Fit4Democracy
Fake News in the Political Discourse

Prof Dr Marten Risius
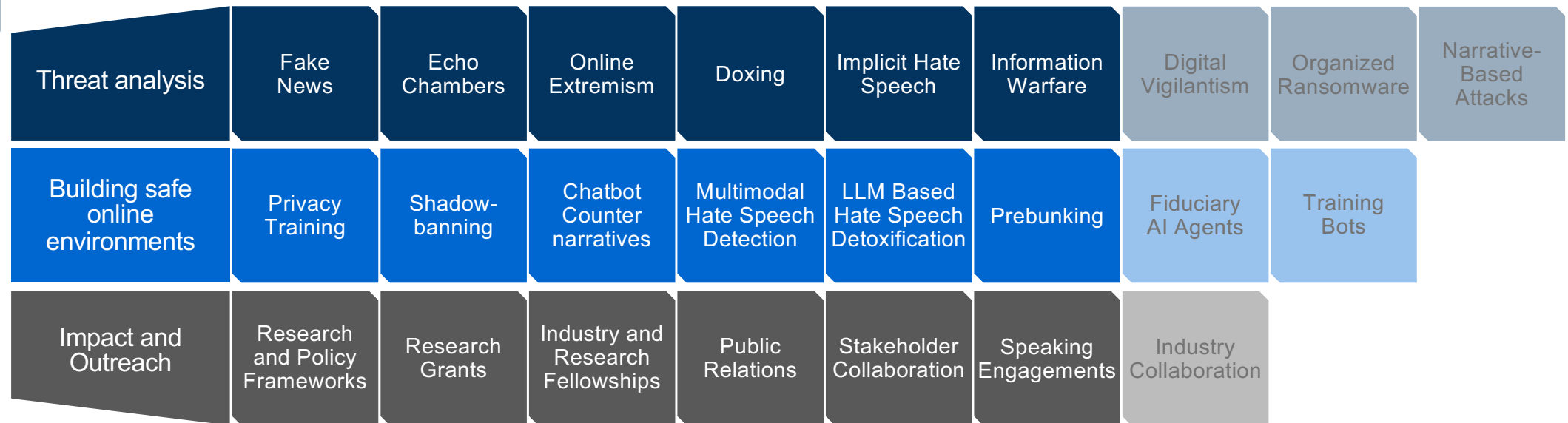
Digital Society and Online Engagement

Institute for Digital Innovation

**Bavarian Distinguished Professorship**

23/09/2025

Who are we?
Online Trust & Safety Topic Overview

HNU

Threat analysis	Fake News	Echo Chambers	Online Extremism	Doxing	Implicit Hate Speech	Information Warfare	Digital Vigilantism	Organized Ransomware	Narrative-Based Attacks
Building safe online environments	Privacy Training	Shadow-banning	Chatbot Counter narratives	Multimodal Hate Speech Detection	LLM Based Hate Speech Detoxification	Prebunking	Fiduciary AI Agents	Training Bots	
Impact and Outreach	Research and Policy Frameworks	Research Grants	Industry and Research Fellowships	Public Relations	Stakeholder Collaboration	Speaking Engagements	Industry Collaboration		

The professorship "Digital Society and Online Engagement" receives generous funding support through the Distinguished Professorship Program in the Bavarian Hightech Agenda by the Bavarian Ministry of Sciences and Arts

