AGENDA

Legend				
Panel	Research Presentations	Lightning Talks	Workshop	Other

THURSDAY SEPTEMBER 25	McCaw Hall Main Stage	Lane/Lyons/Lodato	Barnes/McDowell/Cranston	Ford Gardens
8:00 - 8:30am	Registration			**Breakfast**
8:30 - 8:45am	**Opening Remarks**			
8:45 - 9:30am	**Keynote Speaker: Dave Wilner**			
9:30 - 9:45am	Break			
9:45 - 11:00am	**Content Moderation and AI Alignment**	**Cybercrime, Fraud, and Sextortion**	**Online Privacy / Digital Well-being ***	
11:00 - 11:30am	Break			
11:30 - 12:30pm	**The Role of Tor in Human Rights Abuses of Children**	**Trust and Safety in Video Games**	**Towards Technical Standards: Safety By Design & Open Source**	
12:30 - 1:45pm	**Speed Networking**			**Lunch**
1:45 - 2:45pm	**Inside the Policy Room: How Big AI Labs Shape the Rules of Generative AI**	**Youth Risks 1**	**AI in Content Moderation**	
2:45 - 3:00pm	Break			
3:00 - 4:00pm	**Content Moderation, Language, & Harm Measurement**	**Youth Risks 2 / Multimodal Trust and Safety ***	**Security, Standards, and Law**	
4:00 - 5:30pm				**Happy Hour & Poster Session**

FRIDAY SEPTEMBER 26	McCaw Hall Main Stage	Lane/Lyons/Lodato	Barnes/McDowell/Cranston	Ford Gardens
8:00 - 8:30am	Registration			**Breakfast**
8:30 - 8:45am	**Opening Remarks**			
8:45 - 9:30am	**Fireside Chat With eSafety Commissioner**			
9:30 - 9:45am	Break			
9:45 - 11:00am	**AI Misinformation and Harmful Speech: Compliance and Mitigation**	**Online Harms**	**AI Tools, Applications, and Validation / AI in Search ***	
11:00 - 11:30am	Break			
11:30 - 12:30pm	**Influence, Elections, and Politics**	**Understanding Online Environments**	**Let's Share: A Framework for Researcher Access to Publicly Available Platform Data**	
12:30 - 1:30pm			**Teaching Consortium Lunch (Invite Only)**	**Lunch**
1:30 - 2:30pm	**Research Across Affiliations: Fostering Collaborations Through Understanding**	**Policy, Public Opinion, and AI Use**	**AI and Algorithm Auditing**	
2:30 - 2:45pm	Break			
2:45 - 3:45pm	**Purpose-Driven AI Companions: Legal, Ethical, and Practical Considerations**	**Trust and Safety in Search**	**Trust and Safety: Conflict Edition Online Game**	
3:45 - 4:00pm	**Closing Remarks**			
4:00 - 5:30pm				**Happy Hour**

***Two sessions happening subsequently in the same room and within the same time block.**

Thursday, September 25, 2025

8:00 – 8:30 **Registration and Breakfast**

8:30 – 8:45 **Opening Remarks**
Location: McCaw Hall Main Stage
***Jeffrey T. Hancock**, Harry and Norman Chandler Professor of Communication, Stanford University*
Faculty Director, Stanford Cyber Policy Center and Social Media Lab

8:45 – 9:30 **Keynote Address**
Location: McCaw Hall Main Stage
***Dave Willner**, Co-Founder of Zentropi*

9:30 – 9:45 **Break**

9:45 – 11:00 *The following sessions will be happening simultaneously:*

Content Moderation and AI Alignment
Location: McCaw Hall Main Stage
Research Presentations featuring twelve-minute presentations with time for questions.
Moderated by Samidh Chakrabarti, Zentropi

