

COMM 4010: Trust & Safety in Online Platforms

Cornell University

Department of Communication

Instructor: danah boyd

Course Description

“Trust & Safety” (T&S) is the practice of ensuring that a particular online platform, AI system, or area of the open web is an inviting space where participants can feel comfortable understanding the norms, engaging with the content, and interacting constructively with others. Most technology companies (and many other organizations) have “Trust & Safety” teams to create and manage Trust & Safety policies and practices. These teams support the development of technology and leverage technology to conduct their own work. This work ranges from content moderation to legal compliance to investigations. Although this work began haphazardly, T&S has grown into a wide-reaching discipline with professional associations, conferences, and shared resources.

This course will introduce students to the work of Trust & Safety in online platforms. Students will learn what happens behind the scenes of their favorite platforms while engaging in exercises designed to struggle with the challenges that T&S professionals face on a daily basis. This work forces T&S professionals to reckon with global norms and a range of ethical conundrums. This practice is also continuously evolving alongside the development of new technologies (e.g., AI) and in light of new political, legal, and global pressures.

Learning outcomes. By the end of this course, you should be able to:

1. Map the range of work involved in Trust & Safety.
2. Identify different organizational approaches to T&S.
3. Describe how different theories about T&S connect to the practice itself.
4. Articulate different ethical trade-offs underlying key T&S policies.

You should also expect to come away from this class with a deep appreciation for the moral, ethical, and political challenges underpinning this profession, regardless of whether or not you intend to do work in this area.

Content Warning: Students should be prepared to grapple with topics that are quite upsetting, such as pornography, commercial sexual exploitation of minors, terrorism, rape, hate speech, bigotry, and gore. While students will not be exposed to images of this content, they will be asked to discuss these matters and consider views that challenge their values and commitments.

Attendance:

This is an interactive seminar structured heavily around case studies and virtual visits from practitioners. You are expected to contribute to every class. As such, attendance is compulsory. Please be in your seats and ready to learn **by 11:41** so that class can start on time for the benefit of all (especially our visitors). Arriving late or leaving class early disrupts your peers’ learning

opportunities and is rude to our guests. If getting to class from your previous class is going to be a challenge, please let Professor boyd know this at the beginning of the semester so that we can plan with you accordingly.

This is a pre-professional class. You are to act like this class is your job. To that end, you are expected to provide advance notice (at least 1 hour before) to Professor boyd if you are going to be absent or late to class. If you inform them at least one hour in advance, you will be given the opportunity to make up any missed activities or work. You are allowed 4 *excused* absences during the semester without penalty.

All unexcused absences - and excused absences exceeding 4 - will result in the loss of one point (1%) of your overall grade. If you are more than 6 minutes late to class, this will be recorded. Unless excused, you will lose one point (1%) of your overall grade for every 3 latenesses (and one for every 2 latenesses that exceed 10 minutes). Excess absences and latenesses also affect your participation grade if you do not make up the work.

(Note: it is quite common for students to get sick at the end of the semester or for travel issues during spring break to result in unexpected absences. Keep this in mind if you miss classes earlier in the semester.)

Communication:

I cannot state this loud enough: Up-front communication is both essential AND your responsibility. You are expected to communicate with us in advance of any absences. You are expected to seek clarification to any assignments. If you are struggling, you are expected to reach out via email, ask for an appointment, or come to office hours. If you need accommodations, you are expected to reach out to us for a plan.

This is a pre-professional class. In a professional context, you would be fired for failing to communicate with your boss, failing to do your work without explanation, or failing to show up to class without giving advance notice. Life happens. But the key to being given opportunities to adjust and still succeed is communication.

When it comes to grading, this syllabus serves as a contract between us. If you have concerns about the grading structure laid out in this syllabus, it is your responsibility to reach out to the team before February 1st or at least 3 days in advance of any assignment's due date. If you choose to skip an assignment or turn it in late, your grade will reflect that.

If an unexpected issue arises mid semester that will impact your ability to perform in this class (e.g., illness/injury, death in the family), it is your responsibility to reach out immediately and let us know. We will work with you if we know upfront. We will not accept retroactive explanations for absences or as justifications for performance issues. Taking too many classes is not an acceptable explanation for failure to do the work in this class.

Do not expect to be given last minute opportunities to repair your grade at the end of the semester. If you are concerned that things might get in the way of your performance later on in the semester, take the early additional credit opportunities.

Accommodations:

If you are a student with a disability or who qualifies for academic accommodations, Professor Boyd are happy to work with you. We have provided basic information about the assignments and expectations in this document. Please make sure that you scan the expectations now. Should you need any accommodations for the readings, assignments, or activities in this class, please contact me as soon as possible (or by February 2nd at the latest).

Should a significant change occur over the course of the semester, please contact me as soon as possible. I will not consider alternative accommodations after the details of the assignment have been handed out and discussed in class unless there is an acute medical crisis.

If you are struggling with mental health, make sure to reach out to [Cornell Health](#). If issues arise that are affecting your performance across all classes, you should immediately contact [Student Support and Advocacy Services](#). They will contact me if adjustments are needed.

Given the intensity of the topics that we will be discussing, students should also approach me in advance if they are concerned about the mental health impact of discussing a particular topic on the syllabus. I will do my best to accommodate students if there is one topic that is especially challenging. If multiple topics present an emotional challenge, this might not be the best class for you.

Statement of Inclusivity and Respect

Each student in this course is expected to contribute to an inclusive and respectful class environment. Students of all backgrounds including gender, sexual orientation, race, ethnicity, caste, nationality, and religion are to be treated fairly and with honesty, integrity, and respect. Civil discourse, reasoned thought, sustained discussion, and constructive engagement without degrading, abusing, harassing, or silencing others is required of all students in this class.

