AGENDA

Thursday, September 29, 2022

8:00-8:45 **Registration and Continental Breakfast**

8:45-9:00 **Opening Remarks** (McCaw Hall)
Alex Stamos, Stanford Internet Observatory
Charlotte Willner, Trust & Safety Foundation

9:00-10:00 **Fireside Chat** (McCaw Hall)
Julie Cordua, Thorn
Alex Stamos, Stanford Internet Observatory

10:00-10:30 **Break**

10:30-12:00 **Journal of Online Trust & Safety Panel I** (McCaw Hall)
Presentations of five papers published in the Journal of Online Trust and Safety.
Moderated by Jeffrey Hancock, Stanford University

- Procedural Justice and Self Governance on Twitter - Unpacking the Experience of Rule Breaking on Twitter
  Tom Tyler, Yale Law School
- How to Build a Trust and Safety Team in a Year: A Practical Guide From Lessons Learned (So Far) At Zoom
  Josh Parecki and Chanel Cornett, Zoom
- American Parents' Perceptions of Child and Youth Explicit Image Sharing
  Michael Seto, The Royal's Institute of Mental Health Research
- Assessing the Political Motivations Behind Ransomware Attacks
  Karen Nershi, Stanford Internet Observatory
- Auditing Google's Search Headlines as a Potential Gateway to Misleading Content: Evidence from the 2020 US Election
  Himanshu Zade and Morgan Wack, University of Washington

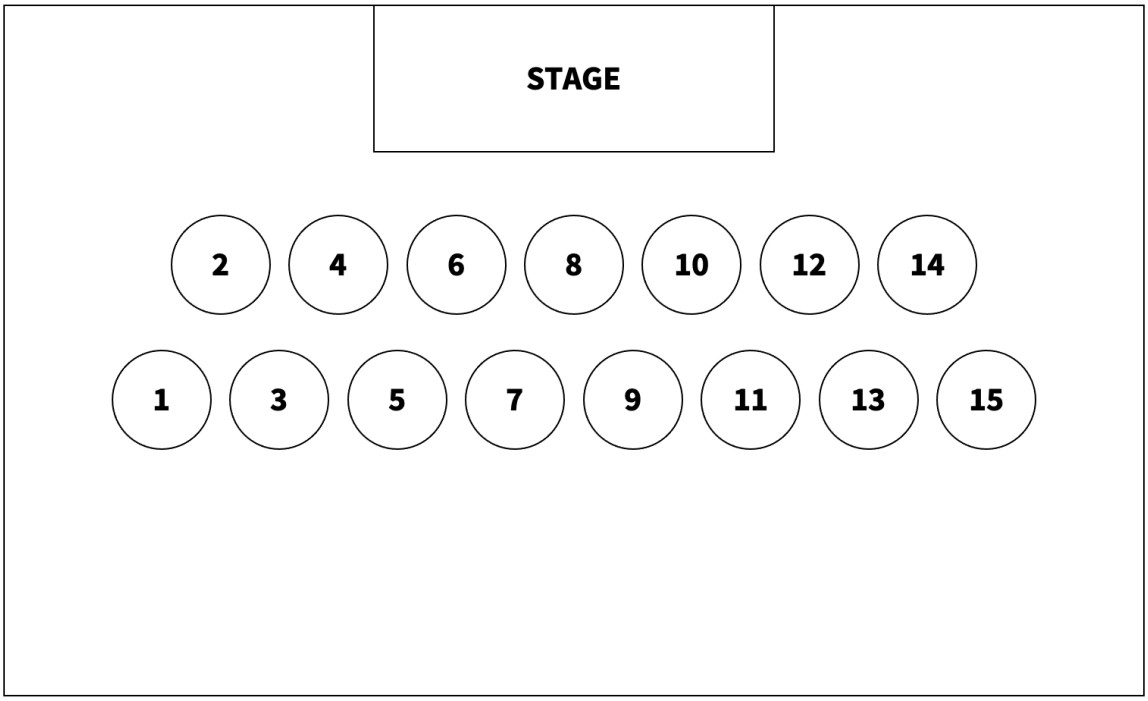
Mainstage sessions will be live streamed and recorded. All other sessions are off-the-record.

12:00-1:00 **Lunch**

Attendees are welcome to sit anywhere in the gardens, lobby or McCaw Hall to eat their lunch. Tables in McCaw Hall will be designated for discussion topics. Feel free to sit at a table to meet people with similar interests.

