Content Moderation in Dating:

Best Practices for Platforms and Users

WebPurify

Table of Contents

Getting Started

WebPurify is a leading content moderation provider and, now in our seventeenth year of business, an early pioneer in the space. Our work spans numerous geographies and industries, including online dating.

Dating platforms present a variety of challenges where user-generated content (UGC) is concerned. Being that the dating industry is a longstanding part of our customer base, WebPurify has developed and refined a dependable approach to the common pitfalls its many players can encounter. Said approach leverages both artificial intelligence (AI) and human moderation solutions and—as always—prioritizes user safety, first and foremost.

The purpose of this guide is to share these best practices and recommendations. Whether you're an engineer or product manager, launching a new app, or simply adding a feature, we hope you'll find it's a go-to resource that helps you protect both your users and your business.

01 What is Content Moderation and How Does it Figure into Online Dating?

WebPurify

Content moderation is the review of any UGC uploaded to websites or mobile apps against predefined criteria using AI models, human teams, or a combination of the two. Aside from screening for illegal material and activities (sale of narcotics, for instance), the criteria typically consists of prohibiting not-safe-for-work (NSFW) subjects such as nudity and hate speech, along with rules unique to a particular company's products or audience.

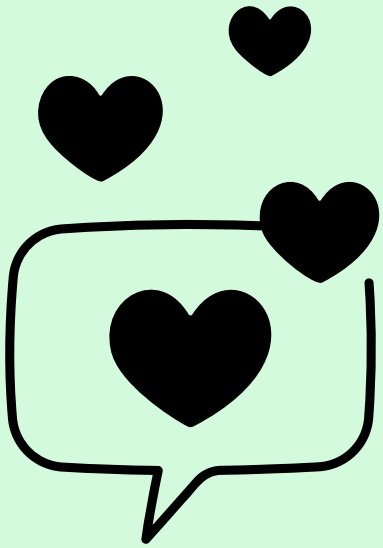
Only recently did moderation graduate from a "nice to have" to something acknowledged as integral to brands' success and customers' safety. Previously, moderation was frequently outsourced to a generalist company or delegated to a company's internal team to complete in spare moments in between doing their "main job." For a time, this can suffice, but a brand looking to properly scale eventually either retains a specialist moderation vendor or builds in house.

Content moderation is especially germane to online dating for two specific reasons:

01

All dating apps are about sourcing and building a romantic connection.

This is impossible without communication through text—and often, voice, photos, and video. Some of these conversations are likely to turn playful, risqué, or outright sexual; user retention and success depend on it.

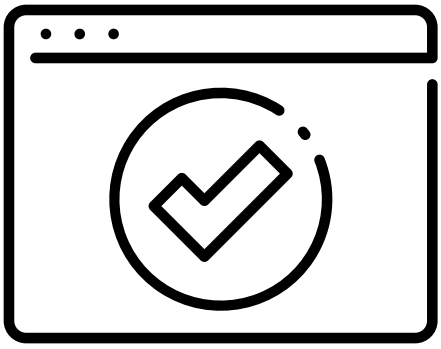


The trick is to identify context, intent, and timing. Simply blocking all nudity or sexual language, as is done in most industries, doesn't apply here. At the same time descriptions or images of sexual violence, or nudity on user profile pages, are never okay. Good content moderation accurately draws and enforces a distinction, allowing users to express themselves as long as they don't cross certain lines.

02

Most online communities have an age requirement of around 18+.

The fallout isn't terrible if this rule is flouted in a community focused on, for example, sports or peer-to-peer e-commerce. But in dating, where the goal is to connect with some level of intimacy, failing to detect underage accounts breaks laws, neglects to protect children, and can be exploited by predators. The stakes are high and accurate moderation can make or break a business's moral reputation and bottom line.



02 What's the Fuss?

WebPurify

"Moderation matters" is a catchphrase we use at WebPurify. It's a blend of inside joke, battle cry, and hashtag. It also happens to be true.

Good content moderation of UGC builds trust within your user base, increases customer retention, boosts app store ratings, and—in an era of screenshots and retweets—vastly reduces the likelihood of a public relations meltdown. It also preempts duplicate or fake accounts and safeguards your users' experience.

After all, while every customer appreciates the availability of a "report content" button to draw admin attention to questionable material, they'd prefer not to see that content at all. Automated, built-in moderation that checks every upload avoids this issue by catching most of the "bad stuff" at the outset.