The topics that we will discuss are intense and the specific incidents are often disturbing. This can trigger all sorts of emotional reactions from students (as it does among professionals). Please be prepared to be especially empathetic towards your peers as each person's reaction can be quite different. And if you are struggling with your own reactions to a discussion, you will be encouraged to call a time-out. We will discuss other processes to ensure that everyone is safe on the first day.

We will have practitioners visit this class at different points. Please treat them with the same level of respect as you would afford your peers.

In-Class Device Policy

Technology can support your educational development; it can also be a distraction, both to you and others in the classroom. In this class, you will need to have your devices with you. However, it is your responsibility to avoid using technology in ways that rupture the academic purpose for which we gather. You are responsible for your behavior and choices – and the impact that they have on those around you. If your use of technology distracts your peers or results in your disengagement from the classroom, you will be treated as being absent from class.

Zoom is a requirement for this class. When we have guest visitors, you will need to have a device with you (or plan to share a computer with a classmate). You will be expected to dial into Zoom from the classroom with your video on (and your audio off and your computer muted). This allows our speakers to see their audience. This will allow you to converse with your classmates in the chat feature. Please do so respectfully and on topic. Using Zoom (or a similar platform) in this way is a common practice in industry and we are going to replicate it. If you have concerns about this norm, please contact us immediately.

Generative AI Policy

Generative AI is a powerful tool that can enhance your creativity, but it also comes with trade-offs. You may use Generative AI for your class work, but you must be transparent about your use. For every assignment, you will be asked to fill out a form concerning your GenAI use - even if you don't use it!

GenAI use comes in many forms. Students use it to synthesize their notes, summarize literature, support their research, produce images, evaluate their papers for grammar/spelling/flow, and much more. You must be clear about all of your uses. This is not just a matter of whether or not GenAI is used to generate text.

GenAI has many problems. It hallucinates references. It bullshits summaries. It produces text that is flat and is often disconnected from the material at hand. When students use GenAI in these ways, the quality is often lacking in ways that cost students significant points. Keep this in mind.

In this statement, you should reflect on how GenAI helped your learning journey, how you ensured the quality of your work, and thoughts you might have on the ethics of using these tools for your work.

Failure to be transparent and reflexive about your use of GenAI – even for doing research for the assignment – will be treated as plagiarism. It is your responsibility to ensure that you've read the materials you've cited and can defend the work you are submitting.

Although this class does not prohibit the use of Generative AI, you are still responsible for every claim and citation in your work, whether you use advanced tools or not. For example, you are expected to fully read any reference/citation you include in your work. You should be prepared to answer questions about or turn over your notes about any reference you include in your materials. A single hallucinated citation from Generative AI will result in a zero for the assignment. So check your work and make sure you can take responsibility for every component

of it. Academic integrity means that you are actually reading and thinking about the texts, not just citing texts based on their title.

Academic Integrity

Each student in this course is expected to hold each other to the highest standards of academic integrity and abide by the Cornell University Code of Academic Integrity. Any work submitted by a student in this course for academic credit will be the student's own work. For collaborative projects, students may be asked to identify who contributed what to the effort. If you have any doubts about what constitutes a violation of the Code of Academic Integrity, or any other issue related to academic integrity, please ask us.

Students agree that by taking this course, all required papers may be subject to submission for textual similarity review to Turnitin.com or a similar service for the detection of plagiarism. All submitted papers will be included as source documents in the Turnitin.com reference database solely for the purpose of detecting plagiarism of such papers. Use of the Turnitin.com service is subject to the Usage Policy posted on the Turnitin.com site.

Any student in this class suspected of plagiarism or cheating on tests or assignments will have a primary hearing. If found guilty, the student will receive a failing grade on the assignment in question and a reduction in the overall course grade, and the results of the hearing will be reported to the CALS Academic Integrity Board. The Department of Communication's policy is based on university-wide policy and procedures. Visit the University Academic Integrity website for further information.

Grading:

In-Class Participation: 20% (20 points)
In-Advance Visitor Questions: 8% (8 points)
Reflection Memos: 20% (20 points)
Knowledge Checks: 12% (12 points)
Case Study Exercise: 20% (20 points)
Final Project: 20% (20 points)

Grading Scale

A+ = 98-100	B+ = 87-89.5 ^[L] _[SEP]	C+ = 77-79.5 ^[L] _[SEP]	D+ = 67-69.5 ^[L] _[SEP]	F = <60
A = 94-97.5 ^[L] _[SEP]	B = 84-86.5 ^[L] _[SEP]	C = 74-76.5	D = 64-66.5	
A- = 90-93.5 ^[L] _[SEP]	B- = 80-83.5	C- = 70-73.5	D- = 60-63.5 ^[P] _[SEP]	

General Note on Assignments and Grading

Keep in mind that there will be many moments in this class where there is no "right" answer. You are going to be asked to debate issues from different vantage points, including taking debate

stances that you might not agree with. You will be graded on your ability to articulate different vantage points in a thoughtful and coherent fashion.

You will receive a grade, along with an explanation for why points were deducted and brief feedback for each assignment. In general, we will grade assignments towards the following rubric:

- F: failure to complete the assignment
- D: assignment is missing significant parts, shows lack of core ideas/concepts, is difficult to read, or is otherwise deeply flawed
- C: assignment is substantially complete, but may miss or mistake important points; writing and argumentation may be unclear
- B: assignment is complete and solid in all respects; covers major relevant points; writing and argumentation are sound; demonstrates understanding of core concepts
- A: assignment is excellent in all respects; covers the material thoroughly and effectively; demonstrates significant mastery of core concepts; advances a clear and convincing line of argument

If you have a concern about a particular grade, please start by sending an email to Professor boyd with a clear articulation of why, in light of this rubric and the articulation of deducted points, you feel as though the grade is inaccurate.

If you choose to work in teams, you will all receive the same grade unless concerns about the collaboration are raised to the teaching team in advance of submitting the assignment.

