Online Trust and Safety Forum

15 May 2024

Centre for Advanced Technologies in Online Safety | CATOS

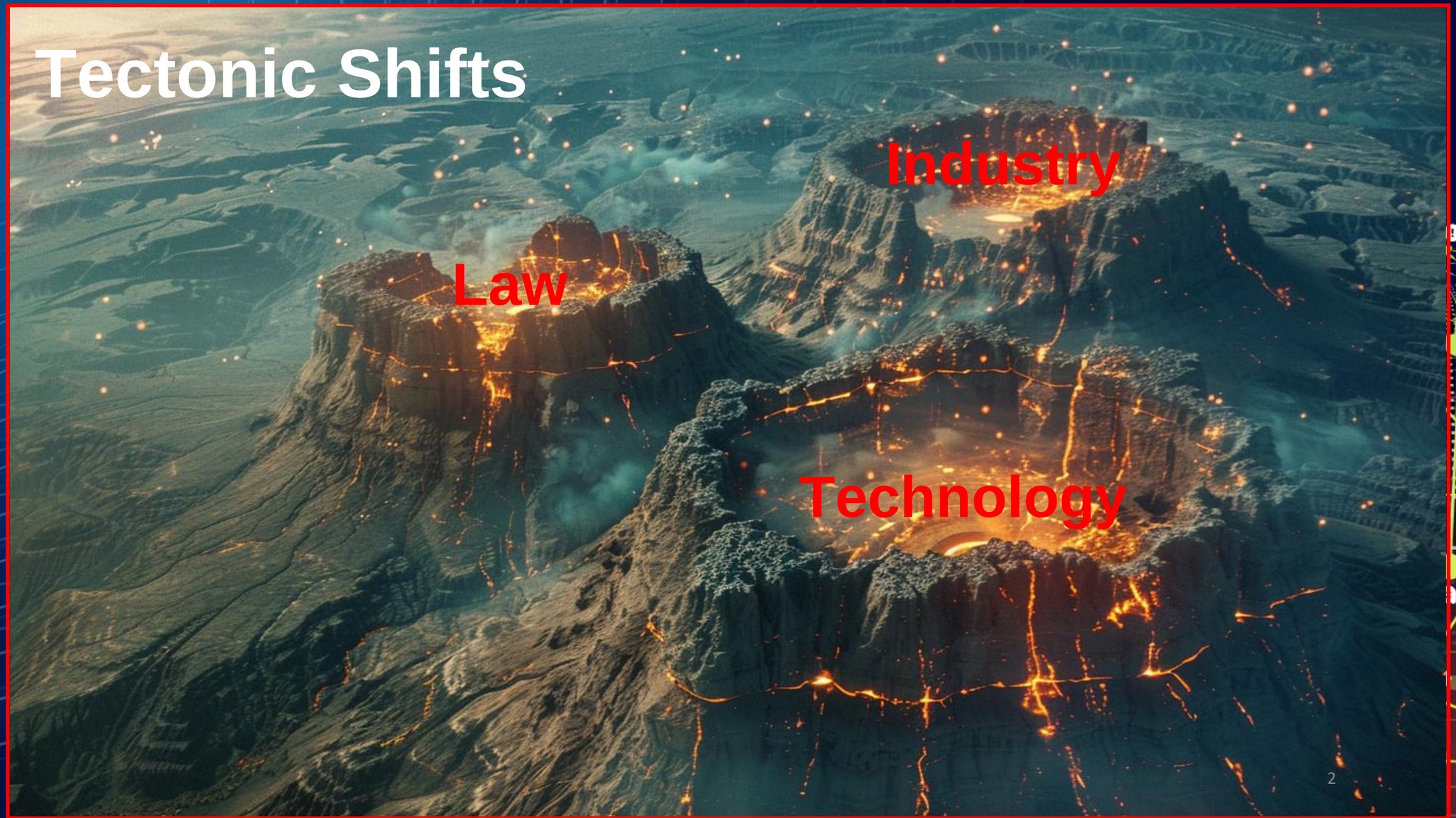
Trust and Safety at the Crossroads

Nate Persily

James B. McClatchy Professor of Law

Co-director, Stanford Cyber Policy Center

Tectonic Shifts



Industry Changes: The Twitter Effect

- Changes at Twitter
  - Disrupts industry equilibrium relating to Trust and Safety
  - Expands Overton window of permissible content moderation

- Reduction in collaboration between platforms on emerging threats

Industry Changes

- Cutting of Trust and Safety teams throughout industry
- **Fracturing** of social media universe
- Rise of **encrypted platforms** (Telegram, Signal, WhatsApp) and **alternative/decentralized social networks** (Mastodon, Bluesky, Threads)
- Retreat from politics/journalism

Technological Change: Rise of AI

- **Basic rule of AI for T and S**:

  Amplifies the abilities of all good and bad actors

  to achieve all their same goals.

- **Lowers the cost** of content production
- **Adversarial**, unpredictable environment
- **New opportunities** for content moderation
- The special challenge of **open models**

Legal Changes: Europe

- Digital Services Act
- Digital Markets Act
- AI Act*
- Product Liability Directive

PRESS RELEASE | 18 December 2023 | Brussels | 4 min read

Commission opens formal proceedings against X under the Digital Services Act

Commission opens formal proceedings against Facebook and Instagram under the Digital Services Act

Legal Changes: US

- Federal Action
  - > Divestiture of TikTok
  - > AI Executive Order
  - > Congressional Investigations
- State Action
  - > Reining in Content Moderation
  - > Privacy
  - > Protecting Children
- Supreme Court cases (*Taamneh, Murthy, Netchoice*)

Three Domains

- Hate Speech
- Child Protection
- Disinformation and Elections

Hate Speech: The State of Research and Policy

- "Tail problem"
- Targeted and networked harassment
- Sought out v. algorithmically recommended
- Increasingly difficult to distinguish from "normal politics" (esp. immigration, Mid Eastern conflict, culture wars)
- Problem = young men
- Trend unclear – "bursty" problem
- Twitter's reinstatement of speakers

- Policy:
  Netchoice Supreme Court cases

The New York Times

Hate Speech's Rise on Twitter Is Unprecedented, Researchers Find

Problematic content and formerly barred accounts have increased sharply in the short time since Elon Musk took over, researchers said.

BROOKINGS

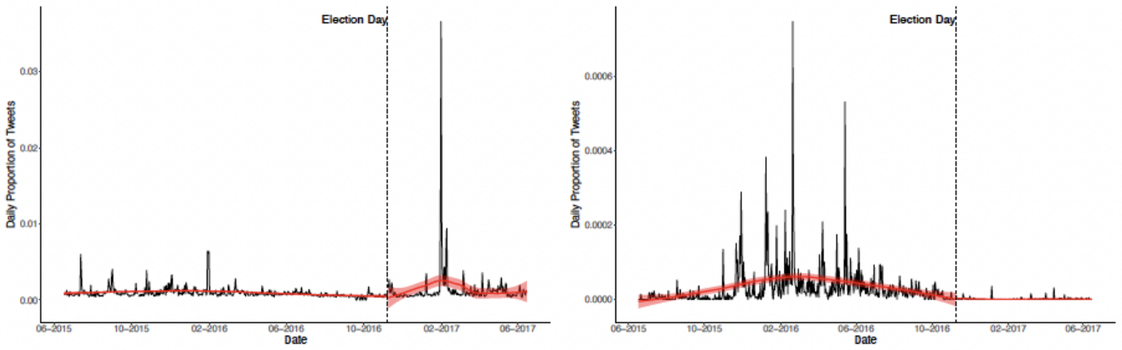
COMMENTARY

Why is Elon Musk's Twitter takeover increasing hate speech?

