



# The unseen side of advertising

Find out why content moderation is essential to ensure ethical and effective ad campaigns and mitigate exposure to brand safety risks.



# Contents

- 03** Why content moderation in advertising matters
- 06** A brief history of brand safety
- 08** The current advertising landscape
- 12** Brand safety in the digital age
- 16** Building a brand safety practice
- 19** Understanding and combating ad fraud
- 23** Collaborating with sales teams and advertisers

# Why content moderation in advertising matters

---

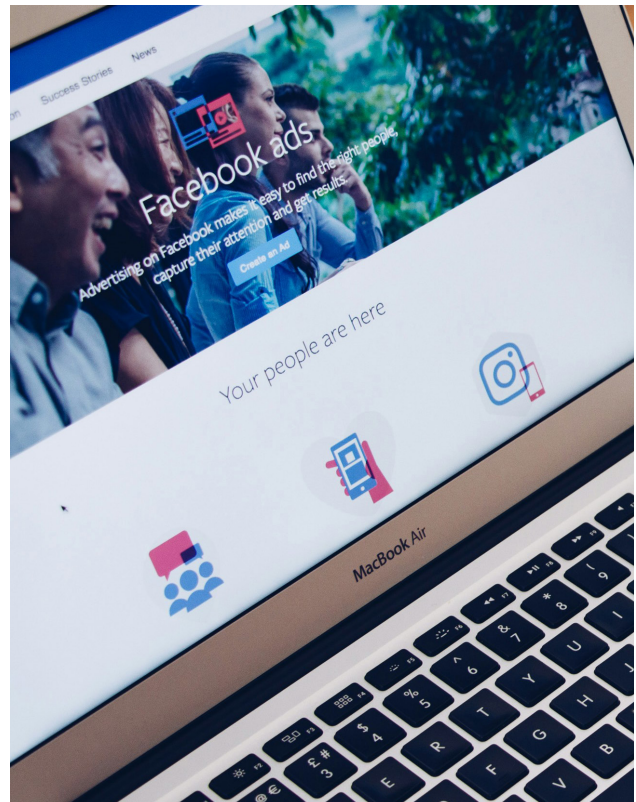
The online advertising market is complex and dynamic – and currently on track to realize a value of \$1 trillion. [One forecast](#) estimates that worldwide digital advertising spend will reach \$835.82 billion by 2026.



It started with the seed of an idea in 1994. The first online advertisement was a simple banner ad for AT&T, displayed on HotWired, an online offshoot of WIRED magazine. Fast forward two years, and DoubleClick demonstrated the ability to track how digital ads were performing.

With the launch of Google in 1998, and its subsequent decision to sell keyword-based ads in 2000, the true potential of the digital advertising landscape came into sharp relief.

Advertising remains the largest means by which much of the internet remains open and free to access. [As Mark Zuckerberg puts it:](#) “If we’re committed to serving everyone, then we need a service that is affordable to everyone.



**“At WebPurify, we are trusted by many of the world’s largest platforms to ensure their ad ecosystems are legally compliant, promoting a safe and positive user experience, and protecting the reputations of the platform and their advertisers”**

**ALEXANDRA POPKEN,**  
VP OF TRUST & SAFETY,  
WEBPURIFY



The best way to do that is to offer services for free, which ads enable us to do.”

Today, social media advertising offers the scale, engagement and sophisticated targeting tools that enable marketers to reach specific, relevant audiences for a product or service.

Although some social media platforms have recently taken steps to diversify their revenue streams – such as Twitter/X’s shift to subscription-based benefits and the shoppable features offered by YouTube and TikTok – advertising remains the primary business model for most major platforms.

The rise of [User-Generated Content \(UGC\)](#) provides an ongoing challenge for platforms that host ad content. Not only can bad actors use advertising as a means to [amplify scams](#) and [disseminate misinformation](#), which harm users, but the placement in which legitimate ads are seen poses risk too. Users may assume that ad adjacency to potentially harmful content implies that the advertiser is endorsing or even sponsoring this content. This perception, valid or not, stands to harm the reputations and bottom lines of both the advertiser and the platform. Brand Safety is, among other things, a form of content moderation that aims to prevent the monetization of harmful content online, and brand safety concerns have cost platforms billions of dollars in recent years.