Moderation also functions as a product feedback loop as much as a UX safeguard. For instance, a new product feature may be a locus for more explicit or spammy photos and need tweaking. Likewise, a pattern of messages or URL posts can help you rapidly identify troublemakers or fake accounts and take action.

In addition, industry data shows a direct relationship between the quality of a company's content moderation and its advertiser revenue. It's hardly a revelation: Big brands want to know their ads won't appear alongside something graphic or hateful. In contrast, a reputation for questionable content sends ad dollars running.

03 How Does Moderation Work?

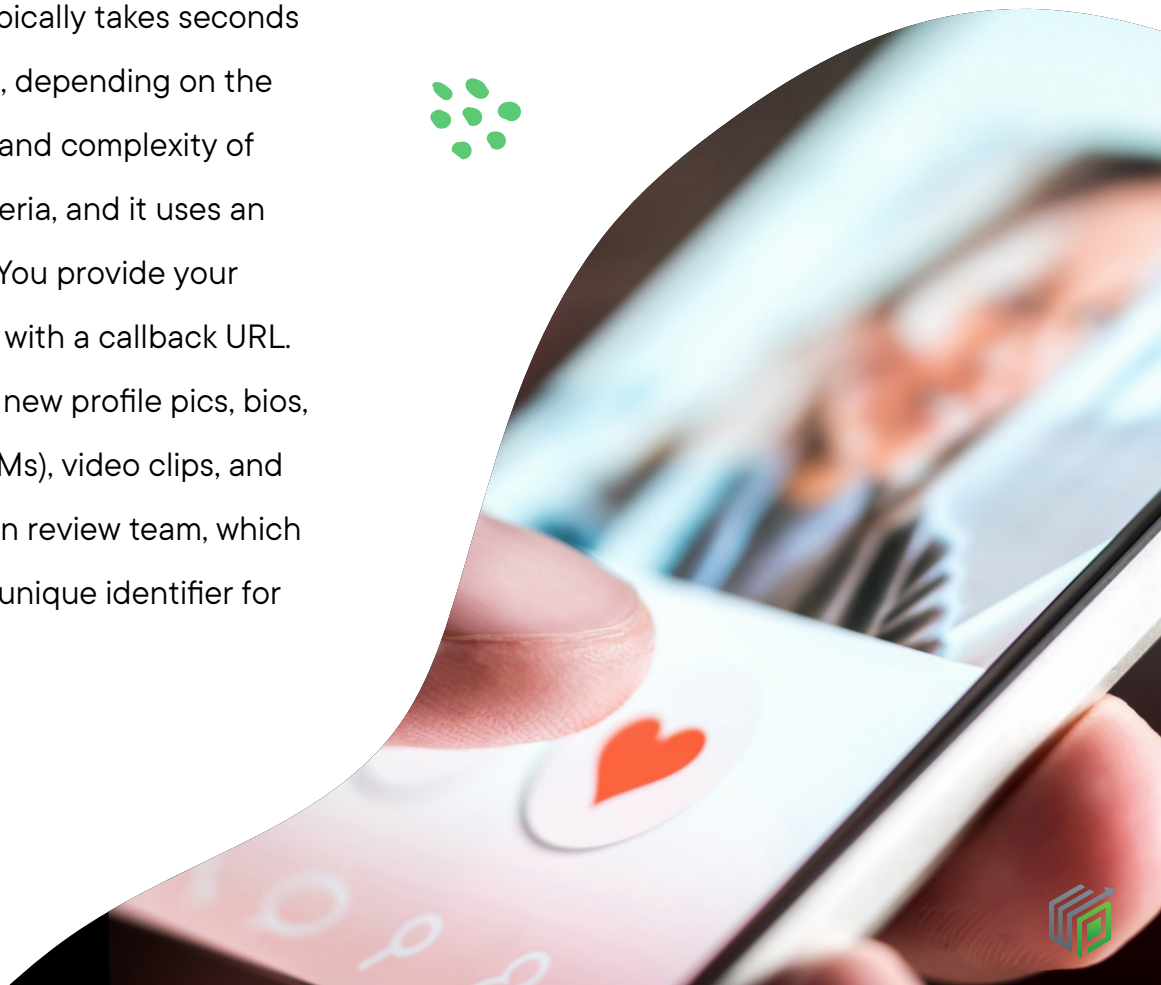
WebPurify

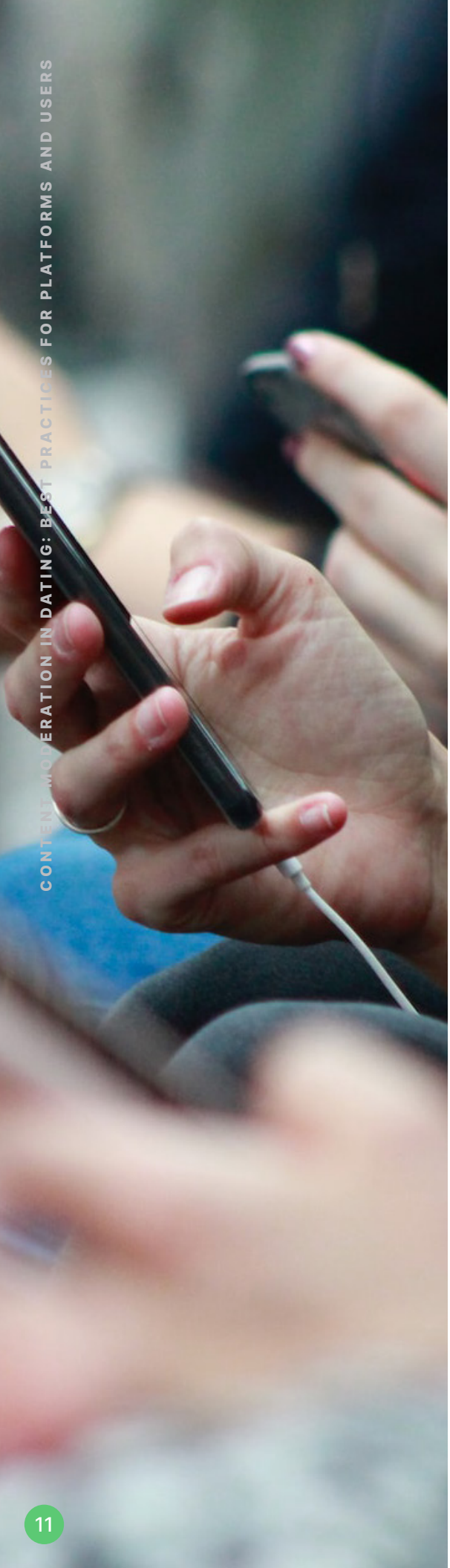
Most moderation solutions use the principles of application programming interfaces (APIs). In lay terms, an API allows two software components to speak with one another. For example, when a user uploads a new profile picture to a dating app, a typical AI moderation setup sends that image data "request" to the moderator's server to start review. The server then returns a result indicating whether the content is appropriate. From beginning to end, this turnaround takes less than one second.

Human review is also fast, but much slower than AI. It typically takes seconds or minutes per item, depending on the size of the content and complexity of the moderation criteria, and it uses an asynchronous API. You provide your moderation vendor with a callback URL. Requests to review new profile pics, bios, direct messages (DMs), video clips, and more reach a human review team, which instantly delivers a unique identifier for each one.