Rashawn Ray and Joy Anyanwu
November 23, 2022

Hate Speech: Impact of 2016 Election

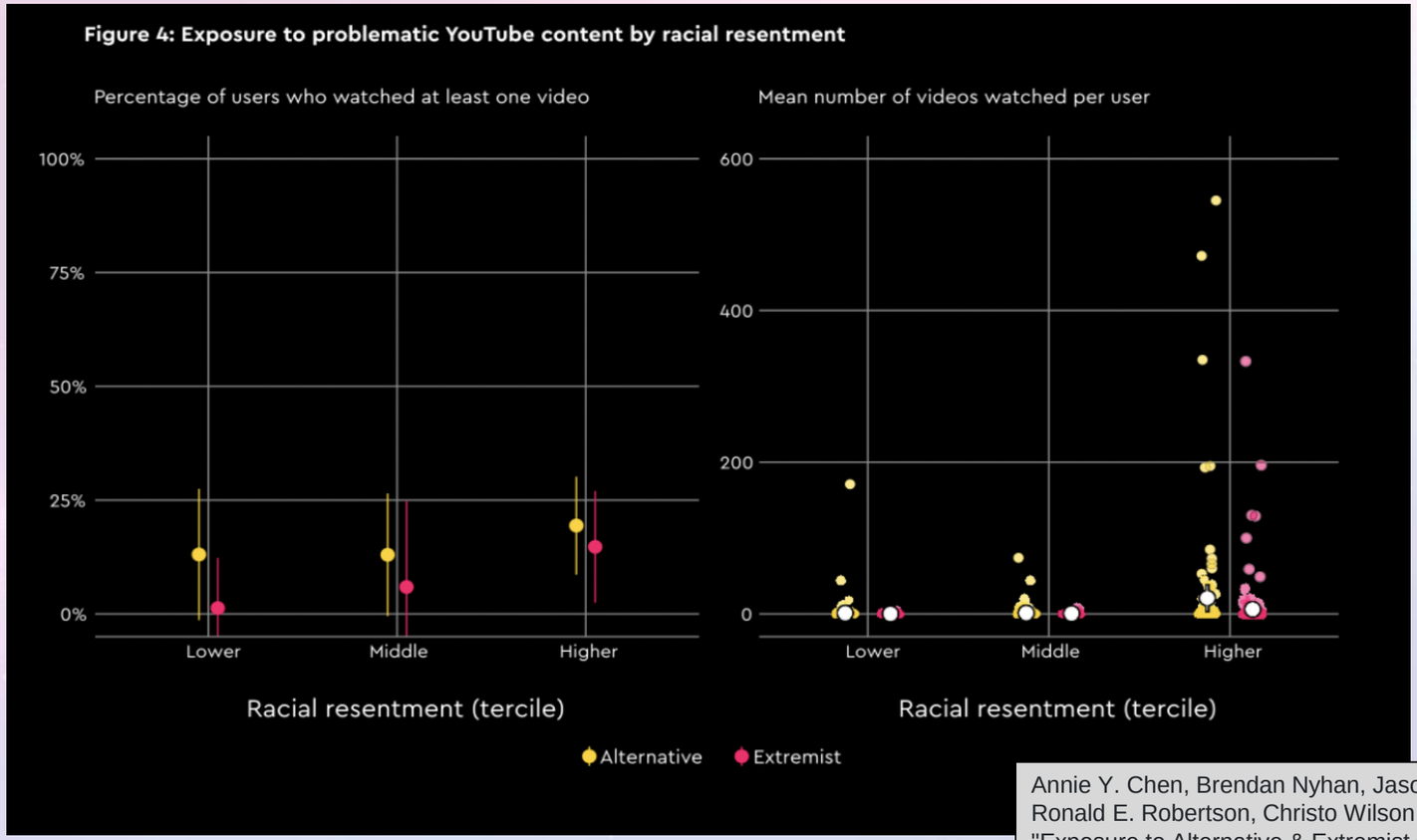
Figure 5: Effect of 2016 Election on Daily Proportion of Hate Speech and White Nationalist Language Tweets
Interrupted Time Series Analysis (Trump, Clinton, and Random Sample Datasets)



"No evidence of a lasting increase in hate speech or white nationalist rhetoric either over the course of the campaign or in the aftermath of Trump's election"

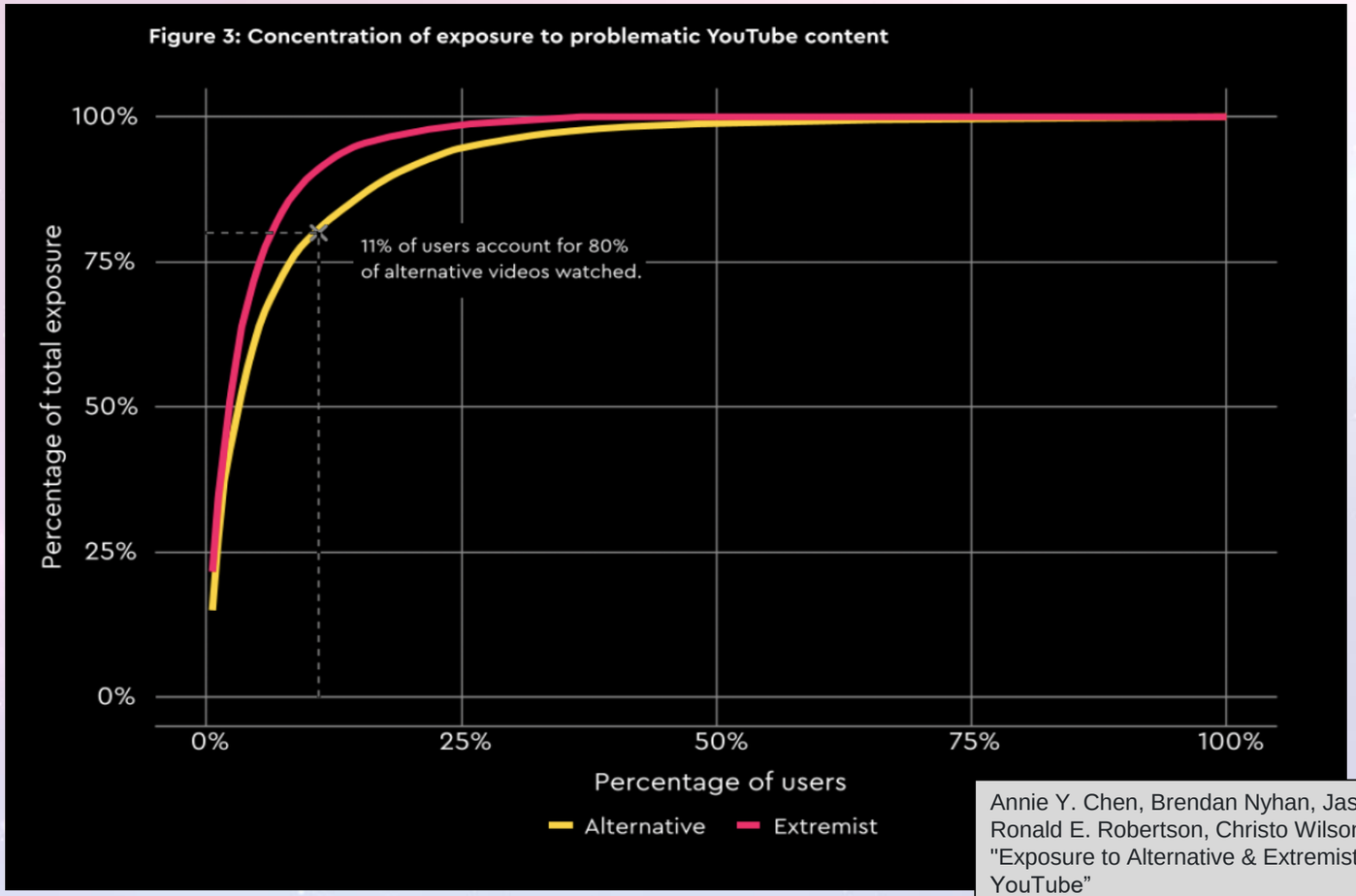
Alexandra A. Siegel, Evgenii Nikitin, Pablo Barberá, Joanna Sterling, Bethany Pullen, Richard Bonneau, Jonathan Nagler and Joshua A. Tucker (2021), "Trumping Hate on Twitter? Online Hate Speech in the 2016 U.S. Election Campaign and its Aftermath", Quarterly Journal of Political Science: Vol. 16: No. 1, pp 71-104. http://dx.doi.org/10.1561/100.00019045

