CSE291 – Sociotechnical Cybersecurity

The Trust & Safety Ecosystem: Harms, Stakeholders, Incentives, Levers

UC San Diego

Housekeeping

- By **10/2 @ 12:30pm PT**
  - Fill out topic preferences document: https://forms.gle/mrNDT2D4JQ7X29Fu7
  - Project spec released for term project in evening
- By **10/8 @ 12:30pm PT**
  - Project intention document due for term project (group selection + general research idea)
- I'm told that EASy requests should be getting approved shortly (we should have plenty of room, though some implications for the course)

News

NEW DECISION SETS OUT NUANCED APPROACH TO ALLOW CONTENT CRITICIZING STATE ACTIONS THROUGH NATIONALITY-BASED CRIMINAL ALLEGATIONS

Meta’s Oversight Board

- Started in October 2020 as a quasi-judiciary on Facebook
  - From Zuck: *“You can imagine some sort of structure, almost like a Supreme Court, that is made up of independent folks who don’t work for Facebook… reflects the social norms and values of people all around the world.”*
- Some notable cases
  - Upholding ban of Donald Trump from Facebook after January 6th, 2021
  - Cartoons depicting dissent and protest on college campuses
- https://transparency.meta.com/oversight/oversight-board-cases/

What happened this week?

All three were removed under hate speech rules

Case 1: *Russians and Americans are criminals*

Case 2: Genocide… all Israelis are criminals

Case 3: All Indians are rapists

"dehumanizing speech or imagery in the form of comparisons, generalizations or unqualified behavioral statements (in written or visual form)" about "criminals."

What happened this week?

All three were removed under hate speech rules

✗ Case 1: *Russians and Americans are criminals*

✓ Case 2: Genocide… all Israelis are criminals

✓ Case 3: All Indians are rapists

Why?

Decision + Remediation

- Amend Hate Speech Community Standard, specifically rule about dehumanizing speech + generalizations, to include exception:
  - *Except when the actors (e.g., police, military army, soldiers, government, state officials) and/or crimes (e.g., atrocity crimes or grave human rights violations, such as those specified in the Rome Statute of the International Criminal Court) imply a reference to a state rather than targeting people based on nationality.*
- Publish results of internal audits to assess the accuracy of human review + performance of a automated systems

Design Exercise

- Let’s say you were building an automated detector for this particular carveout. What factors would you need to consider?

The Trust & Safety Ecosystem

Previously on Sociotechnical Cybersecurity….

- We talked about a history of online communication, and the perfect storm of regulation, culture, and technology that enabled our current communication landscape

- We talked about a myriad of online harms that we now have to figure out what to do with

- We ended our discussion of material with a provocation: **what are we going to do about it?**

Today’s lecture – Understanding the T&S Ecosystem

Learning Objectives

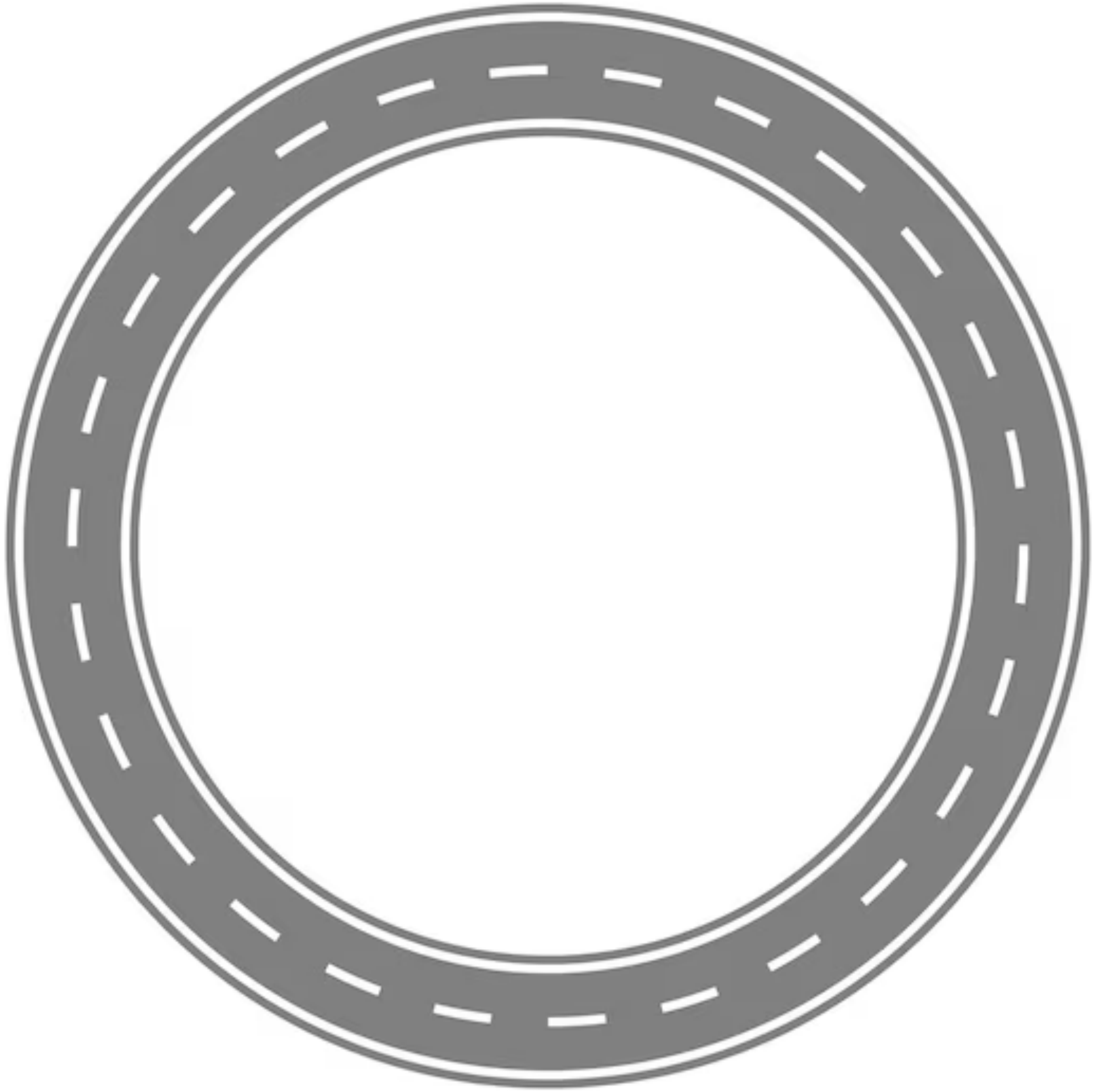
- Learn the motivations behind Trust & Safety and the typical T&S flow
- Know the major stakeholders involved in handling online harms
- Know the major classes of abuse types that stakeholders have to consider, and explain the tradeoffs each stakeholder needs to consider when handling types of harms
- Understand the incentives and levers that T&S teams have at their disposal to address harms

Motivations

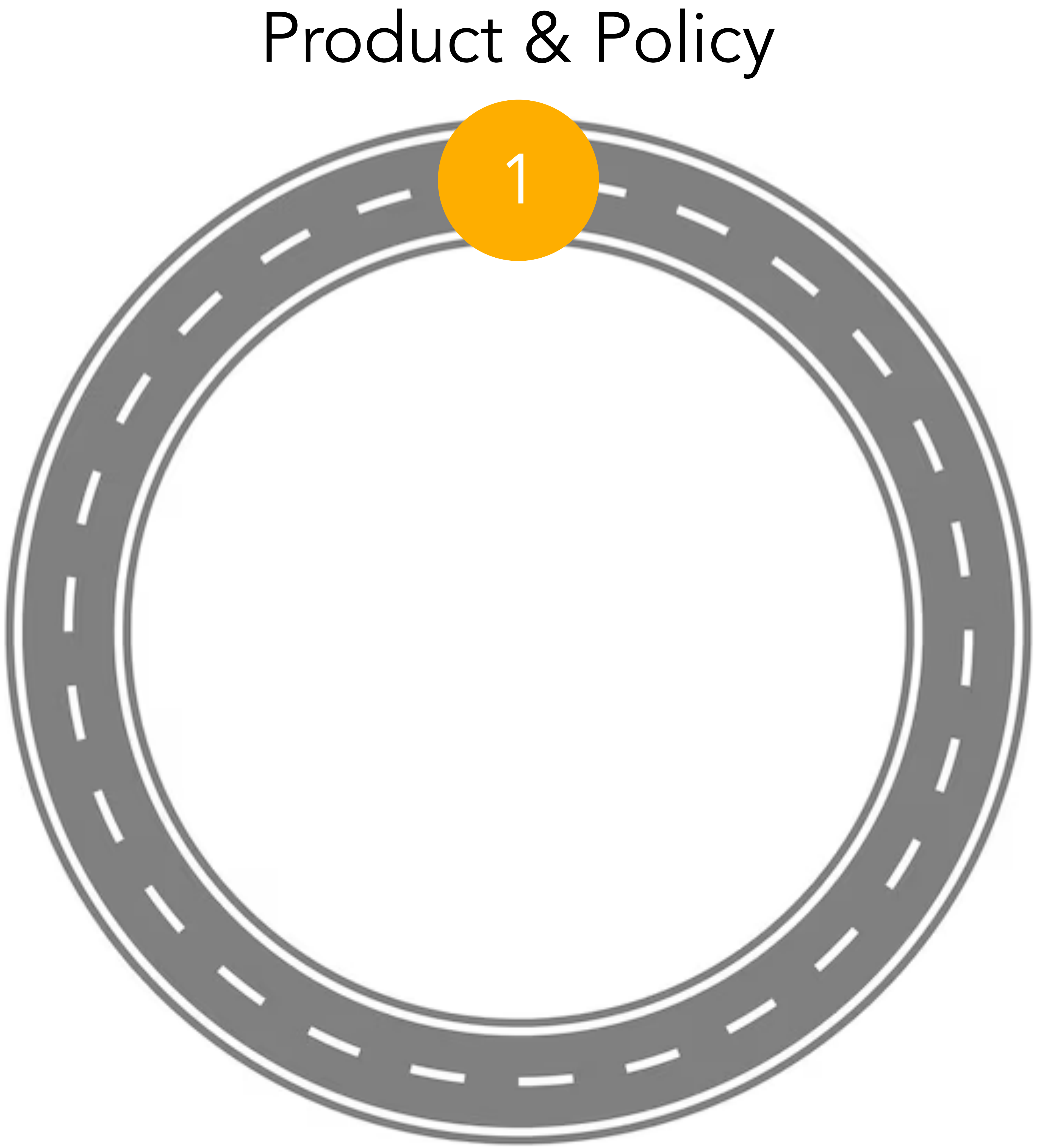
What do Trust & Safety teams do?

