THE LOGIC OF AI & DISINFORMATION ON SOCIAL MEDIA PLATFORMS

Dr. Courtney C. Radsch

Senior Fellow, Center for International Governance Innovation

Senior Technology Policy Advisor, Global Forum for Media Development

Senior US and Tech Policy Advisor, ARTICLE19

Fellow, Center for Media, Data & Society

Visiting Scholar, Center for Media at Risk, Annenberg, Univ. of Pennsylvania

Advisory Council, Ranking Digital Rights

Advisory Board, Dangerous Speech Project

Snapshot of AI & Disinfo on Social Media

Manipulation, Divisiveness, Inauthenticity

Identity groups, societal tensions/divisions, Misogyny, Racism & Ethnic/Religious Hatred

Gendered Disinformation

Rise of Alt- and Far-Right Populism

Overrepresentation of Far-Right Parties in Europe, US

Information Warfare & Foreign manipulation to influence domestic public opinion

Facebook Content Removals Using Detection with AI tools:

- 99.5% of terrorist-related
- 98.5 % of fake accounts
- 96% of adult nudity & sexual activity
- 86% of graphic violence-related

Global Internet Forum to Counter Terrorism (GIFCT) Hash Database

Digital Millennium Copyright Act (DMCA)

Coordinated Disinformation

70+ states/govts engage in information operations/Computational Propaganda

2017-2020: FB identified 150+ coordinated inauthentic behavior campaigns in more than **50 countries**

2020-2021: Twitter removes tens of Thousands of **state-aligned info ops**

commercialization + Industrialization of information & influence operations

AI-FUELED ATTACKS ON PUBLIC INTEREST: ELECTIONS, PUBLIC HEALTH, SAFETY & SECURITY

Individual and Collective harms

- Discrimination, exploitation, intimidation, self-censorship, sabotage, Recruitment, extremism, breakdown of social ties
- Curtail democratic deliberation, representation
- trust deficit & "liar's dividend"
- self-censorship and altered behavior resulting from pervasive surveillance → ↓ diversity → ↓ pluralism
- Moderation & censorship of authentic/legitimate info
- ↑ Polarization → ↓ Pluralism (lack of center)

Targets

- Vulnerable populations
- journalists; Female politicians; public figures
- Gendered Disinformation

Elections, Partisan politics, Protest & Democracy Movements

- **US**: 75% of users exposed to content from clickbait farms in **Macedonia & Kosovo** had never followed the page
- **Germany**: AfD had 5x FB engagement as other parties; 63% of women don't express political views online

Violent conflict

- Ukraine, Ethiopia, Myanmar
- Ethnic, Racial, Religious divisions

COVID19, anti-vaxxer movement

- 12 super-spreaders of disinformation
- Trust & Integrity
- Futility of COMO at scale

IMPACTS OF AI-FUELED DISINFORMATION ON JOURNALISM

- Targeting journalists /Media = denying them **agenda-setting power**
- **Framing** journalists & profession as "fake news"
- **Decertifying** journalists, particularly women, as legitimate actors in the public sphere
- Goal = **undermine viability & sustainability of independent journalism**
  - reinforce distrust in the media, Inoculate those in Power from oversight + accountability
  - delegitimize reporting on certain topics & by "types" of people
    - Controversial issues, social media manipulation, information operations
    - Gender, Minorities & Intersectionality
- Forcing logic of platform AI choices
  - Newsroom resources & priorities
    - FB video emphasis → news orgs pivot
  - Online harassment & Gender-Based Violence

Ukraine: Avg Troll Farm Salary = Avg FT employee ($365/mo)

Russia: Budget of Internet Research Agency est. $400,000/mo

RT online 24hr/da, >3 languages; reaches 700 million people in 100 countries

The Logic of AI in the Information Ecosystem

- AI is shaped by a corporate logic
  - Platforms & profit, whoever can pay + Growth maximization
  - Where resources are devoted: countries, languages, issues, etc.
  - Research priorities: private vs public interest
  - Harm identification, prioritization, reduction
- Algorithmic Intermediation
  - Designed based on this logic
  - Personalization
  - Only "see" what you look for or discover
- Agenda-setting power
- Framing
- Filling information voids

