How to Moderate Images Efficiently:

Save Time, Money, and Resources With These Sample Workflows

twohat

Overview **2**
Introduction 2
What is PopJam? 3
Moderation challenge 4
What is Two Hat's Community Sift? 4
How Much Risk is Too Much? **5**
Develop community guidelines 5
Determine acceptable risk 5
Use Data to Create Efficient Workflows **7**
Risk Levels 7
Trust Levels and User Reputation 8
Pre and Post-Moderation Content Queues 9
Treat New Users Differently **12**
Key Takeaways **14**
Additional Resources **15**
Bonus Workflows **16**

Overview

Introduction

Social platforms face massive challenges when moderating user-generated content. With images and text being uploaded on a scale never seen before, the pressure to retain users, protect brand equity, and (for some platforms) maintain compliance, is high. Platforms struggle with not enough money, staff, and time to efficiently moderate all content submitted by users.

Is there a process that guarantees less work for your moderation team — which translates to not only a smaller, more efficient moderation team, but reduced moderation costs as well? How much do humans *need* to be involved in reviewing content?

The community team at the social app PopJam have years of experience in the industry, and have designed some of the most efficient and cost/time-saving workflows we've seen.

In this paper, we'll share their templates for usable moderation workflows that can be adapted to any community.

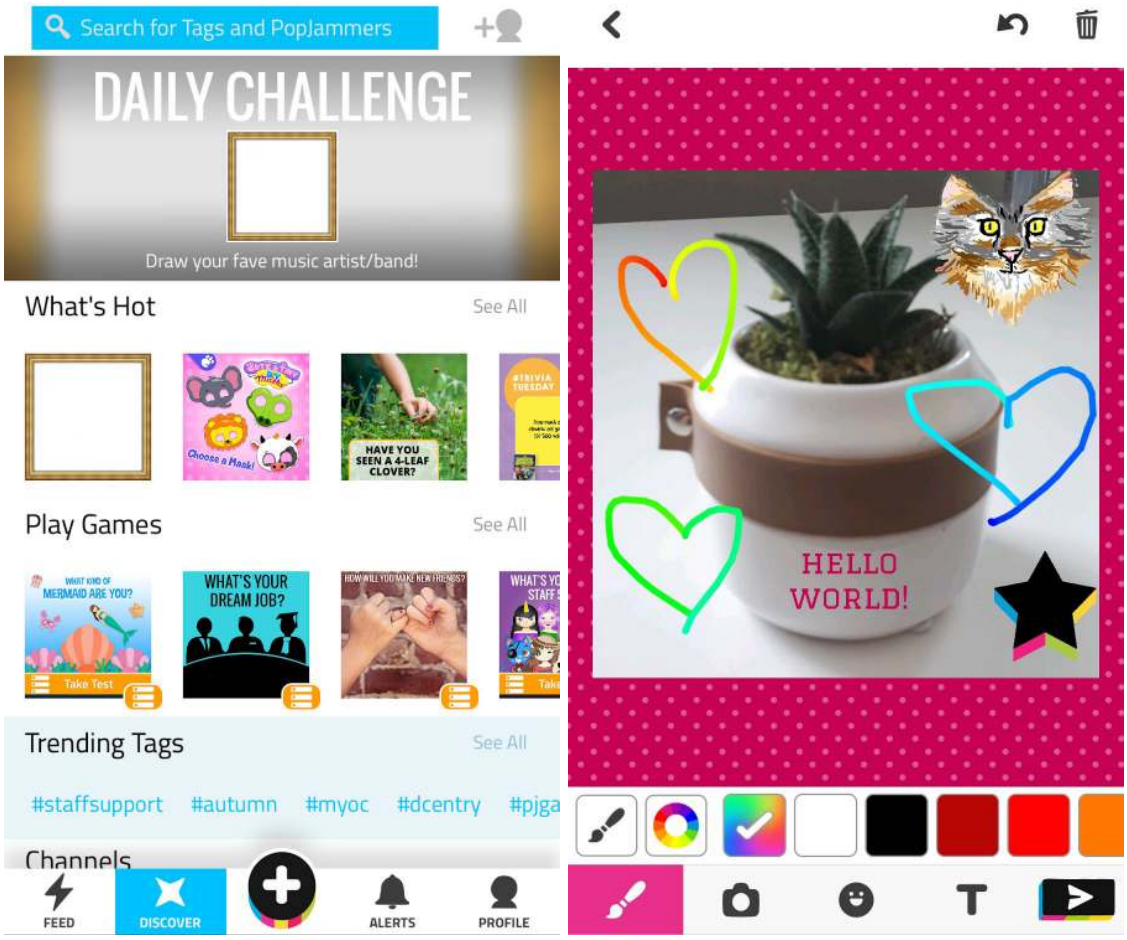
It's important to note that, while the PopJam community is comprised of mostly pre-teens, all of the techniques and workflows presented in this paper are also applicable to platforms that cater to teens and adults. As well, any industry that hosts user-generated content — from edtech platforms to social sharing sites — can benefit from these workflows.