- “Enabling users to have the best experiences possible with a product or service… while preventing, detecting, and responding to abuse.”
- T&S teams have to consider many factors when making decisions, like:
  - The type of product
  - The type of abuse
  - The *values* of a company
  - The demographics of a company’s customers
  - The *countries* in which it operates

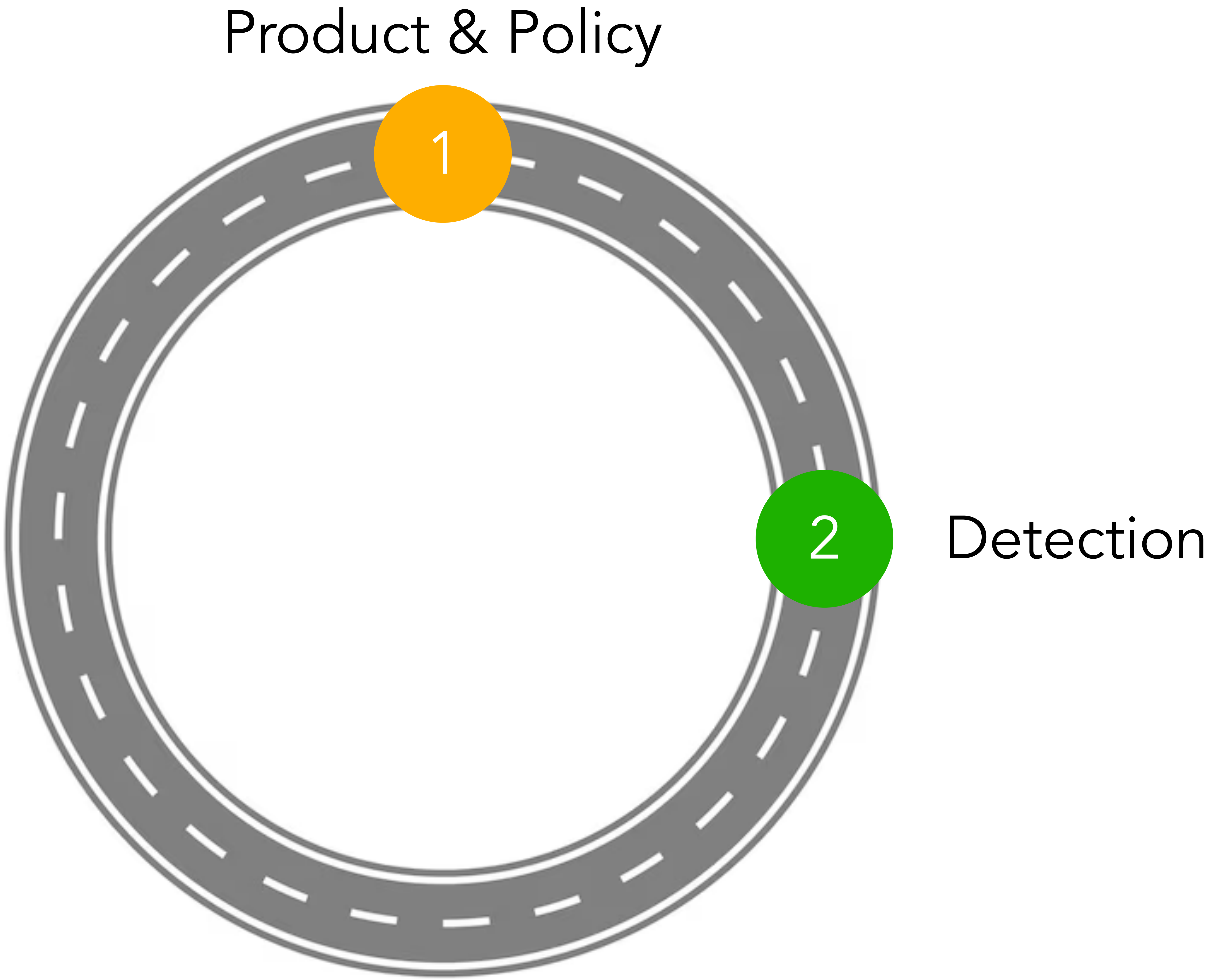
Trust & Safety Lifecycle



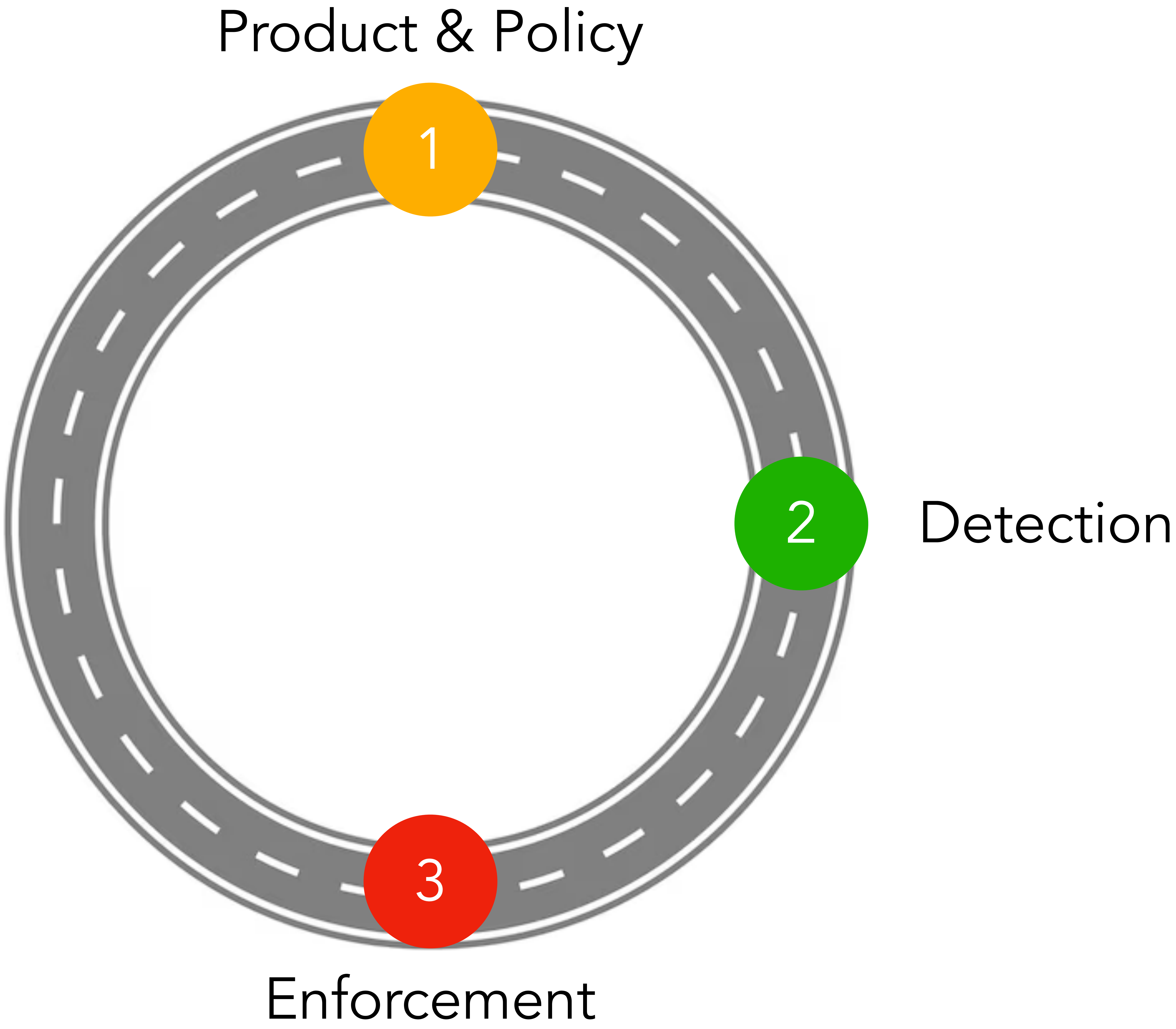
Trust & Safety Lifecycle



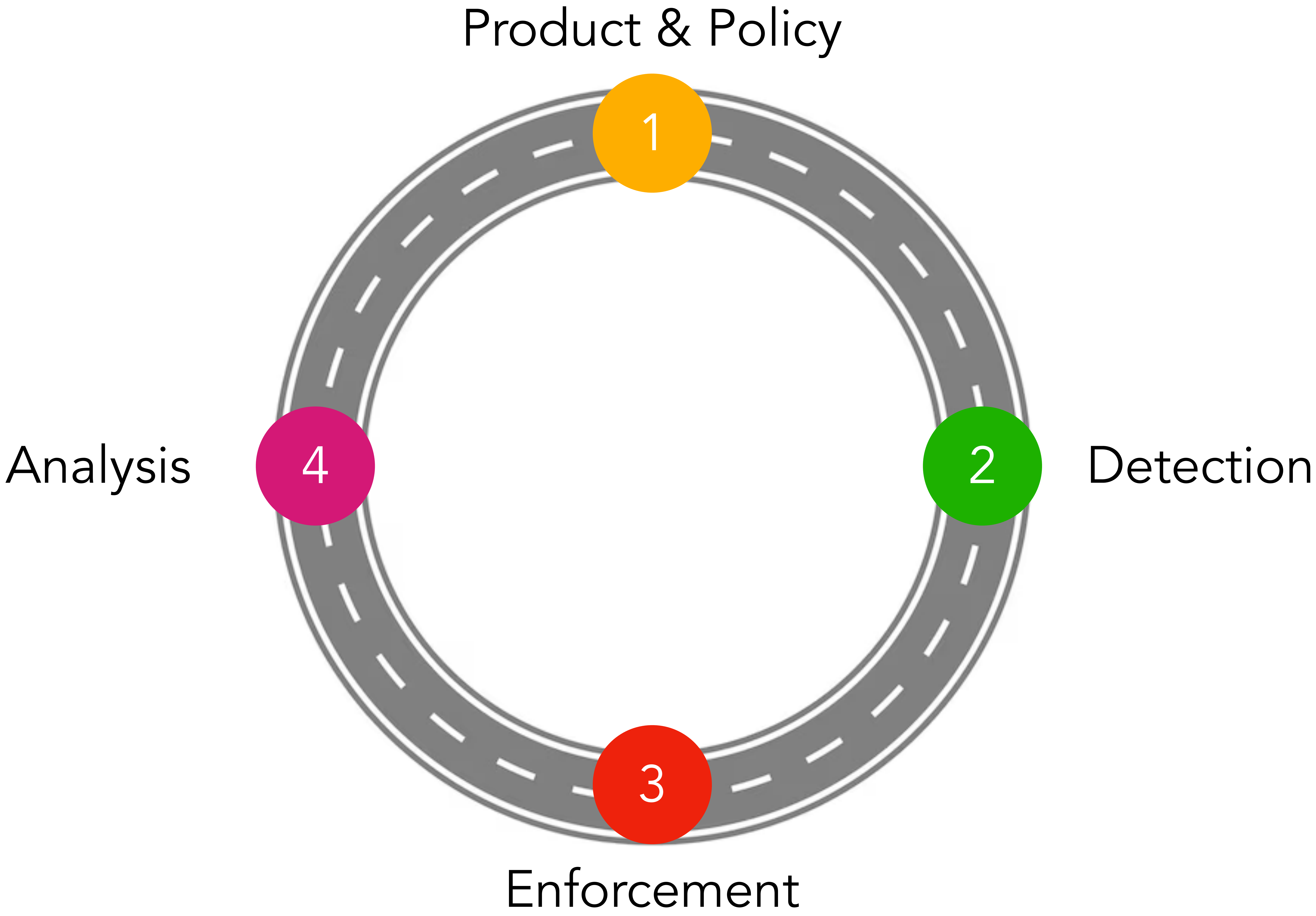
Trust & Safety Lifecycle



