An Introduction to Trust and Safety

CS 152 — Lecture 1

Alex Stamos

Stanford Internet Observatory

January 26, 2023

Stanford | Internet Observatory
Cyber Policy Center

What will we learn today?

Today, we will...

- Discuss the practical aspects of this class
- Explore how Trust and Safety differs from other areas of tech risk
- Start to understand the Trust and Safety lifecycle

TODO

Our Agenda

Mon, Jan 3 - Introduction to Trust and Safety
Wed, Jan 5 - How Tech Companies Work. Designing for Trust, Safety and Privacy
Mon, Jan 10 - Authentication and Identity
Wed, Jan 12 - Guest Lecture: Fraud
Mon, Jan 17 - HOLIDAY
Wed, Jan 19 - Spam, Fraud and Cybercrime
Mon, Jan 24 - Surveillance, Government Oppression and Domestic Abuse
Wed, Jan 27 - Harassment, Bullying and Threatening Behavior
Mon, Jan 31 - Hate Speech and Extremism
Wed, Feb 2 - Incitement and Terrorism
Mon, Feb 7 - Suicide and Self-Harm
Wed, Feb 9 - Child and Adult Sexual Exploitation I - CSAM, NCII and Responses
Mon, Feb 14 - Child and Adult Sexual Exploitation II - Grooming, Sextortion and Trafficking
Wed, Feb 16 - Working with Law Enforcement and CSE Case Studies
Mon, Feb 21 - HOLIDAY
Wed, Feb 23 - Misinformation and Disinformation
Mon, Feb 28 - Case Study: The 2016 and 2020 US Elections
Wed, Mar 2 - Guest Lecture: TBD
Mon, Mar 7 - Content Moderation and Resiliency
Wed, Mar 9 - Sharing Economy, Emerging Issues and Career Advice
Mar 14-18 - Final Presentations, To Be Scheduled

Who should take this class?

Students who:

1. Are interested in the ways technology can be abused to cause harm
2. Want to build consumer internet products more safely
3. Who are interested in careers in Trust and Safety, anti-abuse NGOs, law enforcement or internet policy
4. Who can participate in the group project at a 106B level

We now have a partner class, POLISCI 243C: The Politics of Internet Abuse

Project Teams + Sections

Project teams

- Teams of four CS students, one POLISCI
- Class intro form will be sent out tomorrow on Ed
  - Denote whether you want to be matched into a group OR state your group
- Fill those out by Friday at noon - teams will be announced shortly after

Sections will start next week

- TA facilitated work sessions
- Optional, but highly encouraged
- Starting week 2 - times being finalized (check syllabus + Discord)
- You do not have to officially sign up, but your whole group should be able to attend together

Grading

Course Grading

- Lecture participation and pre-lecture quizzes - 30%
- Project Milestone 1 - 20%
- Project Milestone 2 - 20%
- Project - Final presentation - 30%

Final Project

- Milestone 1: User Studies and Reporting Flow Design (20%) [Due 2/4]
- Milestone 2: Content Moderation App Implementation (20%) [Due 2/25]
- Final: Incident Response and Final Preso (30%) [Due 3/11]

Pre-Reading

Pre-reads and short reading quizzes before **every** class

- Pre-reads will go up a week in advance
- Quizzes will go up after previous class

Readings can be found here

Attendance Policy

- Students are welcome to take the class for credit, even if they can’t attend the lectures synchronously. I will not boot you out for non-attendance, but I also will not make any grading exceptions.
- 30% of your grade will come from class participation. Half from synchronous attendance quizzes that will only be available at the start of the lecture (and gaming this would be an Honor Code violation) and half from asynchronous quizzes on the pre-reads.
- This means that a student who attends zero lectures can get a max of 85% in the class. That’s a B.
- **I strongly suggest that students who can’t make most of the lectures take the class C/NC.** If you take the asynchronous quizzes and your team completes the project, there is really no chance you won’t pass (>60%).

Dealing with difficult content

The subject matter of this course can be difficult intellectually and emotionally. We will read about and discuss difficult topics, including (but not limited to) sexual exploitation of adults and minors, harassment, bullying, hate speech, domestic abuse, terrorism, and more.

If you anticipate acute distress as a result of encountering a particular topic, talk to me ahead of time to arrange an alternative written assignment in lieu of your in-class participation. If you become so distressed that you need to leave during class, feel free to do so. If you need to leave a class, talk to me afterward and we can arrange an alternate assignment. I will not “warn” students about particular topics, because sensitivity to different topics varies from person to person, and because topics may arise unexpectedly in class discussion. Please refer to the course agenda to see the list of course topics.

Additionally, as you may know, there is a difference between being triggered (in the sense of post-traumatic stress disorder) and feeling uncomfortable. One of the goals of this class is to help students develop empathy for victims of online abuse. Feeling uncomfortable (and sometimes even angry or offended) is part of intellectual growth. Feeling triggered or psychologically traumatized is not. Please take care of yourselves and each other, and let me know if I can do anything at all to help.

An example of difficult content from a previous project

Unique Aspects of Trust and Safety

1. The study of how people abuse the internet to cause harm.
2. Often using products the way they are designed to work.
3. Crosses between specialties. Requires understanding of society and humanity.
4. Is dynamic and unpredictable.