Late Assignment Policies:

For every day that an assignment is late, you will receive a deduction of 5% of the grade of that assignment. If an issue related to the deadline arises, you must contact Professor boyd at least 72 hours in advance to get permission for an extension. After that, a doctor's note or a note from student services will be necessary.

Please note that extensions will be rarely given. You will have plenty of advanced notice about the assignments. I encourage you to plan accordingly.

Opportunities for Extra Credit:

There will be occasional opportunities for extra credit throughout the semester, linked to attending external talks or doing extra project work. One of those opportunities is listed below. The maximum semester-length extra credit you can obtain through these opportunities is 5%.

Assignments and Graded Activities

In-Class Participation (20%). You are expected to participate during class in various ways.

- *In-class Activities (10%).* There will be in-class activities for which you will work in groups and document your work on stickies, worksheets, and index cards. You can expect to turn in the outputs from these activities. At a baseline, you (and your teams) will need

to have shown substantive engagement with at least 80% of the semester's activities. You will be docked points for not taking the activities seriously or being distracted by your devices/peers.

- *Contributing to Class Conversation (5%)*: You are expected to raise your hand in response to the professor's questions and offer your insights and ideas. As a baseline, you are expected to offer a contribution in at least 10 different non-visitor classes to receive full points. (If speaking up in class is difficult for you, you should send a brief written reflection to the class conversation via email within 4 hours of class as an alternative to raising your hand.)
- *Contributing to Visitor Visits (5%)*: You are expected to be present, have your camera on, be attentive, and contribute to the conversation in the chat and through questions. At a minimum, you must ask one verbal question of one speaker and contribute meaningfully to the visits through substantive engagement - by asking questions verbally, in chat, or by contributing to discussions in chat threads - of at least 4 other visitors. (Points will be docked for being visually disengaged or distracted.)
- To ensure that you get full points for participation (and leverage these opportunities), I strongly encourage you to go above and beyond these minimums. Show me that you are meaningfully engaged with the material, thinking about it, asking questions, and grappling with the issues at play. You will be given feedback mid-semester about your class participation.

In-Advance Visitor Questions (8%). We expect to have 9 practitioners visit our class via Zoom to share their knowledge and experiences. They will also provide a pre-"read" (which might be a video or podcast). In the weeks that those visitors are coming, students will be expected to submit one thoughtful question each to Canvas by **noon** on the MonClass before the visitor arrives so that those questions can be shared with the visitor. The questions should be incisive and reflective of the material we are grappling with in this class. They should also be personalized to the person/topic (so do a little bit of research about the guest in advance). You should treat these questions with the seriousness of a question you would ask in a job interview. Your question should reflect that you read the related material and are bringing insights to bear on the issue at hand.

- If you use GenAI to help you prepare this question, don't forget to add a brief GenAI statement.
- Grading. Questions submitted for each visitor will be worth one point (1%) of your total grade. You will receive feedback (and be docked points) if your question is not satisfactory.
- You may miss one week's visitor question without penalty. (If we do not have 9 visitors in the end, you will be given freebie points.)
- Example high quality questions: "In the pre-read you gave us, the author said XYZ but this contradicts/complicates <insert different thing you read>. Can you explain why these things differ?" or "Your organization encountered this gnarly situation <insert public example>. Can you walk us through what your thinking was when you navigated trade-offs in that situation?"

Reflection Memos (20%). Students will be asked to submit *five* ~500 word reflection memos covering at least one of the upcoming week's more academic-oriented readings. The purpose of

these memos is to show how you are *thinking* about the readings. You can focus on one reading or leverage multiple for that week. Eligible readings are double starred; they include both required and optional readings (**). Your memos are not supposed to be summaries. Rather, you are to explore the frictions, controversies, and struggles that are presented in the readings and offer new insights and questions. You could apply the reading to a different example, struggle with/debate an argument in the text(s), or discuss how multiple readings speak to one another around the same issue.

- Reflection memos for upcoming readings are due SunClass night by 11:58PM.
- You choose which five weeks to reflect on, but you must do at least one in the first four weeks of class and at least two before Spring Break. I strongly encourage you to do them earlier in the semester.
- Make sure to submit a brief GenAI statement at the bottom of your memo (or indicate that you did not use GenAI). See [an example here](#).
- Grading. Each memo will be graded on a 4-point scale. To receive all four points, your memo must be well-written, thoughtfully engaging of the material for that week, and connecting the ideas from the material to prior weeks' discussions, current events, or notable case studies.
- Correction Option. You may submit more than 5 reflection memos. Your grade will reflect the scores of the top five. You may not retroactively submit reflection memos (after the week's deadline or at the end of the semester) so if you want to ensure the highest grade possible, do these early on.

Two Knowledge Checks (12%). There will be two pre-scheduled short form in-class closed-book knowledge checks to ensure that you've mastered key concepts and information in the class. Each will be worth 6%. In both cases, a set of potential questions will be provided in advance for you to prepare.

[Case Study Exercise \(20%\).](#) Imagine that you are preparing an incident report for an oversight body. Detail the incident, what policies you believe that it violates, and what your recommendations are for remedy. Provide both a "pro" and "con" analysis of the incident. Point to relevant literature. You may fabricate an incident or select from a case you know or read about. You may choose from any social media platform and policy. You should assume that your audience does not know the policies of that platform so spell it out in articulating what the violation is. [[Full assignment details](#)]

- For this exercise, you are strongly encouraged to work in pairs (but are not required to).
- You will be asked to present the case study on one of two days. (On the other day, you will be the audience/judge for the presenters.) You will be expected to produce a 1-page (2-sided) handout for this assignment, a slide, and a brief speech. You will also be prepared to answer questions.
- Grading. This assignment is worth 20 points. You will be graded based on the quality of your submitted materials and your presentation (substantively and as acts of communication), based on strong arguments that manage divergent perspectives, and based on grounding your analysis through literature/examples. To receive full points, you should offer creative remedies that are thoughtful about the implications for precedent and impact.