- Measurement and Metrics for Content Moderation: The Multi-Dimensional Dynamics of Engagement and Content Removal on Facebook
  Laura Edelson, Northeastern University
- Moderation Tools that Can't Tell Use from Mention Misclassify Counterspeech, but Teaching the Distinction Helps
  Kristina Gligoric, Johns Hopkins University
- A Contextual Approach to Alignment Faking
  Eugene Yu Ji, University of Chicago

Cybercrime, Fraud, and Sextortion
Location: Fisher Conference Center Lane/Lyons/Lodato
Lightning Talks featuring seven-minute presentations with time for questions.
Moderated by Tracy Navichoque, Stanford Human-Centered AI

- Reporting NCII Online: Effectiveness, Barriers, and Victim-Survivor Burdens
  Qiwei Li, University of Michigan
- Financial Sextortion, Fraud & Rituals: Inside the Cybercrime Alliance of Yahoo Boys and Sakawa Boys of Africa
  Emmanuel Adinkra, Ghana Internet Safety Foundation

- Sextortion in India: A Cross-Sectoral Study on Emerging Threats, Platform Accountability, and Support Gaps
  Arnika Singh, Social & Media Matters
- Tech-Driven Crime: Online Job Scams and the Road to Forced Labor
  Yulia Sullivan, Baylor University
- Illuminating Fraud: LLM-Driven Insights and Rapid AI Rule Generation in Digital Payments
  Helen Yuan, Google
- Smarter Verification to Combat Evolving Fraud at Scale
  Steven Chua, Google
- Unveiling AI Content Harms in the Global South: A Participatory Inquiry into Deepfake and Disinformation Challenges in South Asia
  Dilrukshi Gamage, University of Colombo School of Computing

Online Privacy
Location: Fisher Conference Center Barnes/McDowell/Cranston
Lightning Talks featuring seven-minute, rapid fire presentations with time for questions.
Moderated by Riana Pfefferkorn, Stanford Human-Centered AI

- The Private Is Political: Identity and Democracy in the Age of Surveillance Capitalism
  Ray Brescia, Albany Law School
- Exploring the unique privacy & safety challenges of closeted dating app users
  Sanjnah Ananda Kumar, Integrity Institute
- User Privacy and Large Language Models: An Analysis of Frontier Labs' Privacy Policies
  Jennifer King, Stanford University

Digital Well-being
Location: Fisher Conference Center Barnes/McDowell/Cranston
Lightning Talks featuring seven-minute, rapid fire presentations with time for questions.
Moderated by Riana Pfefferkorn, Stanford Human-Centered AI

- It Hurts, and We Stand Alone: A Qualitative Study on Digital Hate Targeting Scholars
  Anja Stevic, Social Media Lab at Stanford University
- Measuring online abuse of land and environmental defenders
  Henry Peck, Global Witness

11:00 – 11:30 **Break**

11:30 – 12:30 *The following sessions will be happening simultaneously:*

Panel: The Role of Tor in Human Rights Abuses of Children
Location: McCaw Hall Main Stage
Moderated by Brian Levine, University of Massachusetts Amherst

- *Joanne Pasquarelli, University of Massachusetts*
- *Lloyd Richardson, Canadian Centre for Child Protection*
- *Clay Shields, Georgetown University*

Trust and Safety in Video Games
Location: Fisher Conference Center Lane/Lyons/Lodato
Research Presentations featuring twelve-minute presentations with time for questions.
Moderated by Samantha Bradsaw, American University

- Detecting Prosocial Communication in Game Text Chats: A Sample-Efficient Method Using Self-Anchored Attention
  R. Michael Alvarez, California Institute of Technology
- Mainstreaming Research on Trust & Safety in Gaming
  Dean Jackson, Tech Policy Press & Samantha Bradshaw, American University
- A Field Experiment in Shaping Online Safety Behaviors through Social Norm Reminders on a Popular Social Platform
  Alex Leavitt & Bridget Daly, Roblox

Workshop: Towards Technical Standards: Safety By Design & Open Source
Location: Fisher Conference Center Barnes/McDowell/Cranston
Facilitated by Ethan Breder, Discord and Vinay Rao, ROOST
Technical standards can provide security, reliability, and simplicity to builders of technology. However, they can also add cumbersome requirements, delay innovation, and even exclude certain kinds of products. Effectively balancing these tradeoffs is critical to the maturity of a technology, such as Trust and Safety systems.

