Foundations of Online Trust & Safety

Prof. Mark Schneider
Mas2215@columbia.edu

What rules and expectations should online platforms such as Google, Facebook/Meta, X, OpenAI Instagram, TikTok, Uber use to govern themselves? How do technology companies work to mitigate offline and online psychological, social, and safety harms from arising from their products? How do geopolitical questions and conflicts manifest on online platforms—for instance, how should social media platforms handle gruesome images and unverified information emerging from current wars across the globe? And how do platforms and scholars understand both what the risks are and how to measure and address these risks?

In this class, over 14 sessions, we will examine content moderation harms (election misinformation, violence and incitement, hate, violence incitement, harassment) and how these problems manifest online and impact offline behavior. We will then critically examine the ways that companies have addressed these problems (red-teaming, IPOCs at Meta, human and AI-based content moderation). Third, we will explore the role of context in addressing content moderation harms, drawing on prominent cases including the Ukraine and Gaza Wars, Love Jihad in India, ethnic cleansing in Myanmar, and elections in India, the US, and France where social media played a large role.

Grading and assignments

Students will be evaluated on one term paper, small ad-hoc assignments and class participation. Class participation will aim to demonstrate that students have understood required materials and are able to critically engage with course topics. Students will be asked to submit discussion points on readings for each class.

Class participation -- 15%. Students should attend class and actively participate in discussion with questions to raise about the readings we discuss. Students should participate at least once in each class meeting and complete any online activities to get a strong participation grade.

Responses to Readings – 10%. Students will be asked to respond to the weeks readings critically while raising questions for class discussion that the readings bring up for you. The response should be 250 words and engage the readings, how they connect to each other, and identify one question that this raises for you. Focusing on one reading in depth will be acceptable. I will provide broad questions to think about as you read in advance. You may may also benefit from using the reading worksheet (developed by Kanchan Chandra at NYU) for any empirical readings to help you think through theories and evidence. Please post your responses on Courseworks by 9am the day we meet. You are required to complete responses for 8 weeks of your choosing.

Trust and Safety Risks Memo – 15%. Student will complete a trust and safety memo that evaluates the introduction of a new feature or product change for trust and safety risks and proposes safeguards to address those problems. The memo should both lay out the benefits of the feature in a product sense and evaluate its risks to give you a perspective on both sides of

the debate. It should also consider if there are different risks in different countries or regions (e.g., Global South vs. US) or among different types of users (e.g., genPOP vs. general users; men vs. women; majority group vs. minority groups). The memo should be 6-8 pages double spaced.

Red Teaming Exercise – 5%. The goal of this exercise is to give you some hands-on experience with current safety controls in large language models (LLMs), how they work and, sometimes, how blunt and inefficient they can be. The output of this exercise will fit on roughly 2-3 pages and will include:

- Which model you decided to test (ex. Anthropic's *Claude*, OpenAI's GPT4 through ChatGPT, etc.)
- Which "rule" or policy you managed to break or attempted to bypass (please don't get yourself in trouble and exercise your best judgment in prompt testing!),
- How you went about testing it, and which prompts were most / least successful at bypassing safety controls,
- Screenshots of the resulting test,
- One or two sentences about your personal takeaway from this experience.

This should not be a particularly hard exercise, notably because there are currently huge portions of the Internet doing the same thing, see www.jailbreakchat.com or r/chatGPT on Reddit. Don't spend too much time on it, and remember to have fun: this is about innovative thinking, and espousing a hacker's mindset (something we often have to do in T&S!). And if you're looking for an easy way to test different models at once, the Poe app is free and will allow you to toggle between a handful of models easily.

Present an article as if it were your own – 15%. Students will be asked to present an article as if it were your own. This means that you will choose a required or recommended article assigned for the course and present its arguments, assumptions, methods, evidence, and conclusions, and defend choices made by the author of the article as you would if this was presented at a conference. There will be a discussant who will also provide comments on the paper also like you would see at a conference. The presentation should be about 10 minutes in length. The discussant will comment for about 3 minutes.

Paper Proposal – 5%. You will write a two-page paper proposal with a research question, hypothesis, and proposed method of analysis. This will be due April 1. You will receive feedback from classmates and the professor on your memo.

Course Paper – 35%. The course paper will focus on a dimension or key question in Trust & Safety. Expected paper length will be ~ 15-20 pages (double-spaced). Paper should be submitted one week after the last class session, and paper topics should be discussed together in office hours or over email before then. Students will be required to submit and will be graded on accordingly a) a proposal, b) an outline, and c) the final paper, with the final paper's grade carrying the highest weight.