What is PopJam?

PopJam is a creative community platform for 8-12 year olds. Users upload pictures from their device and decorate them with stickers, text, and drawing tools. They can also play games, watch videos, and complete quizzes, daily challenges, and puzzles which are then shared with the community.

Corporate brands and influential individuals create official channels to share exclusive content and interact with their young fans.

***Below**: The PopJam home screen, and content created with an image uploaded from the user's device, decorated with text, stickers, and drawings.*



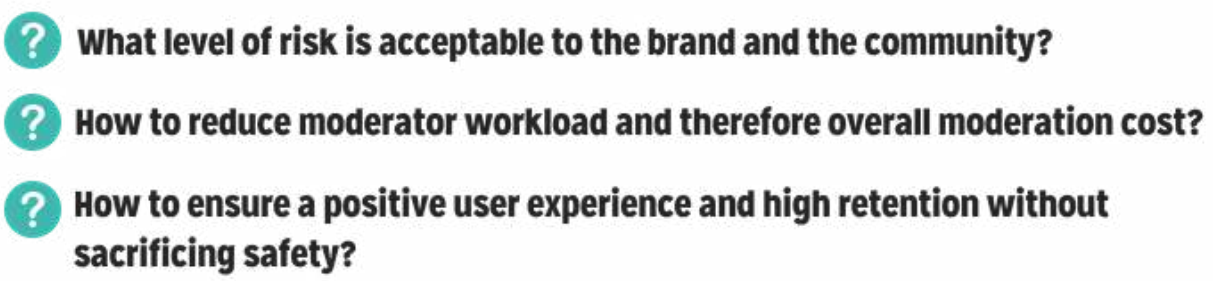
Moderation challenge

PopJam faces similar moderation challenges as other social platforms, including the high human and financial cost of moderating user-generated content on a large scale, with an added factor: Their platform is designed for an under-13 community. **It's critical to their brand reputation that PopJam protects its users from offensive and inappropriate content. Protecting the community is also key to building parental trust, brand loyalty, and user retention.** There is very little room for moderation error in a community of under-13 users.

In addition, well-known children's brands including Barbie, My Little Pony, Cartoon Network, and more have official PopJam channels. Partner confidence in PopJam's moderation tools and processes is critical to the platform's success.

To ensure brand protection and user safety, PopJam enforces strict community guidelines, including no selfies or pictures of faces, no pornographic images, and no bullying, harassment, profanity, or PII (personally identifiable information).

The community team faces several challenges and questions when enforcing these guidelines, including:



What is Two Hat's Community Sift?

The PopJam community team chose to implement Two Hat's Community Sift as its automated content moderation solution. Augmented by PopJam's highly-trained and dedicated moderators, Community Sift provides real-time image and text labeling and filtering for the app.

How Much Risk is Too Much?

Develop community guidelines

The first step when building a moderation process is to create a set of robust and clear community guidelines. PopJam's basic House Rules are outlined here, while a more comprehensive set of community guidelines can be found in their Terms of Use (scroll down to section eight).

The PopJam community team describes their guidelines as follows:

> ***"Our community guidelines leave room for creativity, communication, roleplay, artistic endeavors, silliness and fun but also include zero tolerance for players who are solely on PopJam to continually disrupt, bully, create unusually negative content or otherwise cause unhappiness for themselves and others."***

Determine acceptable risk

After community guidelines comes risk assessment. Determining your platform's risk threshold is key to building moderation workflows. This can start with your brand identity; like PopJam, brands that are associated primarily with children will out of necessity have a high aversion to risk. Community demographic will be a major factor as well.

Some platforms might be more concerned with compliance (COPPA, GDPR, etc) and will devote more resources to eliminating PII. Others may be more focused on bullying or pornography prevention.

