Trust & Safety Reading List

The Trust & Safety Teaching Consortium created this reading list in May 2023. The Consortium is a coalition of academic, industry and non-profit experts in online trust and safety problems. Our goal is to create content that can be used to teach a variety of audiences about trust and safety issues in a wide variety of formats. If you are interested in helping to create teaching content, email trustandsafetyjournal@stanford.edu to join the Consortium.

Module A: Introduction to Trust & Safety

Contributors: Camille François (Columbia University / Niantic Labs); Mariana Olaizola Rosenblat (NYU Stern Center for Business and Human Rights)

Core Content

- Robertson, Ronald E. 2022. "Uncommon Yet Consequential Online Harms." *Journal of Online Trust and Safety* 1 (3). https://doi.org/10.54501/jots.v1i3.87.
- Trust & Safety Professional Association. 2021. "Abuse Types." June 17, 2021. https://www.tspa.org/curriculum/ts-fundamentals/policy/abuse-types/.
- Podcast: Macgillivray, Alex, and Nicole Wong. July, 2020. "Origins of Trust and Safety with Robyn Caplan." https://datasociety.net/library/origins-of-trust-and-safety/.
- Different practitioners' testimonies about building trust and safety teams:
    - Feerst, Alex. 2019. "Your Speech, Their Rules: Meet the People Who Guard the Internet." OneZero. March 2, 2019. https://onezero.medium.com/your-speech-their-rules-meet-the-people-who-guard-the-internet-ab58fe6b9231.
    - Maxim, Karen, Josh Parecki, and Chanel Cornett. 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.

    - Samuelson, Maison. 2022. “How Pinterest Built Its Trust & Safety Team.” Pinterest Engineering Blog. April 7, 2022. https://medium.com/pinterest-engineering/how-pinterest-built-its-trust-safety-team-8d6c026dd4b9.
- Digital Trust and Safety Partnership. 2021. “Trust and Safety Best Practices Framework.” https://dtspartnership.org/wp-content/uploads/2021/04/DTSP_Best_Practices.pdf.

Optional Content

- Video: Alex Stamos. 2018. “2018 CISAC Drell Lecture: The Battle for the Soul of the Internet (Hoover Institute).” https://www.youtube.com/watch?v=NKN6xLhTjIo.
- Digital Trust & Safety Partnership. 2022. “The Safe Assessments: An Inaugural Evaluation of Trust and Safety Best Practices.”
- *For a regulatory perspective*: eSafety Commissioner. 2019. Safety by Design Overview.
    - *Context note for the reading*: Regulators are increasingly interested in shaping platform design for trust & safety issues (see also the “Design Codes” regulations), not just the outcome of trust & safety systems. The Australian e-Commissioner's office is a good example of this.

Module B: Government Regulations

Contributor: Karen Maxim

Core Content

- Section 230 of the CDA: https://www.law.cornell.edu/uscode/text/47/230.
- Jeong, Sarah. “Politicians Want to Change The Internet’s Most Important Law. They Should Read it First.” July 26, 2021. https://www.nytimes.com/2019/07/26/opinion/section-230-political-neutrality.html.
- Keller, Daphne, and Max Levy. “Getting Transparency Right.” Lawfare, July 11, 2022. https://www.lawfareblog.com/getting-transparency-right.

- Trust and Safety Changelog, from SightEngine. https://sightengine.com/trust-and-safety-changelog.
- The GDPR Enforcement Tracker [*Example exercise: Find the 10 largest fines*]. https://www.enforcementtracker.com/.
- Kosseff, Jeff. 2019. "Chapter 7: American Exceptionalism." In The Twenty-Six Words That Created the Internet. Ithaca: Cornell University Press.

Optional Content

- Congressional Research Service. 2019. Data Protection Law: An Overview. https://crsreports.congress.gov/product/pdf/R/R45631.
- Congressional Research Service. 2022. The EU-U.S. Data Privacy Framework: Background, Implementation, and Next Steps. https://crsreports.congress.gov/product/pdf/LSB/LSB10846.
- Congressional Research Service. 2020. EU Data Protection Rules and U.S. Implications. https://crsreports.congress.gov/product/pdf/IF/IF10896.
- Loeb & Loeb LLP. 2023. "Mapping Privacy Requirements." https://www.loeb.com:443/en/insights/publications/2023/03/roadmap-for-2023-privacy-laws-in-effect.

Module C: Metrics and Measurement

Contributors: Inbal Goldberger (ActiveFence); Alex Leavitt (Roblox; UC Berkeley)

Core Content

- Integrity Institute. 2022. Metrics & Transparency Data and Datasets to Track Harms, Design, and Process on Social Media Platforms. https://static1.squarespace.com/static/614cbb3258c5c87026497577/t/617834d31bcf2c5ac4c07494/1635267795944/Metrics+and+Transparency+-+Summary+%28EXTERNAL%29.pdf.
- Freeman, David. 2020. "The Abuse Uncertainty Principle, and Other Lessons Learned from Measuring Abuse on the Internet." Presented at Enigma 2020. https://www.usenix.org/conference/enigma2020/presentation/freeman.

- Metz, Rachel. 2021. “The Facebook Papers: Likes, Anger Emojis and RSVPs: The Math behind Facebook’s News Feed — and How It Backfired.” CNN Business. https://www.cnn.com/2021/10/27/tech/facebook-papers-meaningful-social-interaction-news-feed-math/index.html.
    - *And recommended you read the related Facebook Files coverage of the MSI metric:*
        - Oremus, Will, Chris Alcantara, Jeremy B. Merrill, and Artur Galocha. 2021. “How Facebook Shapes Your Feed.” Washington Post. https://www.washingtonpost.com/technology/interactive/2021/how-facebook-algorithm-works/.
        - Blumfeld, Andrew. Facebook's Meaningful Social Interaction Metric. Telepath. https://telepath.io/blog/facebooks-meaningful-social-interaction-metric.
        - Litt, Eden, Siyan Zhao, Robert Kraut, and Moira Burke. 2020. "What are Meaningful Social Interactions in Today’s Media Landscape? A Cross-Cultural survey." Social Media + Society 6 (3). https://doi.org/10.1177/2056305120942888.
- Chakravorti, Bhaskar, Ajay Bhalla, and Ravi Shankar Chaturvedi. 2018. “The 4 Dimensions of Digital Trust, Charted Across 42 Countries.” Harvard Business Review. https://hbr.org/2018/02/the-4-dimensions-of-digital-trust-charted-across-42-countries.
- Singh, Spandana, and Leila Doty. 2021. “The Transparency Report Tracking Tool: How Internet Platforms Are Reporting on the Enforcement of Their Content Rules.” Open Technology Institute. http://newamerica.org/oti/reports/transparency-report-tracking-tool/.