- You will need to submit a full GenAI statement alongside your case study. (See forms in Canvas.)
- Due: March 24. (Poster sessions March 24/26.)

Build-Your-Own Policy Project (20%). In your final assignment, you are asked to integrate the knowledge in this class and apply it to a task that is commonplace in the T&S world. Working alone or in teams of up to four people, you will be in charge of designing a T&S policy for a fabricated organization. Depending on how many people are on your team will determine the breadth of what your policies must cover. (For example, a single person working alone will be in charge of only one core topical area while a team of four will split responsibilities for covering four areas.) You will need to create a proper “playbook” for your organization, complete with examples. You will also need to provide a report. Your choices should be properly footnoted with references that indicate why you are making the choices you are making. (You will be expected to incorporate existing literature into this assignment.)

- Check-in. By April 16th, please send an email to Professor boyd describing who you are working with and what policy areas your team is intending to cover. This check-in is an important step to make sure you are on track with this assignment.
- Grading. This project is worth 20 points. You will be graded on the quality of your argument and analysis. You will be rewarded for clearly demonstrating your understanding of the trade-offs at stake, leveraging data and citations to defend your claims, and grappling with contextual issues. (See assignment for rubric.)
- There will be a peer evaluation of your policies in-class on May 5. Final policy to be submitted by May 10.
- You will need to submit a full GenAI statement alongside your policy. (See forms in Canvas.)
- [More details about this assignment can be found on a separate page.](#)

One Extra Credit Option

While there may be other extra credit opportunities throughout the semester, there is one way to get all five extra credit points at once:

Create a video response to scholarship. Read a pre-approved social science book (or a pairing of law review articles) on the topic of Trust & Safety and produce a 5-7 minute video book response. This video should be of production quality, reflect on the core contributions of the scholarship, and use the scholarship to discuss one key facet of T&S. Since this extra credit option will take time, you are required to indicate to Professor boyd that you are interested in pursuing this by March 26, 2026. You will need to complete this response video no later than May 5, 2026. Full credit will be given for an engaging, thoughtful, well-produced video that masters the material of the scholarship and reflects your own thinking on the matter. Talk with the teaching team to learn more. (This is a great option for anyone who is interested in graduate school for research or law.)

- You will need to submit a full GenAI statement alongside your video. (See forms in Canvas.)
- *For those who were approved by 3/26, more information for this extra credit opportunity can be found [here](#).*

Useful Resources

- Trust & Safety Professional Association's Curriculum. <https://www.tspa.org/curriculum/ts-curriculum/>
- Trust & Safety Network Starter Kit. <https://www.trustsafety.net/starter-kit>
- Integrity Institute Research. <https://integrityinstitute.org/research>
- ROOST's publicly accessible T&S tools. <https://roost.tools/>

WEEKLY COURSE SCHEDULE AND READINGS

Week 1: Introduction to Trust & Safety

January 20 (Class 1): Introduction to the course and to Trust & Safety as a concept

January 22 (Class 2): A Pre-History of T&S

- Read: Dibbell, Julian (1993). A Rape In Cyberspace. Village Voice, 23 December 1993. <https://www.villagevoice.com/a-rape-in-cyberspace/>
- Encouraged: Watch/Read Transcript: Alex Macgillivray & Nicole Wong (2020). Origins of Trust and Safety with Robyn Caplan. *Data & Society*. <https://datasociety.net/library/origins-of-trust-and-safety/>

Week 2: The Foundation of a Field

January 26 by noon: Submit questions for visitor Del Harvey

January 27 (Class 3): Visitor – [Del Harvey](#) - founded/led the Twitter T&S team (2008-2021), Board Chair of TSPA

- Watch: Del Harvey (2014). Protecting Twitter users (sometimes from themselves). TED. https://www.ted.com/talks/del_harvey_protecting_twitter_users_sometimes_from_themselves
- Prepare questions by reading some of Del's writing: <https://delbius.com/>

January 29 (Class 4): How to do Trust & Safety

- Read: Trust & Safety Professional Association's Curriculum. <https://www.tspa.org/curriculum/ts-curriculum/>
- Bring to class:
 - Questions about the TSPA curriculum!
 - An example T&S case that interests you.
 - You can get some case studies at [T&S Foundation](#).
 - Other good famous cases: Cambridge Analytica, Facebook/Rohingya/Myanmar, Covid-19 Disinformation, Christchurch, Syrian civil war

- Cases from [FB Oversight Board](#).
 - A different case of your choosing!
- ** Optional: Read: Citron, Danielle Keats, and Waldman, Ari Ezra. 2025. "The Evolution of Trust and Safety." Preprint, SSRN /, September 15. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5401604.
- ** Optional: Read: Shulruff, T. (2024). Trust and Safety work: Internal governance of technology risks and harms. *Journal of Integrated Global STEM*, 1(2), 95–105. <https://doi.org/10.1515/jigs-2024-0003>

Week 3: Content Moderation

February 2 by noon: Submit questions for visitor Will Partin

February 3 (Class 5): Content Moderation, an Overview

- Play (for at least 20 minutes): Moderator Mayhem - <https://moderatormayhem.engine.is/>
- Read: Caplan, Robyn (2018). Content or context moderation? *Data & Society*. <https://datasociety.net/library/content-or-context-moderation/>
- ** Optional: Read: Gillespie, Tarleton (2010). The politics of platforms. *New Media & Society* 12(3). <https://doi.org/10.1177/1461444809342738>
- ** Optional: Read: Grimmelmann, James. (2015). "The Virtues of Moderation." *Yale Journal of Law & Technology* 17 (42). <https://scholarship.law.cornell.edu/facpub/1486>
- ** Optional: Read: Gorwa, Robert, Reuben Binns, and Christian Katzenbach. 2020. "Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance." *Big Data & Society* 7 (1): 205395171989794. <https://doi.org/10.1177/2053951719897945>.

