Accountability Content Evaluation System (ACES)

Feasibility Report

Cato Cannizzo, Lara Hattatoglu, Anna Kucinski, Jesse Walling

HCDE 498 / 598: Designing Trustworthy Information Systems

Dr. Kate Starbird

December 15, 2021

Table of Contents

Introduction

With the rise of disinformation in online spaces across the world, from organized misinformation attacks, to taking part in annexations of entire regions (Sinan. 2021) to the prevalence and rise of anti-information groups like anti-vaxxers on Facebook, modern systems of information flow need to be changed. Our accountability content evaluation system (ACES) seeks to improve content evaluation methods. Current content evaluation methods are all flawed, while some work better than others, the current spread of mis- and disinformation shows that these systems all need improvement. Our approach involves platform design that incentivizes users to engage in community-based content moderation.

Literature Review

Prior Work

Facebook's Moderation

Facebook employs the most standard and historic approach to content moderation. They use both algorithmic tools and community-based reviews to remove content that violates community standards. Automated tools are generally used as a first layer of review, detecting harmful images and videos,and prioritizing flagged content. Before a user's content gets published on Facebook, it goes through an automated screening process that uses digital hashes of posted material to proactively identify and block content that

matches existing hashes in the Child Sexual Abuse Material (CSAM) database or is already established as terrorism-related. (“Everything In Moderation,” 2021). Content already posted on the platform is continuously screened and evaluated through AI tools. AI is a fast, low cost and scalable method of content moderation, however, it comes with its limitations involving text. Text can be difficult to evaluate because context is needed to identify whether a specific indicator is being used in a violating manner. AI attempts to provide context by considering other factors to the post such as the profile of the poster, comments, shares, and likes. However, AI is often unable to determine whether a content violates community standards. This warrants a human review.

Facebook’s community based review relies on both users, content moderators, and their oversight board. Users flag inappropriate content on their personal feeds. Flagged posts that don’t get picked up by AI are sent in to one of the 15,000 content moderators hired by Facebook. The oversight board, which consists of 40 members from diverse disciplines and backgrounds, acts as the supreme court of Facebook and oversees appeal cases.

In regards to incentives, it’s difficult to establish exactly what the motivations might be for participating in content moderation. For users, they could be emotionally distraught by a post, or altruistically flag it because they think someone else would react negatively. Companies attempt to frame peer produced content moderation as an act of good citizenship in order to motivate users to moderate posts (Schwarz, Ori. 2019). Overall, there are a few different incentives for moderation along this pipeline - altruism, salary, or

interest. Some people want to make the world a better place, others are paid to moderate, and some are interested in the moderation process.

Reddit's Incentives

In 2020, Reddit implemented a cryptocurrency-backed system to replace karma in select subreddits, namely /r/CryptoCurrency and /r/Fortnite to start. From there it has grown to other communities such as /r/ethtrader. The amount of tokens you have is determined by the upvotes and downvotes leading up to the weekly distributions of subreddit-specific tokens like $MOONS for users of /r/CryptoCurrency (Cochran, 2020). Effectively, these tokens act as a share of ownership in the subreddit.

The more Reddit Community Points you have, the more weight your votes cast count for in polls. This includes improvement proposals for subreddit-specific tokens and how the newfound monetary policy of the subreddit unfolds. Normally, vote weighting would be viewed as a pay to win system where it's possible to rule a subreddit if you have enough money, but any tokens bought with FIAT currency rather than direct subreddit participation do not count towards influencing the weight of votes. As a result, the most active and positively impacting users have the most influence.

Twitter's Birdwatch

In 2021, Twitter implemented a new content management system: Birdwatch. This crowdsourced approach seeks to inform the users instead of moderating posts. The system operates as follows:

> A user posts a tweet containing claims that may be false or taken out of context;
>
> Other users add 'notes' to these tweets to provide context or counter claims;
>
> Further users upvote or downvote notes, selecting a reason from a list of options;
>
> Notes are ranked according to votes. (Benjamin. 2021)

A key aspect of this is that not all users can add notes to a tweet. Only 'diverse' accounts selected by twitter themselves can start this system by adding a note. Research suggests that this management system is effective so far. Users of the system are often providing trustworthy links and are avoiding inflammatory language (Nicolas Pröllochs, 2021). This platform is similar in goal to ACES, but differs slightly in concept.

TikTok Moderation

ACES is centered around the platform design of TikTok. TikTok is a video streaming platform where users can consume content through a personalized feed called the "For You Page." This feed is driven by a powerful algorithm that tracks user's previous interactions with other content and makes recommendations based on that data.

Currently, TikTok handles its content moderation through a mixed approach of both technology and human moderators. Content that goes against community guidelines are removed and repeat offenders are banned from the application. For some content - such as videos under review or videos that may be upsetting, TikTok reduces the discoverability of the video. Similar to Facebook, content moderators are hired from a third party source. Users are able to moderate their own feeds by reporting videos they believe violate community guidelines (Tik Tok Community Guidelines, 2021).

