MITRE’s Experience

MITRE

- Not-for-profit
- Working in the public interest for 60+ years
- Building social media analytics for 14 years (first publication)
- Social technologies patented (4) and licensed (10) for the public interest

MITRE believes protecting the information and news infrastructure is a critical global challenge

A trusted third party

- Aviation Safety Information Analysis and Sharing Program (ASIAS)
- Serves as a central conduit for the exchange of data and analytics
- Establishes key safety benchmarks
- Results are shared with ASIAS participants (airlines)
  https://portal.asias.aero/web/guest

- ATT&CK is a knowledge base of adversary tactics and techniques based on real-world cyber security observations
- Used as a foundation for the development of threat models and methodologies
  https://attack.mitre.org/

MITRE is investing in addressing the intentional spread and amplification of disinformation

Joint Study with the Alliance for Securing Democracy (#1)
Impact of Inauthentic Behavior in Social Media on Anti-Feminist Issues as it Relates to the EU Elections

Gender-based violence (somewhat suspicious)

- La Pildora Roja “The Red Pill”
  - 2 May 2019

- Mereces Saberlo “You Deserve to Know”
  - 26 Feb 2019

- Feminismo Científico “Scientific Feminism”
  - 13 Dec 2018

- Mereces Saberlo
  - 4 Apr 2019

- Spain/ Spanish-centric antifeminist movement

Women’s/Men’s Rights (highly suspicious)

- The Red Pill/MGTOW “Men Going Their Own Way”
  - 29 Apr 2019

- Bot amplification of tweets and URLs

- “Real” gender-based violence data show that women are violent and lie about rape and abuse and “deserve to know” the true statistics

- “Nikola Tesla (hero)” “anti-abortion"

- “distorted wage gap” “anti-Gillette campaign”

- Backlash against feminism

- Misogynist Views
  - 13 Dec 2018

- Redditors

- Pan-EU antifeminist movements

- Men’s Rights Activism (Broader EU Narrative)
  - 15 Jan 2018

- Amplification of anti-feminist disinformation across social media platforms, countries, and languages

Anti-feminist disinformation hate speech across social media platforms, countries, and languages
Difficulty of Monitoring Environment for Harmful Content

For each topic...
- Inauthentic Accounts
- Promoting Crime
- Coordinating Harm
- Platform Abuse
- Privacy Infringement
- Pornography / Sexual Services
- Impersonation of Person / Brand
- Child Privacy / Exploitation
- Self Harm
- Abusive Behavior / Violent Threat

For each topic, modality, and platform!

Categories of Harm

Standardization of categories for Facebook and Twitter based on Terms of Service (see backup slide)

For each modality...

Each category has challenges. What are the definitions? What are the edge cases?

Platform Approach for Disinformation "Game"
Platform to Foster Shared Knowledge and Common Action Against Disinformation

Disinformation spectrum:
- Genuine (free speech)
- Divisive
- Harmful
- Criminal
- Intentional (weaponized bot army)

Inauthentic account/bot spectrum:
- Easy Bots created by anyone
- Curated Bot networks
- Easy “Sock puppets”
- Curated “Sock puppet”
- Human-Like

Influence Operations strategies:
- False news sites, Spread misinformation, Flood channels, Inauthentic accounts, Others...

Cost per Account:
- Easy Bots created by anyone
- Curated Bot networks
- Easy “Sock puppets”
- Curated “Sock puppet”
- Human-Like

Removal of inauthentic accounts will push adversaries to the bottom of the red triangle

Detection of Influence Operations Across Social Media Platforms
- A. Detect inauthentic accounts across social media platforms by topic (as done in example)
- B. Monitor environment for harmful content in text and media (next slide)
- C. Go beyond current terms of service to analyze data to detect influence operations
- D. Setting standards with social media platforms can improve detection

An Alliance with a Response Framework
- Develop industry wide standards
- Share threats across social media platforms
- Coordinate actions
- Share information across infrastructure sectors

# Data Snapshot: Facebook’s Community Standards Enforcement Report

**October 2017 – March 2019**

<table>
<thead>
<tr>
<th>Content Category</th>
<th>Views</th>
<th>Facebook Take Action</th>
<th>Violation Find</th>
<th>Action on Reported Violations</th>
<th>Restore after Removing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adult nudity and sexual activity</td>
<td>0.08%</td>
<td>0.07%</td>
<td>0.13%</td>
<td>0.10%</td>
<td>0.08%</td>
</tr>
<tr>
<td>Bullying and harassment</td>
<td>0.06%</td>
<td>0.07%</td>
<td>0.08%</td>
<td>0.11%</td>
<td>0.08%</td>
</tr>
<tr>
<td>Child nudity and sexual exploitation</td>
<td></td>
<td></td>
<td>0.03%</td>
<td>0.1%</td>
<td>0.97%</td>
</tr>
<tr>
<td>False accounts</td>
<td>5%</td>
<td></td>
<td>0.1%</td>
<td>0.3%</td>
<td>99.9%</td>
</tr>
<tr>
<td>Hate speech</td>
<td>0.04%</td>
<td>0.03%</td>
<td>0.12%</td>
<td>0.15%</td>
<td>0.01%</td>
</tr>
<tr>
<td>Regulated goods: drugs</td>
<td>0.06%</td>
<td>0.05%</td>
<td>0.11%</td>
<td>0.12%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Regulated goods: firearms</td>
<td>0.02%</td>
<td>0.02%</td>
<td>0.05%</td>
<td>0.06%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Spam</td>
<td>0.27%</td>
<td>0.24%</td>
<td>0.27%</td>
<td>0.23%</td>
<td>0.23%</td>
</tr>
<tr>
<td>Terroristic propaganda</td>
<td>0.02%</td>
<td>0.02%</td>
<td>0.12%</td>
<td>0.17%</td>
<td>0.03%</td>
</tr>
<tr>
<td>Violence and graphic content</td>
<td>0.05%</td>
<td>0.05%</td>
<td>0.06%</td>
<td>0.07%</td>
<td>0.07%</td>
</tr>
</tbody>
</table>

**Abusive Behavior / Violent Threat**

- **Good at child porn**
- **Bad at bullying**

**Where do we go from here?**

- **Content categories need to be standardized?**

Transatlantic Vision on Disinformation with the Atlantic Council
Need for Alliance in Addressing Global Disinformation

**Vision**

- Strengthen information/news infrastructure, businesses/economies, and democracies/countries by reducing the effectiveness of disinformation campaigns

**Members**

- Nonprofits
- Analytic Firms
- Fact Checkers
- Academics
- Sectors
- Social Media Platforms
- News Media
- In Discussions
- Others?

**What would the alliance do?**

- Archive data sets that are removed by social media platforms
- Earlier sharing of information on disinformation threats and work together, following cyber security model
- Perform cross social media platform analysis
- Develop industry-wide standards for addressing disinformation
- Use common data sets to assess threats
- Engage globally
- Engage in activities to empower the public
- Other?
Jennifer Mathieu, PhD
jmathieu@mitre.org
781-271-8672
From Facebook and Twitter Terms of Service

Note, Facebook enforcement report has different categories