Trust & Safety Lifecycle



Trust & Safety Lifecycle



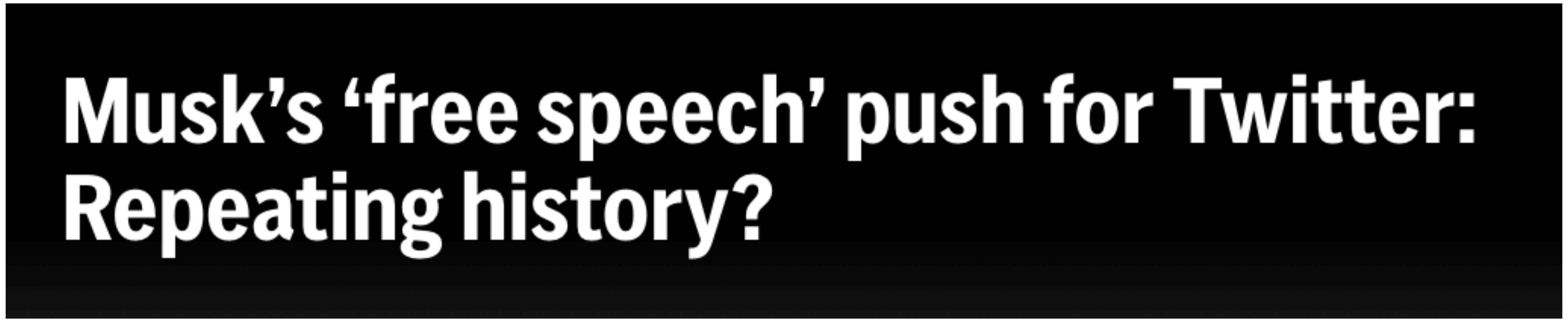
What drives Trust & Safety at companies?

What drives Trust & Safety at companies?

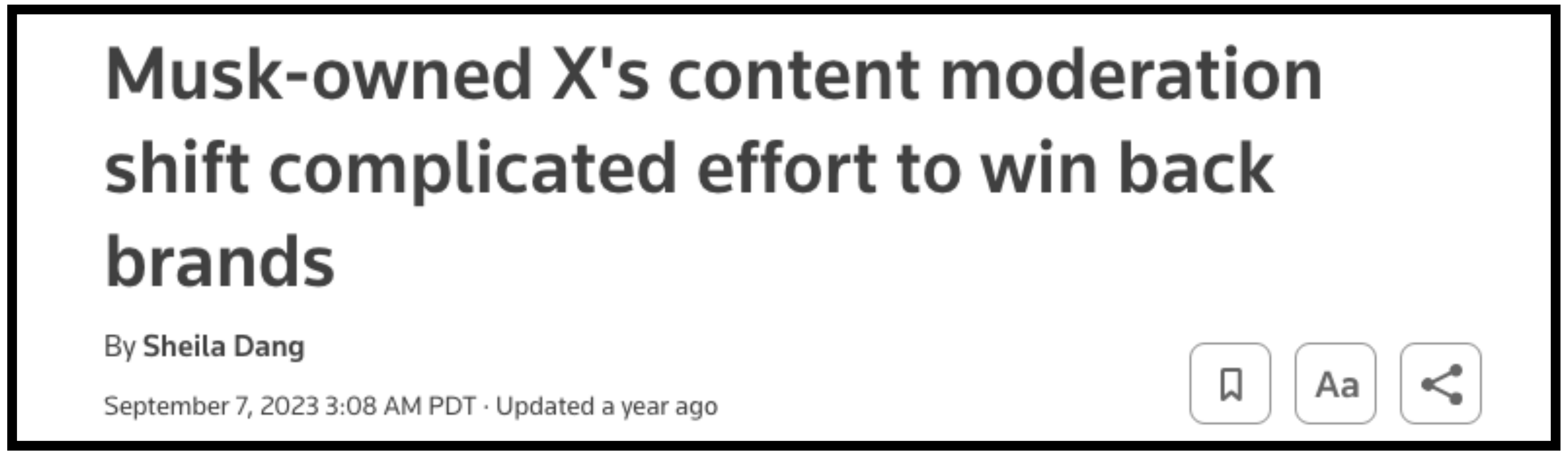
- User safety
  - Why might companies want to keep users safe online?
- Reputation
  - “Nazi Bar” problem – your reputation is **what you allow**
- Regulation
  - GDPR, DSA in the EU
  - Fear of regulation in the US

What drives Trust & Safety at companies?

- Revenue



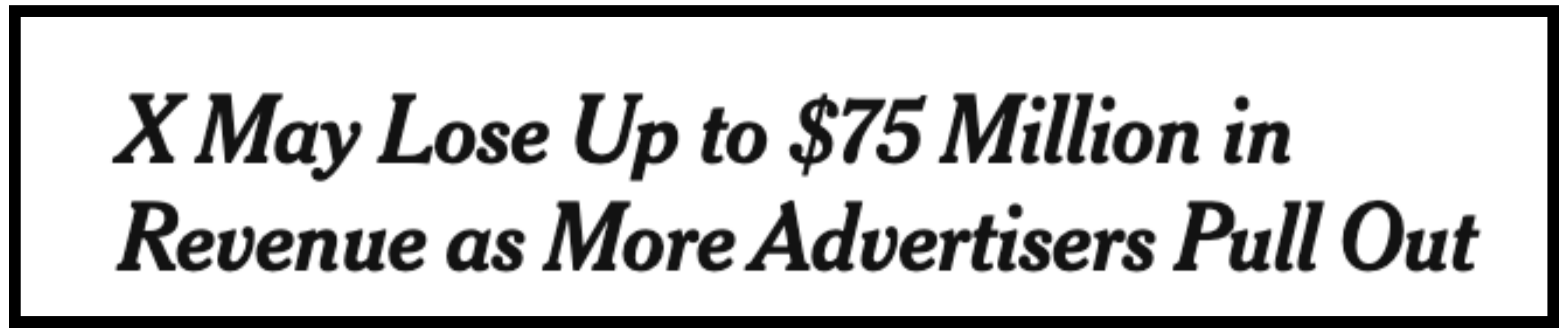
Musk's 'free speech' push for Twitter: Repeating history?