Optional Content

- Sujata Mukherjee, Rottem Sagi. August 2023. “All About Trust: Measurement Challenges and Possibilities.” Marketplace Risk Webinar. https://www.marketplacerisk.com/webinar-series/all-about-trust%3A-measurement-challenges-and-possibilities.
- Baym, Nancy K. 2013. “Data Not Seen: The Uses and Shortcomings of Social Media Metrics.” First Monday. https://doi.org/10.5210/fm.v18i10.4873.

- Christian, Brian. 2020. The Alignment Problem: Machine Learning and Human Values. W. W. Norton.
- Harling, Anna-Sophie, Declan Henesy, and Eleanor Simmance. 2023. “Transparency Reporting: The UK Regulatory Perspective.” Journal of Online Trust and Safety 1 (5). https://doi.org/10.54501/jots.v1i5.108.
- Leerssen, Paddy. 2020. “The Soap Box as a Black Box: Regulating Transparency in Social Media Recommender Systems.” SSRN Scholarly Paper. https://doi.org/10.2139/ssrn.3544009.
- Peters, Kay, Yubo Chen, Andreas M. Kaplan, Björn Ognibeni, and Koen Pauwels. 2013. “Social Media Metrics — A Framework and Guidelines for Managing Social Media.” Journal of Interactive Marketing 27 (4): 281–98. https://doi.org/10.1016/j.intmar.2013.09.007.
- Ruan, Yefeng, Ping Zhang, Lina Alfantoukh, and Arjan Durresi. 2017. “Measurement Theory-Based Trust Management Framework for Online Social Communities.” ACM Transactions on Internet Technology 17 (2): 16:1-16:24. https://doi.org/10.1145/3015771.

Module D: Trust & Safety Ecosystem: Multistakeholderism

Contributors: Devika Malik (Ex-Meta, South Asia); Amanda Menking (Trust & Safety Foundation)

Core Content

- Van der Spuy, Anri. 2017. “What If We All Governed the Internet?: Advancing Multistakeholder Participation in Internet Governance.” Paris: United Nations Educational, Scientific, and Cultural Organization. https://unesdoc.unesco.org/ark:/48223/pf0000259717.
    - Internet Governance Forum: https://www.intgovforum.org/en
    - UN consultation on Digital Global Compact https://www.un.org/techenvoy/global-digital-compact
- Heldt, Amélie, and Stephan Dreyer. 2021. “Competent Third Parties and Content Moderation on Platforms: Potentials of Independent Decision-Making Bodies From A Governance Structure Perspective.” Journal of Information Policy 11 (December): 266–300. https://doi.org/10.5325/jinfopoli.11.2021.0266.

- Savage, John, and Bruce McConnell. 2015. “Exploring Multi-Stakeholder Internet Governance.” East West Institute. https://www.eastwest.ngo/sites/default/files/Exploring%20Multi-Stakeholder%20Internet%20Governance_0.pdf .
- Rachel Griffin. 2022. “Public and Private Power in Social Media Governance: Multistakeholderism, the Rule of Law and Democratic Accountability.” SSRN Electronic Journal. https://sciencespo.hal.science/hal-03940697/document.

Optional Content

- “Trust & Safety and Law Enforcement” (Trust & Safety Professional Association Trust & Safety Curriculum chapter) https://www.tspa.org/curriculum/ts-fundamentals/trust-safety-and-law-enforcement/.
- *A range of different stakeholders’ sites so that students can do desk research and create stakeholder maps as an exercise:*
    - INHOPE’s site: https://www.inhope.org/EN
    - The Global Internet Forum to Counter Terrorism site: https://gifct.org/
    - Oversight Board’s site: https://www.oversightboard.com/
    - Data & Society’s site: https://datasociety.net/
    - World Economic Forum Global Coalition for Digital Safety’s site: https://initiatives.weforum.org/global-coalition-for-digital-safety/home
    - UW Center for an Informed Public’s site: https://www.cip.uw.edu/
    - Cornell’s Citizens and Technology Lab’s site: https://citizensandtech.org/
    - Trust & Safety Professional Association’s site: https://www.tspa.org/
    - Integrity Institute’s site: https://integrityinstitute.org/
- Case studies:
    - TikTok and Oracle: Public Accountability via Private Auditor. Trust and Safety Foundation. https://trustandsafetyfoundation.org/blog/tiktok-and-oracle-public-accountability-via-private-auditor/
    - Public Interest, Local Laws, and Privacy Rules in India. Trust and Safety Foundation. https://trustandsafetyfoundation.org/blog/public-interest-local-laws-and-privacy-rules-in-india/
    - COVID-19 Around the Globe: The Removal of a Facebook Post About an Argentine Treatment. Trust and Safety Foundation. https://trustandsafetyfoundation.org/blog/covid-19-around-the-globe-the-removal-of-a-facebook-post-about-an-argentine-treatment/

- Riley, Chris, and David Morar. 2021. “Applying Multistakeholder Internet Governance to Online Content Management.” R Street. https://www.rstreet.org/research/applying-multistakeholder-internet-governance-to-online-content-management/.
- Danielle. 2018. “Regulating Social Media: A Multistakeholder ‘Content Congress’ - Diplo.” Diplo. October 1, 2018. https://www.diplomacy.edu/blog/regulating-social-media-multistakeholder-content-congress/.
- Panday, Jyoti, Milton Mueller, and Farzaneh Badiei. 2022. “Multistakeholderism & Platform Content Governance.” Internet Governance Project. https://www.internetgovernance.org/wp-content/uploads/MS-Content.docx-1.pdf.

