COMM4010/5010 | Governing the internet

Critical perspectives on online trust and safety

Dr. Yoel Roth *(he/him)*
yoel@upenn.edu

Who decides what's allowed (or banned) online — and how are these choices made and implemented?

Online platforms have become a ubiquitous part of how people socialize, do their jobs, find love and romance, and express their political views; they shape virtually every part of human experience for billions of people around the world. But their governance — the rules that structure what people can and can't do on online platforms — remains obscure and poorly understood.

This interdisciplinary course examines the histories, theories, policies, and technologies of internet governance that make up the emerging field of online trust and safety. Over the course of the semester, we will examine a wide range of topics related to platform governance, each meant to expose how platforms negotiate their position in the world — and in turn, how they govern the conduct of billions of users worldwide. You will learn about the different harms online platforms have to contend with — from governments meddling in elections to abuse and harassment — and the strategies platforms have employed to contend with these challenges. Drawing on a diverse range of scholarly traditions, including empirical studies of platform behavior, legal analyses, historical documents, communication theory, and the history of science, this course aims to provide a deep perspective on how platforms see themselves, the forces that shape platform behavior, and how the decisions underlying platform behavior shape the online (and offline) worlds we experience.

This course is intended for graduate students and advanced undergraduates in communication, political science, sociology, law, computer science, and related fields. While there are no specific prerequisites for this course, some familiarity with at least a few of the following topics will be beneficial: Digital/new media (history, theories, effects); online safety issues; media effects; media industries and news; free speech (First Amendment and related) and human rights law; and social science research methods.

Course structure and expectations

This course is structured as an advanced seminar, with few lectures (other than guest speakers), and a primary focus on group discussion and engagement grounded in the assigned readings. Accordingly, participation is an essential component of what you get out of this class. You are expected to attend class regularly (duh).

Let me be direct: **There is a lot of assigned reading.** You are expected to come to class each week familiar with the readings, and prepared to offer a text-informed perspective on them. Please don't waste the class's time trying to fake your way through a discussion if you aren't prepared (*e.g.* by offering a superficial summary from what ChatGPT was able to burp up); we've all been in classes with *that person*, and it sucks. If you're in a time crunch and don't get to everything (it happens), I've indicated *absolutely essential* readings in the syllabus with a ✷; prioritize these. The syllabus also includes optional readings if you want to go deeper on a specific topic — and I'm happy to suggest additional readings in specific areas if you have a particular interest.

Class discussions will touch on controversial issues, including on divisive political, social, and cultural issues. You can and should feel comfortable expressing yourself and your viewpoints in class — but keep the following in mind: (1) Treating your peers with respect is non-negotiable, even if (and especially if) you disagree. If things get heated, I'll moderate, but it's never a bad idea to pause, take a breath, and approach even intense disputes with curiosity and a desire to understand others' views (rather than jumping to criticism or condemnation). (2) This is an academic course, not a political rally; your focus should be on understanding the theories and research we're discussing. It's fine to excavate how personal politics may influence the subjects we're studying, but try to keep discussions grounded in the assigned texts and research, rather than statements of personal belief and opinion.

Generative AI

Academic integrity in the age of generative AI is a fraught and complicated question — and one which we'll engage with substantively as part of this course. These technologies raise substantial issues: about information authorship, ownership, and attribution; about sustainability and the ecological costs of computing; about bias, diversity, and equity; and, critically, about the safety and accuracy of model outputs.

I recognize that learning about technology requires substantive engagement with technology — and to that end, if you feel that there are areas where generative AI can improve your comprehension of course material or lead you to novel analytic insights, you may thoughtfully employ these tools as part of your work. Equally,

you should remember that cutting corners in this course (or in your other studies) by uncritically using generative AI in place of your own thinking and writing ultimately is just a waste of your time and money. The right balance of these considerations is a personal question that only you can resolve.

That being said, please keep the following principles for acceptable use of generative AI in mind:

- Bottom line up front: You are allowed, but not required, to use generative AI as part of this course. No assignments will require you to use generative AI, although both the midterm and final project have options for engaging with and critiquing generative AI for trust and safety purposes. No part of the course is structured to give an advantage (or disadvantage) to students who do or don't use generative AI.
- I **strongly** urge you not to use generative AI to directly write any of the text you submit to me as part of course assignments. Even if you use generative AI for brainstorming or idea generation, I believe that there is learning value in engaging with model outputs and reworking them into your own words and ideas. To be clear, I don't recommend "paraphrasing whatever ChatGPT came up with" as a strategy for success in this course, but do what you feel is right.
- In the event you choose to directly submit the output of generative AI as part of an assignment (which, again, I urge you not to do), you **must** clearly indicate and cite your use of AI, including an appendix of approximately half a page detailing the model and prompts used, and a brief reflection on your experience).

Plagiarism

Plagiarism, which the University defines as "using the ideas, data, or language of another without specific or proper acknowledgement," is never acceptable. All the work you produce for this class must be original and your own.

I recognize that, especially for graduate students, you may want to work on a long-term, continuing research project as part of this course, which will involve engaging with writing or data you may have produced for other classes or projects; please meet with me to discuss your plans so we can agree on a sufficient scope of novel work for this class.

Accommodations and course content

The University of Pennsylvania provides reasonable accommodations to students with disabilities who have self-identified and received approval from Disability

Services. Students can contact Disability Services and make appointments to discuss and/or request accommodations by calling 215-573-9235.

Content Warning:

As an unavoidable part of studying content moderation and online safety issues, the reading material in this course and course assignments may involve exposure to potentially sensitive content, including (but not limited to): nudity and adult sexual content; violent and gory imagery; hateful and derogatory speech; descriptions and depictions of sexual assault and domestic violence; and discussions of sexual abuse, including child sexual exploitation.