Table Topics:

1. Misinformation
2. Creating a task force on the future of trust & safety
3. Covert influence operations
4. Child safety
5. Trust & safety on the blockchain
6. Trust & safety and search engines
7. Trust & safety and E2EE
8. Harassment
9. Hate speech
10. Trust & safety and video content
11. Trust & safety and video content
12. Extremism
13. Trust & safety and government regulation
14. Trust & safety and government regulation
15. Privacy



1:00-2:30 **Parallel Sessions**

Select one of the below simultaneously held sessions to attend.

Lightning Talks	**Trust and Safety Across Platforms**	**Academic Research Workshop with the Twitter API v2**
McCaw Hall Mainstage	Fisher Conference Center Lane/Lyons/Lodato Rooms	Fisher Conference Center Barnes/McDowell/Cranston Rooms

Lightning Talks (McCaw Hall Mainstage)

Ten speed talks on content moderation, best practices, and online incitement and grooming.

Moderated by Naomi Shiffman, Oversight Board

- The Safe Assessments: An Inaugural Evaluation of Trust & Safety Best Practices
  David Sullivan, Digital Trust & Safety Partnership
- Platform Trajectories and Content Consumption
  Sujata Mukherjee, Google
- Understanding Accounts That Engage in Hate and Harassment on Reddit
  Deepak Kumar, Stanford University
- Do Your Own Research: How Searching Online to Evaluate Misinformation Can Increase Its Perceived Veracity
  Kevin Aslett, University of Central Florida
- Navigating the Online Grooming Landscape
  Amanda Goharian, Thorn
- Architectures of Choice, or Architectures of Control? Dark Patterns and Algorithmic Manipulation
  Jennifer King, Stanford University
- Multilayered Enforcement - Increasing Recall and Limiting Impact on False Positives
  Vanessa Molter, Google
- Countering Online Gender Based Violence: Fighting Deepfakes, Doxxing, and Disinformation
  Katherine Townsend, World Wide Web Foundation
- Quis Custodiet Ipsos Custodes? Content Takedowns and the Censorship of Academic Research
  Aaron Y. Zelin, Brandeis University
- Encrypted Messaging Apps and the Trust and Safety Field
  Inga Trauthig, University of Texas, Austin

Trust and Safety Across Platforms (Fisher Conference Center Lane/Lyons/Lodato Rooms)
This panel will highlight research teams looking at the impact of trust and safety design or moderation interventions across multiple platforms.
Moderated by Kate Starbird, University of Washington

- Disrupting the Highways of Hate: The Impacts of Offline Events and Deplatforming
  Beth Goldberg, Jigsaw; Yonatan Lupu and Richard Sear, George Washington University
- Using Deep Learning to Detect Abusive Sequences of Member Activity on LinkedIn
  James Verbus and Beibei Wang, LinkedIn
- Look Who's Reporting Now: A Content Analysis of Big Tech's 2020 Transparency Reports
  Amanda Reid, University of North Carolina, Chapel Hill
- Examining the Causal Effect of Twitter's Interventions on Donald Trump's Tweets
  Swapneel Mehta, New York University
- Internet Governance Through Site Shutdowns: The Impact of Shutting Down Two Major Commercial Sex Advertising Sites
  Helen Zeng, Carnegie Mellon University

Academic Research Workshop with the Twitter API v2 (Fisher Conference Center Barnes/McDowell/Cranston Rooms)
In this workshop, researchers will learn how to get historical Twitter data using the Twitter API v2, the new Tweet Downloader tool, and the Tweepy package in Python. Code samples will be provided to participants. For questions about developer applications, please reach out to @suhemparack on Twitter.
Facilitated by Suhem Parack, Twitter

2:30-3:00 **Break**

3:00-4:00 **Parallel Sessions**
Select one of the below simultaneously held sessions to attend.

How are Streaming and Short-Form Video Changing Trust & Safety?	**Research Panel on Moderation & Harm Reduction**	**Discussion Session on Younger Users and Child Safety**
McCaw Hall Mainstage	Fisher Conference Center Lane/Lyons/Lodato Rooms	Fisher Conference Center Barnes/McDowell/ Cranston Rooms

How are Streaming and Short-Form Video Changing Trust & Safety? (McCaw Hall Mainstage)
How should we think about the changing abuse types on video content and how best to mitigate their harms. Moderated by Alex Heath, The Verge.