Data review is a huge benefit when assessing your brand's risk acceptance threshold. Data (user-generated content like images and text) labeled with risk levels and topics is incredibly useful.

Some factors to consider:

- How often does your community post high risk content?
- What kind of high risk content do users post? (ie, sexual conversation, profanity, harassment, hate speech, etc)
- How much harm would your brand and/or users experience if even *one* high risk item was successfully shared on your platform?

Once you've determined your platform's risk acceptance and aversion to specific kinds of content, you can start to build a moderation workflow.

In the next section, we'll examine how PopJam used Two Hat's Community Sift's topic and risk level labels to label their content and create efficient workflows.

KEY TAKEAWAY #1: *Know your platform's risk aversion and acceptance before you build a moderation workflow. Start by labeling content based on topic and riskiness.*

Use Data to Create Efficient Workflows

Risk Levels

To understand how PopJam separates user-generated content into content queues for pre or post-moderation, let's first review how Community Sift classifies text and images.

Text is classified by topic (including bullying, sexting, hate speech, profanity, PII, and more), and on a sliding scale of eight risk levels (0=no risk, 7=highest risk).

Community managers then select "Policy Guide" settings that decide which topics and risk levels are acceptable or not acceptable to the community, based on guidelines like those in PopJam's Terms of Use.

Images are classified similarly, with five topics (pornography, gore, weapons, drugs, and extremism) and four risk levels. The percentage indicates the likelihood that an image contains one of the topics:

Risk Level	Probability
	99-100%
	50-99%
	5-50%
	0-5%

PopJam uses OCR (optical character recognition) technology to remove text from images. Text is then sent to the Community Sift text classifier, where it's classified by risk level and topic. Community Sift returns the result to PopJam in real time, and sends the text to queues for moderator review based on escalation policies set by the community team (see Pre and Post-Moderation Content Queues for more).

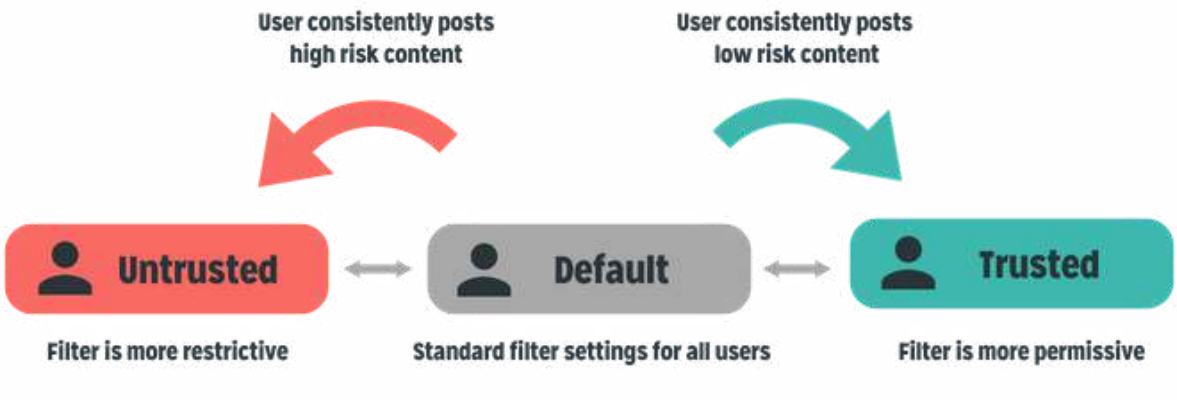
Similarly, images are sent to the Community Sift image classifier, where they are classified by risk level and topic, with results returned to PopJam in real time. Just like text, images are sent to content queues based on specific escalation policies.

Trust Levels and User Reputation

Two Hat's patented User Reputation technology is also key to PopJam's moderation workflows. Every user in Two Hat's Community Sift is immediately set to a "Default" trust level, which means their content is filtered and moderated according to a standard setting that follows the community's guidelines.

However, users who consistently post disruptive, high risk content move to an "Untrusted" level, and are subject to a more restrictive filter.

Conversely, users who consistently post healthy, low risk content move to a "Trusted" level with a more permissive filter. Users automatically move between all three trust levels based on their behavior.



In addition to using topics and risk levels, PopJam uses these trust levels to automatically route user-generated content into their moderation queues.

As we'll see in Treat New Users Differently, PopJam altered the User Reputation flow specifically for their new user workflow.

