# Content Moderation Fails – When Content Moderation Goes Wrong

For such a simple concept, the field of Content Moderation isn't just a can of worms, it's a whole warehouse full of cans of worms.

Content moderation is in the news a *lot* and for many reasons. From a human point of view, there are the overworked and poorly paid content moderators many of whom are suffering from PTSD. You can find out more about the human cost of content moderation in this article.

Then there are the problems with policing huge amounts of data. From content that should never see the light of day - but does - right through to legitimate content that is blocked for the wrong reasons. And all this has to be balanced against what constitutes free speech, personal liberty, and different cultural standards.

The subject of content moderation is a controversial one on so many levels and the problems facing the platforms trying to police this are growing.

## Not all content is equal!

To be clear, the issues and gaps with content moderation are mainly with visual content.

Text moderation is a mature technology and when compared to moderating visual content it is relatively simple to do. Although there are still occasions where "contextual" relevance flags text incorrectly, moderating text content is pretty much sorted.

With visual content the story is different. This article looks at visual content moderation fails and what lessons can be learned from them.

## Visual Content Moderation Fails

Realistically, perfect content moderation is likely to be unachievable. However, there is no doubt that the big platforms can up their game. Vague and fluid rules applied by blunt detection systems that are dealing with millions of items daily is a recipe for disaster.

Without a doubt, the problems facing platforms that host user-generated visual content are growing exponentially. But the response from the platforms needs to be just as rapid, or cases like the ones listed below are going to be on that same exponential path.

### Pink Sisters Australia and NZ

The case of the Pink Sisters is not a "one-off". Rather, it has been chosen as an example of a "case type" that demonstrates the inherent flaws in current moderation procedures.

At the heart of the problem lies one of the prickliest issues facing visual content moderation – that of nudity. It is a reasonable assumption that no right-minded person would agree to open platforms like Facebook being awash with indecent images.

But what constitutes an indecent image?

In this case, Facebook decreed that the images posted by some members of the Pink Ladies were indecent. Some posts were removed and several members of the group had their accounts suspended.

The problem is that the Pink Ladies is a breast cancer support group that often posts pictures of breast reconstructions and mastectomies. This was a closed group and the pictures weren't available outside of it.

As reported by 9News Australia, members of the group were left isolated and distressed at suddenly being cut off from their support group.

This is by no means a unique case and the underlying problems are behind some of the big challenges the platforms face. Breastfeeding mothers, testicular checks, tattoo photos, and many other forms of legitimate photos have all been incorrectly flagged.

## Elsagate

The above example was what can happen when innocently posted and legitimate photos are incorrectly flagged. The Elsagate case is something far more sinister.

The Elsagate fail centers on YouTube for kids. A supposedly highly-moderated channel of kid-orientated content. But it wasn't any clever moderation that exposed the scale of the failure, rather it was a Reddit subgroup that tracked down supposedly kid-safe videos that in reality were full of horrific content.

Amongst the content that "slipped the net" were videos that exposed children to drug and alcohol abuse, cannibalism, gore, rape, and murder. Many of these videos featured these acts being carried out by favorite children's characters including Mickey Mouse, Elsa (from Disney's Frozen), Peppa Pig, and Spiderman.

This wasn't a complex operation that had gone to great lengths to throw moderators of the scent. All it took was some kid-friendly thumbnails and innocuous tags to fool YouTube's moderation controls.

YouTube's response was to label the content as "an extreme needle in a haystack" case. However, critics were not happy with the company's reaction and the incredibly poor moderation procedures that let such blatantly obvious inappropriate content onto a purportedly kid-safe channel.

## The Language Barrier

One example that perfectly demonstrates the limitations inherent in many current visual content moderation tools is the case of the explicit child abuse photos that were uploaded to Facebook from Papua New Guinea.

These appalling images were able to slip through moderation safeguards simply because the language used to caption the images was obscure. In this case, the language used was a local dialect called Tok Pisin, a language not included in Facebook's auto-filter tools.

The takeaway from this case is that even though the vast majority of illicit content is visual, the moderating tools still have an over-reliance on text-based mechanisms to identify inappropriate visual content.

## Misinformation – The Fake News Battle

One of the most public moderating fails of recent times is the ongoing "Fake News" debate. The matter rose to public prominence in the wake of Russian interference in the US election process. This meddling provoked a public backlash and the big platforms moved to improve moderating controls to try and tighten up on the spread of misinformation.

This was tested when the Covid-19 pandemic struck and the kindest thing that can be said is – "could do better."

The pandemic provoked the propagation of a huge surge in posts spreading false information about the pandemic. The problem was compounded by the same pandemic forcing many of the moderating services to halt work because of lockdowns and health concerns. The effect was platforms became over-reliant on creaking auto-filters and algorithms to verify the veracity of content.

As the world locked down on an unprecedented scale, the public turned to social media platforms for information, advice, and guidance. These platforms were rife with false information – conspiracy theories, vaccination disinformation, fake cures, and pandemic deniers, all the major platforms were hotbeds of coronavirus disinformation.

Genuine news and public information channels struggled to get their message across against this wall of misinformation. For many members of the public, this misinformation led to extreme stress and panic attacks.

# Content Moderation Fails – Other Considerations

These are just a few cases from the many thousands that could have been covered. They were selected as case examples that demonstrate some of the specific problems faced as companies struggle to moderate hundreds of millions of images and videos each day.

It is these numbers that are threatening to overwhelm moderating mechanisms. To cope with this volume speed is essential - This is the first consideration: -

- **The Importance of Speed –** Illicit content must be identified and removed quickly. A lot of damage can be done by the wrong content being available for even a short time. Live streams are a particular vulnerability. In one tragic moderation fail a white supremacist terrorist in New Zealand live-streamed himself carrying out a horrific attack on a mosque that left 51 dead. By the

time the stream had been removed and the user's account closed, many people were traumatized by what they'd witnessed on a trusted platform.

For content moderation to be classed as successful, this type of failure cannot be allowed to happen.

It is fair to say that no right-minded person would disagree that content like the above-mentioned should be moderated. But what about files and information shared on private channels and messaging apps? -

- **Sharing illegal content on messaging apps –** This is a controversial area that strays slightly from the scope of the article but is worth an honorable mention.

  Encrypted platforms like WhatsApp are frequently used to share harmful content, with users protected by the promise of end-to-end encryption. However, despite the terms and conditions stating that WhatsApp can't read or listen to messages, they frequently do, as described in this Gizmodo article.

  Whether they are right to do this or not, is a whole different can of worms. The major fail here is a complete lack of industry transparency over the exact terms of end-to-end encryption.

# Looking Ahead

This might all seem like doom and gloom, and certainly stopping every single bit of harmful content will be almost impossible. But we can do much better, and the technology to do this already exists.

Computer vision incorporating cutting-edge AI can scan content in real-time with accuracy levels that seemed impossible a few short years ago. There is no human solution to a problem of this scale, it is simply unfeasible. But it is a void that AI is stepping into and with increasing success.

# Want to Talk?

Visua is leading the charge in transforming how content is moderated. The Visua platform can perform real-time moderation of all visual content. Effectively it assesses content with human knowledge and intuition but at computer speeds.

To learn more about our industry-leading technology you can check out our video on the subject or visit our Visual-AI for content moderation page. If you want to talk more about our content moderation solutions, simply fill out the form below and we will quickly get back to you.