Module E: Content Moderation

Contributor: Joseph Seering (KAIST)

Core Content

- Roberts, Sarah. Behind the Screen: Content Moderation in the Shadows of Social Media. Chapter 2: Understanding Commercial Content Moderation.
- Film: Chen, Adrian, and Ciaran Cassidy, dirs. 2017. The Moderators. Field Of Vision. https://fieldofvision.org/shorts/the-moderators. [*This is a powerful film, but it shows some only semi-blurred nude and violent images. Please watch before assigning, particularly to undergraduates.*]
- Discord Moderator Academy.“Module 441: Community Governance Structures.” https://discord.com/moderation/1500000179202-441-community-governance-structures.
- Seering, Kaufman, Geof Kaufman, and Chancellor, Stevie. “Metaphors in Moderation.” https://journals.sagepub.com/doi/abs/10.1177/1461444820964968?journalCode=nmsa.
- Pan, Christina A., Sahil Yakhmi, Tara P. Iyer, Evan Strasnick, Amy X. Zhang, and Michael S. Bernstein. 2022. “Comparing the Perceived Legitimacy of Content Moderation Processes: Contractors, Algorithms, Expert Panels, and Digital Juries.” Proceedings of the ACM on Human-Computer Interaction 6 (CSCW1): 1–31. https://doi.org/10.1145/3512929.

Optional Content

- Chandrasekharan, Eshwar, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. 2019. “Crossmod: A Cross-Community Learning-Based System to Assist Reddit Moderators.” Proceedings of the ACM on Human-Computer Interaction 3 (CSCW): 174:1-174:30. https://doi.org/10.1145/3359276.
- Chen, Zhaodi, and Dali L. Yang. 2023. “Governing Generation Z in China: Bilibili, Bidirectional Mediation, and Online Community Governance.” The Information Society 39 (1): 1–16. https://doi.org/10.1080/01972243.2022.2137866.
- Grimmelmann, James. 2015. “The Virtues of Moderation.” Cornell Law Faculty Publications, April. https://scholarship.law.cornell.edu/facpub/1486.
- Waters, Michael. 2020. “How a 1980s AIDS Support Group Changed The Internet Forever.” OneZero (blog). December 18, 2020. https://onezero.medium.com/the-long-forgotten-story-of-ben-gardiner-the-aids-activist-whose-network-transformed-the-internet-c14460a73165.
- Gengle, Dean. “Fairwitnessing: The Case for a New Social Role.” https://drive.google.com/file/d/17MhmV8YJQiSxMa1TL7LTmOv9tOf1IbJs/view?usp=sharing (excerpt from http://www.bbsdocumentary.com/software/APPLE/II/COMMUNITREE/#:~:text=communitree.manual.tar).

Module F: Information Environment

Contributors: Kevin Aslett (University of Central Florida); Brian Murphy (Georgetown University)

Core Content:

- Guess, Andy M and Benjamin Lyons. 2020. Misinformation, Disinformation, and Online Propaganda. Social Media and Democracy: The State of the Field, Prospects for Reform. 10. https://doi.org/10.1017/9781108890960.
- Kallas, Kristina. 2016. “Claiming the Diaspora: Russia’s Compatriot Policy and Its Reception by Estonian-Russian Population.” Journal on Ethnopolitics and Minority Issues in Europe 15 (3). 1–25.

https://www.ecmi.de/fileadmin/downloads/publications/JEMIE/2016/Kallas.pdf
.

- King, Gary, Jennifer Pan, and Margaret E. Roberts. 2017. “How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, Not Engaged Argument.” American Political Science Review 111 (3): 484–501. https://doi.org/10.1017/S0003055417000144.
- Roozenbeek, Jon, Jane Suiter, and Eileen Culloty. 2022. “Countering Misinformation: Evidence, Knowledge Gaps, and Implications of Current Interventions.” PsyArXiv. https://doi.org/10.31234/osf.io/b52um.
- Graphika and Stanford Internet Observatory. 2022. Unheard Voice: Evaluating Five Years of Pro-Western Covert Influence Operations. Stanford Digital Repository. Available at https://purl.stanford.edu/nj914nx9540.
- Rossbach, Niklas. 2017. “Psychological Defense: Vital for Sweden’s Defense Capability.” https://www.foi.se/rest-api/report/FOI%20MEMO%206207.

Optional Content

- Starbird, Kate. 2019. “Disinformation’s Spread: Bots, Trolls and All of Us.” Nature 571 (7766): 449. https://doi.org/10.1038/d41586-019-02235-x.
- Van Bavel, Jay J., and Andrea Pereira. 2018. “The Partisan Brain: An Identity-Based Model of Political Belief.” Trends in Cognitive Sciences 22 (3): 213–24. https://doi.org/10.1016/j.tics.2018.01.004.

What classifies as disinformation vs. misinformation vs. propaganda:

- Murphy, Brian J. 2022. “The Impact of Social Media Conveyed Russian-Backed Disinformation in a Polarized America: An Examination of the Executive Branch’s Ethical Responsibility to Respond.” Ph.D., Georgetown University. https://www.proquest.com/docview/2760166084/abstract/9EE61E3EE4184D8BPQ/1,
- Lazer, David M. J., Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, et al. 2018. “The Science of Fake News.” Science 359 (6380): 1094–96. https://doi.org/10.1126/science.aao2998.
- Wardle, Claire, and Hossein Derakhshan. “Thinking about ‘Information Disorder’: Formats of Misinformation, Disinformation, and Mal-Information.” In

Journalism, Fake News & Disinformation: Handbook for Journalism Education and Training, 43–54. Paris: UNESCO, 2018. https://en.unesco.org/node/296051.

