UTS UNIVERSITY OF TECHNOLOGY SYDNEY Behavioral Data Science EPFL

The effectiveness of moderating harmful online content

created by DALL-E 3,
prompt “The effectiveness of moderating harmful online content”

Philipp Schneider
Risk Analytics and Optimization @ EPFL
philipp.schneider@epfl.ch

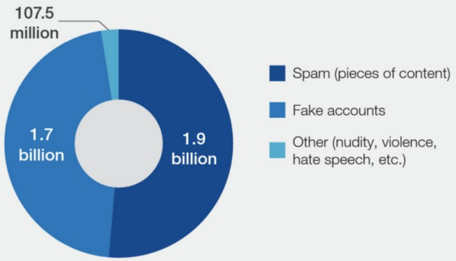
Dr Marian-Andrei Rizoiu
Behavioral Data Science @ UTS
marian-andrei.rizoiu@uts.edu.au

Harmful content in numbers: Trends and statistics

Billions of harmful activities

Facebook content removed or covered*

These figures are for the first quarter of 2020, the most recent available data.

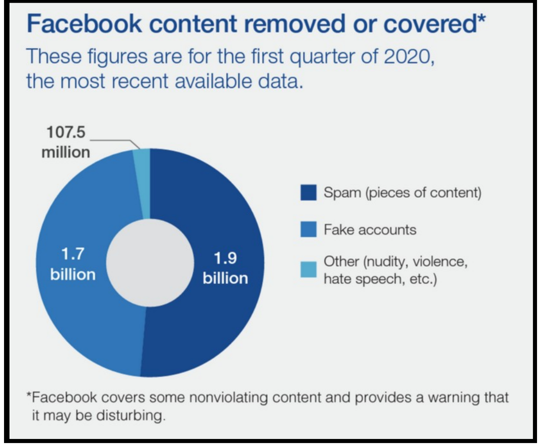


*Facebook covers some nonviolating content and provides a warning that it may be disturbing.

[1]

Harmful content in numbers: Trends and statistics

Billions of harmful activities

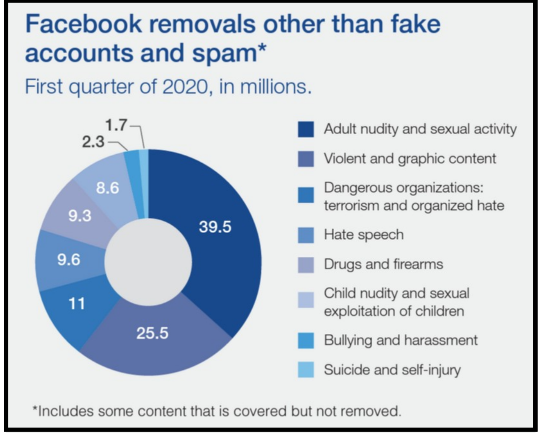

Facebook content removed or covered*

These figures are for the first quarter of 2020, the most recent available data.

*Facebook covers some nonviolating content and provides a warning that it may be disturbing.

[1]

Diverse spectrum of harmful content categories


Facebook removals other than fake accounts and spam*

First quarter of 2020, in millions.

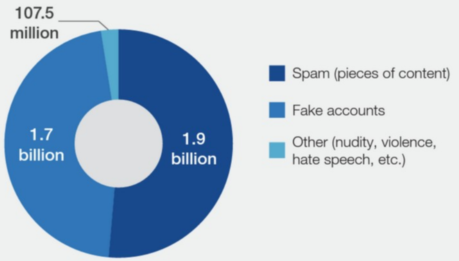
*Includes some content that is covered but not removed.

Harmful content in numbers: Trends and statistics

Billions of harmful activities

Facebook content removed or covered*

These figures are for the first quarter of 2020, the most recent available data.



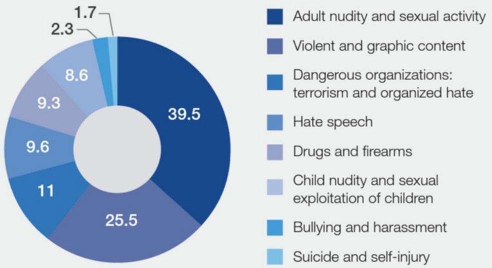
*Facebook covers some nonviolating content and provides a warning that it may be disturbing.

[1]

Diverse spectrum of harmful content categories

Facebook removals other than fake accounts and spam*

First quarter of 2020, in millions.

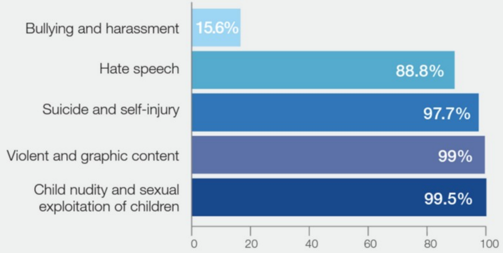


*Includes some content that is covered but not removed.

Algorithms are increasingly vital for detection

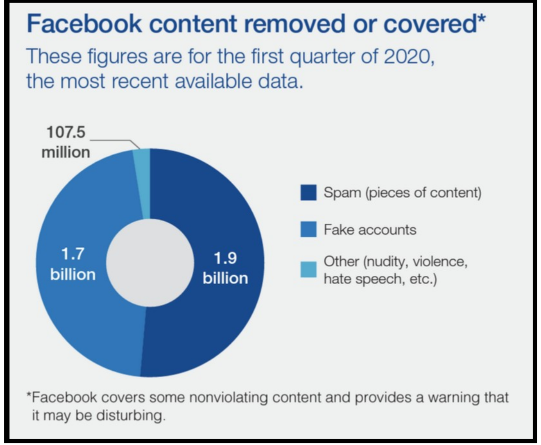
Heavy reliance on artificial intelligence

Percentage of content removed or covered that was flagged by Facebook AI technology before any users reported it (first quarter of 2020).



Harmful content in numbers: Trends and statistics

Billions of harmful activities



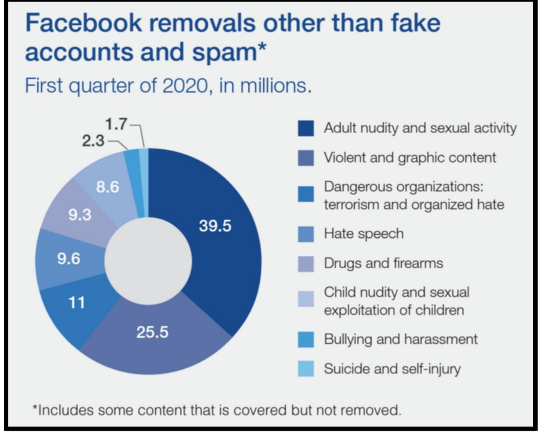
Facebook content removed or covered*

These figures are for the first quarter of 2020, the most recent available data.

*Facebook covers some nonviolating content and provides a warning that it may be disturbing.

[1]

Diverse spectrum of harmful content categories

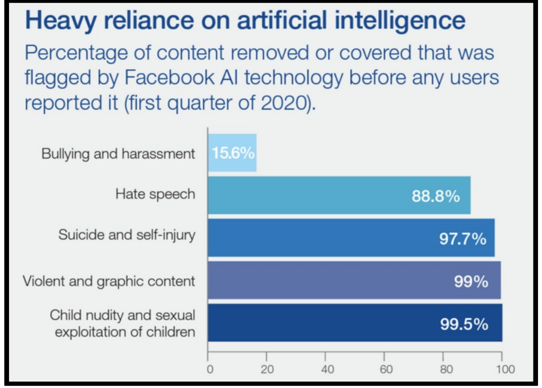


Facebook removals other than fake accounts and spam*

First quarter of 2020, in millions.

*Includes some content that is covered but not removed.

Algorithms are increasingly vital for detection



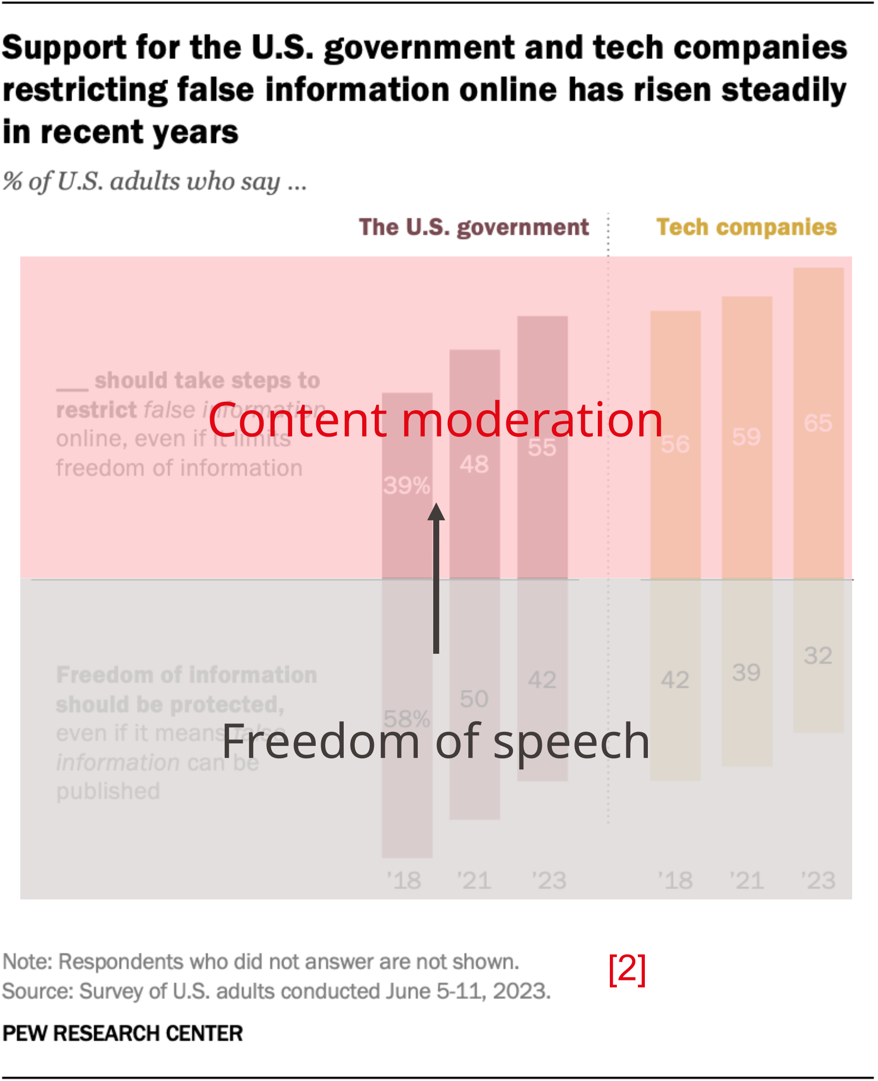
Heavy reliance on artificial intelligence

Percentage of content removed or covered that was flagged by Facebook AI technology before any users reported it (first quarter of 2020).

Harmful content

- **Misinformation:** Dissemination of false or inaccurate information without proper knowledge or verification
- **Disinformation:** Intentionally created with the aim of misleading and disseminating false information (subclass: Illegal content)

Society's perception of content moderation is evolving …



Support for the U.S. government and tech companies restricting false information online has risen steadily in recent years

% of U.S. adults who say …

Note: Respondents who did not answer are not shown.
Source: Survey of U.S. adults conducted June 5-11, 2023.

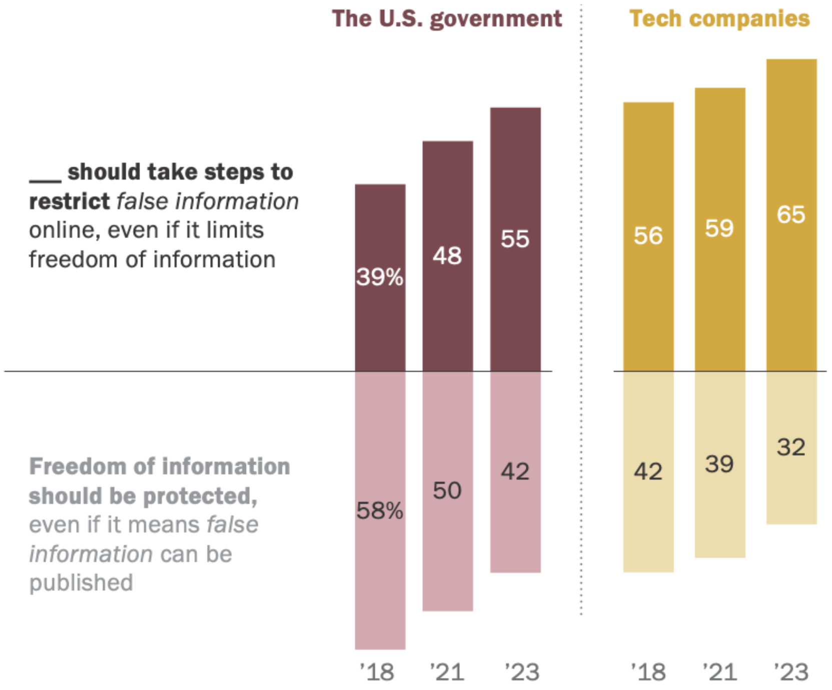
PEW RESEARCH CENTER

[2]

Society’s perception of content moderation is evolving …

Support for the U.S. government and tech companies restricting false information online has risen steadily in recent years

% of U.S. adults who say …

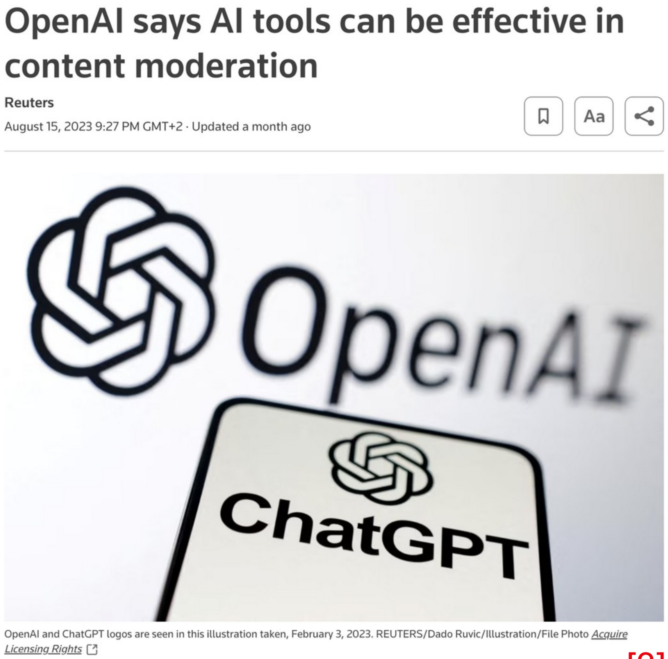


Note: Respondents who did not answer are not shown.
Source: Survey of U.S. adults conducted June 5-11, 2023.