Pre and Post-Moderation Content Queues

PopJam uses multiple workflows to route content into different queues based on variables like risk level, user reputation (ie, Community Sift trust levels), and whether the item was reported. Queues are divided into low, medium, and high risk items, which then determines which content requires moderator resources.

The lowest priority queue contains items that have been posted to the app without moderator review (but can still be post-moderated at any time for quality control). For this queue, the team has determined that the risk of harmful content is so low that they are confident that user reports are sufficient to flag content that requires human review.

For example, images in their “Lowest Risk, Lowest Priority” queue ***cannot*** contain any of the following labels that PopJam have deemed moderate to high risk:

A. Photo content
B. Text that fails the filter
C. An image risk score above 2
D. **OR** was posted by an Untrusted user

That leaves content that is *very* low risk. Items in this queue have a less than half a percentage point chance of being rejected by a moderator.

The results? **Using automation, PopJam reduced their potential moderation workload by 65%.**

With the bulk of submitted content being posted to the app without human moderation, the team can focus their attention on the items that require human review.

Using their risk acceptance thresholds and labeled data review, PopJam was able to determine which items are *potentially* risky enough that they require pre-moderation.

For example, images in their “Highest Risk, Pre-Moderation” queue ***must*** contain:

A. A possible face **OR**
B. Text-on-image or caption content that would fail text filtering **OR**
C. Was posted by an Untrusted User
D. **AND** does not contain animation (sticker or Giphy)

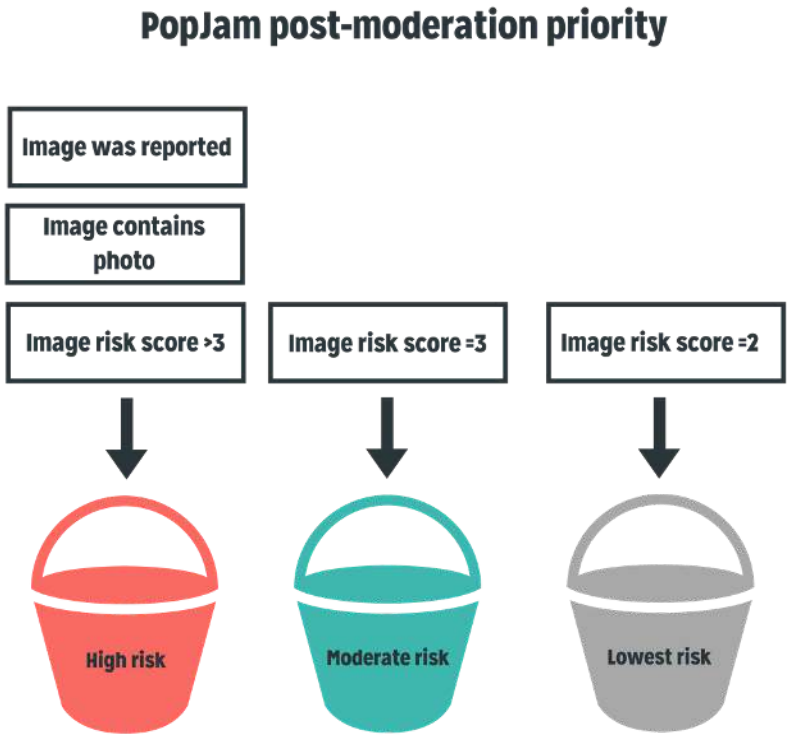
In addition, a very small percentage of images uploaded to the PopJam app contain pornography. In the event that a user attempts to upload a pornographic image, Two Hat's Community Sift image filter rejects and removes these images automatically. This way, they are never seen by the community, re-routed to a content queue, or reviewed by moderators.

Ultimately, the goal of efficient moderation is to push your human resources to the areas of highest risk. Know what is riskiest to your platform, label that in your data, and ensure you have human eyes on it.

