

# CSE291 – Sociotechnical Cybersecurity

*Enforcement Mechanisms, Industry Metrics, and Examples of Research*

UC San Diego

# Housekeeping

- Paper assignments were released **yesterday**
  - If you don't have an assignment and are still enrolled, please let me know ASAP so we can get you a slot
  - More papers will be scheduled to accommodate the growth in class size
- Project specification is out: <https://kumarde.com/cse291-fa24/>, and project intention form is available: <https://forms.gle/TQY9AHQLzfaJC6Tc7>
  - Due by **10/8 @ 12:30pm PT**
- Arshia Arya, Joey Wu, Henry Feng, Ivan Liang are on the docket for next week
  - Paper presentation guidelines are here: <https://kumarde.com/cse291-fa24/projects/cse291-paper-presentation.pdf>

# News

## ***Instagram, Facing Pressure Over Child Safety Online, Unveils Sweeping Changes***

The app, which is popular with teenagers, introduced new settings and features aimed at addressing inappropriate online contact and content, and improving sleep for users under 18.

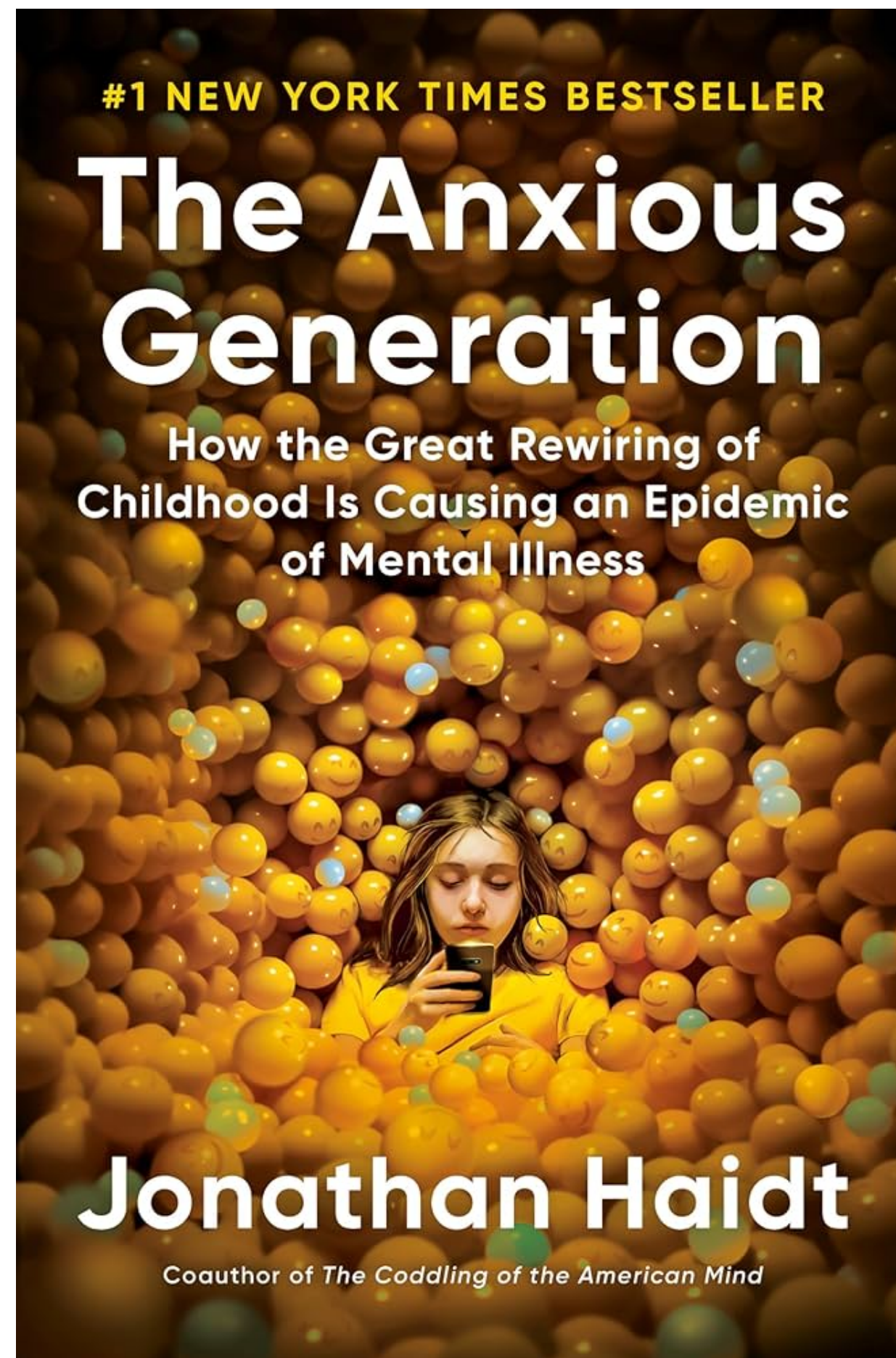
# Instagram's new changes

## Creation of "Teen Accounts"

- "Teens 13 – 17 automatically have a protected experience, with built-in limits on who can contact them and the content they see."
- Here's a list of the changes
  - Users under 18 will be made *private by default*
  - Teen accounts are automatically set so teens can't be messaged by anyone they're not connected to
  - Teen accounts are set to see "less sensitive content" – including potentially offensive comments, message requests w/ strict word settings
  - Notifications automatically muted from 10pm – 7am



# The cultural context



## **Congress's online child safety bill, explained**

What is the Kids' Online Safety Act, and why should you care about it?

- Creates a "duty of care" for users 16 and under
- KOSA passed in senate 91 – 3
  - Heading to the house
- Supporters say its good for kids
- Detractors say its a huge infringement on free speech

# Let's pretend we're teenagers

*What is the first thing I might do if I'm a teenager in this new world?*



# Let's pretend we're teenagers

*What is the first thing I might do if I'm a teenager in this new world?*

How are you ensuring teens don't lie about their age? —

We know teens may lie about their age and that's why we're requiring teens to verify their age in new ways, like if they attempt to use a new account with an adult birthday. We're also building new technology to find teens that have lied about their age to automatically place them in protected settings.

**The fine print:** To prevent teens from lying about their age to circumvent the new settings, Instagram will now require users to verify their age in new ways, such as via a government-issued ID or facial scans. Methods will vary depending on the country, Mosseri said.

- Teens and their parents or guardians have to mutually agree to supervisory relationships for a parent to access control over a teen's account. "We can't verify a parental relationship. There's no good way to do that at scale. So it can be another adult in your life," Mosseri said.



# What's being advocated for?

**Pinterest CEO: To protect our kids online, Congress must make digital IDs the national standard—and require OS makers to share age-validation data with apps**

BY **BILL READY**

Bill Ready is the CEO of Pinterest.

September 23, 2024 at 1:57 AM PDT



**Meta doesn't want to control how teens use the internet – it wants to make app stores do it**



Image: Nick Barclay / The Verge

/ Meta wants Google's and Apple's app stores to handle online age verification.

By [Emma Roth](#), a news writer who covers the streaming wars, consumer tech, crypto, social media, and much more. Previously, she was a writer and editor at MUO.

Nov 15, 2023, 7:31 AM PST

[Link](#) [Facebook](#) [Twitter](#) | [10 Comments \(10 New\)](#)

# How would an OS-level verification system work?

The weird intersection with Digital IDs

- Supporters draw a comparison between age verification for buying alcohol and using Instagram
- *“Congress must make digital IDs the national standard and require OS platforms to send age-validation information to apps”*
- What are some problems we have to consider?