The field of Trust and Safety has developed many best practices, but has yet to formalize these as widely-adopted technical standards. In this workshop, we invite T&S practitioners and researchers to deliberate how we might begin to adopt technical standards, using the rules engine that Discord is open-sourcing in collaboration with ROOST. We will first introduce three abstracted layers of the content moderation stack - detection, review, and enforcement - present in the moderation systems of all social tech. Then, we'll provide specific examples of the configurations Discord uses in its rules engine for each layer. Workshop participants will develop candidates of standardized semantics, and propose the use of them as [must / should / may / should not / must not] attributes of T&S system architecture.

12:30 – 1:45 **Lunch**
Locations: Ford Gardens

12:45 – 1:30 **Speed Networking** happening during lunch
Location: McCaw Hall Main Stage
Led by RT Rogers; Policy Analyst, Cyber Policy Center

1:45 – 2:45 *The following sessions will be happening simultaneously:*

Panel: Inside the Policy Room: How Big AI Labs Shape the Rules of Generative AI
Location: McCaw Hall Main Stage
Moderated by Sarah Shirazyan, Stanford Law School

- *Abby Fanlo Susk, OpenAI*

- *Kevin Klyman, Stanford University*
- *Amre Metwally, Anthropic*

Youth Risks 1
Location: Fisher Conference Center Lane/Lyons/Lodato
Lightning Talks featuring seven-minute, rapid fire presentations with time for questions.
Moderated by Anja Stevic, Social Media Lab at Stanford University

- Leaked - Understanding and Addressing Self-Generated Sexual Content Involving Young People in Thailand
  Jennifer Schatz, Evident
- “If it's harmful, why is it there at all?” How children's voices inform Ofcom's online safety policy
  Michael Allard, Ofcom
- The Four Dimensions of Meaningful Youth Co-Design: Bridging Evidence-to-Practice Gaps for Trust & Safety Teams
  Órla McGovern, Dublin City University, Ireland
- A Trauma-Informed Look into Youth’s Direct Messaging Experiences
  Michal Luria, Center for Democracy & Technology
- Listening to Youth Voices: New Perspectives on Sextortion Risks, Relationships, and Platform Dynamics
  Tim O'Gorman, Thorn
- The REPORT Act: A Landmark Step in U.S. Legislation to Combat Online Child Trafficking
  Avi Jager, ActiveFence

AI in Content Moderation
Location: Fisher Conference Barnes/McDowell/Cranston
Lightning Talks featuring seven-minute, rapid fire presentations with time for questions.
Moderated by Harry Yan, Texas A&M University

- Addressing the Trust Deficit in AI Driven Content Moderation
  Matthew Katsaros, Justice Collaboratory at Yale Law School
- The AI Guardian: Protecting digital communities with context-aware LLM techniques
  Amanpreet Kaur, Google LLC
- Operationalizing Fairness: Practical Bias Mitigation For AI Content Moderation
  Ratnakar Pawar, PlayStation/Sony Interactive Entertainment
- Designing for Thriving: A Psychometric Prototype to Support Wellbeing in Frontline T&S Careers
  Natalie Campbell, TikTok
- Scaling Human Judgment in Community Notes with LLMs
  Haiwen Li, MIT Institute for Data, Systems, and Society

2:45 – 3:00 **Break**

3:00 – 4:00 *The following sessions will be happening simultaneously:*

Content Moderation, Language, & Harm Measurement
Location: McCaw Hall Main Stage
Lightning Talks featuring seven-minute, rapid fire presentations with time for questions.
Moderated by Ravi Iyer, USC Marshall Neely Center