Fake News
Embedded in the broader context of false news

Misinformation
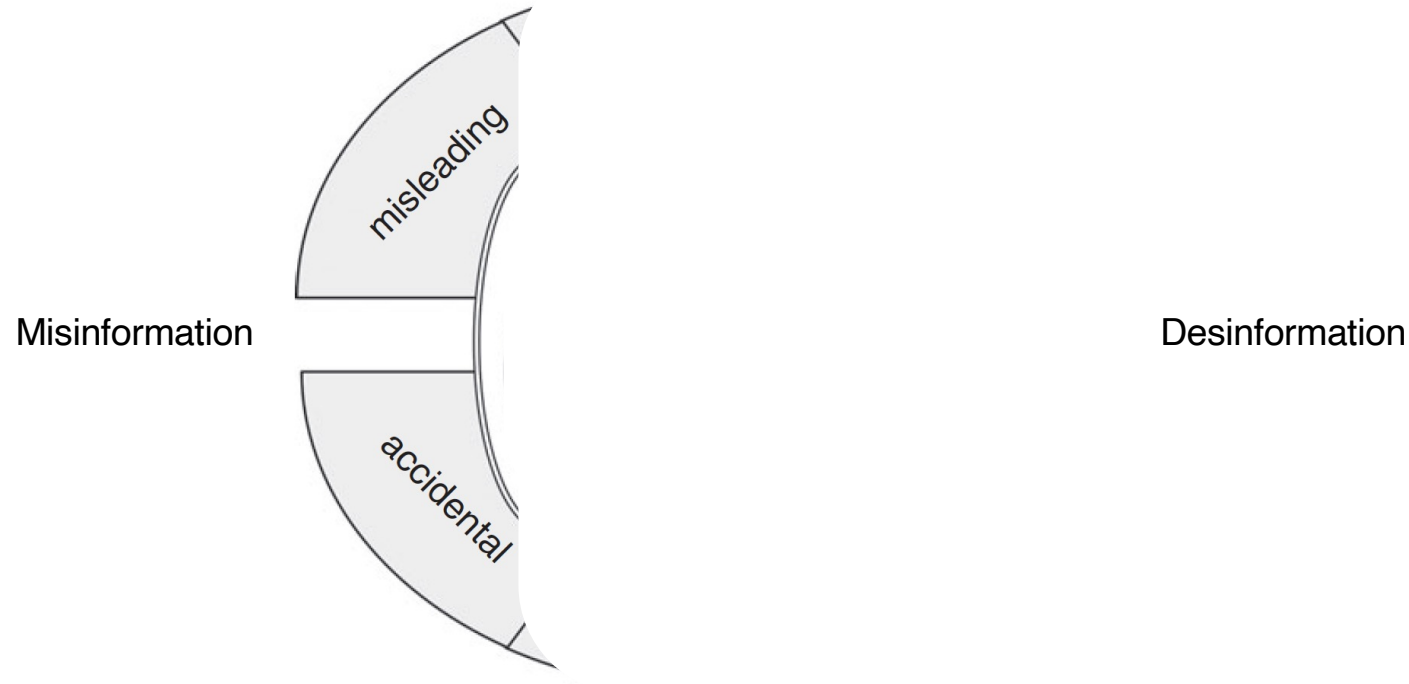
Desinformation

- They **imitate professional news media** formats but differ from them in terms of organizational processes or content:

- (i) the **core** content of the information (including textual information, imagery, audio elements, etc.); (ii) accompanying **meta-information** (headlines/titles, author information, tags, and keywords); and (iii) **contextual aspects** (positioning, references to other articles, framing). All of these elements can be subject to varying levels of "fake".

How to spot Fake News?
Common persuasion techniques to mislead an audience

HNU

01
Emotional Language
Language that contains strong emotional terms – especially those with a negative affect such as fear or outrage. Source

02
False Dichotomy
Presenting a limited number of choices or sides as mutually inclusive, when in reality more options are possible. Source

03
Cherry Picking
Presenting a selection of evidence to support a claim, while ignoring or omitting evidence that contradicts it. Source

04
Fake Experts
Fake experts are often deployed to elicit support for unproven arguments, or to cast doubt on established expert consensus. Credentials are often distorted or generalized to make the messenger appear more knowledgeable.

05
Red Herring
A red herring is used to draw attention away from the topic at hand by bringing up an unrelated topic instead.
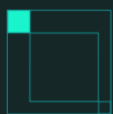
06
Scapegoating
Scapegoating is a psychological mechanism that projects responsibility and blame for a serious and/or complex problem on others, despite it not being realistic for them to be responsible for the entire problem.

07
Ad Hominem
When watching videos or browsing content online, you might come across content that questions (or attacks) the shortcomings of a person/entity, rather than the merits of the argument itself.

08
Polarization
Exaggerating existing differences between two groups to create a sense of hostility towards another group, such as using "us" versus "them" language.

09
Impersonation
Spreading information as another person or organization in order to appear more trustworthy and credible.

10
Slippery Slope
Suggesting that taking a minor action will inevitably lead to major consequences.

11
Decontextualization
Intentionally showing text, audio, or visuals in a different context or removing important background information in order to change its meaning.

Source: https://prebunking.withgoogle.com/

The Evolution of Fake News (1/3)
2016: Trolls and Bots

Fig. 2. Example #BlackLivesMatter related content that was circulated by RU-IRA social media accounts to different audiences.
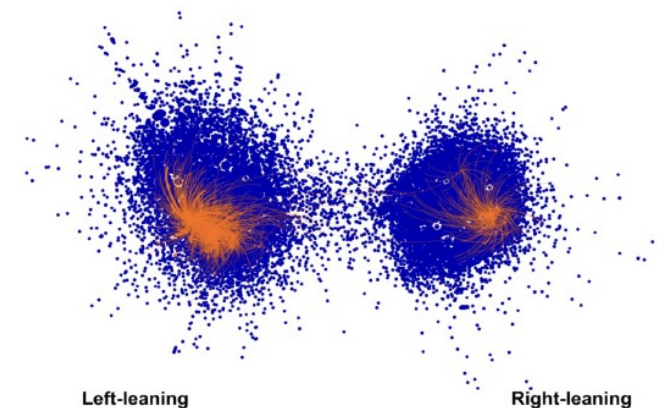

Left-leaning

Right-leaning

Fig. 1. Retweet Network Graph: RU-IRA Agents in #BlackLivesMatter Discourse. The graph (originally published [3]) shows accounts active in Twitter conversations about #BlackLivesMatter and shooting events in 2016. Each node is an account. Accounts are closer together when one account retweeted another account. The structural graph shows two distinct communities (pro-BlackLivesMatter on the left; anti-BlackLivesMatter on the right). Accounts colored orange were determined by Twitter to have been operated by Russia's Internet Research Agency. Orange lines represent retweets of those account, showing how their content echoed across the different communities. The graph shows IRA agents active in both "sides" of that discourse. This graph originally appeared in [87].