Strategies from social media and governmental organizations for identifying, tracking misinformation:

- Content Commons. "A Counter-Disinformation System That Works." February 20, 2020. https://commons.america.gov/article?id=44&site=content.america.gov.
- Robbins, Joseph. "Countering Russian Disinformation." September 23, 2020. https://www.csis.org/blogs/post-soviet-post/countering-russian-disinformation.

Demand-side: Cognitive aspect: Why does information/rumors/conspiracy theories resonate with individuals

- Bryanov, Kirill, and Victoria Vziatysheva. 2021. "Determinants of Individuals' Belief in Fake News: A Scoping Review Determinants of Belief in Fake News." PLOS ONE 16 (6): e0253717. https://doi.org/10.1371/journal.pone.0253717.
- Guess, Andrew, Jonathan Nagler, and Joshua Tucker. 2019. "Less than You Think: Prevalence and Predictors of Fake News Dissemination on Facebook." Science Advances 5 (1): eaau4586. https://doi.org/10.1126/sciadv.aau4586.
- Nyhan, Brendan. 2010. "Why the 'Death Panel' Myth Wouldn't Die: Misinformation in the Health Care Reform Debate:" The Forum 8 (1). https://doi.org/10.2202/1540-8884.1354.
- Carey, James. *Communication as Culture: Essays on Media and Society*. Boston: Unwin Hyman, 1989.

Interventions designed to mitigate misinformation

- Acerbi, Alberto, Sacha Altay, and Hugo Mercier. 2022. "Research Note: Fighting Misinformation or Fighting for Information?" Harvard Kennedy School Misinformation Review, January. https://doi.org/10.37016/mr-2020-87.
- Guess, Andrew M., Michael Lerner, Benjamin Lyons, Jacob M. Montgomery, Brendan Nyhan, Jason Reifler, and Neelanjan Sircar. 2020. "A Digital Media Literacy Intervention Increases Discernment between Mainstream and False News in the United States and India." Proceedings of the National Academy of Sciences 117 (27): 15536–45. https://doi.org/10.1073/pnas.1920498117.

- Linden, Sander van der. 2022. “Misinformation: Susceptibility, Spread, and Interventions to Immunize the Public.” Nature Medicine 28 (3): 460–67. https://doi.org/10.1038/s41591-022-01713-6.
- Voelkel, Jan, Michael Stagnaro, and James Chu. “Megastudy Identifying Successful Interventions to Strengthen Americans’ Democratic Attitudes: Institute for Policy Research - Northwestern University.” Northwestern University, August 2022. https://www.ipr.northwestern.edu/our-work/working-papers/2022/wp-22-38.html.

Supply-side: Influence Operations / Monetary incentives of misinformation

- DiResta, Renée, Shelby Grossman, and Alexandra Siegel. 2022. “In-House Vs. Outsourced Trolls: How Digital Mercenaries Shape State Influence Strategies.” Political Communication 39 (2): 222–53. https://doi.org/10.1080/10584609.2021.1994065.
- Munger, Kevin. 2020. “All the News That’s Fit to Click: The Economics of Clickbait Media.” Political Communication 37 (3): 376–97. https://doi.org/10.1080/10584609.2019.1687626.
- Stukal, Denis, Sergey Sanovich, Richard Bonneau, and Joshua A. Tucker. 2022. “Why Botter: How Pro-Government Bots Fight Opposition in Russia.” American Political Science Review 116 (3): 843–57. https://doi.org/10.1017/S0003055421001507.
- Wilson, Tom, and Kate Starbird. 2020. “Cross-Platform Disinformation Campaigns: Lessons Learned and Next Steps.” Harvard Kennedy School Misinformation Review 1 (1). https://doi.org/10.37016/mr-2020-002.
- Podcast: Hendrix, Justin. 2023. “Examining the Impact of Internet Research Agency Tweets in the 2016 U.S. Election.” The Sunday Show - Tech Policy Press. https://techpolicy.press/examining-the-impact-of-internet-research-agency-tweets-in-the-2016-u-s-election/.
- Video: “Former KGB Agent, Yuri Bezmenov, Warns America About Socialist Subversion,” 1984. https://www.youtube.com/watch?v=Z1EA2ohrt5Q.

Module G: Terrorism, Radicalization, and Extremism

Contributors: Mariana Olaizola Rosenblat (NYU Stern Center for Business and Human Rights); Inga Trauthig (The University of Texas at Austin)

Core Content

- *Online extremism and potential countermeasures:*
  Winter, Charlie, Peter Neumann, Alexander Meleagrou-Hitchens, Magnus Ranstorp, Lorenzo Vidino, and Johanna Fürst. 2020. "Online Extremism: Research Trends in Internet Activism, Radicalization, and Counter-Strategies." *International Journal of Conflict and Violence (IJCV)* 14: 1–20. https://doi.org/10.4119/ijcv-3809.
- *Overview of suggested solutions in academia:*
  Correa, Denzil, and Ashish Sureka. 2013. "Solutions to Detect and Analyze Online Radicalization : A Survey." arXiv. https://doi.org/10.48550/arXiv.1301.4916.
- *Countering Terrorism Online with Artificial Intelligence:*
  United Nations Office of Counter -Terrorism. 2021 . "Countering Terrorism Online with Artificial Intelligence: An Overview for Law Enforcement and Counter-Terrorism Agencies in South Asia and South-East Asia." United Nations Interregional Crime and Justice Research Institute. https://www.un.org/counterterrorism/sites/www.un.org.counterterrorism/files/countering-terrorism-online-with-ai-uncct-unicri-report-web.pdf.
- *Social media's role in the Jan 6th insurrection:*
  Podcast: Jurecic, Quinta, and Jen Patja Howell. 2023. "The Lawfare Podcast: A Jan. 6 Committee Staffer on Social Media and the Insurrection." https://www.lawfareblog.com/lawfare-podcast-jan-6-committee-staffer-social-media-and-insurrection.
- *Gaming sites and extremists:*
  "Gaming the System: How Extremists Exploit Gaming Sites and What Can be Done to Counter Them," NYU Stern Center for Business and Human Rights (May 2023). https://bhr.stern.nyu.edu/tech-gaming-report.

