WebPurify

44 Ways Brands are Using Content Moderation

Introduction

Overview of Content Moderation

Content moderation is the practice of monitoring and managing user-generated content (UGC) to ensure it aligns with a platform's guidelines and community standards. Content moderators don't make the rules, we enforce them. And as more work and personal lives are spent online, moderation has become a cornerstone for maintaining the integrity and safety of online environments. From banking to food service to social media, the importance of content moderation across industries cannot be overstated, as it directly impacts user experience, brand reputation, and legal compliance.

The nature and scope of content moderation has evolved significantly over the years as the explosion of UGC transformed the way we communicate, engage, and do business online. What started as simple text filtering expanded to include sophisticated AI and human-hybrid models capable of detecting nuanced forms of inappropriate content, misinformation, and harmful behavior. Today, UGC is ubiquitous, appearing everywhere from food receipts to live customer service chats, and in 2024, brands that overlook the importance of content moderation do so at their peril.

"In earlier years, most brands that approached WebPurify worried more about inappropriate content than overall optimization of their platforms. We helped them catch the clear-cut bad stuff," explains Bartell Cope, WebPurify's VP of Sales. "Present-day moderation is far more comprehensive, including everything from ensuring a minimum level of quality for uploaded videos, to checking for intellectual property (IP) infringement or vetting the identity of new accounts on e-commerce, social media or dating apps."

Ignoring the many possible touchpoints that exist nowadays can lead to reputational damage and loss of customer trust. As this list will demonstrate, the applications of content moderation are vast and varied, underscoring its importance across all industries, sectors, and digital interactions.

Why We Don't Name Our Clients

All 44 use cases presented in this ebook are a mix of real partnerships and case studies with our clients and other examples that are common within the industry.

However, the nature of content moderation necessitates a high level of discretion, so we refrain from mentioning our clients' names unless explicitly permitted to do so. This approach ensures that we maintain the confidentiality and trust of the brands we protect.

Content moderation is a versatile tool that applies to a vast array of companies. Even if you believe your industry doesn't require such services, this ebook aims to illustrate the many reasons that assumption may not be entirely true. Our list showcases the myriad ways in which companies leverage our services to meet their unique needs, highlighting the flexibility and relevance of content moderation across different verticals.

Content Moderation Case Studies

Click below to learn more.

Cutting Fakes & Embracing Feels

How the team behind the **HUD** dating app worked with WebPurify to keep users safe and grow their platform.

Policing Pixels

Why immersive game creator **ForeVR** partnered with WebPurify to ensure safety in the metaverse.

Creating a Safer Space on Social

How WebPurify helped social network **WeScoop** provide a safer and more enjoyable experience for its users.

Protecting Your Brand

How WebPurify leveraged its text moderation APIs to moderate product customization for **TaylorMade**.

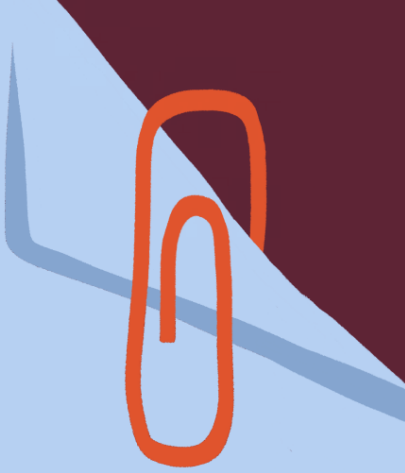
Moderating Engagement Campaigns

How WebPurify designed a custom solution to moderate user-generated videos for a **Pringles** Facebook campaign.

Managing Live Streams

Why **Microsoft** partnered with WebPurify to monitor live stream messages during the flight of Pegasus II.

Content Moderation Use Cases



ONE

Customer Service & Interaction

1.

Offensive messages on customer orders and receipts

In the past, one mishap might mean losing one customer, but in the age of social media, one mishap can be seen by millions. Moderating text on customer receipts or flagging inappropriate order or ticket names is crucial for protecting your brand image from potential harm, whether that's malicious or irresponsible employees making digs at customers, or innocent typos that could become an embarrassing viral moment.

2. Abusive content in live chat and image submissions

Live chat support is an increasingly valuable tool for delivering targeted customer service, but it can also be a venue for abusive content. Moderating live chat text and image submissions helps prevent harmful interactions, ensuring that both your customers and support staff are protected from offensive messages.

> Moderating live chat text and image submissions helps prevent harmful interactions, ensuring that both your customers and support staff are protected from offensive messages.

3. Irresponsible messages or harrassment in internal comms

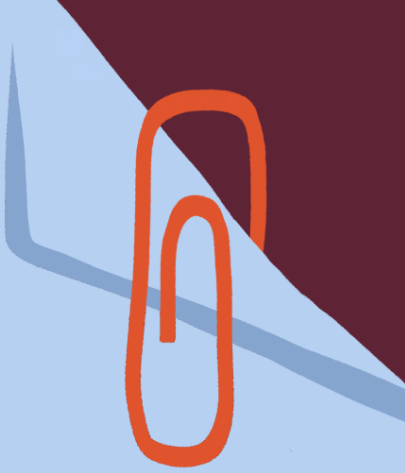
We work with a number of brands, both big and small, to moderate communications between employees, which helps prevent coworker harassment and irresponsible or risky interactions. By monitoring emails, messages, and other forms of internal communication (after, of course, disclosing this to employees during the hiring process), brands can ensure a safe and respectful workplace.

