conectys
global BPO solutions

Content Moderation

How to achieve scale and affordability

A Conectys made ***White Paper***

Introduction

A few months ago, we wrote a paper on the constantly-emerging landscape of content moderation regulations. In it, we actually quoted a Twitter executive based in the United Kingdom, who was bullish that the legal and political regulatory environment would soon catch up to social media platforms.

And now, as 2021 ends, here we are.

Buoyed by the Facebook whistleblower Frances Haugen and other concerns about social media platforms and their role in our day-to-day lives, the regulatory environment seems to be fast approaching how these companies operate. In the USA, even -- which has been a highly-polarized environment for five years, if not much longer -- the need to regulate platform companies around content moderation is now a bipartisan issue. (Now, admittedly, the left and the right side of the ideological spectrum want to more strongly moderate content for entirely different reasons, but it is now a rare topic they can agree on.)

As the regulatory environment finally starts to move faster, there are a number of questions you need to be thinking about. First of all: what would it look like to "solve" this problem? Would platform companies ultimately bear responsibility for all posted on those platforms? If so, they'd need a massive army of moderators, well beyond what they have now. Could those companies, which make billions upon billions a year, afford to increase moderation by potentially 5000% annually?

The big questions are around scale and affordability. The second tier of the question for content moderation right now is about technology vs. humanity. AI is growing by leaps and bounds, but it can't moderate everything, and leaving some aspects of moderation to AI solely can have horrible brand repercussions. So, you clearly need some mix of technology and humans; you need the technology to bolster the human experience of moderation. But where's the exact line?

Those are the two big issues, then:

1. Scale and affordability of content moderation's future
2. The role of technology, only using technology, or a mix of technology and human moderators

In this paper, we're going to address just topic (1). In a future paper -- one to two weeks from now -- we will address topic (2).

In order to understand questions of scale and affordability, it's best to understand the key types of content and the pros and cons of each. Let's start there.

Manual Pre-Moderation

With Manual Pre-Moderation, all user-submitted content is screened before it goes live on your site. Each piece of content is judged by a moderator who takes a decision on whether to publish, reject or edit it, all according to the site guidelines.

Pros	Cons
Highest possible control of content, which leads to quality of the site and better user experience.	Speed of moderators slows down process of content submission, which can negatively impact user experience.
Manual pre-moderation helps with better detection of complex scams that can cause harm to a site's reputation and in turn loss of users.	It's costly and time-consuming to train teams to be able to detect complex scams. Sophisticated processes should be in place for continuous improvement, learning and training of new staff.

Manual Post-Moderation

Manual Post-Moderation allows content to go live on your site instantly to then be reviewed by a moderator after it's been published. The moderator will, in the same way as with manual pre-moderation, review each ad and make a call on whether to keep it on the site, remove it or make edits.

Pros	Cons
Instant user satisfaction as their content is published in one click.	No guarantee of a moderator seeing potentially damage content before users do.
A good option to handle less sensitive content.	General risk of offending site visitors, creating both low retention and bad publicity.

Reactive Moderation

Reactive moderation relies on your users flagging or reporting content on your site. This can be done via report buttons on your site or through customer support tickets. This is a very powerful tool, but for most sites it should only be used as a supplement to one of the other moderation methods.

Pros	Cons
Cost-efficient method that filters out contents that's upsetting enough for people to react to.	No real control of content posted to the site which in turn might not be in line with brand image.
	As content has to first go live and then be found by site communities, unwanted content could potentially be live for days

Distributed Moderation

Distributed moderation is the democratic cousin of reactive moderation. Here you leave the moderation efforts almost entirely to your community. This moderation method relies on rating and voting systems where highly-voted content ends up on top of the page and lowly voted content is hidden or removed. You can either give voting rights to all members or to certain VIP-users, which can be appointed by the community or site owner. While their moderation approach is not entirely distributed, one way to conceptualize this is Reddit.