- Renée DiResta, Stanford Internet Observatory
- Eric Han, TikTok
- Emmanuelle Saliba, Independent investigative journalist

Research Panel on Moderation & Harm Reduction (Fisher Conference Center West)
Moderated by Devra Moehler, Meta

- Prompts and its Impact on Conversation Health
  Calliea Pan, Twitter
- Content Moderator Startle Response: A Qualitative Study
  Timir Bharucha, TaskUs
- Perceptions of Harm and Preferences for Remedies in Social Media Governance
  Sarita Schoenebeck, University of Michigan
- Habituation Study — Reducing the impact of anxiety on content moderators
  Sonal Khaira, Teleperformance

Discussion on Younger Users and Child Safety (Fisher Conference Center Lane/Lyons/ Lodato Rooms)
Facilitated by Riana Pfefferkorn, Stanford Internet Observatory

- Amanda Lenhart, Data & Society Research Institute
- Natalie Shoup, Global Partnership to End Violence Against Children
- Ke (Maddie) Huang-Isherwood, University of Southern California

4:00-5:30 **Happy Hour and Poster Session** (Ford Gardens)
During happy hour, poster presenters will be on hand to discuss their research

- Freedom of Algorithmic Expression
  Inyoung Cheong, University of Washington School of Law
- A Snapshot of the Trust & Safety Field 2022
  Toby Shulruff, Arizona State University
- Computational Approaches to Understanding Credibility in Video-Based Misinformation: An Analysis of Covid-19 Content on TikTok
  Yingdan Lu, Sunny Xun Liu and Jeffery T. Hancock, Stanford University
- Google Says So(S): The Entanglement of Search Engines and Information on Ballot Propositions
  Emma Lurie, Stanford Law School
- Listening to the Global Citizens: A Blockchain Voting Model as a Tool of Global Governance
  Q.J. Yao, Lamar University

- Combating Bias in Algorithmic Decision-Making Systems for Hiring
  Alisar Mustafa, Analytics for Advocacy; Krystal Jackson, Analytics for Advocacy; Michael W. Yang, Independent researcher
- Local Representation: Implications for Press Freedom in Turkey
  Leo Page-Blau, New York University
- Slowcial Media: Promoting slow & mindful consumption through platform design
  Anshuman Dhar, University of Washington; Ed Paradis, Madison Stemmler, & Leo Salemann
- Critical Listening: History, Childhood, and Sonic Safety
  Alexandra Krawetz, independent scholar
- The Fungibility of Non-Fungible Tokens: Vulnerabilities in an Overhyped Market
  Sarah Barrington, University of California Berkeley
- Understanding Dangerous Speech on Social Media in a South Asian Country
  Cathy Buerger, Susan Benesch and Tonei Glavinic, Dangerous Speech Project
- TikTok on the Clock: A Call for Increased Urgency in Researching the Spread of Mis/Disinformation on TikTok
  Kyla Guru and Lila Shroff, Stanford University
- Designing and Building Social Platforms Grounded in Consent
  Jane Im, Nikola Banovic and Florian Schaub, University of Michigan
- Online Hate Speech Literature Between 2018 and 2022: A Mixed Methods Review and Visualization
  Ina Kamenova, University of Massachusetts, Lowell
- Designing Effective Community Guidelines for Social VR
  Rafi Lazerson, University of California Berkeley
- Decoding the Shadowban Mystery: TikTok's Content Governance and How Creators Interpret the Platform's Visibility System
  Diyi Liu, Oxford Internet Institute
- Two Sides of the Same Coin? Comparing Crowdfunding, Cryptocurrency, and Blockchain Use by Extreme Right and Jihadi Groups
  Shahed Warreth, Swansea University

Friday, September 30, 2022

8:00-9:00 **Registration and Continental Breakfast**

9:00-10:00 **What is the Responsibility of Trust & Safety?** (McCaw Hall Mainstage)
Moderated by Elena Cryst, Stanford Internet Observatory
Panelists:
- evelyn douek, Stanford Law School
- Del Harvey, delbius.com
- Mike Masnick, Techdirt
- Brandon Silverman, Former CEO of CrowdTangle

10:00-10:30 **Break**

10:30-12:00 **Parallel Sessions**
Select one of the below simultaneously held sessions to attend.

Lightning Talks	**Research Panel on Trust & Safety and Data Research**	**Discussion on Research Collaborations, Coalitions, and Tools**
McCaw Hall Mainstage	Fisher Conference Center Lane/Lyons/Lodato Rooms	Fisher Conference Center Barnes/McDowell/Cranston Rooms

Lightning Talks on the Mainstage (McCaw Hall Mainstage)
Ten speed talks on gaming, privacy, misinformation, advertisements, and hate speech detection.
Moderated by Brian Fishman, Cinder.
- Does Paying for Access to A Gaming Platform Actually Increase Your Likelihood of Poor Behavior?
  Hill Stark, Spectrum Labs
- Non-Punitive Approaches to Tool Development for Content Moderation
  Joseph Seering, Stanford University
- Privacy and Security Beliefs of Internet Users in Five Developing Countries
  Rebecca Umbach, Google
- Use and Users of Blockchain-based Social Media Platforms
  Anatoliy Gruzd, Toronto Metropolitan University
- Constructive Consensus Building or Coerced Cooperation? The Evolutions of Global Content Moderation Standards
  Courntey Radsch, UCLA
- Ephemeral Fake News: Why Some Misinfo Leaves the Internet and Some Does Not
  Ross Dahlke, Stanford University