4. Grocery pictures for replacement items

In the booming grocery delivery industry, customers and shoppers often exchange images to request or suggest replacement items. Most photos are legitimate, but if you allow image submissions from the public and between strangers you are opening the door to potential misuse. We work with brands to moderate these images and ensure that uploaded content is appropriate and relevant to the request. This helps our clients provide accurate replacements, while improving customer satisfaction.

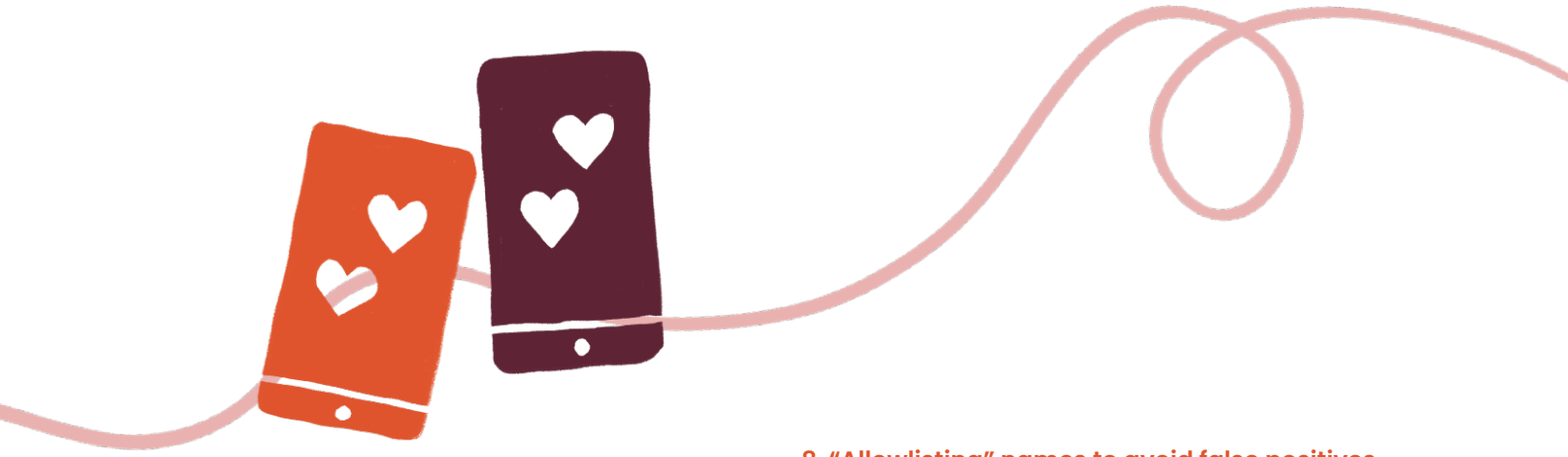
5. Images confirming food delivery

As with the grocery industry, food delivery services grew exponentially during the pandemic, and confirming delivery with images is a common practice. Moderating these images, as well as images in direct messages between drivers and recipients, ensures that they clearly show the delivered items and nothing else (including anything NSFW), avoiding any potential misunderstandings or disputes.



TWO

Personal Profiles & UGC



6.

Dating profile pictures

As in the offline world, first impressions are often visual. From startup dating apps to well-established players in the market, we offer several methods of quickly and accurately screening profile pictures. Some clients may allow suggestive or even explicit images but have clear boundaries demarcating what's considered unacceptable; other platforms may have stricter policies to accommodate users with more reserved values. Clients can tailor how they use our AI models to suit their guidelines, creating a safe space for users seeking genuine connections.

7. Dating profile text

Text in dating profiles must also be reviewed to ensure it adheres to community guidelines. We work with many clients to check profiles for offensive language, inappropriate content (including links directing users off-platform), or misleading information. Combined with profile picture review, checking text also helps us flag potentially fake or fraudulent accounts.

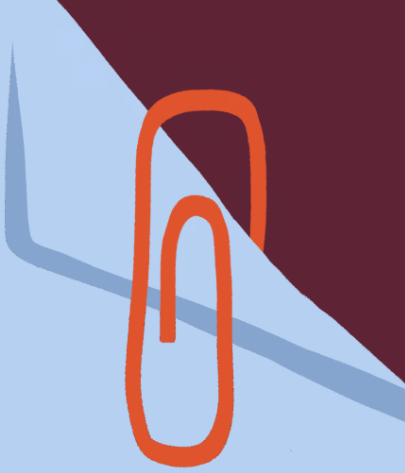
8. "Allowlisting" names to avoid false positives in personalization campaigns

Personalization campaigns often involve people's first and last names, which sometimes trigger false positives in moderation systems due to their similarity with inappropriate words. This is particularly true for global campaigns where a brand's usual list of acceptable first and last names doesn't take names from other cultures and geographies into account. Creating an "allowlist" of permitted names helps prevent these types of errors, ensuring legitimate first and last names are not wrongly flagged or censored.

9.

Webinars and conference calls

Webinars and livestreamed conference calls represent high-risk environments for unfiltered and potentially inappropriate comments. Moderating in real-time is challenging, especially when the audio is messy with multiple people speaking at once. We work with a number of clients, using both AI and human teams, to monitor conversations and ensure professionalism and decorum.



