Syllabus

Course

Title: CS 8803 Online Trust and Safety
Associated Term: Spring 2026
CRN: 35436
Credits: 3

Scheduled Meeting Times

Modality: In-person
Meeting pattern: Tuesday & Thursday, 09:30–10:45 (local Metz time)
Location: Georgia Tech-Europe, 2 Rue Marconi, 57070 Metz, France

Instructor

Instructor: Michael D Bailey
Email: mbailey@gatech.edu
Office: Room 214 (Georgia Tech-Europe)
Office hours: By appointment
Contact & response time: Email is preferred; replies typically within 24–48 hours on business days.

Platforms & Communication

Canvas is the official source of materials, announcements, discussions, schedules, submissions, grades, etc. Materials on Canvas always reflect the most accurate state of the course and supersede anything in this syllabus.

Materials & Tech

Textbook/supplies: None required.
Readings: Links to required readings will be posted in Canvas.

Course Description

Digital platforms now support communication, commerce, creative work, and social interaction at a scale unmatched in earlier eras of technology. These systems enable extraordinary benefits while also creating opportunities for misuse, unwanted behavior, security incidents, and other challenges that organizations must address to maintain user trust. This course introduces the multidisciplinary field commonly referred to as *online trust and safety*, examining how technical systems, policies, operational practices, and legal frameworks interact to shape the user experience.

We will explore how organizations identify and respond to a broad range of issues, from integrity and fraud to user-safety concerns and violations of platform rules, and how approaches

differ depending on a platform's size, function, values, and regulatory environment. A central theme of the course is the distinction between what is unlawful, what may be harmful but lawful, and what organizations choose to restrict under their own terms of service. Because legal definitions, enforcement expectations, and cultural norms vary widely across countries, we will compare U.S. approaches to those in the EU and other regions, highlighting how global platforms navigate conflicting regulations and societal expectations.

Throughout the semester, we will critically examine trust and safety itself as a field: its historical development, the debates surrounding its scope, and the ongoing argument over whether certain practices constitute prudent risk management or excessive restriction of online expression. Students will engage with both proponents and critics of trust and safety practices, assessing arguments about censorship, discrimination, proportionality, and due process in online governance.

This course does not promote any single ideology. Instead, it provides a neutral, analytical setting for evaluating contested questions about digital speech, user protection, and policy design. By integrating case studies, research literature, and comparative legal analysis, students will develop the tools needed to understand and assess trust and safety practices in government, industry, and civil society.

Learning Objectives

- Identify and classify a range of online risks and unwanted behaviors using widely recognized trust and safety concepts and taxonomies.
- Explain how technical, operational, legal, and organizational considerations jointly influence trust and safety decision-making.
- Distinguish among illegal content, harmful-but-lawful content, and policy-violating content, and analyze how different jurisdictions define and regulate these categories.
- Evaluate U.S., EU, and other global regulatory frameworks related to digital safety, intermediary responsibilities, and user rights.
- Analyze the structure and function of trust and safety teams, including policy development, enforcement operations, research, transparency reporting, and measurement.
- Critically assess the tradeoffs among user safety, expression, privacy, growth, and product usability in the design and enforcement of platform policies.
- Compare how platforms of varying size and mission approach risk management, including differences between social media, marketplaces, communication tools, and search or recommendation systems.
- Develop well-reasoned, evidence-based arguments for or against specific policy or product approaches in complex, contested environments.

Prerequisites / Who Should Take This Course

Graduate standing, or

- CS 3235: Introduction to Information Security (or equivalent), and
- CS 3237: Human Dimension of Cybersecurity: People, Organizations, Societies (or equivalent)

Assessment & Grading

In this course, you are evaluated on the clarity, rigor, and evidence of your reasoning, not on which positions you take on contested questions. In particular, you are not required to endorse any government, institutional, or accreditor policy, nor any platform's trust and safety practices. Arguments that defend, reform, or sharply criticize such policies are all welcome, provided they are well-supported, civil, and consistent with applicable law. In evaluating participation and written work, I am interested in how clearly you explain, support, and situate your views, not whether you agree or disagree with positions taken by the instructor, classmates, or assigned authors.

Your overall grade is composed of:

- Participation — 10%
- Discussion Lead — 10%
- Summaries — 20%
- Paper — 60%

Grading scale: A 90–100, B 80–89, C 70–79, D 60–69, F <60. Final scores are rounded up to the nearest whole percent.