The Evolution of Fake News (2/3)
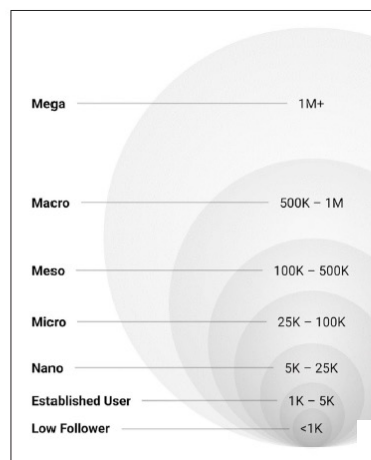## 2020: Participatory Disinformation Campaigns

Figure 1. Audience size classification, design adapted from Mediakix (2019).

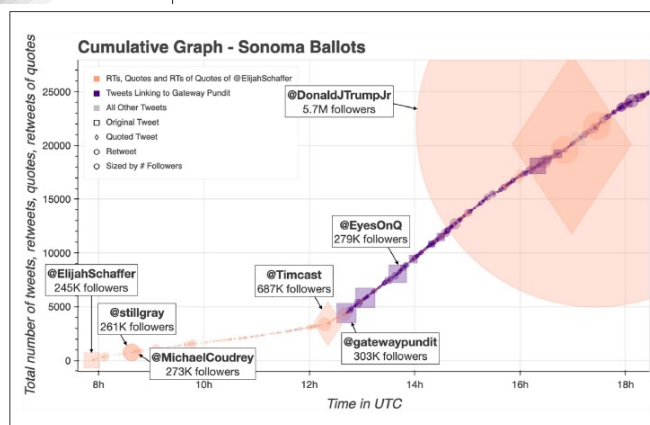Figure 2. Tweet posted by President Trump claiming the 2020 election would be rigged.

Figure 5. Cumulative graph of Sonoma Ballots tweets. The y-axis represents the total number of tweets. The x-axis is time. Individual tweets of influencers (>10,000 followers) are plotted, sized by follower count. The view is focused on the first 10 hr of propagation (aligned with the gray box in Figure 4).
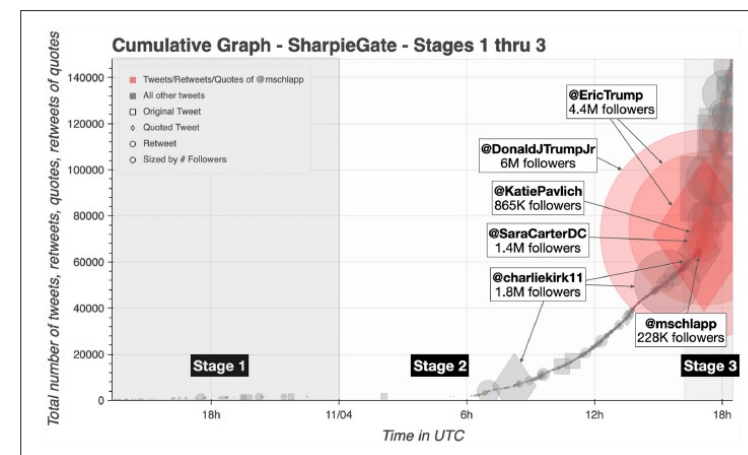
Figure 8. Cumulative graph of Stages 1–3 of SharpieGate. The y-axis represents the total number of tweets. The x-axis is time. Individual tweets of influencers (>20,000 followers) are plotted, sized by follower count. Tweets are colored red if they are tweets, retweets, or quote tweets of @mschlapp.

The Evolution of Fake News (3/3)

2024: Present and Outlook

US conservative influencers say they are 'victims' of Russian disinformation campaign

Tim Pool, Dave Rubin and Benny Johnson addressed allegations that a company they were associated with had been paid to publish videos with messages in favour of Russia

● Russia accused of trying to influence US voters through online campaign

Dave Rubin and Benny Johnson. Composite: Alamy

A number of high-
are "victims" of a
administration acc
influence the outc

Create media in minutes with your own AI Influencer

Pick traits or upload photos, and get your own photorealistic, consistent AI influencer.

Create AI Influencer

14 ways AI can be used and abused in the 2024 election, from deepfakes to foreign interference

Published: October 16, 2024 2.27pm CEST

Artificial intelligence has made its way into most aspects of American life, including elections. Alfieri via Getty

The American public is on alert about artificial intelligence and the 2024 election.

A September 2024 poll by the Pew Research Center found that well over half of Americans worry that artificial intelligence – or AI, computer technology mimicking the processes and products of human intelligence – will be used to generate and spread false and misleading information in the campaign.
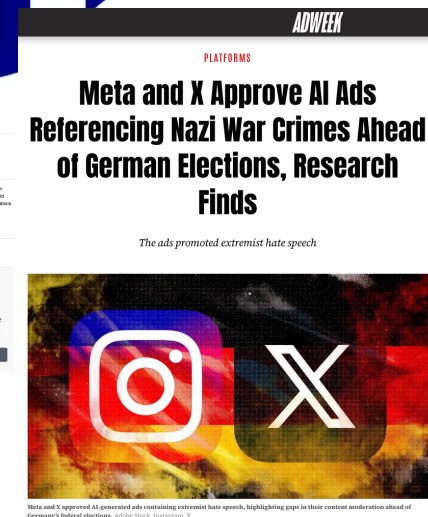
My academic research on AI may help quell some concerns. While this innovative technology certainly has the potential to manipulate voters or spread lies at scale, most uses of AI in the current election cycle are, so far, not novel at all.