Musk-owned X's content moderation shift complicated effort to win back brands

By Sheila Dang

September 7, 2023 3:08 AM PDT · Updated a year ago



X May Lose Up to $75 Million in Revenue as More Advertisers Pull Out

What drives Trust & Safety at companies?

- Revenue
- Crisis
  - https://tsjournal.org/index.php/jots/article/view/81

Jan 2020: *Normal company no one had ever heard of designed for business*

What drives Trust & Safety at companies?

- Revenue
- Crisis
  - https://tsjournal.org/index.php/jots/article/view/81

Jan 2020: *Normal company no one had ever heard of designed for business*

March – May 2020: ***How every single person communicated via video call***

What drives Trust & Safety at companies?

- Revenue
- Crisis
  - https://tsjournal.org/index.php/jots/article/view/81

Jan 2020: *Normal company no one had ever heard of designed for business*

March – May 2020: ***How every single person communicated via video call***

May 2020: ***Zoombombing becomes a thing to deal with***

What drives Trust & Safety at companies?

- Revenue
- Crisis
  - https://tsjournal.org/index.php/jots/article/view/81

zoom

Jan 2020: *Normal company no one had ever heard of designed for business*

March – May 2020: ***How every single person communicated via video call***

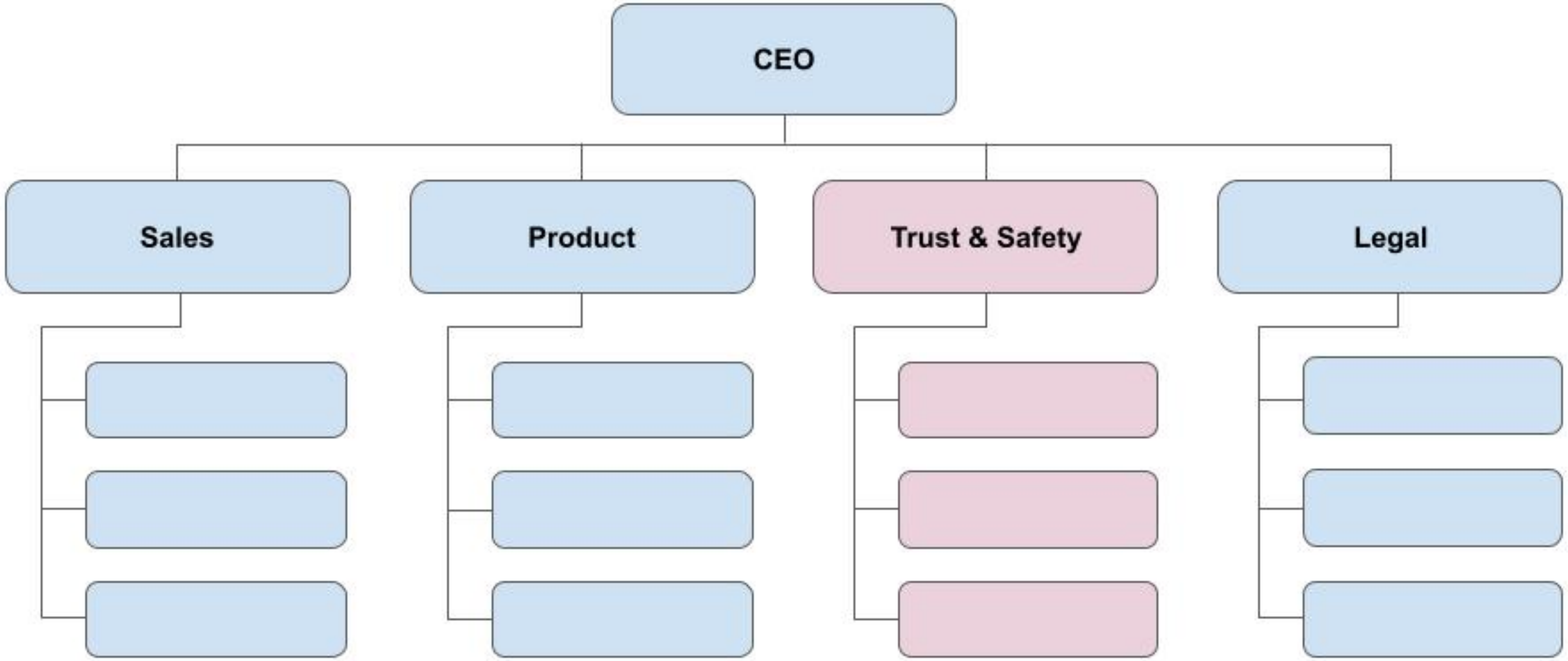
4chan

May 2020: ***Zoombombing becomes a thing to deal with***

May 2020 – Dec 2020: ***Massive increase in T&S development + feature development***

How is T&S situated inside of companies?

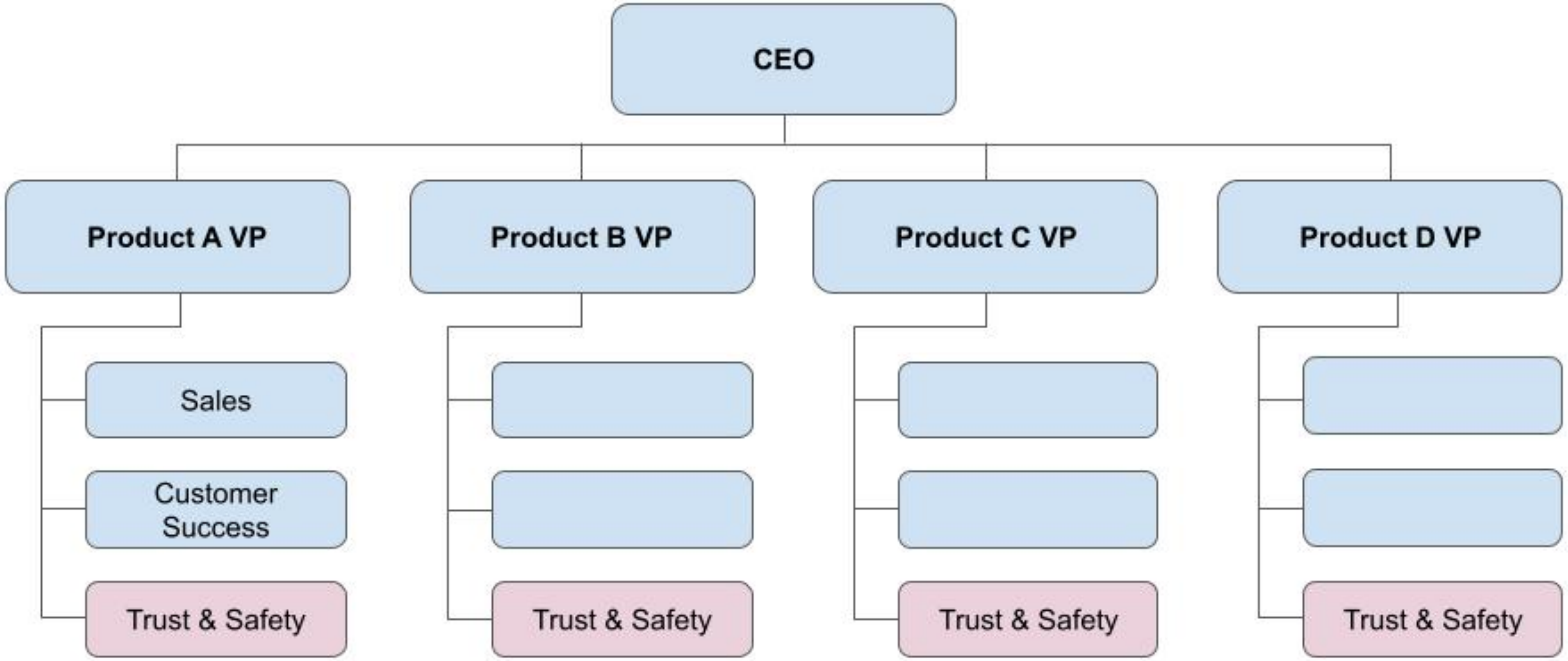
Centralized model



https://www.tspa.org/curriculum/ts-fundamentals/industry-overview/ts-approaches/

How is T&S situated inside of companies?