Take the [‘Stop Hate for Profit’](#) ad boycott, for example. In June 2020, a group of organizations, including the Anti-Defamation League (ADL), Color of Change and Common Sense, persuaded more than 1,200 companies and brands – such as Levi’s, Honda and Starbucks – to pause advertising on Facebook for the whole of July.

Its message to Facebook was to ‘stop valuing profits over hate, bigotry, racism, antisemitism, and disinformation’ – and it reportedly [cost Facebook more than \\$7 billion](#). YouTube has also been hit by high-profile ad boycotts from brands concerned about their messaging being viewed in the same space as [offensive content](#).

More than ever, content moderation plays a critical role in building trust with users and advertisers, preventing costly boycotts, and ensuring ethical and effective ad campaigns. Here, we highlight some of the important ways in which it remains essential in today’s advertising ecosystem.

“It is absolutely essential that ad-supported platforms have robust content moderation in place – both to vet the content of ads, and the context in which they appear,” says Alexandra Popken, VP of Trust & Safety at WebPurify.

“At WebPurify, we are trusted by many of the world’s largest platforms to ensure their ad ecosystems are legally compliant, promoting a safe and positive user experience, and protecting the reputations of the platform and their advertisers.”

**“It is absolutely essential that ad-supported platforms have robust content moderation in place – both to vet the content of ads, and the context in which they appear”**


**ALEXANDRA POPKEN,**  
VP OF TRUST & SAFETY,  
WEBPURIFY




# A brief history of brand safety

---

Brand safety is not a novel concept, but it is one that has gained greater relevance in the digital advertising space.





**B**rand safety can be traced back to print newspapers, where ads were placed according to article themes, and where editors tried to protect against ad misplacement. [Mistakes still occur](#) even today in print media, but the scale of digital advertising increases this risk significantly.

Contextual brand safety started to cause major issues in the digital advertising space circa 2015, explains AJ Brown, COO at the Brand Safety Institute and former Head of Brand Safety and Ad Quality at Twitter.

“The primary focus of brand safety in digital advertising was initially very narrow in scope: how do I as an advertiser ensure my ads are being viewed by humans and how do I prevent my ads from showing up near undesirable content online? We at Twitter began seeing people take screenshots of ads where they didn’t belong, which needless to say upset advertisers. Our peer platforms were running into the same issues, but we were all tackling this problem in silos because there weren’t widely adopted industry standards for this space at the outset. Both for the sake of preserving revenue and ensuring that our platforms were being monetized responsibly, we had to address this.

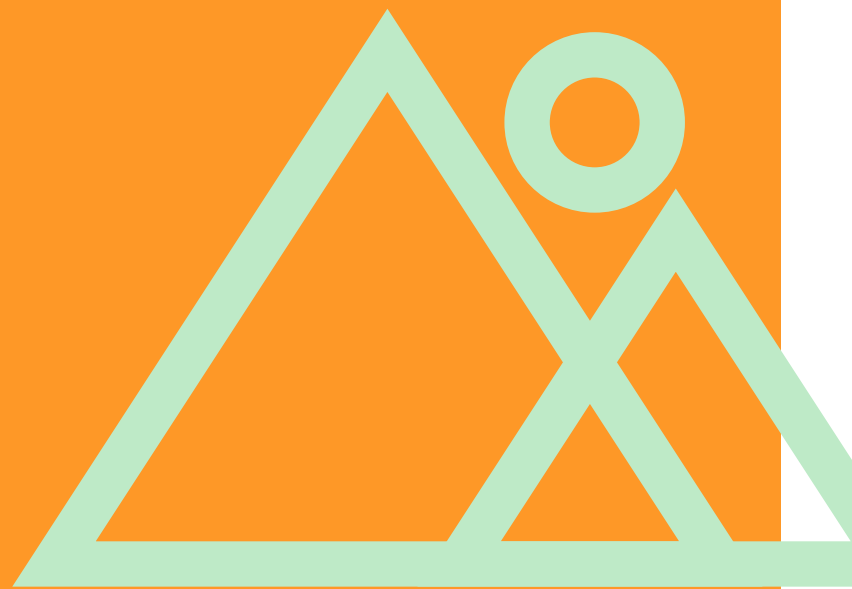
“When the Global Alliance for Responsible Media was founded in 2019, we started trying to coalesce around a set of industry standards for what was and wasn’t acceptable to monetize. Then, during the pandemic, and following the death of George Floyd and the Stop Hate for Profit protests in 2020, the umbrella of brand safety started to expand. The kinds of content permitted on platforms, even if that content wasn’t monetizable, began to factor into advertiser investment decisions. This meant that platforms’ content moderation policies and processes as a whole were now brand safety considerations.