[2]

PEW RESEARCH CENTER

Generative AI’s impact on misinformation: Solution or problem?



OpenAI says AI tools can be effective in content moderation

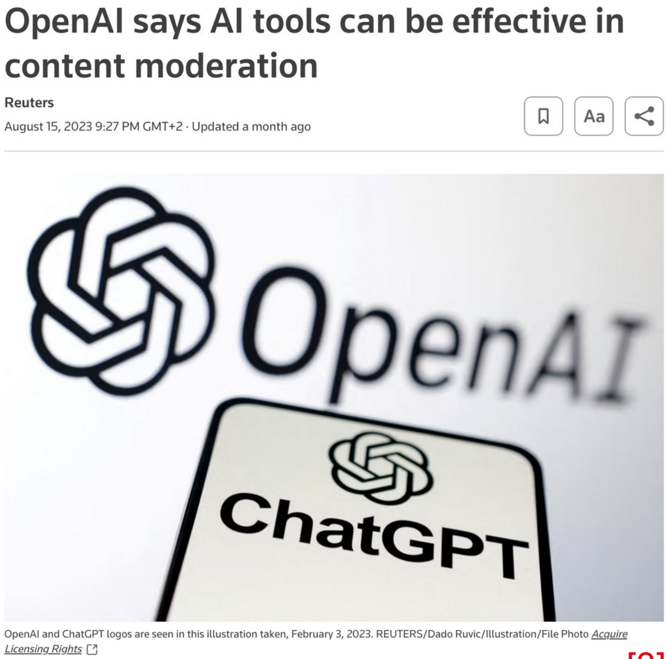
Reuters

August 15, 2023 9:27 PM GMT+2 · Updated a month ago

OpenAI and ChatGPT logos are seen in this illustration taken, February 3, 2023. REUTERS/Dado Ruvic/Illustration/File Photo *Acquire Licensing Rights*

[3]

Generative AI's impact on misinformation: Solution or problem?



OpenAI says AI tools can be effective in content moderation

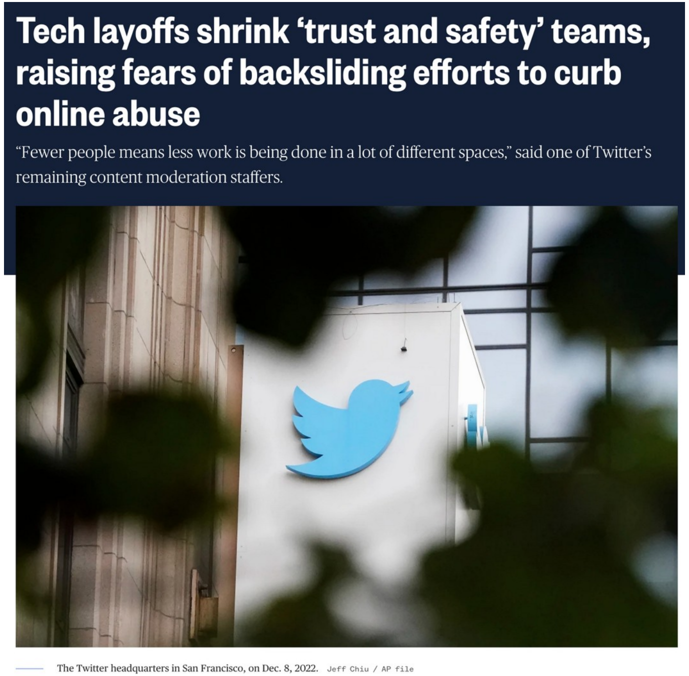
Reuters

August 15, 2023 9:27 PM GMT+2 · Updated a month ago

OpenAI and ChatGPT logos are seen in this illustration taken, February 3, 2023. REUTERS/Dado Ruvic/Illustration/File Photo *Acquire Licensing Rights*

[3]

Tech firms leverage generative AI to combat misinformation, driving workforce adaptations.



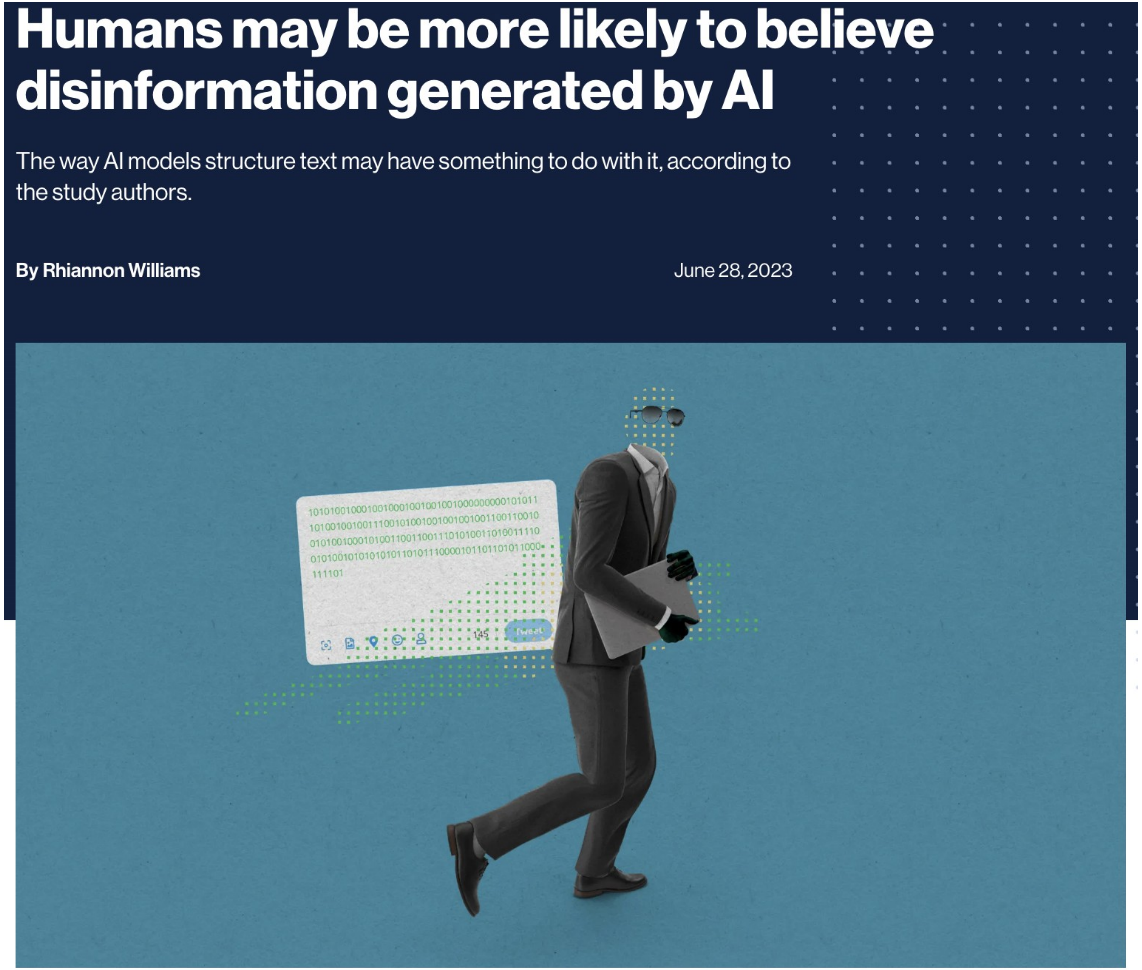
Tech layoffs shrink 'trust and safety' teams, raising fears of backsliding efforts to curb online abuse

"Fewer people means less work is being done in a lot of different spaces," said one of Twitter's remaining content moderation staffers.

The Twitter headquarters in San Francisco, on Dec. 8, 2022. Jeff Chiu / AP file

[4]

Generative AI's impact on misinformation: Solution or problem?



Humans may be more likely to believe disinformation generated by AI

The way AI models structure text may have something to do with it, according to the study authors.

By Rhiannon Williams

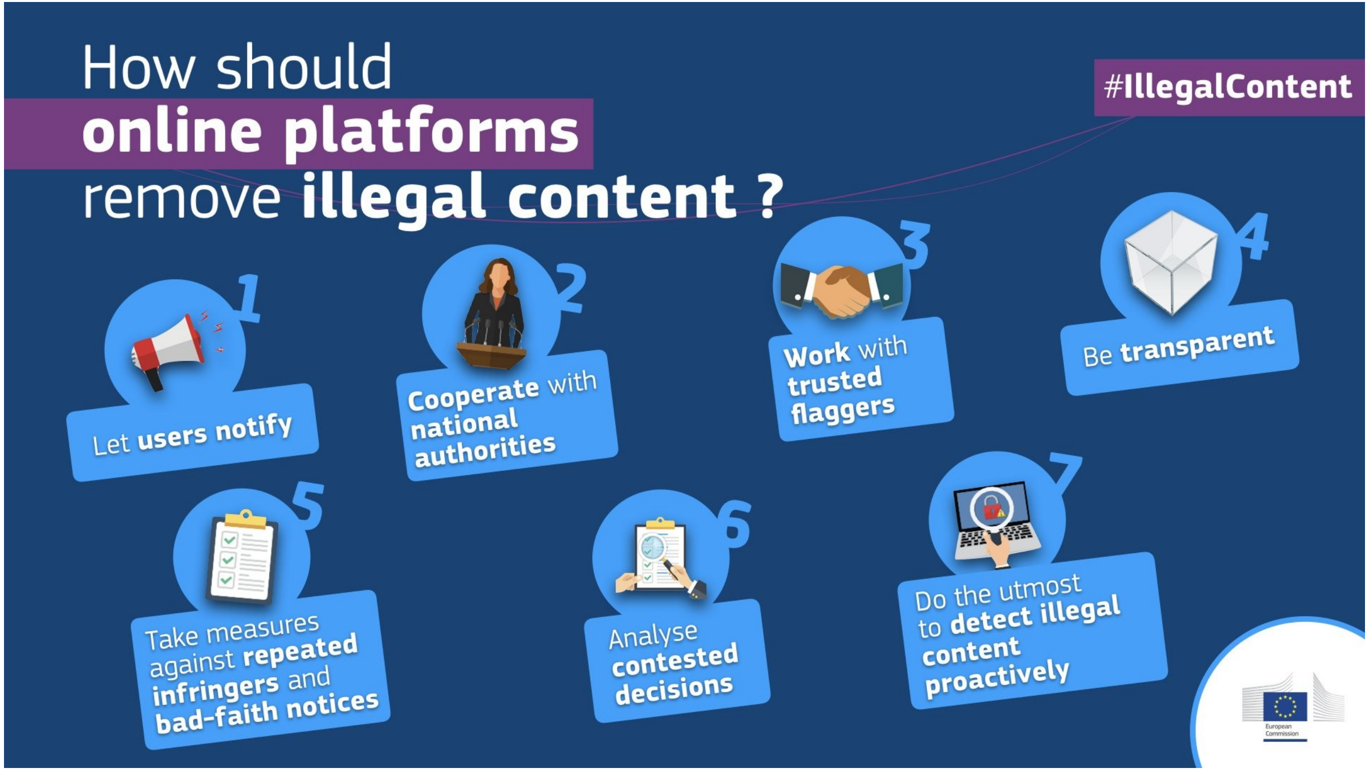
June 28, 2023

STEPHANIE ARNETT/MITTR | ENVATO

[5]

Researchers demonstrate that generative AI (GPT) is capable of generating more persuasive disinformation. [6]