Dispersed model



CEO

Product A VP

Sales

Customer Success

Trust & Safety

Product B VP

Trust & Safety

Product C VP

Trust & Safety

Product D VP

Trust & Safety

https://www.tspa.org/curriculum/ts-fundamentals/industry-overview/ts-approaches/

Stakeholders involved in T&S

- Users
- Policy people
- Product people
    - Engineering
- Legal people
- Comms people
- **Trust and Safety teams are constantly managing tradeoffs between various stakeholder needs.**

Scenario 1

- You run a T&S team inside a Twitter-like platform for online discourse. A user has written a post offering their services to dox any other user on the platform for a fee. What would you do?
    - Take down the post
    - Take down the post and ban the user
    - Keep the post up

Scenario 1

- You run a T&S team inside a Twitter-like platform for online discourse. A user has written a post offering their services to dox any other user on the platform for a fee. What would you do?
  - **Take down the post**
  - Take down the post and ban the user
  - Keep the post up

Scenario 1

Understanding doxxing and harms

- In a paper by Snyder et al. (https://dl.acm.org/doi/pdf/10.1145/3131365.3131385) – they studied the prevalence of doxxing in text-file based sharing platforms (e.g., pastebin)
    - 5530 / 1.7M (0.3%) of files were doxxing files —> **a low prevalence, but high impact attack**
    - Everyone typically agrees that doxxing is bad

Scenario 2

- You run a T&S team inside a YouTube-like platform for streaming content. Someone has uploaded the entirety of the Shrek movie on their channel, and Dreamworks' lawyers have sent a DMCA takedown request. What would you do?
  - Leave "Shrek" alone
  - Take "Shrek" down

Scenario 2

- You run a T&S team inside a YouTube-like platform for streaming content. Someone has uploaded the entirety of the Shrek movie on their channel, and Dreamworks' lawyers have sent a Digital Millennium Copyright Act takedown request. What would you do?
  - Leave "Shrek" alone
    - Your company gets sued and has to appear in court – leaving your lawyers extremely annoyed at you
  - Take "Shrek" down
    - Users are upset that you took the movie down and you take a 3% reduction in DAU

Scenario 3

- You run T&S at a Twitter-like platform for online discourse. A user responded to another, elderly user with the phrase “OK Boomer” in an online discussion, which the elderly user reports. What would you do?
  - Do nothing, this is clearly a joke
  - Remove “OK Boomer” and issue the account a warning to remain civil

Scenario 3

- You run T&S at a Twitter-like platform for online discourse. A user responded to another, elderly user with the phrase “OK Boomer” in an online discussion, which the elderly user reports. What would you do?
    - Do nothing, this is clearly a joke
        - Elderly users leave your platform, feeling unsupported by the decision
    - Remove “OK Boomer” and issue the account a warning to remain civil
        - Younger users believe you to be overreacting to a simple joke, feeling like the platform is far too paternalistic

Scenario 3

- Users of digital platforms often have competing interests – in part because the way they experience the Internet is vastly different from one another
  - So – a defense for one user group may be wrong, or outright *bad* for another user group
- Abuse experiences of users can vary across many demographics

Demographic	Treatment	Reference	Odds of Harassment
LGBTQ+	LGBTQ+	non-LGBTQ+	1.9x
Social Media Usage	Daily	Never	2.5x
	Weekly	Never	2.3x
Age	18 – 24	65 and up	4.0x
	25 – 34	65 and up	3.4x
Year	2017	2016	1.2x
	2018	2016	1.3x

https://trustandsafety.fun/

Types of Harms

Two types of harms

Two types of harms

Perspectival

Harm that occurs when target is exposed to it

- Examples include:
  - Trolling / Bullying
  - Personal attacks

Two types of harms

- Examples include:
  - Doxxing
  - Incitements of violence against an individual
  - Coordinated harassment campaigns

Global

Harm that happens whether the target is exposed or not

Criminal organizations using a platform to organize

Perspectival or Global?

Perspectival

Harm that occurs when target is exposed to it

Global

Harm that happens whether the target is exposed or not

Criminal organizations using a platform to organize

Perspectival or Global?

Perspectival

Global ✓

Harm that occurs when target is exposed to it

Harm that happens whether the target is exposed or not

Explicit threats of violence towards an individual?

Perspectival or Global?

Perspectival

Harm that occurs when target is exposed to it

Global

Harm that happens whether the target is exposed or not

Explicit threats of violence towards an individual?

Perspectival or Global?

[✓ Perspectival]

Harm that occurs when target is exposed to it

Global

Harm that happens whether the target is exposed or not

Doxxing

Perspectival or Global?

Perspectival

Harm that occurs when target is exposed to it

Global

Harm that happens whether the target is exposed or not

Doxxing

Perspectival or Global?

Perspectival

Harm that occurs when target is exposed to it

[✓ Global]

Harm that happens whether the target is exposed or not

Spreading conspiracies that are likely to lead to violence

Perspectival or Global?

Perspectival

Harm that occurs when target is exposed to it

Global

Harm that happens whether the target is exposed or not

Spreading conspiracies that are likely to lead to violence

Perspectival or Global?

✓ Perspectival

Harm that occurs when target is exposed to it

✓ Global

Harm that happens whether the target is exposed or not

Distribution of CSAM

Perspectival or Global?

Perspectival

Harm that occurs when target is exposed to it

Global

Harm that happens whether the target is exposed or not

Distribution of CSAM

Perspectival or Global?

Perspectival

Harm that occurs when target is exposed to it

[✓ Global]