Optional Content

Online extremism outside of main social media platforms/emerging challenges:

- Bodo, Lorand and Inga Kristina Trauthig. 2022. "Emergent Technologies and Extremists: The DWeb as a New Internet Reality?" Global Network on Extremism and Technology. https://gnet-research.org/wp-content/uploads/2022/07/GNET-Report-Emergent-Technologies-Extremists-Web.pdf.
- Elson, Joel S, Doctor, Austin C and Sam Hunter. 2022. "The Metaverse Offers a Future Full of Potential - for Terrorists and Extremists, too." The Conversation. https://theconversation.com/the-metaverse-offers-a-future-full-of-potential-for-terrorists-and-extremists-too-173622.

Additional resources for further readings:

- Center for Countering Digital Hate: https://counterhate.com/
- GIFCT website: https://gifct.org
- Global Network on Extremism and Technology: https://gnet-research.org
- ISD Global toolkits: https://www.isdglobal.org/pub-types/toolkits/?fwp_language=english
- Tech Against Terrorism: https://www.techagainstterrorism.org
- VOX-Pol Network of Excellence (NoE): https://www.voxpol.eu/

Module H: Harassment and Hate Speech

Contributors: Ina Kamenova (University of Massachusetts Lowell); Q. J. Yao (Lamar University, Texas)

Core Material

Harassment:

- Foley, Timothy, and Melda Gurakar. 2022. "Backlash or Bullying? Online Harassment, Social Sanction, and the Challenge of COVID-19 Misinformation." Journal of Online Trust and Safety 1 (2). https://doi.org/10.54501/jots.v1i2.31.
- Harris, Bridget, and Delanie Woodlock. 2022. "Spaceless Violence: Women's Experiences of Technology-Facilitated Domestic Violence in Regional, Rural and Remote Areas." Australian Institute of Criminology. https://doi.org/10.52922/ti78405.

- Slaughter, Autumn, and Elana Newman. 2022. “New Frontiers: Moving Beyond Cyberbullying to Define Online Harassment.” Journal of Online Trust and Safety 1 (2). https://doi.org/10.54501/jots.v1i2.5.
- Atske, Sara. 2021. “The State of Online Harassment.” Pew Research Center: Internet, Science & Tech (blog). January 13, 2021. https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/.

Hate Speech:

- Castaño-Pulgarín, Sergio Andrés, Natalia Suárez-Betancur, Luz Magnolia Tilano Vega, and Harvey Mauricio Herrera López. 2021. “Internet, Social Media and Online Hate Speech. Systematic Review.” Aggression and Violent Behavior 58 (May): 101608. https://doi.org/10.1016/j.avb.2021.101608.
- Kim, Jae Yeon, and Aniket Kesari. 2021. “Misinformation and Hate Speech: The Case of Anti-Asian Hate Speech During the COVID-19 Pandemic.” Journal of Online Trust and Safety 1 (1). https://doi.org/10.54501/jots.v1i1.13.
- Kovács, György, Pedro Alonso, and Rajkumar Saini. 2021. “Challenges of Hate Speech Detection in Social Media: Data Scarcity, and Leveraging External Resources.” SN Computer Science 2 (2): 95. https://doi.org/10.1007/s42979-021-00457-3.
- Matamoros-Fernández, Ariadna, and Johan Farkas. 2021. “Racism, Hate Speech, and Social Media: A Systematic Review and Critique.” Television & New Media 22 (2): 205–24. https://doi.org/10.1177/1527476420982230.
- Tontodimamma, Alice, Eugenia Nissi, Annalina Sarra, and Lara Fontanella. 2021. “Thirty Years of Research into Hate Speech: Topics of Interest and Their Evolution.” Scientometrics 126 (1): 157–79. https://doi.org/10.1007/s11192-020-03737-6.

Optional Content

- Alkiviadou, Natalie. 2019. “Hate Speech on Social Media Networks: Towards a Regulatory Framework?” Information & Communications Technology Law 28 (1): 19–35. https://doi.org/10.1080/13600834.2018.1494417.

- Guiora, Amos, and Elizabeth A. Park. 2017. “Hate Speech on Social Media.” Philosophia 45 (3): 957–71. https://doi.org/10.1007/s11406-017-9858-4.
- Ullmann, Stefanie, and Marcus Tomalin. 2020. “Quarantining Online Hate Speech: Technical and Ethical Perspectives.” Ethics and Information Technology 22 (1): 69–80. https://doi.org/10.1007/s10676-019-09516-z.

Module I: Child and Adult Sexual Exploitation

Contributors: Caroline Humer (International Centre for Missing & Exploited Children); Leslie Taylor (Genpact)

Core Content

- Film: Auroris Media, dir. 2022. Sextortion: The Hidden Pandemic. https://sextortionfilm.com
    - Burnett, Elena, Brian Jarboe, Stuart Rushfel, Sami Yenigun, Brett Neely, and William Troop. 2023. “‘Sextortion’ Documentary May Leave Viewers With Exaggerated Sense Of Risk To Children.” NPR. https://www.npr.org/2023/03/14/1163326306/sextortion-documentary-may-leave-viewers-with-exaggerated-sense-of-risk-to-child.
- Gallay, Amelia. 2021. “Sex Sells, But Not Online: Tracing the Consequences of FOSTA-SESTA.” Berkeley Journal of Criminal Law (blog). Accessed April 6, 2023. https://www.bjcl.org/blog/sex-sells-but-not-online-tracing-the-consequences-of-fosta-sesta.
- Keller, Michael H., and Gabriel J. X. Dance. 2019. “The Internet Is Overrun With Images of Child Sexual Abuse. What Went Wrong?” The New York Times, September 28, 2019, sec. U.S. https://www.nytimes.com/interactive/2019/09/28/us/child-sex-abuse.html.
- Rosenberg, Roni, and Hadar Dancig-Rosenberg. 2021. “Reconceptualizing Revenge Porn.” Arizona Law Review 63: 199. https://heinonline.org/HOL/LandingPage?handle=hein.journals/arz63&div=7&id=&page=.
- Blunt, Danielle, and Ariel Wolf. 2020. “Erased: The Impact of FOSTA-SESTA and the Removal of Backpage on Sex Workers.” Anti-Trafficking Review, no. 14 (April): 117–21. https://doi.org/10.14197/atr.201220148.