February 4 by 11:58PM: Fill out [your case study](#), answering at least 3 of the questions from last week's class.

February 5 (Class 6): Visitor - [Will Partin](#) - YouTube and expert on gaming/Twitch (and Communication PhD from UNC!)

- He will focus on the everyday practices of content moderation. His expertise includes disinformation, hate speech, and gaming/livestreaming. Feel free to scan any of [his academic papers](#).
- Watch (20 mins): <https://fieldofvision.org/shorts/the-moderators>
 - If it doesn't load there, it is also available [on YouTube](#)
- Read: Chen, A. (2014, October 23). "The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed." *Wired*. <https://www.wired.com/2014/10/content-moderation/>
 - A PDF of this article is available in Canvas under "Required Readings Otherwise Unavailable"
- ** Optional Read: douek, everlyn. (2022). "Content Moderation as Systems Thinking." *Harvard Law Review*, 136. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4005326

Week 4: Governance

February 9 by noon: Submit questions for visitor Manuel Parra Yagnam

February 10 (Class 7): Governing Content

- Read: Klonick, Kate (2018). The New Governors: The People, Rules, and Processes Governing Online Speech. *Harvard Law Review* 131(6).
<https://harvardlawreview.org/print/vol-131/the-new-governors-the-people-rules-and-processes-governing-online-speech/>
- ** Optional: Read: Kosseff, Jeff (2022). A User's Guide to Section 230, and a Legislator's Guide to Amending It (or Not). *Berkeley Technology Law Journal* 37(2): 757. <https://doi.org/10.15779/Z38VT1GQ97>
- ** Optional: Read: Ananny, Mike. 2019. Probably Speech, Maybe Free: Toward a Probabilistic Understanding of Online Expression and Platform Governance.
<https://doi.org/10.7916/D8-AM3S-3S09>
- ** Optional: Read: Caplan, Robyn, and Tarleton Gillespie. 2020. "Tiered Governance and Demonetization: The Shifting Terms of Labor and Compensation in the Platform Economy." *Social Media + Society* 6 (2): 2056305120936636.
<https://doi.org/10.1177/2056305120936636>.
- Optional: Check out The Platform Governance Archive:
<https://www.platformgovernancearchive.org/>

February 12 (Class 8): Visitor - [Manuel Parra Yagnam](#) - Facebook Oversight Board staff

- Review: <https://www.oversightboard.com/>
- Read some of Manuel's thoughts on tech governance from his blog: <https://parra-yagnam.com/>
- Read: Klonick, Kate. 2021. "Inside the Making of Facebook's Supreme Court." *The New Yorker*, February 12. <https://www.newyorker.com/tech/annals-of-technology/inside-the-making-of-facebooks-supreme-court>.

Week 5: Disinformation

February 17: No Class

Memos due Sunday night can be delivered by Tuesday night and count.

February 19 (Class 9): Disinformation

- Special Guest: Claire Wardle (from COMM)
- Read: Saltz, Emily, Soubhik Barari, Claire Leibowicz, and Claire Wardle. 2021. "Misinformation Interventions Are Common, Divisive, and Poorly Understood." *Harvard Kennedy School Misinformation Review*, October 27.
<https://doi.org/10.37016/mr-2020-81>.
- ** Optional: Read Camille François's report on "Actors, Behaviors, Content: A Disinformation ABC." https://www.annenbergpublicpolicycenter.org/wp-content/uploads/ABC_Framework_TWG_Francois_Sept_2019.pdf

- ** Optional: Read: Gillespie, Tarleton. 2022. “Do Not Recommend? Reduction as a Form of Content Moderation.” *Social Media + Society* 8 (3): 20563051221117552. <https://doi.org/10.1177/20563051221117552>.
- ** Optional: Read: Marwick, Alice, and Rebecca Lewis. 2017. *Media Manipulation and Disinformation Online*. Data & Society. <https://datasociety.net/library/media-manipulation-and-disinfo-online/>.

Week 6: Children’s Online Safety

February 24 (Class 10): Contact, Content, and Conduct Related Harms. (Bullying, Harassment, Pedophiles, etc.).

- ** Scan-Read: Internet Safety Technical Task Force. 2008. Enhancing Child Safety & Online Technologies: Final Report. <https://cyber.harvard.edu/pubrelease/isttf/> -> Prioritize the literature review!
- ** Scan-Read: Committee on the Impact of Social Media on Adolescent Health, Board on Population Health and Public Health Practice, Health and Medicine Division, and National Academies of Sciences, Engineering, and Medicine. 2024. *Social Media and Adolescent Health*. Edited by Sandro Galea, Gillian J. Buckley, and Alexis Wojtowicz. National Academies Press. <https://doi.org/10.17226/27396>.
- *Note: These two reports are VERY long. Read through them strategically, looking for the substance that can anchor our conversations.*
- ** Optional: Choose one of the relevant chapters from danah’s *It’s Complicated: The Social Lives of Networked Teens*. (Danger and Bullying are particularly relevant.)

