An Introduction to Trust and Safety

CS 152 — Trust and Safety

Alex Stamos

Stanford Cyber Policy Center

March 31, 2026

What will we learn today?

Today, we will…

- Discuss the practical aspects of this class
- Review our learning objectives so you can determine if this class is for you
- Explore how Trust and Safety differs from other areas of tech risk
- Start to understand the Trust and Safety lifecycle and the kinds of roles that exist in industry

Class Logistics

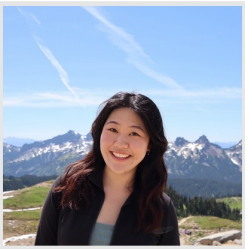
Instructors

Jeff Hancock

Alex Stamos

Course Assistants

Kieran Barrett
Head CA



Tracy Wei
CA

Arjun Jain
CA

Student Introductions

Get into groups of 3 or 4 and introduce yourselves, and why you are interested in trust & safety

Our Agenda

Date	Topic
Tuesday, March 31, 2026	Introduction to Trust and Safety
Thursday, April 2, 2026	How Tech Companies Work: Designing for Trust, Safety, and Privacy
Tuesday, April 7, 2026	AI Safety Part I
Thursday, April 9, 2026	Hate Speech, Incitement, Harassment
Tuesday, April 14, 2026	Intro to the US Legal System, Privacy, Surveillance and Law Enforcement
Thursday, April 16, 2026	Free Speech on the Internet and Trust and Safety
Friday, April 17, 2026	Last day for add/drop, 5pm deadline, Milestone 1 due
Tuesday, April 21, 2026	Project Intro and Group Setup Day
Thursday, April 23, 2026	Adolescent Well Being and Social Media
Tuesday, April 28, 2026	Suicide and Self-Harm
Thursday, April 30, 2026	Terrorism and Violent Extremism
Tuesday, May 5, 2026	Online Child Sexual Exploitation
Thursday, May 7, 2026	Working with Law Enforcement: Investigation Case Studies
Friday, May 8, 2026	Milestone 2 due
Saturday, May 9, 2026	AI Safety Part II
Thursday, May 14, 2026	Fraud, Pig-Butchering, Scams
Tuesday, May 19, 2026	Dating Apps and the Sharing Economy
Thursday, May 21, 2026	Misinformation and Disinformation
Tuesday, May 26, 2026	Content Moderation, Tooling and Resiliency
Thursday, May 28, 2026	Emerging Issues and Career Advice
Tuesday, June 2, 2026	Alumni Panel + Final Project Event at 5pm

Who should take this class?

Students who:

1. Are interested in the ways technology can be abused to cause harm
2. Want to build consumer internet products more safely
3. Who are interested in careers in Trust and Safety, anti-abuse NGOs, law enforcement or internet policy
4. (Taking CS listing) Who can participate in the group project at a 106B level

Project Teams + Sections

Project teams

- Teams of four CS students, one non-CS (on average)
- You will be able to request teams
- Please meet new people/include younger students

Sections will start next week

- TA facilitated work sessions. Optional, but highly encouraged
- Focus of sections will be the project, we will not be discussing practical components in lecture
- Best way to make sure your team is meeting once a week, staying on top of the milestones

Grading

Course Grading

- Pre-read Quizzes - 20%
- Lecture Attendance - 10%
- Project Milestone 1 - 20%
- Project Milestone 2 - 20%
- Project - Final presentation - 30%

Final Project

- **Milestone 1:** Solo - Abuse Study and Project Pitch (20%) - *due April 17*
- **Milestone 2:** Content Moderation Tool Initial Implementation (20%) - *due May 8*
- **Final:** Final Tool and Poster (30%) - *poster session June 5*

An Introduction to Trust and Safety

Alex Stamos

Class Logistics

Learning Goals

What is Trust and Safety Engineering?

Who are the Players?

Questions?