“At that point, brand safety was no longer just a question of ‘Is my ad next to bad things or not?’ – it started to expand to other considerations of the types of companies that I as an advertiser want to do business with. What types of voices am I uplifting or inadvertently suppressing through my advertising strategy? How are my investments directly or indirectly impacting society for better or worse?


“Now, in particular, we’re seeing the focus of brand safety moving beyond just avoiding the negative and making sure ads aren’t next to things that are unsafe or unsuitable, and into how advertising can be a force for good. It’s about media responsibility.”

# The current advertising landscape

Before diving further into brand safety as part of a healthy advertising ecosystem, it's important to ensure that advertising on your platform is legally compliant and high quality.







Laws and regulations for advertising content vary across regions and are constantly evolving. Failure to comply with statutory advertising requirements can result in significant fines and put platforms in legal hot water. In the US, the Federal Trade Commission (FTC) is responsible for protecting the public from deceptive or unfair business practices. The agency also writes and enforces rules for advertising.

Sensitive, regulated products such as gambling, financial services and healthcare are covered by more stringent laws. For example, the advertising of pharmaceutical drugs and direct-to-consumer (DTC) medical devices is regulated at the federal level by multiple agencies, including the Food and Drug Administration (FDA) as well as the FTC. If a company doesn't have proof to support its advertising claims, the FTC can issue a [Notice of Penalty Offenses Concerning Substantiation of Product Claims](#). If the company then continues to engage in that practice, it can face civil penalties of up to \$50,120 per violation.

The global nature of business means that these kinds of country-specific laws and regulations for advertising content also need to be factored into advertising practices and platform content moderation policies. It is AJ Brown's personal view – and not the views of his current or former employers, he emphasizes – that differing laws by country, state, and locality make things 'harder' and 'more expensive' for platforms. But that's sometimes for the better and sometimes for the worse, he says.

"It costs time and money to stay constantly apprised of evolving regulations around the world, to adapt to them and to enforce them in the country or locality in which they're relevant. Oftentimes, the spirit of those laws is substantially similar or the same across jurisdictions, but the individual requirements around the implementation thereof are different enough that companies need to stand up separate dedicated teams and operations to comply with them.

“That’s not to say that country-specific laws aren’t important. Cultural nuances around the world should be accounted for in local laws. But especially when the jobs these laws are serving are similar to other laws around the world, or where their requirements meaningfully deviate from best practices elsewhere, those differences should be scrutinized.”

The more complex and diversified the legal landscape is around the world, the more it entrenches the largest, most established players that have the resources and teams to keep up with those laws, Brown adds. He also points to the politicization of certain areas, in particular around misinformation and freedom of speech, which can impact global content moderation and advertising strategies:

“[Laws regulating topics such as misinformation and freedom of speech] are now creating scenarios in which there are polar-opposite requirements in different parts of the world. It’s not just that you need to publish a different type of transparency report in different markets to show that you safeguard your users or their data effectively, it’s that you could get sued or completely blocked in a country for allowing a piece of content to exist on your service anywhere in the world. Elsewhere, you could also be sued or blocked for removing that same piece of content. Governments are trying to legislate not just for what occurs within their borders, but for what anyone in the world is allowed to say or access on a platform.

**“The more complex and diversified the legal landscape is around the world, the more it entrenches the largest, most established players that have the resources and teams to keep up with those laws”**

**AJ BROWN,**  
COO AT THE BRAND SAFETY  
INSTITUTE AND FORMER HEAD OF  
BRAND SAFETY AND AD QUALITY  
AT TWITTER



“So you’re now forcing platforms to make decisions not just about how to invest enough money to be able to keep up with the evolving legal landscape, but to choose which parts of the world to operate in.”

So, what are the implications for companies that fail to adequately moderate the advertising ecosystem on their platforms? There are legal ramifications for neglecting to take action on illegal ads, of course, but what about advertising experiences that are legal, but are still widely objectionable to users and advertisers?