EU's content removal strategy incorporates a human element



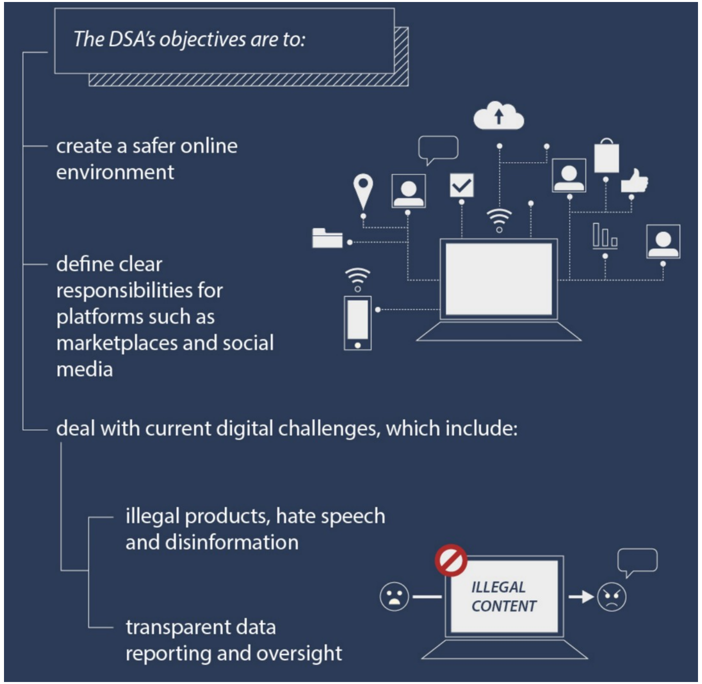
How should **online platforms** remove **illegal content ?**

#IllegalContent

1. Let **users notify**
2. **Cooperate** with **national authorities**
3. **Work** with **trusted flaggers**
4. Be **transparent**
5. Take measures against **repeated infringers** and **bad-faith notices**
6. Analyse **contested decisions**
7. Do the utmost to **detect illegal content proactively**

@DigitalEU

Digital Services Act – Content moderation



The DSA's objectives are to:

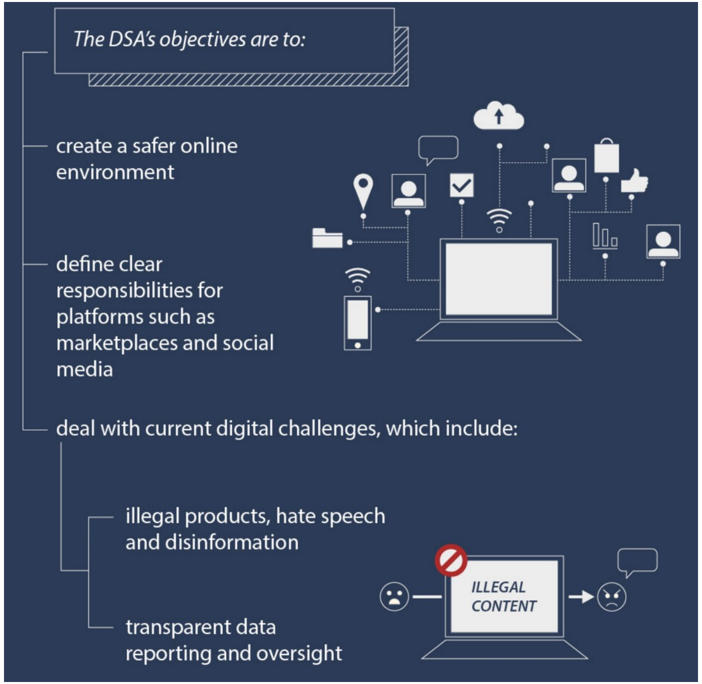
- create a safer online environment
- define clear responsibilities for platforms such as marketplaces and social media
- deal with current digital challenges, which include:
  - illegal products, hate speech and disinformation
  - transparent data reporting and oversight

@DSA-Infographic

Trusted flagger mechanism [7]

"Trusted flagger" (officially appointed entity) reporting problematic content to platforms, who must then remove it within 24 hours.

Digital Services Act – Content moderation



The DSA's objectives are to:

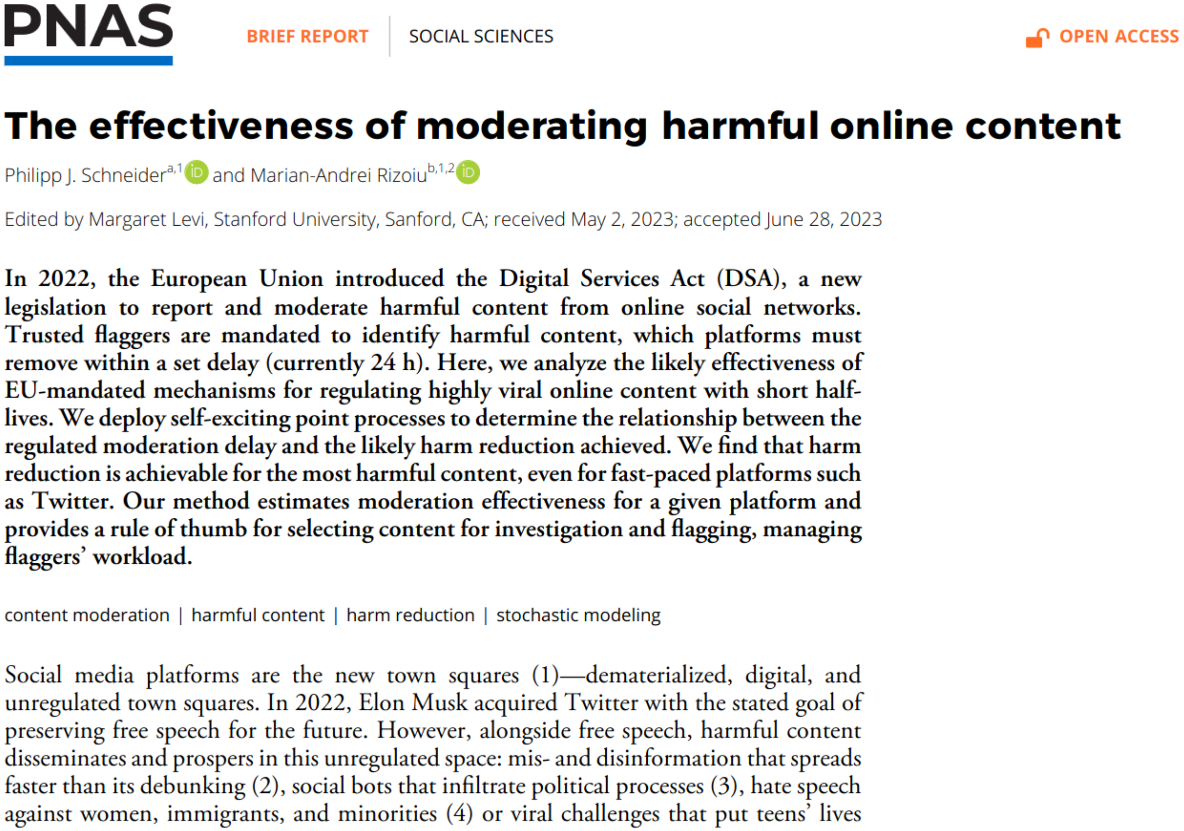
- create a safer online environment
- define clear responsibilities for platforms such as marketplaces and social media
- deal with current digital challenges, which include:
  - illegal products, hate speech and disinformation
  - transparent data reporting and oversight

@DSA-Infographic

Trusted flagger mechanism [7]

"Trusted flagger" (officially appointed entity) reporting problematic content to platforms, who must then remove it within 24 hours.

Can human moderators ever really rein in harmful online content?



PNAS BRIEF REPORT | SOCIAL SCIENCES OPEN ACCESS

The effectiveness of moderating harmful online content

Philipp J. Schneider[a,1] and Marian-Andrei Rizoiu[b,1,2]

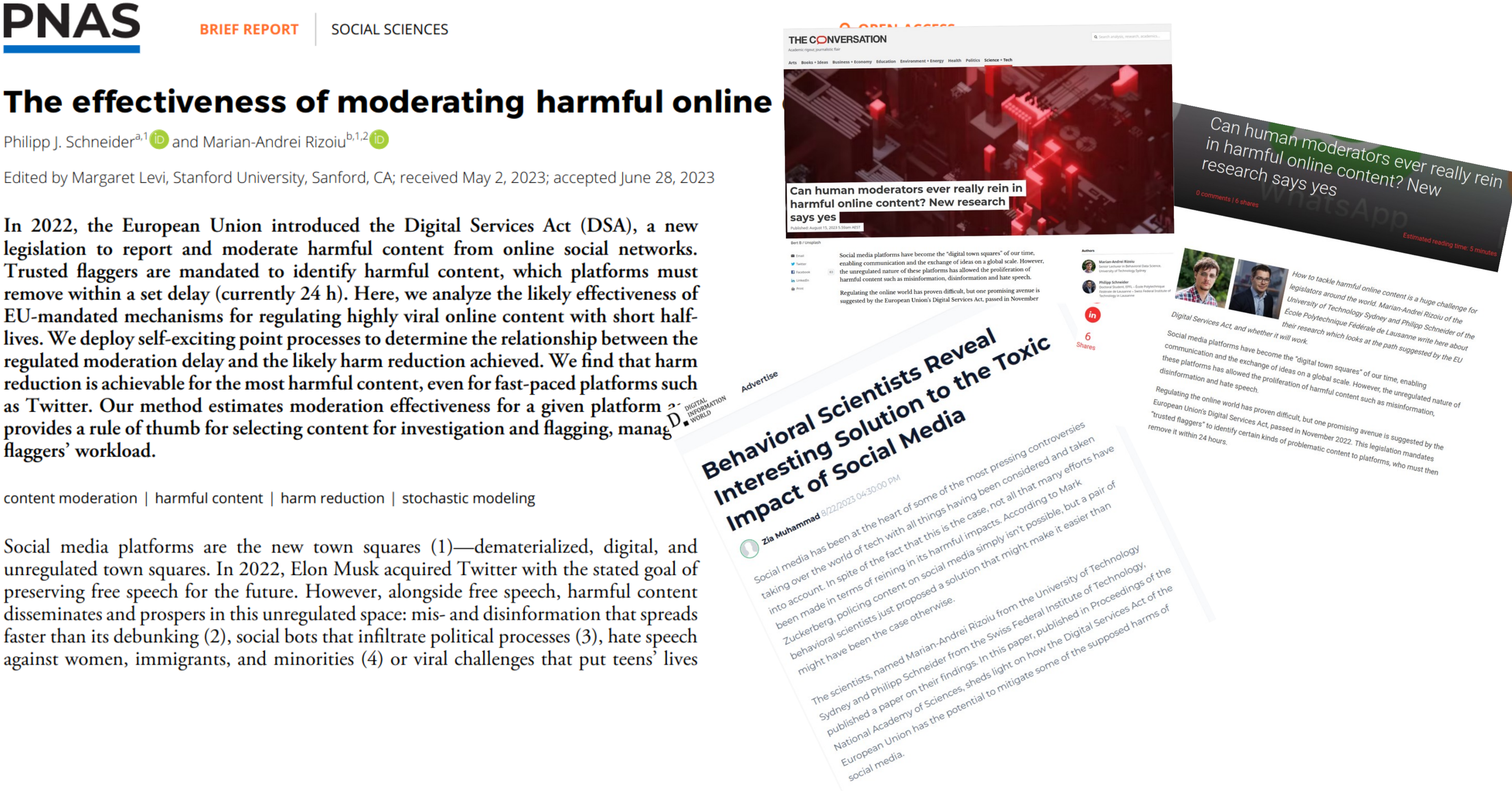
Edited by Margaret Levi, Stanford University, Sanford, CA; received May 2, 2023; accepted June 28, 2023

In 2022, the European Union introduced the Digital Services Act (DSA), a new legislation to report and moderate harmful content from online social networks. Trusted flaggers are mandated to identify harmful content, which platforms must remove within a set delay (currently 24 h). Here, we analyze the likely effectiveness of EU-mandated mechanisms for regulating highly viral online content with short half-lives. We deploy self-exciting point processes to determine the relationship between the regulated moderation delay and the likely harm reduction achieved. We find that harm reduction is achievable for the most harmful content, even for fast-paced platforms such as Twitter. Our method estimates moderation effectiveness for a given platform and provides a rule of thumb for selecting content for investigation and flagging, managing flaggers' workload.

content moderation | harmful content | harm reduction | stochastic modeling

Social media platforms are the new town squares (1)—dematerialized, digital, and unregulated town squares. In 2022, Elon Musk acquired Twitter with the stated goal of preserving free speech for the future. However, alongside free speech, harmful content disseminates and prospers in this unregulated space: mis- and disinformation that spreads faster than its debunking (2), social bots that infiltrate political processes (3), hate speech against women, immigrants, and minorities (4) or viral challenges that put teens' lives

Schneider, P. J., & Rizoiu, M.-A. (2023). The effectiveness of moderating harmful online content. Proceedings of the National Academy of Sciences, 120(34), 1–3. https://doi.org/10.1073/pnas.2307360120



PNAS

BRIEF REPORT | SOCIAL SCIENCES

The effectiveness of moderating harmful online

Philipp J. Schneider[a,1] and Marian-Andrei Rizoiu[b,1,2]

Edited by Margaret Levi, Stanford University, Sanford, CA; received May 2, 2023; accepted June 28, 2023

In 2022, the European Union introduced the Digital Services Act (DSA), a new legislation to report and moderate harmful content from online social networks. Trusted flaggers are mandated to identify harmful content, which platforms must remove within a set delay (currently 24 h). Here, we analyze the likely effectiveness of EU-mandated mechanisms for regulating highly viral online content with short half-lives. We deploy self-exciting point processes to determine the relationship between the regulated moderation delay and the likely harm reduction achieved. We find that harm reduction is achievable for the most harmful content, even for fast-paced platforms such as Twitter. Our method estimates moderation effectiveness for a given platform a provides a rule of thumb for selecting content for investigation and flagging, manag flaggers' workload.

content moderation | harmful content | harm reduction | stochastic modeling

Social media platforms are the new town squares (1)—dematerialized, digital, and unregulated town squares. In 2022, Elon Musk acquired Twitter with the stated goal of preserving free speech for the future. However, alongside free speech, harmful content disseminates and prospers in this unregulated space: mis- and disinformation that spreads faster than its debunking (2), social bots that infiltrate political processes (3), hate speech against women, immigrants, and minorities (4) or viral challenges that put teens' lives

THE CONVERSATION

Academic rigour, journalistic flair

Can human moderators ever really rein in harmful online content? New research says yes

Published: August 15, 2023 5.59am AEST

Bert B / Unsplash

Social media platforms have become the "digital town squares" of our time, enabling communication and the exchange of ideas on a global scale. However, the unregulated nature of these platforms has allowed the proliferation of harmful content such as misinformation, disinformation and hate speech.