Misinformation on TikTok poses a few interesting challenges. First and foremost, misinformation of videos and audios are more difficult to capture in comparison to text. Furthermore, TikTok has a function where users can use audios from other videos and react to them/use them for their own videos. One case study observed how this function is being used to spread misinformation on TikTok. In December, a man under the username alpha_k9 uploaded a video spewing anti-vaccine rhetoric. The video's audio began to be re-used by hundreds of users and soon enough garnered around 4,500 videos that reacted to the audio and were viewed around 16 million times. TikTok's audio sharing feature has enabled misinformation to spread at a faster rate than usual, creating a feedback loop of anti-vaccination rhetoric. When these audio tracks become viral, the original content and claims attached to it become hidden making it more difficult for TikTok to remove and stop the spread of misinformation (Zadronzy, 2021). Our solution will take these unique circumstances into account.

Community Moderation

A key design of our system is community moderation. Literature has shown that community moderators are seldom as accurate as professional moderators. One study focusing on peer grading found that peer graders had three times the level of variance compared to instructors grading (Kulkarni C. Et al. 2013). However, that same article established that peer grading got better over time and dropped to within 10% of instructor grades with minor revisions to their program. For ACES this means that community moderation is feasible, provided proper implementation. We envision this implementation to be a focus on binary metrics and a focus on increasing perspective rather than trying to give posts a quantitative trustworthiness score. This would help our volunteers overcome high levels of variance by focusing on questions where ratings are not required. The same article by Kulkarni C. found that the students remarked peer grading "was an extremely valuable learning activity." ACES passively capitalizes on this finding to combat misinformation by adding a platform based way for the community to evaluate their peers on trustworthiness. The idea here is the more users that critically evaluate the trustworthiness of posts the more those users will further their own online media literacy.

Another notable example of early online community moderation is how AOL utilized volunteers to run their social communication methods like bulletin boards and chat rooms, enforce their Terms of Service, create automated content, and manage other volunteers. Compensation for this work was credit hours taken off of their bill or a free "overhead account." (Postigo H., 2009). Postigo deems this to be a successful way of moderation, but

the system was ultimately dismantled as work-consciousness grew and people found the uncompensated work unacceptable. The Department of Labor (DOL) conducted a study into the system, and an interviewee noted that “90% of what paying AOL users” engaged with was maintained by AOL volunteers (Postigo H., 2009). The DOL found that people were motivated by the credit hours and the personal relationships they had formed through the service. Many of the factors involved with volunteering lead users to believe they were employees, and should be compensated more fairly (Postigo H., 2009). Key points of contention that ACES differs from in its moderation system are increased transparency, non-compulsory actions, and decreased power of the individual. Regarding transparency, AOL notably had their volunteers sign a non-disclosure agreement for what was going on in the system (Postigo H., 2009). This may have fostered secrecy and fear in their volunteers. ACES does not encourage usage of the moderation system to the degree where users feel compelled to participate, as the rewards do not affect a person’s livelihood in a strong way. Further, their individual actions are not large tasks to complete. However, AOL’s volunteer system provided a major incentive, of heavily discounted internet for taking on major tasks, which were monitored by volunteer managers. When AOL lowered the monetary incentive significantly, many were discouraged from participating. Additionally, AOL’s system placed a considerable amount of power on each individual volunteer, which led to imbalances of power and the reliance on these volunteers to maintain the platform (Postigo H., 2009). ACES implements moderation through minimal contribution from each user, minimizing the possibility of one user abusing power.

Incentives

Our approach relies on user participation in content evaluation, thus presenting a challenge of incentivization. Currently, the majority of social media platforms rely on user's altruistic and psychological tendencies to participate in content moderation. There are no formal incentives given to users to encourage them to flag and report posts on their feed. Our approach aims to create alternative incentives in order to encourage users to opt into our system. Our incentives were derived from preliminary research of human psychology and engagement.

Social Validation

The need to be liked and belonged by others is a fairly common and widespread psychological phenomenon. Many social media applications take advantage of this need for validation in order to engage users with their applications. Snapchat uses streaks to make users feel popular and more likable. Instagram takes advantage of this innate desire by letting users like and comment on one another's posts. When an individual receives a lot of likes and comments on their posts, they feel more liked and socially validated. (Yvette, 2021) This validation can be addictive since every time an individual receives that positive social feedback, their brain releases dopamine. Dopamine is what is known as our "pleasure chemical" and is stimulated by unpredictability and reward cues (Gamble, 2019). It rewards us for behavior and motivates us to repeat them. Every time a stimulus releases dopamine, the pathways between the action and our reward center become stronger. This is how addiction begins. Many social media platforms hack user's psychology to keep them

engaged with their application and it works.(Haynes, 2018). One study estimates that over 210 million worldwide people suffer from social media addiction. (Truelist, 2021) Another article claims that it is more difficult to quit alcohol and cigarettes than to quit social media (Gamble, 2019). Proving that social validation is a strong incentive to engagement.

Gamification

The term gamification refers to applying game-like mechanisms to non-game contexts. Gamification has been used as a method to increase motivation and engagement across a wide range of disciplines, such as education and business organizations. Gamification is engaging due to multiple different reasons. When playing a game, there is often an extrinsic motivation to receive rewards or accomplish a task/goal. Extrinsic motivators can be points, milestones, rewards, notifications, and achievements. Gamification also plays into our intrinsic motivation to achieve something. Intrinsic motivators can include "relationships such as competition, collaboration and community feeling, the feeling of accomplishment through progress, achievement and collection, empowerment through autonomy and feedback, unpredictability through surprise, exploration and scarcity and lastly constraints through scarcity, loss and avoidance." (How Does Gamification Drive Engagement, 2021).