I will make every effort to limit exposure to these materials to only what is strictly academically necessary. If one particular topic in the course's syllabus represents a particular sensitivity for you, please raise it with me in advance so we can discuss alternate arrangements. If several topics represent potential areas of heightened sensitivity for you, regrettably, this may not be the right class for you.

The University has significant resources available to support you:

- Student Health and Counseling: https://wellness.upenn.edu/ and 215-746-WELL (9355)
- Free and confidential counseling through the Let's Talk program: https://wellness.upenn.edu/counseling/lets-talk
- Graduate and Professional Students Association support resources: https://www.gapsa.upenn.edu/avenues-for-support-and-counseling

Assignments and grading

Your grade in this class is comprised of three (*graduate students: four*) primary components:

1. **Reading memos:** A short essay (approximately 500 words, or 1 page) engaging with the week's reading materials, *due by 5pm Eastern on the Sunday before class*. You are required to complete 5 (or more, for the overachievers in the class) memos over the course of the semester, and I encourage you to distribute them evenly (and not wait until the end of the semester when you have a bunch of other stuff to do).
    - ***What does "engaging with the week's reading materials" mean?*** In brief, not just a summary you copy-pasted from ChatGPT. You can

endorse, critique, expand, question, connect, or complain — but I expect you to demonstrate an understanding of the week's theme and readings. *Please conclude each essay with 1-2 suggested questions for discussion related to the reading(s) you wrote about*. The most successful essays will connect readings to current events or historical case studies.

- ***How are reading memos graded?*** Memo grades are a single component of your overall course grade, representing an average of the five highest scores your essays receive. (You may submit more than five essays over the course of the semester.) Each essay will receive one of the following scores:
  - ✔+: Advances clear, critical insights that demonstrate a mastery of the assigned readings *(consistent with A-level performance in the course)*
  - ✔: Engages with the assigned text(s), but doesn't go too far beyond summary *(consistent with B-level performance in the course)*
  - ✔-: Superficial engagement with the text(s) that lacks details, textual evidence/references, and/or effort, or submissions that arrive after the 5pm Sunday deadline; let's meet to discuss how to improve *(consistent with C-level performance in the course)*

2. ***Graduate students only:*** **Class discussion facilitation:** You will facilitate the class discussion on 2 weeks of your choice. The role of a facilitator is to offer three key things to the class: (1) A brief introduction of the week's topic and key takeaways from the assigned readings, totaling approximately 15 minutes; (2) an overview of one (or several) instructive case studies related to the week's theme; and (3) a set of questions to spark discussion within the class. An outline of your case study and questions for discussion are *due to me by 5pm Eastern on the Sunday before class*.

3. **Midterm project:** Choose one of the following tracks for your midterm project:

   - ***Policy track:*** Review recent pending cases and decisions from the Facebook Oversight Board, and select one case. In approximately 5 pages (double-spaced), write a preliminary advisory brief for the Oversight Board discussing the case and making a recommendation about how the Board should rule and what actions Facebook should take in response to the incident. The brief should represent *your perspective* on the case — not how you think the Board will rule. (In historical cases, your brief need not correspond with how the Board in fact ruled; there is no right or wrong answer. In the event your

recommendation aligns with the Board's actual ruling, make sure you substantiate your response with unique reasoning, and not just a restatement of the existing public opinion.)

- ***Implementation track:*** Using ChatGPT, Bard, Claude, or another large language model of your choice, develop and deploy a prompt that can be used to enforce a platform policy you develop for dangerous and violent speech. (I will provide prepaid gift cards for paid model usage if you pursue this track.) You will be provided with a sample dataset of posts to moderate. In addition to a dataset labeled in accordance with a moderation taxonomy you develop, you should write a brief (2-3 pages, double-spaced) document that includes: the prompt you used; the moderation taxonomy/definitions you developed; discussion of how you developed and iterated on your taxonomy and prompt; discussion of how you evaluated the effectiveness/accuracy of your prompt and moderation approach; and discussion of issues, caveats, and challenges associated with your chosen approach. Note that there is no "right" answer to this project, and you are not graded on how accurately or "correctly" you label posts; rather, you are evaluated on the rigor of your approach and your analytic perspective on the use of AI in moderation.
- ***Choose your own adventure:*** For students planning to develop a unique, empirical research project for the final paper, you can produce an initial 6-8 page (double-spaced) project proposal and preliminary literature review in lieu of a standalone midterm project. Students pursuing independent projects should plan to review their ideas with me before proceeding too far down this path.

Regardless of which track you choose, your midterm project should represent an initial sketch of the research project you will pursue for your final paper (although you can always change course if your interests shift). Projects are due by 5pm Eastern on Friday, March 1, 2024.

4. **Final paper:** Drawing on your midterm project, develop a paper of approximately 10-15 pages double-spaced (*graduate students: 25-30 pages, double-spaced*), including references, that offers a unique, research-driven perspective on a content moderation topic of your choice.
    - **Undergraduate guidelines:** Successful papers will present a straightforward, analytical accounting of a content moderation issue; precisely describe the scene before evaluating it. Drawing on the theories and frameworks we covered in class, as well as additional research you conduct into related content moderation issues and approaches, cultural context, legal paradigms, etc, make a cogent

argument discussing how this content moderation case should impact platform governance, policy, and product development. The most effective papers will demonstrate an understanding of the competing values and considerations faced by technology companies and users, and offer a persuasive, data-driven perspective on these issues.

- **Graduate student guidelines:** You may choose to pursue an expanded version of the undergraduate paper assignment, or a unique research project of your choosing. Papers should take care to position topics within a broad and well-researched historical and theoretical context. Some degree of novel, empirical analysis is essential (recognizing that there are unavoidable limits to what you can execute during one class in one semester); document and draw on data (qualitative and/or quantitative) to support your arguments and ground theoretical/conceptual analyses in real-world evidence and applications. Don't feel constrained by this assignment; if you have a project related to this course's themes that would advance your overall research interests or goals (but isn't exactly related to a content moderation issue we cover in class), discuss it with me and we can figure out a productive path for your work in this course.