Office hours can be scheduled over email (mas2215@columbia), and can be held over Zoom or in-person throughout the duration of the class.

Section I | Social Media Harms: What We're Up Against

Session 1: Introduction + The Challenge of Content Moderation

Learning Goals: Introduce the broad challenge and debates facing content moderation and introduce the class. Start the overview of content harms at a high level.

- "Scaling Trust on the Web", report of the Atlantic Council's Taskforce for a Trustworthy Future Web, June 2023 (~150p), available here. Read the executive summary and familiarize yourself with the key findings.
- Maxim, K., Parecki, J., & Cornett, C. (2022). How to Build a Trust and Safety Team In a Year: A Practical Guide From Lessons Learned (So Far) At Zoom. *Journal of Online Trust and Safety*, *1*(4). https://doi.org/10.54501/jots.v1i4.81

Session 2: Principles and Competing Incentives for Content Moderation

Learning Goals: Introduce the broad incentives and principles underlying content moderation. Illustrate these principles with examples.

- Douek, Evelyn, Content Moderation as Administration (January 10, 2022). forthcoming Harvard Law Review Vol. 136, Available at SSRN: https://ssrn.com/abstract=4005326 or http://dx.doi.org/10.2139/ssrn.4005326
- Robyn Caplan, Content or Context Moderation? Artisanal, Community-Reliant, and Industrial Approaches (Data and Society 2018)
- Liu, Yi, Pinar Yildirim & Z. John Zhang. 2022. "Implications of Revenue Models and Technology for Content Moderation Strategies." *Marketing Science* 41 (4): 831–847.

Session 3: Trust and Safety Foundations

Learning Goals: Learn the ways that trust and safety teams operate and how this has evolved over time in large companies.

- Data & Society podcast on the origins of the Trust & Safety field, July 8th 2020, Alexander MacGillivray & Nicole Wong in conversation with Robyn Caplan: https://datasociety.net/library/origins-of-trust-and-safety/ *Note*: Transcript can be downloaded if you're rather read through.
- Gillett, Rosalie, Zahra Stardust, and Jean Burgess. "Safety for whom? Investigating how platforms frame and perform safety and harm interventions." *Social Media+ Society* 8.4 (2022): 20563051221144315l; available here.
- Maschmeyer, L., Deibert, R. J., & Lindsay, J. R. (2020). A tale of two cybers - how threat reporting by cybersecurity firms systematically underrepresents threats to civil society. *Journal of Information Technology & Politics*, *18*(1), 1–20. https://doi.org/10.1080/19331681.2020.1776658

- On *middleware:* Daphne Keller, "The Future of Platform Power: Making Middleware Work". Journal of Democracy, vol. 32, no. 3, July 2021, pp. 168-72., online here

Session 4: Taking Cultural and Socio-Political Context into Account

Learning Goals: Understand content moderation challenges across different contexts with a focus on the global south vs. US comparison. We'll also have an activity evaluating content in different country contexts.

Guest Panel of Civic Integrity Alumni: Diane Change (Cohere, ex-Meta Civic Integrity); Kyle Gibson (Meta AI and Global Elections); Theodora Skeadas (NDI, ex-Twitter).

- Benesch, Susan. 2012. *Dangerous Speech: A Proposal to Prevent Group Violence.* Washington, DC: World Policy Institute.
- Stecklow, Steve. 2018. "Why Facebook Is Losing the War on Hate Speech in Myanmar." *Reuters Investigates,* August 15, 2018.

Session 5: Misinformation and Disinformation

Learning Goals: Understand misinformation and disinformation and some ways to combat it.

- Berinsky, Adam J. 2017. "Rumors and Health Care Reform: Experiments in Political Misinformation." BJPS.
- Blair, Robert A., Jessica Gottlieb, Brendan Nyhan, Laura Paler, Pablo Argote, and Charlene J. Stainfield. 2023. "Interventions to counter misinformation: lessons from the global north and applications to the global south." Prepared for USAID.
- Amar, Priyadarshi, Sumitra Badrinathan, Simon Chauchard, and Florian Sichart. 2025. "Countering misinformation early: Evidence from a classroom-based field experiment in India." *American Political Science Review*: 1-21.
- Recommended: Bennett, W. L., & Livingston, S. 2018. "The disinformation order: Disruptive communication and the decline of democratic institutions." *European Journal of Communication*, 33 (2), 122–139.

Session 6: Elections Risks and Social Media

Learning Goals: Examine case studies of social media during elections.