When the human team completes their review, they ping the callback URL with the moderation results, and you receive these plus the corresponding ID, so your system can tie each result to a specific piece of content.

The API setup isn't always used for human review, though. For instance, an API isn't necessary if your company decides to recruit, train, and build your own in-house team. That's also true if your moderation vendor's human team reviews your content using your own in-house, moderation tool.





Chances are, your first priority is creating a great dating app, not designing a system for moderators to access. That's why we recommend you look for moderation vendors that provide their own content review tools, like WebPurify. While requiring an API, these systems are designed to increase moderation speed and accuracy while reducing personnel fatigue. For instance, one of our review tools automatically deconstructs videos into a mosaic of still images so moderators can quickly scan a clip that would otherwise take 30 seconds to watch. We also built image magnification tools for ease of drilling down on detailed content, and a proprietary QC tool that allows managers to dip into content queues and spot-check work.

AI models are trained on huge datasets, refined over time, and scan content for anything that looks like the "bad" data they've ingested in the past, such as nudity, guns, gore, etc. A respectable UGC moderation solution allows you to customize both the categories of illicit content you want checked and the probability thresholds. These are cut-off points at which content is rejected or accepted. For example: "If the likelihood of nudity is equal to or greater than 70 percent, reject. If under 70 percent, accept." Customizing text moderation most often involves building lists of words and phrases that are blocked or allowed in a model's logic.