Projects are due by 5pm Eastern on Monday, May 6, 2024.

Each of these components is factored into your overall grade:

Component	Undergraduate students	Graduate students
Reading memos	25%	15%
Discussion facilitation *(graduate students only)*	-	10%
Midterm project	25%	25%
Final paper	50%	50%

Readings

W1:// Origins

1/23/2024

Why bother restricting what people can say online? In our first week, we'll engage broadly with the harms (physical, psychological, emotional, economic, societal, informational, and interpersonal) that can result from networked interactions, and consider how one of the earliest online communities, LambdaMOO, navigated abusive behavior by a dedicated bad actor.

Read

- ✷ Julian Dibbell (1993). A Rape In Cyberspace. *Village Voice*, 23 December 1993. https://www.villagevoice.com/a-rape-in-cyberspace/
- World Economic Forum (2023). Toolkit for Digital Safety Design Interventions and Innovations: Typology of Online Harms. https://www.weforum.org/publications/toolkit-for-digital-safety-design-interventions-and-innovations-typology-of-online-harms

Watch

- Del Harvey (2014). Protecting Twitter users (sometimes from themselves). TED. (video)

Optional

- Philip Elmer-Dewitt (2001). Battle for the Soul of the Internet. *Time*, 24 June 2001. https://content.time.com/time/magazine/article/0,9171,164784,00.html
- Tarleton Gillespie (2010). The politics of platforms. *New Media & Society* 12(3). https://doi.org/10.1177/1461444809342738
- José van Dijck (2013). *The Culture of Connectivity: A critical history of social media*. Oxford University Press. *(Especially chapters 1 and 2)*
- Langdon Winner (1988). Do Artifacts Have Politics?. In *The Whale and the Reactor: A Search for Limits in an Age of High Technology*. University of Chicago Press.

W2:// Governance outside of government

1/30/2024

Platforms exist as a strange duality: they're software companies that build (sometimes) fun, (sometimes) useful, and (sometimes) profitable tools for talking with friends and sharing photos of what you had for lunch; but also, they're essential parts of our civic and economic lives, and seem to operate like mini-governments. We'll examine three essential perspectives on how platforms navigate this duality: (1) A regulatory account of one of the most important laws in the history of modern platforms, Section 230 of the Communication Decency Act; (2) a legal framework for understanding

	platforms as quasi-legal "governors" of speech and conduct; and (3) a recognition of the mix of public, financial, and regulatory pressures that shape how platforms make governance decisions.
Read	• Daphne Keller (2019). Facebook Restricts Speech by Popular Demand. *The Atlantic*, 22 September 2019. https://www.theatlantic.com/ideas/archive/2019/09/facebook-restricts-free-speech-popular-demand/598462/ • ✷ Kate Klonick (2018). The New Governors: The People, Rules, and Processes Governing Online Speech. *Harvard L. Rev.* 131(6). https://harvardlawreview.org/print/vol-131/the-new-governors-the-people-rules-and-processes-governing-online-speech/ • Jeff Kosseff (2019). *The Twenty-Six Words That Created the Internet*. Cornell University Press. *(Chapter 7 only)*
Watch	• Radiolab (2018). Post No Evil. (podcast) • *Optional:* David Kaye (2019). Speech Police: The Global Struggle to Govern the Internet. Carnegie Council for Ethics in International Affairs, 6 June 2019. (video)
Optional	• Chinmayi Arun (2022). Facebook's Faces. *Harvard L. Rev. Forum* 236. https://harvardlawreview.org/forum/no-volume/facebooks-faces/ • John Perry Barlow (1996). A Declaration of the Independence of Cyberspace. Electronic Frontier Foundation, 8 February 1996. https://www.eff.org/cyberspace-independence • Frank Easterbrook (1996). Cyberspace and the Law of the Horse. *University of Chicago L. Forum* 1996. https://chicagounbound.uchicago.edu/uclf/vol1996/iss1/7/ • Daphne Keller (2018). Internet Platforms: Observations on Speech, Danger, and Money. Hoover Institution Aegis Paper Series, 1807. https://www.hoover.org/research/internet-platforms-observations-speech-danger-and-money • Lawrence Lessig (1999). The Law of the Horse: What Cyberlaw Might Teach. *Harvard L. Rev.* 113(2). https://cyber.harvard.edu/works/lessig/LNC_Q_D2.PDF • Chris Reed (2018). Why judges need jurisprudence in cyberspace. *Legal Studies* 38(2). https://www.cambridge.org/core/journals/legal-studies/article/abs/why-judges-need-jurisprudence-in-cyberspace/BE9DB598337A1F16AC138FC4BCA1F210

W3:// What is content moderation?

2/6/2024

Meeting via Zoom.