Hate Speech: YouTube Exposure



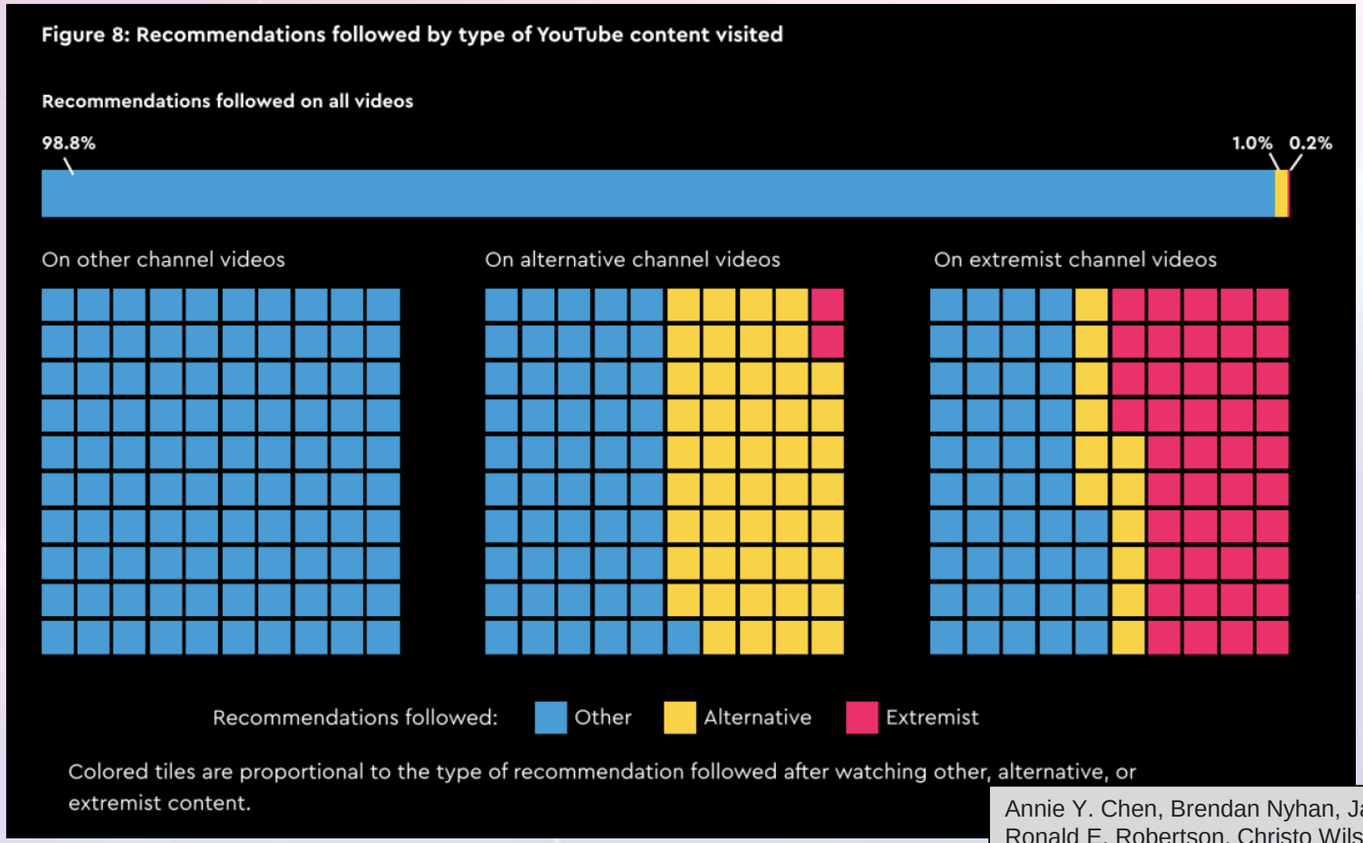
Annie Y. Chen, Brendan Nyhan, Jason Reifler, Ronald E. Robertson, Christo Wilson (2021), "Exposure to Alternative & Extremist Content on YouTube" https://www.adl.org/resources/report/exposure-alternative-extremist-content-youtube

Hate Speech: YouTube Exposure



Annie Y. Chen, Brendan Nyhan, Jason Reifler, Ronald E. Robertson, Christo Wilson (2021), "Exposure to Alternative & Extremist Content on YouTube" https://www.adl.org/resources/report/exposure-alternative-extremist-content-youtube

Is the algorithm to blame?



Annie Y. Chen, Brendan Nyhan, Jason Reifler, Ronald E. Robertson, Christo Wilson (2021), "Exposure to Alternative & Extremist Content on YouTube" https://www.adl.org/resources/report/exposure-alternative-extremist-content-youtube

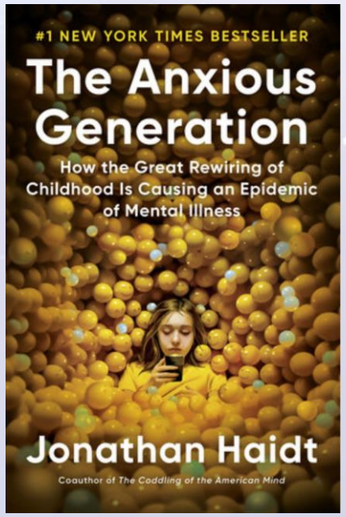
Areas of Concern: Child Protection

- Mental Health
- Child Sexual Abuse Material (CSAM)

Whistleblower discusses how Instagram may lead teenagers to eating disorders.

Assessment of the Impact of Social Media on the Health and Wellbeing of Adolescents and Children

SHARE



What do kids think about social media?