I've identified four roles AI is playing or could play in the 2024 campaign – all arguably updated versions of familiar election activities.

I. Voter information

The 2022 launch of ChatGPT brought the promise and peril of generative AI into public consciousness. This technology is called "generative" because it produces text responses to user prompts: It can write poetry, answer history questions – and provide information about the 2024 election.

Rather than search Google for voting information, people may instead ask generative AI a question. "How much has inflation changed since 2020?" for example. Or, "Who's running for U.S. Senate in Texas?"

ADWEEK

PLATFORMS

Meta and X Approve AI Ads Referencing Nazi War Crimes Ahead of German Elections, Research Finds

The ads promoted extremist hate speech

Meta and X approved AI-generated ads containing extremist hate speech, highlighting gaps in their content moderation ahead of Germany's federal elections. Adobe Stock, Instagram, X

Case Example
The Russian "Doppelgänger" Influence Operation

- Doppelgänger infiltrated European (social) media landscape by a network of cloned websites, fake articles, and amplification on social media

- State-backed Russian companies purchased domain names similar to legitimate news outlets (e.g., welt.pm instead of welt.de)

- Campaign relied on AI to create disinformation stories promoting Russian narratives

- Fake news stories then spread and amplified by bot networks on social media (particularly X)
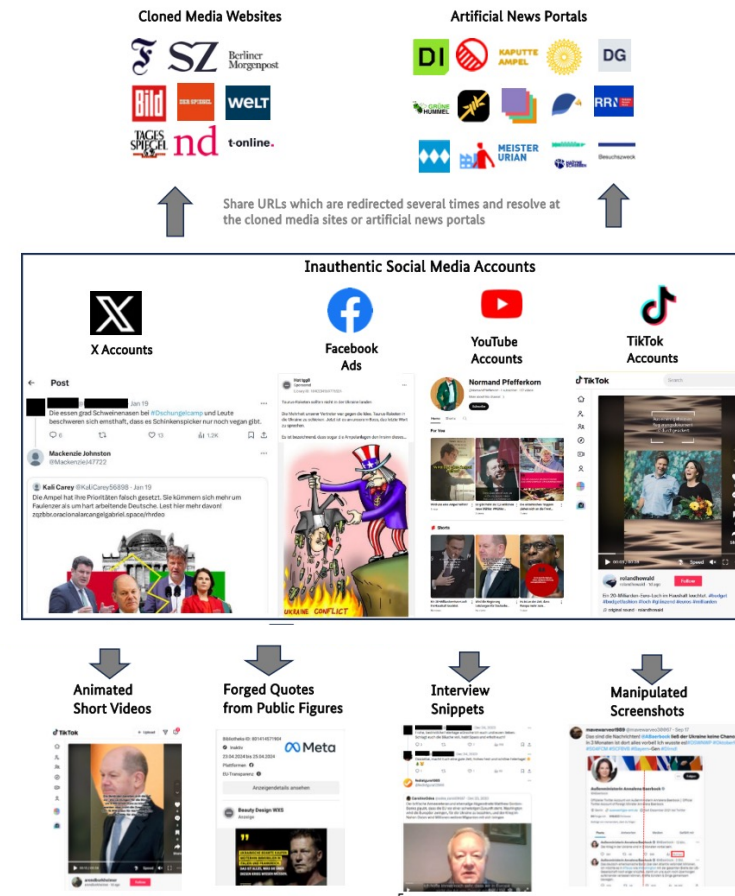
Disinformation Campaigns as Information Warfare
Maskirovka and Reflexive Control

Pentagon ran secret anti-vax campaign to undermine China during pandemic

1. Fully employed agents of influence

2. Locally recruited agents of influence

3. Unwitting agents of influence

Creating participatory disinformation campaigns
Amplifying content

"Seeder"

"Pinger"

MAGA/conservative U.S. influencers

Kremlin-aligned influencers

Elon Musk

Public figures on X

Coordinated retweeting network

Core group (pingers/seeders)