**Californians can now add their mobile driver's license to Google Wallet**

# Recap

# Previously on Sociotechnical Cybersecurity....

- We talked about the T&S ecosystem in industry, the various stakeholders involved, and the complications of regulating T&S
  - *Can we recall two factors that drive T&S at companies*
- We discussed a framing for types of harms and how the harm determines the mechanism
  - *What are the two types of harms?*

# Today's lecture – Enforcement, Metrics, and Research

## Learning Objectives

- Understand the enforcement mechanisms that T&S teams have at their disposal to address harms
- Identify the metrics used by T&S teams and how these metrics are typically collected
- A potpourri of styles of research projects
- *Next time: Papers!*



# Enforcement Levers

# Studying Platform Enforcement

- Our team has been studying how platforms claim to enforce the rules they lay out in community guidelines
- We studied the guidelines of 10 major social media platforms: Facebook, TikTok, Nextdoor, Tinder, BeReal, X, YouTube, Snapchat, Twitch, and LinkedIn
- Though *affinity diagramming*, we identified and synthesized into major themes



# A Taxonomy of Enforcement Mechanisms

- Platform Actions
  - Actions that affect offending content
  - Actions that affect offending community / group
  - Actions that affect the offending account
  - Actions that affect the offending entity
- User-driven Actions