- Governing Babel: The Debate over Social Media Platforms and Free Speech
  John Wihbey, Northeastern University
- Beyond What's in Store for Us: Designing Personalized Content Moderation Systems for Individual Wellbeing
  Rayhan Rashed, University of Michigan
- Low Resources and Low Priority: A Comparative Case Study Analysis of Content Moderation Across Four Languages in the Global Majority
  Dhanaraj Thakur, Center for Democracy & Technology
- Perceived Legitimacy of Layperson and Expert Content Moderators
  Cameron Martel, Johns Hopkins Carey Business School
- Conducting Multilingual Participatory Trust & Safety Research in Wikimedia Communities
  Claudia Lo, Wikimedia Foundation

Youth Risks 2
Location: Fisher Conference Center Lane/Lyons/Lodato
Lightning Talks featuring seven-minute, rapid fire presentations with time for questions.
Moderated by Fangjing Tu, Social Media Lab at Stanford University

- Responding to Online Sexual Abuse and Exploitation of Children (OSAEC) in the Philippines: The role of Technology in Public Awareness Raising and Support for Reporting at Community Level
  Maggie Brennan, Dublin City University
- Perceived Online Risks and Digital-Safety Knowledge in Quetzaltenango: A Multi-Stakeholder Urban–Rural Baseline Study
  Julio Estuardo Santos Velásquez, Iuris Digital
- Connected and Protected: Insights from FOSI's 2025 Online Safety Survey
  Alanna Powers, Family Online Safety Institute (FOSI)

Multimodal Trust and Safety
Location: Fisher Conference Center Lane/Lyons/Lodato
Lightning Talks featuring seven-minute, rapid fire presentations with time for questions.
Moderated by Fangjing Tu, Social Media Lab at Stanford University

- Warning Labels in the Age of Deepfakes: Timing, Frequency, and Unintended Consequences
  Kazutoshi Sasahara, Institute of Science Tokyo
- Designing for facilitation in ephemeral social media spaces
  Nazanin Sabri, University of California San Diego
- Anatomy of a Site Policy Update: Deepfakes, Dual Use, and GitHub's Approach to Platform Governance
  Margaret Tucker, GitHub

Security, Standards, and Law

Location: Fisher Conference Center Barnes/McDowell/Cranston

Lightning Talks featuring seven-minute presentations with time for questions.

Moderated by Daphne Keller, Stanford University

- Toward International Standards for Trust and Safety
  David Sullivan, Digital Trust and Safety Partnership
- The Abusability of Modern Authentication Mechanisms: A Case Study of Passkeys
  Alaa Daffalla, Cornell University
- ExpProof : Operationalizing Explanations for Confidential Models with ZKPs
  Chhavi Yadav, Carnegie Mellon University
- Five Truths about Trust and Data Transfer
  Lisa Dusseault, Data Transfer Initiative
- Actionable Content Under Color of Law
  Susan Benesch, Dangerous Speech Project and Berkman Klein Center for Internet & Society
- Should Social Media Platforms Permit Violating Content that is "Newsworthy"?
  Ricki-Lee Gerbrandt, University College London

4:00 – 5:30 **Happy Hour and Poster Session**

Location: Ford Gardens

A combined happy hour and poster session. Poster presenters will be on hand to discuss their research.