CS152 Team 36

Cryptocurrency Fraud on Twitter

Julia Steinberg, Chase Small, Jasper den Otter, Matt Frank, and Miles McCain

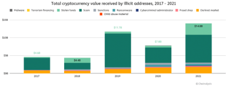
What's the problem?

The cryptocurrency ecosystem is a **hotspot of abuse**. The experimental technology has grown into a sprawling ecosystem of competing tokens, "smart contract"-powered financial instruments, NFTs, DAOs, and more.

Between its rapid growth, minimal oversight, and culture of hype, it is unsurprising that scams and fraud plague the cryptocurrency ecosystem. In 2021, at least **$10 billion in illicit transactions** took place via cryptocurrencies. We differentiate between **first-order** and **second-order** scams.

First-order scams	Second-order scams
• Money multiplication schemes • Unfulfilled purchases of goods • Romance scams • Phony or fake cryptocurrency exchanges • Certain online gambling platforms • Wallet phishing (e.g. via fake "airdrops")	• "Pump and dump" scams • Artificially expensive NFTs • Fraudulent NFTs (i.e. where minter is not the original creator)

How do you regulate a community that is perhaps defined by its lack of regulation? How do you fight scams in an ecosystem that eschews the idea of a central authority? Fortunately, we find that much "consumer-facing" abuse takes place **centrally on Twitter**. We therefore focus on how Twitter can fight cryptocurrency abuse.



"Crypto Crime Trends for 2022: Illicit Transaction Activity Reaches All-Time High in Value, All-Time Low in Share of All Cryptocurrency Activity," Chainalysis, January 6, 2022, https://blog.chainalysis.com/reports/2022-crypto-crime-report-introduction/.

Evaluation

Semantic Search Engine

- Our semantic similarity engine is **not intended** to detect all forms of cryptocurrency abuse.
- Rather, its purpose is to detect and flag **known scam content** while being robust to small variations in phrasing and vocabulary.
- Many cryptocurrency scams rely on high message volume; by preventing scams from reusing messages, we hope to make it **prohibitively expensive** to operate a large-scale scam operation.
- The semantic search engine shows **promising performance**, shown in the confusion matrix below.

N=156	Flagged	Not Flagged
Scam Content	100%* (76)	0%* (76)
General Crypto Content	2.5% (2)	97.5% (78)

* As this is a semantic search system intended to detect repeat scams, these results are **not intended** to indicate perfect generality.

Known Abuse Detection

- Our bot runs every link and Bitcoin address through databases of known harmful content (Bitcoin Abuse API and the Safe Browsing V4 API).
- While these databases reduce the burden of human moderators by blocking known scams, they are **easily bypassed** by adversarial actors.
- Of all the links and addresses in our cryptocurrency abuse dataset, **none were blocked by these external APIs**.
- Our bot is also largely unable to distinguish between abusive content and *warnings* about abusive content.

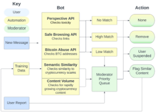
General Toxicity Detection

- While our bot removes the most toxic messages automatically, it is **oversensitive to profanity**.
- Messages like "fuck I forgot" and "haha fuck me" are removed, despite not being highly toxic.

Obama's Twitter account being used to peddle a cryptocurrency scam. Aspirationally, our system would have proactively flagged this content and prevented it from being posted.

Technical Backend

- To prevent **message obfuscation**, we normalize all formatting and special unicode characters in every message; we also run all images through OCR (Tesseract).
- We then detect general toxicity with the **Perspective API**, deceptive links with the **Safe Browsing API**, and scam Bitcoin addresses with the **BitcoinAbuse.com API**.
- Our bot also maintains its own database of **previously seen** abusive cryptocurrency addresses and cashtags.
- We compare every message to previously detected abuse using our own **Semantic Similarity Engine**, powered by Facebook's **FastText**.
- We automatically flag **high-velocity cryptocurrency content** for human review.



Looking Forward

Whether we like it or not, cryptocurrencies are here to stay. Given the complexity of "Web3" technology—and the inflexibility of blockchains—the victim pool and potential impact for crypto abuse are large.