THREE

E-commerce

10. Product listings

Product listings are the backbone of e-commerce, and their accuracy and appropriateness determine a platform's success. We moderate listings for a large number of e-commerce brands to ensure all product descriptions, titles, and images adhere to the platform's standards. This includes filtering out offensive or misleading content and verifying that all information is truthful and relevant.

11.

Generative AI images in reviews

With the rise of generative AI, even product and user reviews can be of questionable authenticity. Moderating AI-generated images and videos is just as important in e-commerce as it is in other industries, ensuring customers can be confident reviews are genuine and free from offensive or misleading information.

12. User-uploaded images to customize products

Customization gives users the option to upload images and personalize products. Whether it's adding photos to custom-made t-shirts, patterns and designs to shoe brands, or creating bespoke book covers, our AI and human teams are able to moderate these user-uploaded images before the order is processed and paid for. Verifying that each purchase is appropriate removes the risk of offensive material appearing on branded products and protects against copyright infringement.

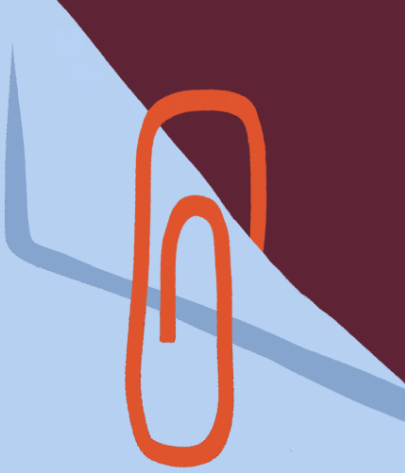
13.

Customer engagement campaigns displayed on big screens

We probably don't need to explain the obvious risks around customer engagement campaigns that feature UGC displayed on big screens in public spaces and sporting events. Surprisingly, many brands underestimate the public's capacity for bad behavior, or aren't aware that it's possible to moderate these campaigns in real-time.

> From in-store selfie contests to sporting events and political rallies, WebPurify offers a hybrid AI-human moderation system to screen content as it's submitted.

This can also be used to ensure customer engagement is on a brand's terms (for example, if it only wants customers to upload selfies taken in a specific store or city).



FOUR

Gaming

14. Abusive or inappropriate content within in-game chat

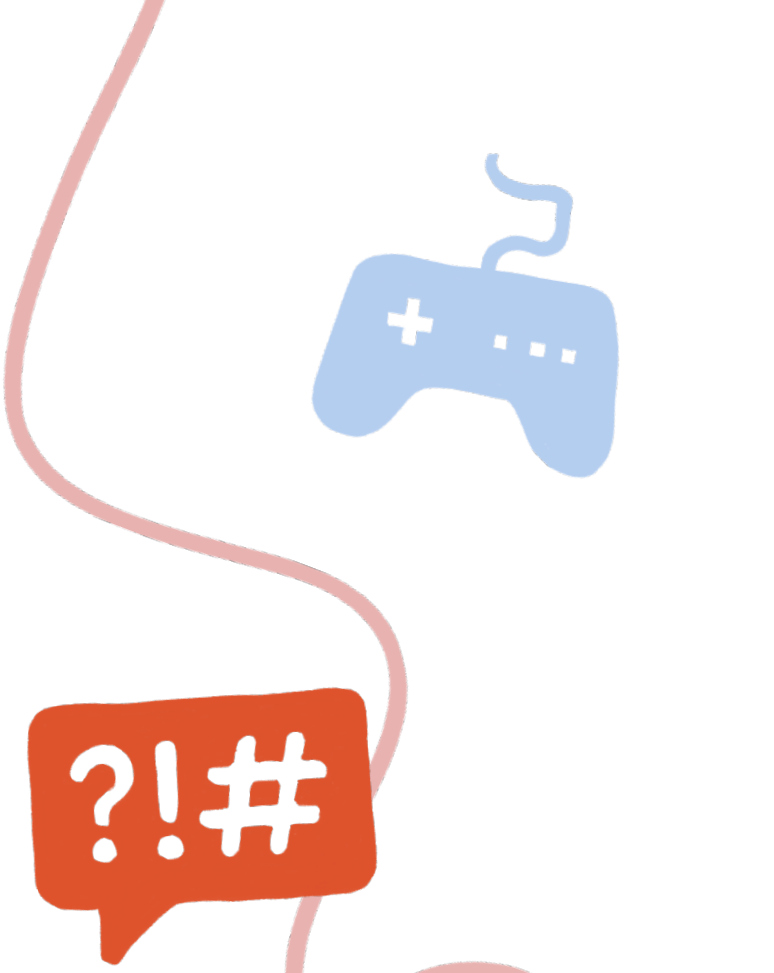
Real-time communication is one of the best-loved features of modern online gaming. Rally your team, razz the competition; it's good fun until it turns negative. We work with developers to moderate in-game chats, preventing the spread of abusive language and other inappropriate content. Filtering out harmful messages and maintaining a respectful environment ensures that all players can enjoy a game without encountering toxic behavior.