Harm that happens whether the target is exposed or not

Taxonomies of abuse

Tagging taxonomies with our harm framing

Violent + Criminal Behavior

Regulated Goods and Services

Offensive Content

User Safety

Scaled Abuse

Deceptive + Fraudulent Behavior

Taxonomies of abuse

Tagging taxonomies with our harm framing

Violent + Criminal Behavior

mostly global

Regulated Goods and Services

Offensive Content

User Safety

Scaled Abuse

Deceptive + Fraudulent Behavior

Taxonomies of abuse

Tagging taxonomies with our harm framing

Violent + Criminal Behavior

mostly global

Regulated Goods and Services

mostly global

Offensive Content

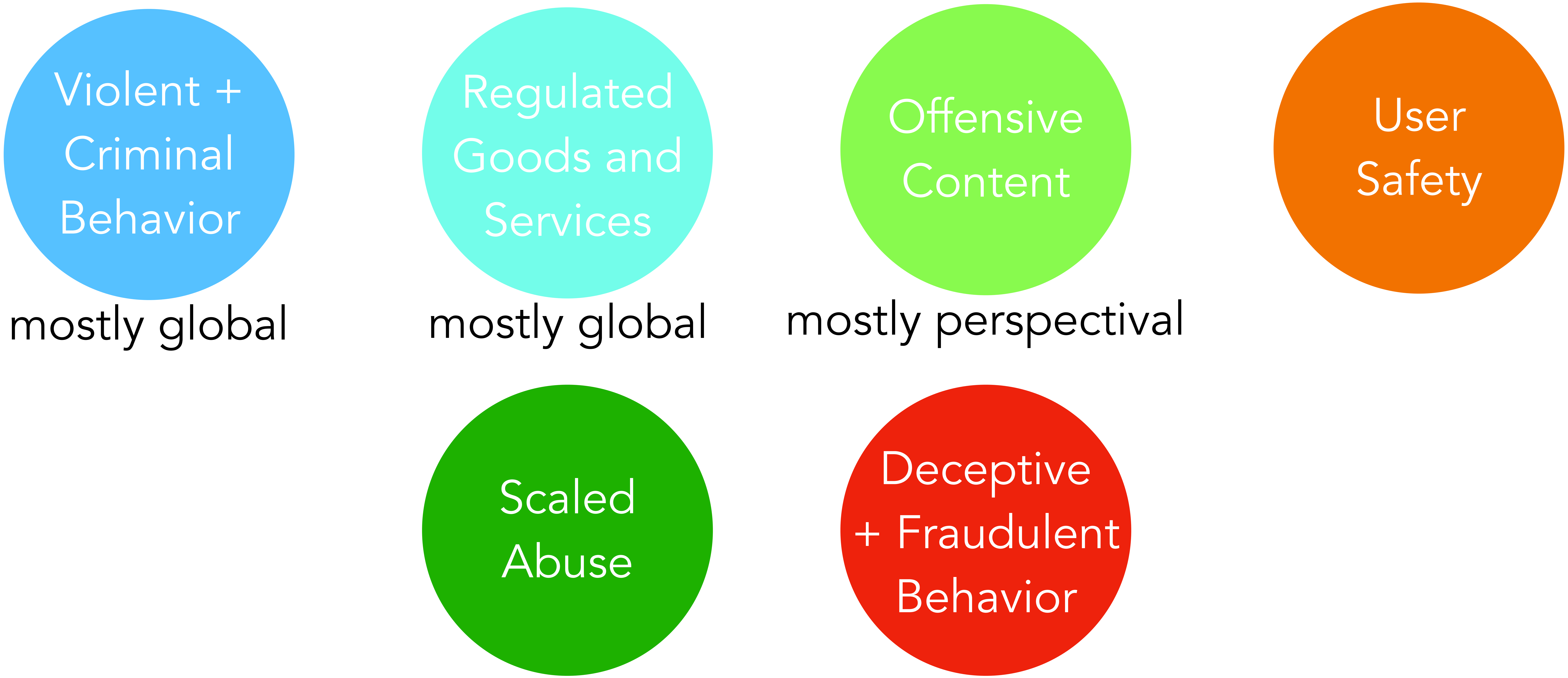
User Safety

Scaled Abuse

Deceptive + Fraudulent Behavior

Taxonomies of abuse

Tagging taxonomies with our harm framing



Violent + Criminal Behavior

mostly global

Regulated Goods and Services

mostly global

Offensive Content

mostly perspectival

User Safety

Scaled Abuse

Deceptive + Fraudulent Behavior

Taxonomies of abuse

Tagging taxonomies with our harm framing

Violent + Criminal Behavior

mostly global

Regulated Goods and Services

mostly global

Offensive Content

mostly perspectival

User Safety

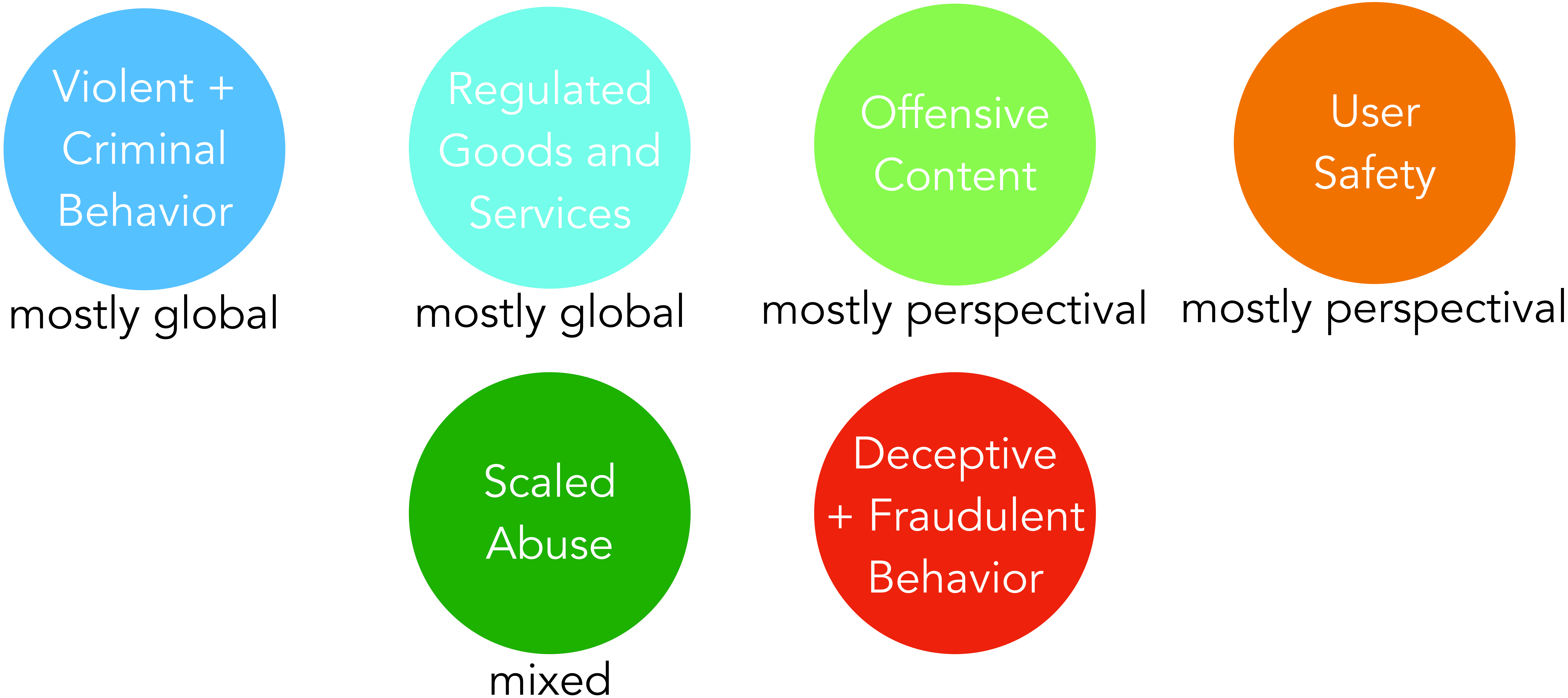
mostly perspectival

Scaled Abuse

Deceptive + Fraudulent Behavior

Taxonomies of abuse

Tagging taxonomies with our harm framing



Violent + Criminal Behavior

mostly global

Regulated Goods and Services

mostly global

Offensive Content

mostly perspectival

User Safety

mostly perspectival

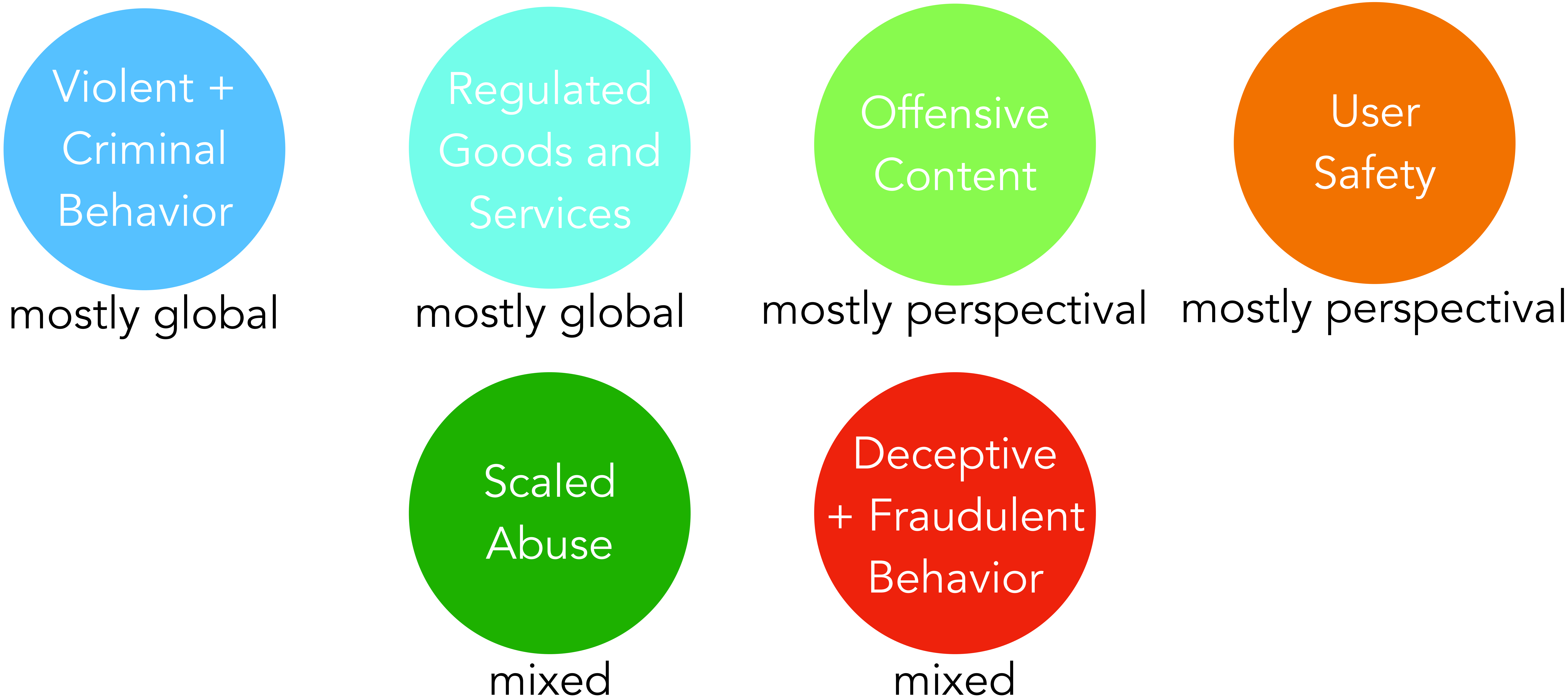
Scaled Abuse

mixed

Deceptive + Fraudulent Behavior

Taxonomies of abuse

Tagging taxonomies with our harm framing



Violent + Criminal Behavior

mostly global

Regulated Goods and Services

mostly global

Offensive Content

mostly perspectival

User Safety

mostly perspectival

Scaled Abuse

mixed

Deceptive + Fraudulent Behavior

mixed

How do we handle different types of harms?

How do we handle different types of harms?

- Global harms require global coordination
  - Community Rules
  - Automated tools / consistent moderation practices
  - ***Top-down approach***

How do we handle different types of harms?