Regulating the online world has proven difficult, but one promising avenue is suggested by the European Union's Digital Services Act, passed in November

Authors

Marian-Andrei Rizoiu
Senior Lecturer in Behavioral Data Science, University of Technology Sydney

Philipp Schneider
Doctoral Student, EPFL – École Polytechnique Fédérale de Lausanne – Swiss Federal Institute of Technology in Lausanne

Can human moderators ever really rein in harmful online content? New research says yes

0 comments | 6 shares

Estimated reading time: 5 minutes

How to tackle harmful online content is a huge challenge for legislators around the world. Marian-Andrei Rizoiu of the University of Technology Sydney and Philipp Schneider of the École Polytechnique Fédérale de Lausanne write here about their research which looks at the path suggested by the EU Digital Services Act, and whether it will work.

Social media platforms have become the "digital town squares" of our time, enabling communication and the exchange of ideas on a global scale. However, the unregulated nature of these platforms has allowed the proliferation of harmful content such as misinformation, disinformation and hate speech.

Regulating the online world has proven difficult, but one promising avenue is suggested by the European Union's Digital Services Act, passed in November 2022. This legislation mandates "trusted flaggers" to identify certain kinds of problematic content to platforms, who must then remove it within 24 hours.

Behavioral Scientists Reveal Interesting Solution to the Toxic Impact of Social Media

Zia Muhammad 8/22/2023 04:30:00 PM

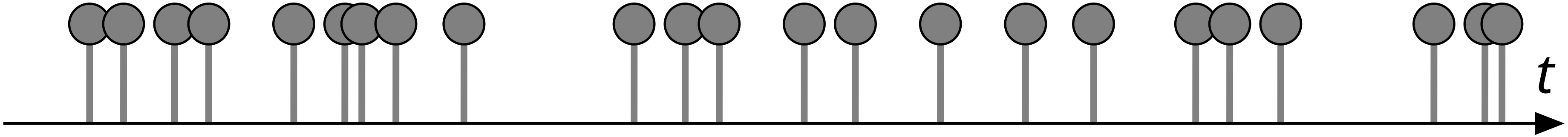
Social media has been at the heart of some of the most pressing controversies taking over the world of tech with all things having been considered and taken into account. In spite of the fact that this is the case, not all that many efforts have been made in terms of reining in its harmful impacts. According to Mark Zuckerberg, policing content on social media simply isn't possible, but a pair of behavioral scientists just proposed a solution that might make it easier than might have been the case otherwise.

The scientists, named Marian-Andrei Rizoiu from the University of Technology Sydney and Philipp Schneider from the Swiss Federal Institute of Technology, published a paper on their findings. In this paper, published in Proceedings of the National Academy of Sciences, sheds light on how the Digital Services Act of the European Union has the potential to mitigate some of the supposed harms of social media.

The effectiveness of moderating harmful online content

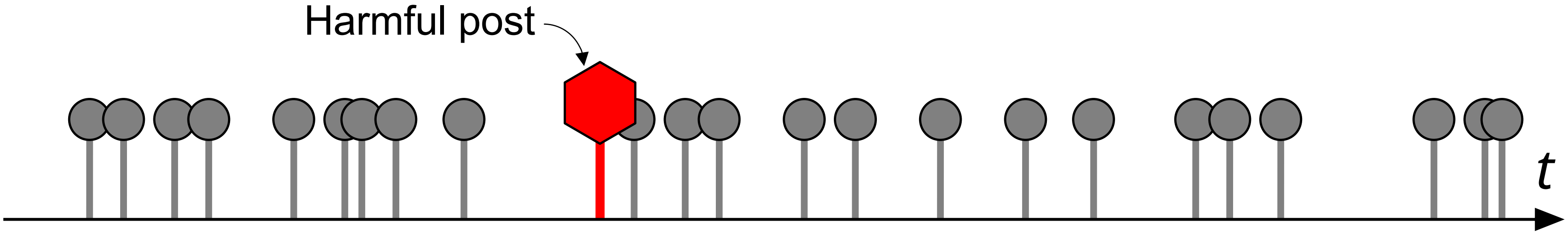
Schneider, P. J., & Rizoiu, M.-A. (2023). The effectiveness of moderating harmful online content. Proceedings of the National Academy of Sciences, 120(34), 1–3. https://doi.org/10.1073/pnas.2307360120