	By the early 2010s, content moderation and trust and safety emerged as specific fields of practice within platforms. What does this work look like? What are its goals and politics? We'll examine the broad contours of how platforms wrestle with governance as an essential part of their business practices, and why this work always, inevitably, involves making value-laden tradeoffs.
Read	• ✷ Tarleton Gillespie (2021). *Custodians of the Internet*. Yale University Press. • Mike Masnick (2022). Hey Elon: Let Me Help You Speed Run The Content Moderation Learning Curve. *Techdirt*, 2 November 2022. https://www.techdirt.com/2022/11/02/hey-elon-let-me-help-you-speed-run-the-content-moderation-learning-curve/
Watch	• Alex Macgillivray & Nicole Wong (2020). "Origins of Trust and Safety with Robyn Caplan." (podcast) • *Play:* Mike Masnick, Randy Lubin, & Leigh Beadon (2023). Moderator Mayhem. https://moderatormayhem.engine.is/
Optional	• Robyn Caplan (2018). Content or context moderation?. Data & Society Research Institute. https://datasociety.net/library/content-or-context-moderation/ • Alex Feerst (2023). A Natural History Of Trust & Safety. *Techdirt*, 7 June 2023. https://www.techdirt.com/2023/06/07/a-natural-history-of-trust-safety/ • Jonathon Penney (2022). Understanding Chilling Effects. *Minnesota L. Rev.* 106(3). https://minnesotalawreview.org/article/understanding-chilling-effects/

W4://

2/13/2024

The dirty work of internet sanitation

From content *moderation* to content *moderators*: Who are the people responsible for "internet sanitation," and what are the impacts this work has on their safety and wellbeing? We imagine that content moderation is the product of shadowy algorithms and machine learning models; but oftentimes, it's real people, spread around the world, keeping the worst bits of the internet at bay. This week engages with their experiences, and the costs platforms externalize in order to keep their users safe.

Read	• ✷ Casey Newton (2019). The Trauma Floor: The Secret Lives of Facebook Moderators in America. *The Verge*, 25 February 2019. https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona

	• Billy Perrigo (2023). OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic. *Time Magazine*, 18 January 2023. https://time.com/6247678/openai-chatgpt-kenya-workers/ • Sarah Roberts (2021). *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press. *(Chapter 2 only)*
Watch	• Hans Block and Moritz Riesewieck, *dirs*. (2018). The Cleaners. PBS Independent Lens. • Adrian Chen & Ciaran Cassidy, *dirs*. (2017). The Moderators. Field Of Vision. (video)
Optional	• Adrian Chen (2014). The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed. *Wired*, 23 October 2014. https://www.wired.com/2014/10/content-moderation/

W5:// Identities, bodies, and communities

2/20/2024

Facebook's stated mission, for years, has been to "connect every person on the planet." This vision of a global community encompassing billions of people is, in important ways, the root of trust and safety's enduring challenges. We examine the fraught concepts of "community" and "identity" in the context of social networks and content moderation from three vantage points: how people see themselves and their audiences when interacting online; how communities constitute themselves through technology; and how those same technologies shape what is knowable and doable.

Read

- Rena Bivens (2015). The gender binary will not be deprogrammed: Ten years of coding gender on Facebook. *New Media & Society* 19(6). https://doi.org/10.1177/1461444815621527
- André Brock Jr. (2020). *Distributed Blackness: African American Cybercultures*. NYU Press. *(Chapters 1 and 3 only; chapters 4 and 5 highly recommended)*
- Alice Marwick and danah boyd (2011). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media and Society* 13(1). https://doi.org/10.1177/1461444810365313
- ✷ Safiya Umoja Noble (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press. *(Chapter 1 only)*

Watch	•
Optional	• danah boyd (2007). Viewing American Class Divisions Through Facebook and MySpace. Apophenia Blog Essay, 24 June 2007. https://www.danah.org/papers/essays/ClassDivisions.html • Simone Browne (2010). Digital epidermalization: Race, identity and biometrics. *Critical Sociology* 36(1). https://doi.org/10.1177/0896920509347144 • Ian Hacking (2006). Making up people. *London Review of Books* 28(18). https://www.lrb.co.uk/the-paper/v28/n16/ian-hacking/making-up-people • Donna Haraway (1991). A cyborg manifesto. In D. Haraway, *Simians, cyborgs, and women: The reinvention of nature*, Routledge. • Thomas Nagel (1974). What is it like to be a bat? *Philosophical Review* 83(4). https://www.jstor.org/stable/2183914 • Lisa Nakamura (2001). Race in/for cyberspace: Identity tourism and racial passing on the internet. In D. Trend, *Reading digital culture*, Wiley. https://smg.media.mit.edu/library/nakamura1995.html • A. R. Stone (1995). *The war of desire and technology at the close of the mechanical age*. MIT Press. *(Especially chapter 1)*

W6:// Safety and extremism

2/27/2024

What happens online doesn't stay online — but how, exactly, do online interactions translate into offline harms? Starting with legal analyses of how concepts like "hate speech" can be understood in digital contexts, we examine the ways that abuse, harassment, and violent extremism manifest on social media, and why platforms struggle to adapt to novel malign uses of their products.

Read

- Danielle Citron (2016). *Hate Crimes in Cyberspace*. Harvard University Press. *(Chapters 1, 2, and 3 only; chapters 5 and 6 highly recommended)*
- ✷ Nina Jankowicz, Jillian Hunchak, Alexandra Pavliuc, Celia Davies, Shannon Pierson, & Zoë Kaufmann (2021). Malign Creativity: How Gender, Sex, and Lies are Weaponized Against Women Online. The Wilson Center. https://www.wilsoncenter.org/publication/malign-creativity-how-gender-sex-and-lies-are-weaponized-against-women-online
- ✷ J. Nathan Mathias (2017). The Real Name Fallacy. *Coral by Vox Media*, 3 January 2017. https://coralproject.net/blog/the-real-name-fallacy/