Media*

Fact-checkers*

Kremlin-aligned media

Donald Trump Jr
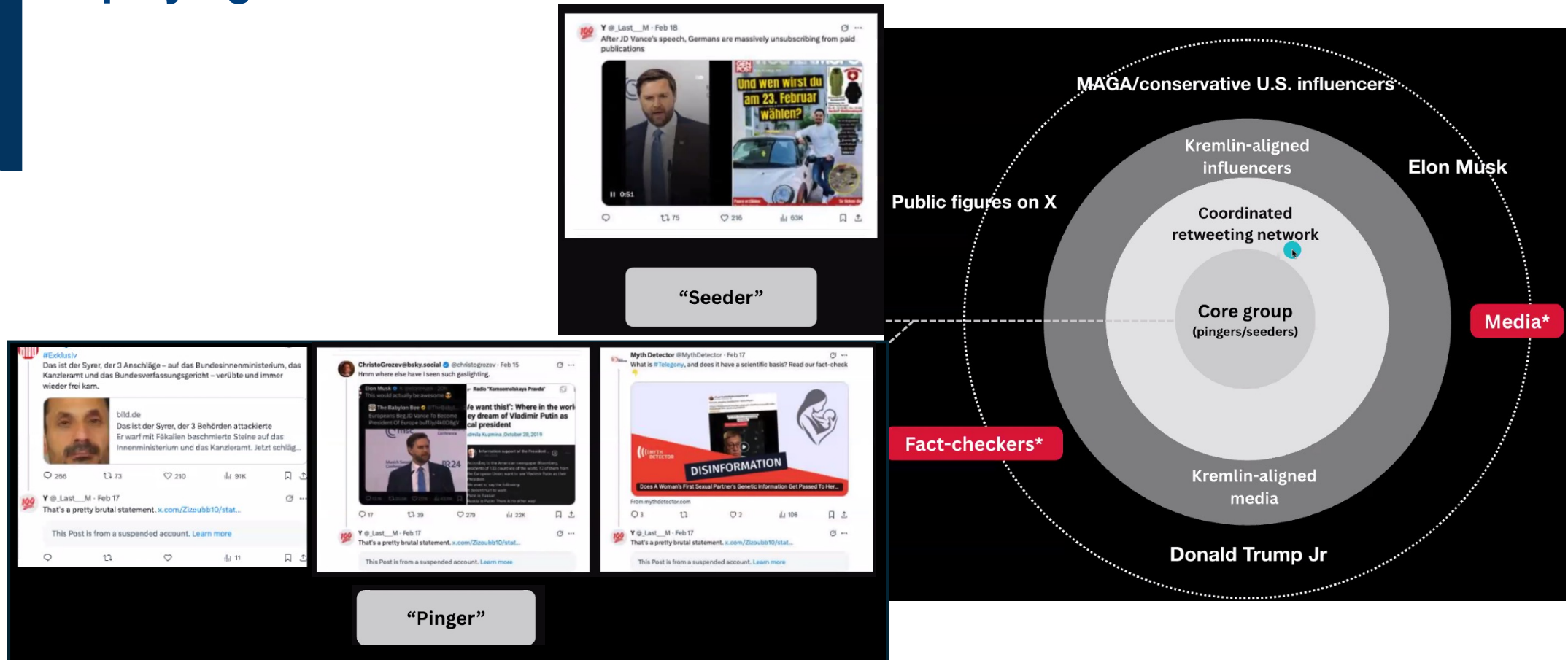
What is new about the online information environment (1/2)?
Traditionally: News media as gatekeeper
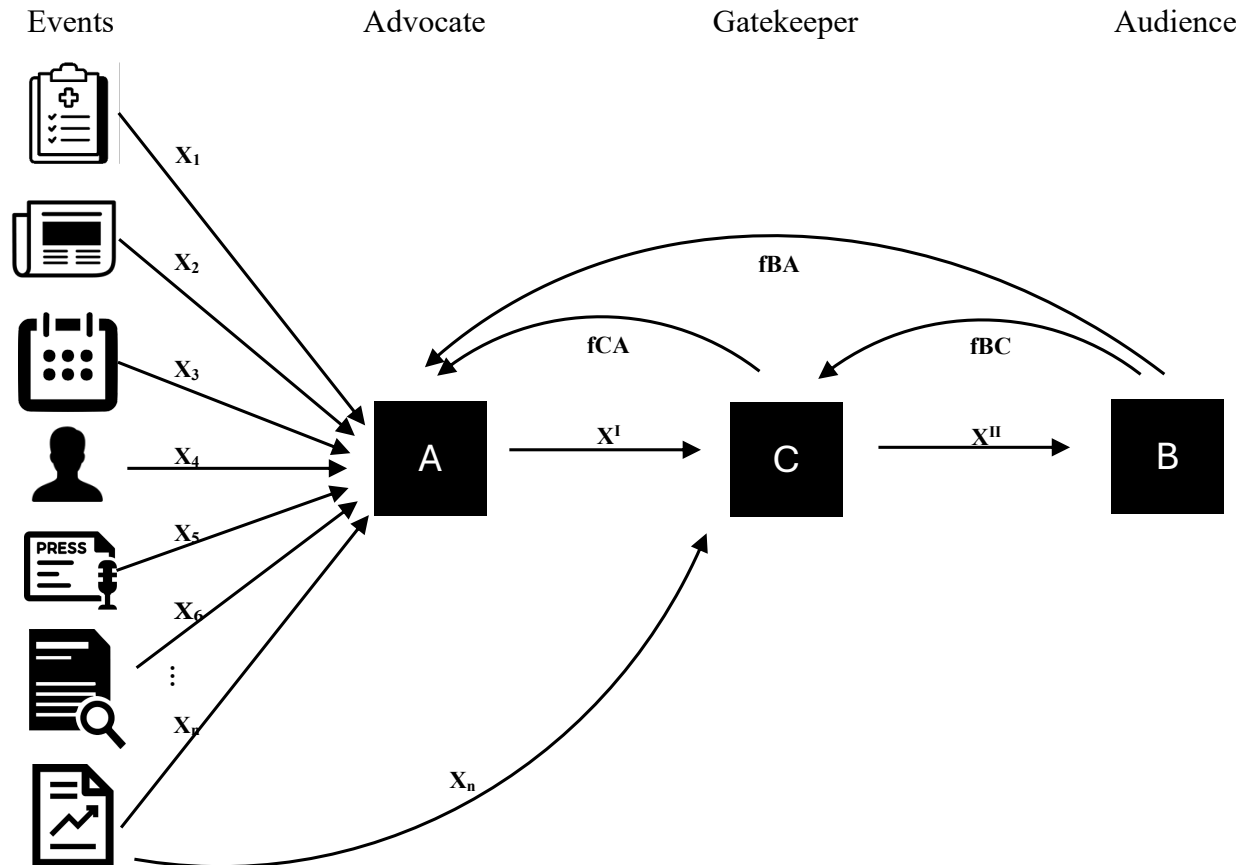
What is new about the online information environment (2/2)?
New: Platforms as the new gatekeeper