Poster Titles and Presenters:

- Agency of Expression: Perceptions of Human and AI Speakers in Online Environments. *Gowri Swamy, University of California Berkeley*
- Detecting AI Impersonation: Benchmarking the Persuasion Gap Between Genuine Human and LLM-Simulated Responses. *Emily McKinley, University of California Davis*
- Digital Architects: Future-Forward Security for LGBTQ+ Youth. *Marisa Hall, UC Berkeley School of Information*
- Digital well-being features on social media platforms – an LLM-assisted analysis of social media company announcements. *Yuning Liu, Harvard University*
- Evaluating User Engagement with Posts from Popular Alt-right Accounts. *Raghav Jain, SimPPL*
- Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers. *Jared Moore, Stanford University*
- "It Feels Good to Be Let In": The Social Psychology Behind Voluntary AI Safety Commitments. *Divya Ganesan, Stanford Center for International Security and Cooperation (CISAC)*
- Labeling messages as AI-generated does not reduce their persuasive effects. *Isabel Gallegos, Stanford University*
- Mitigating the Diffusion of AI CSAM: A Holistic Approach. *Tracy Y Wei, Stanford University*
- Navigating Reddit's Filters: The Role of Ideology and Rules in Shaping Discourse. *Lucen Li, University of California Davis*
- Not the Average User: A Latent Profile Analysis of Privacy Protection Appraisals in Targeted Advertising. *Laurent Wang, University of California Santa Barbara*
- Social Media News Consumption and Anxiety Among Adolescents Preceding the 2024 U.S. Presidential Election. *Samantha Vigil, University of California Davis*

- Systematic Bias or Congeniality? YouTube RecommendationAlgorithms have systematic right-leaning bias: A Longitudinal Audit. *Miner Ye, University of California Davis*
- The AT Protocol and the Sociotechnical Architecture of Moderation in a Post-Platform Environment. *Emerson Johnston, Stanford University*
- The Evolution of Online Harms Research: A Decadal Analysis of Trends, Methodologies, and Key Contributions (2015–2025). *Angela Ng, ETH Zürich*
- Third Party App Usage As An Indicator Of Coordination. *Manita Pote, Indiana University Bloomington*
- Understanding the Influence of Visual Features and Formats on Credibility Perceptions of Social Media Posts. *Salman Khawar, University of California Davis*
- Weathering Trust: Credibility Formation in YouTube's Weather Information Ecosystem. *Julie Vera, University of Washington*
- What do we mean when we talk about child safety? *Emily Fowler, Oxford Internet Institute*
- What is the purpose of a system? AI and Criminal Innovation. *Gabriel Toscano, Duke University Sanford School of Public Policy*
- Who Speaks, Who Quits, and Who Gets Hurt: Structural Trust and Safety in Social Media Platforms. *Ata Uslu, Network Science Institute at Northeastern University*
- Measuring the Impact of School Phone Policies: Toolkit for Assessing Phones in Schools (TAPS). *Jason Lu, The Anxious Generation; Sarah Wu, Social Media Lab at Stanford University*
- Can Large Language Models Debate Like Community Note Writers on X. *Swapneel Mehta, SimPPL*

Friday, September 26, 2025

8:00 – 8:30 **Registration and Breakfast**

8:30 – 8:45 **Opening Remarks**
Location: McCaw Hall Main Stage
***Ronald Robertson**, Co-Editor of Journal of Online Trust and Safety & Research Scientist at Cyber Policy Center, Stanford University*
***Rosie Ith**, Managing Editor of Journal of Online Trust and Safety & Program Manager at Cyber Policy Center, Stanford University*

8:45 – 9:30 **Fireside Chat**
Location: McCaw Hall Main Stage
***Julie Inman Grant**, Australia's eSafety Commissioner & **Jeffery Hancock**, Harry and Norman Chandler Professor of Communication, Stanford University & Faculty Director, Stanford Cyber Policy Center and Social Media Lab*

9:30 - 9:45 **Break**

9:45 – 11:00 *The following sessions will be happening simultaneously:*

Panel: AI Misinformation and Harmful Speech: Compliance and Mitigation
Location: McCaw Hall Main Stage
Moderated by Florence G'sell, Stanford University
- *Daphne Keller, Stanford University*
- *Alexios Mantzarlis, Cornell University*
- *Alexandra Sanderford, Anthropic*