15. Custom avatars in video games

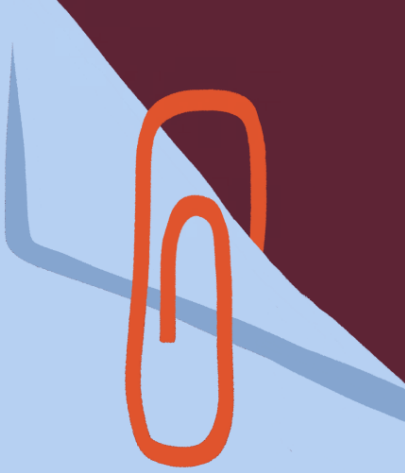
As with physical products, customization is a popular feature in many video games, allowing players to create unique avatars. Most players are respectful of a game's community standards, but you don't, for example, want someone using an image of Hitler for their gamertag. Similar to our work with dating profiles, we work with gaming platforms and developers to moderate custom avatars for offensive or inappropriate elements before they're put into use.

16. Custom items in video games

Game developers often allow players to create custom items, such as weapons, clothing, accessories, and even multiplayer maps. These also require diligent moderation, which involves setting up AI systems to screen items for inappropriate or offensive content, as well as verifying that they do not infringe on copyright or violate community guidelines.



Anything that scores in the gray area between pass or fail is passed on to our human moderators, who can make a more nuanced decision.



FIVE

Media & Entertainment

17.

Offensive content in movies

Content moderation in the film industry involves reviewing movies to identify and timestamp offensive footage such as inappropriate language, explicit scenes, violence or other sensitive material. Our work here enables streaming platforms to provide accurate viewer discretion advisories and to help studios make edits for different audiences.

18. Ad content

Online platforms across every industry rise and fall on their ad revenue, yet this crucial income-generator is often the last thing brands think about when it comes to content moderation. Not only should platforms ensure advertisements comply with their community guidelines, but also that legal standards are enforced. WebPurify's content moderation services can scrutinize advertisements for misleading claims, offensive language, and illegal content, among other criteria, which can be as specific as our clients' parameters and appetite for risk dictate.

19.

Inappropriate gestures or logos in amusement park photos

The photos taken at amusement parks are intended to capture moments of fun, but they sometimes feature inappropriate gestures or reveal more than they should. Our sophisticated AI model can moderate these photos at scale and speed, ensuring that only appropriate images are shared with or sold to visitors.

20. Offensive content in adult material

You might be surprised to discover that content moderation in adult entertainment is critical both for compliance and user safety.

> Our work in this industry involves screening image and video content for actual or apparent children, violence, abuse, scatological/bestial content, hidden cameras, scenarios that depict nonconsentual sex, deepfakes, and copyright violations.

It's difficult but important work, and our rigorous moderation ensures that adult content adheres to legal standards and ethical guidelines.

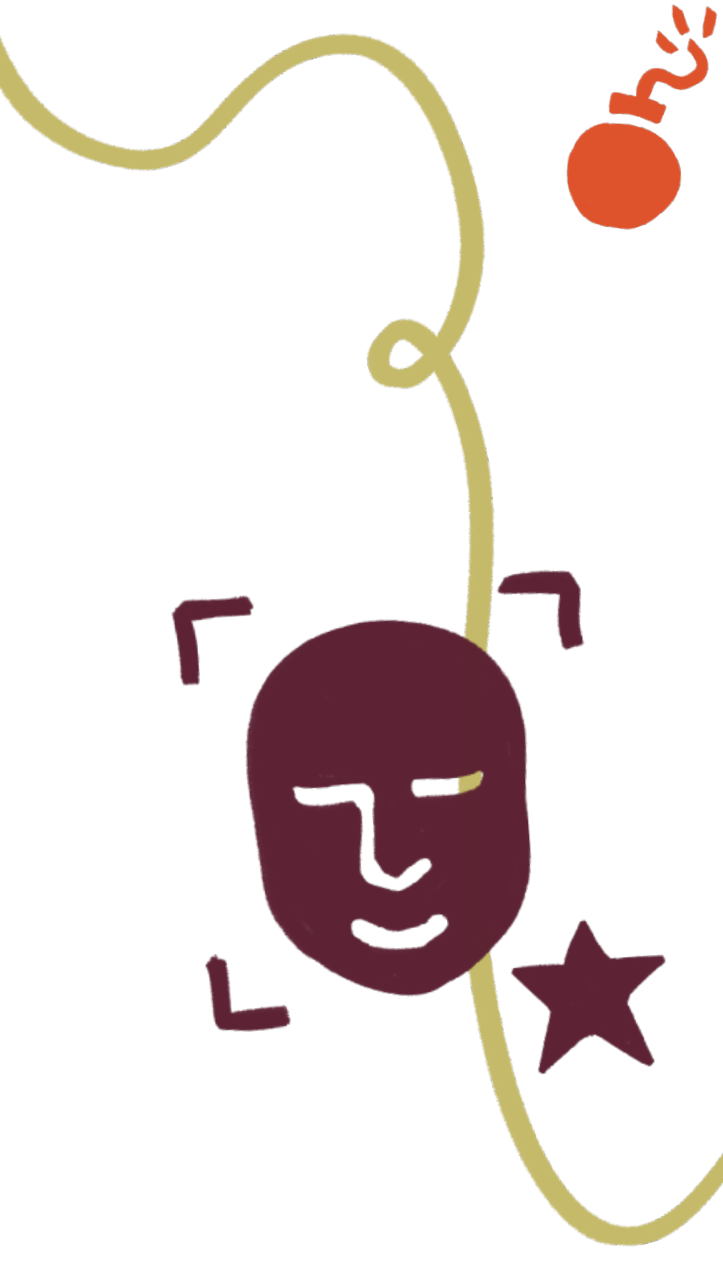
21. Compliance with payment processors on adult sites

Adult content platforms must adhere to stringent rules set by their payment processors. In other words, if these websites and their content creators want to get paid, they must consistently demonstrate the legality and ethical standards of their content.