- We could partner with **independent coin validation companies** like CertiK to automatically apply warning labels to untrustworthy projects.
- Much of cryptocurrency abuse is **intersectional**: consider sextortion, CSAM sharing, and traditional money laundering. Our bot focuses primarily on financial scams; a clear next step is to handle **different sides** of cryptocurrency abuse.
- More broadly, there are already standards for the regulation of **traditional scams** — enforcement must extend to cryptocurrencies as well.

A scammer targeting one of our group members, attempting to get them to sign up for a relatively unknown cryptocurrency exchange.

Our Policy

While we allow legitimate discussion of cryptocurrencies, we have a strict policy against cryptocurrency scams. You may not use our platform's services to deceive others into sending or investing money, including cryptocurrency.

What is a cryptocurrency scam?

- We disallow both **first-order** and **second-order** cryptocurrency scams.
- First-order cryptocurrency scams solicit cryptocurrency payments under false pretenses.
- Second-order scams are superficially legitimate investments or purchases (e.g. one receives an asset in return), but have deceptive foundations.

What violates this policy?

- Phishing scams, including wallet links (e.g. for "airdrops")
- Money-multiplication schemes (e.g. asking someone to send you 0.1 BTC in exchange for 1000 BTC)
- Promoting fake cryptocurrency exchanges
- Coordinated promotion of coins or NFTs (e.g. "pump and dumps")
- Posting links meant to capture the financial information of other individuals

What is not in violation of this policy?

- Giving financial advice without solicitation of personal information, including speculation about future performance
- General discussion of cryptocurrency markets
- Asking for donations via cryptocurrency (provided the recipient is clearly and truthfully disclosed)
- Commentary *about* cryptocurrency scams, including warning others

What happens if I violate this policy?

- If you post an address or "cashtag" known to be abusive, we will automatically remove your post or apply a warning label.
- If you repeatedly violate the policy, you may be suspended from the platform.
- If your content violates our policies, you will be notified automatically.

Differences with the Project This Year

1. **We expect more because of generative AI.** You can use AI coding tools to be more productive, so the bar for final deliverables is higher. We expect nicer interfaces, more complete solutions.
2. **We will focus on software engineering practices.** You will learn how to work together with GitHub, CI/CD, tickets, and cloud infra.
3. **We are widening the scope** of types of trust and safety problems to include more AI-specific issues.

Pre-Reading

Pre-reads and short reading quizzes before **every** class

- Pre-reads will be finalized two weeks in advance (although most are avaiable now)
- Quizzes will go up after previous class

Readings can be found at cs152.stanford.edu/syllabus

Attendance Policy

- Students are required to attend lecture in person.
- There are no videos available this quarter. The resolution of real-life is amazing.
- 30% of your grade will come from attending the lectures and the pre-read quizzes
- We will grant two missed lectures and two missed pre-read quizzes per student, no questions asked. No other exceptions will be granted without an OAE letter.
- Handling a blizzard of special exceptions from 125 students is the worst part of this job and we appreciate your understanding.

Word of the Day

Hedgehog

Enrollment

- We have a long waitlist, so I guess we are doing something ok
- We want everybody to be able to take the class, there is usually enough churn that the waitlist will make it in
- If we get to the add/drop deadline and there is still a waitlist, we will possibly expand the class, as long as we can balance the teams