Online Harms
Location: Fisher Conference Center Lane/Lyons/Lodato
Research Presentations featuring twelve-minute presentations with time for questions.
Moderated by Angela Lee, Social Media Lab at Stanford University
- Social Media and Mental Health: Evidence from 40M Privately-Insured Individuals
  Jacob Shapiro, Princeton University
- Untrustworthy Website Exposure and Election Beliefs: Selective Exposure and Ideological Asymmetry
  Ross Dahlke, University of Wisconsin-Madison
- Regulating Image-Based Abuse: Insights from eSafety's Reporting and Removal Scheme
  Mariesa Nicholas, eSafety Commissioner
- The Evolving Impact of Social Risk Factors on Cybercrime: A Regression Analysis of Victim Reports to the FBI IC3 and FTC Consumer Sentinel
  Eliana Caceres, University of New Hampshire

- Uncovering and Overcoming Offenders' Tactics for Distributing CSAM on File and Image Hosting Services
  Kelly Barker, Canadian Centre for Child Protection

AI Tools, Applications, and Validation
Location: Fisher Conference Center Barnes/McDowell/Cranston
Lightning Talks featuring seven-minute, rapid fire presentations with time for questions.
Moderated by Izzy Benjamin Gainsburg, The Politics and Social Change Lab (PASCL)

- Multilingual Inconsistency as a Trust & Safety Risk in Humanitarian AI Systems: A Human Rights-Based Evaluation Platform to Document and Analyze LLM Inconsistencies Across Languages
  Roya Pazkad, Taraaz and Mozilla Foundation
- From Classifiers to Confidence Intervals: An AI-First Framework for Harm Measurement
  Nick Preston, OpenAI
- Using ChatGPT with Twine to Create Case Stories for Learning and Research
  Laura McLester, University of Alabama in Birmingham
- Governing Model Context Protocol: A Call for Policy Research
  Gabriel Nicholas, Anthropic
- Ethics in Action: A Practical Toolkit for T&S to Build Ethical Tech
  Manuela Travaglianti, McCoy Family Center for Ethics in Society at Stanford University

AI in Search
Location: Fisher Conference Center Barnes/McDowell/Cranston
Lightning Talks featuring seven-minute, rapid fire presentations with time for questions.
Moderated by Izzy Benjamin Gainsburg, Associate Director of the Politics and Social Change Lab (PASCL)

- Teaching AI to Respond Safely: A Quantitatively-Validated Response Framework for Sensitive Search Queries
  Samar Elshafiey, Google
- Public Use of Large Language Models to Obtain Health Information
  Vishala Mishra, Duke University

11:00 – 11:30 **Break**

11:30 – 12:30 *The following sessions will be happening simultaneously:*

Influence, Elections, and Politics
Location: McCaw Hall Main Stage
Lightning Talks featuring seven-minute, rapid fire presentations with time for questions.
Moderated by Nate Persily, Stanford Law School

- Persuading Voters using Human-AI Dialogues
  Hause Lin, Massachusetts Institute of Technology
- Islamic Seminary Students' Engagement with Social Media and Formation of their Religiopolitical Worldviews
  Muhammad Rizwan Safdar, University of the Punjab, Lahore

- Towards Generalizable AI-Assisted Misinformation Inoculation: Protecting Confidence against False Election Narratives with Articles and Chatbots
  Mitchell Linegar, Caltech
- Voter ID Misinformation: Cutting Through the Noise on Voter ID Laws
  Paul Spencer, Disability Rights California
- CandiData24: Collecting Social Media Data for American Political Candidates & Electeds
  Kaitlyn Dowling, National Conference on Citizenship
- The Logic of Diffusion: Structural Drivers of Digital Authoritarianism
  Kassadi Smith, University of Central Florida

Understanding Online Environments
Location: Fisher Conference Center Lane/Lyons/Lodato
Lightning Talks featuring seven-minute, rapid fire presentations with time for questions.
Moderated by Will Schulz, Social Media Lab at Stanford University