Regulatory Environment (R)

Events (E) Opinion Leader (O) Intermediary (I) Audience (A)

X_1, X_2, X_3, X_4, X_5, X_6, X_n

AO, GO, AG, AA, X^I, X^{II}, X^{III}
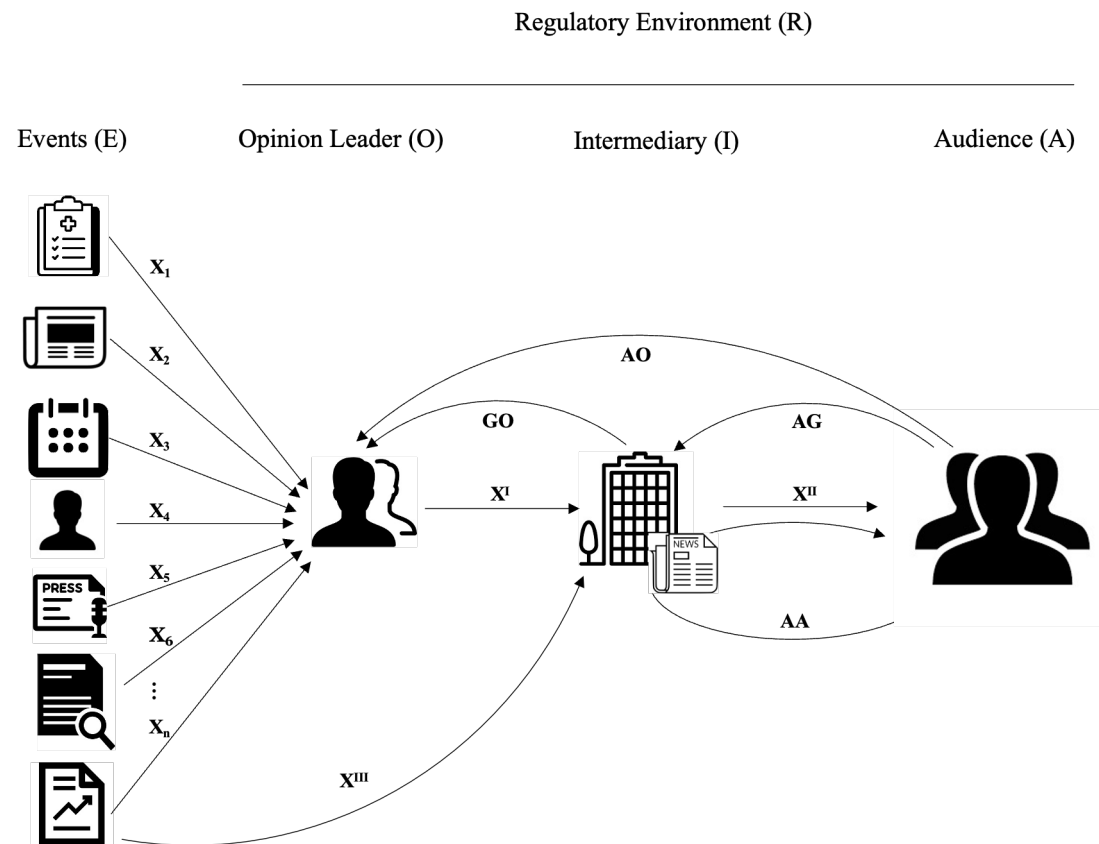
Key changes

1) Platforms as Intermediaries
2) News organizations to OL
3) Audience Engagement
4) Regulatory Environment