Human moderation teams also train on datasets, from buckets of NSFW content to custom libraries that feature specific, unique, client-defined violations of criteria. The best teams use a proprietary tool in content review. It is increasingly common for employees to collaborate with AI, which often skims content first, quickly denies egregious uploads, and escalates borderline content to a human for closer inspection. This adds a layer of certainty to your moderation process while helping to improve the AI by allowing moderators to call misses or false positives to the quality assurance team's attention.

04 Timing Matters

WebPurify

Ideally, you'd consider content moderation during the planning phase of building your product—for two reasons.

01

IF YOU'RE BUILDING IN HOUSE

You can integrate your moderation and platform features so that content review is natural and unobtrusive. You also save time by avoiding having to retroactively add moderation functionality into your solution after launch.

02

IF YOU'RE SHOPPING FOR AN EXTERNAL MODERATION PROVIDER

Starting the conversation early lets you identify planned features that might create a challenge for your vendor. You can then allow sufficient time to manage expectations or make adjustments. For example, content criteria your team sees as straightforward may be more nuanced in your vendor's experience—or you might envision faster turnaround for new profile photos than is practical or within budget.

Planning ahead is ideal, yet most brands, even the giants, still play catch-up with UGC safety. Many WebPurify clients engage us long after their platform is live, and their user base has grown beyond the capabilities of their initial homegrown moderation setup. Your team may be in the same boat, and that's okay. New types of content (dating in the Metaverse, anyone?), new hot-button issues, and news stories that lend a new, negative meaning to certain cultural references, plus changing laws, affect every industry, including online dating. Keeping up can feel like a scramble, and it's well worth engaging a third-party expert, even simply as a consultant, to advise your in-house team.

Anticipate Industry Pain Points

WebPurify consults and moderates for dating platforms that collectively serve more than 65 million active users. Over the years, patterns of specific priorities and risks have emerged. We recommend that you always consider the following:

What will we allow on public user profiles?

Public profiles should reflect the strictest version of your moderation criteria (and strictness depends on your type of dating community and user base). While allowing self-expression, be mindful that public-facing profiles are the first things new users see. The tone they set can dictate whether people abandon or adopt your platform. Individuals viewing potential matches usually haven't yet "opted in" to your content by connecting with someone specific, so they shouldn't see or read anything you wouldn't allow in your platform's greater community.

How rigorously, and in what manner, will we vet new accounts?

In WebPurify's experience, online dating services need to vet new accounts in four ways:

01 Ensure a person's face is clearly in frame for the main profile photo.

02 Ensure the person is 18 years old or above.

03 Screen for duplicate accounts.

04 Confirm the person creating the account is the person in the photo.

AI models trained on human faces, age, and gender can easily address the first and, generally, second criteria. If your company is especially stringent, you might require an ID upload, which AI can also review. Limiting accounts to one per email address somewhat controls for duplicates, but this is the weakest of safeguards, and users usually clear this low hurdle by creating multiple personal emails with different providers.

Enforcing the fourth criterion is perhaps the most rigorous way to check new accounts. Asking a user to submit two or three photos at once, usually recreating two or three simple gestures—like a thumbs-up, a salute, and a peace sign—remains the most reliable approach, and confirms that the individual behind the screen is likely who they say they are. At present, AI is not sophisticated enough to vet this type of submission, however, so you'll need to rely on human review.

How rigorously will we vet account updates after initial setup?

User profiles are not static. People change, their dating strategy or relationship status changes, and their bios follow suit. Almost everyone who uses a dating app has, at some point, sought feedback from a friend, and then rewritten headlines or reshuffled photos.

Moderating every change and edit is ideal, but reputation scoring your users' accounts can help economize where both time and budget is concerned. AI at volume adds up, so many companies moderate a profile's first ten or so changes. If no violations are found, they spot-check that user thereafter. In other words, some amount of trust has been established, requiring less oversight. All the same, you need to find the right balance between efficiency and risk tolerance.

Will we permit unsolicited photos or videos in private messaging?

Like spot-checking profile edits, moderating direct messages between users is a calculated risk. Many moderation vendors, WebPurify included, do not store the content they moderate, so text strings are processed without being downloaded and images or videos are reviewed where the content is hosted. This preserves privacy and largely anonymizes data, even when reviewing DMs.