Content dynamics of harmful content



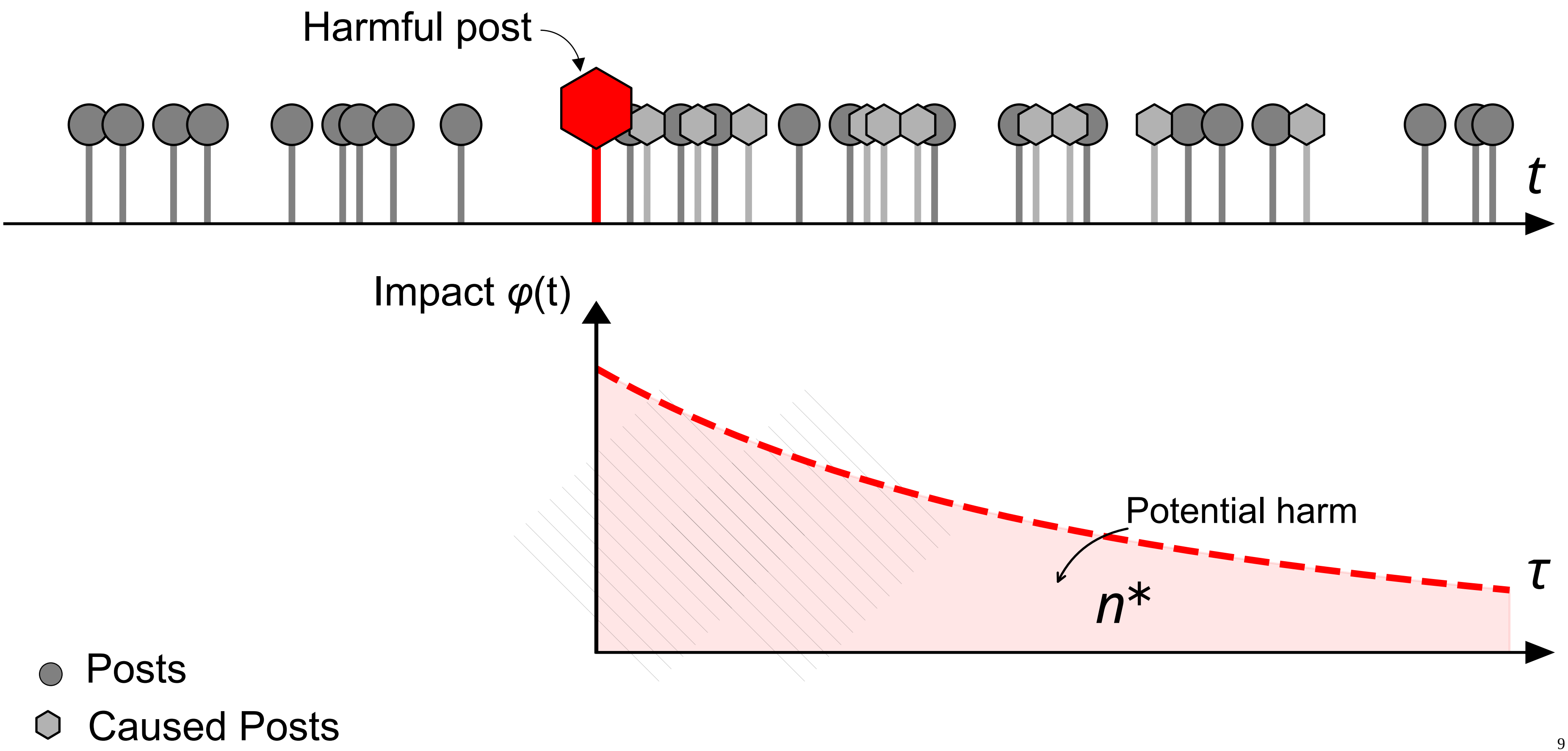
● Posts

Content dynamics of harmful content

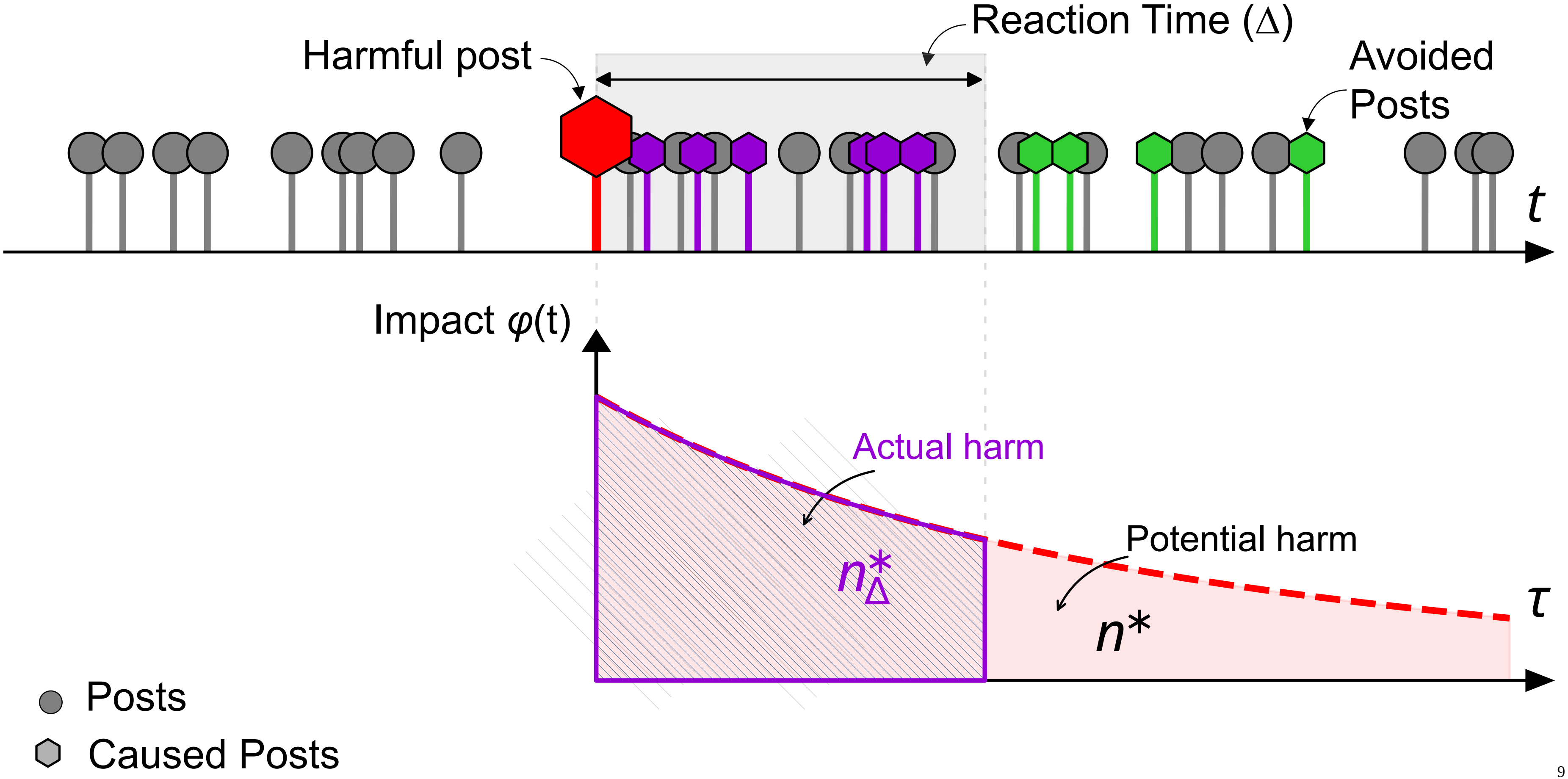


● Posts

Content dynamics of harmful content



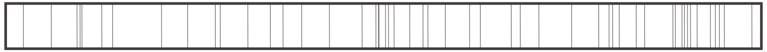
Content dynamics of harmful content



Real-world events occur in groups

Arrival Processes

Homogenous Poisson point process

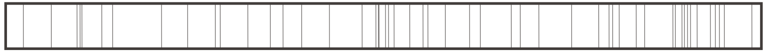


$$\lambda(t) = \mu$$

Real-world events occur in groups

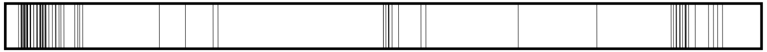
Arrival Processes

Homogenous Poisson point process



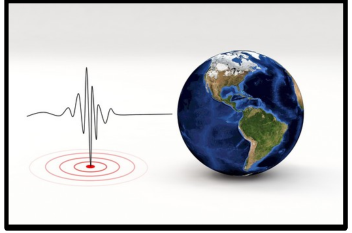
$$\lambda(t) = \mu$$

Self-exciting point process



Applications

Seismology



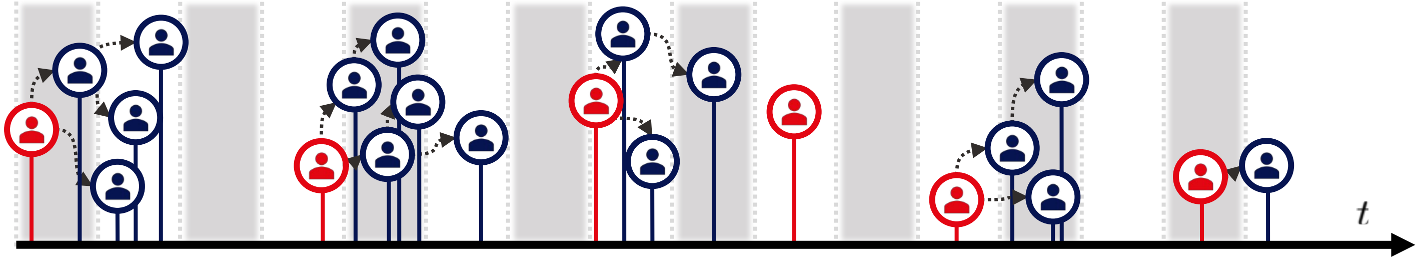
Finance

Social Media

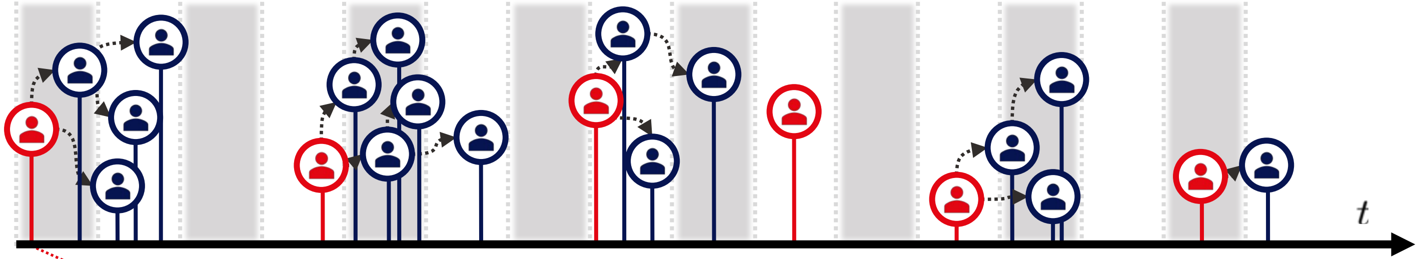
Self-excitation in social media

● Original posts

● Retweets / Likes

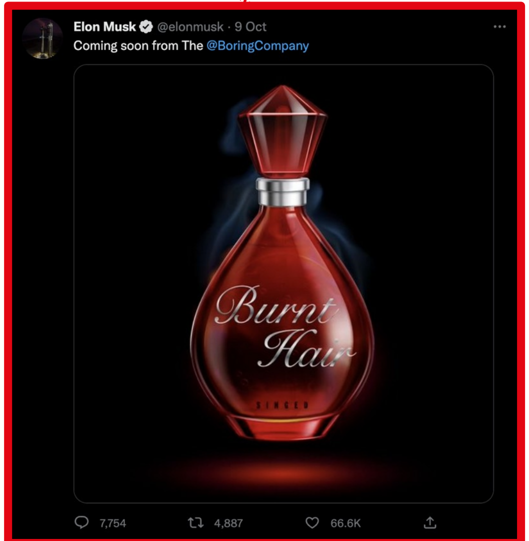


Self-excitation in social media



Original posts

Retweets / Likes



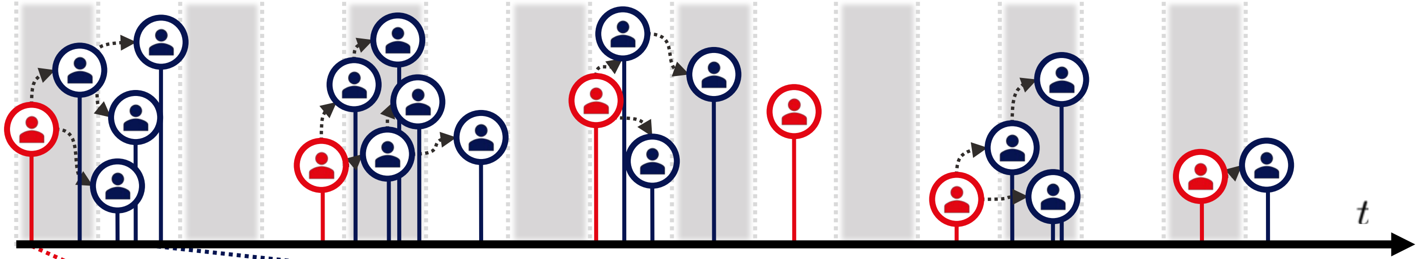
Elon Musk @elonmusk · 9 Oct

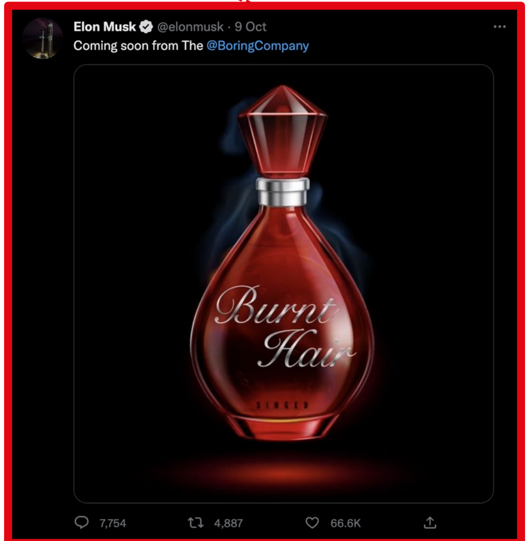
Coming soon from The @BoringCompany

Self-excitation in social media

Original posts

Retweets / Likes

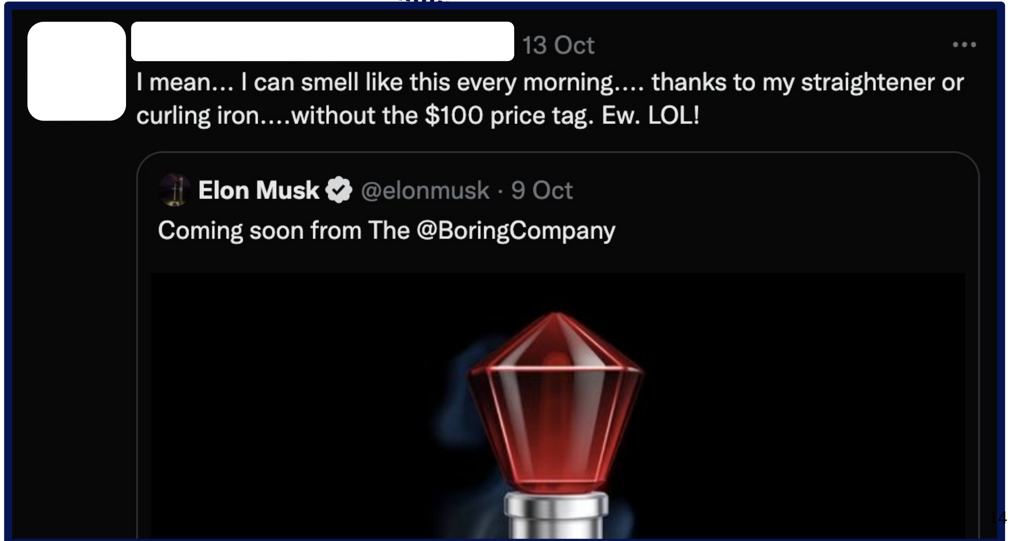




Elon Musk @elonmusk · 9 Oct

Coming soon from The @BoringCompany

7,754 4,887 66.6K



[redacted] [redacted] 13 Oct

I mean... I can smell like this every morning.... thanks to my straightener or curling iron....without the $100 price tag. Ew. LOL!

Elon Musk @elonmusk · 9 Oct

Coming soon from The @BoringCompany

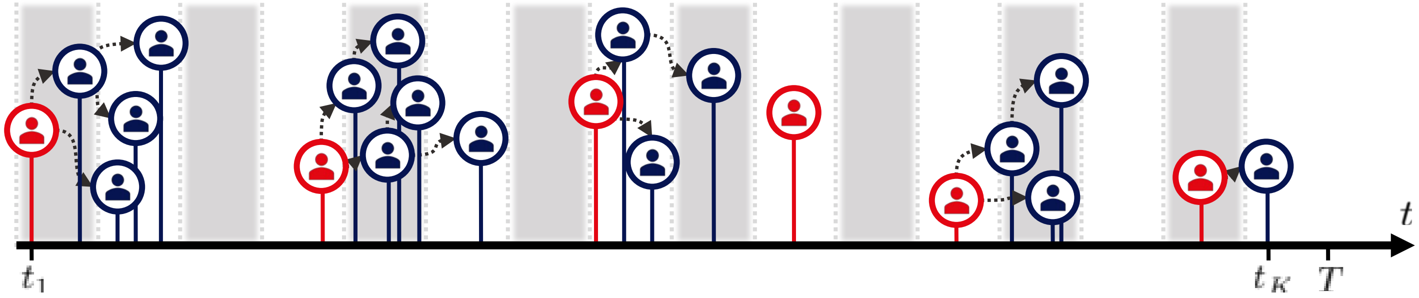
Self-excitation describes the clustering effect

Intensity function

$$\lambda(t|\mathcal{H}_t) = \overbrace{\mu}^{\text{Base intensity (exogenous)}} + \sum_{i:t_i<t} \phi(t - t_i)$$

Event history

$$\mathcal{H}_T = \{t_1, \dots, t_K\} \subset (0, T$$



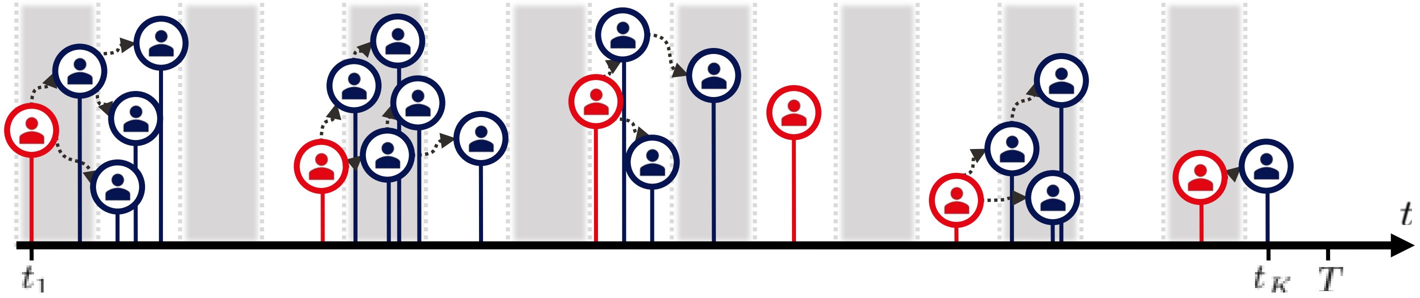
Self-excitation describes the clustering effect

Intensity function

$$\lambda(t|\mathcal{H}_t) = \overbrace{\mu}^{\substack{\text{Base intensity}\\ \text{(exogenous)}}} + \overbrace{\sum_{i:t_i<t} \phi(t-t_i)}^{\substack{\text{Self-excitation}\\ \text{(endogenous)}}}$$