Borderline Content
Platforms exacerbating user engagement

Source: https://www.facebook.com/notes/751449002072082/
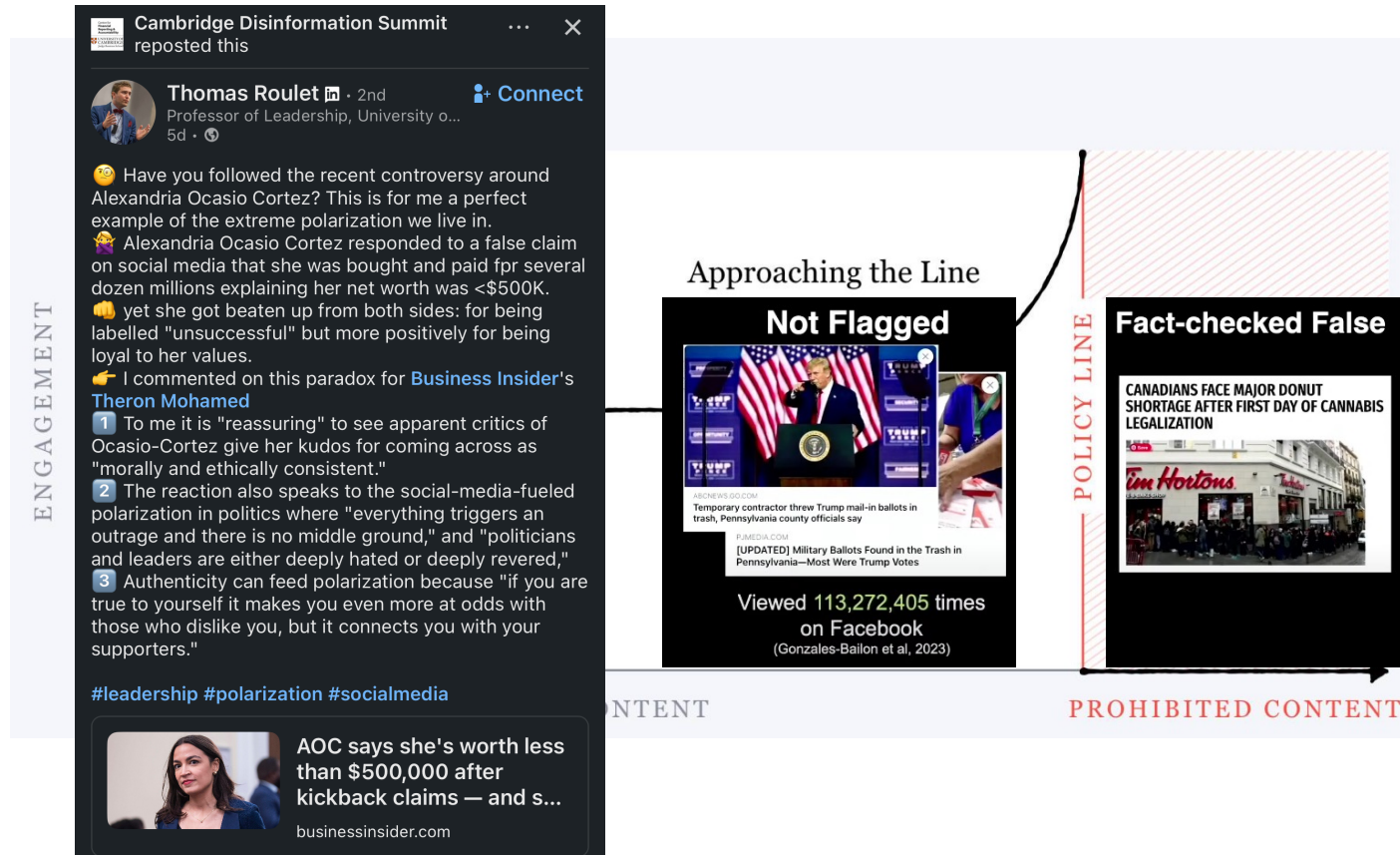
Fake News are not THE Problem
Impact of borderline content

Media consumption, all (ages 18+)

Legend: TV: Not news | TV: news | N/A TV: Fake news | Mobile: Not news | Mobile: news | Mobile: Fake news | Desktop: Not news | Desktop: news | Desktop: Fake news

Allen et al, *Science Advances*, 2020
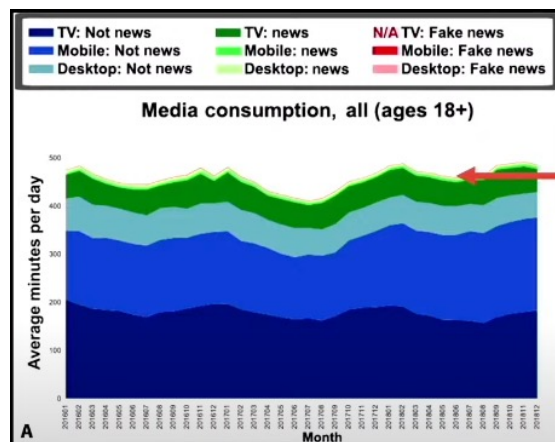
Fake News makes up less than .1% of overall information consumption

Aggregate Impact of Misinformation vs. Vaccine-Skeptical Content

Misinformation content lowers vaccine intentions by predicted .05pp (0.07,.02) per user

Vaccine-skeptical content lowered vaccine intentions by predicted 2.28pp (3.4, 1) per user

A ~50X difference!

Fact Check Rating: Misinformation | Not Flagged

Impact Per User
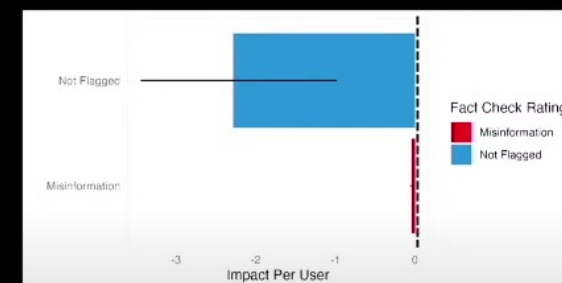
Countermeasures
Fact Checking is not the panacea

HNU

Digital Literacy
Guess et al (2020)

Source Credibility Labels
Aslett et al (2022)

Why Prebunking — Online manipulation is a complex problem and multiple approaches are needed to curb its worst effects. Prebunking provides an opportunity to **get out ahead of online manipulation**, providing a layer of protection before individuals encounter malicious content.

