WHITEPAPER

Content Moderation: How to Protect Your Business and Customers

Content Moderation: How to Protect Your Business and Customers

Executive Summary

In the past 20 years or so since the internet gained mass adoption by the general public, its users have developed a sense of what to expect from their daily interactions with it. Nearly everyone has favorite sites that they enjoy visiting. From YouTube to Facebook, to Amazon and the New York Times, the internet has become embedded into our daily lives as a convenient tool to satisfy our desires for entertainment, social interaction, news and commerce. It may not always be a perfect experience, but with few exceptions, we feel safe using it.

But what if the internet wasn't safe? What if there was a dark underbelly to it that most users couldn't begin to fathom possibly? What if this dark, alternate universe's internet went beyond questionable language, but also entailed aspects that were vastly more sinister? Would users still feel comfortable using it? Would they feel comfortable allowing their children to do so?

Consider: In 1995, an estimated 7,000 images of child pornography were believed to be in circulation around the world. In 2012 - less than 20 years later – data collected by law enforcement revealed that this figure had risen dramatically to over 150 million images - in the United Kingdom alone! (1)

By nearly everyone's standards and morals, that's not dark. That's pure evil. And unfortunately, in a modern world where the internet reaches almost everywhere, child pornography is not the only troubling category to be found. All too frequently, content ranging from illegal drug sales, hate speech, murder, fraud, recruitment for terrorist groups and threats against minority groups and women have exploded throughout the internet.

For businesses, the proliferation of such content in these and other dicey categories (misogyny, racism, homophobia, etc.) can do tremendous harm to a business if it is seen to be complacent in addressing the problem in the eyes of consumers. Social media, especially, has given consumers a voice to publically shame, criticize and rebuke companies that, in users' opinions, are not doing enough to combat the problem.

One way that organizations address the issue is through content moderation. Social Media Today defines this as, "the practice of monitoring submissions and applying a set of rules that define what is acceptable and what is not. Unacceptable content is then removed." (2)

Businesses who moderate site content are often reluctant to admit to such culling, as it can lead to censorship concerns from consumers. For users who expect and are increasingly reliant upon technology in their daily lives, it can come as a shock that much of the content moderation work is churned out by AI and humans.

An October 2014 article from Wired, "The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed," explored the toll that such work can have on the employees who work in the field. [3] For companies in the business process operations sector, such as TaskUs, this type of work is more than a simple day-to-day functional concern, but rather one that has daily implications for employees' well-being.

The cross section of consumers' interests, business' functionality and employees' mental health requires thought leadership so that each of these concerns are more firmly addressed, resulting in a safer internet environment for users.

Over the course of this article, TaskUs will provide this thought leadership and offer the following to readers:

We turn now to the first item on the agenda—a brief overview of the business practices and strategies related to content moderation.

Social media, especially, has given consumers a voice to publically shame, criticize and rebuke companies that, in users' opinions, are not doing enough to combat the problem.

Overview of the Business Practices and Strategy Related to Content Moderation

As the internet has expanded its effective reach, the sheer volume of content available to read, to watch and to interact with each day is unfathomable. From a practical standpoint, it is impossible to interact with everything held within it. Google alone, as of June 29, 2016, has indexed an estimated 3.89 billion pages. (4) This is the equivalent of roughly one indexed page for every two people on the planet!

In November 2015, Facebook estimated that 500 million end users watched an estimated 8 billion videos.

Each day.

To put that into perspective, at only 3 seconds of viewing time per video, Facebook users are watching the equivalent of 760 years of videos.

Every day. (5)

It is hard to imagine, but YouTube's video viewership is broader than Facebook. The video behemoth has "only" been existence for a decade approximately, yet analytics show that over 1 billion people around the world utilize the site. That is roughly 1/3 of planet Earth's internet users. Perhaps more impressively, "every day, people watch hundreds of millions of hours on YouTube and generate billions of views." (6)

It is a lot.

In fact, no business in existence can completely wipe out the scourge of illegal or questionable content on its sites. Current technology and the fiscal constraints of any one business make it prohibitively impossible to accomplish such.

If it is impossible to eliminate such content, should companies throw in the towel, make the best of a bad situation and hope that consumers understand?

No.

As we will demonstrate in the next section, the risks to both people and businesses are simply too high. Organizations must protect both to the best of their abilities.

So how do businesses and organizations grapple with the magnitude of content that is consumed and produced by internet users every day? Outside of doing nothing, there are essentially five approaches to content moderation utilized today. (2)

1. **Pre-Moderation**: With this type of moderation, all submitted content is pre-screened by content moderators before publication and "going live." This strategic approach can be beneficial for sites that focus, for instance, on children and families. However, this diminishes the immediate gratification that many Web users expect. Additionally, such an approach is more challenging to scale as a site grows.