# Actions that affect offending content

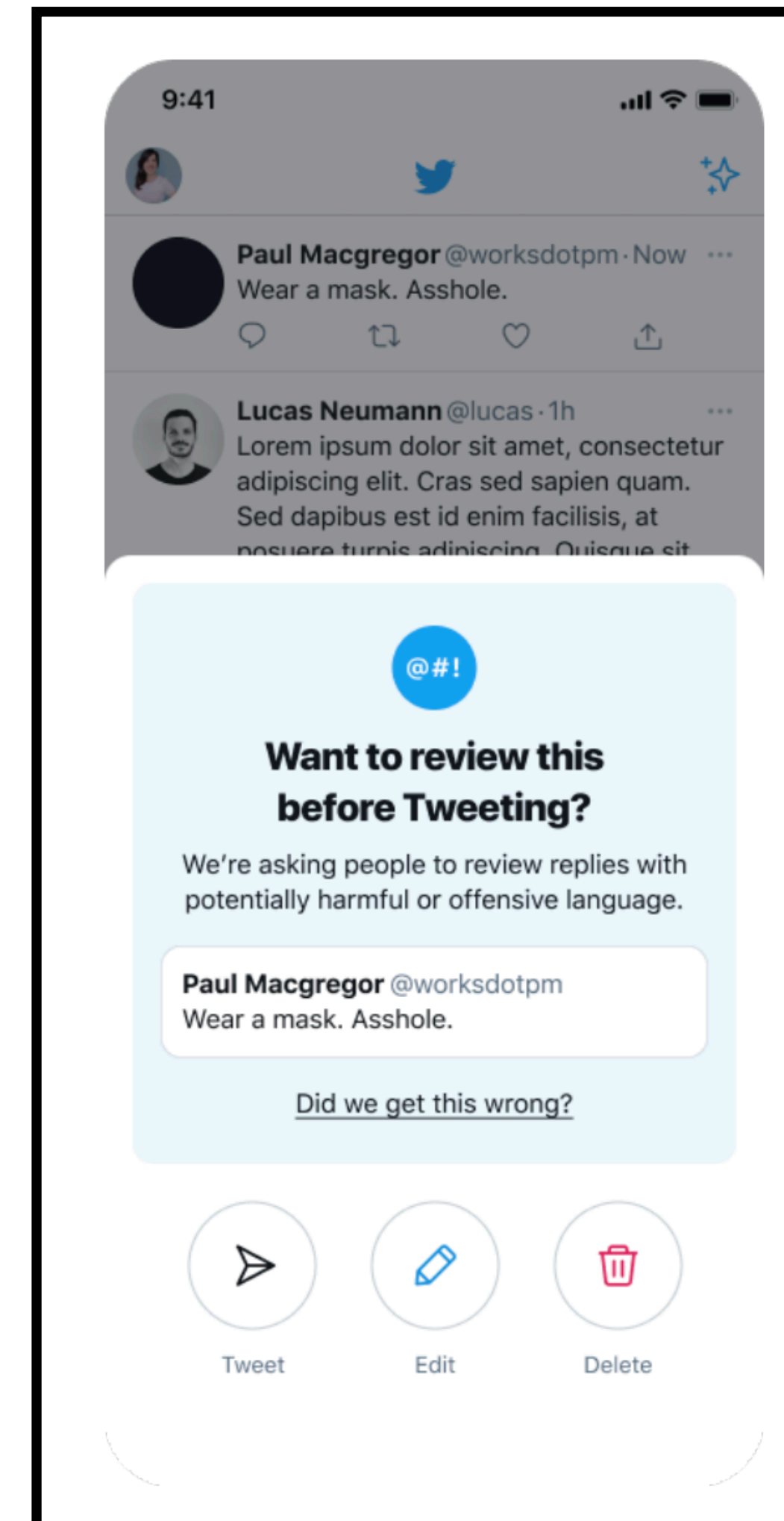
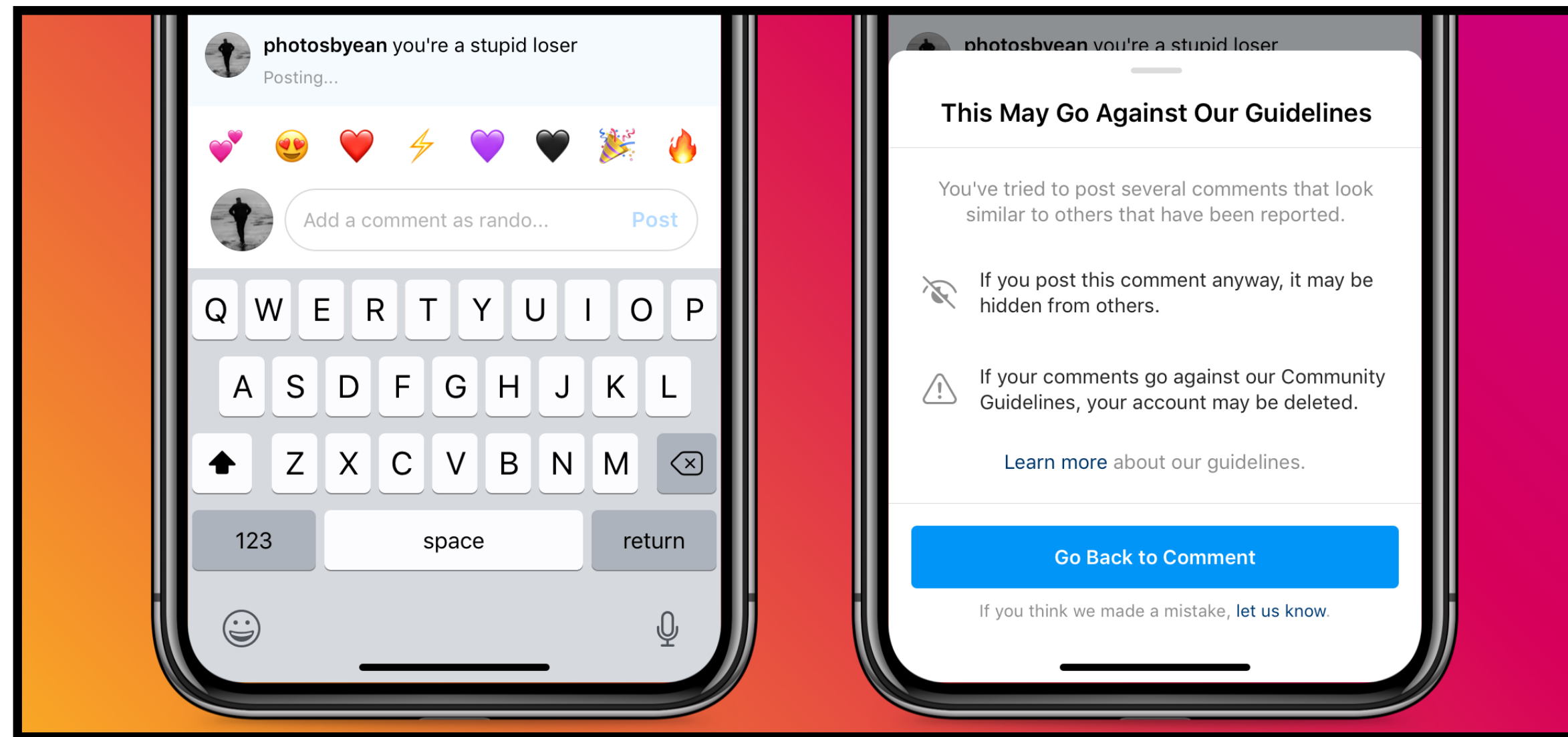
- Remove content
- Limit content visibility
- Limit content interaction
- Label content
- Restrict content monetization

