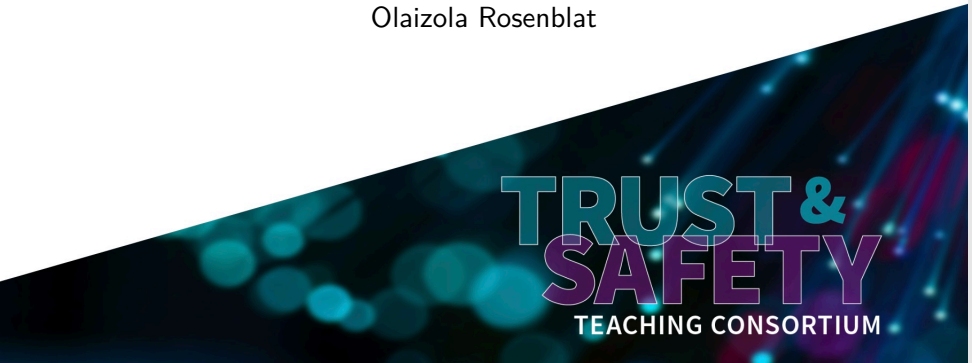
Introduction to Trust & Safety

Peter Carragher, adapted from Camille François and Mariana Olaizola Rosenblat



Introduction to Trust & Safety

Peter Carragher, adapted from Camille François and Mariana Olaizola Rosenblat

Learning objectives

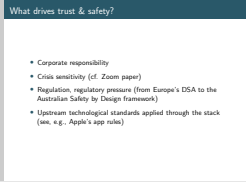
- Learn about the purpose and history of trust & safety
- Learn about approaches to trust & safety

Learning objectives

Today we will:
- Learn about the purpose and history of trust & safety
- Learn about approaches to trust & safety

What drives trust & safety?

- Corporate responsibility
- Crisis sensitivity (cf. Zoom paper)
- Regulation, regulatory pressure (from Europe's DSA to the Australian Safety by Design framework)
- Upstream technological standards applied through the stack (see, e.g., Apple's app rules)



1. "Trust and safety is the study of how people abuse the internet to cause real human harm, often using [online] products the way they are designed to work" (Journal of Online Trust & Safety).
2. Trust and Safety is also a practice and a field within technology companies that is concerned with the reduction, prevention, and mitigation of online harms. Per the Trust & Safety Professional Association: "As internet communities, online services, and the use of digital technologies to mediate our daily lives and interactions have continued to grow, technology companies have needed to determine the kinds of content and behaviors that are appropriate and those that are not. The teams that handle this responsibility often fall under the general term 'trust and safety.'" Source: App store review guidelines, https://developer.apple.com/app-store/review/guidelines/

High-level taxonomy of relevant abuses

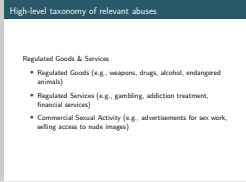
- Dangerous Organizations (e.g., extremist groups, criminal organizations)
- Violence (e.g., explicit threats, bomb-making instructions)
- Child Abuse & Nudity (e.g., child sexual abuse material, solicitation of minors)
- Sexual Exploitation (e.g., non-consensual sex acts, sextortion)
- Human Exploitation (e.g., human trafficking, forced marriage)

High-level taxonomy of relevant abuses
Violent & Criminal Behavior
- Dangerous Organizations (e.g., extremist groups, criminal organizations)
- Violence (e.g., explicit threats, bomb-making instructions)
- Child Abuse & Nudity (e.g., child sexual abuse material, solicitation of minors)
- Sexual Exploitation (e.g., non-consensual sex acts, sextortion)
- Human Exploitation (e.g., human trafficking, forced marriage)

1. Adapted from TSPA page on "Abuse Types" in T&S: https://www.tspa.org/curriculum/ts-fundamentals/policy/abuse-types/ Summarized by ChatGPT

High-level taxonomy of relevant abuses

- Regulated Goods (e.g., weapons, drugs, alcohol, endangered animals)
- Regulated Services (e.g., gambling, addiction treatment, financial services)
- Commercial Sexual Activity (e.g., advertisements for sex work, selling access to nude images)



1. Adapted from TSPA page on "Abuse Types" in T&S: https://www.tspa.org/curriculum/ts-fundamentals/policy/abuse-types/ Summarized by ChatGPT

