新闻的可能性是普通追踪小组成员的2.5倍。超级传播者贡献了其普通关注者所接触虚假新闻的近四分之一（24.4%），并且是11.3%关注者的唯一虚假新闻来源。

## 超级传播者的人口社会学特征

- 与所有参照群体相比，超级传播者中女性（59%）、老年人（平均年龄58.2岁）和共和党人（64%）的比例显著更高。其中性别差异主要源于共和党超级传播者中女性的超高代表性。
- 与普通样本和SS-NF群体相比，超级传播者中高加索人（白人）的比例也显著更高。
- 超级传播者在佛罗里达州、亚利桑那州和德克萨斯州的代表性过高。
- 超级传播者来自教育程度略低的地区，平均教育年限比普通样本和SS-NF群体少0.3年。相对于基于教育水平的预期收入，超级传播者的年收入平均比SS-NF和普通虚假新闻分享者群体高出约2500美元。但研究者也强调这些差异的幅度较小。

## 超级传播者的技术手段

- **自动化程度低：**没有证据表明超级传播者广泛使用自动化工具。使用Botometer等工具检测发现，超级传播者中可被视为机器人的比例不超过7.1%，与SS-NF群体无显著差异。发帖时间模式和自动化应用使用方面也未发现显著差异。
- **主要依赖手动转推：**超级传播者行为的最大差异在于其极高的转推率。普通超级传播者发布的推文中，有74.7%是转推，远高于SS-NF群体的59.9%和普通样本的32.7%。这表明其巨大的分享量主要通过手动且持续的转推行为产生。