# Actions that affect offending account

- Push a nudge / notification
- Issue a warning / strike
- Limit account visibility
- Force identity verification
- Limit account abilities
- Temporarily suspend account
- Terminate account



# Interventions are on the rise



## Strike system

With the new Rule[0] revision, we'll also be introducing a strike system in an attempt to improve the content quality and encourage people to read and follow the new rule. Authors of posts that will be removed for violating the new revision of Rule[0] will receive 1 strike for every post removed. Please note that the strike system currently only applies to Rule[0]. The following punishments will be given for receiving strikes:

- Strike 1 - 1 day tempban
- Strike 2 - 3 day tempban
- Strike 3 - 7 day tempban
- Strike 4 - 30 day tempban
- Strike 5 - permanent ban

# Actions that affect offending entity

- Prevent entity from using the service
- Report entity to law enforcement
- Proactively ban the entity

# User-driven actions

- User is encouraged to block, silence, or hide content
- User is encouraged to label / identify content
- User is encouraged to contact external entity
- User is encouraged to engage in interpersonal off-platform mediation

# Enforcement levers are vast and the design space is growing

- Strikes and warnings are growing in popularity
  - Twitch, Discord have adopted a public “strike” system which is auditable and verifiable
- Design exercise – break into groups of 3
  - **Brainstorm 3 new enforcement mechanisms that you think might be useful in a Trust & Safety context**

# Metrics



# What's a metric?

- "A measurement system that quantifies static or dynamic characteristics"
  - How do we quantify some experience in a way that is consistent and comparable?
- In practice, metrics...
  - Have a definition that can be counted + measured
  - Are utilized for tracking the effect of a team, product, policy over time
  - Are most useful when they can be measurably changed by meaningful actions
    - New system —> metrics change

# Some popular metrics in digital platforms

- DAU / MAU – Daily/Monthly Active Users
  - Keeps track of number of people who have active accounts on the platform
- Clickthroughs
  - Keeps track of how many times people interact with a button
- Engagement
  - Keeps track of platform specific features: likes, comments, shares, etc.

# The duality of metrics

- Goodhart's Law
  - "When a measure becomes a target, it ceases to be a good measure"
- If measures are used for *control* or to promote *scarcity*, it can lead to bad social outcomes
  - YouTube algorithm optimizing for screen time —> leads to radicalization
  - Facebook algorithm optimizing for engagement —> leads to upranking contentious / harassing content

# Why measure metrics for T&S

- Goal of T&S is to protect users from harm
  - In theory, there's a relationship between high user trust & higher engagement, but it's *hard to demonstrate*
- T&S is a cost center —> the levers that T&S use also reduce engagement
  - How do we show that T&S contributes to the success of a company?

# Some examples of metrics T&S use





*Let's talk about measurement*

- How prevalent are different types of abuse on a platform?
- How much harm to users has been prevented?
- To what degree do users trust a platform and feel safe interacting there?
- Are users leaving the platform because of under or over enforcement?
- How effective are our content moderation processes (e.g., response speed, accuracy, etc.)?

*What else?*

# What metrics are used in practice?

*Gleaned from Transparency Reports*

							
Report Frequency	Quarterly	Last Report in 2021	Quarterly	Annually	Quarterly	Quarterly	Quarterly
Ability to Compare Reporting Periods		✓		✓*	✓*	✓	✓
Removed Content Items	✓		✓	✓	✓	✓	✓
Breakdown by Policy Violation	✓		✓	✓	✓	✓	✓
Number of Content Appeals	✓		✓			✓	✓*
Number of Restored Accounts	✓		✓			✓	✓*
Number of Restored Content Items	✓		✓			✓	✓*
Accounts Actioned	✓		✓	✓		✓	✓*