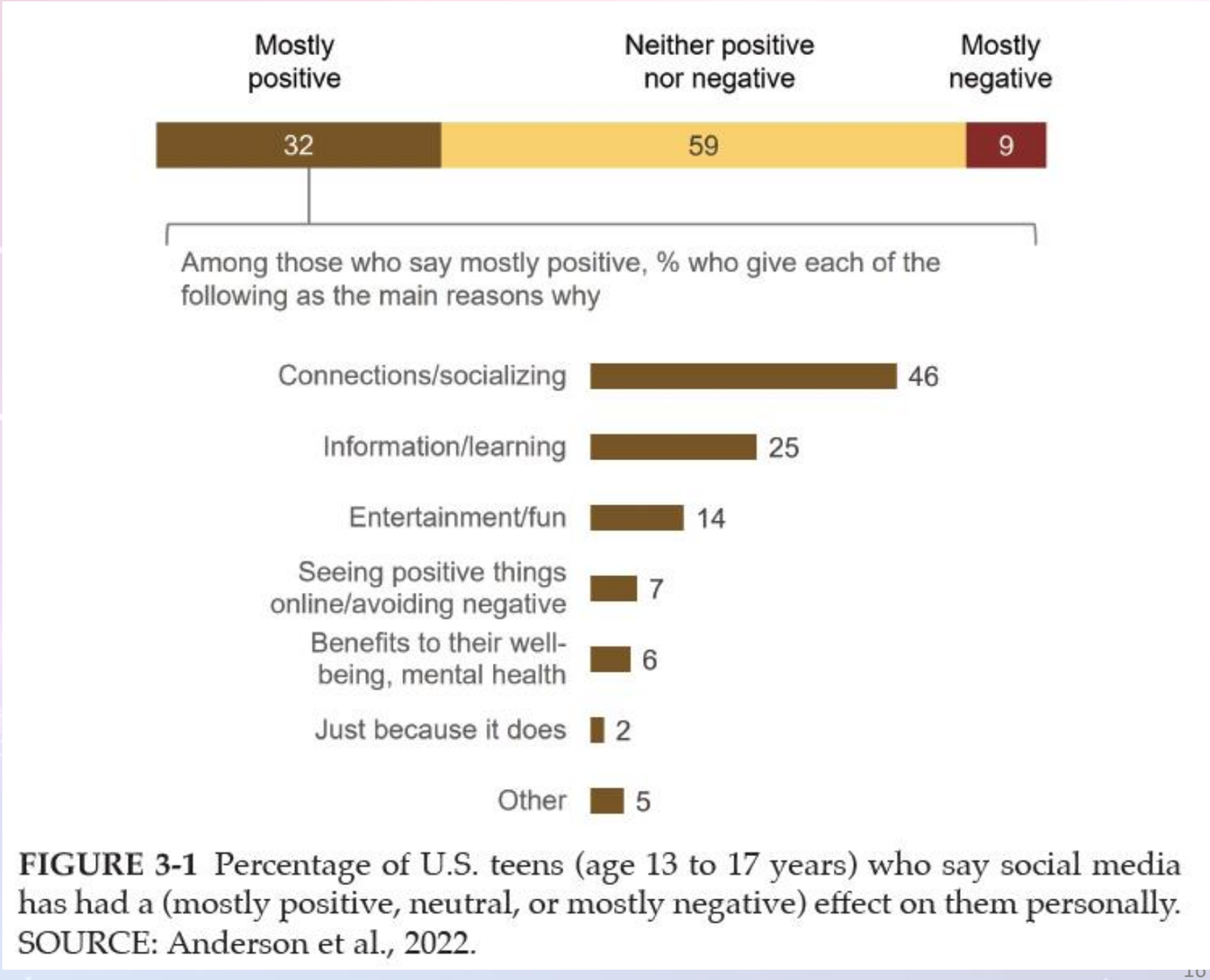


FIGURE 3-1 Percentage of U.S. teens (age 13 to 17 years) who say social media has had a (mostly positive, neutral, or mostly negative) effect on them personally.
SOURCE: Anderson et al., 2022.

National Academies of Sciences, Engineering, and Medicine. 2024. *Social Media and Adolescent Health*. Washington, DC: The National Academies Press. https://doi.org/10.17226/27396.

Vibrant Scholarly Debate As to Social Media's Effect on Kids

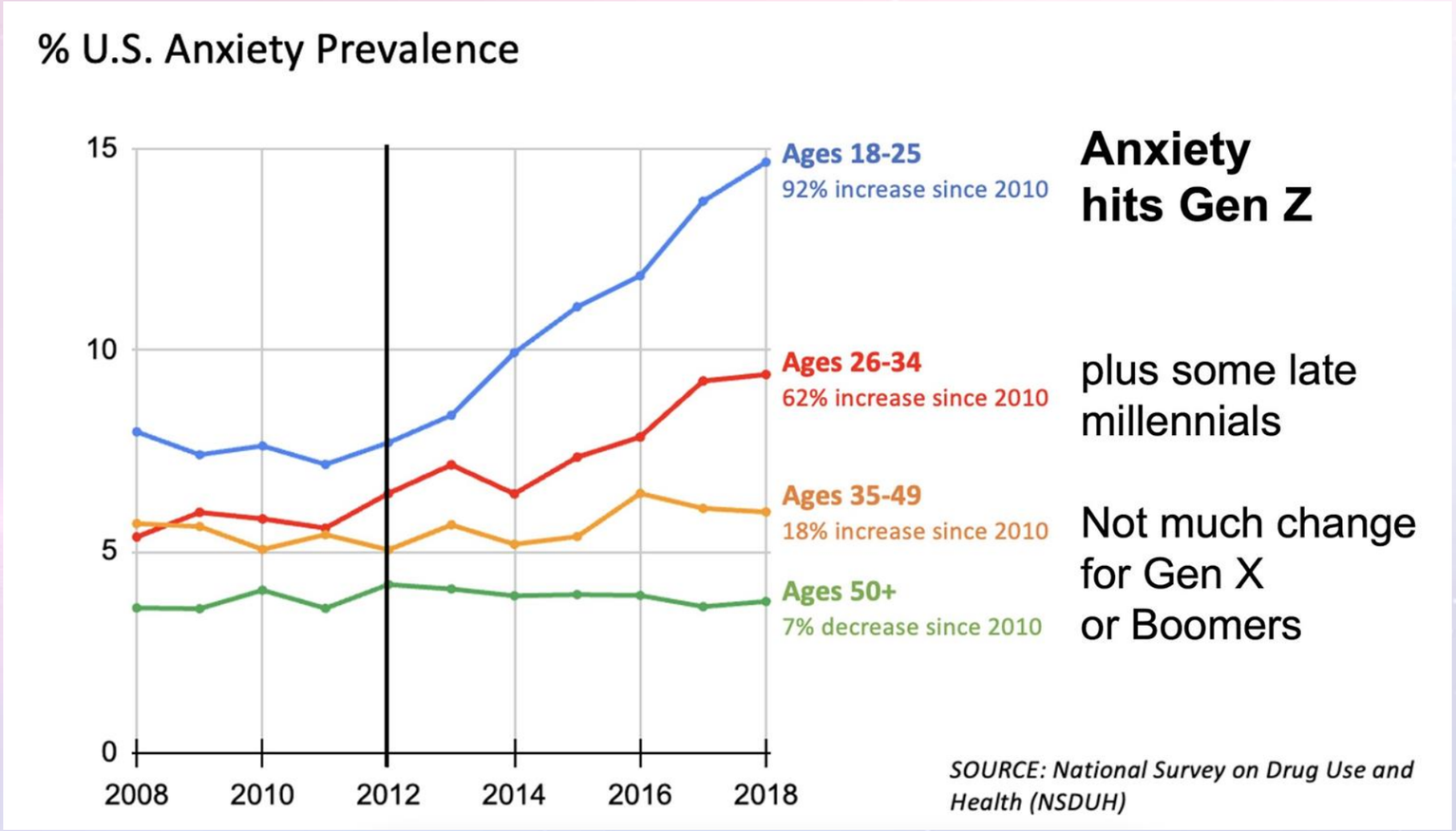
The Alarmists

- **Rise in depression** coincides with rise in smartphones/social media/self-facing cameras
- Particularly pronounced among **teen girls**
- **Mechanism** is self-comparison, FOMO, bullying