- Social Media Algorithms Can Shape Affective Polarization via Exposure to Antidemocratic Attitudes and Partisan Animosity
  Tiziano Piccardi, Stanford University
- Fixing the Feeds: A Policy Road Map to Address Algorithmic Recommender System Harms
  Alissa Cooper, Knight-Georgetown Institute
- Scaling Community-Driven Solutions to Digital Manipulation Campaigns
  Samuel Woolley, University of Pittsburgh
- Declining Information Quality under New Platform Governance
  Burak Ozturan, Network Science Institute, Northeastern University
- Influence Outshines Credibility: How News is Amplified on Truth Social
  Swapneel Mehta, SimPPL
- Complaint Narratives as Signals of Emerging Regulatory Risk on Digital Platforms
  Maria Manuela Palacio, Stanford Law School

Workshop: Let's Share: A Framework for Researcher Access to Publicly Available Platform Data
Location: Fisher Conference Center Barnes/McDowell/Cranston
Facilitated by Leticia Bode, Knight-Georgetown Institute (KGI); Naomi Shiffman, Meta Oversight Board; and Angie Holan, International Fact Checking Network
Publicly available platform data is central for understanding our online information environment. While regulations like the EU's Digital Services Act require digital platforms to make platform data accessible, some platforms have narrowed or limited access to data that was previously available for research – negatively impacting our understanding of the information environment. The workshop has three main purposes:

1. Introduce participants to a new framework for high-influence publicly available platform data, developed by the Knight-Georgetown Institute's Expert Working Group on Publicly Available Platform Data;

2. Discuss how a shared framework for minimum standards of high-influence publicly available platform data could help enable more consistent and equitable access to data across platforms; and
3. Identify strategies to support and operationalize efforts to expand independent research with publicly available platform data in diverse information and policy contexts.

This workshop will feature presentation of the framework and collaborative discussion of strategies to advance independent access to publicly available platform data.

12:30 – 1:30 **Lunch**
Location: Ford Gardens

12:30 – 1:30 **Invite-Only Teaching Consortium Working Lunch (Sign-Up Here)**
Location: Barnes/McDowell/Cranston

1:30 – 2:30 *The following sessions will be happening simultaneously:*

Panel: Research Across Affiliations: Fostering Collaborations Through Understanding
Location: McCaw Hall Main Stage
Moderated by Amanda Menking, Trust and Safety Foundation

- *Julia Kamin, Prosocial Design Network*
- *Jolguer Perez, TaskUs*
- *Sonja Schmer-Galunder, University of Florida*
- *Andrew Smart, Google Research*

Policy, Public Opinion, and AI Use
Location: Fisher Conference Center Lane/Lyons/Lodato
Research Presentations featuring twelve-minute presentations with time for questions.
Moderated by David Evan Harris, Haas School of Business at UC Berkeley

- Unpacking Public and Expert Opinion on AI
  Colleen McClain, Pew Research Center
- Case Studies in Research Informed US State Technology Policy
  Ravi Iyer, Marshall Neely Center, University of Southern California
- How Terrorist Groups are Adapting and Exploiting AI
  Robert Demgenski, ActiveFence
- Problematic Media Use as a Content-Agnostic, Enforceable Harm
  Jenny Radesky, University of Michigan Medical School

AI and Algorithm Auditing
Location: Fisher Conference Center Barnes/McDowell/Cranston
Lightning Talks featuring seven-minute, rapid fire presentations with time for questions.
Moderated by Matthew DeVerna, Cyber Policy Center at Stanford University

- Democratizing AI Auditing: BiasTestGPT for User-Driven Detection of Social Bias in Language Models
  Rafal Kocielnik, California Institute of Technology