High-level taxonomy of relevant abuses

- Hateful Content (e.g., slurs, support for supremacy movements, mockery of victims)
- Graphic & Violent Content (e.g., imagery of fatal incidents, dismembered bodies, animal cruelty)
- Nudity & Sexual Activity (e.g., pornography, explicit art)

High-level taxonomy of relevant abuses
Offensive & Objectionable Content
- Hateful Content (e.g., slurs, support for supremacy movements, mockery of victims)
- Graphic & Violent Content (e.g., imagery of fatal incidents, dismembered bodies, animal cruelty)
- Nudity & Sexual Activity (e.g., pornography, explicit art)

1. Adapted from TSPA page on "Abuse Types" in T&S: https://www.tspa.org/curriculum/ts-fundamentals/policy/abuse-types/ Summarized by ChatGPT

High-level taxonomy of relevant abuses

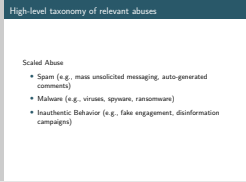
- Suicide & Self Harm (e.g., intention to self harm, encouraging self harm, instructions)
- Harassment & Bullying (e.g., hateful conduct, dogpiling, blackmail threats, doxxing)
- Dangerous Misinformation & Endangerment (e.g., conspiracy theories, false safety info, dangerous challenges)

High-level taxonomy of relevant abuses
User Safety
- Suicide & Self Harm (e.g., intention to self harm, encouraging self harm, instructions)
- Harassment & Bullying (e.g., hateful conduct, dogpiling, blackmail threats, doxxing)
- Dangerous Misinformation & Endangerment (e.g., conspiracy theories, false safety info, dangerous challenges)

1. Adapted from TSPA page on "Abuse Types" in T&S: https://www.tspa.org/curriculum/ts-fundamentals/policy/abuse-types/ Summarized by ChatGPT

High-level taxonomy of relevant abuses

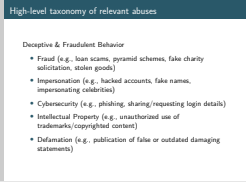
- Spam (e.g., mass unsolicited messaging, auto-generated comments)
- Malware (e.g., viruses, spyware, ransomware)
- Inauthentic Behavior (e.g., fake engagement, disinformation campaigns)



1. Adapted from TSPA page on "Abuse Types" in T&S: https://www.tspa.org/curriculum/ts-fundamentals/policy/abuse-types/ Summarized by ChatGPT

High-level taxonomy of relevant abuses

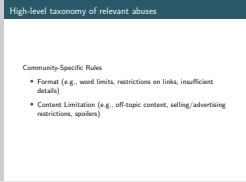
- Fraud (e.g., loan scams, pyramid schemes, fake charity solicitation, stolen goods)
- Impersonation (e.g., hacked accounts, fake names, impersonating celebrities)
- Cybersecurity (e.g., phishing, sharing/requesting login details)
- Intellectual Property (e.g., unauthorized use of trademarks/copyrighted content)
- Defamation (e.g., publication of false or outdated damaging statements)



1. Adapted from TSPA page on "Abuse Types" in T&S: https://www.tspa.org/curriculum/ts-fundamentals/policy/abuse-types/ Summarized by ChatGPT

High-level taxonomy of relevant abuses

- Format (e.g., word limits, restrictions on links, insufficient details)
- Content Limitation (e.g., off-topic content, selling/advertising restrictions, spoilers)



1. Adapted from TSPA page on "Abuse Types" in T&S: https://www.tspa.org/curriculum/ts-fundamentals/policy/abuse-types/ Summarized by ChatGPT

High-level taxonomy of relevant abuses

High-level taxonomy of relevant abuses

Key point: Relevant abuse types depend on your audiences, feature, product, etc.

1. Adapted from TSPA page on "Abuse Types" in T&S: https://www.tspa.org/curriculum/ts-fundamentals/policy/abuse-types/ Summarized by ChatGPT

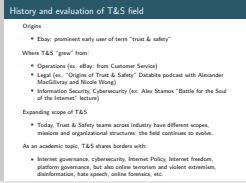
History and evaluation of T&S field

- Ebay: prominent early user of term “trust & safety”

- Operations (ex. eBay: from Customer Service)
- Legal (ex. “Origins of Trust & Safety” Databite podcast with Alexander MacGilivray and Nicole Wong)
- Information Security, Cybersecurity (ex: Alex Stamos “Battle for the Soul of the Internet” lecture)

- Today, Trust & Safety teams across industry have different scopes, missions and organizational structures: the field continues to evolve.

