Harmful Information Overview

Summary

What exactly is “harmful information”?

Attacks causing harms that stem from the use and/or abuse of information systems as they are designed and/or intended to be used.

- For example, a system might be designed to allow its users to spread hate speech to any user even though that behavior is against a community standards policy
- Major categories of abuses fall into misinformation and harassment

Learning Objectives

- Vocabulary and Classification of Harmful Information
- Approaches to Identifying, Prioritizing, and Mitigating Harmful Information

Pre-Readings

- Sarah Jeong, Charlie Warzel, Brianna Wu, Joan Donovan. New York Times. “Everything is GamerGate” [https://www.nytimes.com/interactive/2019/08/15/opinion/gamergate-twitter.html] - **Read all of the four essays.**
- IFTF “State-Sponsored Trolling: How Governments Are Deploying Disinformation as Part of Broader Digital Harassment Campaigns”. Read pages 3 to 21 & 45 to 51. [http://www.iftf.org/statesponsoredtrolling]
- Cindy Otis. USA Today. “Americans could be a bigger fake news threat than Russians in the 2020 presidential campaign” [https://www.usatoday.com/story/opinion/2019/07/19/disinformation-attacks-americans-threaten-2020-election-column/1756092001/]
- InterAction “Disinformation Toolkit.” [https://staging.interaction.org/documents/disinformation-toolkit/]

- Reply All podcast. “#112 The Prophet” Listen to or read transcript. [https://www.gimletmedia.com/reply-all/112-the-prophet]
- **(Optional)** Tahmina Ansari. First Draft. “This Muslim journalist embraced social media until it ‘ruined’ his life” [https://firstdraftnews.org/this-muslim-journalist-embraced-social-media-until-it-ruined-his-life/]

Resources

*** Mitigation Framework**

Activities

Beyond Hacking

How might Twitter be used to harm an organization even when the site is used as designed (ie not being “hacked”)?

Vocabulary and Classification Problems

Which category (Usually Acceptable, Sometimes/Borderline Acceptable, Always Unacceptable) do the following terms belong to: Propaganda – Disinformation – Misinformation – Malinformation – Internet Shutdowns - Harassment – Trolls – Bots – Doxxing – Mobbing – Swatting – Leaks – Sockpuppets – Astroturfing – Clickfarms - Deceptive Advertising – Exclusionary Advertising – Dog Whistles – Subtweeting – Parody News – Clickbait

Discussion

How much does intent behind harmful information matter for the organization?

How much does the truth of the harmful information matter for the organization?

Input

Harmful information threats can be considered as security risk management problems, however:

- Resources will be limited compared to scope of the problems
- Harms & risks are ill-defined for prioritization
- Less agreement on what ”best practice” looks like

Threats may have greater impact depending on the context, content, audience, medium, and the capabilities of an attacker to gain legitimacy, impersonate, link, amplify, collect, and suppress.

Harms may result regardless of Intent and Factual Accuracy of an attack. Intent is useful for anticipating escalation & future attacks. Truth will be leveraged by most attackers.

Organizations have to consider Direct Targeting and Indirect Threats to Individuals, Groups, Organizations, and Beyond the Organization as well as their own Ingestion and Generation of harmful information. Focus on Harms to Self-Determination, Reputation, Economic Situation, and Operations.

Practical “Solutions” for Civil Society:

1. Increase understanding / practices around holistic security
2. Integrate risk mitigation into existing systems and processes
3. Strengthen external relationships and collaboration

Deepening

Step through the Harmful Information Mitigation Framework to Case Study 1 or 2

- What are the harms or risks you find most important to address? (top 3)
- Which mitigations would you prioritize for implementation? (top 3)

Synthesis

- Is the juice worth the squeeze?

Assignments

Last update: April 27, 2020