Design Preview

The primary solution space we chose was moderation, with a focus on user input and consensus. We deemed moderation essential for social platforms, namely because it can

protect against direct harm. A key example of this was early on in the COVID-19 pandemic, where drinking bleach was interpreted to protect against the disease as mentioned in Sarah Shirazyan's guest lecture. So, one of our main motivating factors was the wellbeing of others. This is both directly impacted, like the bleach example, and indirectly impacted - through anti-vaccination measures or similar as discussed earlier. After looking at the moderation space, we wanted to extend the moderation process to users and empower them to stay informed about the content they are seeing. Ultimately, we reached this conclusion parallel to the discussion around media literacy from Danah Boyd. This deviated from our original plans of full community content moderation and pivoted slightly to media literacy with an option to elevate the content to TikTok for moderation if deemed appropriate.

Rethinking Centralized Moderation

Currently, social media platforms follow a centralized moderation process. Many use third party content moderators, sometimes paired with AI moderation tactics. With such a large amount of information posted per day, this presents a couple problems, and opportunities for improvement.

Increasing Transparency

First, the current moderation process reduces transparency for the users. On TikTok, users typically either see a video or don't see one. There isn't any information around the validity, perspectives, or sentiment of content.

When bias from filter bubbles or echo chambers take hold, we believe it's important to at the very least let people know of the contentious content the algorithm has presented to them. This is where ACES consensus comes in, establishing broad differences in perspective on a specific piece of content. ACES isn't meant to be a fact checking framework, instead it acts as a basis of epistemological media literacy.

Multi-Dimensional Moderation

It is impossible to know how many posts Facebook moderates in a given day, but it's very likely a large amount due to the volume of users posting on the platform. Currently, their third party moderators are one of the primary sources of moderation, much like TikTok. These employees typically evaluate posts on the basis of violating platform policies, not on the content or sentiment. Facebook has made some progress in the media literacy space by providing labels, but TikTok is lacking in this field. Instead they solely rely on AI and third party moderators going through posts. We ask why not let user evaluators do some of the work too? Similar to how third party employees moderate posts, it is entirely possible that an AI method could be adopted to assist ACES through detecting posts gaining irregularly large traction, more cross-topic cluster spreading, or more.

Discussing Consensus

Consensus does not have to be one side or the other. This is the primary reason we decided against a pure moderation standpoint; it would be very difficult to agree on group moderation without a specific sub-group being unhappy. So, what exactly does consensus

mean then? To us, it's the process of gathering information around what people think about the content and informing others. In a way, it's shared epistemological media literacy. As a result, we opted for the stages system, evaluating sentiment at every step along the way. This is also why we pivoted away from a purely moderation perspective to include media literacy.

Stakeholders

TikTok (Platform Owner)

With third party moderators upholding policy enforcement and a growing need for labeling and media literacy in the social media space, ACES would positively impact TikTok. Assuming TikTok takes the first step of acknowledging the importance of media literacy, they need a way to implement it. Hiring more third party moderators is an option, but that would limit perspectives and interpretations to one person rather than many. Implementing ACES would not undermine content policies either, and in fact would help TikTok establish a community perspective around whether or not to moderate a given post.

Opt-in Evaluators

The opt-in evaluators are incentivized to watch content and provide input at different stages of our system. This process comes with an inherent risk of seeing offensive or disturbing material, in which case they would be updated on the progress and conclusion of moderation from TikTok. It is unknown exactly how effective AI moderation is, but this

has the potential to reduce the mental strain on opt-in evaluators as well. The opt-in evaluators want to earn rewards for participating in the TikTok community process.

TikTok Users

ACES provides transparency and understanding of content to general users. The only avenue for any sort of sentiment around specific TikToks are the comments section, which tend to be difficult to interpret or read through on a broader scale. Showing a summary statement based on collective feedback will foster understanding and perspective as well.

Feasibility Calculations

In order to evaluate the feasibility of ACES on a larger scale, some assumptions must be made. It is difficult to find the exact count of daily active users (DAUs), but we have found that there are around 1 billion monthly active users (MAUs) globally (Wang, 2021). Based on this information, here are our assumptions:

- Between 50 and 200 million daily active users (est.)
- 25% of daily active users post content (est.)
    - They post between 1 and 3 times per day (Johnson, 2021)
- Each stage will have 35 opt-in evaluators and 35 topic cluster members
    - This results in an ~90% confidence level and ~10% margin of error[1]
    - This assumes any given topic cluster has at least 100,000 users; very small topic clusters would require less evaluators to remain statistically significant

[1] This was determined using a SurveyMonkey sample size calculator. The population input is considered to be the size of the topic cluster.

    - Total evaluators for all stages, per TikTok: 280
    - Total *opt-in* evaluators for all stages, per TikTok: 140
- 40% of posts are valid for evaluation (i.e. not recommendations, cooking, skits, etc.)