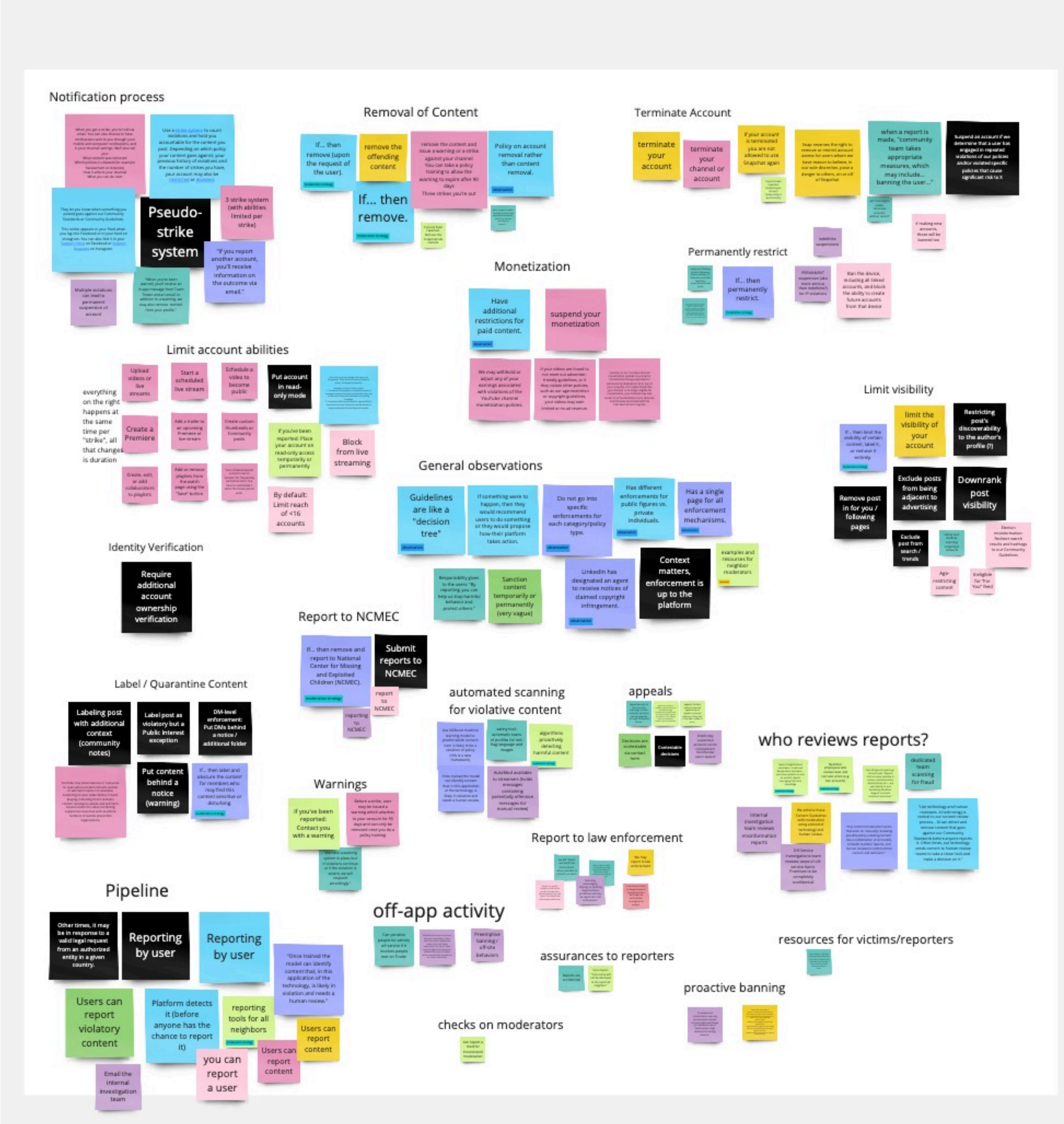
- Perspectival harms require *flexibility*
  - Lawful but awful speech categories are harder to moderate
  - *"User Control"*
  - ***Bottom-up approach***

- Global harms require global coordination
  - Community Rules
  - Automated tools / consistent moderation practices
  - ***Top-down approach***

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

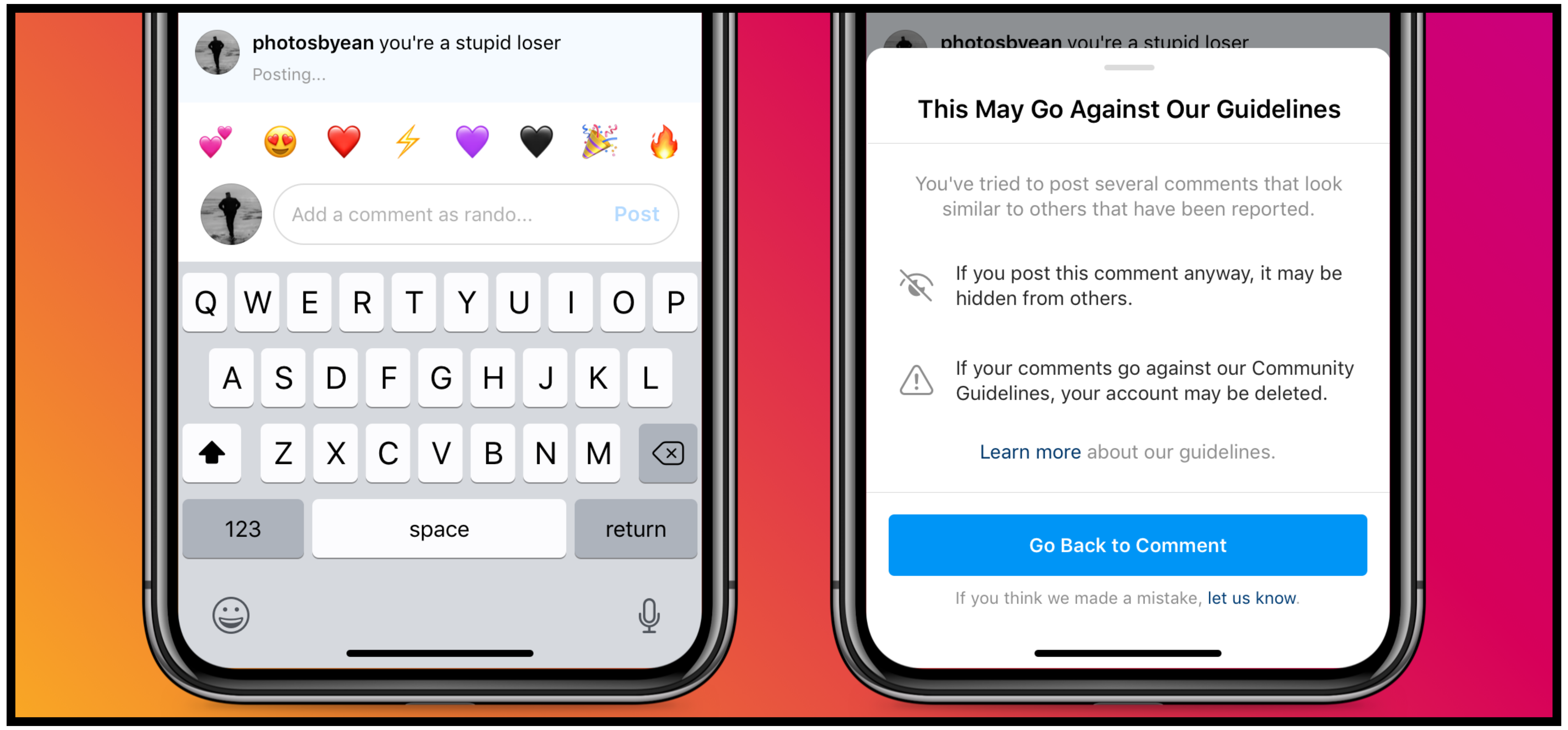
Actions that affect offending content

- 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



photosbyean you're a stupid loser

Posting...

photosbyean you're a stupid loser

This May Go Against Our Guidelines

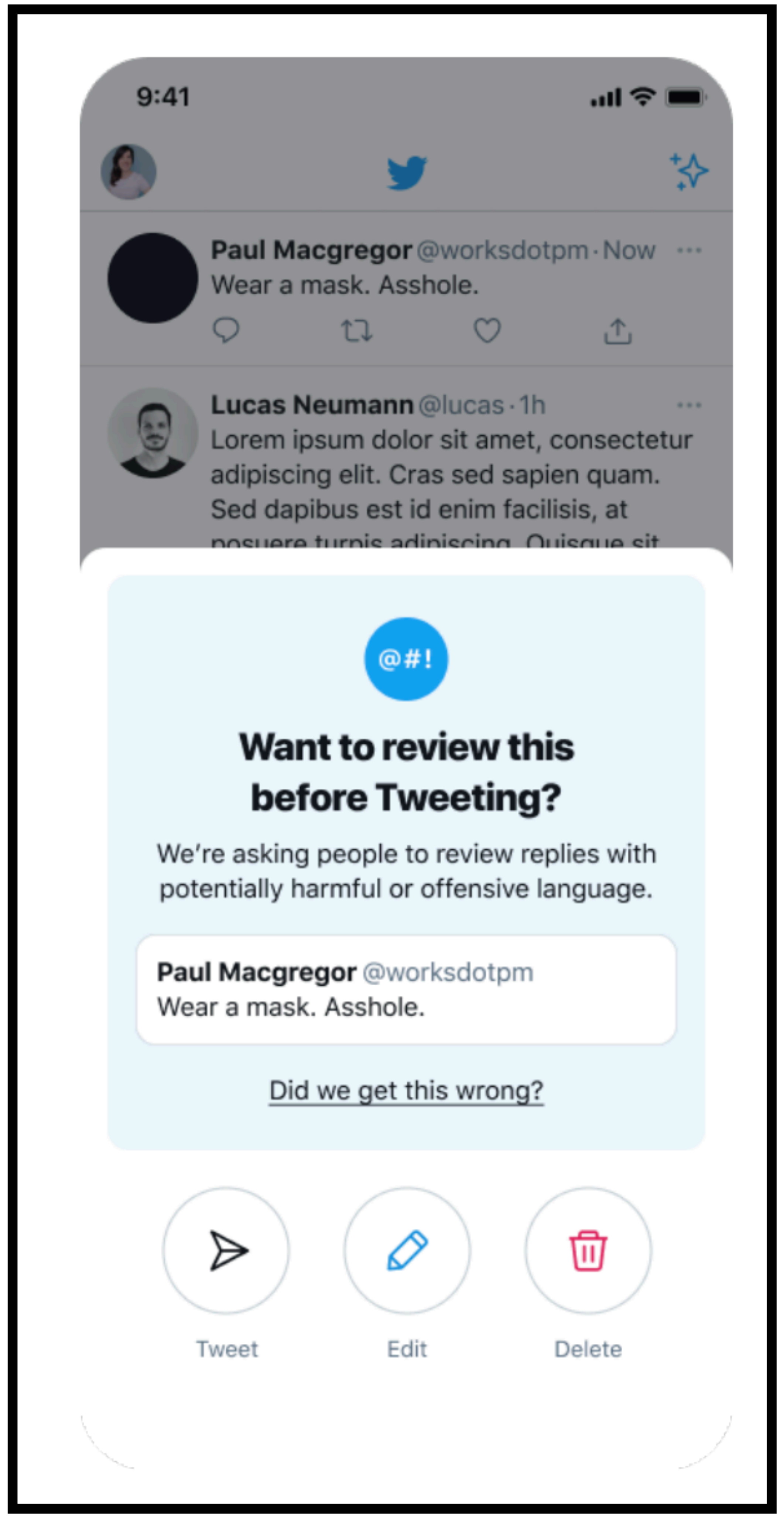
You've tried to post several comments that look similar to others that have been reported.