## 2. Nature、Science论文逐篇总结

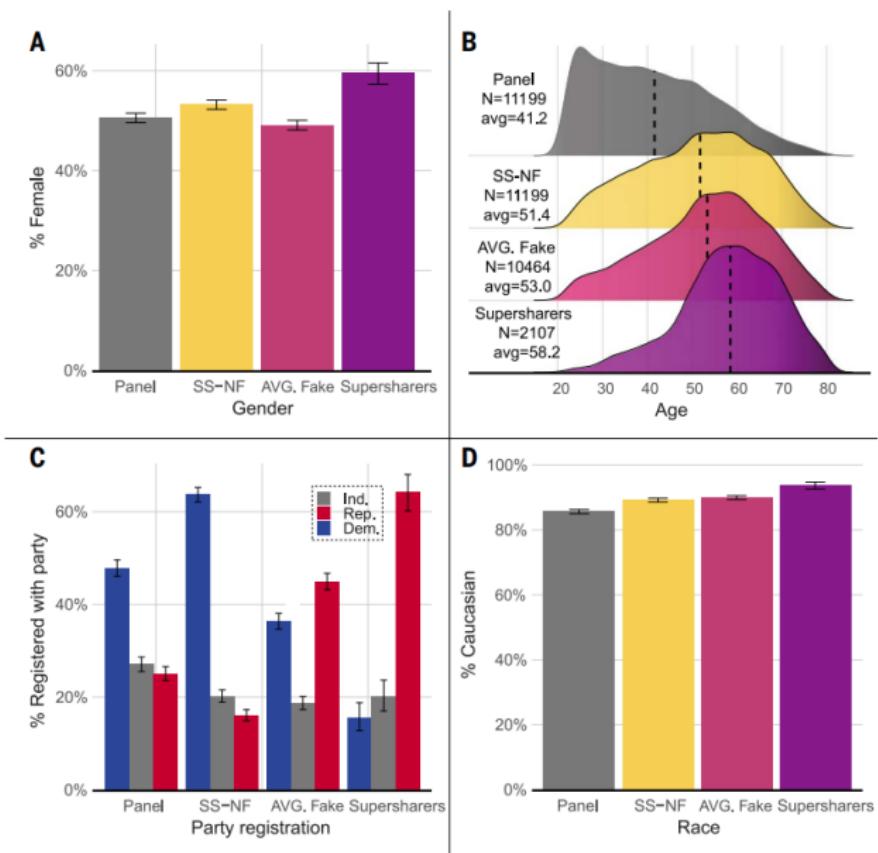


图 39: 超级传播者与三个参照群体（SS-NF、普通虚假新闻分享者和随机抽样的小组成员）之间的人口统计学差异

## 研究结论

本研究首次系统揭示，极少数普通用户通过“低技术”持续转推方式，在推特平台上充当假新闻的主要放大器，严重威胁数字民主的平等话语权。仅依靠自动化检测与事实核查难以对其行为形成有效制约，平台应将干预重点聚焦于这一小群体，例如对其转发权限进行限制或实施定向警示，以最小范围的用户管理实现最大的风险缓释。

# 目录

1 文献检索结果

2 Nature、Science论文逐篇总结

3 其他3本期刊论文选择性摘读

## TL;DR

超过三分之一的论文( $37\% = \frac{10}{27}$ )使用了NLP方法或者复杂网络方法：

只用到NLP方法 (4篇): ([Kim et al., 2025](#)), ([Roozenbeek et al., 2022](#)), ([Jones et al., 2017](#)), ([Sultan et al., 2024](#))

只用到复杂网络方法 (2篇): ([Ferraz de Arruda et al., 2022](#)), ([Del Vicario et al., 2016](#))

同时使用了NLP和复杂网络方法 (4篇): ([Bovet, A., & Makse, H. A., 2019](#)), ([Mosleh, M., & Rand, D. G., 2022](#)), ([Broniatowski et al., 2023](#)), ([Smith et al., 2021](#))

# From subcritical behavior to a correlation-induced transition in rumor models (Nature Communications 2022) I

**研究方法：**本研究综合运用了数学建模、蒙特卡洛模拟和解析理论分析。首先，研究者定义了标准的Maki-Thompson (MT)模型和一个包含遗忘机制的改进模型，这两个模型都将个体划分为无知者(X)、传播者(Y)和压制者(Z)三种状态。接着，通过在不同网络结构（如随机正则网络和幂律无标度网络）上进行蒙特卡洛模拟，研究者考察了模型的相变行为、次临界特性以及谣言生命周期等动态指标。为了进行精确的临界点估计和次临界行为表征，研究采用了准静态算法 (quasi-stationary algorithm) 和生命周期方法 (lifespan method)。在解析层面，研究者运用了基于分支过程的近似方法来估计相变临界点，并进行了均场方程的渐近分析以揭示一阶均场近似无法捕捉相变的原因。此外，还针对次临界区的谣言生命周期进行了理论推导。

**NLP方法：**无

**复杂网络方法：**

- **模型构建与网络表示：**将谣言传播过程定义在由邻接矩阵 $A$ 表示的无向网络上。个体间的接触遵循网络结构。

# From subcritical behavior to a correlation-induced transition in rumor models (Nature Communications 2022) II

- 网络拓扑的应用与分析:

- 随机正则网络 (**Random Regular Networks**): 用于数值模拟Maki-Thompson (MT)模型及其改进模型的相变行为和次临界特性，其中网络中所有节点度相同。研究者在此类网络上观察到MT模型的相变以及特定的次临界行为。
- 无标度网络/幂律网络 (**Scale-Free/Power-Law Networks**): 采用无关联构型模型 (Uncorrelated Configuration Model) 生成幂律度分布 ( $P(k) \sim k^{-\gamma}$ ) 的网络，研究网络异质性对临界点的影响，特别是分析了临界点如何随网络规模 $N$ 和幂指数 $\gamma$ 变化(例如，在 $\gamma < 3.0$ 时临界点趋于消失，而 $\gamma > 3.0$ 时收敛到非零值)。改进模型的次临界行为也在幂律网络上进行了数值研究。

- 基于网络结构的解析方法:

- 分支过程近似 (**Branching Process Approximation**): 该方法应用于局部树状网络 (locally tree-like networks)，通过计算一个初始传播者平均能产生的新传播者数量的期望值 $q_k(i)$ （并对其在度分布上求平均得到 $q(1)$ ）来估计相变临界点（当 $q(1) > 1$ 时发生传播）。此方法明确考虑了局部反馈效应，即传播事件会增加初始传播节点被压制的概率，而这是传统一阶均场近似所忽略的关键因素。

# From subcritical behavior to a correlation-induced transition in rumor models (Nature Communications 2022) III

- **均场方程的渐近分析 (Asymptotic Analysis of Mean-Field Equations):** 在分析改进模型的一阶均场方程 (形如  $dx_i/dt = \dots, dy_i/dt = \dots, dz_i/dt = \dots$ ) 时, 考虑了网络结构 (例如, 在正则网络中, 邻接矩阵的主特征值  $\Lambda_{max}$  被用于求解传播者密度  $y_i$  的近似表达式)。这部分分析旨在说明为何此类一阶近似无法预测模型中的相变。
- **次临界行为的近似分析:** 在推导次临界区谣言到达吸收态的平均时间  $\langle T_{abs} \rangle$  时, 假设了局部树状网络结构, 并考虑了节点的平均度  $\langle k \rangle_k$ 。

# Influence of fake news in Twitter during the 2016 US presidential election (Nature Communications 2019) I

**研究方法:** 本研究采用了多方面结合的研究方法。首先，通过Twitter API收集了2016年美国总统选举前五个月内提及两位主要候选人的1.71亿条推文。其次，基于OpenSources.co等外部资源对推文中链接的新闻源进行分类，区分为虚假新闻、极端偏见新闻和不同政治倾向的传统新闻。接着，利用非官方Twitter客户端的使用情况来评估自动账户（机器人）在新闻传播中的作用。然后，构建了各类新闻的转发网络，并使用集体影响力（Collective Influence, CI）算法识别关键传播者。最后，通过对各类新闻传播者和候选人支持者活动的时间序列进行相关性分析和多变量因果网络重构，探究它们之间的动态关系和影响机制。研究还使用了机器学习算法对用户进行立场分类（支持克林顿或特朗普）。

## NLP方法：

- 用户立场分类：研究提及基于推文内容，使用监督机器学习算法对用户进行分类，判断其为希拉里·克林顿的支持者还是唐纳德·特朗普的支持者。该算法在一个通过主题标签共现网络获得的数据集上进行训练。这涉及到对推文文本内容的分析和特征提取。

# Influence of fake news in Twitter during the 2016 US presidential election (Nature Communications 2019) II

- 数据收集的关键词选择：在数据收集阶段，使用了与两位主要候选人相关的关键词（如hillary, clinton, trump, realdonaldtrump）来筛选推文。
- URL提取与解析：从推文中提取URL，并对短链接进行解析以获取最终目标URL的主机名。

## 复杂网络方法：

- 信息流网络构建与分析：为每种新闻类别（如虚假新闻、极端偏见新闻、传统新闻等）构建了转发网络（retweet networks）。在这些网络中，用户是节点，转发行构成有向边，表示信息的流动方向。
- 网络结构特征量化：分析了这些转发网络的结构特性，包括节点数、边数、平均度 ( $\langle k \rangle$ )、出度和入度分布的异质性（通过标准差与平均度的比值  $\sigma(k_{out})/\langle k \rangle$  和  $\sigma(k_{in})/\langle k \rangle$  来衡量）。这些分析揭示了不同类型新闻传播网络的连接密度和结构差异。
- 关键传播者识别：应用了集体影响力（Collective Influence, CI）算法来识别在信息转发网络中最具影响力的用户（即“超级传播者”）。该算法基于网络的最优渗透理论，能够识别那些移除后能最大程度瓦解网络连接的节点。