How many posts are there per day given specific extremes?

```
[Total posts per day] = [DAUs] * [% post daily] * [post frequency modifier]
```

```
[Evaluable posts] = [Posts per day] * [40% evaluability]
```

Estimate Range	**Total Posts Per Day**	**Evaluable Posts**
Low (50M DAUs; low post freq.)	12.5 M	5 M
Medium (125M DAUs; mid post freq.)	62.5 M	25 M
High (200M DAUs; high post freq.)	150 M	60 M

What percent of DAUs would be needed to opt-in for moderation and how frequently would they need to evaluate posts each day?

```
[% DAUs opt-in] = 100 *
   ((140 / [posts eval. Per day each evaluator]) * [Evaluable posts]) / [DAUs]
```

Posts / Day (Evaluable)	**% DAUs opt-in [250 posts / day]**	**% DAUs opt-in [150 posts / day]**	**% DAUs opt-in [100 posts / day]**	**% DAUs opt-in [50 posts / day]**	**% DAUs opt-in [25 posts / day]**

12.5M (5M)	6% (3M)	10% (5M)	14% (7M)	28% (14M)	56% (28M)
62.5M (25M)	11% (14M)	19% (23M)	28% (35M)	56% (70M)	112% (140M)
150M (60M)	17% (34M)	28% (56M)	42% (84M)	84% (168M)	168% (336M)

Based on the table above, remaining statistically significant in ACES evaluations is at least relatively feasible. That being said, the more posts there are relative to daily active users and the less opt-in evaluators there are, the worse it will perform. As the platform scales, more and more evaluators will be needed, even if they do 100 posts per day. These estimates are more realistic than those presented in our slides to maintain statistical significance for hypothetical situations. Additionally, our estimated percentage of content creators relative to daily active users and the amount of evaluable posts is on the high end to establish worst case scenarios. In reality, these opt-in percentages would be lower due to AI assistance and further elimination of evaluation when a post doesn't move past stages 1 or 2.

Methods

Literature review

After establishing the general direction of our project, our team conducted a literature review on existing platforms and how they approached content moderation. The purpose of this review was to develop a deeper understanding of our problem space in order to identify potential design opportunities that we want our deliverable to tackle. Our paper

focused on the content moderation process of three separate platforms: Reddit, Twitter, and Facebook. For each platform, we summarized the content moderation process, existing debates surrounding the process, and some interesting ideas that arose from this research. Using the data collected, we were able to establish the first ideation of our project deliverable highlighted in our project proposal.

Design Scope and Features

After gathering some background information of our problem space, we began to ideate our design solution. These ideations were done through brainstorming sessions and paper sketches of the content moderation process. Our preliminary idea was to create a four-tiered tunnel system where users participate in different stages of content evaluation. Originally, our platform design was not separate from any one platform. It was a separate identity that could be universally used regardless of the user's choice of social media and can fairly arbitrate without having to adhere to platform policy. It was to be a web browser plugin to make a system accessible across a variety of social media platforms. However, after being advised by our professors to scope down our idea, we abandoned the web browser plug-in idea and chose TikTok as our platform of choice. In addition, we placed greater emphasis on the incentivization aspect of our design solution. After a few more iterations, we narrowed down our design features and content moderation process to a four question system.

Medium-Fidelity Prototypes

In order to visualize our crowd-sourced content moderation process and incentives, we created a medium fidelity prototype on Figma. We utilized a design kit containing high fidelity mockups of the TikTok and added our designs to it. After creating the designs, we prototyped the design to enable interaction and fully immerse the user into the experience.

Usability Tests

Our final stage of the design process was testing our prototypes with users. We conducted four user tests, which included desirability testing and usability testing. We began the user tests by asking the users about their backgrounds using social media and their experiences with TikTok specifically. This included their opinions on incentives that are similar to what would be used in ACES. We also asked about their familiarity and opinion on cryptocurrency. We then asked desirability questions about ACES incentives. We presented the concept of TikCoin and asked their opinion on moderating for ad removal, gifts, and boosting visibility. This included asking about specific rates of moderation to reward, and honesty on moderation. Following the first round of desirability questions, we introduced the full ACES system and conducted a guided usability test with our prototype, inquiring about stage lengths, clarity of instructions, and usefulness of moderation information. We completed the test by asking about the desirability of moderation grades and statistics. Questions asked during our tests can be found here.

Design presentation

Our main design to combat misinformation and disinformation on TikTok is our crowdsourced accountability content evaluation system (ACES). The system uses the native audience of a specific post combined with volunteer moderators to help give a diverse and accurate account of the information present in a post. This system was divided into two design spaces, ACES where the trustworthiness of a post is evaluated, and “the incentives” where we list active measures to encourage users to volunteer for the system.

ACES

ACES in essence is a series of four stages where users individually rate the trustworthiness of the post. These stages would build off each other, so stage two only gets enacted if stage one reaches certain criteria, and stage three and four directly incorporate answers from the previous stage. To ensure minimal impact to the main focus of TikTok each stage is designed to only take a few seconds to answer after watching the target post. A full prototype of the system can be found here.

Stage One

Stage one is a simple yes or no question, asking if the viewer trusts the information presented in the post. The goal of this stage is to limit the number of posts analyzed by our system, a preliminary check to ensure the post is untrustworthy. To do this we compare the answers from the native audience of the post with the answers from our system's volunteers, specifically non-native audiences. If both groups don't trust the post, or only one group trusts the info it immediately moves on to the next stage. However if both groups do trust the information, it gets discarded by the system. There is also a not applicable option included in case the post was wrongly flagged for the ACES system. Since this stage is only built upon in stage two it can progress with far fewer responses than the other stages.