AI Policy

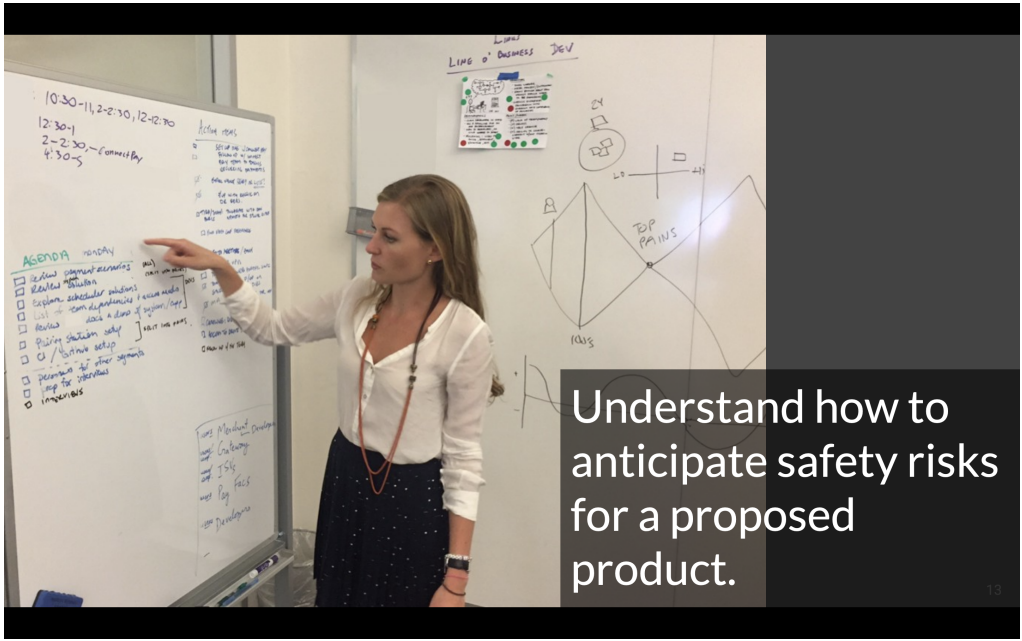
- Students **can** use AI tools as a research aid
- Students **can not** use AI to write any prose in the class (no use to generate Milestone 1 or the written parts of the final project).
- Google credits will be provided for use as part of the final project
- AI code generation is allowed, but must be disclosed on the final poster and final submission
- You really should understand what the code does, the goal is to let AI make you a faster coder so that you learn more and your project is more impactful

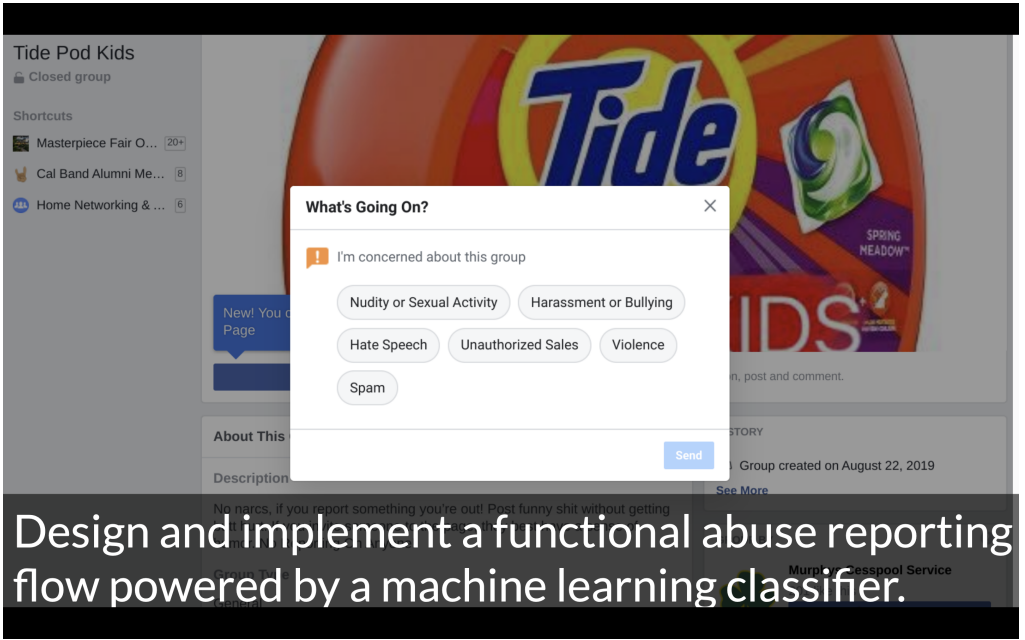
Getting Help

- Go to section!
- Ask most questions on the Ed Discussion Board
- Chat on the Discord (invite coming next week)
- Alex's Office Hours - Book on Calendly

Learning Goals

Recognize the pattern of how long-existing societal challenges (hate speech, disinformation, child abuse) can be changed or amplified by modern communication platforms.