We monitor for:

- CSAM
- Violence
- Non-negotiable red flags

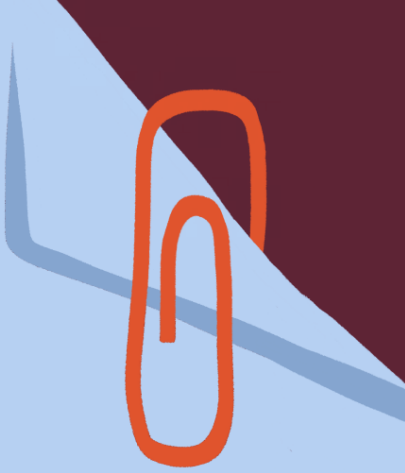
Maintaining compliance lets brands keep payment services, ensure performers are paid, and avoid operational disruption.



22.

Deepfakes and celebrity face swaps

In an age of highly advanced digital manipulation, deepfakes and face swaps pose significant risks to platforms allowing UGC, especially when they involve celebrities. Our AI image model is finely tuned—and constantly evolving—to detect synthetic media, and we work with numerous platforms to flag images suspected of being created with GenAI, ensuring platforms do not propagate misleading or unauthorized content.



SIX

Content Management & Data Labeling

23.

Stock image quality control

We're not always looking for the worst the internet has to offer.

Content moderation serves as an excellent layer of quality control for large data libraries, such as stock images. Our AI solution can check large volumes of content for resolution, correct labeling, possible IP infringement, licensing and appropriate quality thresholds, filtering out any images that don't meet the platform's standards.

24. Annotating stock photos and videos for categorization

Accurate categorization of stock photos and videos is essential for efficient search and retrieval. Content moderators can annotate these visuals, ensuring they are correctly tagged with relevant keywords and categories, making it easier for users to find what they're looking for.

25. Generative AI in interactive consumer campaigns

While many consumer brands are turning to generative AI to create engaging and interactive content for marketing campaigns, it's important this content aligns with brand guidelines, is appropriate for the target audience, and avoids potentially offensive or misleading information.

Our moderation solutions provide the necessary oversight, while enabling brands to take advantage of the time and cost benefits of using GenAI.

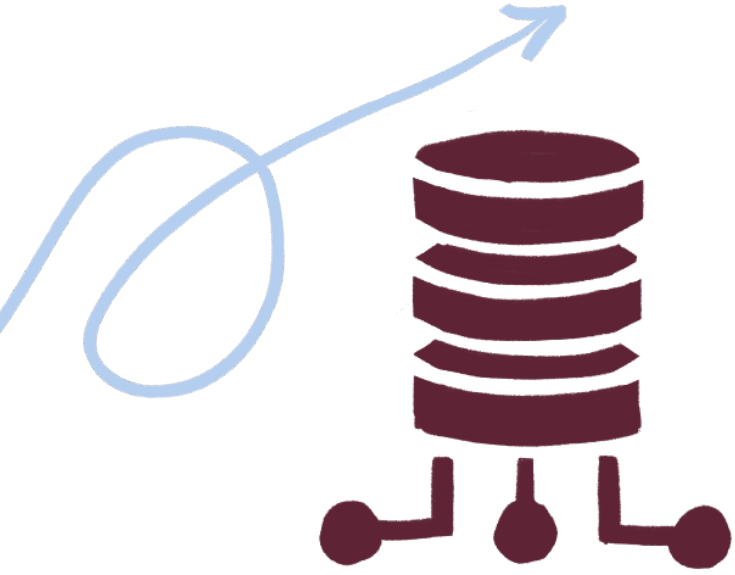
26.

Dataset labeling for AI model training

Training AI models requires large, accurately labeled datasets. Our content moderators play a crucial role in labeling data for clients, ensuring each item is correctly classified. It's a meticulous process, but one that enhances the accuracy and reliability of models, resulting in better performance and more trustworthy outcomes for a range of applications.

27. Dataset sorting for AI models

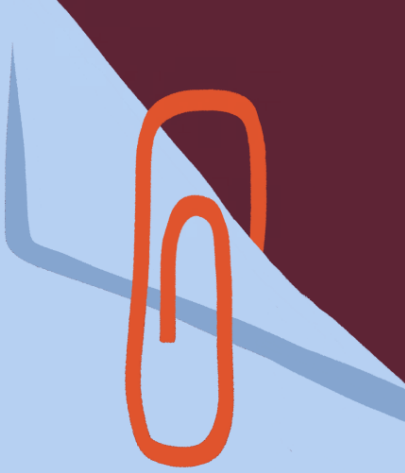
Similar to dataset labeling, improving AI models also involves sorting vast amounts of data to identify useful patterns and insights. Content moderators help organize and clean data, removing irrelevant or incorrect information. This sorting process refines the dataset, contributing to the development of more robust and effective AI models. The widespread availability of generative



28.