- Advances in Data Science for Web Safety and Integrity: Cross-Platform Multi-Modal Misinformation
  Srijan Kumar, Georgia Institute of Technology
- Guilty until Proven Innocent: Autoethnographies of Powerlessness in Instagram and TikTok Account Deletions
  Carolina Are, Northumbria University
- The Role of Ad Transparency in Global Electoral Integrity
  Lisa Reppell, International Foundation for Electoral Systems
- How good are they actually? Evaluating hate detection tools across languages
  Bertie Vidgen, Rewire

Trust & Safety and Data Research (Fisher Conference Center Lane/Lyons/Lodato Rooms)
Moderated by David Thiel, Stanford Internet Observatory

- The Intersection of T&S and Algorithmic Impact
  Amar Ashar, Spotify; Henriette Cramer, Spotify; Olya Gurevich, Spotify
- Sustained Exposure to Fact-Checks can Inoculate Citizens Against Misinformation in the Global South
  Jeremy Bowles, Stanford University
- Emerging safety issues on Web 3.0. and implications for Internet Trust in Africa
  Arthur Gwagwa, Utrecht University
- Digging into the (Internet) Archive: Examining the NSFW Model Responsible for the 2018 Tumblr Purge
  Renata Barreto and Claudia Von Vacano, University of California Berkeley
- How to Make Evidence-Based Trust and Safety Routine and Boring by 2050
  J Nathan Matias, Cornell University

Discussion on Research Collaborations, Coalitions, and Tools (Fisher Conference Center Barnes/McDowell/Cranston Rooms)
Facilitated by Amanda Menking, Trust & Safety Foundation

- J Nathan Matias, Cornell University
- Ryan Williams, University of Texas at Austin
- Beth Goldberg, Jigsaw; Yonatan Lupe and Richard Sear, George Washington University
- Amanda Menking, Trust & Safety Foundation

12:00-1:30 **Lunch**
Attendees are welcome to enjoy their lunch in the gardens or bring their lunch to breakout.

Breakout on Teaching Trust & Safety (Fisher Conference Center Lane/Lyons/Lodato Rooms)
Join instructors in a lively discussion around developing and teaching trust and safety courses to postsecondary students. Do you teach a class that relates to online harm? If so, we encourage you to prepare a minute of remarks about your class.

- Alex Stamos, Stanford Internet Observatory
- Camille Francois, Columbia University School of Public and International Affairs
- Shelby Grossman, Stanford Internet Observatory
- Miles McCain, Stanford Internet Observatory

1:30-3:00 **Journal of Online Trust & Safety Panel II** (McCaw Hall Mainstage)
Presentations of five papers published in the Journal of Online Trust and Safety.
Moderated by Daphne Keller, Stanford Law School

- How can Platform Engagement with Academics and Civil Society Representatives Inform the Development of Content Policies?
  Sarah Shirazyan, Meta
- Online Civility Through Platform Architecture: An Experiment Promoting Civil Conversation Between Neighbors on Nextdoor
  Paul Meosky, Yale
- Election Fraud, YouTube, and Public Perception of the Legitimacy of President Biden
  Megan Brown, New York University
- Creating, Using, Misusing, and Detecting Deep Fakes
  Hany Farid, University of California, Berkeley
- The Demographics of U.S. Facebook Engagement with an Influence Operation in the Alternative Media Ecosystem
  Andrew Beers, University of Washington

3:00-3:30 **Closing Keynote** (McCaw Hall Mainstage)
Ethan Zuckerman, University of Massachusetts at Amherst
Introduction by Jeffrey Hancock, Stanford University

3:30-5:30 **Happy Hour**

Acknowledgements

Thank you for joining us at the first annual Trust and Safety Research Conference. A detailed conference agenda is available on the conference website at http://tsresearchconference.org/schedule.

Recordings of the mainstage sessions will be posted to the Stanford Internet Observatory YouTube Channel and website after the event.

Conference proceedings have been published through the Journal of Online Trust and Safety and are available at http://tsjournal.org.

The Stanford Internet Observatory wishes to acknowledge the following organizations for their support in putting together this event.

Organizing Partner: Trust & Safety Foundation

Livestream Sponsor: YouTube | Google

Event Production: Stanford Law School Program Group; Frances C. Arrillaga Alumni Center; Stanford Event Services; Stanford Video; Melons Catering

Stanford | Internet Observatory
Cyber Policy Center

io.stanford.edu