February 26 (Class 11): Legal Approaches to Protecting Children from the Internet

- ** Read: Angel, María P., and danah boyd (2024). “Techno-Legal Solutionism: Regulating Children’s Online Safety in the United States.” In *Proceedings of the Symposium on Computer Science and Law*, 86–97. Boston MA USA: ACM. <https://doi.org/10.1145/3614407.3643705>.
- ** Scan-Read: Marwick, A., Smith, J., Caplan, R., & Wadhawan, M. (2024). *Child Online Safety Legislation (COSL)—A Primer*. The Bulletin of Technology & Public Life. <https://doi.org/10.21428/bfcb0bff.de78f444>

Week 7: Sexual Exploitation

March 2 by noon: Submit questions for visitor Alex Leavitt

March 3 (Class 12): Visitor - [Alex Leavitt](#) - T&S research for Roblox; former Meta (PhD in Communication from USC)

- Check out their publications and commentary: <https://alexleavitt.net/>
- Compare <https://www.theguardian.com/technology/2021/sep/14/facebook-aware-instagram-harmful-effect-teenage-girls-leak-reveals> vs. <https://about.fb.com/news/2021/09/research-teen-well-being-and-instagram/> (also, explore the 2 "annotated research decks" linked in the FB post)
- Family Safety - <https://about.roblox.com/safety> (click through to each sub-page, to see how it addresses child safety from the parent's perspective)

- Alex is an example of a T&S professional who is constantly tracking the different studies and reports that are coming out. Here are some of the materials that they sent in to explore. These are optional:
 - Children's Rights - <https://www.un.org/en/global-issues/child-and-youth-safety-online> and https://www.unicef.org.uk/wp-content/uploads/2019/10/UNCRC_summary-1_1.pdf
 - Online Experiences - https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/bade/documents/products-and-services/en-us/digitalsafety/2026_Global_Online_Safety_Survey_Results.pdf (spec. pages 4-5, 9-18)
 - Doctor's POV: <https://publications.aap.org/pediatrics/article/157/2/e2025075320/206129/Digital-Ecosystems-Children-and-Adolescents-Policy>
 - Social Media Bans - compare <https://www.bbc.com/news/articles/cwyp9d3ddqyo> vs. <https://thehill.com/opinion/technology/5726492-social-media-teen-mental-health/>
 - Sexual Content - https://info.thorn.org/hubfs/Research/Thorn_24_YouthMonitoring_Report.pdf (spec. pages 8, 13, 16-17, 22, 26, 31, 33, 35)

March 5 (Class 13): Children's Sexual Exploitation, CSAM, and Revenge Porn.

- ** Read: Farid, Hany. 2018. "Reining in Online Abuses." *Technology & Innovation* 19 (3): 593–99. <https://doi.org/10.21300/19.3.2018.593>.
- ** Optional: Read: Citron, D. K., & Franks, M. A. (2014). Criminalizing Revenge Porn. *Wake Forest Law Review*, 49, 345–391. https://digitalcommons.law.umaryland.edu/fac_pubs/1420/
- ** Optional: Read: Dvoskin, B., & Kadri, T. (upcoming). Safe Sex in the Age of Big Tech Feminism. *Harvard Journal of Law and Technology*. <https://doi.org/10.2139/ssrn.5170919>

Week 8: Hateful Words, Violent Acts

March 9 by noon: Submit questions for visitor Yoel Roth

March 10 (Class 14): Free Speech or Violence?

- Watch: Benesch, Susan (2018). What is Dangerous Speech? Dangerous Speech Project. <https://dangerousspeech.org/what-is-ds/>
- ** Read: Schissler, M. (2025). Beyond Hate Speech and Misinformation: Facebook and the Rohingya Genocide in Myanmar. *Journal of Genocide Research*, 27(3), 445–470. <https://doi.org/10.1080/14623528.2024.2375122>
- ** Optional: Read: Wilson, Richard Ashby, and Molly K. Land. 2021. "Hate Speech on Social Media: Content Moderation in Context." *Connecticut Law Review* 52 (3): 1029–76.
- ** Optional: Read: Peterson-Salahuddin, Chelsea. 2024. "Repairing the Harm: Toward an Algorithmic Reparations Approach to Hate Speech Content Moderation." *Big Data & Society* 11 (2): 20539517241245333. <https://doi.org/10.1177/20539517241245333>.

March 12 (Class 15): Visitor – [Yoel Roth](#) - former Twitter, current Match Group T&S (Phd in Communication from UPenn)

- Read: Roth, Yoel. 2022. “I Was the Head of Trust and Safety at Twitter. This Is What Could Become of It.” *New York Times*, November 18.
<https://www.nytimes.com/2022/11/18/opinion/twitter-yoel-roth-elon-musk.html>.
- Read through Yoel’s website for what interests you: <https://yoyoel.com/> → He has a LOT of interesting talks on his site, along with articles that might be of interest.

Week 9: National Security and International Politics

March 16 by noon: Submit questions for visitor Michael Golebiewski

March 17 (Class 16): Securitization, Terrorism, and Warfare

- First Knowledge Check (20 minutes)
- Read: Koerner, Brendan. (2016). “Why ISIS is winning the social media war.” *Wired*.
<https://www.wired.com/2016/03/isis-winning-social-media-war-heres-beat/>
- Read: Klonick, Kate. (2019). Inside the team at Facebook that dealt with the Christchurch shooting. *New Yorker*. <https://www.newyorker.com/news/news-desk/inside-the-team-at-facebook-that-dealt-with-the-christchurch-shooting>
- ** Optional: Read: Chachko, Elena. (2021). “National security by platform.” *Lawfare*.
<https://www.lawfaremedia.org/article/national-security-platform>
- ** Optional: Read: Sullivan, Gavin. (2025). Algorithmic governance of ‘terrorism’ and ‘violent extremism’ online. *London Review of International Law*, 13(1), 47-75.
<https://academic.oup.com/lril/article/13/1/47/8152419>

March 19 (Class 17): Visitor - [Michael Golebiewski](#) - Microsoft (Bing, AI)

- His expertise is about how to build teams to deal with T&S problems in search and AI, including scams, disinformation, algorithmic manipulation. His team also builds tools to improve T&S.
- As you’re preparing questions, also think more broadly about the role that T&S can and should play vis-a-vis search engines uniquely. He’s also a good one to ask about AI in various ways.
- Read: Golebiewski, Michael, and danah boyd (2019). “Data Voids: Where Missing Data Can Easily Be Exploited.” *Data & Society*. <https://datasociety.net/library/data-voids/>
- Example articles shared by Michael that go deeper on the “data voids” issue:
 - [One-third of Americans unaware ObamaCare, ACA are the same](#) (real world well documented case of "obama care" vs "affordable care act" being used to divide an audience)
 - [Online searches to evaluate misinformation can increase its perceived veracity | Nature](#)
 - [PolitiFact | How days of vote counting became go-to ‘evidence’ for false election fraud claims](#) (real time news and unknowns exploited to create voids)
 - [AI and Data Voids: How Propaganda Exploits Gaps in Online Information | Lawfare](#) (data voids and propagation into chat bots)

- [Misinformation does not reduce trust in accurate search results, but warning banners may backfire](#)

Week 10: Poster Sessions

March 24 by 8AM: Poster, slide, and handout are due on Canvas.