Limited Pluralism & New Gatekeepers

A few platforms determine the **logic** of the public sphere & freedom of expression

- Dominance of a few social media intermediaries = limited pluralism; they decide how information gets to be seen and shared – lack of alternative logics
- Susceptibility to influence by a handful of influential governments & their priorities
- Prone to manipulation: driven by political, economic, social factors

Gatekeeping: of expression [channeling, censoring, adding value, infrastructure, user interaction, editorial; relevance, recommending, trending topics] + scientific research & innovation

- Content moderation – challenging and prone to error & abuse at scale
- Shifts in who the gatekeepers are and what their logic of gatekeeping is
  - Journalism norms -> Platform Norms = AdTech

The Political-Economic Logic of AI & Disinfo

Monetization opportunities+ AI → Rise of troll farms, spammers, clickbait farms, plagiarism, micro-preneurs

- Without integrity evaluation, monetization attracts & rewards low-quality sources
- Drown out organic, legitimate news, opinion, and engagement
- Ease of account creation, content posting & regurgitation across platforms

- Certification and Equivalency
- Deep & Shallow Fakes:
  - Machine learning techniques → ↑ sophistication = more realistic, resistant to detection
  - cheaper, easier to create → more ubiquitous

Case Study: Philippines

- 2016: Duterte subject of 68% of all election-related discussions (46% for closest rival)
- Maria Ressa @ Rappler reports on info ops, FB manipulation → harassment & legal peril
  - Receives 2021 Nobel Prize
- firm he hired made $8 million from FB & Google before being taken down in 2019

Zoom In: Economic Logic of Adtech & Disinfo

- Pre-2015: FB click → Publishers (ads served by Google) → FB Instant Articles (FB serves ads)
- **2015**: Instant Articles in US & Europe
- **2016**: Instant Articles in Global South
- **2018**: $1.5 B paid out
- **2019**: FB begins checking publishers for content originality & demonetizing
- **2021**: ads embedded in live videos

Case Study: Myanmar

- **2015**: 6 of top 10 websites w/most FB engagement = legitimate publishers
- **2016**: Instant Articles rolled out
- **2017**: 2 of 10 top publishers = legitimate
- 2017 Military crackdown & anti-Muslim propaganda
- **2018**: 0 of top 10 publishers = legitimate
- 2018 UN determines violence = genocide
- 2021 Military Coup & Information Warfare, Social Media bans Ineffective
- **2021**: YouTube channels converted to FB articles & redistributed; some posted "live"

LEVERAGING LOGIC OF PUBLISHING & PLATFORMS
CORPORATE/PROFIT MOTIVE, NETWORK EFFECTS, DIFFUSION, VIRALITY, ENTRENCHMENT

- 1:1 + few:many + many:many
- Pervasive privatized surveillance + datafication → iterative collection and processing of massive amounts of data about people → big data, classification, machine learning, neural networks → prediction, promotion, discrimination
- Rapid diffusion with limited constraints vs. slow, often ineffective mitigation
- Virality → "information cascades"
- filter bubbles → confirmation bias
- echo chambers → polarization, ↓ diversity
- predictive engagement algorithms based on look-alike audiences → spread of disinformation + extremism

AI ENABLED DISINFORMATION: IMPACT ON THE INFORMATION ECOSYSTEM

- Inputs: data (sources/what counts as data/language), diversity & pluralism, Fairness, Bias
- Identification, targeting, surveillance
- Scope, Scale & Speed + sophistication, nuance
- Neural networks + machine learning + social media network effects
  - "generative adversarial networks"
- Algorithms identify, make visible, amplify content/accounts + predict engagement with content + Targeted Advertising & Promotion
- Algorithmic impacts: bots + human attraction to negative/novel information
  - People & their engagement behavior
  - Bots: Exposure + can manipulate predictive content engagement algorithms
  - Skewing prevalence, drowning out, filling information voids

THANK YOU

DR. COURTNEY C. RADSCH

CRADSCH@GMAIL.COM

WWW.MEDIATEDSPEECH.COM

@COURTNEYR