Watch	• Kevin Roose (2019). The Making of a YouTube Radical. *New York Times*, 8 June 2019. https://www.nytimes.com/interactive/2019/06/08/technology/youtube-radical.html • Steve Stecklow (2018). Why Facebook is losing the war on hate speech in Myanmar. *Reuters*, 15 August 2018. https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/
	• Susan Benesch (2018). What is Dangerous Speech? Dangerous Speech Project. (video) • *Optional:* Evelyn Douek, Quinta Jurecic, & Jeff Kosseff (2021). Finstas, Falsehoods and the First Amendment. Lawfare Podcast, 14 October 2021. (podcast)
Optional	• Matt Goerzen (2019). The Ironic Hedge: Political Uses of Irony Online. Data & Society, prepublication draft. • Sarah Jeong (2018). The Internet of Garbage (v. 1.5). *The Verge*. https://www.theverge.com/2018/8/28/17777330/internet-of-garbage-book-sarah-jeong-online-harassment • Erin Kissane (2023). Meta in Myanmar. https://erinkissane.com/meta-in-myanmar-full-series ○ *See also (but with a very significant grain of salt):* BSR (2018). Human Rights Impact Assessment: Facebook in Myanmar. https://about.fb.com/news/2018/11/myanmar-hria/ • Adrienne LaFrance (2020). The Prophecies of Q. *The Atlantic*, June 2020. https://www.theatlantic.com/magazine/archive/2020/06/qanon-nothing-can-stop-what-is-coming/610567/ • Whitney Phillips (2015). *This Is Why We Can't Have Nice Things: Mapping the Relationship between Online Trolling and Mainstream Culture*. MIT Press. *(Especially chapters 5, 7, and 9)* • Amanda Taub & Max Fisher (2018). Where Countries Are Tinderboxes and Facebook Is a Match. *New York Times*, 21 April 2018. https://www.nytimes.com/2018/04/21/world/asia/facebook-sri-lanka-riots.html • Charlie Winter, Peter Neumann, Alexander Meleagrou-Hitchens, Magnus Ranstorp, Lorenzo Vidino, & Johanna Fürst (2020). Online Extremism: Research Trends in Internet Activism, Radicalization, and Counter-Strategies. *International Journal of Conflict and Violence* 14(2). https://doi.org/10.4119/ijcv-3809

W7:// Sex

3/12/2024

"The internet is for porn," sang a cast of muppets in the musical Avenue Q.

	And certainly, from the earliest networked platforms like the French Minitel and USENET, communicative technologies have been inextricably linked with that most human desire to find love and sex. Yet, despite the internet's ribald roots, modern platforms struggle to navigate appropriate governance of sex and sexuality, instead ending up mired in endless debates about breastfeeding images, female-presenting nipples, and — gasp! — twerking. Is there a path for principled policymaking about sex?
Read	• Alice Marwick (2008). To catch a predator? The MySpace moral panic. *First Monday* 13(6). https://firstmonday.org/ojs/index.php/fm/article/view/2152/1966 • Roni Rosenberg & Hadar Dancig-Rosenberg (2021). Reconceptualizing Revenge Porn. *Arizona L. Rev.* 63. https://arizonalawreview.org/pdf/63-1/63arizlrev199.pdf • Yoel Roth (2015). "No Overly Suggestive Photos of Any Kind": Content Management and the Policing of Self in Gay Digital Communities. *Communication, Culture, & Critique* 8(3). https://doi.org/10.1111/cccr.12096 • ✷ Jillian York (2021). Silicon Values: The Future of Free Speech Under Surveillance Capitalism. Verso. *(Chapters 6 and 7 only)*
Watch	• Paul Detrick (2019). The War on Backpage.com is a War on Sex Workers. *Reason*. (video)
Optional	• John Edward Campbell (2004). *Getting It On Online: Cyberspace, Gay Male Sexuality, and Embodied Identity*. Harrington Park Press. • Yasmin Ibrahim (2017). Facebook and the Napalm Girl: Reframing the Iconic as Pornographic. *Social Media + Society* 3(4). https://doi.org/10.1177/2056305117743140 • Elena Pilipets (2020). Nipples, memes, and algorithmic failure: NSFW critique of Tumblr censorship. *New Media & Society* 24(6). https://doi.org/10.1177/1461444820979280

W8:// Kids, youth culture, wellbeing, and exploitation

3/19/2024

Protecting kids and teenagers from the harms of technology is an age-old preoccupation (see also: moral panics about books, television, and movies). How much do we know about how kids use technology, and what its effects are on their health and wellbeing? What are the true threats facing children, such as sexual exploitation, and how are these concepts mobilized to advance political agendas? How do seemingly well-intentioned regulations focused on child safety wind up negatively impacting the most vulnerable kids? This week begins to unpack perhaps the most fraught of all content governance questions, arriving, as danah boyd puts it, at an unsatisfying

	conclusion: It's complicated.
Read	• ✷ danah boyd (2014). *It's Complicated: The Social Lives of Networked Teens*. Yale University Press. *(Chapters 1, 2, 4, and 5)* • Mike Masnick (2023). APA Report Says That Media & Politicians Are Simply Wrong About Kids & Social Media; Media Then Lies About Report. *Techdirt*, 12 May 2023. https://www.techdirt.com/2023/05/12/apa-report-says-that-media-politicians-are-simply-wrong-about-kids-social-media-media-then-lies-about-report/ • David Thiel & Renee DiResta (2023). Addressing Child Exploitation on Federated Social Media. Stanford Internet Observatory. https://doi.org/10.25740/vb515nd6874
Watch	• *Optional:* Alex Winter (2022). The YouTube Effect.
Optional	• American Psychological Association (2023). Health Advisory on Social Media Use in Adolescence. https://www.apa.org/topics/social-media-internet/health-advisory-adolescent-social-media-use • Shelley L. Craig, Andrew D. Eaton, Lauren B. McInroy, Vivian W. Y. Leung, & Sreedevi Krishnan (2021). Can Social Media Participation Enhance LGBTQ+ Youth Well-Being? Development of the Social Media Benefits Scale. *Social Media + Society* 7(1). https://doi.org/10.1177/2056305121988931 • Bree Holtz & Shaheen Kanthawala (2020). #T1DLooksLikeMe: Exploring Self-Disclosure, Social Support, and Type 1 Diabetes on Instagram. *Frontiers in Communication* 5. https://doi.org/10.3389/fcomm.2020.510278 • Lisa Miller (2023). Tate-Pilled: What a generation of boys have found in Andrew Tate's extreme male gospel. *New York Magazine*, 14 March 2023. https://nymag.com/intelligencer/article/andrew-tate-jail-investigation.html

W9:// Politics, polarization, and misinformation

3/26/2024

Infamously, Facebook began as a website for Harvard students to rate the attractiveness of their peers — and somehow, over the subsequent decade, morphed into a powerful and pervasive force in political discourse in the United States and globally. We examine the impacts that social media platforms (and the governance of those platforms) can have on political discourse, from propaganda and fake news to polarization — and question whether the received wisdom about social media's negative impact on political life is really supported by the evidence.