If you post this comment anyway, it may be hidden from others.

If your comments go against our Community Guidelines, your account may be deleted.

Learn more about our guidelines.

If you think we made a mistake, let us know.



Paul Macgregor @worksdotpm · Now
Wear a mask. Asshole.

Lucas Neumann @lucas · 1h
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Cras sed sapien quam. Sed dapibus est id enim facilisis, at posuere turpis adipiscing. Quisque sit

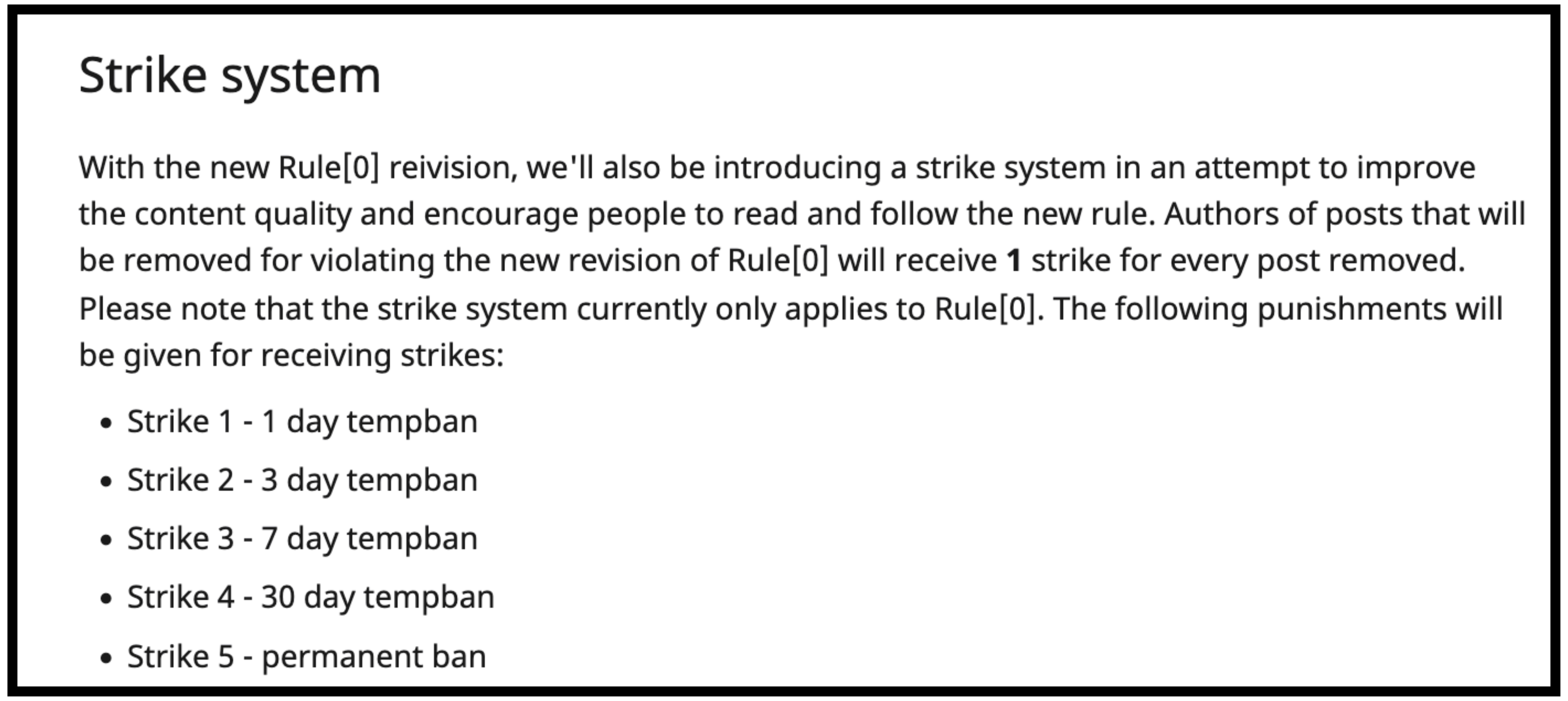
Want to review this before Tweeting?

We're asking people to review replies with potentially harmful or offensive language.

Paul Macgregor @worksdotpm
Wear a mask. Asshole.

Did we get this wrong?

Tweet
Edit
Delete



Strike system

With the new Rule[0] reivision, 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

Interventions are on the rise

- In paper: Reconsidering Tweets: Intervening during Tweet Creation Decreases Offensive Content, Katsaros et al. talk about the efficacy of interventions
- Study ran an RCT with 219,052 users on the platform that posted toxic content
  - 50% were in treatment, 50% were in control
  - Treatment users received a prompt saying "do you really want to tweet this?"
- Study found that 27% users revised tweets on first exposure, but limited effect over multiple exposures
- 3% edits Tweets were *more offensive* after prompt

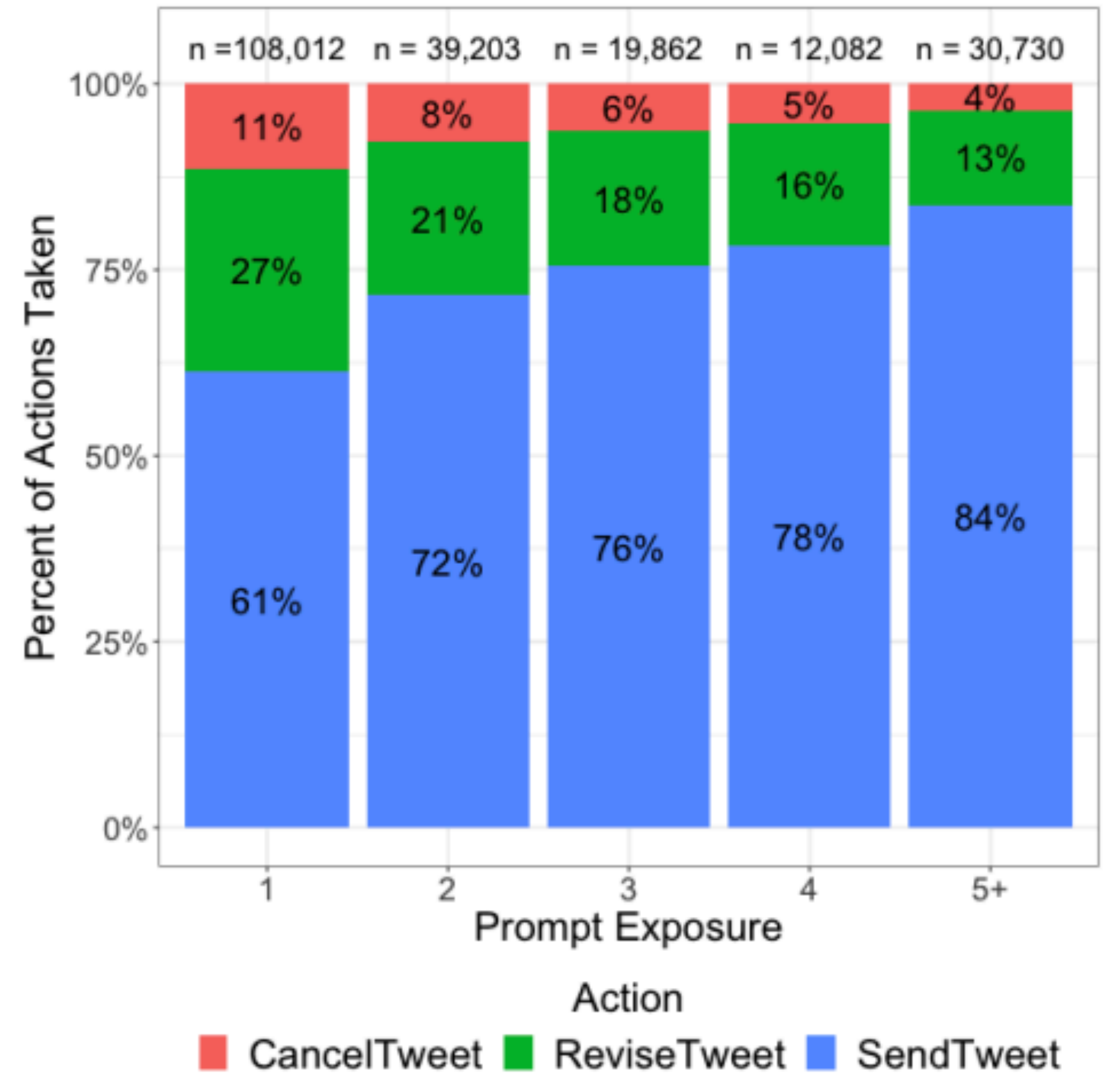


Figure 2: Distribution of Actions Taken by Prompt Number

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 enforcement mechanisms that you think might be useful in a Trust & Safety context**