## Influence of fake news in Twitter during the 2016 US presidential election (Nature Communications 2019) III

- 网络可视化：通过可视化展示了顶级新闻传播者构成的转发网络，揭示了不同媒体类别传播者之间的聚集情况和相对重要性。

# Measuring exposure to misinformation from political elites on Twitter (Nature Communications 2022) I

**研究方法：**本研究首先利用专业事实核查网站PolitiFact的数据，为816名精英（公众人物和组织）基于其公开声明的真实性计算了“虚假性评分”。随后，通过平均用户在Twitter上关注的这些精英的虚假性评分，为每位用户分配一个“精英错误信息接触评分”。研究收集了一个包含5000名Twitter用户的样本数据，分析了他们的关注行为、分享内容、以及推文中的语言特征。研究通过关联分析验证该评分的有效性，即检验其与用户分享新闻质量（由专业事实核查员和普通民众评价）、估算的意识形态、语言的攻击性（使用Google Jigsaw Perspective API）和道德愤怒表达程度之间的关系。此外，研究还构建并分析了用户的共关注、共分享和共转发网络，以探究错误信息接触的社群结构和意识形态关联。最后，研究将此方法开发成开源R包和API供研究社区使用。

## NLP方法：

- 使用谷歌Jigsaw Perspective API（Google Jigsaw Perspective API）来计算用户推文中语言的平均攻击性程度（language toxicity），以探究精英错误信息接触评分与用户推文攻击性之间的关系。

## Measuring exposure to misinformation from political elites on Twitter (Nature Communications 2022) II

- 使用一种已发表的估计器（Brady et al. 2021）来衡量用户推文中道德愤怒（moral outrage）语言的平均表达水平，以探究精英错误信息接触评分与用户表达道德愤怒之间的关系。

### 复杂网络方法：

- 构建了用户的共分享网络（co-share network），其中节点代表被至少20位用户分享的网站域名，边权重基于共同分享这些域名的用户数量。该网络用于识别被不同错误信息接触程度和意识形态用户群体优先分享的域名集群。
- 对共分享网络进行了社群检测分析（community-detection analysis），以识别出不同的域名集群，并分析这些集群在用户平均错误信息接触评分和估算意识形态上的差异。
- 类似地，还构建并分析了共关注网络（co-follower network）和共转发网络（co-retweet network），以补充对错误信息接触和意识形态在网络结构中关联性的理解。

# Differential impact from individual versus collective misinformation tagging on the diversity of Twitter (X) information engagement and mobility (Nature Communications 2025) I

**研究方法：**本研究利用了Twitter (X)的大规模数字痕迹数据，识别了两类被错误信息标记的用户：一类是被其他个体用户通过引用PolitiFact事实核查文章进行标记的用户（个体化标记），另一类是其推文收到通过Twitter的Community Notes平台集体审核并发布的标记的用户（集体性标记）。研究追踪了7733名用户在被标记前后发布的712,948条推文（包括发帖、转推和引用）。主要结果变量为“政治多样性”（衡量用户是否引用了与其自身政治立场相反的信源）和“内容多样性”（衡量用户发布的推文是否讨论了其历史推文中不常见的新颖话题，通过SentenceBERT模型提取推文语义向量并计算距离得到）。为处理观察性数据中的因果推断问题，研究主要采用了中断时间序列分析（Interrupted Time Series, ITS）和延迟反馈分析（Delayed Feedback, DF）两种准实验方法。此外，研究还对标记信息的语言特征（攻击性、情感、长度、可读性、延迟）进行了比较分析，并进行了一系列控制分析和稳健性检验。

## NLP方法：

# Differential impact from individual versus collective misinformation tagging on the diversity of Twitter (X) information engagement and mobility (Nature Communications 2025) II