Event history

$$\mathcal{H}_T = \{t_1, \dots, t_K\} \subset (0, T$$



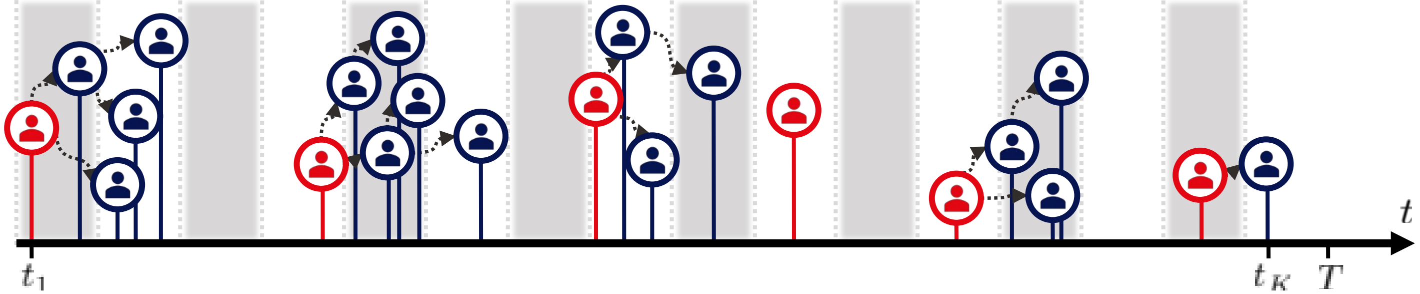
Self-excitation describes the clustering effect

Intensity function

$$\lambda(t|\mathcal{H}_t) = \overbrace{\mu}^{\substack{\text{Base intensity}\\ \text{(exogenous)}}} + \overbrace{\sum_{i:t_i<t} \underbrace{\phi(t-t_i)}_{\text{Kernel function}}}^{\substack{\text{Self-excitation}\\ \text{(endogenous)}}}$$

Event history

$$\mathcal{H}_T = \{t_1, \ldots, t_K\} \subset (0, T$$



What are the key metrics of this study?

Potential harm: Number of harmful offspring the post generates

Content half-life: Amount of time required for half of all the post’s offspring to be generated

What are the key metrics of this study?

Potential harm: Number of harmful offspring the post generates

Content half-life: Amount of time required for half of all the post’s offspring to be generated

- Twitter – 24 min
- Facebook – 105 min
- Instagram – 20 h
- LinkedIn – 24 h
- YouTube – 8.8 d
- Pinterest – 3.75 mo

[8]

What are the key metrics of this study?

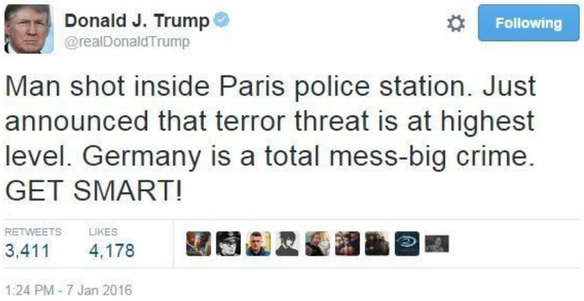
Potential harm: Number of harmful offspring the post generates

Content half-life: Amount of time required for half of all the post’s offspring to be generated

- Twitter – 24 min
- Facebook – 105 min
- Instagram – 20 h
- LinkedIn – 24 h
- YouTube – 8.8 d
- Pinterest – 3.75 mo

[8]

News around terrorist attack



Donald J. Trump
@realDonaldTrump

Man shot inside Paris police station. Just announced that terror threat is at highest level. Germany is a total mess-big crime. GET SMART!

RETWEETS 3,411 LIKES 4,178

1:24 PM - 7 Jan 2016

@BBC

- High potential harm / virality
- Short content half-life

What are the key metrics of this study?

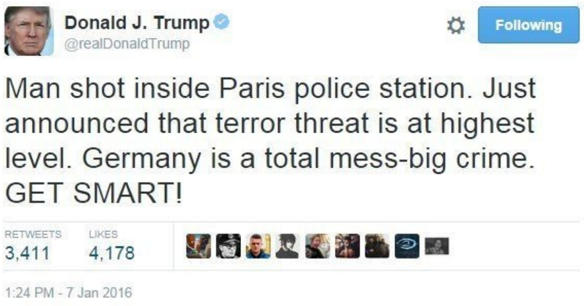
Potential harm: Number of harmful offspring the post generates

Content half-life: Amount of time required for half of all the post's offspring to be generated

- Twitter – 24 min
- Facebook – 105 min
- Instagram – 20 h
- LinkedIn – 24 h
- YouTube – 8.8 d
- Pinterest – 3.75 mo

[8]

News around terrorist attack



Donald J. Trump
@realDonaldTrump

Following

Man shot inside Paris police station. Just announced that terror threat is at highest level. Germany is a total mess-big crime. GET SMART!

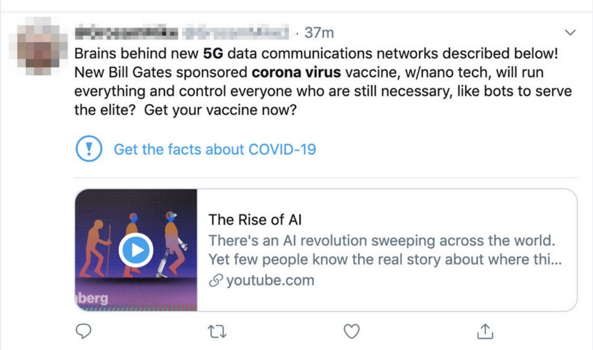
RETWEETS 3,411 LIKES 4,178

1:24 PM - 7 Jan 2016

@BBC

- High potential harm / virality
- Short content half-life

Anti-vaccine conspiracies



[redacted] · 37m

Brains behind new **5G** data communications networks described below! New Bill Gates sponsored **corona virus** vaccine, w/nano tech, will run everything and control everyone who are still necessary, like bots to serve the elite? Get your vaccine now?

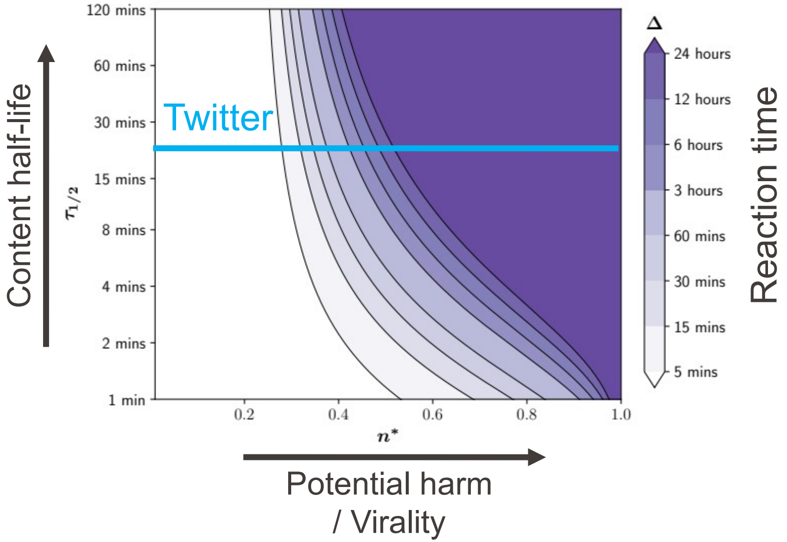
Get the facts about COVID-19

The Rise of AI
There's an AI revolution sweeping across the world. Yet few people know the real story about where thi...
youtube.com

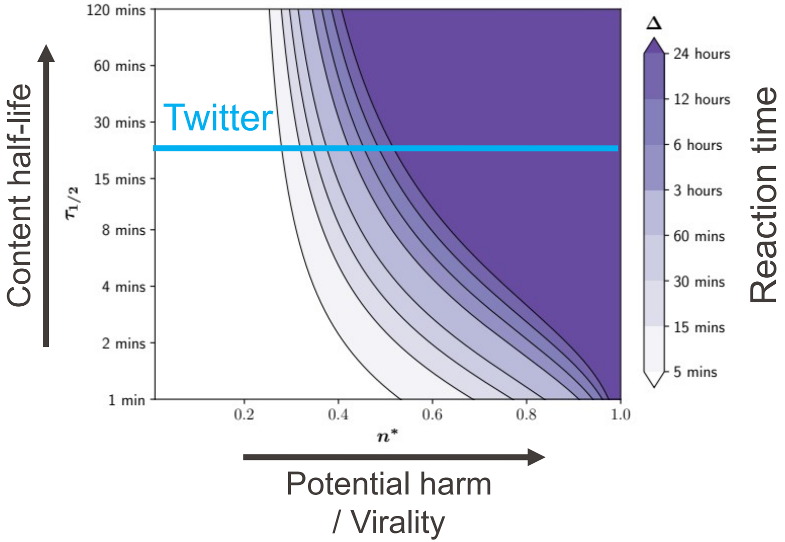
@Vox

- Low potential harm / virality (before COVID-19)
- Long content half-life

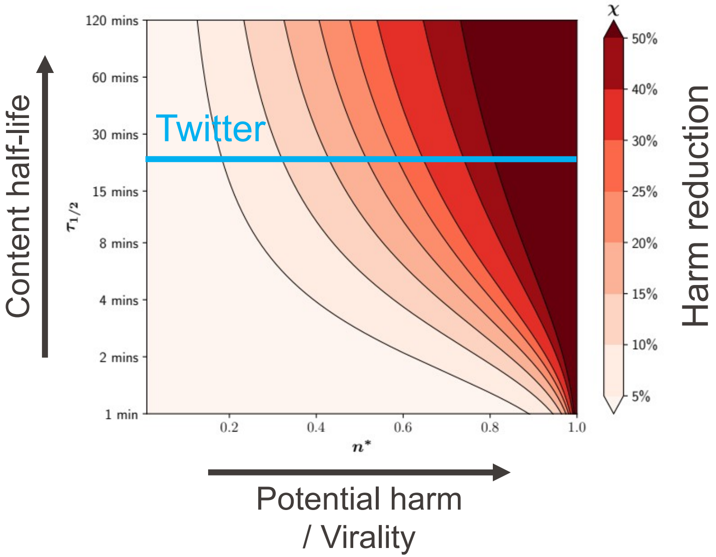
What is the **reaction time** to obtain **20% harm reduction**?



What is the **reaction time** to obtain **20% harm reduction**?



What is the **achieved harm reduction** when **removing** content after **24 hours?**



Application to real-world discussions

Twitter datasets (1 July to 31 December 2022)

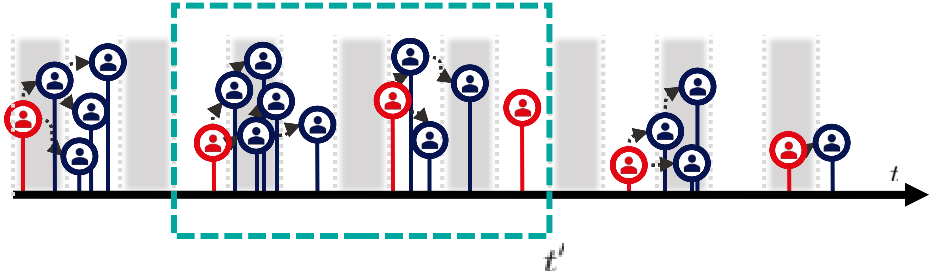
- #climatescam (479,051 posts) – Controversial opinions regarding climate change [9]
- #americafirst or #americansfirst (278,899 posts) – Debates over key US political topics such as immigration and foreign policies [10]

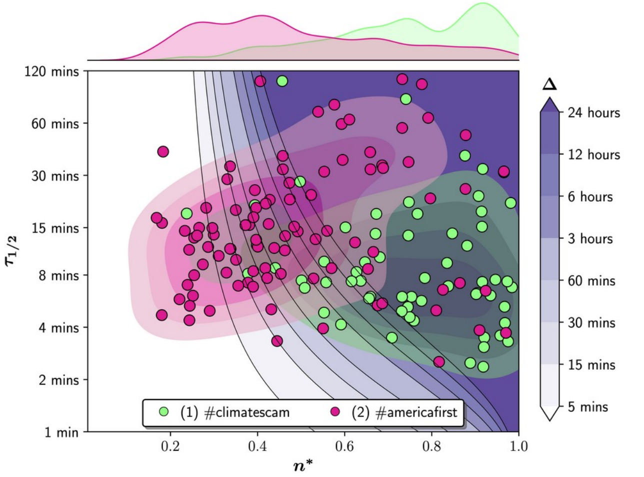
Application to real-world discussions

Twitter datasets (1 July to 31 December 2022)

- #climatescam (479,051 posts) – Controversial opinions regarding climate change [9]
- #americafirst or #americansfirst (278,899 posts) – Debates over key US political topics such as immigration and foreign policies [10]