The Skeptics

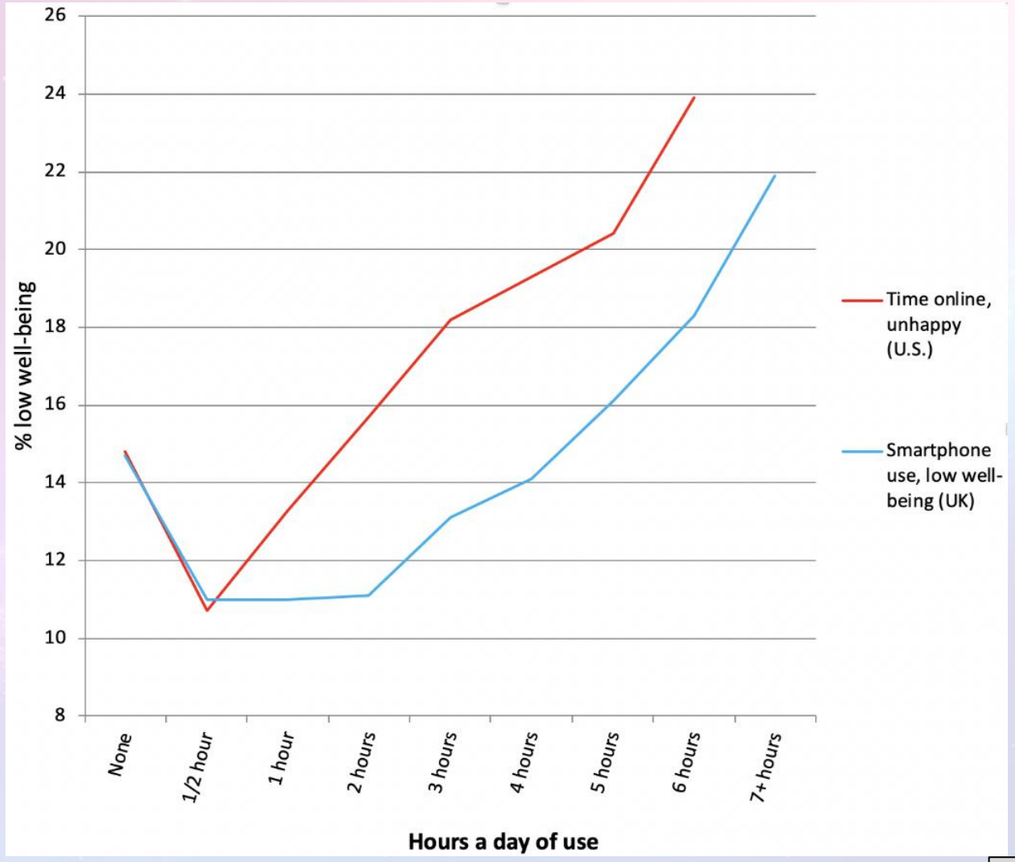
- **Heterogenous** effects
- **Other sources** of recent rise in depression (pandemic, etc.)
- Effect on **sleep** and exercise
- Panic feeds on itself
- Social media is not one thing

Child Protection: Mental Health



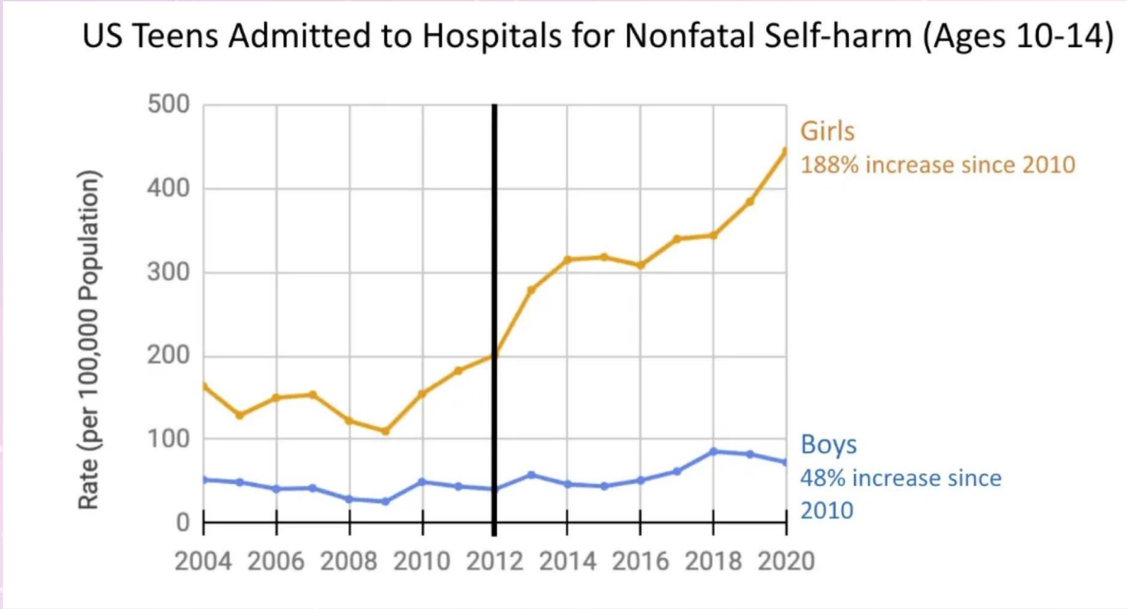
Jon Haidt. “Yes, Social Media Really Is a Cause of the Epidemic of Teenage Mental Illness”

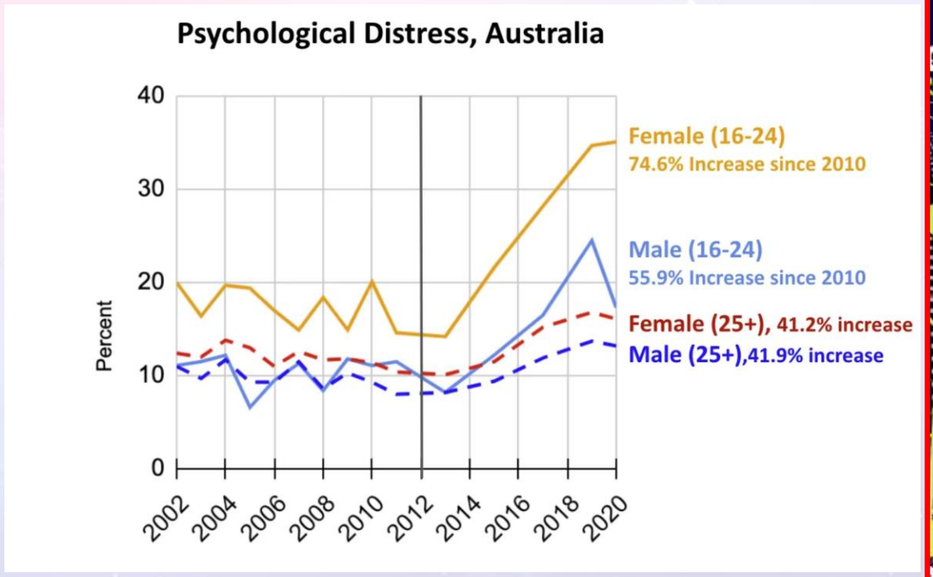
Time Online vs. Wellbeing



Haidt, J., Rausch, Z., & Twenge, J. (ongoing). Social media and mental health: A collaborative review. Unpublished manuscript, New York University.