- 使用基于Transformer的句子嵌入模型（SentenceBERT，具体为Twitter4SSE模型）将每条推文内容转换为高维语义向量，用于计算用户推文的“内容多样性”。该方法通过量化新推文与用户历史推文平均语义向量的距离，来评估用户是否在探索新颖主题。
- 使用BERTopic进行主题建模（Topic Modeling），从未经处理的错误信息标记推文中提取潜在主题。此方法结合了Twitter4SSE的语义嵌入、UMAP降维和HDBSCAN聚类技术。该方法用于识别个体化标记和集体性标记所针对的错误信息在主题上的差异，并在控制分析中控制这些主题差异。
- 使用谷歌Jigsaw Perspective API测量错误信息标记信息的“攻击性”（toxicity）。
- 使用VADER（Valence Aware Dictionary and sEntiment Reasoner）进行情感分析，以估计错误信息标记信息的“情感色彩”（sentiment scores）。
- 使用Flesch-Kincaid阅读易读性公式（Flesch-Kincaid Reading Ease score）评估错误信息标记信息的“可读性”。
- 在稳健性检验中，使用ChatGPT (gpt-4)模型来注释个体化标记中引用的PolitiFact链接是否用于纠正原始发帖人而非支持他们。

**复杂网络方法：**无

# Psychological inoculation improves resilience against misinformation on social media (Science Advances 2022) I

**研究方法：**本研究包含七项预注册的实验。其中六项为随机对照研究（总样本n=6464，为美国全国配额样本），一项为在YouTube平台上进行的具有生态效度的田野研究（n=22,632）。研究人员开发了五个简短的动画免疫视频，分别针对五种常见的错误信息操纵技巧：情绪操控语言、不连贯论证、错误二分法、寻找替罪羊和人身攻击。在随机对照研究（研究1-6）中，参与者被随机分配观看其中一个免疫视频或一个中性的控制视频（内容无关，时长和风格相似）。观看视频后，参与者对10条合成的社交媒体帖子（模拟Twitter和Facebook帖子，随机呈现操纵性或中性版本）进行评价。测量的结果变量包括：操纵技巧识别能力、识别信心、内容可信度辨识以及分享意愿的质量（以“辨识度”即中性帖与操纵性帖评分之差为主要指标）。研究6还检验了结果变量呈现顺序的影响。在YouTube田野研究（研究7）中，两个免疫视频（情绪操控语言和错误二分法）作为广告投放给YouTube用户，并在观看广告后24小时内，通过嵌入YouTube平台的单项测试题（识别标题中的操纵技巧）评估其效果，并与未观看广告的控制组进行比较。

## NLP方法：

## Psychological inoculation improves resilience against misinformation on social media (Science Advances 2022) II

本研究在为情绪操控语言免疫视频（研究1和研究6）验证其刺激材料（manipulative stimuli）是否确实捕捉到了预期的（操纵性）情绪维度时，使用了情感分析（sentiment analysis）方法。这一步骤是为了确保实验材料的有效性，即被设计为具有情绪操控性的刺激内容确实比中性对应内容更具情绪性。

- 对情绪操控语言研究中的刺激材料进行了情感分析，以验证操纵性刺激材料相较于中性刺激材料在情绪维度上的差异，从而确保实验材料的构念效度。

复杂网络方法：无

# The efficacy of Facebook's vaccine misinformation policies and architecture during the COVID-19 pandemic (Science Advances 2023) I

**研究方法：**本研究使用CrowdTangle工具收集了2019年11月15日至2022年2月28日期间，Facebook上216个公共主页（Pages）和100个公共群组（Groups）（均包含反疫苗和支持疫苗两类）的公开数据，包括帖子数量和用户互动量（点赞、评论、分享、情感反应的总和）。研究采用比较性中断时间序列（Comparative Interrupted Time-Series, CITS）设计，以Facebook于2020年12月开始明确移除COVID-19疫苗错误信息及相关账户的政策为干预时间点，比较政策实施前后反疫苗和支持疫苗两类主页/群组中内容数量及互动量的变化趋势。研究还分析了帖子内容的特征变化，包括：通过结构主题模型（STM）分析帖子主题的比例变化；分析帖子中链接到低可信度和政治极化外部网站的比例；以及探究Facebook平台架构特征（如主页间的链接、群组与主页间的协同发帖行为、用户对帖子的情感反应分布）如何影响信息传播和政策效果。