Some platforms do not review DMs with AI at all, though, unless a user flags a message. This doubles down on respect for private communications, but passes the responsibility for initial content review to users.

Another approach is to allow users to toggle "accept images or video" on or off in the user interface. This requires one user to grant permission to another, and vice versa, before they receive anything beyond text communications. It affords your customers a degree of consent in their interactions and is becoming increasingly popular.

How are we embedding features that encourage safe courtship?

WebPurify strongly recommends weaving best practices for safe dating into your company's terms of service or any tutorials you serve to new users during sign-up. Highlighting practical, common-sense habits for socializing online communicates that user well-being is a priority. Given that the Apple App and Google Play stores are crowded with dating options, thoughtful sign-up experiences that blend best practices with indications you're truly looking out for your users can help edge out your competitors.

You can also give a moment's pause by integrating a warning message into your moderation AI when someone sends an image or videos involving nudity by DM. This is very similar to the "double check" confirmation pop-up you see in bank transactions. It might read: "We think this content might contain nudity. Once sent, it cannot be canceled. Please be sure you trust this individual. Do you want to send?"

On the heavily sexualized internet, with the sending of nudes so normalized, this messaging serves as a circuit breaker. It interrupts engagement at some level, but it protects younger users who likely haven't considered the long-term ramifications of private photos being made public.

What safeguards do we have in place for CSAM?

Child sexual abuse material (CSAM) is unacceptable and illegal in any context. Training AI on illegal content is rightly difficult, so we recommend using ready-made models for detecting and reporting CSAM along with a human moderation team. Presently, there are a number of CSAM detection tools on the market.

AI models that check for both nudity and human faces that appear underage in the same video frame or image can be a good start, although they are less accurate than CSAM-specific detection. Escalating any suspected offenses to humans for moderation is absolutely required.

And because CSAM is particularly serious, WebPurify recommends you vet moderation vendors by inquiring about their use of tools like PhotoDNA or other safety multipliers.

What safeguards do we use to detect and flag scams and criminal behavior?

Good scammers, experienced lawbreakers, and terrorists use dating apps to hide nefarious activity. A community of singles looking to meet one another is a natural place to blend in or to take advantage of folks who've let their guard down.

For example, people who know one another might create profiles and connect/match to discuss incriminating subjects through DMs. Or a bad actor might use code words in a bio to advertise illegal items or services for sale. Other scams target the unsuspecting through romance fraud or, even worse, luring or extorting vulnerable people into sex work using promises, flattery, and lines of questioning designed to gain personal information or solicit money.

Pushing back against such malicious activity requires robust AI to flag problematic phrases and words. Yet AI models cannot stay current with every new term, so it's important to have block and allow list functionality, which facilitates quick manual updates. Since images and video can also contain text (including personally identifiable information and QR codes), the ability to detect this (namely, OCR) should be built into whatever AI you select.

It's worth remembering that human moderation isn't always reactive. Bad behavior requires an opposite party (a "mark") and is often stamped out using techniques similar to secret shopping. Several dating apps rely on human teams to observe and report inappropriate behavior, or even to entertain it without veering into entrapment, so that offending accounts can be banned.

How can users challenge a moderation decision in the case of false positives?

Sometimes AI triggers a false positive. One of your human moderators might miss a cultural or current event context that renders something that is actually innocuous seemingly offensive. Or one user might falsely flag another's post as a form of trolling, revenge, or practical joke. Even for platforms with clear-cut terms of service, some room should be allowed for learning curves. What seems to be a clumsy advance to one person may be reported as lewd by another. A new user interpreting the overall "energy" of an app may be off the mark and cause unintended trouble.

Permitting offending users to challenge moderation decisions—whether made by AI, humans, or members of the community—opens the door to redress, helps improve your moderation models and training, and offers a chance to educate unsuspecting offenders about your platform. Customers who feel enfranchised and safe, from one another and from arbitrary content decisions, are always more loyal users.

How quickly is content reviewed? What does a user experience during that time?

This question is of paramount importance in online dating. Whether signing up for the first time or editing their profile (for the eleventh time), users expect instant gratification. As the product owner, though, you need to have oversight in place before uploaded UGC becomes viewable by everyone.

When you consider your budget, audience, and app structure, you might find the best way to proceed is decided for you.