Pros	Cons
A great way to moderate content where communities/users are very invested in the site.	Limited control over what is being moderated and when it is carried out.
A good correlation between the site and site community's perception of high-quality content is needed. If that exists, this model is great.	If the site owner can be held accountabl for site content in any way, it is better to use the distributed moderation methods as a support for the main moderation method.

Automated Moderation

Automated moderation is becoming increasingly popular with more sophisticated filters and tools being developed. The most basic version is a filter which catches words from a list and acts on preset rules to either highlight, replace or ban the word or content piece. Filters do

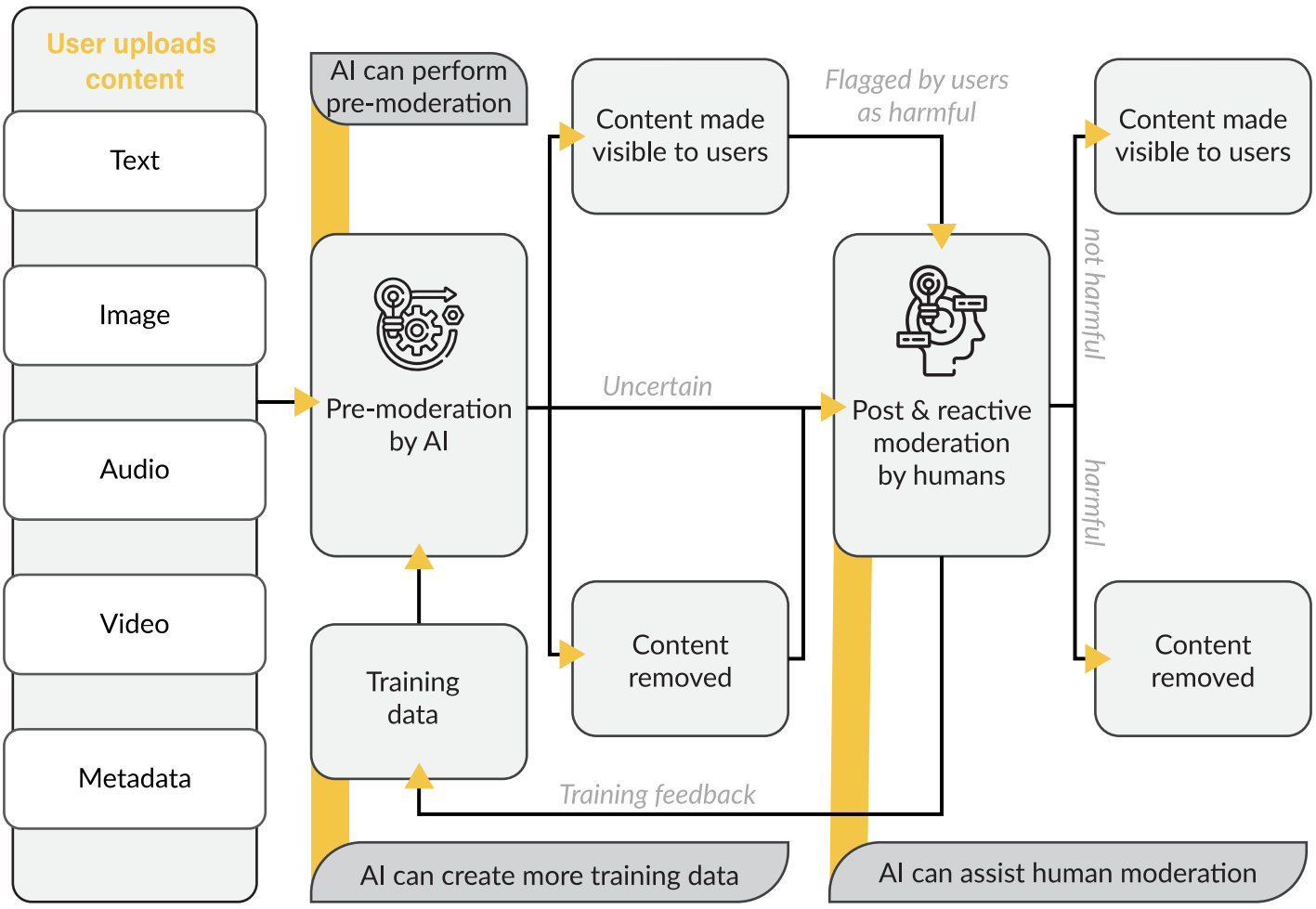
require setup by someone with great knowledge of moderation and industry trends. They need continuous review to ensure that the rules it builds on are up to date and accurate.

