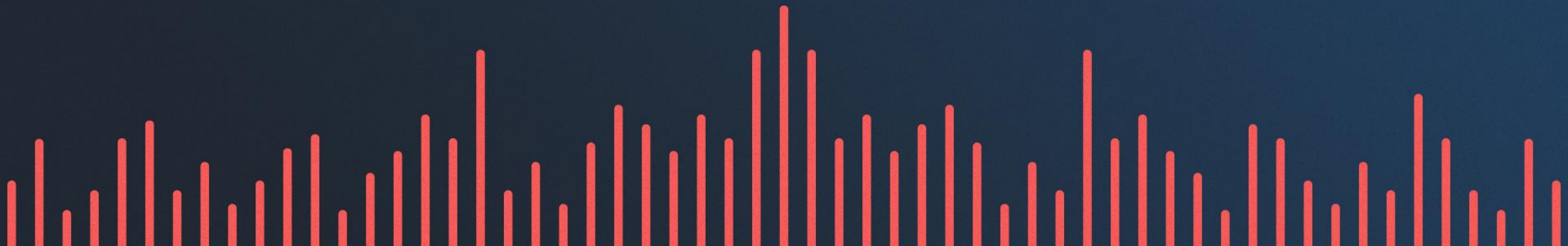
Challenges inhibiting platforms from effective moderation

Mike Pappas, CEO of Modulate

The Problem Statement

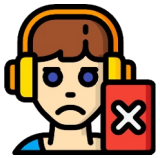


The Problem of Toxicity

77% of players face severe harassment in games (up 20% over last 3 years)

30% of players stop playing certain games due to toxicity

20% of players encounter white supremacism (more than 2x in 1 year!)[1]



320%

"A user who experiences toxicity is 320% more likely to quit."

Riot Games

326%

"Users who engage in chat are 326% more likely to keep playing after the first day."

Two Hat Security

[1]https://www.adl.org/resources/report/hate-no-game-hate-and-harassment-online-games-2022

Voice toxicity is most potent

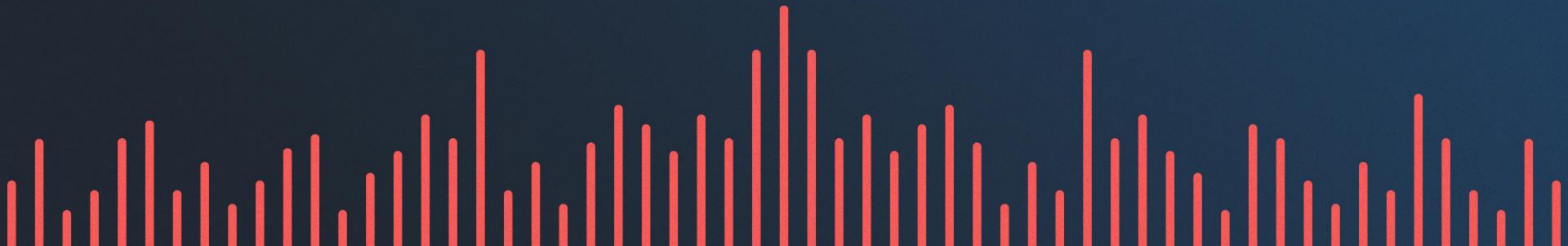
- Voice carries emotion[1] and builds[2] – or destroys – interpersonal bonds[3]

- Users expect text moderation – they have learned to circumvent it
  - Sometimes by d1$tört!on
  - Sometimes by being slightly less obvious in what they say

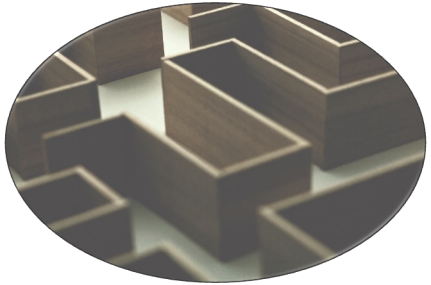
- Voice toxicity is more overt
  - Distorting your speech means nobody can tell you're being hateful
  - Since voice has been unmoderated, people are ***obvious*** and unconcerned with consequences

[1]https://academic.oup.com/hcr/article-abstract/33/4/427/4210727
[2]https://psycnet.apa.org/record/2017-43854-002
[3]https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1013&context=sighci2017

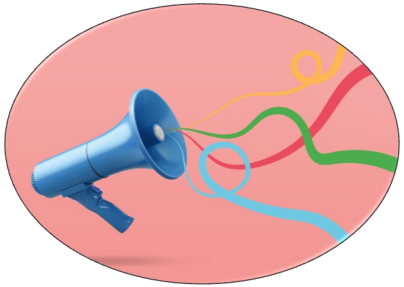
Why haven’t platforms done something about this?



Three Key Challenges



Moderation Is Hard



Voice Analysis Is Harder

Regulatory Uncertainty Favors Inaction

Moderation is hard

- What counts as toxicity?
  - The N-word is generally recognized as bad…except as a reclaimed slur, where it becomes essential to authentic interaction
  - Some players enjoy aggressive trash talk; others (especially the young and vulnerable) prefer not even to hear simple curses

- How do you find it?
  - ESRB rates games based on ***scripted content***…
  - …but online toxicity is user-generated and unpredictable.
  - Requires real-time analysis, not just up-front design + care.