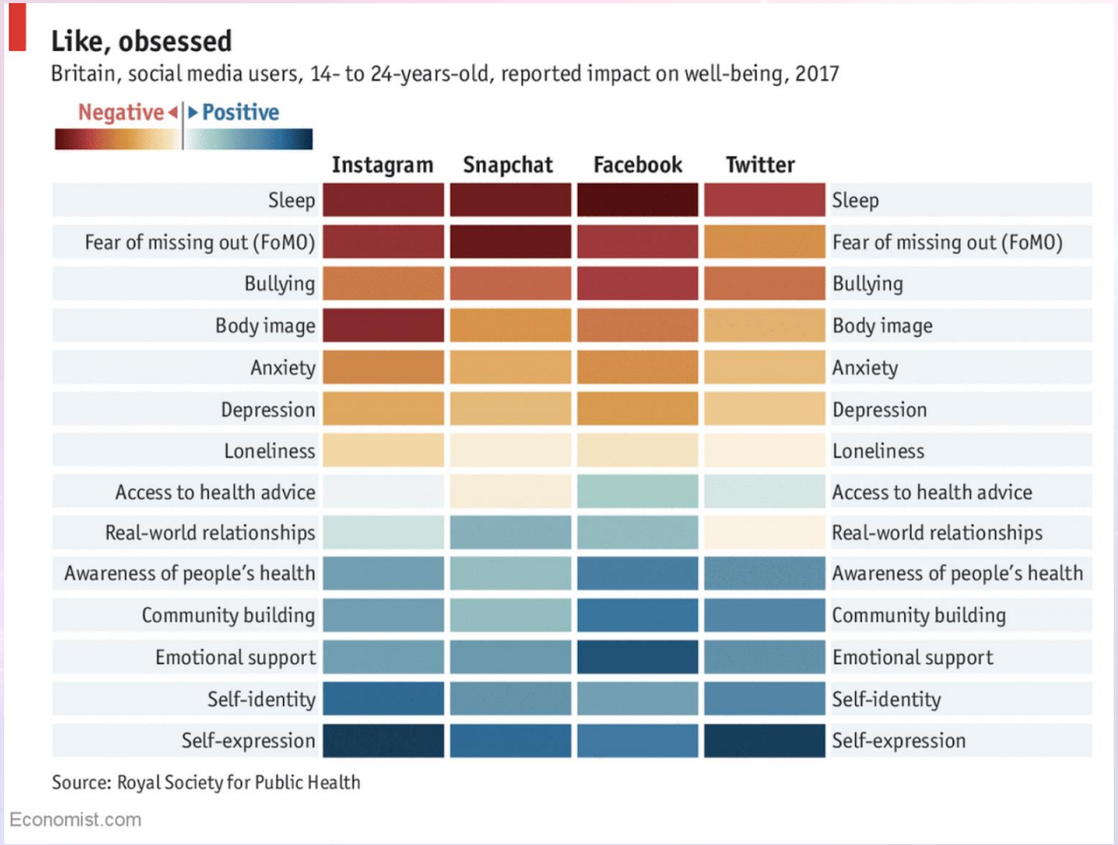
The Gender Dimension





Jon Haidt. “Yes, Social Media Really Is a Cause of the Epidemic of Teenage Mental Illness”

Child Protection: App Impact on Wellbeing



Haidt, J., Rausch, Z., & Twenge, J. (ongoing). Social media and mental health: A collaborative review. Unpublished manuscript, New York University.

Child Sexual Abuse Material (CSAM)

- Increase in CSAM reports:
  31.8 million reports from platforms in 2022

- AI undermines CSAM enforcement architecture
  > Open source tools reduce cost of production to zero
  > Overwhelms reporting and enforcement systems
  > Synthetic images of real victims
  > Provides new defenses for possession

- Note: Non-obscene virtual child pornography protected by First Amendment in US.

Cyber Tipline

Categorization of CyberTipline Reports	2020 Reports	2021 Reports	2022 Reports
Child Pornography (possession, manufacture, distribution)	21,669,264	29,309,106	31,901,234
Misleading Words or Digital Images on the Internet	8,689	5,825	7,517
Online Enticement of Children for Sexual Acts	37,872	44,155	80,524
Child Sex Trafficking	15,879	16,032	18,336
Unsolicited Obscene Material Sent to a Child	3,547	5,177	35,624
Misleading Domain Name	3,109	3,304	1,948
Child Sexual Molestation	11,770	12,458	12,906
Child Sex Tourism	955	1,624	940
Grand Total	**21,751,085**	**29,397,681**	**32,059,029**

Alex Stamos. “Current and Future Work in Online Child Sexual Exploitation”

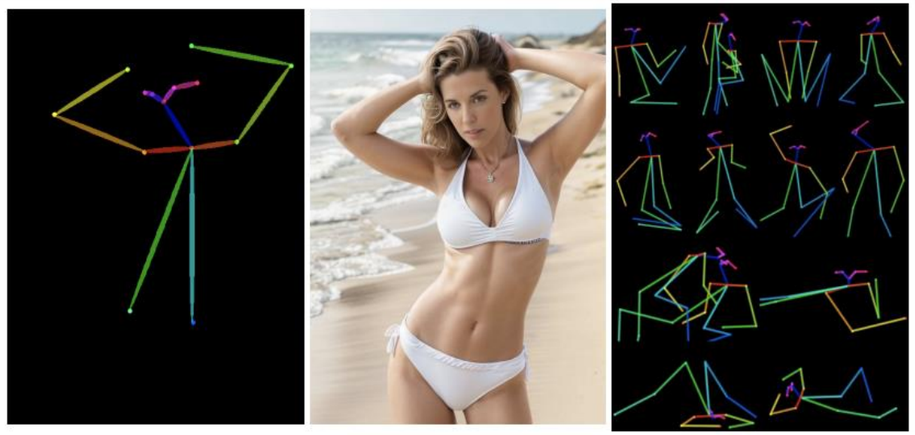


Figure 3: Left: An OpenPose “skeleton” pose. Center: A scene using that pose, generated by Stable Diffusion in conjunction with ControlNet. Right: A variety of different OpenPose poses, some of which are potentially usable for creating explicit content. OpenPose skeletons from CivitAI.

Generative ML and CSAM: Implications and Mitigations

David Thiel, Melissa Stroebel and Rebecca Portnoff

June 24, 2023

THORN Stanford | Internet Observatory
Cyber Policy Center

Disinformation

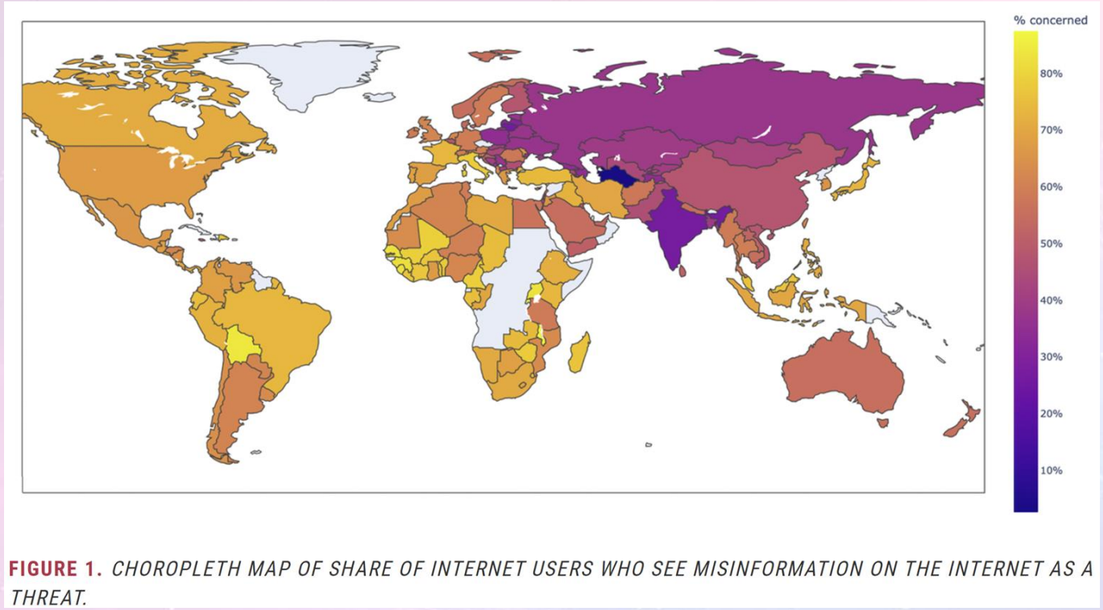
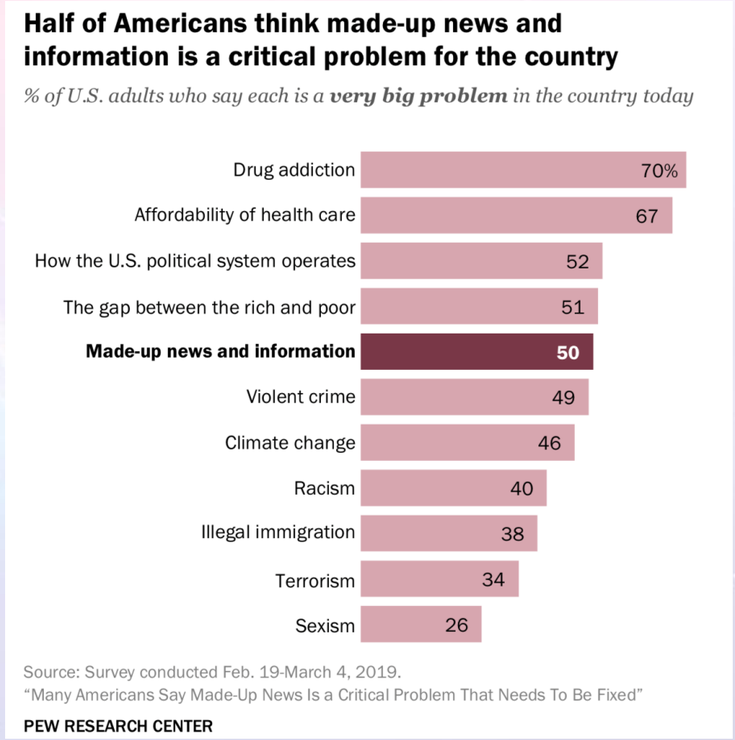


FIGURE 1. *CHOROPLETH MAP OF SHARE OF INTERNET USERS WHO SEE MISINFORMATION ON THE INTERNET AS A THREAT.*



Knuutila et al. Who is afraid of Fake News?, Harvard Misinformation Review 2024.