## NLP方法：

- 使用‘langdetect’ Python包识别英文帖子，以确保后续主题模型分析的语言一致性。
- 对Facebook帖子的文本内容（包括消息、图片文本、链接文本和描述）进行预处理，如转换为小写、移除数字、标点符号、符号和网址链接，为主题模型分析做准备。

# The efficacy of Facebook's vaccine misinformation policies and architecture during the COVID-19 pandemic (Science Advances 2023)

## II

- 应用结构主题模型（Structural Topic Models, STM），使用R语言的‘stm’包，通过Mimno和Lee算法自动选择主题数量，从帖子文本中提取潜在主题，并分析这些主题在反疫苗和支持疫苗内容中以及政策实施前后的比例变化，以评估错误信息主题内容是否减少。

### 复杂网络方法：

本研究结合了复杂网络分析的视角和方法来探究Facebook系统架构对其错误信息政策效果的影响。具体如下：

- 分析了Facebook公共主页之间的链接网络：通过提取帖子中指向其他Facebook页的链接，构建了主页间的连接网络，并使用力导向布局算法（force-directed layout algorithm）进行可视化，以展示反疫苗和支持疫苗主页集群的连接密度和结构，评估平台政策是否削弱了反疫苗主页集群的灵活性。

## The efficacy of Facebook's vaccine misinformation policies and architecture during the COVID-19 pandemic (Science Advances 2023) III

- 研究了不同主页与群组间的协同链接分享行为（coordinated link-sharing）：通过识别“近似同时”（near-simultaneous，本研究定义为33秒内）分享相同URL的行为作为潜在协同活动的信号，分析这种协同行为在政策实施前后的变化，并探讨这种行为如何帮助内容规避移除。

## The spreading of misinformation online (PNAS 2016) I

**研究方法:** 本研究采用了定量分析和数据驱动建模相结合的方法。首先，通过Facebook Graph API收集了关于科学新闻和阴谋论两大类公共主页在五年（2010-2014）内的所有帖子及其用户互动数据。接着，对收集到的数据进行统计分析，比较两类信息在级联规模、生命周期、树高度等方面的特征。然后，引入用户极化和边同质性的概念，分析同质性集群（回声室）的形成及其在信息传播中的作用。最后，基于分析结果，提出了一个考虑同质性和极化的数据驱动的谣言传播渗透模型，并通过模拟验证了模型的有效性。

**NLP方法:** 无

**复杂网络方法:**

- **分享树 (Sharing Tree) 的构建与分析:** 将新闻的分享过程构建为有向无环图（具体为有根树），其中节点代表分享用户，边代表分享行为。通过分析分享树的规模（节点数）、高度（从根到叶节点的最长路径长度）和最大度等网络拓扑属性，来刻画信息传播的级联动态。

## The spreading of misinformation online (PNAS 2016) II

- 同质性集群 (**Homogeneous Clusters**) / 回声室 (**Echo Chambers**) 的识别与分析：基于用户对不同类型内容的偏好（用户极化  $\sigma$ ），定义了分享网络中连接两个节点的边的同质性 (edge homogeneity)。通过分析平均边同质性，揭示了信息传播主要发生在观点相似的用户群体内部，形成了高度同质化的集群，即回声室。
- 渗透模型 (**Percolation Model**) 的构建与应用：提出了一种基于小世界网络 (small-world network) 的渗透模型来模拟谣言传播。该模型考虑了用户的观点 ( $\omega_i$ )、新闻内容的倾向性 ( $\theta_j$ )、分享阈值 ( $\delta$ ) 以及网络中同质连接的比例 ( $\phi_{HL}$ )，用以解释和预测传播级联的规模。该模型可以视为一种分支过程 (branching process)。