Suppose your app algorithmically or randomly matches users at a relatively low daily volume. In that case, users will probably be open to slower moderation turnaround time—and a new photo can immediately go live and stay unchecked for several minutes without issue. However, if your product works rapid-fire, with lots of swiping or even live video chats, you'll need AI to check in real-time that everyone's behaving.

If you prefer to have human moderation of all new content but a quick turnaround time for each upload, we suggest you code accordingly. Design your app so that only the user who makes the update sees their change instantly, and their profile refreshes for the wider community after it's been checked by your human team. It's a simple, effective sleight of hand. It's also required for most teams since turning around all profile updates (text, images, etc.) with humans only takes at least a few minutes per upload—or requires a cost-prohibitive moderator headcount.

Do we communicate our content policy in our terms of service and when an account is deactivated, or content is deleted?

Consistent, transparent communication is crucial. Your terms of service and privacy policy must be readily accessible and understandable to the average reader. If legalese is necessary, it should follow clearly worded policies. This protects your team by making reasons for disciplinary action clear, objective, and defensible, no matter how much someone might cry foul. Although it requires added work, we recommend including your terms of service in any new-user tutorial, along with your "dating safely" online tips.

Our customers often ask what they should tell a user, if anything, when they upload something inappropriate that is rejected. There's no right answer. Moderation tools, WebPurify included, return information about a string of text or an image, etc. What you do with that information depends on the rules you set and your product's code.

Common approaches include:

- Letting text go through, but with a symbol "censoring" the bad words (ex: a ### sign)
- Triggering an in-app message stating that the last upload violated your terms of service and linking to those terms
- If the offense was egregious, saying nothing in the moment, suspending the account, and then sending an email from your support team explaining why the account is suspended

Are our criteria realistic for AI? Do we need human moderation? When or for what?

The term “realistic criteria for AI” is subjective. In general, though, the more nuanced and complex criteria are, the less likely AI will be an accurate enforcer.

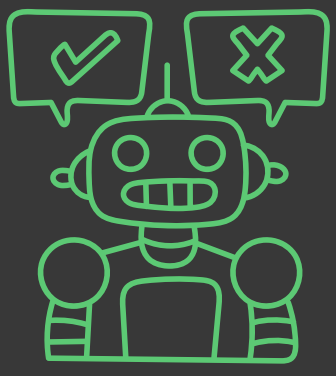
WebPurify has found that you can almost always use AI for text and to check any images or video for basic NSFW content for which AI models are long established: nudity, weapons, alcohol, drugs, overlaid text (watermarks), etc.

You need humans, alone or alongside AI, to review anything:

- Artistic: drawings, distorted photos, 3D graphics etc.
- Sarcastic or contextual: irresponsible drinking, fighting, racist tropes, undertones, etc.
- Impending: AI models can’t understand when an image that isn’t offensive shows a situation that’s about to be offensive—a revving car heading into a crowd of people, or a person winding up a fist to punch someone in the back of the head, etc.

Criteria that are specific to your brand, but not innately difficult for AI to learn, such as “no mentions of your competitor’s products and no display of their logos,” don’t necessarily require humans. Whether you use humans or train an AI model in these cases typically depends on how quickly you need moderation in place and your budget. A hybrid approach remains the gold standard since you can quickly introduce new criteria to your human team while, over many weeks, train a new AI model to eventually take over, if desired.