Read

- Josh Goldstein, Jason Chao, Shelby Grossman, Alex Stamos, & Michael Tomz (2023). Can AI Write Persuasive Propaganda? Forthcoming, preprint on SocArXiv, 8 April 2023. https://osf.io/preprints/socarxiv/fp87b/
- Sandra Gonzalez-Bailon, David Lazer, et al (2023). Asymmetric ideological segregation in exposure to political news on Facebook. *Science* 381(6656). https://www.science.org/doi/10.1126/science.ade7138
- Ferenc Huszár, Sofia Ira Ktena, Conor O'Brien, Luca Belli, Andrew Schlaikjer, & Moritz Hardt (2021). Algorithmic amplification of politics on Twitter. *PNAS* 119(1). https://doi.org/10.1073/pnas.2025334119
- ✷ Claire Wardle (2017). Fake News. It's Complicated. First Draft, 16 February 2017. https://firstdraftnews.org/articles/fake-news-complicated/

Watch

- Barack Obama (2022). Speech at Stanford Cyber Policy Center. (video)

Optional

- Chris Bail (2021). *Breaking the Social Media Prism: How to Make Our Platforms Less Polarizing*. Princeton University Press.
- Yochai Benkler, Robert Farris, & Hal Roberts (2018). *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics*. Oxford University Press.
- Election Integrity Partnership (Center for an Informed Public, Digital Forensic Research Lab, Graphika, & Stanford Internet Observatory) (2021). The Long Fuse: Misinformation and the 2020 Election. https://www.eipartnership.net/report
- Daniel Kreiss & Shannon McGregor (2023). A review and provocation: On polarization and platforms. *New Media & Society* OnlineFirst. https://doi.org/10.1177/14614448231161880
- Nandita Krishnan, Jiayan Gu, Rebekah Tromble, & Lorien Abroms (2021). Research note: Examining how various social media platforms have responded to COVID-19 misinformation. *Harvard Kennedy School Misinformation Review* 2(6). https://misinforeview.hks.harvard.edu/article/research-note-examining-how-various-social-media-platforms-have-responded-to-covid-19-misinformation/
- Whitney Phillips (2019). The Toxins We Carry. *Columbia Journalism Review*, 2 December 2019. https://www.cjr.org/special_report/truth-pollution-disinformation.php
- David Scales, Jack Gorman, & Kathleen Hall Jamieson (2021). The Covid-19 Infodemic — Applying the Epidemiologic Model to Counter Misinformation. *New England Journal of Medicine* 2021(358). https://www.nejm.org/doi/full/10.1056/NEJMp2103798

W10:// Bots, troll farms, and disinformation

4/2/2024

Since the US government's bombshell revelations in 2017 that agents of the Russian government had engaged in a multi-pronged campaign to interfere in American elections in 2016, discussions of bots, trolls, and disinformation have become inseparable from how we think about social media's impact on politics. We examine disinformation as a unique class of content moderation problem — one that, arguably, doesn't involve the "content" part of "content moderation" very much at all — and assess how and why coordinated manipulation campaigns upended public trust in social media platforms.

Read

- ✷ Adrian Chen (2015). The Agency. *New York Times Magazine*, 2 June 2015. https://www.nytimes.com/2015/06/07/magazine/the-agency.html
- ✷ Camille François (2019). Actors, Behaviors, Content: A Disinformation ABC. Annenberg Public Policy Center, Transatlantic Working Group. https://cdn.annenbergpublicpolicycenter.org/wp-content/uploads/2020/05/ABC_Framework_TWG_Francois_Sept_2019.pdf
- Kathleen Hall Jamieson (2018). *Cyberwar: How Russian Hackers and Trolls Helped Elect a President (2nd edition)*. Oxford University Press. *(Introduction, Part 1, and Afterword only)*
- Samanth Subramanian (2017). Welcome to Veles, Macedonia, Fake News Factory to the World. *Wired*, 15 February 2017. https://www.wired.com/2017/02/veles-macedonia-fake-news/
- Harry Yaojun Yan, Kai-Cheng Yang, Filippo Menczer, and James Shanahan (2020). Asymmetrical perceptions of partisan political bots. *New Media & Society* 23(10). https://doi.org/10.1177/1461444820942744

Watch

- *Optional:* Karim Amer & Jehane Noujaim (2019). *The Great Hack*. Netflix.