- Sheikh, Shahana. 2024. "How Technology Is (and Isn't) Transforming Election Campaigns in India." *Carnegie Endowment for International Peace*, March 7, 2024
- McGregor, Shannon C., and Daniel Kreiss. 2024. "Influencers, Algorithms, and the New Campaign Playbook." *Political Communication* 41 (3): 257–281.
- Liu, Yuxin, M. Amin Rahimian, and Kiran Garimella. 2025. "Structural Dynamics of Harmful Content Dissemination on WhatsApp." *arXiv* preprint arXiv:2505.18099. https://arxiv.org/abs/2505.18099
- Coppock, Alexander, Kimberly Gross, Ethan Porter, Emily Thorson, and Thomas J. Wood. "Conceptual replication of four key findings about factual corrections and

misinformation during the 2020 US election: evidence from panel-survey experiments." *British Journal of Political Science* 53, no. 4 (2023): 1328-1341.

- Recommended: Biswas, Ahana; Javadian Sabet, Alireza; Lin, Yu-Ru. 2025. "Toxic politics and TikTok engagement in the 2024 U.S. election." *Harvard Kennedy School Misinformation Review* 6 (4).

Session 7: Child Safety

Learning Goals: Examine the risks of child safety on social media and existing enforcements against this content.

- Stanford Internet Observatory (2024). "The Strengths and Weaknesses of the Online Child Safety Ecosystem." *Read selectively: ecosystem overview, CyberTipline workflow, and detection challenges.*
- Keum, Brian TaeHyuk, Yu-Wei Wang, Julia Callaway, Israel Abebe, Tiana Cruz, and Seini O'Connor. "Benefits and harms of social media use: A latent profile analysis of emerging adults." *Current Psychology* 42, no. 27 (2023): 23506-23518.
- Thorn & All Tech Is Human (2023). "Safety by Design for Generative AI: Preventing Child Sexual Abuse."
- Case Study (choose one):
    - Kashmir Hill, "A Dad Took Photos of His Naked Toddler for the Doctor. Google Flagged Him as a Criminal." (NYT, 2022)
      False positives, context, consequences.
    - Roblox investigation ("pedophile hellscape" report)
      Unique product surface + child safety challenges in gaming environments.

Class 8: Violent Extremist and Violence Incitement

Learning Goals: Understand how violent extremist and terrorist groups use social media and the impacts of efforts to combat this.

- Paresh Dave, "Inside Two Years of Turmoil at Big Tech's Anti-Terrorism Group", Sept. 30 2024, Wired Magazine, online here
- Tamar Mitts, *Safe Havens for Hate: the challenge of moderating online extremism*, Princeton University Press (2025) – Introduction (pp. 15-33)
- Thomas, Daniel Robert, and Laila A Wahedi. "Disrupting hate: The effect of deplatforming hate organizations on their online audience." Proceedings of the National Academy of Sciences of the United States of America vol. 120,24 (2023): e2214080120. doi:10.1073/pnas.2214080120
- Schissler, Matt. "Beyond hate speech and misinformation: Facebook and the Rohingya genocide in Myanmar." *Journal of Genocide Research* 27, no. 3 (2025): 445-470.

Session 9: Content Moderation and AI

Learning Goals: Evaluate changes to content moderation in the current era when AI has replaced much of human content moderation.

- Nafia Chowdhury, March 19, 2022, Automated Content Moderation: A Primer, Stanford Program on Platform Regulation, available here (~12p).
- Best Practices for AI Automation in Trust and Safety, DTSP, September 2024. Available here
- Thomas H. Costello, Gordon Pennycook, and David G. Rand. "Durably reducing conspiracy beliefs through dialogues with AI." Science 385(6714).
- Barrett, Paul M., and Justin Hendrix. 2024. "Is Generative AI the Answer for the Failures of Content Moderation?" *TechPolicy.Press*, April 3, 2024.

Session 10: Content Moderation Decisions and Escalations

Learning Goals: Understand the structure of escalations decisions and their limitations at scale.

- Katsaros, M., Kim, J., & Tyler, T. (2023). Online Content Moderation: Does Justice Need a Human Face? International Journal of Human–Computer Interaction, 40(1), 66–77. https://doi.org/10.1080/10447318.2023.2210879
- Center for Democracy & Technology. "The Digital Services Act: An Overview." 2022.
- **Case Studies:** Select cases from the Facebook Oversight Board:
    - Case 2021-014-FB-UA, "regarding a post discussing the situation in Ethiopia" (online)
    - Case 2021-001-FB-FBR, President Trump's Access to Facebook & Instagram (online)
    - Case 2023-023-FB-UA, Altered Video of President Biden (online)
    - 2024 - Explicit AI Images of Female Public Figures - multiple case decision (online)
- In October 2024 the EU established the Appeals Centre Europe, an out-of-court dispute settlement body set up under the EU's Digital Services Act and backed by Meta. News coverage on this announcement here and here; Appeals Center Europe's website

Session 11: Class Workshop on Paper Proposals

Submit your memo on Courseworks Monday at 11:59 PM.