Voice moderation is especially difficult

- Voice is expensive to process
  - Transcription at $0.40/hr, plus top platforms with 100M hrs/mo of chat…$0.5B/yr is too steep a price for nearly any platform to pay

- Voice is complex
  - Even if you could transcribe, you're losing emotion, nuance, and cadence

- Voice is identity
  - COPPA and others treat voice clips alone as PII, raising platform concerns about collection even when it's for a good reason

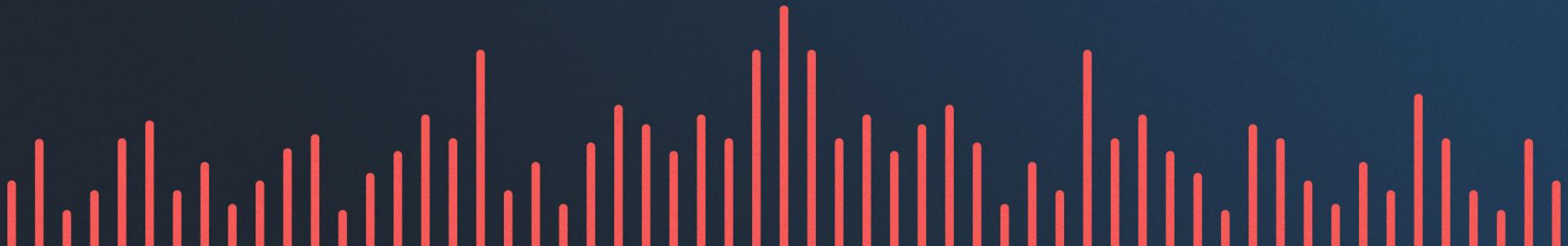
Navigating regulation is a nightmare

- Historically, privacy regulation has had more “teeth” than safety, leading platforms to prefer maximum privacy (no data collected) even when it comes at a cost of safety

- Risk of COPPA violations means this ***adds*** liability…
  - …though incoming regulation (EU’s DSA, CA’s AADA, etc) may shift the balance
  - …and Epic/FTC settlement emphasizes that COPPA is about ***protecting kids***, not merely limited to privacy

- Platforms fear sharing data about prevalence and impact, believing it will make them “stand out” as a target for consumer backlash or regulatory attention

So how do we solve this problem?



Removing chat is unacceptable

Gamers are actually ***more*** inclusive than non-gamers by ~20%[1]

Suicide risk among LGBTQ+ teens drops for those with an affirming online community[2]

Riot found that only 1% of players are consistently toxic - most players mean well and just sometimes make mistakes.[3]

[1]https://assets.evilgeniuses.gg/dei/EG_YouGov_GamingForAll.pdf
[2]https://www.thetrevorproject.org/survey-2021/?section=AffirmingSpaces
[3]https://www.gamesindustry.biz/articles/2018-03-13-toxic-players-and-the-power-of-positive-engagement

Player reports are insufficient

Only 8% of players ever report harassment.[1]

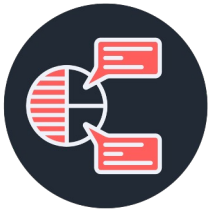
Insidious harms like child predation or radicalization are almost never reported – no knowing victim until it's far too late.

Riot acknowledges that only 10.8% (13M out of 120M) of reports from their players are actionable today across all Riot titles.[2]

Discord finds ~15% of harassment reports are actionable.[3]

[1]https://www.adl.org/media/15349/download
[2]https://www.riotgames.com/en/news/an-update-on-player-dynamics?previewId=et5cddv
[3]https://www.gamesindustry.biz/articles/2021-04-06-discord-releases-july-to-december-transparency-report

How ToxMod Works



1. Triage

ToxMod triages voice chat data to determine which conversations warrant investigation and analysis

2. Analyze

ToxMod analyzes the tone, context, and perceived intention of those filtered conversations using its advanced machine learning processes

3. Escalate

ToxMod escalates harmful language and severe toxicity directly to moderators so they can mitigate bad behavior

Triage: Identify which clips are high risk

- Don’t wait for a user report; actually watch for risky behavior
- But how does this impact privacy?

Triage: Identify which clips are high risk

- Don’t wait for a user report; actually watch for risky behavior

- But how does this impact privacy?
  - ***We know how to strike this balance in the physical world***

Analyze: Look for *harm*, not keywords or phrases

4

Will you quit ******* feeding[1] - We're losing this game because of YOU![2]

Stop yelling at me[3]! I'm leading the team in kills right now!

[1] Transcription Analysis

[2] Prosody Analysis

[3] Conversational Context

[4] Player Relationships

Escalate: Prioritize the worst of the worst

- Let each platform set priorities by Code of Conduct
- Platforms engage directly with users, now backed with clear evidence and context regarding the offense
- Typically finds 10x more than player reports – and **>98% of ToxMod reports are actionable**

Summary

- Online toxicity is a real problem impacting children and other vulnerable demographics
- We're not just talking about friendly trash talk – white supremacy and other extreme ideologies are on the rise
- Platforms aren't ignoring this out of greed. They are facing serious technical, logistical, and regulatory challenges that make answers non-obvious.
- **Solving this problem requires a compromise between safety and privacy to reach true wellbeing.**

Thank you for listening

Any questions?