Optional

- Ahmer Arif, Leo Stewart, & Kate Starbird (2018). Acting the Part: Examining Information Operations Within #BlackLivesMatter Discourse. Proceedings of the ACM on Human-Computer Interaction 2, CSCW. https://doi.org/10.1145/3274289
- Katie Benner, Mark Mazzetti, Ben Hubbard, & Mike Isaac (2018). Saudis' Image Makers: A Troll Army and a Twitter Insider. *New York Times*, 20 October 2018. http://www.nytimes.com/2018/10/20/us/politics/saudi-image-campaign-twitter.html
- Joseph Bernstein (2021). Bad News. *Harper's Magazine*, September 2021. https://harpers.org/archive/2021/09/bad-news-selling-the-story-of-disinformation/

- Joan Donovan & Brian Friedberg (2019). Source hacking: Media manipulation in practice. Data & Society. https://datasociety.net/library/source-hacking-media-manipulation-in-practice/
- Miriam Elder & Charlie Warzel (2018). Stop Blaming Russian Bots For Everything. *Buzzfeed News*, 28 February 2018. https://www.buzzfeednews.com/article/miriamelder/stop-blaming-russian-bots-for-everything
- Ben Nimmo (2020). The Breakout Scale: Measuring the impact of influence operations. Brookings Institution. https://www.brookings.edu/articles/the-breakout-scale-measuring-the-impact-of-influence-operations/
- Office of the Director of National Intelligence (2017). Intelligence Community Assessment: Assessing Russian Activities and Intentions in Recent US Elections. 6 January 2017. https://www.dni.gov/files/documents/ICA_2017_01.pdf
- Stanford Internet Observatory (2019-2022). Platform takedown reports (various). https://cyber.fsi.stanford.edu/io/research/takedowns
- Thomas Rid (2020). *Active Measures: The Secret History of Disinformation and Political Warfare*. Macmillan.
- Gavin Wilde (2023). It's time to focus on information warfare's hard questions. *CyberScoop*, 5 January 2023. https://cyberscoop.com/russia-information-operations-facebook/
- Kamya Yadav, Martin Riedl, Alicia Wanless, & Samuel Woolley (2023). What Makes an Influence Operation Malign? Carnegie Endowment for International Peace, Partnership for Countering Influence Operations. https://carnegieendowment.org/2023/08/07/what-makes-influence-operation-malign-pub-90323

W11:// Beyond leave-up/take-down

4/9/2024

If you believe that content moderation is unjustifiable censorship, what do alternative approaches to governance look like (assuming you don't think the internet should just devolve into a complete abject hellscape)? We examine technologies and policies meant to move beyond the "leave-up/take-down binary" (as Evelyn Douek has described it), and consider how counterspeech, behavioral nudges, algorithmic friction, and more can help manage the harms of online speech without relying on content takedowns. Or can they?

Read

- ✷ Adrian Chen (2015). Unfollow: Conversion via Twitter. *The New Yorker*, 15 November 2015. https://www.newyorker.com/magazine/2015/11/23/conversion-via-twitter-westboro-baptist-church-megan-phelps-roper

- Renee DiResta (2018). Free Speech is not the Same as Free Reach. *Wired*, 30 August 2018. https://www.wired.com/story/free-speech-is-not-the-same-as-free-reach/
- Matthew Katsaros, Kathy Yang, & Lauren Fratamico (2022). Reconsidering Tweets: Intervening during Tweet Creation Decreases Offensive Content. Proceedings of the International AAAI Conference on Web and Social Media 16(1). https://doi.org/10.1609/icwsm.v16i1.19308
- ✷ Daniel Robert Thomas & Laila Wahedi (2023). Disrupting hate: The effect of deplatforming hate organizations on their online audience. *PNAS* 120(24). https://doi.org/10.1073/pnas.2214080120

Watch

-

Optional

- Jennifer Allen, Cameron Martel, & David Rand (2022). Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in Twitter's Birdwatch crowdsourced fact-checking program. Proceedings of the CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3491102.3502040
- Cody Buntain, Martin Innes, Tamar Mitts, & Jacob Shapiro (2023). Cross-Platform Reactions to the Post-January 6 Deplatforming. *Journal of Quantitative Description: Digital Media* 1. https://doi.org/10.51685/jqd.2023.004
- Evelyn Douek (2021). More Content Moderation Is Not Always Better. *Wired*, 2 June 2021. https://www.wired.com/story/more-content-moderation-not-always-better/
- Tarleton Gillespie (2022). Do Not Recommend? Reduction as a Form of Content Moderation. *Social Media + Society* 8(3). https://doi.org/10.1177/20563051221117552
- Eric Goldman (2021). Content Moderation Remedies. *Michigan Technology L. Rev.* 28(1). https://repository.law.umich.edu/mtlr/vol28/iss1/2/
- Shagun Jhaver, Christian Boylston, Diyi Yang, & Amy Bruckman (2021). Evaluating the Effectiveness of Deplatforming as a Moderation Strategy on Twitter. Proceedings of the ACM on Human-Computer Interaction 5, CSCW2. https://doi.org/10.1145/3479525

W12:// Saving us with/from AI

4/16/2024

Generative AI is having a real *moment* in the tech industry... but what are the actual impacts of these technologies on communication, connectivity, and safety? Even as the proponents of these technologies (most of whom work at the companies profiting from their development) argue that generative AI can at once transform the world for the better *and* represents a potential

	existential risk, we already have to contend with the very real, very immediate harms AI produces. We examine AI's promise for moderation, its perils for the internet (and humanity writ large), and how much we still don't know about this burgeoning class of technologies.
Read	• ✷ Emily Bender, Timnit Gebru, Angelina McMillan-Major, & "Shmargaret Shmitchell" (2021). On the dangers of stochastic parrots: Can language models be too big?. FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. https://doi.org/10.1145/3442188.3445922 • Matt Burgess (2023). The hacking of ChatGPT is just getting started. *Wired*, 13 April 2023. https://www.wired.com/story/chatgpt-jailbreak-generative-ai-hacking/ • Nafia Chowdhury (2022). Automated Content Moderation: A Primer. Stanford Cyber Policy Center Program on Platform Regulation. https://cyber.fsi.stanford.edu/news/automated-content-moderation-primer • Brian Christian (2020). The Alignment Problem: Machine Learning and Human Values. W. W. Norton. *(Chapter 1 only)*
Watch	•
Optional	• Dario Amodei, Chris Olah, Jacob Steinhardt et al (2016). Concrete Problems in AI Safety. https://doi.org/10.48550/arXiv.1606.06565 • Thiago Dias Oliva, Dennys Marcelo Antonialli, & Alessandra Gomes (2021). Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online. *Sexuality & Culture* 25(2). https://link.springer.com/article/10.1007/s12119-020-09790-w • Vinodkumar Prabhakaran, Margaret Mitchell, Timnit Gebru, & Iason Gabriel (2022). A Human Rights-Based Approach to Responsible AI. 2022 ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization. https://arxiv.org/abs/2210.02667 • Spandana Singh (2019). Everything in Moderation: An Analysis of How Internet Platforms Are Using Artificial Intelligence to Moderate User-Generated Content. New America Open Technology Institute. https://www.newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-user-generated-content/