In November 2015, Facebook estimated that 500 million end users watched an estimated 8 billion videos.

2. **Post-Moderation**: Conversely, post-moderation results in most content going live immediately. All submitted/posted content – which begins as live content – is then queued for review after being posted. Such an approach, while admirable, can be cost prohibitive for many businesses. The volume of material is simply too large for any one team to handle.

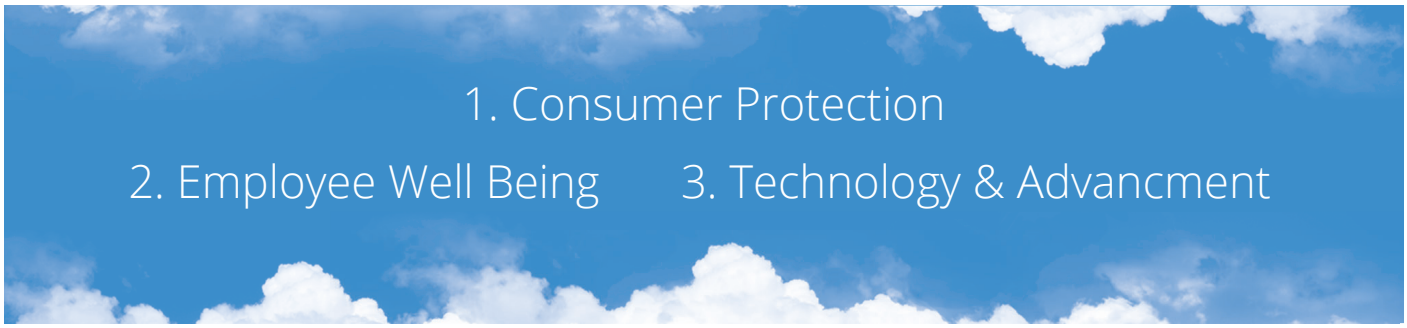
3. **Reactive Moderation**: This strategic approach to moderation brings users into the process by letting them "flag" content for moderator review that they deem questionable or unworthy for inclusion onto a site. Doing this provides for a certain amount of scalability, in addition to helping curtail costs. On the downside, users can have agendas and/or values that differ from an organization's intent for the flagging tool. For instance, political posts could be flagged by users on the opposite side of the issue in question, despite being fully compliant with both "house rules" and the law. This can lead to unnecessary work and expense for the business to absorb.

4. **Distributed Moderation**: This strategic approach to content moderation relies on a rating system to "vote" on content that does/does not meet site expectations. Content that does not meet a threshold is removed/not seen as widely as it otherwise would without such moderation. Reddit is an example of a site that uses a version of this tactic. Unfortunately, this is not a perfect strategic approach either, as users can and often do have interests that compete against an organization's interests or the law.

5. **Automated Moderation**: As it implies, an automated approach to moderation is instantaneous. Such moderation can block certain IP addresses, keywords and similar content, before it ever goes live. Unfortunately, such an approach often lacks context. For instance, automatically blocking keyword "breast" would likely have the unintended consequence of blocking educational or informational material on "breast exams."

Identification of the Issues and Related Recommendations

There are 3 primary issues that parties interested in content moderation should consider in today's market:



1. **Consumer Protection**: At TaskUs, we believe that content moderation is important work that provides an extraordinarily valuable service that helps to protect our partners' businesses that in turn, protects its customers.

Protecting consumers does not always mean only blocking porn or questionable language. When successful, it can be the difference between life or death.

The content that our moderators review – and the potential risks to the public, to consumers, and to businesses – runs the gamut from pornography, child abuse, financial fraud, spam and threats of violence. It is ugly, but critically necessary work in today's world.

For one of our partners, a team members actively discovered a school shooting threat that had been posted on the service. Lives were potentially - and presumably - in very real danger. As such, we didn't have days or even hours to solve the situation, we had minutes at best! The TaskUs team dropped everything and raced against the clock to involve key stakeholders. Our team activated our direct lines of communication protocol and escalated the issue immediately! We worked in rapid tandem with our partner to get the critical information they needed, while at the same time worked swiftly with law enforcement to provide potentially life-saving details to the volatile environment that was taking place in that moment. It was an all-hands on deck moment for the entire team!

Thankfully, this specific school shooting threat was a hoax. Our partner received public praise for how it handled this situation – something that we are very proud of as a company within content moderation space. With our help, our partner remains a safe place for teens.

What is our scope?

Each month, the broader TaskUs team reviews over 24 thousand videos, 5.3 million posts, and 68 million videos. Per review, our average response time is between 7.5 and 60 seconds, depending on the type of content.

While consumers can sometimes have very real concerns about content moderation related to censorship, the human cost to not moderate is simply too high to brush aside and ignore.