A GLOBAL PROBLEM
#1
Global experts identified mis- and disinformation as the greatest threat to global stability going into 2024. World Economic Forum

MEANINGFUL RESULTS
5%
The share of individuals who could correctly identify a manipulation technique increased by 5 percentage points on average after viewing a prebunking video. Science

Accuracy Nudges
Pennycook et al (2021)

Fact-Checking

Debunking / Warning Labels
Martel & Rand (2023)

From Fake News to Narrative Campaigns
Weaponization of Fact Checkers

Conclusion (1/2)
False news is a complex issue not merely resolved by fact checks

- **Misconception 1:** Disinformation is simply (a single piece of) false information
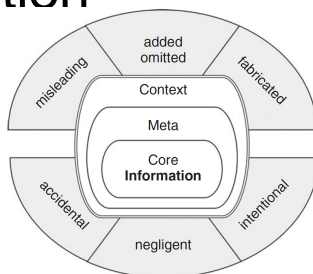
If it were, platforms could simply add 'true' and 'false' labels, a tactic that has often been suggested. But disinformation often layers true information with false — an accurate fact set in misleading context, a real photograph purposely mislabelled. The key is not to determine the truth of a specific post or tweet, but to understand how it fits into a **larger disinformation campaign**.

- **Misconceptions 2:** Disinformation stems mainly from agents producing false content
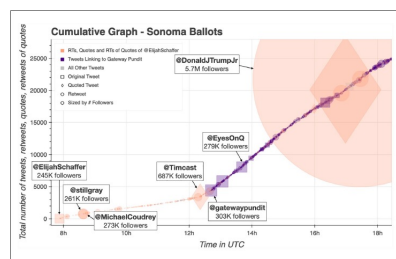

Another misconception is that disinformation stems mainly from agents producing false content (paid 'trolls') and automated accounts ('bots') that promote it. But effective disinformation campaigns **involve diverse participants**; they might even include a majority of **'unwitting agents'** who are unaware of their role, but who amplify and embellish messages that polarize communities and sow doubt about science, mainstream journalism and Western governments.
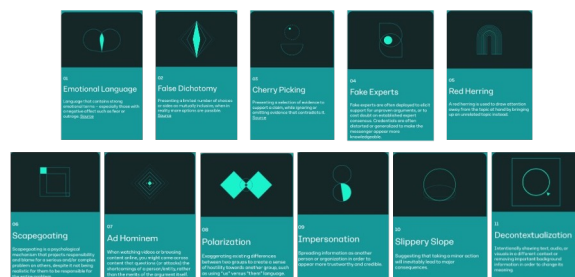
Conclusion (2/2)
Mis- and Disinformation as a Broader Societal Issue

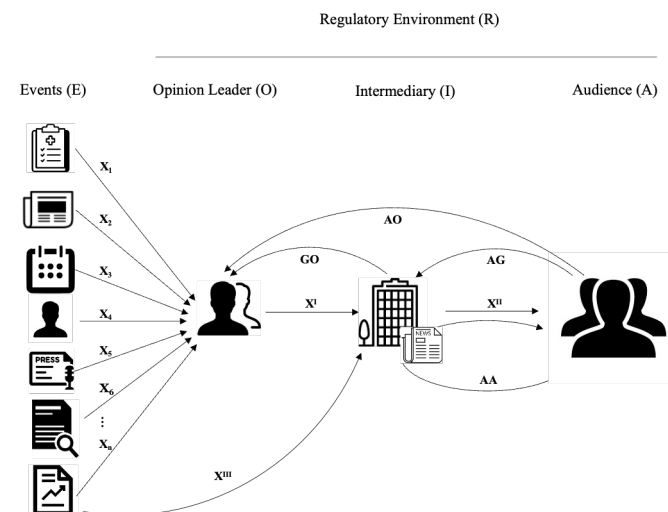Consider platform interests

Natural Engagement Pattern
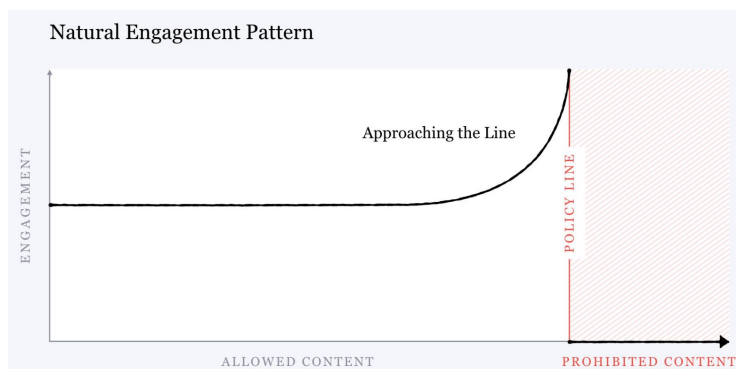
Prebunking/Attitude Inoculation

Platforms as critical infrastructure
Middleware solutions
New platforms with "European" values

"Fake News" are here to stay

Thank you...

Empower trust and safety researchers

- Further information:
- https://cornelltech.github.io/TeachingTrustSafety/