# Distress and rumor exposure on social media during a campus lockdown (PNAS 2017) I

**研究方法：**本研究采用了混合方法设计，结合了问卷调查（研究1）和大数据分析（研究2）来探究同一次校园枪击封锁事件。

**研究1：**在事件发生一周后对该大学在封锁期间的学生进行了匿名在线问卷调查（ $n=3,890$ ），收集了他们关于信息获取渠道、接收冲突信息的情况、对不同渠道的信任度以及急性应激水平（使用标准化量表测量）的数据。通过回归分析检验冲突信息暴露、不同沟通渠道使用及其信任度与急性应激之间的关系。

**研究2：**采用大数据方法，通过Twitter API收集了该大学两个官方Twitter账户的部分关注者在封锁事件前后约5小时内发布的推文数据。对这些推文进行手动编码以识别谣言，并使用R脚本和LIWC词典自动标记提及封锁事件和包含负面情绪词汇的推文，以分析谣言传播的时间进程和社区层面困扰情绪的模式，并将其与官方校园警报的发布时间进行对照。

## NLP方法：

- **文本情感分析：**使用R脚本和Linguistic Inquiry and Word Counter (LIWC) 的负面情绪词典，自动标记推文中包含的负面情绪词汇，以量化社区层面的事件相关负面情绪。

## Distress and rumor exposure on social media during a campus lockdown (PNAS 2017) II

- **关键词提取/事件相关推文识别：**使用R脚本和一个包含17个词条的自定义词汇列表（包括特定情境词如lockdown, #[university name]strong等）来自动识别和标记与枪击和封锁事件相关的推文。
- **词频与N-gram分析（用于验证谣言代理指标）：**在研究1中，为了验证“接收冲突信息”作为谣言暴露的代理指标，研究者对一个开放式问题的回答进行了词语提及分析（如“rumor”）和N-gram分析（如二元“multiple shooter”，“rumor spread”；三元“rumor [of] multiple shooters”），使用了名为Meaning Extraction Helper的文本分析程序。

复杂网络方法：无

# Automatic detection of influential actors in disinformation networks (PNAS 2021) I

**研究方法：**本研究提出并验证了一个端到端的自动化框架，用于检测和表征影响行动（IO）。该框架按顺序包括五个主要阶段：1) 目标数据收集：基于关键词、账户、语言和时空范围，通过Twitter API收集与潜在IO相关的社交媒体内容。2) 叙事检测：利用主题建模算法（如MALLET）从收集的数据中自动生成叙事，并由分析师识别感兴趣的IO叙事。3) IO账户分类：采用半监督机器学习方法（结合使用scikit-learn和Snorkel库），基于账户的行为特征、画像特征、使用的语言以及推文中的1-gram和2-gram等特征，对参与选定叙事的账户进行IO分类评分。训练数据结合了已知的IO账户（来自Twitter的选举诚信数据集）和通过启发式规则（Snorkel的软标签函数）标记的账户。4) 网络发现：基于账户间的互动模式（如转推）构建参与IO叙事的社交网络（叙事网络）。5) 影响评估：采用基于网络潜力结果框架（network potential outcome framework）的因果推断方法，量化每个账户对叙事在整个网络传播的独特因果贡献。该方法考虑了社会混杂因素（如社群成员身份、受欢迎程度），并通过拟合泊松广义线性混合模型（Poisson GLMM）来估计缺失的潜力结果。