## THE CURRENT ADVERTISING LANDSCAPE

An audience being turned off is the first indication that more needs to be done. If users come to view the ads on a platform as low quality and intrusive, then they may well go elsewhere – and be less likely to come back. Platforms may measure this by tracking user reports on ads, among other internal approaches, and there are also industry best practices for ad experiences.

“The [Coalition for Better Ads](#) has standards that stipulate ad formats across different internet services that are overly intrusive – like popup ads and auto-playing ads with sound,” Brown explains. “Speaking for myself, if an ad meaningfully interferes with my experience accessing important information, regardless of the content of the ad or who it’s coming from, I’m going to leave the site – [the publisher] has lost a customer.”

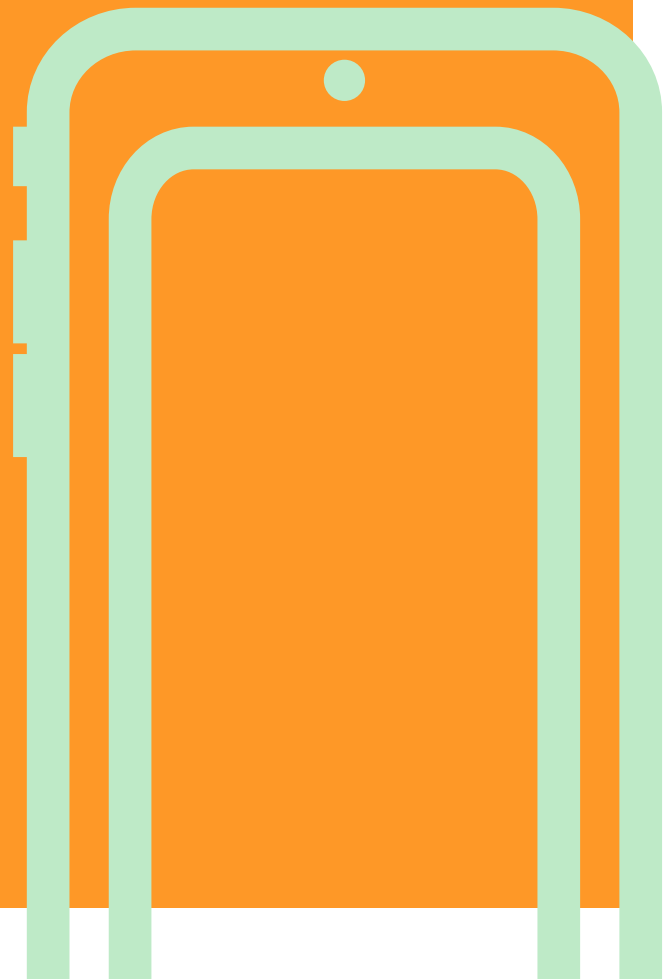
The health of a platform’s advertising ecosystem can also be damaged by low-quality ad content, even if presented in a non-intrusive manner. Accepting low-quality ads impacts your ability to attract large brands with deep pockets, Brown says: “Even if you have contextual brand safety nailed and you have fantastic consumer policies, if you’re taking money from anyone who’s willing to give it to you and putting low-quality advertisers into the same auctions as the big-name brands, those big brands may choose not to do business with you.

“[Big brands] aren’t going to love the fact that people might be scrolling and see one ad that’s well crafted and has a high-quality message, and then the next ad they see is a crypto scam. So, the types of advertising you do and don’t permit could also become a brand safety problem for those large advertisers. Moderating the content of ads and the conduct of advertisers is just as important as moderating the context in which they appear.”

# Brand safety in the digital age

---

We've covered the importance of content moderation in ensuring that ads are legal and high quality, but what about moderating the context in which ads are served? How do you prevent advertising messaging from being seen alongside inappropriate content that could damage brand reputations, and how do you ensure that ads and their surroundings contribute to a positive user experience?

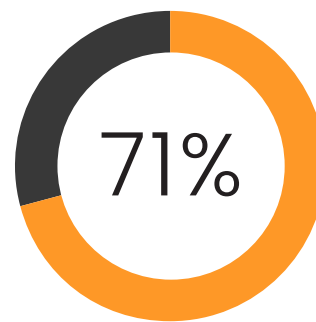


Users really do notice this kind of thing. [A survey of US consumers](#) highlighted that 71% of consumers would feel less favorable towards a brand that advertises near inappropriate content.