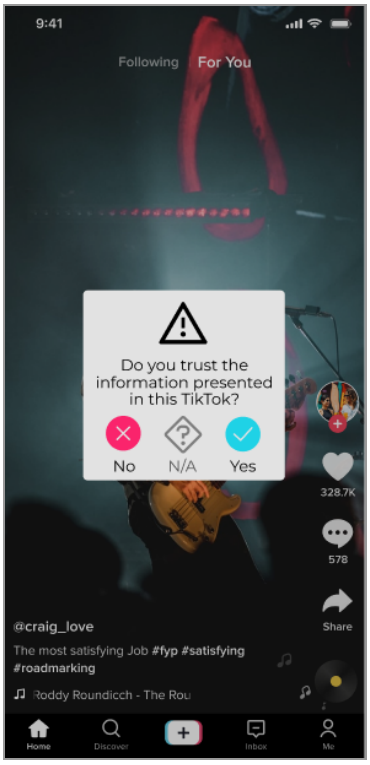


Figure 1. Stage 1 of ACES

Stage Two

Stage two includes the same initial question as stage one, then adds a checklist of standard reasons why for the viewer to choose from. These reasons change depending on if the viewer trusts or distrusts the information given. The goal of this stage is to collect community data on why a post is trusted or untrusted. This data is used to help inform the viewers in the next stage as well as provide a community perspective if TikTok decides to directly moderate the post. Once an appropriate number of responses have been collected, ACES progresses to stage three. For a one hundred thousand person topic cluster this would be the 70 people discussed in feasibility calculations.

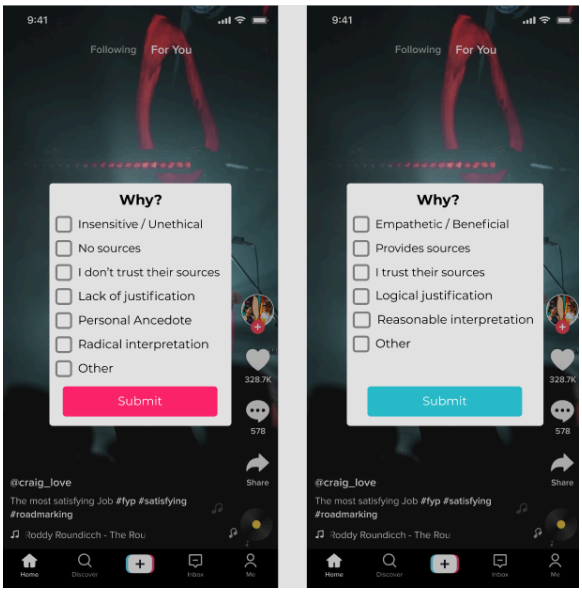
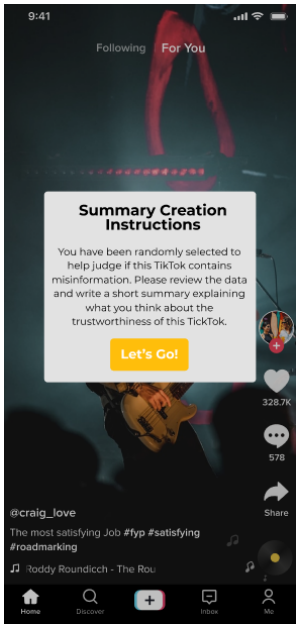


Figure 2. Stage 2 of ACES

Stage Three

Stage three asks the viewer to write a short summary declaration. The goal of which is to get crowd sourced statements to help inform new viewers of a post about the potential issues with the post. The stage also presents viewers with the information gathered in stage two to help inform the viewer about the community thoughts of a post. Again once enough declarations are collected the process moves forward to stage four.



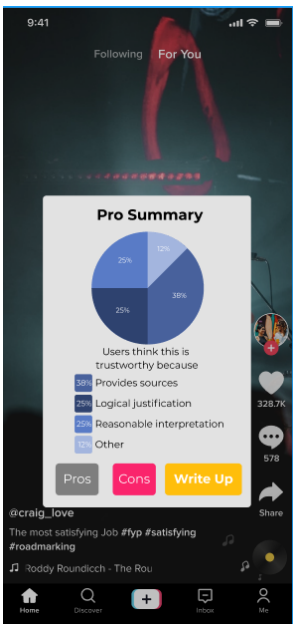
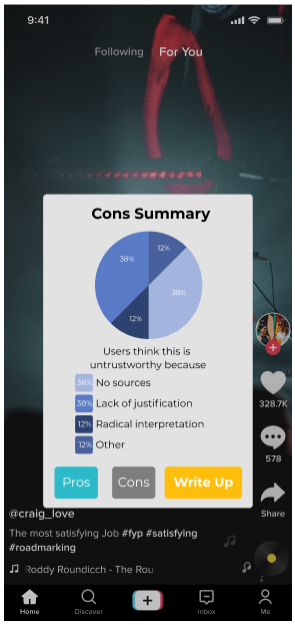
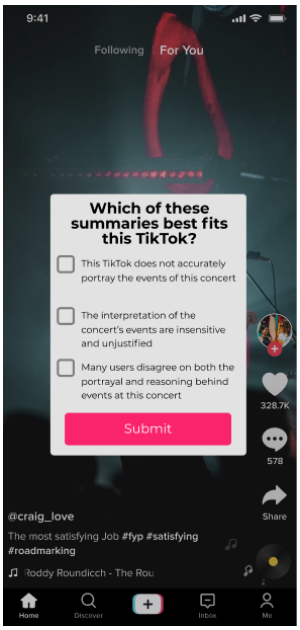


Figure 3. Stage 3 ACES multi Screen-display





Stage Four

Stage four presents the viewer with four randomly selected summary declarations and asks the viewer to select the best declaration. During this stage each declaration is shown the same number of times, however not all declarations are shown at the same time. At the end of this stage the most selected declaration is attached to a warning banner on the post.