March 24 (Class 18): First Poster Session

March 26 (Class 19): Second Poster Session

March 26 by 5PM: Deadline to communicate your interest in Bonus Book Assignment to Prof boyd.

Optional Readings for Memos This Week:

- ** Swartz, Lana (2022). “Theorizing the 2017 Blockchain ICO Bubble as a Network Scam.” *New Media & Society* 24 (7): 1695–1713. <https://doi.org/10.1177/14614448221099224>.
- ** Guarna, Tomás. 2026. “From the Virtual Community to ‘Trust and Safety’: eBay (1995–2007) and the Rise of Platform Governance.” *Big Data & Society* 13 (1): 20539517251403973. <https://doi.org/10.1177/20539517251403973>.
- ** Beers, Andrew, Tom Wilson, and Kate Starbird. 2022. “Demographics of an International Influence Operation Affecting Facebook Users in the United States.” *Journal of Online Trust and Safety* 1 (4). <https://doi.org/10.54501/jots.v1i4.55>.

March 30-April 3: Spring Break

Week 11: The Business of T&S

March 27 by 5PM (or April 6 by noon): Submit questions for visitor Camille François

April 7 (Class 20): Tooling for T&S groups (tech and policy) and the politics of vendors.

- ** Read: Wright, Lucas. 2025. “The Salesforce of Safety: Software Vendors as Infrastructural/Professional Nodes in the Field of Online Trust and Safety.” *Platforms & Society* 2 (November): 29768624251390683. <https://doi.org/10.1177/29768624251390683>.
- ** Read: Douek, Evelyn. (2020). “The rise of content cartels.” *Knight First Amendment Institute*. <https://knightcolumbia.org/content/the-rise-of-content-cartels>
- ** Optional: Read: Francois, C., Shen, J., Roth, Y., Lai, S., & Povolny, M. (2025). Four Functional Quadrants for Trust & Safety Tools: Detection, Investigation, Review & Enforcement (DIRE). In M. L. Daniel, A. Menking, M. T. Savio, & J. Claffey (Eds.) (In Press), *Trust, Safety, and the Internet We Share: Multistakeholder Insights*. Taylor & Francis. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5369158

- ** Optional: Read: Denyer Willis, G. (2023). ‘Trust and safety’: Exchange, protection and the digital market–fortress in platform capitalism. *Socio-Economic Review*, 21(4), 1877–1895. <https://doi.org/10.1093/ser/mwad003>

April 9 (Class 21): Visitor - [Camille François](#) - led T&S for Niantic (Pokemon Go); former Google/Jigsaw; testified on Russian election interference for Senate Intelligence

- ** Optional: Read: François, Camille, and Evelyn Douek. 2021. “The Accidental Origins, Underappreciated Limits, and Enduring Promises of Platform Transparency Reporting about Information Operations.” *Journal of Online Trust and Safety* 1 (1). <https://doi.org/10.54501/jots.v1i1.17>.
- In preparing questions, feel free to scan any of her papers: <https://scholar.google.com/citations?user=76PkeF8AAAAJ&hl=en&oi=sra>
- (You also read her paper on ABCs earlier if you want to revisit that.)
- *On April 2nd, Camille and collaborators wrote a new essay relevant to our class:* <https://www.techpolicy.press/why-ai-model-cards-are-an-urgent-necessity-for-child-safety/>

Week 12: Generative AI

April 13 by noon: Submit questions for visitor Dave Willner

April 14 (Class 22): Unique GenAI challenges. Deep Fakes. IP.

- Read: Lehrer, Eli. 2025. “AI Safety Requires Pluralism, Not a Single Moral Operating System | TechPolicy.Press.” Tech Policy Press, December 10. <https://www.techpolicy.press/ai-safety-requires-pluralism-not-a-single-moral-operating-system/>.
- ** Read: Gillespie, Tarleton. 2020. “Content Moderation, AI, and the Question of Scale.” *Big Data & Society* 7 (2): 2053951720943234. <https://doi.org/10.1177/2053951720943234>.
- ** Optional: Read: Gillespie, Tarleton, Ryland Shaw, Mary L. Gray, and Jina Suh. 2026. “AI Red-Teaming Is a Sociotechnical Problem.” *Communications of the ACM* 69 (2): 88–95. <https://doi.org/10.1145/3731657>.
- Optional Read: Case study “How Meta updated its approach to direct disclosure based on user feedback” <https://partnershiponai.org/meta-framework-case-study/>

April 16 (Class 23): Visitor: [Dave Willner](#) - former OpenAI/AirBnb/Meta, now [Zentropi](#)

- Read: Willner, Dave. 2026. “AI is Removing Bottlenecks to Effective Content Moderation at Scale.” *Tech Policy Press*, January 29.
- Read: Willner, Dave, and Samidh Chakrabarti. 2024. “Using LLMs for Policy-Driven Content Classification.” *Tech Policy Press*, January 29. <https://www.techpolicy.press/using-llms-for-policy-driven-content-classification/>.
- Watch part of: <https://www.youtube.com/watch?v=XKJWUCliqpM>

April 16 by 5PM: Check-in for final project due

Week 13: (April 21/23) – The Quicksand Situation

April 20 by noon: Submit questions for visitor Renee DiResta

April 21 (Class 24): Gamergate. Revisiting Disinformation. Politicization.