Consider: In 2010, a 12-year-old girl named Amanda Todd believed that she was chatting online with a boy her own age. In reality, the person on the other side of the screen was a pedophile. Over the course of this online conversation, she exposed her chest to this predator who took a picture of what he saw. This picture was distributed on a Facebook page called, “Controversial Humor”. By October 2012, Amanda had taken to YouTube to explain the harassment and emotional trauma that she had experienced, along with contemplating suicide.

Amanda, in fact, did commit suicide.

For Amanda and her family, it’s tragically too late, but as more young people use social media as a primary form of self-expression, a proportional response to content awareness and moderation is critical in order to save lives and to protect consumers.

Each month, the broader TaskUs team reviews over 24 thousand videos, 5.3 million posts, and 68 million videos. Per review, our average response is between 7.5 and 60 seconds, depending on the type of content.

2. **Employee Well Being**: As the previously mentioned Wired article indicated, providers of content moderation services must be actively aware of the human toll that such service offerings can have on the employees that do such work. More so, understanding both how workers view their roles, and safeguarding them should be primary concerns for businesses.

TaskUs sought to gain a richer understanding of our content moderators in June 2016.

Why?

Because we believe that our employees are the most important aspect of our business. We sent out a survey request to which 270 employees responded. For subjective questions, respondents were asked to answer on a scale from 1 (at the low end) to 5 (at the high end). The results were quite insightful, and painted a clearer picture of an often-misunderstood industry and of its workers.

81.2% of the survey's respondents were between the ages of 20 and 30. Bucking the norm of high attrition of workers in this segment, the survey revealed that 83.3% of respondents have been TaskUs employees for longer than 6 months, while 52.2% of workers have been with the company for longer than a year.

TaskUs' low employee attrition is especially important in light of the tremendous growth of user generated content (UGC) on the internet.

Consider: UGC dominates web content - 25% of the internet's search results for the world's 20 largest brands are related to UGC! Social UGC alone has increased by 176 million individual pieces of content over the past year alone. 2-4 billion photos are posted by consumers to social media daily. Year-to-year Pinterest pins are up 75% Each minute: Facebook users share 2.5 million pieces of content; Twitter users post roughly 300,000 tweets, Instagram sees 220,000 new posts and YouTube users generate 72 hours worth of UGC! (11)

Such explosive growth requires a stable, effective and scalable workforce. High attrition - such as the BPO industry's 80% average restricts the ability to fulfill a rapidly growing content moderation workforce. In sharp contrast, TaskUs has a scant 35% annual attrition rate. It's more than capable of helping its partners to handle the transformational growth of UGC.

Additionally, TaskUs surveyed teammates to find out whether or not they agreed with the following statement: "I am proud of my work as a content moderator at TaskUs because I know my work is keeping inappropriate content from being seen by millions of people."

A staggering 94.05% of respondents (with a weighted response average score of 4.65 out of 5) agreed. 91.08% of respondents (weighted average of 4.53) affirmed that they enjoy their work as a content moderator.

Respondent Neil Ongpauco, who had a stated tenure of between 12 and 24 months wrote, "I love working at TaskUs as a Content Moderator!"

However, job satisfaction is only one component to exhibiting a well-rounded focus on improving employees' lives. 85.13% of respondents (4.31 weighted) reported that they felt that TaskUs supported their physical and mental well-being. An impressive figure to be sure.

So what does TaskUs offer its offshore employees to help them to feel valued and to give a clear indication that their work was important to the company?

We provide all regular employees with HMO health insurance. Medical insurance is also available for employees' families and LGBT partners.

Each of TaskUs' sites in Manila has a nurse on staff 24 hours per day, seven days per week. Further, TaskUs employs a psychologist that rotates amongst our sites to provide care for workers that may need such assistance, and to promote physical and mental health.

Our Manila-based offices are inspired by the spacious, open-floor plans of the startups located in Silicon Valley. They include: sleeping quarters, gyms, recreation areas and showers. We provide 120 days of maternity leave for new mothers, in addition to our paid vacation and sick leave policies.

Additionally, our TaskUs Scholars Program provides access to higher-quality private school education for the children of our Filipino employees than is available through public education programs. For the upcoming 2016-2017 school year, 112 employees' children were awarded scholarships totaling over $123,000 to cover the cost of private tuition, school books, backpacks and classroom supplies.

TaskUs considers itself a steward of not only its Filipino workers physically well-being, but also their emotional well-being. A phenomenon exists in the country called “Overseas Filipino Workers.” Because of financial constraints, these workers leave their family and friends behind in search of employment opportunities outside of the country. Once employed, OFWs send money back to their loved ones back home. The Philippine Statistics Authority estimates that 2.44 million of its population work outside of the country. For a country of 98.39 million people, that equates to 2.5% of the population. TaskUs is proud of the fact that we help to keep families and friends together by providing excellent jobs that otherwise would not exist.