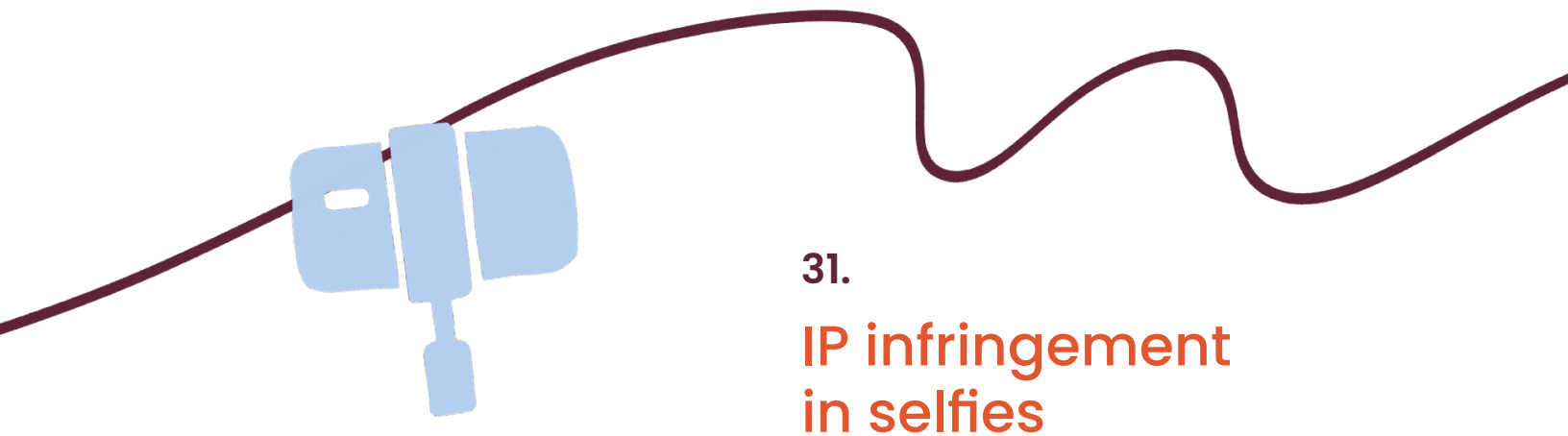
Reducing AI-generated scams

AI tools means that anyone can produce fake user profiles or even write scripts for interactions in dating apps. We work with numerous platforms to moderate this content and ensure people are who they say they are, and that interactions are genuine. Crucially, our moderation systems are fast and accurate, meaning users experience no lag when posting messages (even in running DM conversations) or images.



SEVEN

Proactive Moderation



29.

Misinformation and disinformation

Our hybrid model of AI and human moderators not only responds to reported inaccuracies but actively scans content in real-time to identify and flag potentially false or misleading information before it gains traction. This preemptive measure helps prevent the spread of harmful narratives, protecting users from being misled and upholding a platform's credibility.

30. Inappropriate content ('search and destroy' or 'secret shopping')

Proactive moderation goes beyond moderating content as it's uploaded or reported by community members and actively seeks out inappropriate content on platforms, a method often termed 'search and destroy' or 'secret shopping.' Our moderators and AI tools continuously monitor and scan platforms for content that violates guidelines. This includes targeted searches and 'hunting' for specific terms or behaviors known to be problematic. Inappropriate content is swiftly removed, ensuring it never reaches a wider audience.

31.

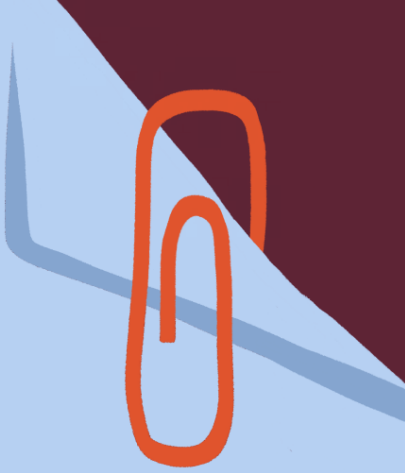
IP infringement in selfies

Ensuring that selfies and other user-uploaded images do not infringe on IP rights is a critical aspect of content moderation. Our moderators review these images before they go live, identifying and flagging any elements that may violate copyright or trademark laws. This proactive approach not only helps platforms avoid potential legal issues but also respects the rights of content creators and brands.

32. Verifying accounts

While initial vetting during an online account sign-up is crucial, follow-up vetting is also essential for ongoing compliance and security. Our moderators regularly verify that accounts are still being used by the individuals who originally signed up. This includes periodic checks and biometric verifications to prevent unauthorized access or misuse.

Continuous vetting is a key strategy for safeguarding a platform's trustworthiness and security.



EIGHT

Security & Surveillance

33.

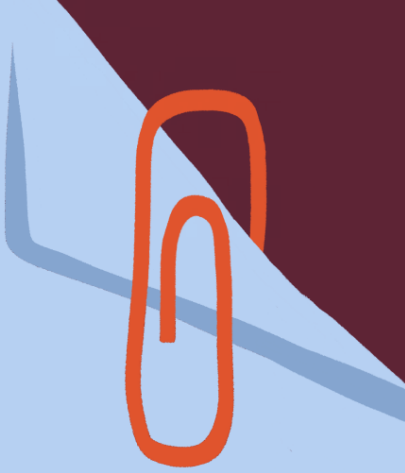
Suspected weapons in security camera footage

For our clients in security and surveillance, speed and accuracy is of the utmost importance.