- Is Intimacy the New Attention? An Audit of Expressed Intimacy Across LLM Generations
  Pearl Vishen, University of California Davis
- The Exclusion of Disabled Workers by Digitized Hiring Assessments
  Ariana Aboulafia, Center for Democracy & Technology
- Algorithmic Amplification of Out-group Hate on YouTube
  Claire Wonjeong Jo, University of California Davis
- Safety by Design? Researching User Behaviour Online
  Jonathan Porter, Ofcom
- Catching Dark Signals in Algorithms: Unveiling Audiovisual and Thematic Markers of Unsafe Content Recommended for Children and Teenagers
  Haoning Xue, University of Utah

2:30 – 2:45 **Break**

2:45 – 3:45 *The following sessions will be happening simultaneously:*

Panel: Purpose-Driven AI Companions: Legal, Ethical, and Practical Considerations
Location: McCaw Hall Main Stage
Moderated by Jeff Hancock, Stanford University

- *Nathanael Fast, University of Southern California*
- *Meetali Jain, Tech Justice Law Project*
- *Ryn Linthicum, Anthropic*

Trust and Safety in Search
Location: Fisher Conference Center Lane/Lyons/Lodato
Research Presentations featuring twelve-minute presentations with time for questions.
Moderated by Ronald Robertson, Cyber Policy Center at Stanford University

- Do Age-Verification Bills Change Search Behavior? A Pre-Registered Synthetic Control Multiverse
  David Lang, Stanford University
- Redesigning Google's CSAM Prevention and Deterrence Intervention on Search
  Rebecca Umbach, Google
- From Blue Links to Black Boxes: Understanding User Behaviors and Regulatory Implications of GenAI search
  Doris Li, Ofcom

Workshop: Trust and Safety: Conflict Edition Online Game
Location: Fisher Conference Center Barnes/McDowell/Cranston
Facilitated by Mark Silverman, International Committee of the Red Cross (ICRC) & Mike Masnick, Copia Institute .
The International Committee of the Red Cross (ICRC) has partnered with the Copia Institute to develop a conflict-focused edition of Copia's Trust and Safety Tycoon game. In this simulation, players lead the Conflict and Crisis Team at a fictional social media company (situated within a broader Trust and Safety Organization). Through realistic, conflict-based scenarios, players must make decisions that can

influence local communities, aid organisations, and the company's reputation and performance—either for better or worse.

This workshop is designed with three key aims:

1. Group gameplay: Participants will play the game together to understand its dynamics and ground the workshop in the real-world implications of digital technology for conflict-affected populations.
2. Explore applicability: Discuss how the game can be effectively used to foster dialogue and learning among technology firms, policymakers, and researchers to better support people impacted by conflict.
3. Identify research gaps: Highlight short and long term research, policy, or technical areas where more work is needed to understand and address the opportunities, risks, and harms digital technologies pose to those affected by conflict—and to act upon the findings.

3:45 – 4:00 **Closing Remarks**

Location: McCaw Hall Main Stage

***Jeffrey T. Hancock**, Harry and Norman Chandler Professor of Communication, Stanford University*
Faculty Director, Stanford Cyber Policy Center and Social Media Lab

4:00 – 5:30 **Happy Hour** in Ford Gardens

Recording & Photography Policy

This is a fully in-person conference. The event will **not** be recorded or live-streamed.

Recording of any kind — including audio, video, or live-streaming — is **not** permitted in any session or space at the event. This policy is in place to foster open, honest, and trust-based dialogue among participants.

Photography **is** permitted, but we ask that photos only be taken and shared with the consent of those being photographed. Please note we will have a professional photographer at the event. If you would like access to photos post-event, please contact tsrconference@stanford.edu.

Social media sharing (quotes, impressions, general photos) is welcome with consent, unless a session or speaker explicitly requests otherwise. Please be mindful of context when posting.

Resources

Thank you for joining us at the fourth annual Trust & Safety Research Conference.

Conference proceedings is now published in the *Journal of Online Trust and Safety,* available at http://tsjournal.org.

Event Sponsors

The Stanford Cyber Policy Center thanks the Omidyar Network and Knight Foundation for their generous support of the Center and this conference.