Each minute:

- Facebook users share 2.5 million pieces of content
- Twitter users post roughly 300,000 tweets
- Instagram sees 220,000 new posts and
- Youtube users generate 72 hours worth of UCFC!

3. **Technology and Advancements**: As with other business sectors, technology promises to change the content moderation industry in ways that are both obvious and entirely unexpected. For one, despite the rapid increase of internet content available, TaskUs expects that technology, not additional workers, will pick up the slack in the future.

However, advancements in technology are not likely to negate the need for content moderators anytime soon. In some cases, the technology is not capable enough yet to replace a human worker. Context matters in such instances.

Consider: The news media regularly reports that terrorist organizations, such as ISIS, utilize the internet to recruit, to communicate to each other and even to claim responsibility for various attacks. After such terrorism instances, politicians frequently mention the need to work with tech companies to detect this activity online. However, this is not as simple as it would seem.

With certain forms of content – copyrighted video, for instance – sophisticated software can spot patterns and seek out infringing content. As The Guardian reported in 2014, the number of terrorists (relative to the broader population) is small enough to avoid pattern recognition in most instances. It also reports that terrorists (in their capacity as people) are adaptable, whereas copyrighted materials (music, for example) are static in nature. As a feature, adaptability lends itself to more easily avoiding detection. (7)

The future of content moderation is likely to bring with it exciting new technologies. One exciting next-gen company looking to deliver future tech is Clarifai. It is a new offering that points to a future with technology that does not

rely on relatively simple pattern recognition of stationary media (such as pictures) but instead analyzes video content using artificial intelligence and contextual clues. (8)

Another company causing technology waves in the content moderation sector is Image Analyzer. In 2015, it released version 6.0 of its product that it claims has, "the ability to identify up to 99 percent of sexually explicit images and videos. The adjustable engine sensitivity allows users to reduce moderation queues by 50 to 95 percent, depending on their preferred levels of risk." (9) Logically, smaller queues will result in lower organization costs and less reliance on human workers to review this content.

Whatever the future might bring, TaskUs and our thought leadership will guide the way. We believe in the promise of and the strategic implications of strengthening and supporting our existing human teams with powerful AI technology to broaden our coverage scope capabilities. We believe in providing exceptional working conditions for all of our teams. We believe in the fairness of a quality work-life balance. We believe in living out our values day-to-day. We believe in protecting our customers, their users and our employees equally.

TaskUs is poised not just for the future of content moderation, but also for the future business world that the Millennial workforce demands today.

Citations

1. BUNI, CATHERINE, and SORAYA CHEMALY. “The Unsafety Net: How Social Media Turned Against Women.” The Atlantic. Atlantic Media Company, 9 Oct. 2014. Web. 25 June 2016.
2. GRIMES-VIORT, BLAISE. “6 Types of Content Moderation You Need to Know about.” Social Media Today. Social Media Today, 07 Dec. 2010. Web. 25 June 2016.
3. CHEN, ADRIAN. “The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed.” Wired.com. Conde Nast Digital, 23 Oct. 2014. Web. 25 June 2016.
4. “The Size of the World Wide Web (The internet).” WorldWideWebSize.com. Ed. MAURICE DE KUNDER. N.p., 29 June 2016. Web. 29 June 2016.
5. CONSTINE, JOSH. “Facebook Hits 8 Billion Daily Video Views, Doubling From 4 Billion In April.” TechCrunch. AOL, Inc., 4 Nov. 2015. Web. 27 June 2016.
6. “YouTube Statistics.” YouTube. YouTube, n.d. Web. 29 June 2016.
7. WEITZNER, DANIEL J. “Spotting Terrorist Behaviour Online Is Harder Than Finding Child Abuse Images.” The Guardian. Guardian News and Media, 04 Dec. 2014. Web. 25 June 2016.
8. NOVET, JORDAN. “Deep Learning Startup Clarifai Goes Beyond Image Recognition, Now Offers Video Analysis.” VentureBeat. N.p., 28 Apr. 2015. Web. 25 June 2016.
9. Image Analyzer 6.0: Market Leader in Explicit Image and Video Scanning. PRWeb. Image Analyzer, 22 Apr. 2015. Web. 25 June 2016.
10. TABLE 1.1 Distribution of Overseas Filipino Workers by Sex and Region: 2015(n.d.): n. pag. Philippine Statistics Authority. Philippine Statistics Authority, 2015. Web. 1 Nov. 2016.
11. Dhamdhere, Prasad. “The Ultimate List of User Generated Content Statistics.” Social Annex Blog. Social Annex, 1 Sept. 2016. Web. 02 Nov. 2