## NLP方法：

# Automatic detection of influential actors in disinformation networks (PNAS 2021) II

- **目标数据收集：**基于关键词从Twitter API收集数据。
- **叙事检测：**使用主题建模算法（具体提及MALLET）从收集到的Twitter数据中自动生成不同的叙事。这些叙事由词袋模型表示。研究提及更复杂的NLP技术如Transformer模型也可用于识别显著叙事。
- **IO账户分类的特征工程：**分类器的特征空间包括语言学特征，如账户使用的语言种类以及推文中使用的1-gram和2-gram词组。
- **叙事内容分析示例：**通过词云图展示了与2017年法国选举中“离岸账户”叙事相关的最频繁词汇和表情符号，分别针对英文和法文语料库。

## 复杂网络方法：

- **网络构建 (Network Discovery)：**基于账户间的互动（如转推），构建了参与特定IO叙事的社交网络，称为叙事网络 (narrative network)。网络中的节点是账户，边代表互动或影响的强度。
- **社群发现 (Community Detection)：**利用基于马尔可夫链蒙特卡洛 (MCMC) 的模块度最大化方法（具体为随机块模型/blockmodel）对叙事网络进行社群划分，以识别网络中不同的行动团体或叙事簇。

# Automatic detection of influential actors in disinformation networks (PNAS 2021) III

- **影响评估中的网络结构：**在因果推断影响评估中，叙事传播通过网络中的n跳暴露( $n$ -hop exposure)来建模，即一个账户对叙事的影响会通过网络路径传递给其他账户。影响矩阵 $A$ 被联合估计，并用于量化叙事在网络中的传播。
- **网络可视化：**使用网络图来可视化叙事网络中的社群结构、IO分类器得分以及基于因果推断的影响力得分，其中节点大小表示影响力，颜色表示IO分类器得分。
- **传统网络中心性度量：**在结果比较中，提及了PageRank中心性作为传统的影响力度量之一，并将其与本研究提出的因果影响度量进行了对比。

# Susceptibility to online misinformation: A systematic meta-analysis of demographic and psychological factors (PNAS 2024) I

**研究方法:** 本研究采用系统性个体参与者数据元分析 (systematic individual participant data meta-analysis) 的方法。研究者检索并筛选了符合纳入标准 (如参与者为美国成年人、采用新闻标题真实性判断范式、包含真假新闻等) 的已发表和未发表研究。在获得原始数据后, 对来自31项实验的11561名美国参与者所做的256337个独立选择进行了汇总分析。核心分析方法是应用信号检测理论 (Signal Detection Theory, SDT), 通过贝叶斯广义线性混合效应模型 (Bayesian generalized linear mixed-effects model) 来区分辨别能力 (区分真假新闻的能力) 和反应偏差 (将新闻判断为真或假的倾向)。模型中包含了人口统计学因素 (年龄、性别、教育、政治身份)、心理因素 (分析性思维、意识形态一致性、动机性反思) 以及一些元分析变量 (新闻主题、研究平台、是否展示来源) 作为预测变量。由于熟悉度数据在许多研究中缺失, 对其的分析采用了单独的完整案例SDT模型。

## NLP方法:

在预处理阶段使用了自然语言处理 (NLP) 工具 (GPT-4) 来辅助数据处理。

## Susceptibility to online misinformation: A systematic meta-analysis of demographic and psychological factors (PNAS 2024) II

- **新闻标题政治倾向分类:** 由于原始研究中16/31的研究缺少新闻标题的政治倾向信息，研究者使用GPT-4对所有新闻标题的政治倾向进行分类，分为强烈共和党、中度共和党、倾向共和党、中立、倾向民主党、中度民主党、强烈民主党七个类别。此分类用于后续计算意识形态一致性。研究者首先通过让GPT-4对已知分类的标题进行编码来评估其性能，达到了88%的评分者间一致性和0.78的Cohen's Kappa系数。
- **新闻标题主题分类:** 使用GPT-4对新闻标题的主题进行分类，包括是否与政治议题、COVID-19或一般健康相关。具体做法是向GPT-4提出针对性的问题，如“该标题是否与政治议题或讨论相关？”，并要求以0（否）或1（是）作答。

复杂网络方法：无

欢迎交流