Have empathy for a broad cross-section of the people who use your products and the risks they face.

Dealing with difficult content

The subject matter of this course can be difficult intellectually and emotionally. We will read about and discuss difficult topics, including (but not limited to) sexual exploitation of adults and minors, harassment, bullying, hate speech, domestic abuse, terrorism, and more.

If you anticipate acute distress as a result of encountering a particular topic, talk to me ahead of time to arrange an alternative written assignment in lieu of your in-class participation. If you become so distressed that you need to leave during class, feel free to do so. If you need to leave a class, talk to me afterward and we can arrange an alternate assignment. I will not "warn" students about particular topics, because sensitivity to different topics varies from person to person, and because topics may arise unexpectedly in class discussion. Please refer to the course agenda to see the list of course topics.

Additionally, as you may know, there is a difference between being triggered (in the sense of post-traumatic stress disorder) and feeling uncomfortable. One of the goals of this class is to help students develop empathy for victims of online abuse. Feeling uncomfortable (and sometimes even angry or offended) is part of intellectual growth. Feeling triggered or psychologically traumatized is not. Please take care of yourselves and each other, and let me know if I can do anything at all to help.

Pay Attention to the Topics

Date	Topic
Tuesday, March 31, 2026	Introduction to Trust and Safety
Thursday, April 2, 2026	How Tech Companies Work: Designing for Trust, Safety, and Privacy
Tuesday, April 7, 2026	AI Safety Part I
Thursday, April 9, 2026	Hate Speech, Incitement, Harassment
Tuesday, April 14, 2026	Intro to the US Legal System, Privacy, Surveillance and Law Enforcement
Thursday, April 16, 2026	Free Speech on the Internet and Trust and Safety
Friday, April 17, 2026	Last day for add/drop, 5pm deadline, Milestone 1 due
Tuesday, April 21, 2026	Project Intro and Group Setup Day
Thursday, April 23, 2026	Adolescent Well Being and Social Media
Tuesday, April 28, 2026	Suicide and Self-Harm
Thursday, April 30, 2026	Terrorism and Violent Extremism
Tuesday, May 5, 2026	Online Child Sexual Exploitation
Thursday, May 7, 2026	Working with Law Enforcement: Investigation Case Studies
Friday, May 8, 2026	Milestone 2 due
Saturday, May 9, 2026	AI Safety Part II
Thursday, May 14, 2026	Fraud, Pig-Butchering, Scams
Tuesday, May 19, 2026	Dating Apps and the Sharing Economy
Thursday, May 21, 2026	Misinformation and Disinformation
Tuesday, May 26, 2026	Content Moderation, Tooling and Resiliency
Thursday, May 28, 2026	Emerging Issues and Career Advice
Tuesday, June 2, 2026	Alumni Panel + Final Project Event at 5pm