Optional Content

- dir. 2018. Ella Cobbs: Human Trafficking in America. https://www.ted.com/talks/ella_cobbs_human_trafficking_in_america.
- Goldstein, Jessica. 2020. "'Revenge Porn' Was Already Commonplace. The Pandemic Has Made Things Even Worse." The Washington Post. https://www.washingtonpost.com/lifestyle/style/revenge-porn-nonconsensual-porn/2020/10/28/603b88f4-dbf1-11ea-b205-ff838e15a9a6_story.html.
- Rock, Amy. 2022. "Human Trafficking on College Campuses: What It Looks Like and Resources for Police, Students - Campus Safety." Campus Safety Magazine. https://www.campussafetymagazine.com/podcast/human-trafficking-college-campuses/.
- Video: Amanda Todd Story: https://www.youtube.com/watch?v=vOHXGNx-E7E.
- Human Trafficking Fact Sheet DHS 2022 - https://www.dhs.gov/sites/default/files/publications/ht_101_one-pager_.pdf.
- Harris, Sam, and Gabriel Dance.2020. "The Worst Epidemic." Accessed April 6, 2023. https://wakingup.libsyn.com/213-the-worst-epidemic.

Module J: Suicide, Self-Harm, & Well-Being

Contributors: Katherine Keyes (Columbia University); Alex Leavitt (Roblox; UC Berkeley); Elena Cryst (Stanford Internet Observatory)

Core Material

- Gonzalez, Robbie. 2019. "Screens Might Be as Bad for Mental Health as ... Potatoes." Wired. https://www.wired.com/story/screens-might-be-as-bad-for-mental-health-as-potatoes/.
- Nesi, Jacqueline, Taylor A. Burke, Alexandra H. Bettis, Anastacia Y. Kudinova, Elizabeth C. Thompson, Heather A. MacPherson, Kara A. Fox, et al. 2021. "Social Media Use and Self-Injurious Thoughts and Behaviors: A Systematic Review and Meta-Analysis." Clinical Psychology Review 87 (July): 102038. https://doi.org/10.1016/j.cpr.2021.102038.

- Orben, Amy, and Andrew K. Przybylski. 2019. “The Association between Adolescent Well-Being and Digital Technology Use.” Nature Human Behavior 3 (2): 173–82. https://doi.org/10.1038/s41562-018-0506-1.
- Robinson, Jo, Georgina Cox, Eleanor Bailey, Sarah Hetrick, Maria Rodrigues, Steve Fisher, and Helen Herrman. 2016. “Social Media and Suicide Prevention: A Systematic Review: Suicide Prevention and Social Media.” Early Intervention in Psychiatry 10 (2): 103–21. https://doi.org/10.1111/eip.12229.
- Facebook Safety Center: Suicide Prevention. https://www.facebook.com/safety/wellbeing/suicideprevention

Optional Material

- Boyd, Danah, Jenny Ryan, and Alex Leavitt. 2011. “Pro-Self-Harm and the Visibility of Youth-Generated Problematic Content.” I/S: A Journal of Law and Policy for the Information Society 7: 1. https://kb.osu.edu/bitstream/handle/1811/72981/ISJLP_V7N1_001.pdf?sequence=1.
- Feiner, Jessica Bursztynsky, Lauren. 2021. “Facebook Documents Show How Toxic Instagram Is for Teens, Wall Street Journal Reports.” CNBC. September 14, 2021. https://www.cnbc.com/2021/09/14/facebook-documents-show-how-toxic-instagram-is-for-teens-wsj.html.
- Meier, Adrian, and Leonard Reinecke. 2021. “Computer-Mediated Communication, Social Media, and Mental Health: A Conceptual and Empirical Meta-Review.” Communication Research 48 (8): 1182–1209. https://psyarxiv.com/573ph/.
- Moon, Khatiya C., Anna R. Van Meter, Michael A. Kirschenbaum, Asra Ali, John M. Kane, and Michael L. Birnbaum. 2021. “Internet Search Activity of Young People With Mood Disorders Who Are Hospitalized for Suicidal Thoughts and Behaviors: Qualitative Study of Google Search Activity.” JMIR Mental Health 8 (10): e28262. https://doi.org/10.2196/28262.

Module K: Authenticity, Identity, and Platform Manipulation

Contributors: Lee Foster (Alperovitch Institute for Cybersecurity Studies at Johns Hopkins School of Advanced International Studies); Tabea Wilke (Twincler)

Core Material

- FireEye. 2018. “Operation Leveraging Inauthentic News Sites and Social Media Aimed at U.S., U.K., Other Audiences.” Mandiant. https://www.mandiant.com/sites/default/files/2021-09/rpt-FireEye-Iranian-IO%20%281%29-1.pdf.
- Video: Katya Sedova. 2022. “AI-Powered Disinformation, Present and Future.” Towards Data Science. https://www.youtube.com/watch?v=AAe_ZfW0pP8.
- Fredheim, Rolf, and Sebastian Bay. 2020. “Social Media Manipulation 2021/2022: Assessing the Ability of Social Media Companies to Combat Platform Manipulation.” NATO Strategic Communications Centre of Excellence. https://stratcomcoe.org/publications/social-media-manipulation-20212022-assessing-the-ability-of-social-media-companies-to-combat-platform-manipulation/242.
- Graphika Team. 2020. “Step into My Parler: Suspected Russian Operation Targeted Far-Right American Users on Platforms Including Gab and Parler, Resembled Recent IRA-Linked Operation That Targeted Progressives.” Graphika. https://public-assets.graphika.com/reports/graphika_report_step_into_my_parler.pdf.
- Stubbs, Jack. 2020. “Exclusive: Russian Operation Masqueraded as Right-Wing News Site to Target U.S. Voters - Sources.” Reuters, October 1, 2020, sec. 2020 Candidate Slideshows. https://www.reuters.com/article/usa-election-russia-disinformation-idUSKBN26M5OP.