Machine learning can optimize this process even further, using an algorithm to learn from data and develop more sophisticated decisions over time. However, even machine-learning requires ongoing monitoring and tuning so you have to keep your smartest moderators close at hand.

Pros	Cons
Faster and more cost efficient than manual moderation.	Automated moderation still needs human involvement.
Time to site is instant which adds to user experience, which works especially well in marketplaces and communities for time-sensitive goods.	To optimize filters and rules, staff needs to be kept up to date on trends on the site as well as the industry.

Omni-Model Solution Design

Across the five models, there's no perfect solution. Here's one way to think of the entire process, visually:



Creating a content moderation process with a foundation of transparency, including regular review and reporting, is critical to the sustainability of the moderation process.

Many leading platforms do publish regular transparency reports, disclosing the actions that are taken with regard to the content on their platforms. Twitter, for example, hosts a web-based transparency center that includes current and historical reports data on rules enforcement, government requests for information, and platform manipulation.

Every site and every audience is different, though, and each will have unique guidelines and specifications for acceptable behavior by their users. A one-size-fits-all solution cannot completely address the complexities of content moderation. Automated solutions must also be customizable to the needs of each site.

When trying to implement or upgrade content moderation in a digital environment, it is best to consider an automated solution that works in real time, across multiple languages, and can evaluate the context of content to ensure accurate, effective, customizable moderation.

Can Platform Companies afford to hire more moderators?

The math is tricky and not always transparent, but the short answer is: yes.

Use Facebook as one example.

If we want to improve how moderation is carried out, Facebook needs to bring content moderators in-house, make them full employees, and double their numbers, according to a report from New York University's Stern Center for Business and Human Rights.

"Content moderation is not like other outsourced functions, like cooking or cleaning," says report author Paul M. Barrett, deputy director of the center. "It is a central function of the business of social media, and that makes it somewhat strange that it's treated as if it's peripheral or someone else's problem."

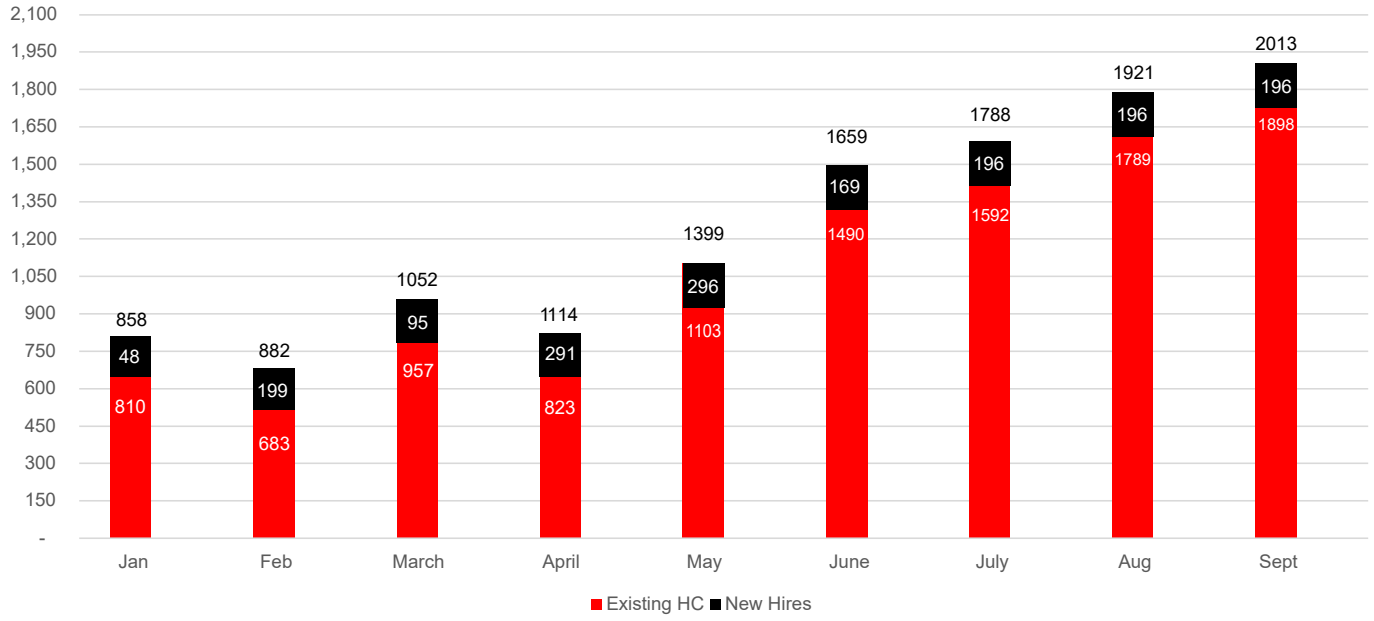
Why is content moderation treated this way by Facebook's leaders? It comes at least partly down to cost, Barrett says. His recommendations would be very costly for the company to enact—most likely in the tens of millions of dollars (though to put this into perspective, it makes billions of dollars of profit every year).