Participation (10%)

Each class meeting is worth an equal share of the 10%. Substantive participation includes asking critical questions, offering arguments for or against positions in the readings or in class, and responding respectfully to others' contributions. Thoughtful dissent and the articulation of minority or unpopular views are explicitly valued and count positively toward participation

Discussion Lead (10%)

Your 10% is divided equally across the number of lead slots you're assigned. As a leader you prepare prompts, guide the conversation, and situate the paper(s) in context.

Summaries (20%)

Assessments are equal fractions of the total for each of the required papers. Summaries articulate the main finding or argument, discuss how the argument was supported, and comment on the implications of the finding. Students are expected to critically engage with the work and offer their opinion.

Paper (60%)

Students will produce a research paper on modeled on the types of papers read in class. The grade reflects the quality of the final paper and steady progress through required milestones

(problem definition, literature review, method, implementation/analysis, draft, peer review, and camera-ready).

Submission, Regrades, Late Work

- Where to submit: All work is submitted on Canvas (including summaries right before class on the assigned day).
- Late work: Not accepted. (Exceptions only for institute-approved/documented reasons with advance notice when possible.)
- Regrades: Via email to instructor within 3 business days of grade release. Provide a clear, specific rationale.

Attendance, Excused Absences, and Accommodations

Attendance: Required and part of your grade via Participation.

Excused absences & documentation:

As per Georgia Tech policy, you are permitted to be absent from class to participate in athletic events, official field trips, and religious observances. For planning purposes, please provide me with written notice of your upcoming absence at least two weeks before the event, and ideally within the first two weeks of class. When I receive this notice, you and I will discuss opportunities to make up work you will miss in your absence.

Please see for more information about receiving official notice from the Registrar about the nature and timing of your upcoming Institute-approved absence.

Accessibility & accommodations:

If you are a student with learning needs that require special accommodation, contact the Office of Disability Services at (404)894-2563 or http://disabilityservices.gatech.edu/, as soon as possible, to make an appointment to discuss your special needs and to obtain an accommodations letter. Please also e-mail me as soon as possible in order to set up a time to discuss your learning needs.

Device Policy

Laptops/tablets are welcome for course work. Be considerate and please avoid side-tasking and distractions to yourself or others.

Recording Policy

I may record class sessions or segments for instructional purposes. Students may not record without prior permission from the instructor. Any shared recordings and materials are for class use only; you do not have permission to post or share these materials.

Digital Learning Day / Disruptions

If campus calls a Digital Learning Day or we must pivot, we will meet on Zoom at the normal class time; slides/materials and any updates will be posted on Canvas. No remote proctoring will be used.

Academic Integrity

Georgia Tech aims to cultivate a community based on trust, academic integrity, and honor. Students are expected to act according to the highest ethical standards. For information on Georgia Tech's Academic Honor Code, please visit:

https://www.policylibrary.gatech.edu/student-life/student-conduct

Any student suspected of cheating or plagiarizing on a quiz, exam, or assignment will be reported to the Office of Student Integrity, who will investigate the incident and identify the appropriate penalty for violations.

All students must follow the academic integrity and Georgia Tech Honor Code.

Generative AI and Large Language Models (LLMs)

We treat AI-based assistance (e.g., OpenAI ChatGPT, Google Gemini, Microsoft Copilot, and Anthropic's Claude) the same way we treat collaboration with other people (e.g., a classmate, mentor). As a rule, these interactions serve as a potentially useful way to learn. However, we all know our submitted work must be our own and that we must follow course guidelines on acceptable collaboration. Submitting wording, structure, or code you did not write yourself is plagiarism (D.3). If an assignment forbids outside help, then using AI is unauthorized collaboration (D.2) even if you cite the source.

When collaborating with AI is explicitly allowed you must, of course, still cite your sources (D.3). When you reproduce AI wording or a close paraphrase, add a footnote or endnote at the end of the sentence/paragraph such as:

> *1. ChatGPT, response to "Explain Fitts's law with everyday examples," OpenAI, August 19, 2025. (If you edited the AI text, indicate this in the note: "…edited for clarity.")*

When AI only informed or inspired your ideas Include a brief acknowledgment, as footnote, such as:

> Acknowledgment: I consulted ChatGPT (OpenAI) on August 19, 2025, to brainstorm usability test scenarios; the writing and analysis are my own.