The challenge that platforms face in moderating content to align with brand values is that ‘inappropriate content’ can be hard to define. Sensitivities vary widely across both advertisers and users. While there are some obvious content categories where advertising support should be off the table, others require more nuanced tools.

“News is probably the most noteworthy example of this,” AJ Brown explains. “There are certain advertisers that don’t want to associate with news at all, and some that don’t want to associate with news on specific subjects. Then there are those that are very keen to invest in, support and uplift quality journalism in all its forms.

“But it is not incumbent upon an industry body or the platform to make a determination that all advertisers are either going to serve on all types of news, only some types of news, or no news.” Advertisers need the ability to make these kinds of brand suitability decisions for themselves.



of consumers would feel less favorable towards a brand that advertises near inappropriate content

Brand safety and suitability concerns extend beyond social media platforms and user-generated content. Programmatic advertising, which delivers ads at scale across thousands of websites, presents additional challenges for brands to track where their ads appear, and incidents of brands being exposed to risk have been reported in programmatic as well – such as [travel-related ads](#) appearing alongside news reports of a fatal air accident.

In 2019, the Global Alliance for Responsible Media (GARM) defined the [Brand Safety Floor + Suitability Framework](#), which provides both ‘a common understanding of where ads should not appear’ and a ‘common way of delineating different risk levels for sensitive content’.

The brand safety floor enables organizations to identify content that is not appropriate for advertising support through a framework of industry-consistent definitions, including but not limited to:

- The glamorization of illegal arms for the purpose of harm to others
- Pirating, Copyright infringement, & Counterfeiting
- Harassment or bullying of individuals and groups
- Explicit or gratuitous depiction of sexual acts, and/or display of genitals, real or animated

Platforms are expected to ‘leverage their community standards and monetization policies’ to ensure that the GARM brand safety



## Controlling brand safety & suitability: where the responsibility lies

“It is the responsibility of a platform to uphold brand safety standards, and that onus should not fall on your advertisers,” says Brown. “Advertisers should understand what your brand safety policy is, but it’s not their job to operationalize it. Providing platform-wide brand safety protections is your job as an ad-supported service, preventing all ads from appearing adjacent to content that should never be monetized.

“It’s also the platform’s responsibility to build brand suitability controls that allow advertisers to tailor how their ads appear, above and beyond that platform-wide brand safety ‘floor’. Lastly, platforms need to pursue independent validation of their efforts. This entails independent audit and accreditation of a platform’s brand

floor is upheld. While this dictates the type of content that platforms should not put any ads near, the brand suitability framework points to potentially sensitive content that is appropriate for advertising as long as it is supported by 'enhanced advertiser controls' that allow advertisers to opt-in or opt-out of being near that content.

The framework provides high-, medium- and low-risk brand suitability tiers for common content categories. Full nudity is considered high risk, for example, with dramatic depiction of sexual acts presented in the context of entertainment identified as medium risk, and educational treatment of sexual subjects seen as low risk. Platforms can leverage this framework to build brand-suitability tools that allow advertisers to tailor their experience beyond the brand safety floor, as opposed to platforms having to impose policies that are implemented for everyone, such as for news or educational content.

Whether brand-unsafe or brand-unsuitable, content needs to be competently categorized and remediated in order to effectively operationalize these industry frameworks. So how is content moderation leveraged to ensure a safe and suitable advertising platform?

safety practices, through accreditors such as the Media Ratings Council and the Trustworthy Accountability Group, as well as integrating with third-party measurement partners that allow advertisers to validate whether they were able to effectively execute their brand suitability preferences on your platform.

**“The onus is on the advertiser to understand the platform’s policies, what tools are available, and how to use them effectively”**

“A common issue that I would run into with clients involved alleged brand safety incidents that were actually brand suitability issues, where the advertiser wasn’t taking advantage of the tools that were being offered to them. This could involve showing up on a surface on the platform that they didn’t want to be on, like search results, or showing up adjacent to content that is brand safe, but that may have been unsuitable for their brand. Unfortunately, news often falls into this bucket. Platforms shouldn’t demonetize that sort of content categorically, but they should offer tools to avoid unsuitable content should an advertiser wish to do so. It’s the advertiser’s job to use those tools.”

