

Building Blocks of Policy Development

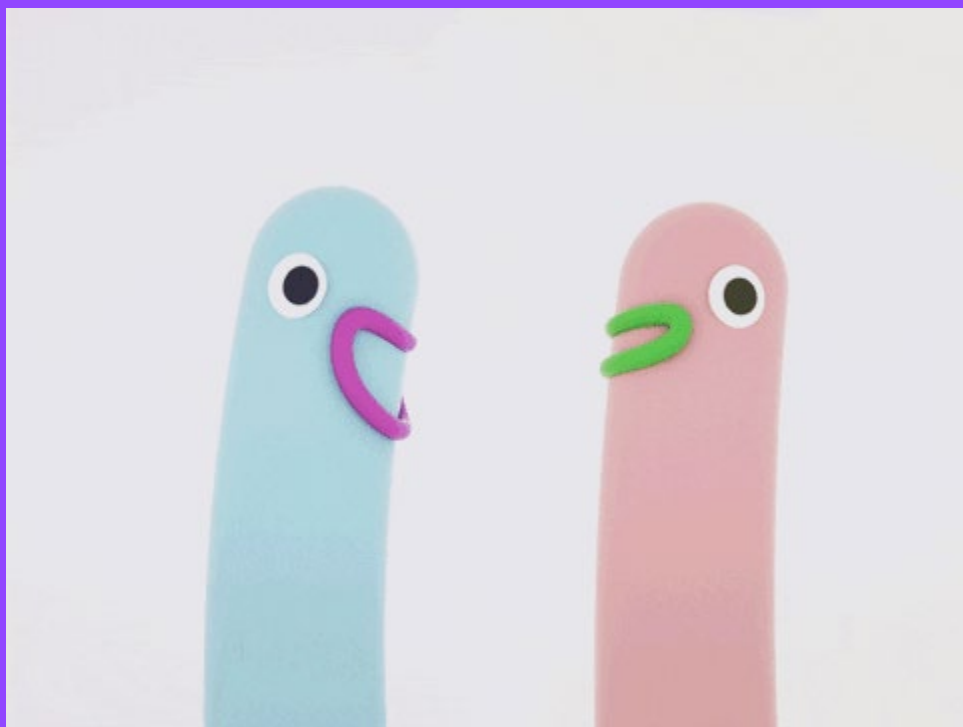
Connie Chung and Robert Lewington
Twitch Trust and Safety



Goals

1. Understand different stages of policy development
2. Know what your community needs next
3. See how to use principles from the FPA Disruption and Harms in Online Gaming Framework









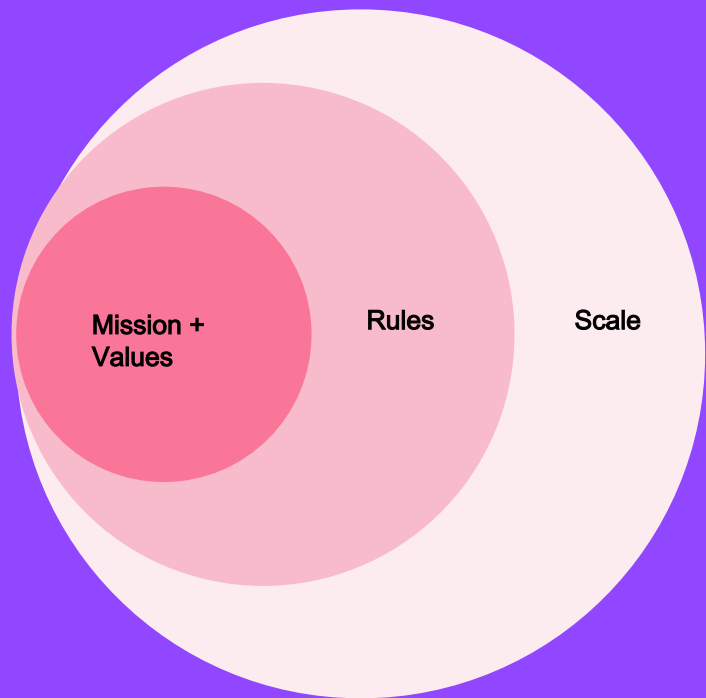
1.



2.



STAGES OF POLICY DEVELOPMENT



1. What community do you want?



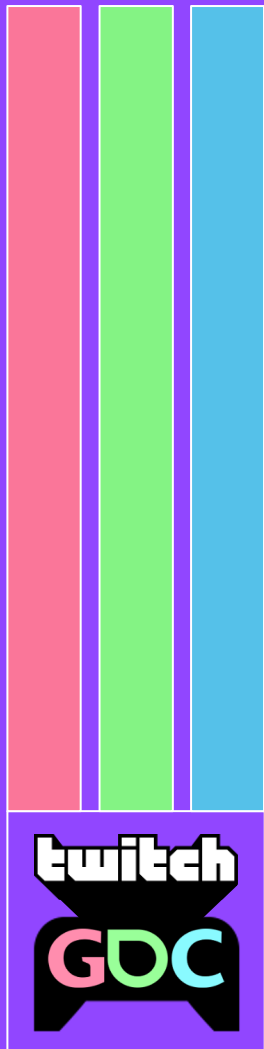
2. How do you translate norms to rules?



3. How do you scale your rules and values?



1. What Community Do You Want?



Company Missions






Empower communities to create together.

We serve the public conversation.



Put smiles on the faces of everyone we touch.

What community do you want?