Disinformation Basics

1. **Scale**:
   - Large amount but small *share* of content for **most** people
   - Nevertheless, can cause significant harm

2. **Direction**: “Disinformation comes from the top”

3. **Persistence:**
   - Difficult to combat in timely fashion
   - Difficult to dislodge false beliefs over long-term

4. **Perverse Effects**: Difficult to inoculate and combat without generating widespread skepticism
5. **Intense political pressure** to avoid biased censorship.

The Family of Disinformation Phenomena

Disinformation

Misinformation

Malinformation

False Narratives

vs

Opinion

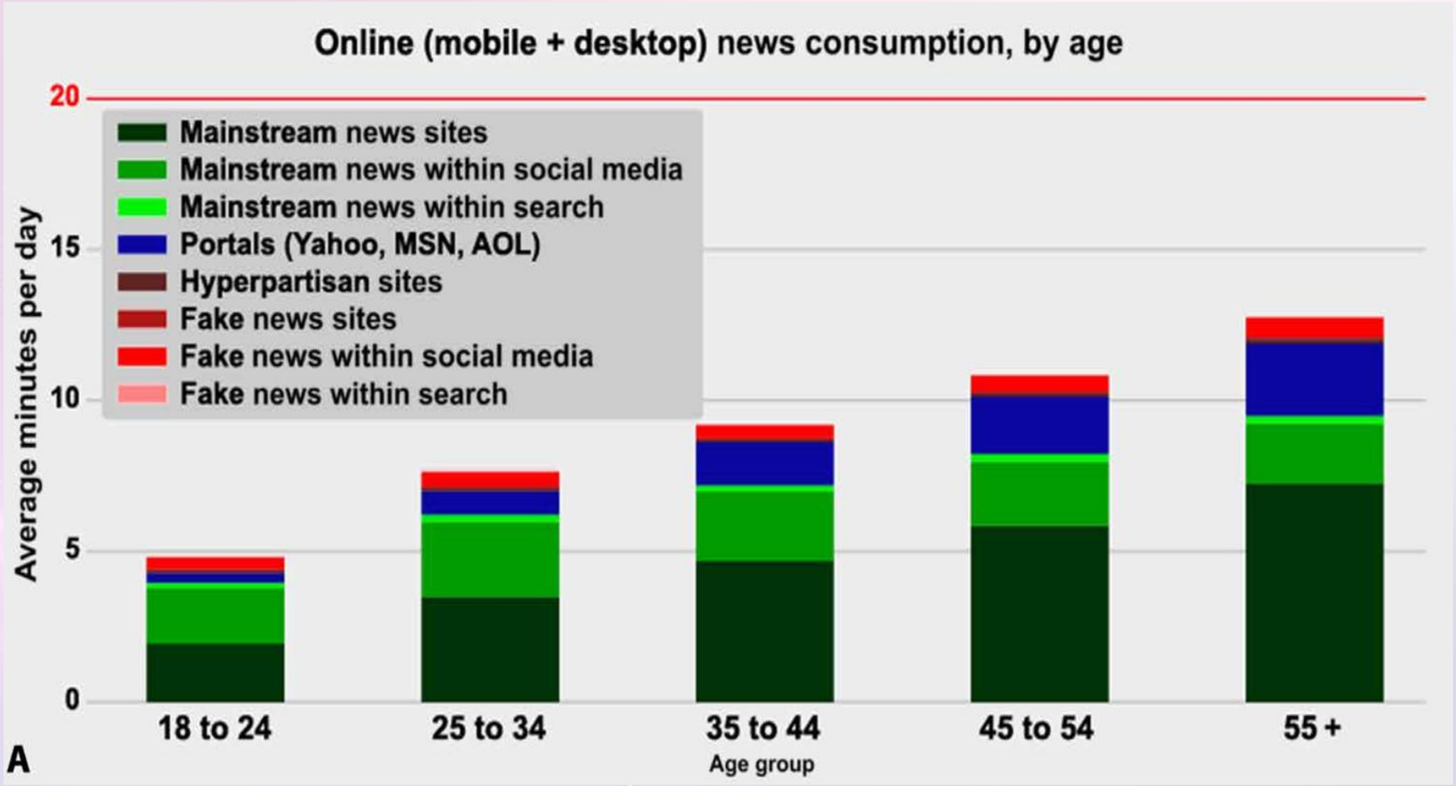
Satire/Comedy

Fiction

Asking Questions?

Reporting on Falsehoods

Disinformation: Average Consumption



Allen, J., Howland, B., Mobius, M., Rothschild, D., & Watts, D. J. (2020). Evaluating the fake news problem at the scale of the information ecosystem. *Science advances*, *6*(14).

Policy Responses to Disinformation

1. Deletion
2. Demotion
3. Disclosure
4. Delay
5. Dilution
6. Distraction & Diversion
7. Deterrence
8. Digital Literacy*

Online Trust and Safety Forum

15 May 2024

Centre for Advanced Technologies in Online Safety | CATOS

The New Frontier: The Impact of AI

https://studio.infinity.ai/feed/4385

AI: Difference in Kind or Degree?

- Lowers cost of production.
- May optimize for targeting and reach.
- Key question is not *amount* of synthetic media, but share of information diet and likelihood to persuade.
- Liar’s dividend

Online Trust and Safety Forum

15 May 2024

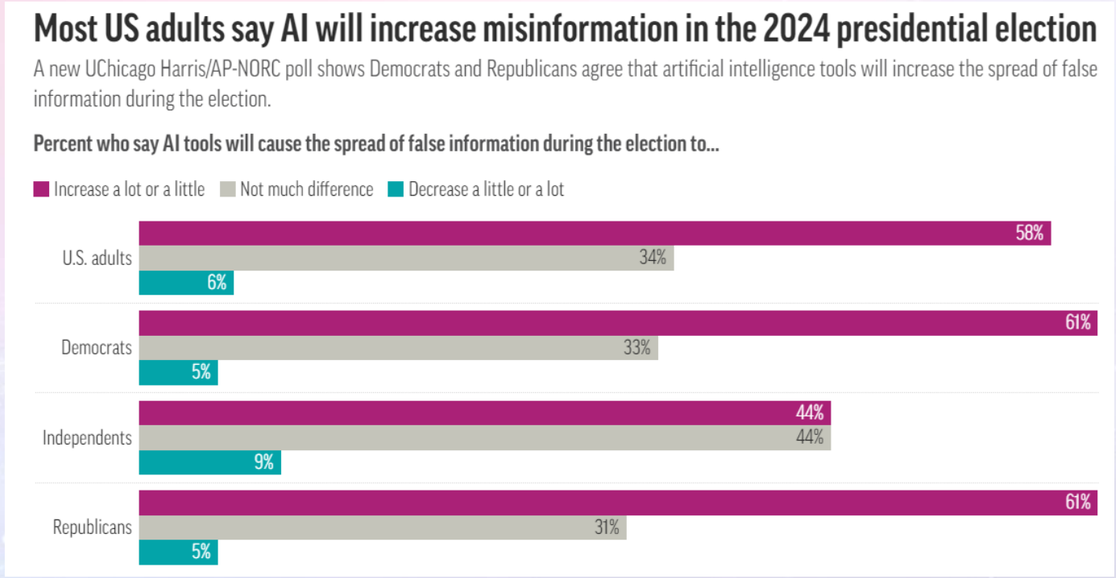
Centre for Advanced Technologies in Online Safety | CATOS

