Content Moderation

ICANN 72 Policy Forum, August 2021

Sai Prasad Chandrasekaran

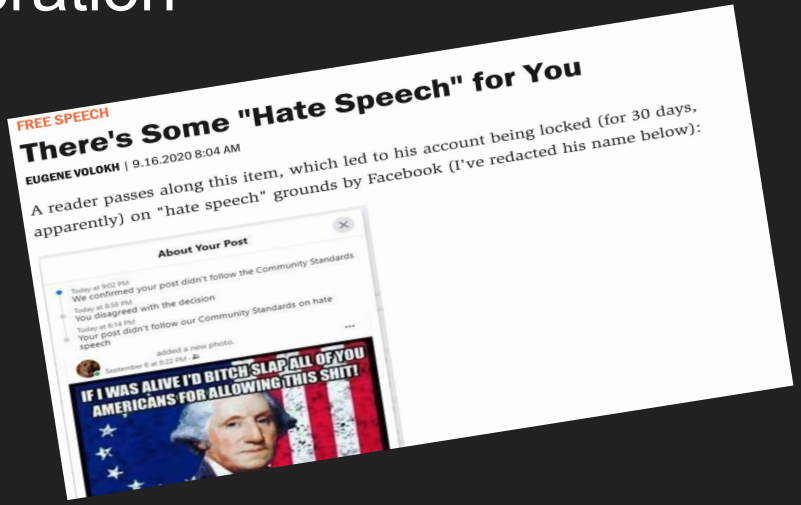
Bio

- Education : Graduate student of Cybersecurity at Indiana University

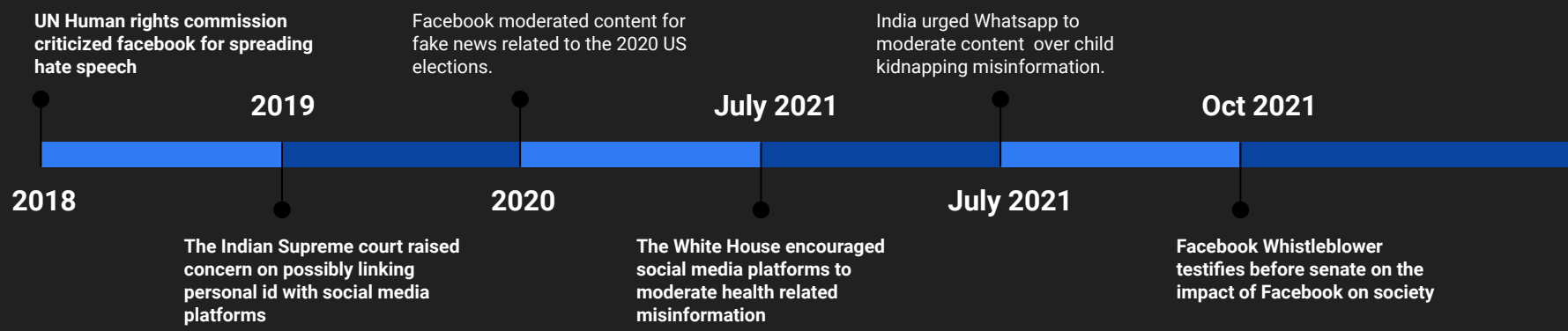
- Interests : Likes to explore ways to improve user privacy choices, manage risks and occasionally geek on privacy laws and judicial decisions

- Fun Fact: I love cricket (the game, not the insect :))

The buzz around content moderation



A timeline of content moderation



Content Moderation: A Privacy Risk Assessment

Privacy Risk[1]	Description	Threat
Breach of Confidentiality	Exposure of content in an encrypted system is a breach of user's confidentiality	Identity theft Loss of anonymity
Surveillance	Eavesdropping on private content by malicious or state actors	Loss of freedom Impedance of free speech Chilling Effects
Exposure / Intrusion	Leak of sensitive personal information to unintended recipients	Possibility to blackmail and mental harm especially when certain health information is subject to taboo (substance abuse, sexual health etc.)
Disclosure	Leaking of certain types of information might inflict emotional and physical harm	Disclosure of certain types of information is irreversible like health data

1 - Based on Daniel Solove's taxonomy of Privacy

A Privacy Preserving Approach to Content Moderation

Technical

Analysis of meta data at the client side ensures not storing information on the server side and guarantees E2EE

Use of cryptographic techniques such as secure multi party computation can flag for instance images such as true (match) or false (no match). The false images can remain encrypted and hidden from the service provider

Process

User reporting is a crucial aspect to tackling content moderation. Creation of a reporting and feedback channel for the user provides a legitimate basis for the service provider to access content

Principle

Content moderators must provide “algorithmic transparency” to their users and in the public domain to win trust

Combating Content Moderation: Multi-stakeholder Approach

- **Individuals, Families and Communities:** Verify the accuracy of the information with trustworthy and credible sources
- **Educators and Educational Institutions:** Media, science, digital, data and health literacy program implemented across all settings
- **Tech Platforms**: Address Information deficits and prioritize early detection of super spreaders and repeated offenders in a **privacy preserving manner**
- **Governments**: Convene local, federal, private, non-profit and federal partners to find common ground for appropriate legal and regulatory measures