<https://www.activefence.com/research/transparency-report/>



# How do we get metrics for T&S?

*Let's talk about measurement*

## Behavioral Logs

*Telemetry collected from users and queried by T&S teams*

## Entity Labeling

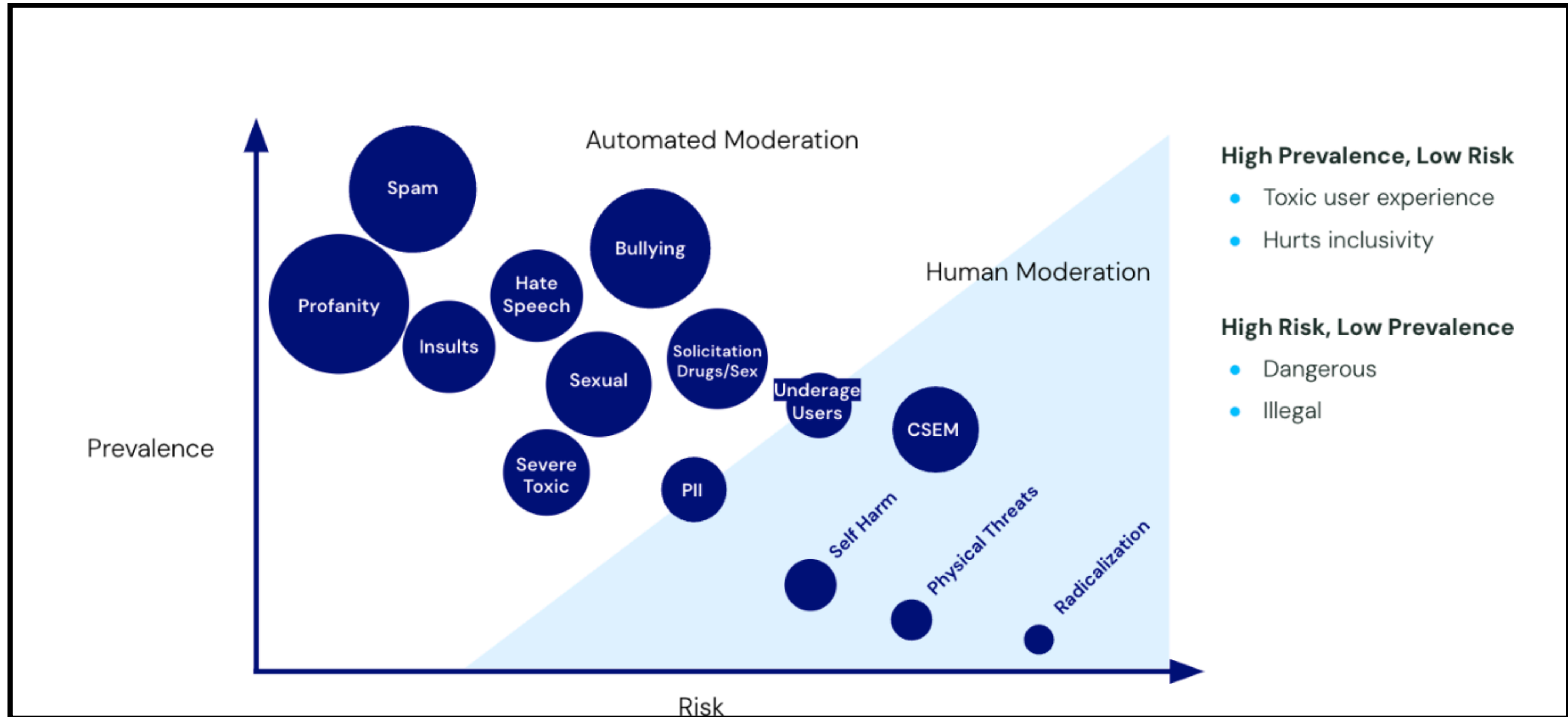
*Automated / Manual systems that detect specific harms or targets*

# Tradeoffs between both types of metrics

*Let's talk about measurement*

- Behavioral logs typically come from normal operation of the platform (e.g., telemetry) – but as a result can be limited in their purview
- Metrics based on labeling tend to cost more and can be error-prone and biased
  - Small sample labeled
  - Labeler consistency
  - The world changes! So your models have to as well.
- But, labels can provide more nuance and context for solving more subjective issues in T&S contexts