# Building a brand safety practice

---

A logical starting point for a platform looking to maintain brand safety is human- and keyword-based enforcement, suggests Brown.





**“W**hen starting out, it’s often easiest to identify your highest-risk surfaces, and implement small-scale systems to reduce as much risk as possible, with the understanding that it won’t be perfect.” However as your practice scales, Brown points out, keywords are a blunt tool and human teams can be costly.

“[These approaches] don’t work for every situation, especially when you’re at a major social platform. When you need to decide which of the millions of user profiles or conversations on your platform should have ads inserted, it’s impossible to hire enough humans to handle that scale, and blocking one wrong keyword at that scale can be catastrophic. Machine Learning usually proves more effective in these kinds of environments.”

Despite its bluntness, uses still exist for keyword-based enforcement in a mature brand safety practice, Brown says. “If I never want ads to appear next to the word ‘Nazi’, for example, I can easily ensure this if I’m using a keyword list. In doing so, I risk catching quite a bit in that broad net I’m casting that isn’t the kind of content I’m trying to avoid, such as news, or people who are speaking out against that terminology. Keyword lists allow you to make binary decisions to exclude content from monetization at scale, and sometimes that’s exactly what you want, but they have ramifications and should be used sparingly.”

**“When you need to decide which of the millions of user profiles or conversations on your platform should have ads inserted, it’s impossible to hire enough humans to handle that scale, and blocking one wrong keyword at that scale can be catastrophic”**

**AJ BROWN,**  
COO AT THE BRAND SAFETY  
INSTITUTE AND FORMER HEAD OF  
BRAND SAFETY AND AD QUALITY  
AT TWITTER





When it comes to using Machine Learning to ensure brand safety, Brown warns that this isn’t as simple a task as it may seem. “I’d caution anyone who thinks that they can build a single all-encompassing brand safety model. Models can be very good with narrowly defined missions: Is this a cat? Does this image contain

nudity? Does this image contain violence or blood? But with brand safety, if we're going off of the GARM framework, there are at minimum 12 different kinds of unsafe content. You can't ask a single machine learning model to reliably tell you whether a given piece of content is either adult or misinformation or violent or profane or sensitive. Some of these areas are highly subjective, but even if they weren't, you're also asking the model to achieve a very broad mission and it's going to be very unlikely that it can do that reliably at scale. There are entire companies dedicated to this kind of work, and they leverage hundreds if not thousands of ML models to do that, and many of them still supplement those models with humans and keyword lists to reach optimal outcomes."

So if no approach is perfect, what does this mean for the future of human involvement? Brown believes there will always be a place for human moderation in brand safety and suitability operations. "For some of the most fundamental, worst-of-the-worst types of content on the internet, there are robust automated detection systems, such as PhotoDNA. By comparison, brand safety, and brand suitability even more so, are incredibly subjective. Subjectivity requires a greater degree of human involvement.

"We're dealing with human sentiment fundamentally, because developing a safety and suitability strategy for your brand is also the process of anthropomorphizing your



**"We're dealing with human sentiment fundamentally, because developing a safety and suitability strategy for your brand is also the process of anthropomorphizing your brand"**

**AJ BROWN,**  
COO AT THE BRAND SAFETY  
INSTITUTE AND FORMER HEAD OF  
BRAND SAFETY AND AD QUALITY  
AT TWITTER



brand. You're answering questions about what your brand is and isn't comfortable with, as if it's a person with its own unique point of view and values. And that's really hard to develop clear and consistently enforceable criteria for, especially when you add in things like cultural and linguistic nuance. You'll always need a degree of human involvement when answering these very human questions."

# Understanding and combating ad fraud

---

Fraud is another longstanding and growing risk in the digital ad ecosystem.



**D**igital advertising fraud is an attempt to deceive marketers, ad platforms and users for financial gain. With click fraud, a prevalent form of ad fraud, fraudsters game the system in a variety of ways in order to generate fake clicks and impressions, often using bots to mimic human behavior. Invalid, or nonhuman, ad traffic is one of the longest-standing brand safety concerns in digital advertising, arguably predating contextual brand safety concerns in this space.