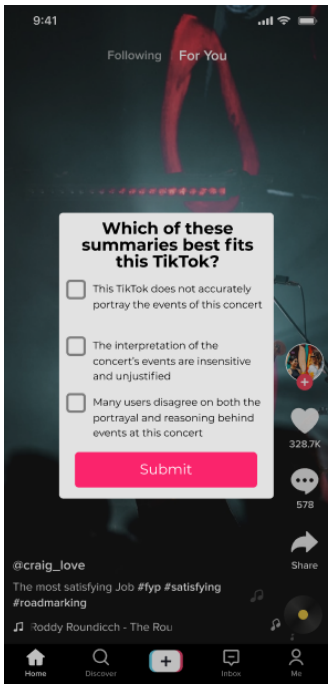


Figure 4. Stage 4 of ACES

Stage X

Stage X stands in for what actions TikTok might take after ACES has collected community feedback on a post. For instance, if a post was viewed as untrustworthy by both the native audience and volunteers it can be banned from sharing and from showing up in anyones 'For You Page' on top of the warning banner appearing. However if community opinion is more mixed the post can just have the warning banner and the sharing or downloading capabilities removed. Additional interventions can include the removal of the TikTok's audio sharing feature for a specific post, with the use of AI to remove that audio clip from any other post. The specific actions taken based on the result of ACES need more data driven research before recommendations can be made.

Adversarial Impact

A large part of the design of these stages was to limit adversarial impact, in other words bad actors who are responsible for disinformation instead of misinformation. Literature analysis of Twitter's Birdwatch program that closely mirrors our own solution suggested the implementation of community sources summaries itself did not protect against organized disinformation attacks, (Garfiel B. 2021). Our primary design implementation to protect against this is imposed randomness. native viewers of a post take part in the moderation of the post, they are randomly selected and matched one-to-one with randomly selected volunteer moderators. This also means that most native views will not interact with the moderation system at all, helping to prevent brigading. Additionally with multiple stages involved with the moderation a single bad actor can not play a substantial role in the entire process. Stage three is currently the most influential stage but a group attempting to influence the moderation by writing a purposely dishonest declaration is circumnavigated by stage four only showing four random posts. This means that even if the group had several members voting for declarations in stage four most of their votes would either be discarded or forced to vote for a different declaration.

While this system design approach should limit disinformation attacks additional background systems could be implemented to help protect the integrity of the ACES system. One such system would be a hidden demoting process for the selection of users into the ACES system. The demoting process would reduce the chances of both native

viewers and volunteers being selected for the ACES system. The demoting can be based automatically on a variety of factors:

for stage two this is making sure the user is not just selecting the same answers each time;

for stage three this is prioritizing users who provides in depth summaries with links;

more generally if the user consistently deviates from the community consensus.

Regardless of the specific actions judged for demotion this process would be hidden to prevent gamification from users.

Incentivization

Overview

All of our incentives are displayed on the user's profile under "Moderator Stats". We chose the profile page since social validation from peers is the primary incentive to completing these tasks and earning these awards. Essentially, the user can "show off" their achievements to their followers.

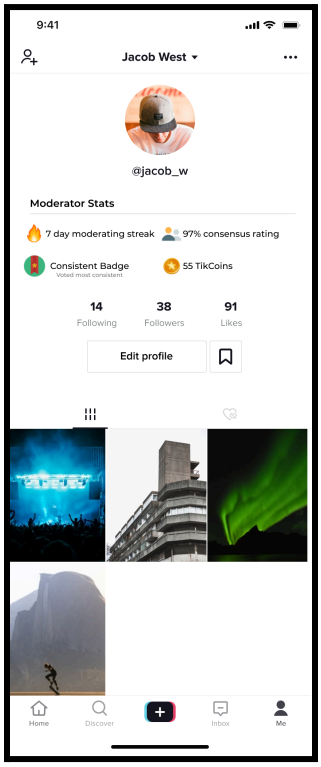


Figure 5. Stage 4 of ACES

TikCoin

TikCoin is an in-app currency system where users can earn TikCoins through moderation and use those to buy features that
enhance a user's experience on the app. A user would earn TikCoin for each moderation task completed. TikCoin could be exchanged for TikTok Gifts and for ad removal. These features were chosen as they were the incentives the users were most interested in. Users can view their TikCoins on their profile page. Their coin balance would not be visible to other users.

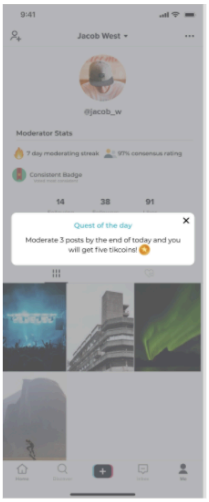
TikCoin was updated based on user feedback we received. Notably, we removed the cryptocurrency aspect and chose the specific rewards of ad removal and gifts.

