Welcome!

This is the second of four events we are holding on encryption, privacy and safety.
Today’s Agenda

10:00-10:20  Introductions and Welcome Remarks

10:20-12:00  Presentations x 10

12:00-12:20  Break / Grab Lunch

12:20-2:00  Lunch & Working Sessions
         ●  Research: Future research directions and collaborations
         ●  Privacy: Red-lines and slippery slopes
         ●  Design: Product designs that reduce individual harm

2:00-3:00   Wrap up discussion
Our goals

Make new introductions

Spark research collaborations

Set an agenda for moving forward
Conversation is twisted by abuse context collapse

Encryption technology is a very important part of maintaining security and personal privacy. It protects personal information by using a scrambled code and allows users to send sensitive financial or other information over the internet.

At the same time, it is equally important to acknowledge that law enforcement has historically been able, under strict legal requirements, to intercept and obtain communications in order to investigate crimes.

The single most important criminal justice challenge in the last ten years is, in my opinion, the use of mobile devices by bad actors to plan, execute, and communicate about crimes. Just as ordinary citizens rely on digital communication, so do people involved in terrorism, cyber fraud, murder, rape, robbery, and child sexual assault.

For this reason, lawful, court-ordered access to these communications has become essential for us to prevent crime, to hold people accused of crimes accountable, and to exonerate the innocent.
Law enforcement access is not the only option!

How do we know? All kinds of abuse mitigations currently happen in “unencrypted space” without equivalent of lawful access.

- Spam/malware detection
- Fake account detection
- Grooming detection
- Voluntary reporting by victims
- Rate limiting/group size limiting

Almost all proactive online policing is about harm reduction, not investigation and prosecution.
So how do we move forward?

Break the problem into addressable steps:

1. Tease out different types of abuse
2. Look into factors that make each abuse more harmful
3. Consider mitigations that target those factors
4. Research liberty/privacy protections and transparency measures
5. Weigh mitigations by liberty and privacy impact
6. Perform fundamental research necessary for mitigations
7. Gather empirical evidence of efficacy of mitigations
<table>
<thead>
<tr>
<th>Abuse</th>
<th>Prevalence</th>
<th>Impact</th>
<th>Illegal</th>
<th>Victim in convo?</th>
<th>Image based</th>
<th>Amplification -&gt; harm</th>
<th>Metadata useful to LE</th>
<th>ML on content possible</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spam</td>
<td>Very High</td>
<td>Low</td>
<td>No</td>
<td>Yes</td>
<td>Sometimes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Malicious documents/links</td>
<td>High</td>
<td>High</td>
<td>Yes</td>
<td>Yes</td>
<td>Sometimes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Targeted harassment</td>
<td>High</td>
<td>Moderate-High</td>
<td>Rarely</td>
<td>Yes</td>
<td>Sometimes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>CSAM trading</td>
<td>High</td>
<td>High</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Sextortion/Grooming</td>
<td>Moderate</td>
<td>High</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Live abuse</td>
<td>Low</td>
<td>High</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Maybe</td>
</tr>
<tr>
<td>Incitement to violence</td>
<td>Low</td>
<td>High</td>
<td>Sometimes</td>
<td>No</td>
<td>Sometimes</td>
<td>Yes</td>
<td>Maybe</td>
<td>Yes</td>
</tr>
<tr>
<td>Disinformation</td>
<td>Moderate</td>
<td>Low</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Sometimes</td>
<td>Yes</td>
<td>Maybe</td>
</tr>
<tr>
<td>NCII</td>
<td>Moderate</td>
<td>High</td>
<td>Sometimes</td>
<td>Sometimes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Only with priming from victim</td>
</tr>
<tr>
<td>Criminal conspiracy</td>
<td>Low</td>
<td>High</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Potential Mitigation</td>
<td>Description</td>
<td>Potentially targeted abuses</td>
<td>Impact on liberty/privacy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------------------------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------</td>
<td>--------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Limits to group size</td>
<td>Capping group sizes, or reducing certain invite mechanisms (like external URLs) at certain sizes</td>
<td>Disinformation, incitement, spam, malware</td>
<td><strong>Low.</strong> Would handicap individuals (such as independent journalists) who rely on platforms to reach broad audiences</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Limits to forwarding/amplification</td>
<td>Setting limits on the amount of times a user can forward a message or a message can be forwarded</td>
<td>Disinformation, incitement, spam, malware</td>
<td><strong>Low.</strong> Would impinge on speech whether it abuses ToS or not. Might create traceability.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Client-side image fingerprinting</td>
<td>Push image hashsets to clients to look for banned images</td>
<td>CSAM trading, NCII, disinformation, incitement</td>
<td><strong>Low-High.</strong> Depends on what actions are automatically taken. Is the image prevented from being sent, user prompted to report or automatically reported?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Client-side ML detection and guided reporting</td>
<td>Use ML on the client (with server-side model training) to detect potential abuse, prompting the user to securely report the conversation/group</td>
<td>Targeted harassment, disinformation, incitement, spam, malware, CSAM trading, NCII, sextortion/grooming</td>
<td><strong>Moderate.</strong> Has a human user in the loop, but could be used by authoritarians to prompt users to report for political content.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Detection and removal of fake accounts with metadata based ML</td>
<td>Using ML to identify inauthentic behavior patterns and registration data</td>
<td>All</td>
<td><strong>Low.</strong> Will have irreducible FPs, might require retention of more metadata</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Historical metadata available with process</td>
<td>Storing conversational metadata and providing it with lawful process (e.g. who is in which groups, when messages are sent)</td>
<td>CSAM trading, sextortion, criminal conspiracies, incitement,</td>
<td><strong>High.</strong> Creates potential for law enforcement overreach, corporate surveillance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Personal opinions on the way forward