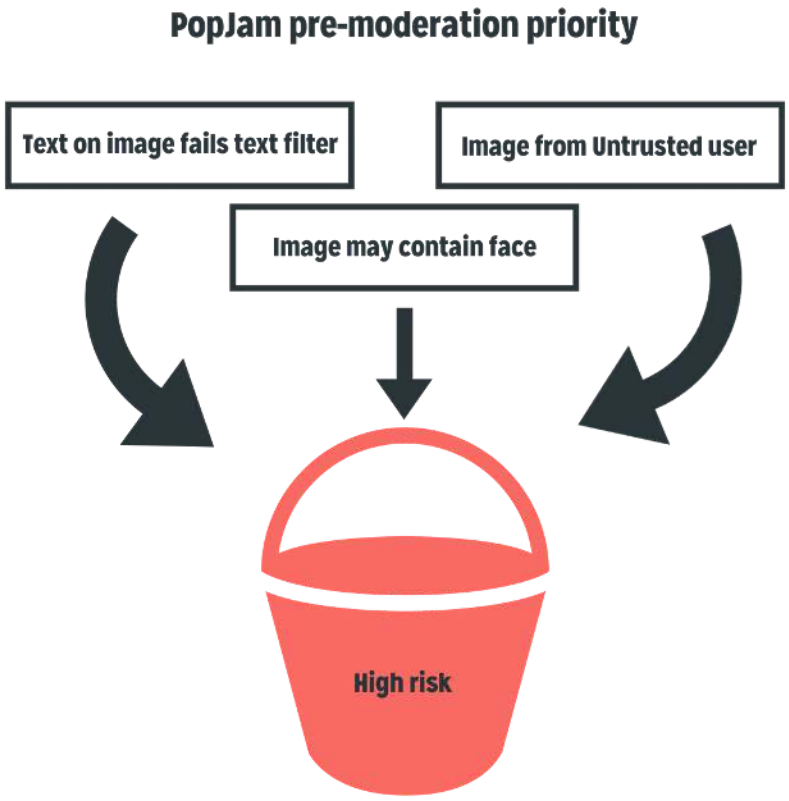
KEY TAKEAWAY #2: *Save human and financial resources by routing your lowest risk content to a queue that doesn't require moderator approval before being posted.*

KEY TAKEAWAY #3: *Use risk thresholds to determine which items require moderator approval before being posted.*

***Below**: PopJam separates post-moderation content into three "buckets" - high, moderate, and lowest risk. Risk levels help them determine which queues needs to be reviewed by moderators first.*



Below: *PopJam routes their riskiest content to a pre-moderation queue. Moderators must review images before they're published on the app.*



Treat New Users Differently

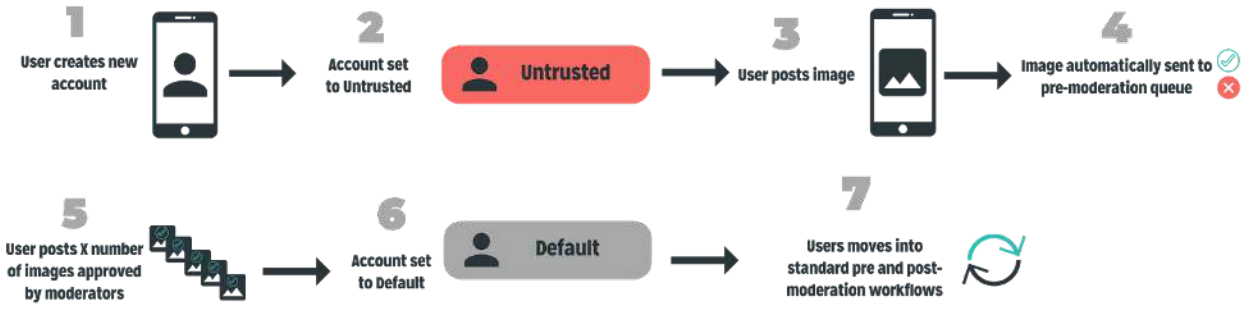
Based on their years of experience moderating online communities, the PopJam community team created a special workflow for new users.

New users have little invested in their accounts, and so have nothing to lose when posting inappropriate content that could result in being removed from the app. It's unusual for a user to spend weeks or months building up content including posts, likes, and followers on their PopJam profile — then risk losing that work by posting an image that breaks community guidelines. **The team knew that users who are willing to break the rules are far more likely to disrupt the community when they first create an account.**

In the typical Community Sift trust level system, all new users are automatically set to the "Default" trust level, meaning their filter settings are based on standard community guidelines, with the opportunity to move to a more restrictive or permissible filter, based on behavior.

However, in PopJam all new users are automatically "Untrusted". Any image that they post is placed in a pre-moderation queue for moderators to review before the image is published for the community to see. **New users remain in this "Untrusted" state until they have had a predetermined number of images approved by moderators without a rejection.** Once they have reached that threshold, their account is moved to the "Default" level, and their posts are subject to the standard moderation workflows.