Using advanced AI tools, our system scans live and recorded footage to detect firearms, knives, and other dangerous objects. When the AI identifies a potential weapon, the footage is flagged for immediate review by human moderators who verify the threat's validity and can then dispatch authorities to the scene.

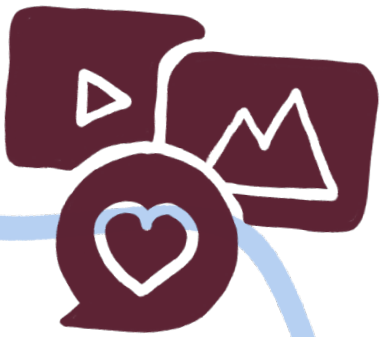
34. Dash or police cam footage

Dashcams and body-mounted cameras worn by police capture real-time events that require careful moderation to ensure accuracy and compliance. Reviewing and filtering out sensitive or inappropriate content ensures that only relevant material is used for evidence or public dissemination, while flagging potential breaches in police protocol to the appropriate oversight boards. It's a fine line, and our clients rely on this process to maintain transparency, accountability, and trust in law enforcement activities while protecting the privacy and rights of the individuals involved.



NINE

Specialized Use Cases



36. Backend communication moderation with Twilio

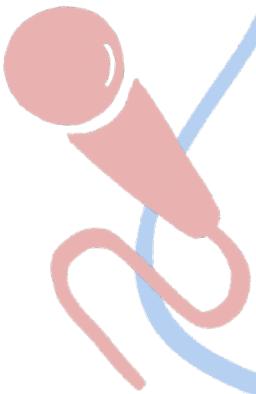
Twilio's communication infrastructure can be integrated with WebPurify for backend moderation, ensuring that all forms of communication, including SMS, voice, and video, adhere to community standards and guidelines. By leveraging Twilio's API, our moderation tools actively monitor and filter communications in real-time, ensuring they adhere to community standards and guidelines. This integration allows us to detect and prevent the spread of abusive or inappropriate content before it reaches users.

37. Content moderation SOPs and playbooks

Retaining our Trust & Safety consulting services enables brands to develop and implement Standard Operating Procedures (SOPs) and playbooks for content moderation. These detailed guidelines provide a structured approach, ensuring consistency and compliance across a platform. SOPs and playbooks can be tailored to each brand's specific needs, helping to streamline moderation efforts and enhance the overall effectiveness of your strategy.

35. Turnkey live team moderation with Cloudinary

Among other partnerships, we integrate with cloud-based image and video management platform Cloudinary to enable turnkey live team moderation. This means we can deliver real-time content moderation within its UI. This integration provides clients with easy access to Cloudinary's powerful media management capabilities, while ensuring images and videos stored therein are moderated efficiently and effectively.



38.

Moderation stress and pen tests

Stress testing and penetration testing (aka pen testing) are a common way for platforms to evaluate the strength of their existing moderation defenses.

By getting into the mindset of a 'bad actor' and simulating attacks, WebPurify attempts to bypass moderation defenses to identify vulnerabilities or weaknesses in brands' existing moderation models. Rigorously testing said defenses enables platforms to strengthen security measures, identify weaknesses and withstand real-world challenges.

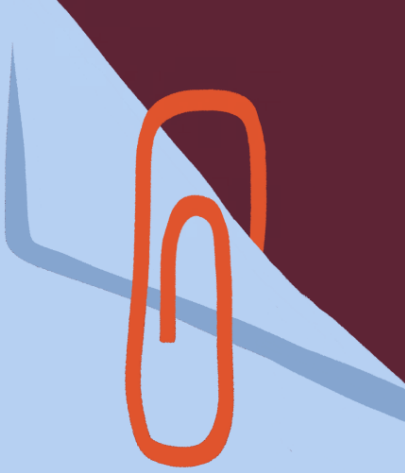
39. Profile verification through selfies and gesture matching

Impersonation, account hijacking, and unauthorized access are some of the biggest issues that plague platforms with user accounts. We work with clients to verify the legitimacy of user profiles by requiring new account holders to upload selfies while making specific gestures with their hand. Our moderators and AI systems review the selfies and gestures to confirm that the person creating or accessing the account is who they claim to be. Our clients in the dating industry, social media, and even financial services, rely on this verification process to maintain the authenticity and trustworthiness of their platform.

40.

Retroactive moderation

Retroactive moderation involves bulk moderating old posts and profiles to identify and address content that may have been overlooked when moderation safeguards were less stringent or incomplete. (Our service can do this in an automated manner, crawling a platform's pages and moderating any content encountered.) This might happen when a client acquires another website, or when a platform grows and wants to tackle issues it might not have had the resources to solve in the past.



TEN

Virtual Reality / Augmented Reality

41. Live moderators in-app to 'usher' VR experiences

In virtual reality (VR) environments, our live moderators are on the lookout for bad behavior, but they also serve as in-app ushers, guiding new users through their experiences and helping them get the hang of the controls. This hands-on assistance helps make user adoption smoother and faster, ensuring that newcomers feel comfortable and confident navigating the VR space.

> By providing real-time support, our live moderators effectively serve as ambassadors of the game and encourage greater engagement within the platform and higher rates of user retention.