Statistics

Users would be able to view details about their moderation style, including how often they matched the eventual moderation consensus of a post. A user would be able to see these statistics on their profile page and it would be visible to others.

Quests and Achievements

ACES users would be able to complete sets of moderation tasks to receive bonus TikCoin, badges, and features. A quest would pop up on the user's profile page and ask if they would like to begin a quest to complete a certain number of moderation tasks, either general or at a specific stage, in order to receive a bonus. To complete an achievement, users would moderate a certain amount of posts to receive a badge that is visible on their profile.



Quest Implementation of ACES

Streaks

A streak is a type of achievement. In order to get a streak, a user must complete at least one moderation task a day. They lose a streak when they don't moderate for a day they are on TikTok. They are notified when their streak is started, after moderating once, and are subsequently notified when they continue their streak through the following days by moderating for the first time that day. Users don't lose their streaks when they do not visit the app one day or if a moderation activity does not appear while they are scrolling for the day. Their streak number would be visible on their profile page as a badge with a number inside.

User Test Insights

The purpose of our usability tests were to understand whether our incentives were incentivising and unravel users' feelings, thoughts, reactions and pain points regarding the content moderation stages. User insight was imperative to establishing the feasibility and effectiveness of our design solution as well as potential questions and recommendations for future work. User tests were based on a preliminary model of the ACES system. We used the insights from the user tests to finalize different elements of ACES. The following is the feedback we received from the user research we conducted, and the number of users who resonated with a piece of feedback is denoted by "(x/4)":

General

Overall, feedback on the ACES System was positive. Users mentioned that they were displeased with the current moderation in place on TikTok and desired reformation. There were a few notable pieces of feedback:

- **Excessive Moderation is Demotivating** - When we asked about how many "one question surveys" (encoded for moderation tasks), some users noted that they would do one survey for a reward (2/4). However, when we asked about 5 or 10 questions, users became more hesitant and noted that it would be annoying if questions were popping up frequently (2/4).
- **Presence of Moderation Task Alters Opinion** - Users noted that while it depended on the type of TikTok they were shown, the presence of a moderation task or banner would likely make them question the TikTok they just viewed (2/4).

Stages 1, 2, and 4

We received no critical feedback about these stages. Users noted that the tasks were short and that they would likely do them if they popped up on TikTok.

Stage 3

Users had varied responses to the task in stage 3, noting concerns for duration, data, and consistency.

- **Previous Stage Data Yielded Mixed Opinions** - When shown the previous responses as a pie chart, some users appreciated the information (2/4) while others didn't find it important (1/4). One user noted that the information was too complex and speculated that not many people would pay attention to it. Another user appreciated the breakdown of the information because it helped articulate their thoughts.
- **Long Duration** - Some users mentioned that the task presented would take them a long time to complete (2/4).
- **Concern for Consistent Summaries** - Users were concerned about the consistency of the summaries written by the community (2/4).

Stage X

Response was mostly positive for stage X of moderation, with users noting that they appreciated the detailed explanation of the problematic aspects of the TikTok, and the rationale behind the moderation decision (2/4). There were some concerns for the efficacy of the banner.

- **Dismissal of the Banner** - One user noted that they would likely stop clicking on the warning banner after seeing it a few times.

TikCoin

Users were disinterested in the concept of TikCoin being a cryptocurrency (3/4), but noted that gaining TikCoin to use to improve their experience may incentivize them to use the moderation system (2/4). One user noted that implementing a cryptocurrency into an app that children used concerned them. The user also noted that gifting was also a point of contention for them for similar reasons.

Rewards

We inquired about three types of rewards: ad removal, gifts, and post visibility. Users were most interested in the rewards of ad removal (4/4) and gifts (2/4). However, we had a disproportionate amount of users in our test that did not post on TikTok frequently.

Statistics

Users were interested in knowing their own moderation statistics regarding their percentage of aligning with consensus (1/4) but noted their opinion on other people's statistics would depend on their own statistics (1/4).

- **Minimal Judgment of Other Users** - One user noted that their opinions of other users would not be changed significantly if they saw their consensus score, unless it was especially high or low. They also noted that they would need context from their own score.
- **Interest in User Statistics of Particular Topics** - All users noted that they would be more interested in knowing how people moderated on specific issues and TikToks (4/4).

Streaks

We asked about how users used streaks on other social media apps, including Snapchat and Duolingo.

- **Non-Intentional Streak Maintenance** - Some users would not maintain streaks on purpose, but instead because they would use the app daily regardless and would complete the same actions or messages (2/4).

Conclusion

Moderation solutions for social media need to adapt to the changing attacks on our information spaces. As social media companies grow, so do the unique nuances in their diaspora of internet cultures. This makes traditional moderation harder and harder as it must adapt to the specific subgroup it is moderating. Incentivising a community to perform its own moderation with tools that provide them the ability to foster perspective answers this concern. This draws in the second key design space of ACES, incentivization. While the gamification, and the incentives listed in our design are not new our literature review showed very little integration of these systems into moderation. Our contribution to this space is envisioning how these systems can be used to foster a grass roots moderation for posts regardless of post status. We see this as an improvement on Twitter's Birdwatch system whose hands off approach makes their volunteers target popular posts disproportionately and almost never extends to fringe communities.