Actually, though, there is a way to do this at scale and cost.

What Conectys did with one social media giant

The client was a global leader in social media platform engagement, available in 155 countries and 75 languages overall. 83% of all users have posted at least once, and the user base grows annually roughly 5x.

We were brought in on moderation capabilities that have expanded to 5 sites, 15 languages, and over 1,800 moderators. Headcount growth among Conectys agents has scaled rapidly with this client.



From January 2020 to September 2020, we went from 858 headcounts associated with the account to 2,013, for a growth of 135%.

To help this client get moderators at scale and cost, we worked with 15 local agencies and 30 job boards across the globe. This pipeline allows us to add between 350-700 headcount per month when necessary. Conectys prides itself on flexibility with clients, which is normally displayed in terms of seasonal demand. But when a client is high-growth such as this client, we can grow upwards for a long period of time as well.

We have 25 in-house talent acquisition (TA) specialists, plus a flex force of 20 talent acquisition specialists that we can use during aggressive ramp-ups or seasonal demand increases. The flex force is managed by six local TA managers. Recruitment manuals and processes are globally-designed, but locally-customized, which allows for a homogenous approach to quality and delivery across the world.

In 2020, we began with a standardized Assessment Center for this client, which screens in only

the talent that matches certain client needs and competency models. This allows us to move faster on top talent, and retain them significantly longer.

Our median time to hire on this account for moderators was about 5.1 days, with an expert hiring timeline of 4.1 days.

This same social media platform client had a deep focus on well-being of moderators, a common trend of late, and we developed this model to support that need:

We gamified the moderator process on this account, creating a “Guardians of the Internet” style competition that showed content moderators how important their role was to this company, what they needed to do successfully, and encouraged contribution of best practices to a knowledge bank for future moderator learning.

During an account that existed largely in COVID lockdown, the gamification aspect kept moderators engaged and mentally healthy in a period where social interaction was lacking for many of us.

We also built out new office looks for this client in Poland, Davao, and Iloilo, both in the interest of social distance and safety requirements, but also vibrant color schemes and a chance to interact with fellow moderators when on-site.

Scale and Affordability is Possible

Moderation can be done at a global scale in an affordable way, and big companies don't need to bring it in-house. They can stay ahead of the regulatory environment, cost-contain, and have moderators not subjected to horrible content all day. Above is some of the roadmap. ***To learn more,*** *contact us.*

Contact us at sales@conectys.com

Or learn more from our Blog.