Students with Documented

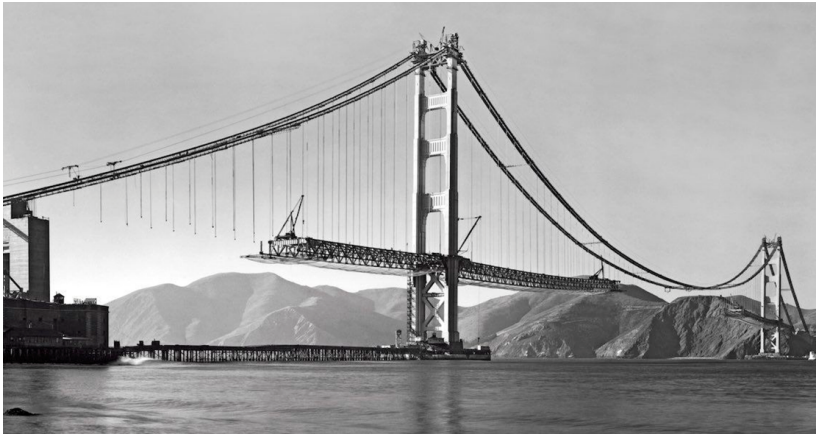
Students who may need an academic accommodation based on the impact of a disability must initiate the request with the Office of Accessible Education (OAE). Professional staff will evaluate the request, review appropriate medical documentation, recommend reasonable accommodations, and prepare an Accommodation Letter for faculty dated in the current quarter in which the request is being made. The letter will indicate how long it is to be in effect. Students should contact the OAE as soon as possible since timely notice is needed to coordinate accommodations. The OAE is located at 563 Salvatierra Walk (phone: 650-723-1066, oae.stanford.edu).

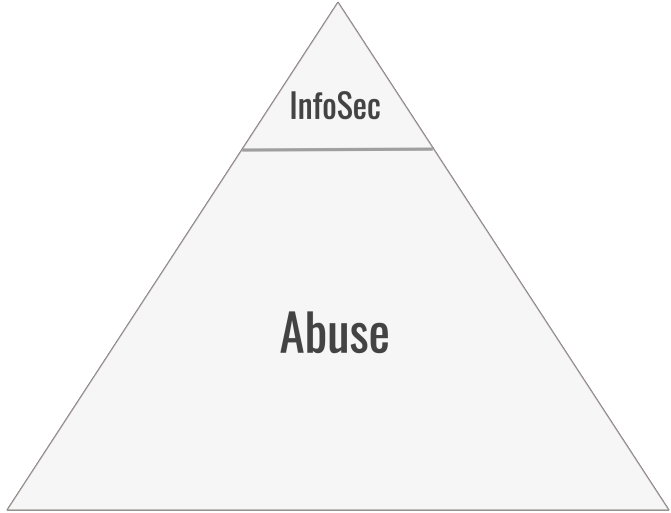
This class has no timed exams, and the project deadlines are firm as they apply to entire teams, so most OAE accommodation requests relevant to other courses are not relevant here. This is something we can discuss once we have your OAE letter.

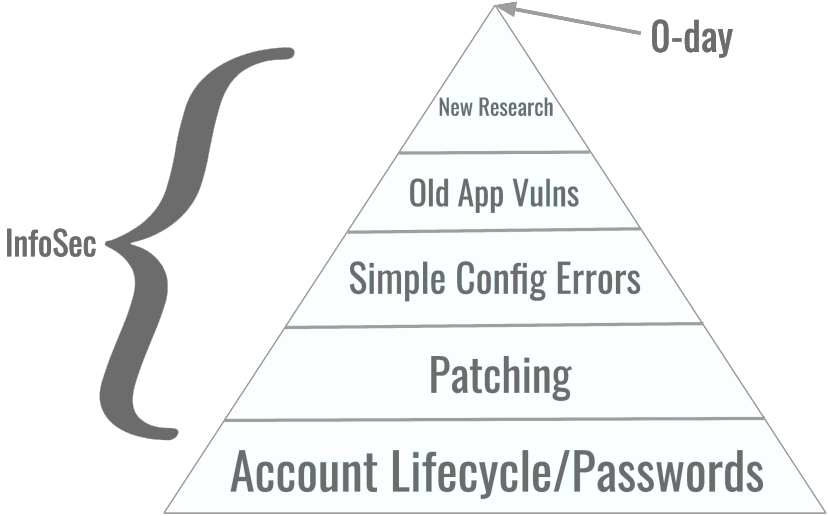
What is Trust and Safety Engineering?