Public perception

Do you expect artificial intelligence (AI) to increase or decrease your trust in the following, or do you think it will have no impact? Advertisements I see for candidates in upcoming U.S. elections		
Increase trust	320	15%
No impact	1114	51%
Decrease trust	768	35%

Do you expect artificial intelligence (AI) to increase or decrease your trust in the following, or do you think it will have no impact? The outcome of U.S. elections		
Increase trust	332	15%
No impact	1098	50%
Decrease trust	773	35%

Do you think misinformation spread by artificial intelligence (AI) will have an impact on who wins the upcoming 2024 U.S. presidential election?		
Yes, definitely	510	23%
Yes, probably	657	30%
No, probably not	344	16%
No, definitely not	135	6%
Don't know / No opinion	557	25%



Morning Consult/Axios. August 10-13, 2023.

Public Perception

The Washington Post
Democracy Dies in Darkness

AI deepfakes threaten to upend global elections. No one can stop them.

As more than half the global population heads to the polls in 2024, AI-powered audio, images and videos are sowing confusion and clouding the political debate

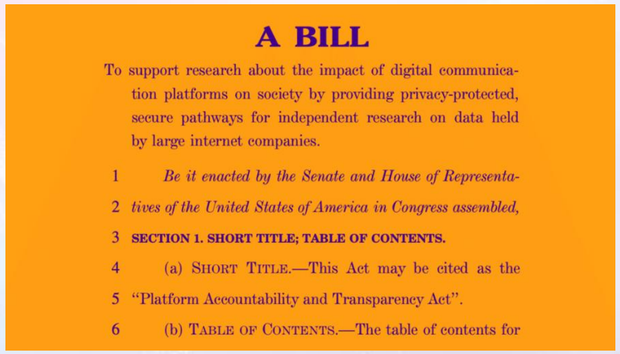
By Pranshu Verma and Cat Zakrzewski
April 23, 2024 at 7:26 a.m. EDT

What is to be done?

- EU AI Act
- Bans on risky or political uses?
- Disclosure, Watermarking, Digital Signatures, Authentication
- Development of robust auditing ecosystem

Transparency…in general

- Unsustainable equilibrium: **Platforms know more about us than we do about them**
- **API apocalypse**
- **Danger of legislating in the dark**
- Voluntary disclosures have not worked
- Three components
  - > Public disclosures
  - > Researcher immunity
  - > Privacy-protected researcher access
- **Digital Services Act**
- **Platform Accountability and Transparency Act** (proposed)



A BILL

To support research about the impact of digital communication platforms on society by providing privacy-protected, secure pathways for independent research on data held by large internet companies.

Be it enacted by the Senate and House of Representatives of the United States of America in Congress assembled,

SECTION 1. SHORT TITLE; TABLE OF CONTENTS.

(a) SHORT TITLE.—This Act may be cited as the "Platform Accountability and Transparency Act".

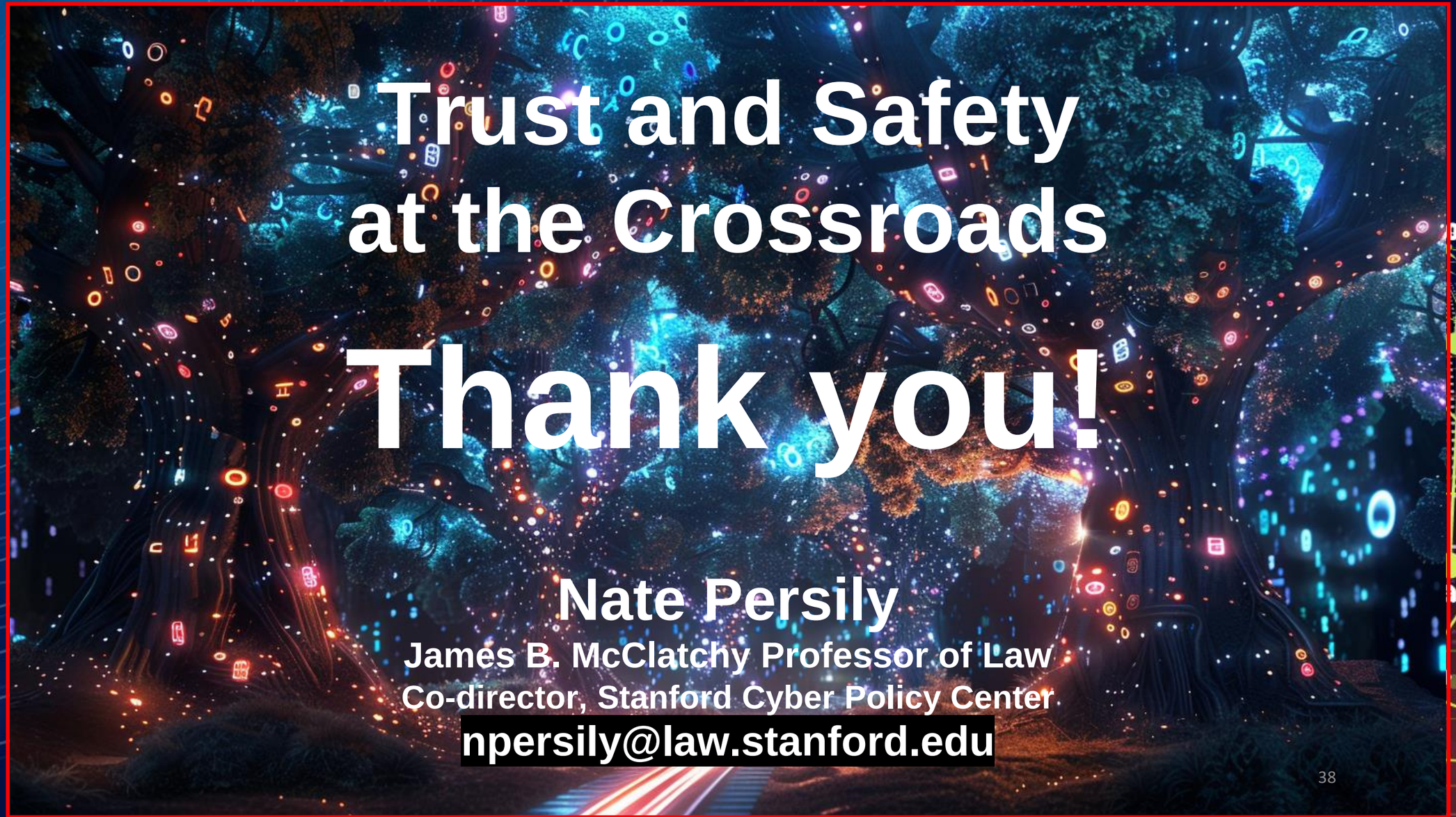
(b) TABLE OF CONTENTS.—The table of contents for

Online Trust and Safety Forum

Centre for Advanced Technologies

- E.O. Wilson

“We have Paleolithic emotions, medieval institutions, and godlike technology.”



Trust and Safety at the Crossroads

Thank you!

Nate Persily
James B. McClatchy Professor of Law
Co-director, Stanford Cyber Policy Center
npersily@law.stanford.edu