- Internet governance, cybersecurity, Internet Policy, Internet freedom, platform governance, but also online terrorism and violent extremism, disinformation, hate speech, online forensics, etc.


History and evaluation of T&S field

Origins

- Ebay: prominent early user of term “trust & safety”

Where T&S “grew” from:

- Operations (ex. eBay: from Customer Service)
- Legal (ex. “Origins of Trust & Safety” Databite podcast with Alexander MacGilivray and Nicole Wong)
- Information Security, Cybersecurity (ex: Alex Stamos “Battle for the Soul of the Internet” lecture)

Expanding scope of T&S

- Today, Trust & Safety teams across industry have different scopes, missions and organizational structures: the field continues to evolve.

As an academic topic, T&S shares borders with:

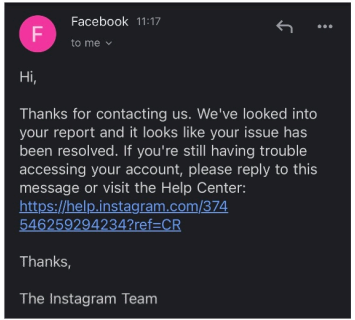
- Internet governance, cybersecurity, Internet Policy, Internet freedom, platform governance, but also online terrorism and violent extremism, disinformation, hate speech, online forensics, etc.

1. Cryst, E., Grossman, S., Hancock, J., Stamos, A., & Thiel, D. (2021). Introducing the Journal of Online Trust and Safety. Journal of Online Trust and Safety, 1(1). Retrieved from https://tsjournal.org/index.php/jots/article/view/8

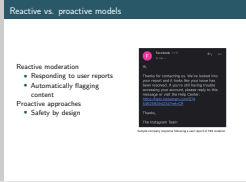
Reactive vs. proactive models

- Responding to user reports
- Automatically flagging content

- Safety by design



Sample company response following a user report of T&S violation



Managing trade-offs

- Privacy and safety may sometimes be at odds
- Restrictive platform features that require less moderation v. open features that require more moderation
- More false positives or more false negatives

Managing trade-offs
- Privacy and safety may sometimes be at odds
- Restrictive platform features that require less moderation v. open features that require more moderation
- More false positives or more false negatives

Gaining senior management support

- Educating senior management about issues
- Making the case for why trust & safety is related to core product mission
- Involving senior management in important edge case decisions

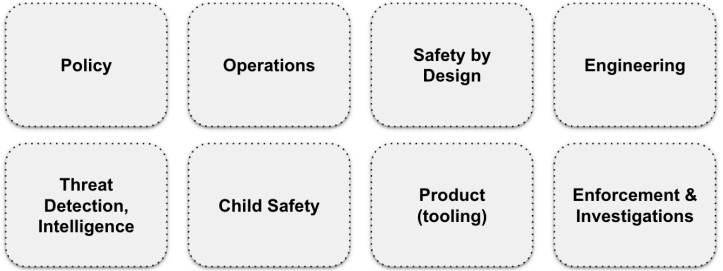
Gaining senior management support
- Educating senior management about issues
- Making the case for why trust & safety is related to core product mission
- Involving senior management in important edge case decisions

Building a T&S team

- Components of a T&S team
- Types of T&S teams
- Types of T&S professionals

Building a T&S team
Reactive moderation
• Components of a T&S team
• Types of T&S teams
• Types of T&S professionals

Sample functions



Sample functions

Discussion: Contrasting perspectives on building T&S teams

- What resonated from the stories collected and shared by Alex Feerst?
- What are themes echoed by Feerst, by Zoom and Pinterests' papers, and by Nicole Wong + Alexander MacGilivray?
- What are significant differences emerging from these testimonies?

Discussion: Contrasting perspectives on building T&S teams
- What resonated from the stories collected and shared by Alex Feerst?
- What are themes echoed by Feerst, by Zoom and Pinterests' papers, and by Nicole Wong + Alexander MacGilivray?
- What are significant differences emerging from these testimonies?

Overview of automated technologies

- Digital hash technology
- Image recognition tools
- Metadata filtering
- Natural language processing (NLP) classifiers

Overview of automated technologies
- Digital hash technology
- Image recognition tools
- Metadata filtering
- Natural language processing (NLP) classifiers

1. "These tools can be deployed across a range of categories of content and media formats, as well as at different stages of the content lifecycle, to identify, sort, and remove content." Source: https://www.newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-user-generated-content/how-automated-tools-are-used-in-the-content-moderation-process/