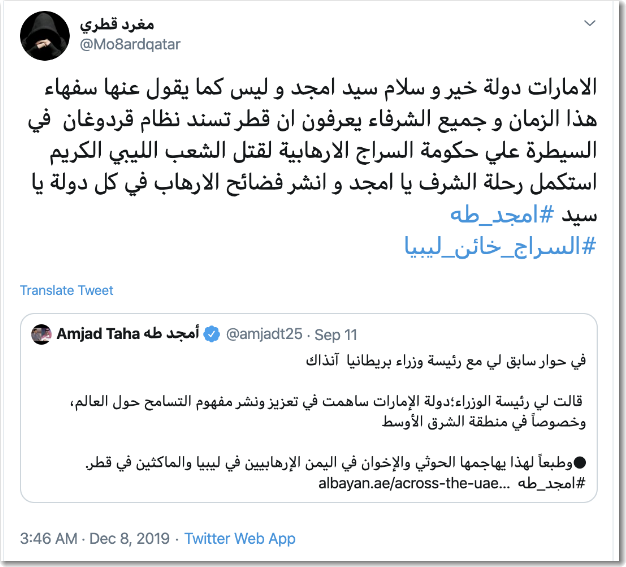
The Biggest Challenges in Trust and Safety

1. Scale
2. Non-diverse studies and solutions
3. Measurement and definition challenges
4. Privacy vs Safety
5. Information sharing and division of responsibility
6. Government vs private action
7. Fairness in ML solutions
8. Freedom of expression

Other concerns

- Four
- Five
- Six

Some problems arise



مغرد قطري
@Mo8ardqatar

الامارات دولة خير و سلام سيد امجد و ليس كما يقول عنها سفهاء
هذا الزمان و جميع الشرفاء يعرفون ان قطر تسند نظام قردوغان في
السيطرة علي حكومة السراج الارهابية لقتل الشعب الليبي الكريم
استكمل رحلة الشرف يا امجد و انشر فضائح الارهاب في كل دولة يا
سيد #امجد_طه
#السراج_خائن_ليبيا

Translate Tweet

Amjad Taha أمجد طه @amjadt25 · Sep 11

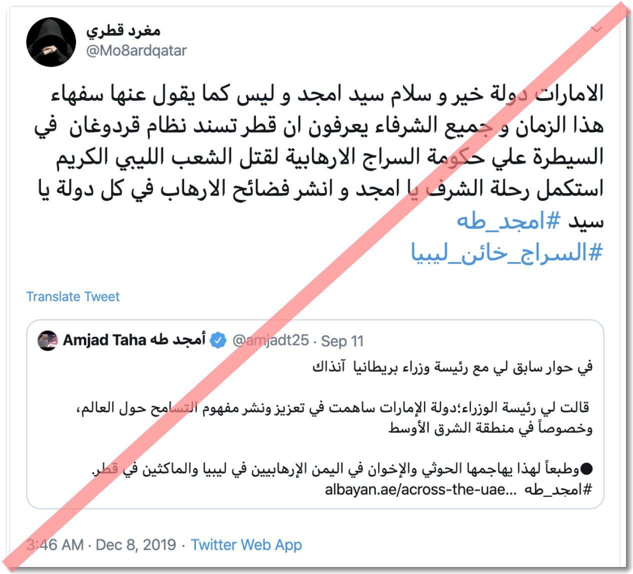
في حوار سابق لي مع رئيسة وزراء بريطانيا آنذاك

قالت لي رئيسة الوزراء؛دولة الإمارات ساهمت في تعزيز ونشر مفهوم التسامح حول العالم،
وخصوصاً في منطقة الشرق الأوسط

●وطبعاً لهذا يهاجمها الحوثي والإخوان في اليمن الإرهابيين في ليبيا والماكثين في قطر.
#امجد_طه albayan.ae/across-the-uae...

3:46 AM · Dec 8, 2019 · Twitter Web App

Information not always true 😭



مغرد قطري
@Mo8ardqatar

الامارات دولة خير و سلام سيد امجد و ليس كما يقول عنها سفهاء
هذا الزمان و جميع الشرفاء يعرفون ان قطر تسند نظام قردوغان في
السيطرة علي حكومة السراج الارهابية لقتل الشعب الليبي الكريم
استكمل رحلة الشرف يا امجد و انشر فضائح الارهاب في كل دولة يا
سيد #امجد_طه
#السراج_خائن_ليبيا

Translate Tweet

Amjad Taha أمجد طه @amjadt25 · Sep 11

في حوار سابق لي مع رئيسة وزراء بريطانيا آنذاك

قالت لي رئيسة الوزراء؛دولة الإمارات ساهمت في تعزيز ونشر مفهوم التسامح حول العالم،
وخصوصاً في منطقة الشرق الأوسط

●وطبعاً لهذا يهاجمها الحوثي والإخوان في اليمن الإرهابيين في ليبيا والماكثين في قطر.
#امجد_طه albayan.ae/across-the-uae...

3:46 AM · Dec 8, 2019 · Twitter Web App