CENTER FOR MEDIA @RISK

TRUST & SAFETY

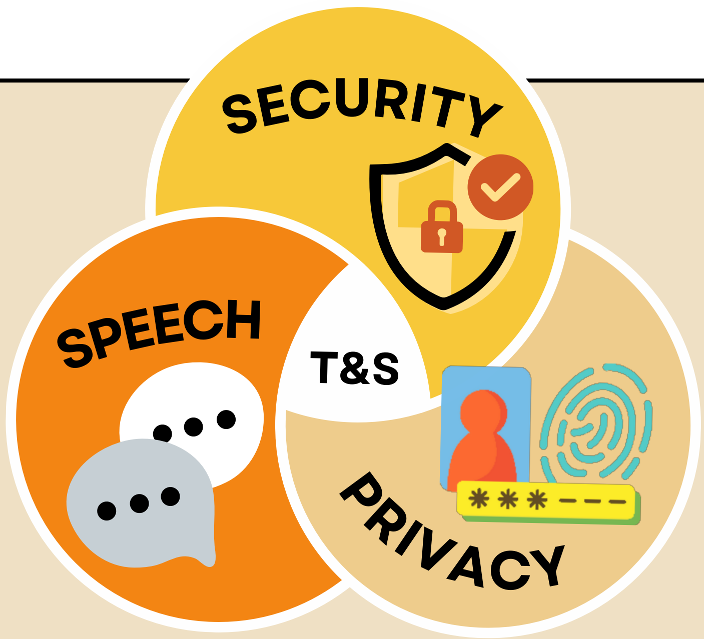
EXPLAINING THE INDUSTRY

DEFINITIONS & DEBATES

In a sentence...

Trust and Safety (T&S) is the precarious task of trying to balance **speech, security and privacy**, elements which can never truly be balanced

The industry grew out of a corporate attempt to **protect platform users from experiencing and witnessing harm**



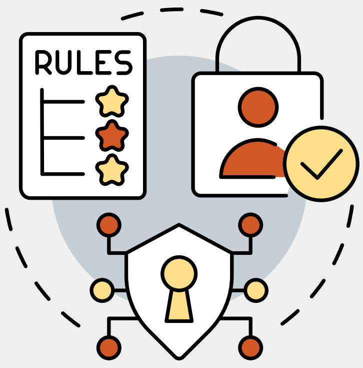
> *The professionalization of Trust & Safety inside the companies ... has also affirmed specific approaches to content moderation—reifying who counts as users, what registers as legitimate harm and what reads as a reasonable intervention.*
>
> *Tarleton Gillespie (2023)*

T&S teams construct a set of guidelines for:

- Interpreting platform policy
- Applying platform policy
- Justifying the policy's evidentiary standards

T&S is reactive: it's an effort to respond to events, controversies and scandals that ripple across platforms

POLICIES & PROBLEMS



Tech companies take a highly formalistic approach to T&S

When a policy is incorporated into a company's rules, it becomes an artifact that everything else must be evaluated against

Harm and hate are often understood through the lens of what's deemed **universally harmful**

So, exploitation of children is an area where there's general agreement, and many policies and institutions focus on preventing harm to minors online

Online harms often accumulate over time, which poses a challenge for business models

> *Trust, to some extent, is a perception, but its basis is safety. Only when people feel confident and comfortable about the safety of their presence and activities, in other words, there is no negative implication or loss to themselves, then they trust the platform and other people on the platform.*
>
> *Kenny Shi (2016)*

FRAMEWORKS & FUTURES

Diversify T&S teams:
Diverse T&S teams are essential to taking on the challenges of regulating a fundamentally vast and ungovernable thing: the internet

Understand T&S within the framework of capitalism:
Online platforms need advertisers to generate profits and survive; corporate sponsors value inclusive and safe environments for users

Learn from the past:
Position new technical problems–from content moderation to platform governance–in their historical lineage