Optional Content

- Threat Analysis Group, Mandiant, and Google Trust & Safety. 2023. “Fog of War: How the Ukraine Conflict Transformed the Cyber Threat Landscape.” Google. https://services.google.com/fh/files/blogs/google_fog_of_war_research_report.pdf.
- Saurwein, Florian, and Charlotte Spencer-Smith. 2021. “Automated Trouble: The Role of Algorithmic Selection in Harms on Social Media Platforms.” Media and Communication 9 (4): 222–33. https://www.cogitatiopress.com/mediaandcommunication/article/view/4062/4062#.
- Bradshaw, Samantha, Hannah Bailey and Howard, Philip. 2020. “Industrialized Disinformation: 2020 Global Inventory of Organized Social Media Manipulation.” Oxford Internet Institute & Computational Propaganda Research Project. https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/12/2021/02/CyberTroop-Report20-Draft9.pdf.
- Revelli, Alice, and Lee Foster. 2022. “‘Distinguished Impersonator’ Information Operation That Previously Impersonated U.S. Politicians and Journalists on Social Media Leverages Fabricated U.S. Liberal Personas to Promote Iranian Interests.” Mandiant (blog). 2022. https://www.mandiant.com/resources/blog/information-operations-fabricated-personas-to-promote-iranian-interests.
- Subramanian, Samanth. 2017. “Meet the Macedonian Teens Who Mastered Fake News and Corrupted the US Election.” Wired. https://www.wired.com/2017/02/veles-macedonia-fake-news/.
- Renee DiResta, Dr. Kris Shaffer, Becky Ruppel, David Sullivan, Robert Matney, Ryan Fox, Dr. Jonathan Albright, Ben Johnson: “Tactics & Tropes of the Internet Research Agency”, Report on the Internet Research Agency commissioned by the U.S. Senate Intelligence Committee. https://digitalcommons.unl.edu/senatedocs/2/
- Mandiant. 2022. “Pro-PRC DRAGONBRIDGE Influence Campaign Leverages New TTPs to Aggressively Target U.S. Interests, Including Midterm Elections.” https://www.mandiant.com/resources/blog/prc-dragonbridge-influence-elections.
- Graphika and Stanford Internet Observatory. 2022. “Bad Reputation: Suspected Russian Actors Leverage Alternative Tech Platforms in Continued Effort to

Covertly Influence Right-Wing U.S. Audiences." https://fsi9-prod.s3.us-west-1.amazonaws.com/s3fs-public/2022-12/graphika-stanford-report-bad-reputation_2.pdf.

- Wilke, Tabea. 2020. "Information Threats - Challenges for the European Information Space." https://www.twincler.com/whitepaperinformationthreats/.
- Gleicher, Nathaniel. 2019. "Removing Coordinated Inauthentic Behavior From Georgia, Vietnam and the US." Meta. https://about.fb.com/news/2019/12/removing-coordinated-inauthentic-behavior-from-georgia-vietnam-and-the-us/.
- Graphika and DFRLab. 2019. "#OperationFFS: Fake Face Swarm - Facebook Takes Down Network Tied to Epoch Media Group That Used Mass AI-Generated Profiles." https://graphika.com/reports/operationffs-fake-face-swarm.
- Emerging Technology from the Arxiv. 2014. "The Hidden World of Facebook 'Like Farms.'" MIT Technology Review. https://www.technologyreview.com/2014/09/19/171293/the-hidden-world-of-facebook-like-farms/.

Module L: Types of Attack Surfaces

Contributor: Alex Leavitt (Roblox; UC Berkeley)

Core Content

- Electronic Frontier Foundation: Surveillance Self-Defense https://ssd.eff.org/
  - *Read through:*
    - *All of the Basics pages*
    - *The first 2 Tool Guides pages*
    - *The first 3 Further Learning pages*
    - *Any other pages that seem interesting*
- Podcast: Wall Street Journal. 2021. The Facebook Files, Part 4: The Outrage Algorithm. Wall Street Journal. https://www.wsj.com/podcasts/the-journal/the-facebook-files-part-4-the-outrage-algorithm/e619fbb7-43b0-485b-877f-18a98ffa773f.
- Bradshaw, Samantha, Hannah Bailey, and Philip Howard. 2020. "Industrialized Disinformation: 2020 Global Inventory of Organized Social Media Manipulation." Oxford Internet Institute & Computational Propaganda Research Project.

https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/12/2021/02/CyberTroop-Report20-Draft9.pdf.
- Davis, Antigone. 2021. “Our Approach to Safer Private Messaging.” Messenger News (blog). https://messengernews.fb.com/2021/12/01/our-approach-to-safer-private-messaging/.
- Ribeiro, Filipe N., Koustuv Saha, Mahmoudreza Babaei, Lucas Henrique, Johnnatan Messias, Fabricio Benevenuto, Oana Goga, Krishna P. Gummadi, and Elissa M. Redmiles. 2019. “On Microtargeting Socially Divisive Ads: A Case Study of Russia-Linked Ad Campaigns on Facebook.” In Proceedings of the Conference on Fairness, Accountability, and Transparency 140–49. FAT* ’19. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3287560.3287580.