Fraudsters have become more adept over time, with malicious activity techniques growing ever more sophisticated to inflate impressions and deceitfully drive ad revenue. These can include:



### **Ad stacking**

multiple fake ads hidden behind a single visible ad

### **Pixel stuffing**

ads placed inside a 1x1 pixel iframe that is imperceptible to users

### **Ad injection**

unauthorized ads inserted into web pages

### **Domain spoofing**

a fake website or email domain that impersonates a known organization or person, or simply seems authentic

### **Cookie stuffing**

multiple affiliate cookies placed on a user's device without their consent

### **"Made for Advertising" (MFA) websites**

also known as Made for Arbitrage sites, MFA sites are often low-quality, templated web pages with abnormally high ad loads, ads that auto-refresh to inflate impression counts, and other design elements that encourage accidental ad clicks, resulting in higher bounce rates and lower time on page

It is challenging to mitigate the problem of ad fraud. Display, programmatic, video, in-app and social media advertising are all susceptible, and the techniques used are highly developed and multifaceted. Getting on the front foot is vital though: according to [a report by Juniper Research](#), \$170 billion per year of ad spend is expected to be lost to ad fraud by 2028.

If you're on top of your data, then you may be able to drill down to find the signs of ad fraud in a marketing campaign, such as:

### **1. A spike in traffic with no legitimate reason**

An unusual and sudden increase in clicks might be a sign of malicious bot activity.

### **2. High Click-Through Rate (CTR) but low conversions**

A surge in CTR that doesn't deliver a corresponding increase in conversions or engagement can indicate suspicious behavior.

### **3. Unfamiliar sources of clicks**

An abnormally high volume of traffic from locations that are not included in your target audience, or multiple clicks from the same IP address should raise concerns.

[A September 2023 survey](#) from Integral Ad Science found that **60% of US digital media professionals agree that social media is most vulnerable to ad fraud**. In addition to the malicious tactics described above, Jason Adatao, former Manager of Global Ad Fraud and Risk Operations at Twitter, categorizes other common types of ad fraud into these buckets: content risk, security risk, and payment risk.

Content risk might include ad copy that's promoting a scam; for example, a get-rich-quick scheme. Security risk, on the other hand, includes behaviors like account takeovers where the perpetrator might hack into a high-profile user's account to publish unauthorized posts on their behalf, thus reaching a wider audience. Finally, payment risk is where a fraudster pays for ads using a stolen credit card or intentionally charges back. "In some cases, the three blend together, which is why it's a good idea to make sure that your enforcement isn't siloed on that front," explains Adatao. Moderation work is often focused on content (and context in the case of brand safety), but evaluating behavior is equally important when combatting fraudsters and other bad actors.

**\$170 billion per year of ad spend is expected to be lost to ad fraud by 2028**



'Cloaking' is another pernicious and lingering issue in the ad space, and for content moderation in general, explains Adauto. "For example, you might engage with an online ad from your home in the US that looks like a normal e-commerce website. But if you have a content moderator based in another country, they might see something completely different given the IP variance. Through this practice, bad actors may find success in circumventing detection by deceiving users into engaging with malicious websites.

"So, as a content moderator, if I'm reviewing this specific piece of content, it largely depends on my IP address location, and in some cases device ID. This is something that often flies under the radar, and is a really hard challenge to solve."

So, how do you identify risky behavior in the first place? AI, human moderation, or a hybrid solution can be leveraged to detect and prevent the problem. While AI is at the forefront of tech in general right now, says Adauto, a certain level of manual review is always going to be required.

"A hybrid approach ensures you're treating your users with care and that your policies are fairly applied in practice," he says. "With user-generated content, you can easily build machine learning models on the backend that are intended to scan for certain pieces of content, but if they're not being fine-tuned day-to-day, they aren't going to catch

**"With user-generated content, you can easily build machine learning models on the backend that are intended to scan for certain pieces of content, but if they're not being fine-tuned day-to-day, they aren't going to catch everything"**