- Read: Massanari, Adrienne L. *Gaming Democracy: How Silicon Valley Leveled up the Far Right*. MIT Press. Chapter 2: From #Gamergate to the Alt-Right (pp. 27-64) [Available on Canvas under Required Readings Otherwise Unavailable]
- Read: essays on #Gamergate:
<https://www.nytimes.com/interactive/2019/08/15/opinion/what-is-gamergate.html>
- ** Optional Read: Donovan, Joan, and danah boyd. 2021. “Stop the Presses? Moving From Strategic Silence to Strategic Amplification in a Networked Media Ecosystem.” *American Behavioral Scientist* 65 (2): 333–50.
<https://doi.org/10.1177/0002764219878229>.
- ** Optional Read: Napoli, Philip M. 2025. “In Pursuit of Ignorance: The Institutional Assault on Disinformation and Hate Speech Research.” *The Information Society* 41 (1): 1–17. <https://doi.org/10.1080/01972243.2024.2419105>.
- ** Optional Read: Lewis, Rebecca. 2018. *Alternative Influence: Broadcasting the Reactionary Right on YouTube*. Data & Society. https://datasociety.net/wp-content/uploads/2018/09/DS_Alternative_Influence.pdf

April 23 (Class 25): Visitor: [Renee DiResta](#) - Georgetown University; advisor to Congress, State Dept

- She has written extensively about a range of topics. Scan and read some of her essays in popular press:
 - <https://www.reneediresta.com/writing/>
 - <https://substack.com/@reneediresta>
- [“For Expertise to Matter, Nonpartisan Institutions Need New Communication Strategies”](#) is a recent report that shows some of her broader thinking about how communication hits social media.
- [“Process as Punishment”](#) is her commentary on the scrutiny by Jim Jordan and others.
- She has *lots* of videos [on YouTube](#) (and elsewhere).
- ** Optional Read: DiResta, Renee. 2024. *Invisible Rulers: The People Who Turn Lies into Reality*. PublicAffairs.

Week 14: The Future of T&S

April 28 (Class 26): Mental health challenges of doing T&S.

- Read: Theodora Skeadas, Jenn Louie, Ryan Roberts, 2026. [“Examining Moral Injury in the Trust and Safety Profession.”](#) TechPolicy.press. March 30.
- ** Optional Read: Gauthier, Gabrielle M, Eesha Ali, Amna Asim, Sarah Cornell-Maier, and Lori A Zoellner. 2025. “I’ve Seen Enough: Measuring the Toll of Content Moderation on Mental Health.” Preprint, arXiv, November 12.
<https://arxiv.org/abs/2511.09813>.

April 30 (Class 27): Legal evolution. Bureaucratization. The future

- Read: Newton, Casey (2025). [Is anyone left to defend trust and safety?](#) AND [Trust and safety workers on why they're not speaking out.](#)
- Read: Keller, Daphne. 2024. "The Rise of the Compliant Speech Platform." Lawfare, October 16. <https://www.lawfaremedia.org/article/the-rise-of-the-compliant-speech-platform>.
- ** Optional Read: Moran, Rachel Elizabeth, Joseph Schafer, Mert Bayar, and Kate Starbird. 2025. "The End of Trust and Safety?: Examining the Future of Content Moderation and Upheavals in Professional Online Safety Efforts." Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, April 26, 1–14. <https://doi.org/10.1145/3706598.3713662>.
- ** Optional Read: Kelley, Patrick Gage, Steven Rousso-Schindler, Renee Shelby, Kurt Thomas, and Allison Woodruff. 2026. ["How Generative AI Empowers Attackers and Defenders Across the Trust & Safety Landscape."](#) In Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26). Association for Computing Machinery, New York, NY, USA, Article 1316, 1–21.

Week 15: Wrapping Up

May 5 by 11:00AM: Submit Draft of Final Project

May 5 (Class 28): Wrapping Up

- Second Knowledge Check (final 20 minutes of class)
- Peer evaluation of T&S Policies

May 10 by 11:58PM: Submit Final Project

ADDITIONAL READINGS FOR THE CURIOUS (more to be added)

douek, everlyn. (2021, June 2). More Content Moderation Is Not Always Better. *Wired*. <https://www.wired.com/story/more-content-moderation-not-always-better/>

Francois, Camille and Shen, Juliet and Roth, Yoel and Lai, Samantha and Povolny, Mariel, Four Functional Quadrants for Trust & Safety Tools: Detection, Investigation, Review & Enforcement (DIRE) (July 17, 2025). Trust, Safety, and the Internet We Share: Multistakeholder Insights, Edited Volume, Taylor & Francis (forthcoming), Available at SSRN: <https://ssrn.com/abstract=5369158>

Hay Newman, Lily and Burgess, Matt (2024). The Pig Butchering Invasion Has Begun. *Wired*. <https://www.wired.com/story/pig-butchering-scam-invasion/>

Swartz, Lana (2022). "Theorizing the 2017 Blockchain ICO Bubble as a Network Scam." *New Media & Society* 24 (7): 1695–1713. <https://doi.org/10.1177/14614448221099224>.

Guarna, Tomás. 2026. “From the Virtual Community to ‘Trust and Safety’: eBay (1995–2007) and the Rise of Platform Governance.” *Big Data & Society* 13 (1): 20539517251403973.
<https://doi.org/10.1177/20539517251403973>.

How have platform policies changed over time? The Platform Governance Archive tracks this:
<https://www.platformgovernancearchive.org/>

Mini-Documentary by the NYTimes about a worker doing the work of managing CSAM: [“The Human Toll of One of the Internet’s Worst Jobs”](#)