Be careful what you share with AI systems. You do not know what an AI service will do with uploaded content. Uploading class documents constitutes an intellectual-property violation (D.9). Further, you may also facilitate cheating for others if the system retains or learns from restricted material; do not upload, request, or regenerate improperly acquired or restricted

content (D.1) such as assignments, answer keys, prior exams, slides, or classmates' work. Share only your own materials or those you are explicitly given permission to share.

Do not request fabricated data, citations, results, screenshots, or logs as such fabrication and misrepresentation are serious violations (D.4/D.6). Remember that AI can, and often does, hallucinate data, facts, and references, but you are ultimately responsible for everything you submit.

Course-specific collaboration policy

You may not use outside assistance for your paper summaries and your discussion lead work. After the submission deadline, you may share your work collaboratively with others to learn how to engage critically with the topic and improve future attempts.

You may collaborate to explore topics of interest in preparation for your paper topic and for searching for appropriate literature. You should cite any prompts or individual conversations appropriately. However, the submitted work should be yours. Neither other students nor AI agents should write the problem statements nor review the literature on your behalf. After these tasks, remaining activities such as establishing research methods, implementation or analysis, writing paper drafts, and peer reviews must be completed independently.

Sensitive Content & Neutrality Statement

This course does not adopt or promote any perspective or ideology. Instead, it engages with issues that governments, industry, and civil society are actively debating today. We will approach these topics in a neutral, academic setting, focused on critical analysis rather than advocacy. Because these debates are live and consequential, some material will inevitably touch on difficult, sensitive, or personally challenging subjects. Except for activities that are clearly unlawful, this course will not tell you what to think or what solutions to adopt; instead, it will give you tools to frame, analyze, and debate the issues as future professionals.

- Content Warnings & Alternatives: Whenever possible, I will provide advance notice when sensitive material is covered. If a student feels unable to engage with specific content, they should work with me to identify alternative, but relevant assignments or activities. Alternative assignments are offered to support individual well-being, not to restrict what others may read or discuss. Primary materials remain available for all students, and class discussion will not avoid difficult topics solely because they are sensitive or controversial.
- Open Dialogue: In line with widely adopted principles of university free expression, such as the University of Chicago's "Report of the Committee on Freedom of Expression", our classroom is a place where even strongly contested or offensive ideas may be introduced for discussion and critique. You are free to question, reject, or defend any position raised in course materials or discussion, as long as you do so without threats, harassment, or disruption of others' ability to participate

- Respect & Inclusion: We will not all agree on every point, but mutual respect is always expected. This includes listening attentively, avoiding personal attacks, and recognizing that others may come from different lived experiences and cultural perspectives.
- Professional Preparation: These challenges mirror those faced by professionals worldwide. By practicing critical, respectful engagement here, you will be better prepared for leadership in industry, government, or civil society.

Student–Faculty Expectations Agreement

At Georgia Tech, we believe that it is important to strive for an atmosphere of mutual respect, acknowledgement, and responsibility between faculty members and the student body. The student-faculty expectations articulate some basic expectations that you can have of me and that I have of you. In the end, simple respect for knowledge, hard work, and cordial interactions will help build the environment we seek. Therefore, I encourage you to remain committed to the ideals of Georgia Tech while in this class.

Campus Resources (Graduate)

A list of resources for graduate students is provided by the Office of Graduate and Postdoctoral Education. Information for current graduate students includes Academic Resources (Communication Center, Language Institute, Library, Catalog, Registrar, research support, Advocacy & Conflict Resolution, and guidance for unexpected situations affecting academic performance), Student Resources (Campus Services, Child Care/Family programs, Health & Wellness, Career Services, Student Resource Guide), and Professional Development (Career Center programming and other professional development resources and events).

https://grad.gatech.edu/student-resources
https://grad.gatech.edu/academic-resources

Student Well-Being: Georgia Tech is concerned about your overall physical, social, and mental well-being. A comprehensive list of wellness-related resources is maintained by the Office of the Vice President for Student Engagement & Well-being (see the Student Resource Guide).

https://students.gatech.edu/

Subject to Change

The syllabus/schedule may evolve; I will post changes via Canvas Announcements with at least 48 hours' notice when feasible. I expect the most adjustments in week 1 as we learn more about each other's interests; this is a graduate seminar.