- Read your classmates memos on Course Works.

Session 12: Influence Operations

Learning Goals: Examine influence operations and their role in spreading disinformation online.

- Camille Francois and Herb Lin, The strategic surprise of Russian information operations on social media in 2016: Mapping a blind spot (online)
- Kate Starbird, Ahmer Arif, and Tom Wilson. "Disinformation as Collaborative Work: Surfacing the Participatory Nature of Strategic Information Operation," University of Washington, 2019 (online)

- Thomas Rid, "The Lies Russia Tells Itself", Sept. 30th, Foreign Affairs (here); and the related indictment of "Doppelganger" by the U.S. Department of Justice, Sept. 4th 2024, full affidavit here (277 p., but worth a skim for the many fascinating operational details.)

Class 13: Red Teaming
Guest Speaker: TBA.

Learning Goals: Learn the method of red teaming and conduct a class activity on red teaming around a new product.

- Frontier Model Forum, *What is Red Teaming?* Available online here.
- Data & Society Policy Brief: AI Red-teaming is not a one-stop solution to AI harms, Sorelle Friedler, Ranjit Singh, Borhane Blili-Hamelin, Jacob Metcalf, Brian J. Chen. Available here (~10p).
- Weidinger, Laura, John Mellor, Bernat Guillen Pegueroles, Nahema Marchal, Ravin Kumar, Kristian Lum, Canfer Akbulut et al. "STAR: SocioTechnical Approach to Red Teaming Language Models." (2024), pre-print.
- Enigma Talk, "Neither band-aids nor silver bullets: how bug bounties can help the discovery, disclosure and redress of algorithmic harms", Sasha Constanza-Chock and Camille François (online, full paper). (*skim*)

Class 14: What has AI done to T&S, and closing session on the future of the field

Learning Goals: Discuss the current state of content moderation and connect course themes in a broader discussion.

- Alyssa Boicel , "Using LLMs to Moderate Content: Are They Ready for Commercial Use?", Tech Policy Press, Apr 3, 2024, online here.
- Evans, Benedict. 2021. "Is Content Moderation a Dead End? *BenedictEvans.com*, April 13, 2021.

Additional resources: resiliency
The course may touch upon difficult material: if you anticipate strong distress as a result of encouraging one of the topics listed in the syllabus (harassment, hate speech, sexual abuse of minors and adults, violent extremism, etc.), please talk to your instructor and we will arrange an alternative path. As a reminder, Columbia's Student Health resources are available to you.
Additional resources: newsletters, podcasts, websites of interest

- Excellent Trust & Safety podcasts include Stanford's weekly *Moderated Content* with Evelyn Douek: https://law.stanford.edu/directory/evelyn-douek/moderated-content/
- For newsletters: Casey Newton's Platformer often covers T&S issues with a lot of nuance and understanding; Ben Whitelaw's *Everything in Moderation* is a great read on all things content moderation.
- The Trust & Safety Professional Association, the Digital Trust & Safety Partnership and the Integrity Institute are professional associations focused on T&S professionals, and dedicated to creating standards in the field, and connectivity amongst its members.
    - Of note, TSPA has a public job board: https://www.tspa.org/explore/job-board/

- Finally, the Facebook Oversight Board cases are often detailed, thoughtful reads on specific moderation decisions. I recommend familiarizing yourself with one or two of them at least. Case library: https://www.oversightboard.com/decision/

Generative AI Policy

Students are encouraged to use generative AI tools (such as Open AI's *ChatGPT, Anthropic's Claude, Mistral's LeChat, Google's NotebookLM, Poe,* etc.) responsibly (to prepare and improve upon work, never to author it) and transparently (disclosing both systems and prompts used when Generative AI is utilized in a short section at the end of the assignment). For additional resources on this topic, please see Columbia CTL's resource page on AI tools in the classroom, consider subscribing to newsletters covering the latest advancements and limitations of these tools (ex. *Ben's Bite*, or *The Neuron*), and engaging with expert commentary specialized on how to responsibly use generative AI tools in academic contexts (ex. Mushtaq Bilal).