**JASON ADAUTO,**  
FORMER MANAGER OF  
GLOBAL AD FRAUD AND RISK  
OPERATIONS AT TWITTER



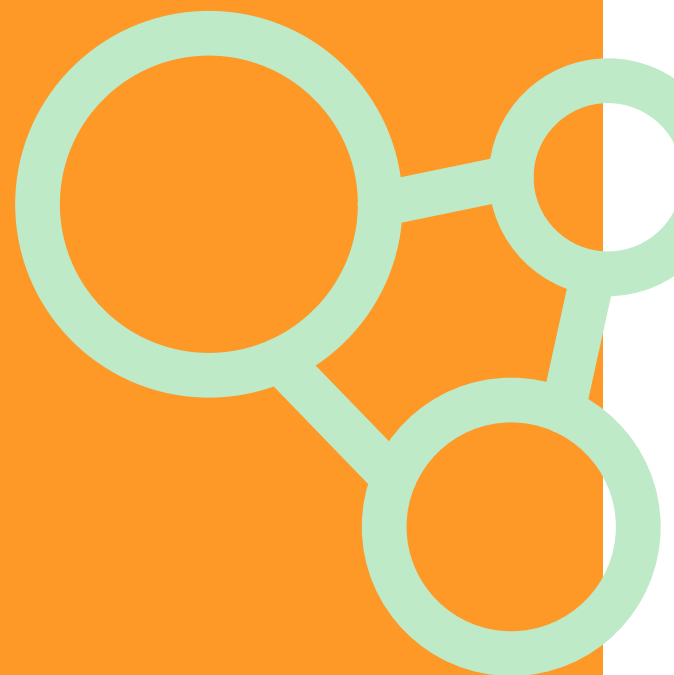

everything. That's where a level of manual review is also required."


Equipping moderators with the skills they need to carry out risk reviews is certainly possible, he adds: "You need to make sure your house is in order. For example, creating an internal knowledge base with strong guidance and metrics helps empower your front-line teams to make decisions in line with your platform policies. There are a suite of account heuristics that can supplement the team and allow them to evaluate content and behavior holistically."

# Collaborating with sales teams and advertisers

---

Content moderation in the ad space should not be a siloed process, as the relationship between sales teams, advertisers and moderation teams is a symbiotic one.





**S**ales teams build and maintain relationships with advertisers; they can explain the content moderation policies and are able to react quickly when clients raise concerns. Advertisers want to be confident their brands are not going to be seen in ‘unsafe’ spaces while maximizing ROI, and need the content moderation and sales teams to ensure this happens.

Despite the fact that content moderation often involves removing or demonetizing content, the value of content moderation in an advertising-based business model is generally accepted. As Dr Yi Liu from the University of Wisconsin–Madison, co-author of the 2021 paper, *Implications of Revenue Models and Technology for Content Moderation Strategies*, [told us](#), platforms that operate under an advertising revenue model do have some motivation to moderate content.

“Their aim is to strike a balance between those who staunchly advocate for freedom of speech and those who are sensitive to extreme content. The majority of users occupy the middle ground, where explicit violent imagery or hate speech is generally unwelcome. Employing content moderation tools to prune such posts may result in the loss of some users, but it enhances the platform’s overall appeal to a broader audience, including advertisers.”

Balancing compliance with positive advertiser experiences is key to sustainable relationships and revenue. Failure to moderate effectively

can leave brands exposed to reputational risk, thus jeopardizing revenue. Overzealous moderation of certain content, however, can leave the platform exposed to accusations of bias or censorship, and risk regulatory scrutiny.

Fail to get the balance right, and you risk an exodus of advertisers – and ultimately users. As Leonardo Madio and Martin Quinn report in their [‘Content moderation and advertising in social media platforms’](#) working paper:

“Concerns escalated in 2022 when Elon Musk, taking over Twitter, relaxed the platform’s content moderation policies. The world’s biggest media buyer, GroupM, classified Twitter as a ‘high-risk platform’ for brands and many luxury brands (e.g., Balenciaga) either paused their ad purchases or quit the platform.”

“At the end of the day, platforms get to decide where their red lines are regarding user safety, brand safety, and ad quality,” says Brown. “Regulation notwithstanding, platforms determine what speech and behavior is acceptable on their services, and what gets to be monetized. Those decisions aren’t going to please everyone. Some decisions may cost them users, others may cost them advertisers. Content moderation in all its forms is a balancing act, and it’s up to platforms to determine the role that content moderation will play in cultivating an online environment that is both appealing to people and viable as a business.”