# Where's the line between automated / human labeling?



# How else can we get data?

*We can ask people!*

- Surveys, focus groups, interviews are all part of the “health” conversation
- Some difficult-to-measure ideas are often the most practically measured through surveys
  - Trust, sentiment, overall vibe-checks
- Drawbacks
  - Hard to time surveys with interventions, so harder to draw causal effects
  - Difficult to design (tons of pitfalls)
  - Humans are messy and aren’t really consistent in their adjudication
- But, it’s one more data point and can be useful in conjunction with other metrics

# Break Time + Attendance



**Codeword:**  
Pumpkin-Spice

<https://tinyurl.com/cse291attendance>

# Doing Research in Sociotechnical Cybersecurity



# Why do we do research at all?

*Isn't this their problem?*

- Companies are governed by their own incentives
  - Product growth, internal politics, re-orgs, etc.
- Whimsy of the market means T&S is often short-changed
- T&S teams are always on fire (and therefore not forward looking)
- As an academic, you can ask questions *across the ecosystem* instead of simply fixing one platform's issues

# Drawbacks of research in this space

*Data.... data*

- Access to data can be quite tricky
  - Some platforms offer open APIs, others have been restricting them due to massive LLM training controversies
- Researchers often don't have platform context
  - Other signals that might be useful
- But... more often than not, researchers are on the same page at T&S teams

## **Twitter's \$42,000-per-Month API Prices Out Nearly Everyone**

Tiers will start at \$500,000 a year for access to 0.3 percent of the company's tweets. Researchers say that's too much for too little data.

**Reddit  
Is Killing  
Third-Party  
Applications  
(And Itself)**

# Nonexhaustive List of Research Styles

*4 main types of research*

- Measurement
- Design / Systems
- Causal Inference
- Human Subjects (impossible for our 10 weeks together)

# Measurement Research

*How do we measure Internet problems?*

- Typical construction: “I have a question about X ecosystem. I have devised a system to collect data to measure that ecosystem. I make a lot of assumptions about how the data ought to look. I analyze the data and check my understanding”
- Pros
  - Construction is similar across different projects – similar techniques but vastly different areas
  - Provides more basic understanding of problems
- Cons
  - Takeaways are not always obvious (framing is important)
  - Often requires a lot of assumptions and a lot of data (and you never have perfect data)
  - You could fail

# Design / Systems Research

*What can we try that hasn't been tried?*

- Typical construction: "I have observed some problem X. I think X can be solved with Y design / system. I built Y system using some clever ideas, and I evaluated Y system to see how well it solves X problem."
- Pros
  - You finish with an end-product – something you have actually built
  - Forces you to think about practicalities in building the thing (software issues, performance, scale, etc.)
- Cons
  - In T&S, design work is hard to immediately get deployed – transfer into industry isn't as neat
  - Evaluations can be very hard to pull off well
  - You could fail (your system could not work)

# Causal Inference

## *A child of measurement*

- Typical construction: “I have observed phenomena X, which I think has an impact on Y outcomes. I design or find a natural experiment E, which I think tests X in isolation, and I figure out if X had any effect on Y.”
- Pros
  - Causal effects can be a stronger way of demonstrating a phenomena’s effect
  - Often on the cutting edge of what statistics / econometrics research tools are using
- Cons
  - Experiments can be very fickle and hard to reproduce (c.f., reproducibility crisis in psychology)
  - Sometimes results are hard to understand due to a number of hidden assumptions
  - You could fail



# Human Subjects Research

*Surveys, interviews, and many more*

- Typical construction: “I want to understand how people experience X harm/phenomena, but it’s hard to measure with existing metrics. I carefully design a survey or interview experiment Y, pilot that experiment with test participants, iterate on my research instrument, and then deploy it to the world. I analyze the data and try to understand what’s going on.”
- Pros
  - Grounds your work in lived experience – one of the most important and overlooked aspects in computer science
  - Results are often much more nuanced + complex and match the reality of how people behave on the Internet
- Cons
  - Humans are messy: small effect sizes with modeling
  - Very hard to do well, and they take a longer time (hence, not in these 10 weeks)
  - You could fail

**Through line in all of research:  
You could fail**

# Group Time