42. Secret shoppers to identify and report harassment

Some of our clients deploy our moderators as 'secret shoppers,' who operate covertly within VR environments to identify and report harassment. These moderators blend in with regular users, observing interactions and flagging inappropriate behavior (sometimes after intentionally putting themselves in positions that might invite an unscrupulous user to act out). This proactive approach helps to deter potential harassers and ensures that all users can enjoy their VR experiences without fear of abuse.

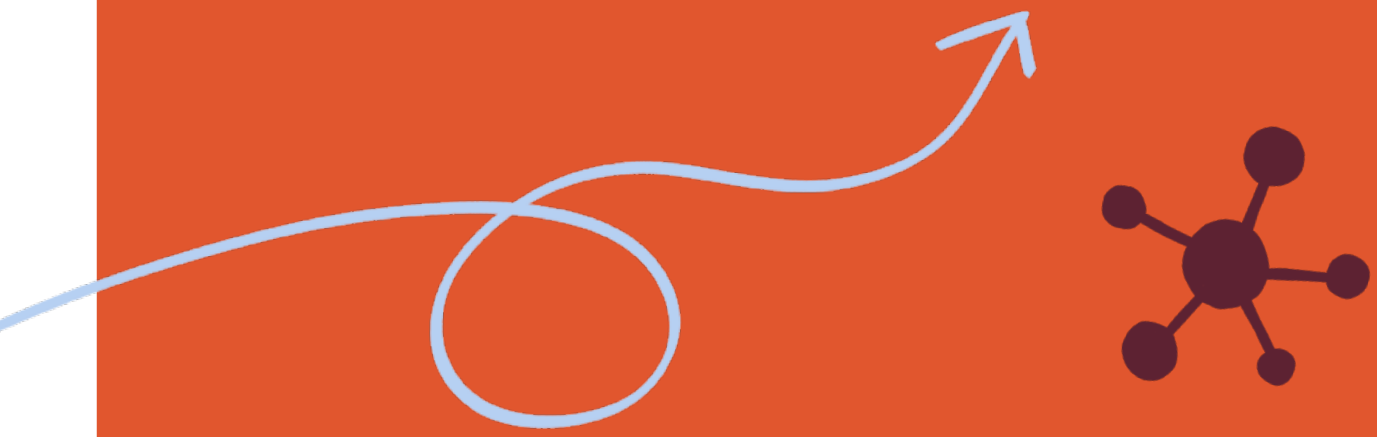
43. Dynamic focus groups for UX feedback

Our moderators can also serve as focus groups within VR/AR applications, providing excellent feedback to developers. After using the apps and taking detailed notes, our moderators give product teams valuable real-time evaluation and insights into the user experience (UX). Clients tell us this helps them identify areas for improvement, enhance usability and make informed decisions about updates.

44.

User surveys for feedback

In VR/AR environments, moderators can conduct surveys to gather user feedback from within the virtual space. Interacting with users and collecting their opinions and suggestions enables moderators to help platforms understand their customers' needs and preferences. This feedback is crucial for improving the VR/AR experience, helping platforms to tailor content and features to better meet user expectations and desires.



Future Trends in Content Moderation

We've explored 44 diverse and innovative ways that brands are leveraging content moderation with WebPurify to enhance their operations and boost user experiences. From customer service interactions to e-commerce, gaming, media, security, and VR/AR, content moderation plays a crucial role in maintaining safety, compliance, and brand reputation across a whole spectrum of use cases. And it's constantly evolving. New needs and use cases for content moderation are cropping up every week.

Looking ahead, content moderation will continue to evolve alongside advancements in technology and shifts in user expectations. AI and machine learning will become even more integral to the process, offering increasingly sophisticated and efficient ways to detect and manage the different types of inappropriate content, which themselves are growing harder to detect by the day.

As AI technologies advance, they also create new opportunities for human moderation, from reviewing data to train AI models to overseeing AI-generated output. Human moderators play a vital role in refining the accuracy of AI systems, ensuring nuanced judgment and ethical oversight remain part of the content moderation process.

AI and automation technologies are becoming more adept at detecting nuanced forms of inappropriate content, misinformation, and harmful behavior in real-time. As AI systems become even more advanced, they will not just identify bad content but predict potential future risks, allowing for better preemptive action. This proactive approach to content moderation will be crucial in ensuring platforms remain safe and trustworthy, particularly with the rise of synthetic media.

Conclusion

Emerging Use Cases

In the future, we will also see the emergence of new use cases for content moderation, and before long our list of 44 here won't seem so exhaustive. As VR/AR technologies become mainstream, there will be a growing need to moderate immersive environments to prevent new forms of harassment and ensure user safety in ways we haven't yet imagined. And as technology such as haptic suits becomes more commonplace in VR environments, we must learn how to effectively moderate new elements of risk.

Similarly, as the Internet of Things (IoT) expands, content moderation will extend to smart devices and interconnected systems, safeguarding users from malicious content and security breaches. Moreover, the rise of decentralized platforms and blockchain technology will introduce unique moderation challenges and opportunities, requiring innovative solutions tailored to these new digital ecosystems.

Building Trust & Community

At the heart of all these advancements is the goal of building trust and fostering healthy online communities. Effective content moderation will continue to be a cornerstone in achieving this, helping platforms to uphold their standards and provide safe spaces for users to interact and engage. As brands leverage these cutting-edge moderation technologies and strategies, they will be better equipped to navigate the complexities of life online, ensuring their platforms remain resilient, respectful, and relevant amid the rapid pace of change.