Solutions need to be introduced to handle not only platform guideline violations but also disinformation attacks and the general spread of misinformation. Social media companies have too many posts everyday for third party moderation to be their sole solution. Additionally third party moderators rarely provide communities with reasons for removal that foster understanding and perspective in the community. Even currently implemented community moderation systems like Birdwatch do not safeguard against disinformation attacks. ACES is envisioned as working within the existing moderation sphere and so does

not hold the answers for every edge case. In fact the actions suggested after an ACES evaluation in stage X are envisioned to be changed once a more holistic concept of its integration is made. ACES should be viewed as a framework to build moderation systems off of. It is intended to be a theoretical system that shows how a moderation system can tackle platform violations, disinformation attacks, and educate against misinformation.

Plaintext Links

Prototype :
https://www.figma.com/proto/PXJ4JOPXHzGGHVjwBeovZN/Low-Fidelity-Prototypes?scaling=scale-down&page-id=0%3A1&starting-point-node-id=5%3A3&show-proto-sidebar=1&node-id=5%3A86

User Test Questions:
https://docs.google.com/document/d/1j3lcQmBDugjlGOOT4fQAR__2YnVAr2_rXdcIlpktfoQ/edit?usp=sharing

References

On TikTok, Audio Gives New Virality to Misinformation. (2021, July 14) *NBCNews.com* NBCUniversal News Group. Retrieved on December 16, 2021, from https://www.nbcnews.com/tech/tech-news/tiktok-audio-gives-new-virality-misinformation-rcna1393.

StriveCloud (2021, Aug. 04) *How Does Gamification Drive Engagement* Retrieved on December 16, 2021, from https://strivecloud.io/blog/app-gamification/how-gamification-drives-engagement/.

Garfield B. (2021, Oct. 08) *Who watches the Birdwatchers? Sociotechnical vulnerabilities in Twitter's content contextualisation* Solent. Retrieved on December 16, 2021, from https://pure.solent.ac.uk/en/publications/who-watches-the-birdwatchers-sociotechnical-vulnerabilities-in-tw

Wang E. (2021, Sept. 27) *TikTok hits 1 billion monthly active users globally - company*. Reuters. Retrieved on December 16, 2021, from https://www.reuters.com/technology/tiktok-hits-1-billion-monthly-active-users-globally-company-2021-09-27/

Johnson A. *How Often Should You Post On Social Media As A Content Creator In 2021?* Audiosocket. Retrieved on December 16, 2021, from fromhttps://www.audiosocket.com/social-media-guides/how-often-should-you-post-on-social-media-as-a-content-creator-in-2021/

Sample Size Calculator (2021, Dec. 16) SurveyMonkey. Retrieved on December 16, 2021, from https://www.surveymonkey.com/mp/sample-size-calculator/

Kulkarni, C., Wei, K. P., Le, H.,Chia, D., Papadopoulos, K., Cheng, J., Koller, D., Klemmer, S. R. (2013). *Peer and self assessment in massive online classes*. AMC Transactions on Computer-Human Interaction, Vol. 20, No 6. https://dl.acm.org/doi/10.1145/2505057

New America. 2021. *Everything in Moderation*. [online] Accessed 1 November 2021, from https://www.newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-user-generated-content/case-study-facebook/

Ori, S., 2019. *Facebook Rules: Structures of Governance in Digital Capitalism and the Control of Generalized Social Capital.* Research Gate. Accessed 1 November 2021, from https://www.researchgate.net/publication/331314053_Facebook_Rules_Structures_of_Governance_in_Digital_Capitalism_and_the_Control_of_Generalized_Social_Capital

Nicolas Pröllochs. (2021) *Community-Based Fact-Checking on Twitter's Birdwatch Platform.* Cornell University. Retrieved 1 November 2021, from https://arxiv.org/abs/2104.07175

Cochran, A. (2021). *What are Reddit Community Points?*. Medium. Retrieved 1 November 2021, from https://medium.com/@adamscochran/what-are-reddit-community-points-13e2a839849b

Garfield Benjamin. (2021) *Who watches the Birdwatchers? Sociotechnical vulnerabilities in Twitter's content contextualisation*. Solent University. Retrieved 1 November 2021, from https://pure.solent.ac.uk/en/publications/who-watches-the-birdwatchers-sociotechnical-vulnerabilities-in-tw

Gamble, L. (2019, September 23). *The psychology behind Social Media*. Medium. Retrieved December 16, 2021, from https://medium.com/@l.gamble/the-psychology-behind-social-media-a4aafba46d57

Postigo, H. (2009). *America Online volunteers: Lessons from an early co-production community*. International Journal of Cultural Studies, *12*(5), 451–469. https://doi.org/10.1177/1367877909337858

Haynes T. (2018, May 1). *Dopamine, smartphones & you: A battle for your time*. Science in the News. Retrieved December 16, 2021, from https://sitn.hms.harvard.edu/flash/2018/dopamine-smartphones-battle-time/

Seeking validation online: The effects of social media. Watts Basketball. (2021, June 17). Retrieved December 16, 2021, from https://wattsbasketball.com/blog/seeking-validation-online

Seiter, C. (2020, June 30). *The Psychology of Social Media: Why we like, comment, and share online*. Buffer Resources. Retrieved December 16, 2021, from https://buffer.com/resources/psychology-of-social-media/

TikTok. (n.d.). Retrieved December 16, 2021, from https://www.tiktok.com/community-guidelines?lang=en