With this workflow, users who break community guidelines realize very quickly that they will be prevented from posting inappropriate content — which in turn leads to fewer high risk items that require costly human moderation.

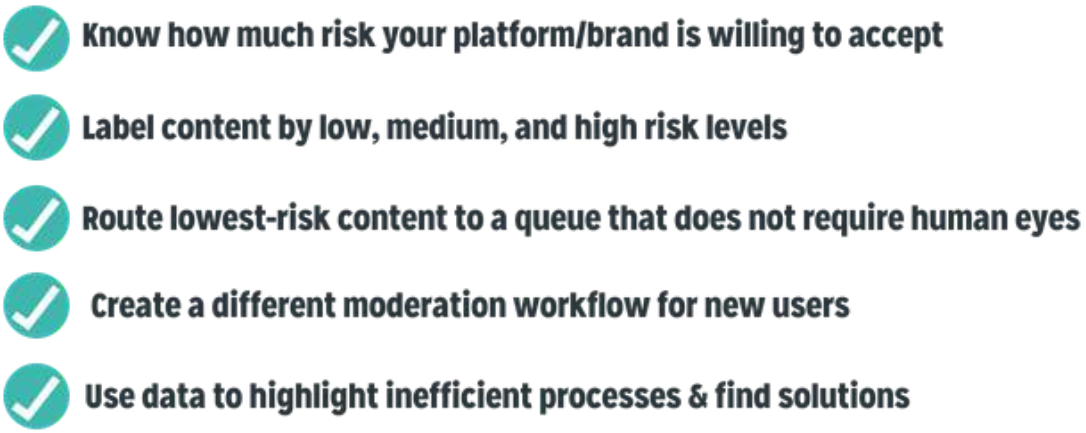


KEY TAKEAWAY #5: *Spare yourself the headache. Use a combination of well thought-out community guidelines and a special workflow for new users to set the community tone early in the user experience.*

Key Takeaways

By implementing a proactive chat and image filter, assessing their risk acceptable level, and using a data-driven approach, **the PopJam community team reduced their moderation workload by 65%**. This leaves moderators free to review the content that matters, and to interact and engage with the PopJam audience in retention-boosting activities.

Key takeaways for those who are building their own moderation workflows include:



To learn more about how you can use Two Hat's Community Sift chat and image filter to proactively moderate your online community, visit our website, or get in touch at demo@communitysift.com.

Additional Resources

 Download PopJam

 PopJam Community Guidelines & TOU

 Image Moderation: Lessons From the Experts at PopJam (on-demand webinar)

 4 Step Beginner's Guide to COPPA Compliance

 What is GDPR?

 Interview with SuperAwesome's Head of Trust and Community Rebecca Newton

 Top six reasons you should combine automation and manual review

 Five best practices to optimize your image moderation workflows

 Image Moderation 101 (on-demand webinar)

Bonus Workflows

PopJam's Community Technology and BI Task Leader Lynn Snyder stresses the importance of using data to reduce moderation workload and sort content into specific buckets for quick and efficient review.

Here are three ultra-efficient workflows used by PopJam. Each of these three solutions was built after the team identified an inefficient process:

1. Supervisors can locate disruptive accounts efficiently in escalation queues for certain thresholds of topic-related behaviour points. Content queues include:

    a. Escalated bullying content.
    b. Users whose reputation changes from "Default" to "Untrusted".
    c. Escalated custom topic. PopJam uses the "Custom" topic option in Two Hat's Community Sift to label specific words and phrases known to be associated with negative behaviours across all topics.

    These queues serve up very specific high risk content to moderators, preventing them from wasting valuable time.

2. An "all content" button on the user's profile page in Community Sift calls up every image and line of text that a user has created, in chronological order. This allows moderators to quickly review posts in context. For example, a user may have recently written "commit suicide", which would normally be cause for alarm. However, using the "all content" button, a moderator may observe a drawing that contains drawn text, surrounded by a circle with a line through it — completely changing the context of the text and image.

3. Front-facing markers on images in the queues indicate whether the post was made by a staff member or a PopJam partner. This tells moderators, for example, if a selfie was posted by an influencer or a brand. Normally selfies are against PopJam guidelines, but in certain cases these are allowed.

BONUS TAKEAWAY: *Use analytics to determine where your moderators are spending the most time. Can those tasks be streamlined or even eliminated using automation?*