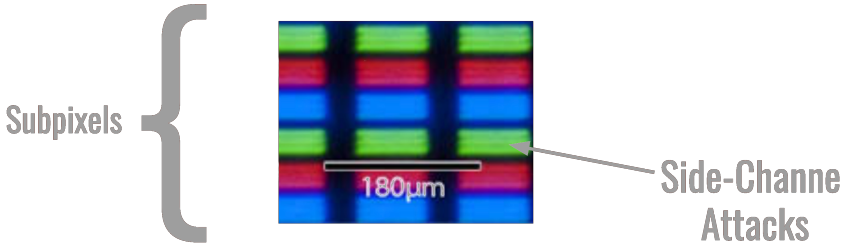
Unique Aspects of Trust and Safety

1. The study of how people abuse the internet to cause harm.
2. Often using products the way they are designed to work.
3. Crosses between specialties. Requires understanding of society and humanity.
4. Is dynamic and unpredictable.









The Biggest Challenges in Trust and Safety

1. Scale
2. Non-diverse studies and solutions
3. Measurement and definition challenges
4. Privacy vs Safety
5. Information sharing and division of responsibility
6. Government vs private action
7. Fairness in ML solutions
8. Freedom of expression

Who are the Players?

Areas of Trust and Safety Work

Policy & Research

Defining abuse types, building measurements and metrics, performing field studies, interviewing users and victims, working with gov affairs teams, provides data and ideas to product

Product & Eng

Red teaming product designs, designing UX that encourage good behavior, building detection and moderation mechanisms, building and training ML

Operations

Defining appropriate behavior, building operational pipelines, sorting and handling billions of events, implementing constant improvement through self-testing and QA

Investigations

Investigates worst cases or most effective bad guys, handles incoming LE requests and external referrals, applies lessons learned to future operational and product

Jobs in Trust & Safety

Figure 1: Discord job screenshot

- Identify and assess the capabilities and activities of threat actors involved in Minor Safety and Exploitative Content behaviors and produce intelligence reports to help initialize or support law enforcement investigations or activities and strengthen the company's defenses
- Proactively identify currently undetected abuse by leveraging internal data, open-source intelligence, trusted partner information, and third-party private intelligence

Jobs in Trust & Safety

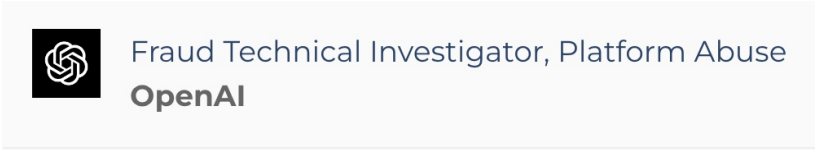


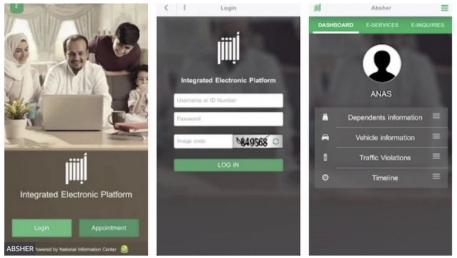
Figure 2: OpenAI job screenshot

- Discover, triage, investigate, and report on abusive and fraudulent behaviors on our platform.
- Respond to real-time abuse incidents by stabilizing the situation and implementing mitigations.
- Develop new methods to expand and automate our detection coverage.
- Collaborate with engineering, policy, and research teams to enhance our tools and understanding of abusive content.

A Real and Very Political Example: Saudi Arabia

Saudi app used to track women 'not against' Google rules

5 March 2019



Google has remained silent over reports it told a US congresswoman that a controversial app was not in breach of its terms and conditions.

If you worked at Google or Apple, would you allow this app in your app store? What policy language would you use to justify allowing or prohibiting it?

Questions?