- Lawmakers and law enforcement will have to expand their thinking on harm reduction versus traditional investigation and prosecution.
- Cryptography and civil liberties communities should not be blasé about the ability to hold back the tide without mitigations.
- We need to separate privacy from liberty concerns.
- There are some obvious ways forward that are reasonably low risk. My favorite is client-side ML for abuses targeted at the user and enhanced reporting.
- Lots of fundamental research needed, will require industry-academic cooperation.
Presenters (hyperlinks to slides)

- Kate Klonick
- Jon Millican
- Matthew Green
- Matthias Marx
- Aviv Ovadya
- Erica Portnoy
- Nicholas Weaver
- Hany Farid
- Jon Callas
- Farzaneh Badii
Mitigating Abuse in an End-to-End World

Wrap-up Discussion
Research Projects and Collaborations

● NCMEC measurement and data science problem
  ○ Thorn

● Identity: detecting when a child is interacting with an adult
  ○ Deeptrust Alliance

● Client-based image and video hashing
  ○ Privacy-preserving
  ○ Resilience to attacks
  ○ Classes and resistance to attacks
  ○ Auditable
  ○ Cryptographic techniques for server-side detection of image hashing in an e2e world
  ○ Scaling

● Detection of incitement to real-world harm

● Bridging the gap between policy seeking to end harms and law enforcement
  ○ Throwing sand on the slippery slope
Research Projects and Collaborations (continued)

- What techniques are other governments and law enforcement agencies using to address these problems?
  - Tools apart from content-scanning
  - Threat modeling, for individuals and geopolitically

- What are the known effective techniques for combatting specific types of abuse?
  - What is the status quo? What are companies doing right now?

- Detecting documents and manuals associated with CSAM

- Identifying bad actors online

- A holistic view: policy in addition to technical measures
  - Including survivors’ voices

- Testable UX interventions
  - What are the effects of overzealous client-side scanning?

- Peer-to-peer systems, decentralized systems
Research Projects and Collaborations (continued)

- Are there legal changes that would help companies do more?
  - E.g. legal requirement to report, which might incentivize companies not to look
  - Structuring policy to treat age groups differently
    - Current incentive is not to determine age
Mitigating Abuse in an End-to-End World
A workshop of the Stanford Internet Observatory

Saturday, January 11, 2020
Craig Newmark Journalism School • City University of New York

Agenda

9:30-10:00  Coffee and Registration

10:00-10:15  Introductions and Welcome Remarks
Alex Stamos, Director, Stanford Internet Observatory

10:00-11:30  Presentations
Kate Klonick, Assistant Professor, St. John’s University Law School
Jon Millican, Software Engineer, Facebook
Matthew Green, Associate Professor, Computer Science, Johns Hopkins University
Matthias Marx, Research Associate, University of Hamburg
Aviv Ovadya, Founder, The Thoughtful Technology Project
Erica Portnoy, Staff Technologist, Electronic Frontier Foundation
Nicholas Weaver, Researcher, International Computer Science Institute at the University of California, Berkeley
Hany Farid, Professor, University of California, Berkeley
Jon Callas, Senior Technology Fellow, American Civil Liberties Union
Farzaneh Badii, Research Scholar, Yale Law School

11:30-12:00  Q&A and Working Sessions discussion

12:00-12:20  Break / Grab Lunch

12:20-2:00  Lunch & Working Sessions

- Research: Future research directions and collaborations - This group will discuss the potential avenues for future research and collaboration.
  Suggested opening questions:
  - How can be empirically understand the impact of online harms and the effectiveness of specific responses?
  - What emerging research areas (like homomorphic algorithms or MPC) are applicable to the problem set?
  - How will we seed useful research in this area?
- **Privacy: Red-lines and slippery slopes** - This group will discuss the definitions and framing around the privacy benefits we want from E2EE and the compromises that might be acceptable in preventing certain harms.
  
  *Suggested opening questions:*
  - How do we conceptualize the difference in privacy from external forces versus the limitation of choices internal to a product?
  - What are red-lines different groups would have for mitigations? Can those red-lines be enforced via cryptographic design?
  - What transparency mechanisms might be available to ensure that protections for specific abuses aren’t expanded into other areas, like political speech?

- **Design: Product designs that reduce individual harm** - This group will discuss the current state-of-the art of our understanding of how online communications products can be used to cause harm and currently available mechanisms or design criteria that can be deployed today.
  
  *Suggested opening questions:*
  - What do we know about the relationship between specific design choices in discovery and amplification and online harms?
  - Do we have any data of the impact of recent responses by WhatsApp and other companies?
  - Are there good models available for emerging E2EE products to emulate?

2:00-3:00  **Wrap up discussion**