Company			
Mission	We serve the public conversation	Empower communities to create together.	Put smiles on the faces of everyone we touch
Politics	When content is “in the public interest,” make an exception on enforcement + label	No public interest exception for politicians	Animal Crossing: New Horizons asks people to “refrain from bringing politics into the game”

2. How do you translate norms to r





Initial Considerations

1. Policy goals

1. Current and desired community needs

1. Comparable communities

Considerations While Building

1. Clarity and consistency

1. International norms

1. Trade-offs

NO DONUTS

Captures



Excludes

NO SWEET, FRIED DOUGH CAKE

Captures



Excludes



NO SWEET, FRIED DOUGH CAKES SHARE

Captures



Excludes



3. How do you scale your rules and



Charting the Problem vs Disruptive Behavior Framework



Diagram 3. An overview of how disruptive behaviour can be expressed.

Understanding the Problem Space

1.



Community

Conducted extensive interviews with a diverse range of streamers:
What constitutes harassment from them? How does it manifest? Etc.

2.



Academic

Consulted with academics and civil society experts.
Free speech, inclusivity in gaming, online abuse, ethnic studies, and women's studies.

3.

Safety Advisory Council

Alex Holmes • CohlhCarnage •
Cupahnoodle • Emma Llanso •
FerociouslySteph • Dr. Sameer Hinduja •
T.L. Taylor • Zizaran •

Advisory

Leveraged our Safety Advisory Council to assess/test drafted policies
Creators, academics and experts on abuse types such as bullying

4.



Data & Testing

Modelled and tested the impact of our proposed policy changes
Operational metrics, consistency of application

FINDINGS

- Our community sees hate and harassment as intertwined and streamers are worried about positive as opposed to punishing norms
- Academics we consulted emphasized the importance of clarity in standards, exceptions
- Themes ID'ed from cases included wanted sexual advances



Defining the Solution Strategy Key Points

1.

twitch

© 2019 Twitch Interactive, Inc.

Impact over 'intent'

- “**Words and actions have meaning** **and impact** ...even if the target of your behavior or comments isn't bothered by them, others in the community may”
- Under the new policy, our Safety team look at the **content** of statements or actions **rather than** relying solely on perceived **intent**.



Disruption and Harms in Online Gaming Framework

“It is important not to conflate why a player does something disruptive or harmful (**their intent**) in an online game space with the impact the action has on a targeted player.”

“...we refer less to the intended game experience or **the intentions of players** who are attempting to disrupt it, and more to the harm as it is experienced by the affected players. In other words, the focus here should be on **im pact m ore than intention.**”



Defining the Solution Strategy The Key Points

2.

twitch

Hateful Conduct

Clarity

- **More detail** - e.g. hate speech and symbols always prohibited, and we've added explicit language banning hate groups, membership in hate groups, and sharing of hate group propaganda.
- Specifically prohibited black/brown/yellow/red face unless they are being used in an explicitly educational context.
- Again, **not new**, but the new guidelines make the standard clearer for everyone.

Protected classes

- We've added caste, color, and immigration status to this list to ensure we are evolving with our global community and **providing sufficient protection for under-represented groups.**



Disruption and Harms in Online Gaming Framework

“Note: Some conduct can be misinterpreted as abuse, but can stem from a **misunderstanding of the...rules or expectations.**”

“A player’s **understanding** of what is expected of them and their peers in a game space **influences their behaviour**”



Defining the Solution Strategy The Key Points

3.

Harassment



Harassment + Sexual Harassment

- Sexual harassment was always prohibited, but we heard **from our community** that the guidelines didn't fully cover this.
- Separated sexual harassment into its own category + **much lower tolerance** for objectifying or harassing behavior inc.:

“

- *Repeatedly commenting on someone's perceived attractiveness, even in what you believe to be a positive or complimentary manner, is prohibited if there is indication that it's unwelcome (i.e. you've been asked to stop, timed -out, or channel -banned)*
- *Lewd or explicit comments about anyone's sexuality or physical appearance is prohibited.*
 - *Note that we do not make an exception for public figures*
- *Sending unwanted/unsolicited links to nude images or videos is prohibited*

”



Disruption and Harms in Online Gaming Framework

“ADL's 2020 survey Free to Play? Hate, Harassment and Positive Social Experiences in Online Games: found that 81 percent of adult online gamers in the U.S. experienced harassment in online games, an increase from 74 percent in the 2019 edition of the survey.

“Sixty-eight percent experienced severe harassment, such as physical threats, identity-based discrimination, sustained harassment, **sexual harassment** and stalking.”



Measuring Impact

QUANTITATIVE



>700%
More
enforcements
Harassment

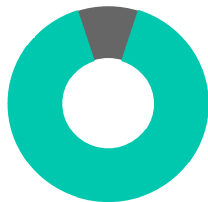
Of which
>30% are
Sexual Harassment
Enforcements

>45%
More
enforcements
Hateful
Conduct

QUALITATIVE



Report Validity



Harassment

+95%



Hate

+31%



North America

Harassment
+3.7%
Vs. market size

Hate
+4.7%
Vs. market size

EMEA

Harassment
+2.9%
Vs. market size

Hate
+14%
Vs. market size

APAC

Harassment
-3.9%
Vs. market size

Hate
-5.7%
Vs. market size

Latin America

Harassment
-2.8%
Vs. market size

Hate
-13%
Vs. market size

30%* of Hateful Conduct

*based on sampling



Enforcements

Were from streamers playing
FPS/Battle Royale

What we learn about global community?



Thank you!!!