Optional Content

- McDonald, Allison, Catherine Barwulor, Michelle L. Mazurek, Florian Schaub, & Elissa M. Redmiles. 2021. “‘It's Stressful Having All These Phones: Investigating Sex Workers' Safety Goals, Risks, and Practices Online.” In 30th USENIX Security Symposium. 375-392. USENIX. https://www.usenix.org/system/files/sec21-mcdonald.pdf.
- DiResta, Renee. 2018. “Free speech is not the same as free reach.” Wired. https://www.wired.com/story/free-speech-is-not-the-same-as-free-reach/.
- Ndahinda, Felix M., & Aggée S. Mugabe. 2022. “Streaming Hate: Exploring the Harm of Anti-Banyamulenge and Anti-Tutsi Hate Speech on Congolese Social Media.” Journal of Genocide Research. 1-25. http://www.francegenocidetutsi.org/ExploringTheHarmOfHateSpeechJOGR19052022.pdf.
- Napier, Sarah, Coen Teunissen, & Hayley Boxall. 2021. “How Do Child Sexual Abuse Live Streaming Offenders Access Victims?” Trends and Issues in Crime and Criminal Justice (642). 1-18. https://search.informit.org/doi/abs/10.3316/agispt.20220310063199.
- Whittaker, Jack M., Matthew Edwards, Casandra Cross, & Mark Button. 2022. “I Have Only Checked after the Event.” Consumer Approaches to Safe Online Shopping. Victims & Offenders. 1-23.

https://www.tandfonline.com/doi/pdf/10.1080/15564886.2022.2130486?needAccess=true&role=button.

Module M: Emerging Technologies and Career Advice

Contributors: Michael Swenson (Meta); Amar Ashar (Berkman Klein Center)

Core Content

- Video: American Enterprise Institute. 2023. Is Artificial Intelligence Effective at Content Moderation? [Video]. YouTube. https://www.youtube.com/watch?v=6jT9LBea8hc.
- Moody's Analytics. 2023. "Navigating the AI Landscape: Insights From Compliance and Risk Management Leaders." https://www.moodys.com/web/en/us/site-assets/ma-kyc-navigating-the-ai-landscape-report.pdf.
- Goldstein, Josh, A., Sastry Girish, Micah Musser. Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023. "Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations." arXiv preprint. https://arxiv.org/abs/2301.04246.
- Shulruff, Toby. 2022. "Trust & Safety: A Snapshot of the Field 2022." https://sites.google.com/asu.edu/trustandsafety/report?authuser=0.
- Trust & Safety Professional Association. 2023. "Trust & Safety Curriculum: Key Functions and Roles." https://www.tspa.org/curriculum/ts-curriculum/functions-roles/.
- Video: Trust & Safety Professional Association. 2022. "Careers in Trust & Safety AMA | Dona Bellow and Christine Lehane." YouTube. https://www.youtube.com/watch?v=WGfy5R7S99E.

Optional Content

- Gillespie, Tarleton; 2020. "Content Moderation, AI, and the Question of Scale." Big Data & Society 7(2). https://journals.sagepub.com/doi/full/10.1177/2053951720943234.
- Udupa, Sahana, Antonis Maronikolakis, Hinrich Schütze and Axel Wisiorek. "Ethical Scaling for Content Moderation: Extreme Speech and the

(In)Significance of Artificial Intelligence." Harvard Kennedy School Shorenstein Center Discussion Papers. https://shorensteincenter.org/ethical-scaling-content-moderation-extreme-speech-insignificance-artificial-intelligence/.
- "Challenges of Incorporating Algorithmic 'Fairness' Into Practice." https://www.microsoft.com/en-us/research/uploads/prod/2020/10/FAT_2019-tutorial_-algorithmic-bias-in-practice.pdf.
- Altman, Sam. 2023. "Planning for AGI and Beyond." https://openai.com/blog/planning-for-agi-and-beyond/.
- Anthropic. 2023. "Core Views on AI Safety: When, Why, What, and How." https://www.anthropic.com/index/core-views-on-ai-safety.
- National Institute of Standards and Technology. 2019. "Artificial Intelligence Risk Management Framework (AI RMF 1.0)." https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

Supplementary Module N: Copyright Safe Harbors and the DMCA

Contributor: Justin Francese (University of Oregon)

Core Content

- Keller, Daphne. 2020. "Intermediary liability 101: An update for 2020." Center for Internet and Society. https://cyberlaw.stanford.edu/blog/2020/01/intermediary-liability-101-update-2020.
- Digital Trust and Safety Partnership. 2023. "Copyright" and "Safe Harbors" in the Digital Trust and Safety Partnership's Trust and Safety Glossary of Key Terms. https://dtspartnership.org/wp-content/uploads/2023/07/DTSP_Trust-Safety-Glossary_July-2023.pdf.
- Harvard's Berkman Klein Center for Internet & Society. 2023. "Copyright Claims Based on User Content." Digital Media Law Project. http://www.dmlp.org/legal-guide/copyright-claims-based-user-content.
- Keller, Daphne. 2021. "Empirical evidence of over-removal by internet companies under Intermediary Liability Laws: An updated list." Center for

Internet and Society. https://cyberlaw.stanford.edu/blog/2021/02/empirical-evidence-over-removal-internet-companies-under-intermediary-liability-laws.
- Urban, Jennifer M., Karaganis, Joe., Schofield, Brianna. L. 2016. "Notice and Takedown in Everyday Practice." SSRN Electronic Journal. https://doi.org/10.2139/ssrn.2755628.

Optional Content

- Bridy, Annemarie., Keller, Daphne. 2017. "US Copyright Office Section 512 Study: Comments in Response to Second Notice of Inquiry." SSRN Electronic Journal. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2920871.
- Engstrom, Evan., Feamster, Nick. 2017. "The Limits of Filtering:: A Look at the Functionality & Shortcomings of Content Detection Tools 27." ENGINE. https://www.engine.is/the-limits-of-filtering.
- Kulk, Stefan. 2020. "Internet Intermediaries and Copyright Law. Towards a Future-proof Legal Framework." SSRN Electronic Journal. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3584157.
- Grimmelmann, James. 2022. Chapter 7: Copyright. Part F: Section 512. In Internet law: Cases & problems. essay, Semaphore Press. https://internetcasebook.com/ (purchase required).
- Sutton, Maria. 2014. "Copyright Law as a Tool for State Censorship of the Internet." Electronic Frontier Foundation. https://www.eff.org/deeplinks/2014/12/copyright-law-tool-state-internet-censorship.