Reporting Metasystem Design and Penalization Strategy Best Practices

Fair Play Alliance

GAME DEVELOPERS CONFERENCE
MARCH 18–22, 2019 | #GDC19
Intel & Spirit AI

George Kennedy III
Technical Lead
Intel

Peter Alau
Director of Business Development
Spirit AI
Introduction

- All games of the future will be voice chat based.
- We’re working to leap frog the technology so that harassment we see in text chat isn't as prevalent in voice chat.
- We want to make sure that voice chat is manageable in the future.
- It’s difficult to identify (comic-con vs comma-con)
- Voice chat adds a new vectors (pitch, loudness, tone)
## Multi-Pronged Approach Required

<table>
<thead>
<tr>
<th>Detection</th>
<th>Measurement</th>
<th>Policies &amp; Codes of Conduct</th>
<th>Behavior Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Focus on machine learning (ML) approaches to detecting behavior within platforms.</td>
<td>Build on existing models with community and industry partners to define and test effective measurement and severity ratings.</td>
<td>Working with community moderators, platforms, and researchers to define and test best practices for developing, deploying, and enforcing codes of conduct.</td>
<td>Leveraging social science research to create and test strategies for culture change – exploring aspects of communication, timing, incentives, and enforcement.</td>
</tr>
<tr>
<td>Explore both client based/personalized and game developer side deployments.</td>
<td>Incorporate relevant metrics into ML solutions.</td>
<td>Working with codes of conduct to customize ML solutions.</td>
<td>Incorporating results into policies, metrics, and ML solutions.</td>
</tr>
</tbody>
</table>

**Tech Research Focus**
Why bother?

Racism, misogyny, death threats: Why can’t the booming video-game industry curb toxicity?

Right-Wing Hate Groups Are Recruiting Video Gamers

November 5, 2018 • 10:37 AM ET
Heard on All Things Considered
Why is this hard?

- Sentence fragments give misleading indicators.
  - See: Self Harm
  - See Also: Consensual conversation
- Synonyms and Homonyms
- Changing Language to avoid detection
Simplest ID method

Analyze reactions and relationships
Weight for guilds and other affinity groups
Consider non-reactions as reactions
Look for unexpected trailing indicators

“Why did you ask that?”
“Does that work for you”
What are/aren’t we detecting

- Community Sentiment
- Friendly Behavior
- Harassment and Bullying
- Hate Speech
- Grooming
- Bot Detection
- Fraud Detection
- Self Harm
- Unsolicited Content

- Roleplaying
- Sarcasm
  - humans have the same problem
NLP in 30 Second.
Outcomes from ML

- Measurable reactions from policy changes
- Easier identification of your best and worst
- Problem Severity
- Community standards mostly created by the community
Adding Voice Detection

Background:
Building functions that maps inputs to outputs
\[ y = f(x), \] solve for \( f \)

\( f \) can be applied to never before seen \( x \) and predict \( y \).
Without machine learning, engineers write \( f \) for every case of \( x \) and \( y \).

Audio Challenges

Audio challenges:
• Not all content is going to be recorded directly from the user’s mic.
• Most reports that require moderator action will be recorded from a “3rd person” perspective.
  • Will simultaneously have in game audio, simultaneous speakers.
  • Variation in recording quality.
• Some language is unique to games
  • Characters
  • Abilities

Easy to translate: - low noise

Hard to translate: - high noise, multiple audio sources
Detection Solution Architecture

How detection tools interact with a game.

Lessons Learned:

| Results are best if labelers are from the community in which the ML model will be used. | Inter-annotator agreement and model accuracy is improved. |
| Selected outcomes should be decided by the community moderators. | Community moderators have most insight into right incentive mechanisms. |

Why:
Deployment Models

User in Game Audio

Client Platform
Data Processing

Cloud Platform
Behavior Detection

User is responsible for the cost of compute.

Platform owner is responsible for the cost of compute.

Behavior detection consumes CPU, memory, GPU, and networking resources (varies w/ implementation)

Overall Benefits:
- Voice and text detection
- Data center deployment can support any platform
- Workload optimization on server

Benefits of Client Offload:
- Reduced costs of detection because it’s running on user owned compute
- Less traffic sent over network
- Workload optimization on client
Thank you!