‘On-the-fly’ (real-time) parameter estimates



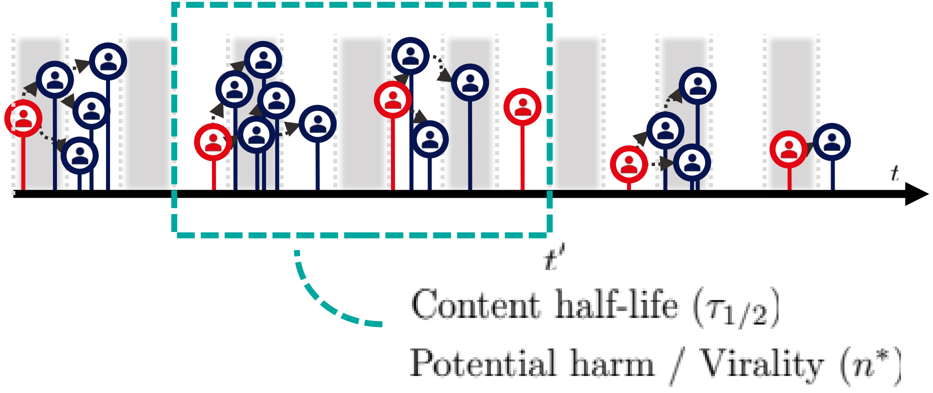


Application to real-world discussions

Twitter datasets (1 July to 31 December 2022)

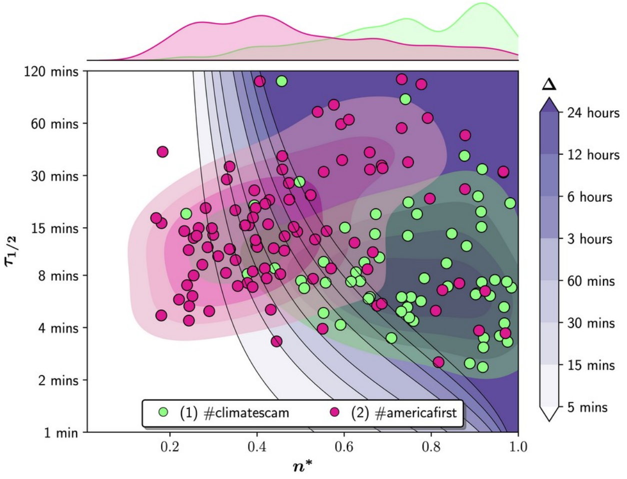
- #climatescam (479,051 posts) – Controversial opinions regarding climate change [9]
- #americafirst or #americansfirst (278,899 posts) – Debates over key US political topics such as immigration and foreign policies [10]

‘On-the-fly’ (real-time) parameter estimates



Content half-life ($\tau_{1/2}$)

Potential harm / Virality (n^*)

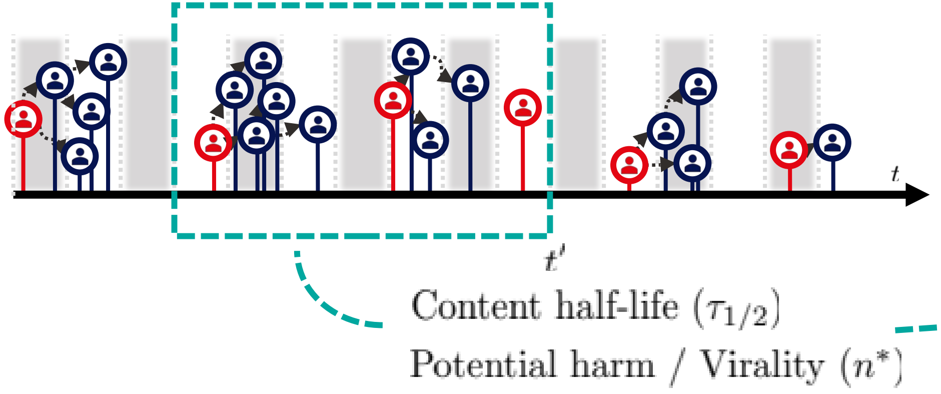


Application to real-world discussions

Twitter datasets (1 July to 31 December 2022)

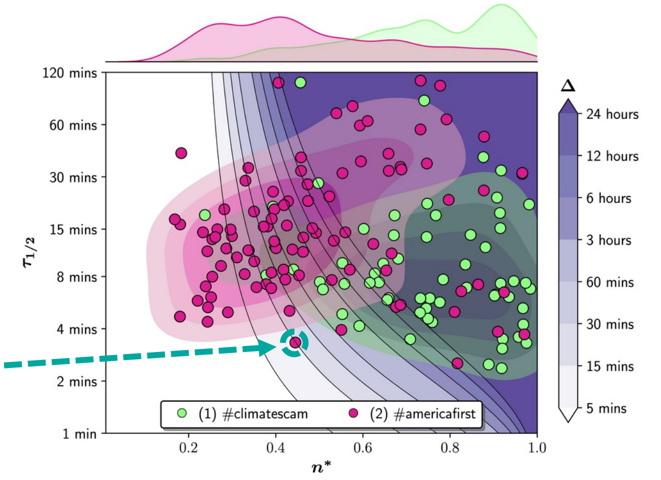
- #climatescam (479,051 posts) – Controversial opinions regarding climate change [9]
- #americafirst or #americansfirst (278,899 posts) – Debates over key US political topics such as immigration and foreign policies [10]

‘On-the-fly’ (real-time) parameter estimates



Content half-life ($\tau_{1/2}$)

Potential harm / Virality (n^*)

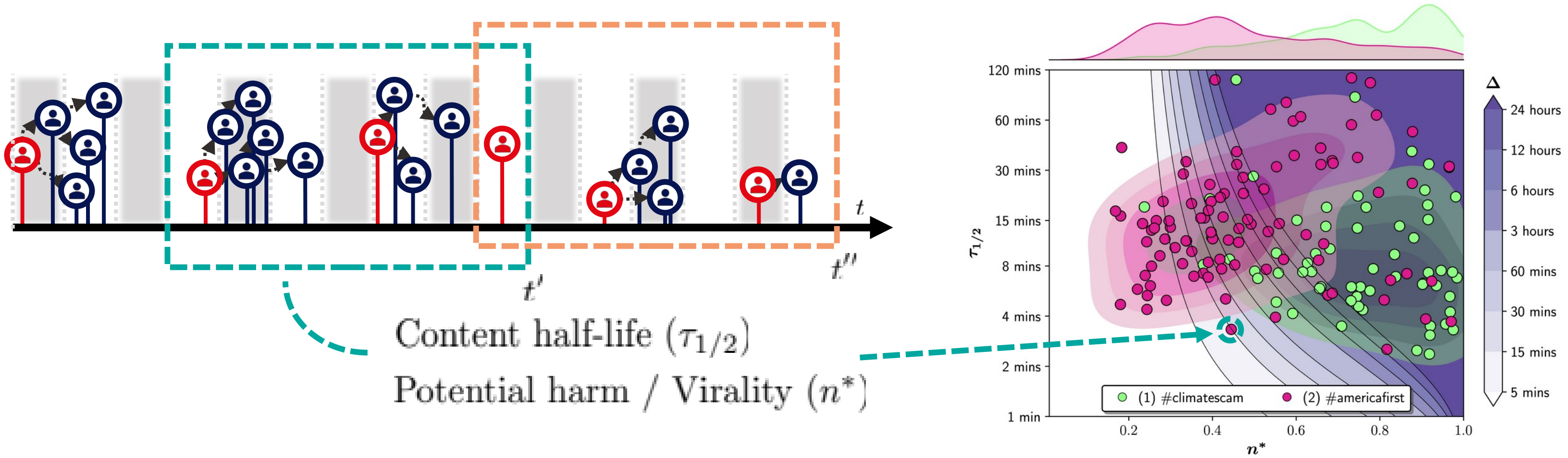


Application to real-world discussions

Twitter datasets (1 July to 31 December 2022)

- #climatescam (479,051 posts) – Controversial opinions regarding climate change [9]
- #americafirst or #americansfirst (278,899 posts) – Debates over key US political topics such as immigration and foreign policies [10]

'On-the-fly' (real-time) parameter estimates

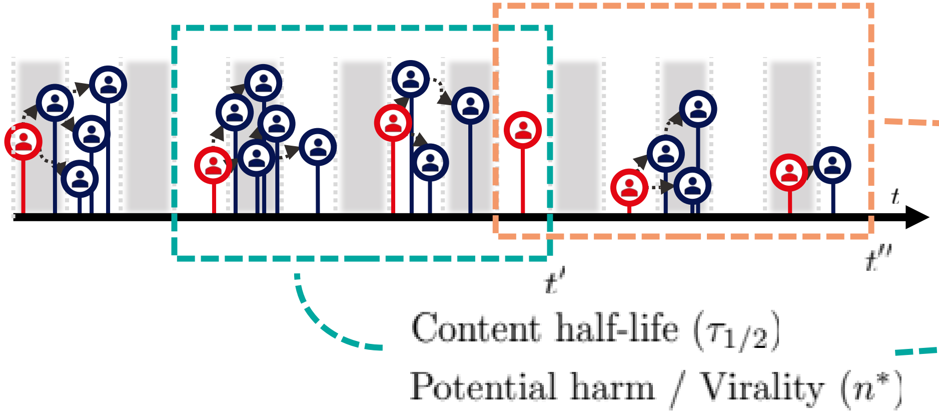


Application to real-world discussions

Twitter datasets (1 July to 31 December 2022)

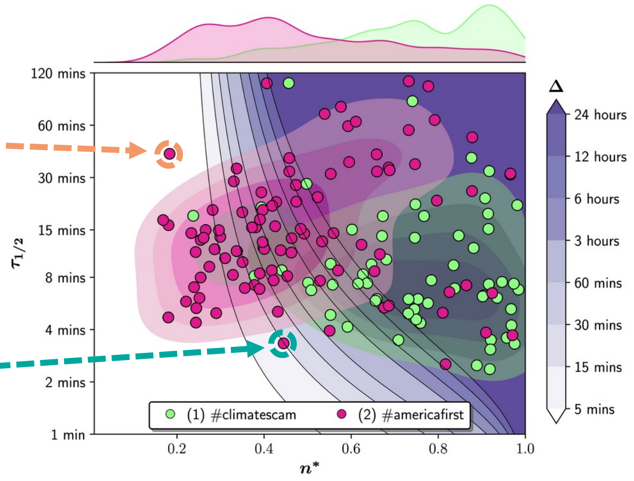
- #climatescam (479,051 posts) – Controversial opinions regarding climate change [9]
- #americafirst or #americansfirst (278,899 posts) – Debates over key US political topics such as immigration and foreign policies [10]

'On-the-fly' (real-time) parameter estimates

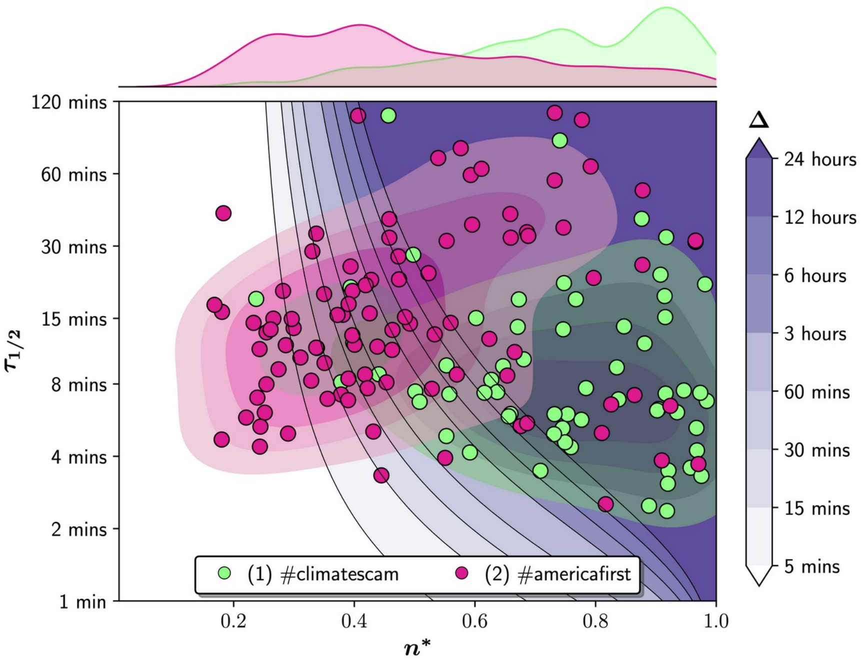


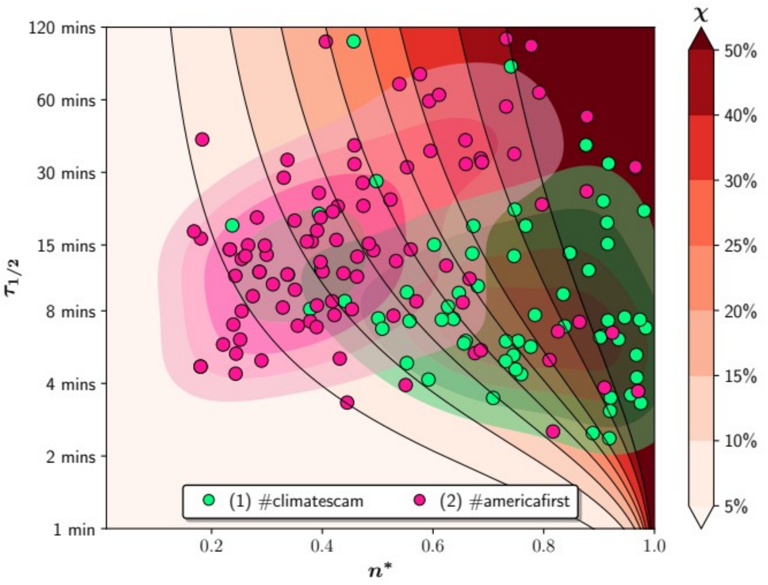
Content half-life ($\tau_{1/2}$)