W13:// Moderation breaks down

4/23/2024

Is content moderation a doomed proposition? After 15+ years of commercial content moderation across platforms of every shape and size, we're still

	wrestling with many of the same questions and concerns — about fairness, censorship, legitimacy, bias, and the harms that persist despite (or perhaps because of) moderation. We examine moderation's failure conditions, and how internet governance struggles to contend with the constantly shifting terrain of online communication.
Read	• Julia Angwin & Hannes Grassegger (2017). Facebook's Secret Censorship Rules Protect White Men From Hate Speech But Not Black Children. *ProPublica*, 28 June 2017. https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms • ✷ Evelyn Douek (2020). The Rise of Content Cartels. Knight First Amendment Institute at Columbia University. https://knightcolumbia.org/content/the-rise-of-content-cartels • Mike Masnick (2019). Masnick's Impossibility Theorem: Content Moderation At Scale Is Impossible To Do Well. *Techdirt*, 20 November 2019. https://www.techdirt.com/2019/11/20/masnicks-impossibility-theorem-content-moderation-scale-is-impossible-to-do-well/ • Yoel Roth (2023). Content Moderation's Legalism Problem. *Lawfare*, 24 July 2023. https://www.lawfaremedia.org/article/content-moderation-s-legalism-problem
Watch	• Cambridge Disinformation Summit (2023). Platform accountability vs free speech rights (panel discussion). (video)
Optional	• Ysabel Gerrard (2018). Beyond the hashtag: Circumventing content moderation on social media. *New Media & Society* 20(12). https://doi.org/10.1177/1461444818776611 • Oliver Haimson, Daniel Delmonaco, Peipei Nie, & Andrea Wegner (2021). Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas. Proceedings of the ACM on Human-Computer Interaction 5, CSCW2. https://doi.org/10.1145/3479610 • Yoel Roth (2023). Trump Attacked Me. Then Musk Did. It Wasn't an Accident. *New York Times*, 19 September 2023. https://www.nytimes.com/2023/09/18/opinion/trump-elon-musk-twitter.html

W14:// Futures

4/30/2024

Where do we go from here? At what appears to be the twilight of the great Web 2.0 platforms, a new crop of online services has emerged — with their own ideas about how governance and safety should work. And, in

governments around the world, a new crop of regulations threaten to reshape the internet as we know it, casting aside the freewheeling legacy of Section 230 in the United States. What comes next for trust and safety?

Read

Regulatory futures

- Anu Bradford (2023). After the Fall of the American Digital Empire. Knight First Amendment Institute at Columbia University. https://knightcolumbia.org/content/after-the-fall-of-the-american-digital-empire

Platform futures

- ✷ Mike Masnick (2019). Protocols, Not Platforms: A Technological Approach to Free Speech. Knight First Amendment Institute at Columbia University. https://knightcolumbia.org/content/protocols-not-platforms-a-technological-approach-to-free-speech

Governance futures

- ✷ Daphne Keller (2022). Lawful but Awful? Control over Legal Speech by Platforms, Governments, and Internet Users. University of Chicago Law Review Blog. https://lawreviewblog.uchicago.edu/2022/06/28/keller-control-over-speech/
- Aviv Ovadya (2021). Towards Platform Democracy: Policy Beyond Corporate CEOs and Partisan Pressure. Harvard Kennedy School Belfer Center for Science and International Affairs. https://www.belfercenter.org/publication/towards-platform-democracy-policymaking-beyond-corporate-ceos-and-partisan-pressure

Watch

Optional

- Daphne Keller (2021). The Future of Platform Power: Making Middleware Work. *Journal of Democracy* 32(3). https://www.journalofdemocracy.org/articles/the-future-of-platform-power-making-middleware-work/
- Kate Klonick (2021). Inside the Making of Facebook's Supreme Court. *The New Yorker*, 12 February 2021. https://www.newyorker.com/tech/annals-of-technology/inside-the-making-of-facebooks-supreme-court
- Mike Masnick (2022). Everyone's Mad At Cloudflare; Is There Room for Principled Takes on Moderation? *Techdirt*, 6 September 2022. https://www.techdirt.com/2022/09/06/everyones-mad-at-cloudflare-is-there-room-for-principled-takes-on-moderation-when-everyone-wants-what-feels-right-to-them/

- Matthew Prince & Alissa Starzak (2022). Cloudflare's abuse policies & approach. Cloudflare blog, 31 August 2022. https://blog.cloudflare.com/cloudflares-abuse-policies-and-approach/
- Ethan Zuckerman (2020). The Case for Digital Public Infrastructure. Knight First Amendment Institute at Columbia University. https://knightcolumbia.org/content/the-case-for-digital-public-infrastructure
- Ethan Zuckerman & Chand Rajendra-Nicolucci (2023). From Community Governance to Customer Service and Back Again: Re-Examining Pre-Web Models of Online Governance to Address Platforms' Crisis of Legitimacy. *Social Media + Society* 9(3). https://doi.org/10.1177/20563051231196864