Potential harm / Virality (n^*)



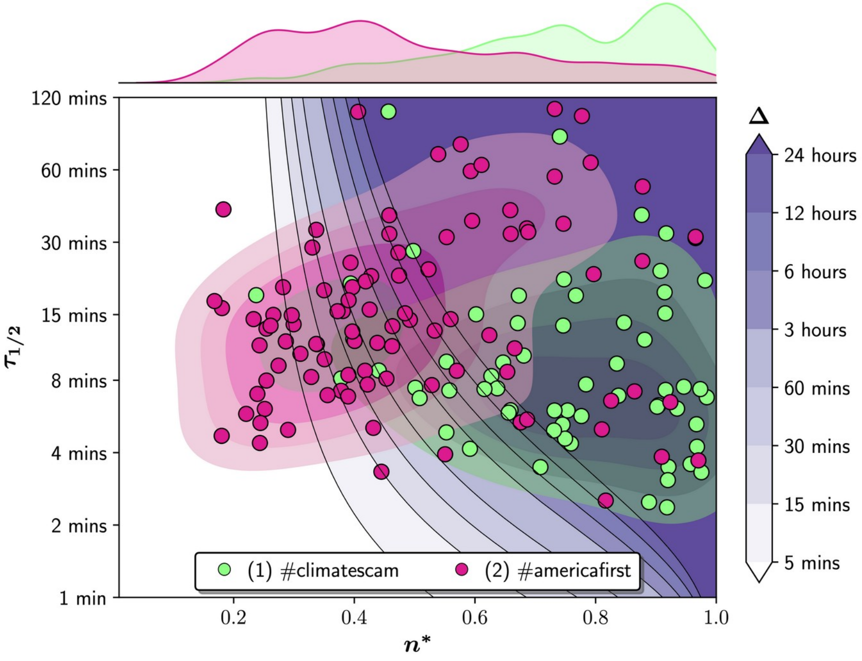
Effectiveness of EU-regulated moderation

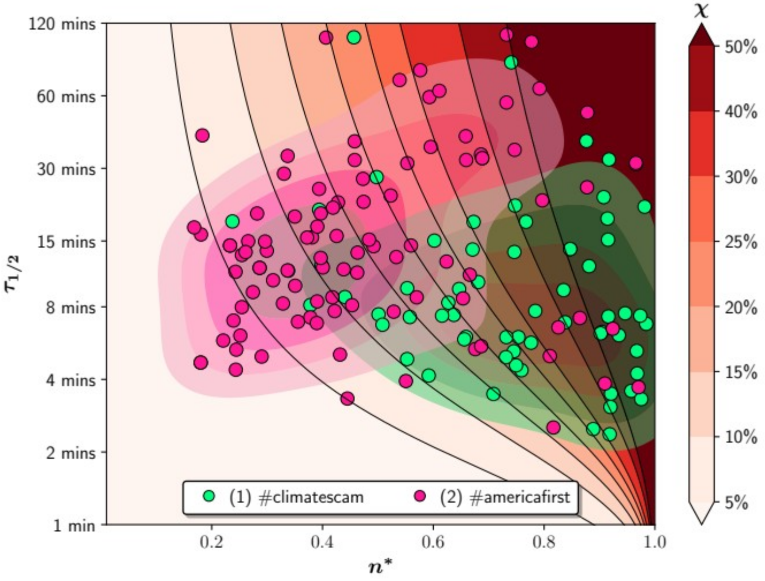




- Real-world potentially problematic content exhibits widely highly variable dynamics
- Harm reduction via manual flagging efforts is achievable

Effectiveness of EU-regulated moderation



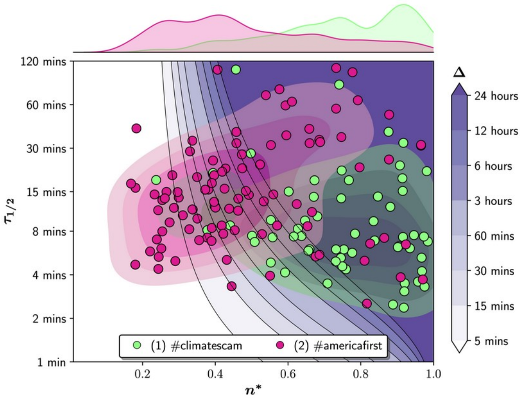


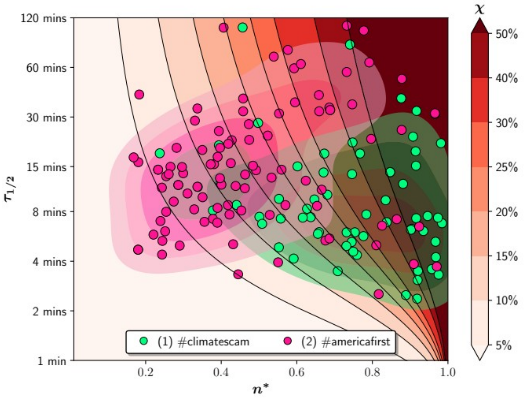
- Real-world potentially problematic content exhibits widely highly variable dynamics
- Harm reduction via manual flagging efforts is achievable

Topics	Potential harm / Virality	Content half-life	Harm reduction
#climatescam	0.75	7.48 min	29.18%
#americafirst	0.44	13.97 min	13.29%

Conclusion

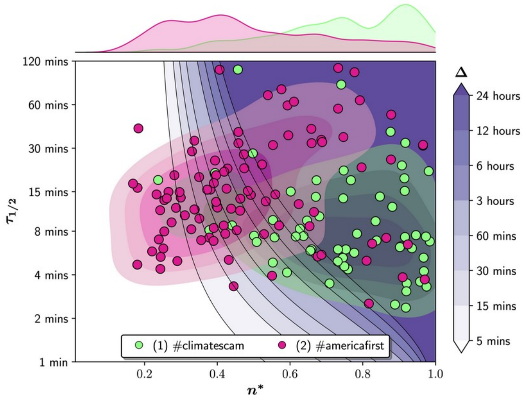
- Harm reduction is achievable with manual flagging efforts, even for fast-paced platforms such as Twitter
- Framework for policymakers to draft mechanisms for content moderation by indicating where to focus human fact-checking efforts and how quickly to react

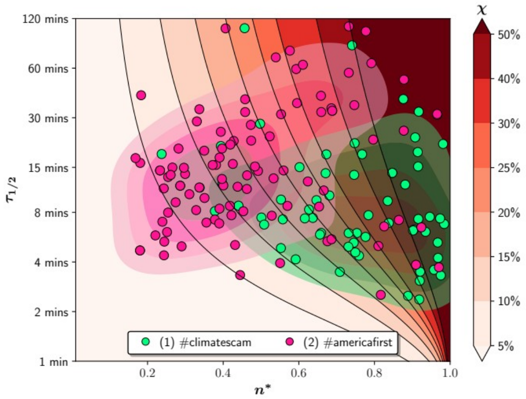




Conclusion

- Harm reduction is achievable with manual flagging efforts, even for fast-paced platforms such as Twitter
- Framework for policymakers to draft mechanisms for content moderation by indicating where to focus human fact-checking efforts and how quickly to react





Future work / Open questions

- How to select discussion topics?
- What is the 'actual' reaction time?
- What metrics measures the effectiveness in more granularity?
  - Post-based potential harm?

References

- [1] P. M. Barrett, *Who Moderates the Social Media Giants?* (NYU Stern Center for Business & HumanRights, 2020).
- [2] C. St. Aubin, J. Liedke, *Most Americans favor restrictions on false information, violent content online*. Pew Research Center (2023). https://www.pewresearch.org/short-reads/2023/07/20/most-americans-favor-restrictions-on-false-information-violent-content-online/ (Accessed 27 September 2023).
- [3] J. Singh, *OpenAI says AI tools can be effective in content moderation*. Reuters (2023). https://www.reuters.com/technology/openai-says-ai-tools-can-be-effective-content-moderation-2023-08-15/ (Accessed 27 September 2023).
- [4] J. J. McCorvey, *Tech layoffs hit “trust and safety” teams, raising fears of backsliding efforts to curb online abuse*. NBC News (2023). https://www.nbcnews.com/tech/tech-news/tech-layoffs-hit-trust-safety-teams-raising-fears-backsliding-efforts-rcna69111 (Accessed 27 September 2023).
- [5] R. Williams, *Humans may be more likely to believe disinformation generated by AI*. MIT Technology Review (2023). https://www.technologyreview.com/2023/06/28/1075683/humans-may-be-more-likely-to-believe-disinformation-generated-by-ai/ (Accessed 27 September 2023).
- [6] G. Spitale, N. Biller-Andorno, F. Germani, *AI Model GPT-3 (dis)informs us better than humans*. Science Advances, 9(26), eadh1850 (2023).
- [7] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act).OJ L277, 1–102 (2022).
- [8] S. M. Graffius, *Lifespan (half-life) of social media posts: Update for 2023* (2023). https://dx.doi.org/10.13140/RG.2.2.19783.98722 (Accessed 31 July 2023).
- [9] O. Milman, *#ClimateScam: Denialism claims flooding Twitter have scientists worried*. Guardian (2022). https://bit.ly/guardian- climatescam-twitter (Accessed 31 July 2023)
- [10] D. L. Linvill, P. L. Warren, *Troll factories: Manufacturing specialized disinformation on Twitter*. Polit. Commun. 37, 447–467 (2020).

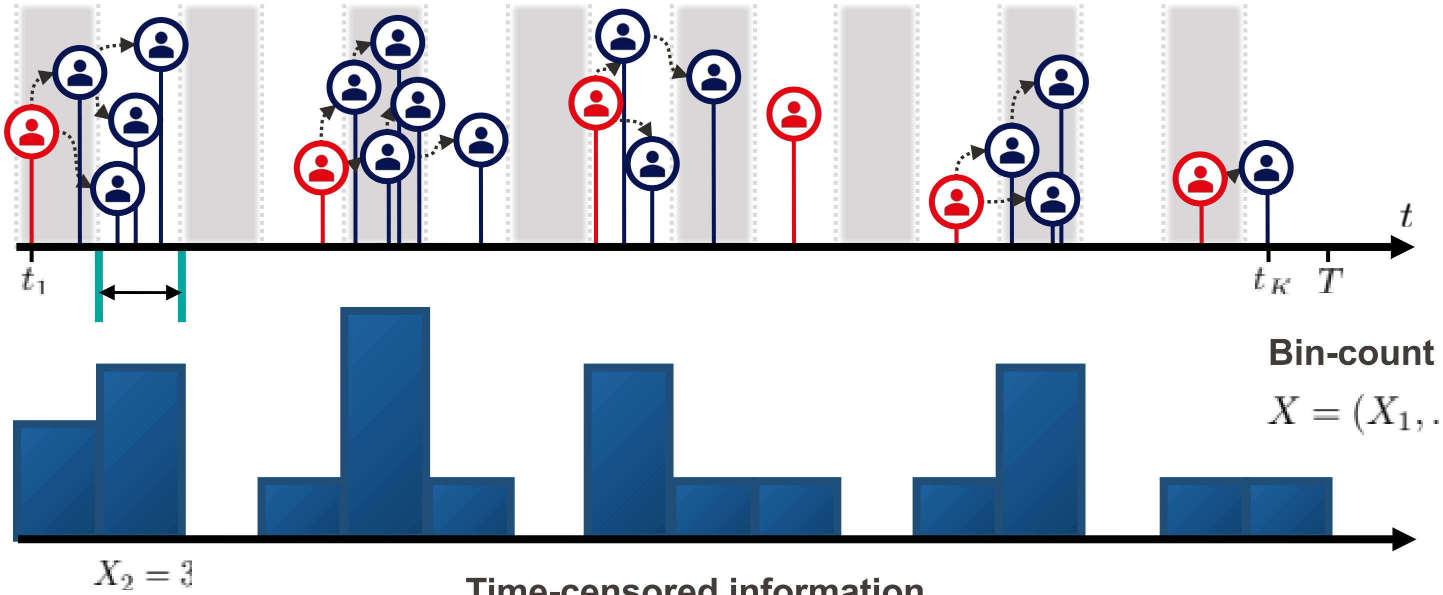
Applicability to time-censored information

Intensity function

$$\lambda(t|\mathcal{H}_t) = \overbrace{\mu}^{\text{Base intensity}} + \overbrace{\sum_{i:t_i<t} \underbrace{\phi(t-t_i)}_{\text{Kernel function}}}^{\text{Self-excitation}}$$

Event history

$\mathcal{H}_T = \{t_1, \dots, t_K\} \subset (0, T$



Bin-count vector

$X = (X_1, \dots, X_L)$

Time-censored information