05 Choosing Your Approach: AI

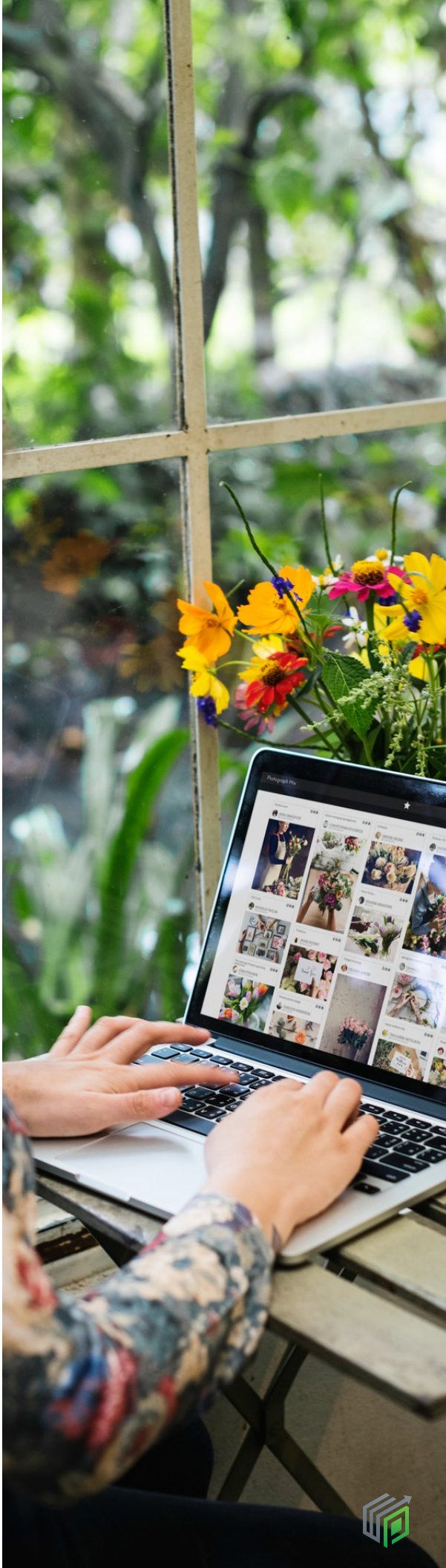


WebPurify

Almost all moderation solutions make use of AI. Many more than most people realize also rely on human teams. Some companies use one or the other, but a growing number find a hybrid approach most effective and economical. The best approach for you depends on the type of content users are allowed to share, necessary speed of review, total volume of content (including any backlog), and your criteria.

Simple moderation of images, focused on basic NSFW parameters, can usually be left to AI. This is also true of text, especially if you're concerned with basic bad or hateful words. Additionally, the only timely way to handle a large backlog of content is to leverage AI.

There are many good options for well-trained AI models that detect guns, nudity, hate symbols, alcohol, drugs, middle fingers, and the like. A good provider will allow you to pick and choose the categories you check. So, if your community of singles is geared toward casual encounters, you might not screen for partial nudity (swimsuits, underwear, etc.). If you're helping to match people in recovery or practicing a sober lifestyle, you might reject any photos containing alcohol, pills, or drug paraphernalia.



If your volume of new UGC is large, AI is essential to keeping costs reasonable and moderation scalable. WebPurify's experience is that human moderators cannot outpace AI or come close to its processing rate, even if they are only enforcing the most basic of criteria, reviewing ten images at a time and watching videos at four times the speed, without audio. It's practical, even essential, to rely on AI for a first pass, then escalate anything it flagged as dubious to humans thereafter. Even then, the math can be dizzying when you're reviewing thousands of hours of video per day, so if you're retaining an external moderation provider, ask about an AI discount in exchange for a commitment to a monthly minimum volume. A reputable vendor will be open to this.

06 Choosing Your Approach: Humans

WebPurify

On the flip side, if content involves nuance, context, art, current events, or behavior in virtual reality, it needs human review. The same goes for audio files or audio in video since AI audio moderation cannot understand tonality and, in many cases, transcribes audio and then feeds that transcription through AI text review. This extra step increases overall turnaround time while adding another way in which things can get lost in translation … quite literally.

Product teams understandably resist hearing that their moderation needs make human oversight, with its longer frames and bigger costs, unavoidable but it's important to be clear-eyed about AI's limitations. AI can be used as a first pass in these situations, allowing content to be displayed immediately, lightening a human team's load, and reducing overall costs.

Content moderator is now a bona fide career path in the larger world of trust and safety, and a difficult but essential job. Whether hiring and training a team in-house or outsourcing everything, your company is obligated to either provide a healthy work environment or be certain the vendor you're working with is doing the same. Moderation teams with ready access to mental health resources, mandated rest periods, advanced training for difficult subject matter, and rewards for top performance have been shown to review content more accurately and have high employee retention rates. (See WebPurify's guide, Mental Wellness of Content Moderators.)

Be sure to ask external moderation providers whether staff is shared across clients or if a dedicated team is assigned to your projects only, so they get to know your app, community, and content criteria in detail. Don't settle when it comes to team selection. Try to avoid companies that outsource or crowdsource. Finding economical solutions overseas is fine and even prudent, but you want to avoid vendors that cobble together remote workers here and there. Stick to players that use a full-time staff, all under one roof. Ask about their security and QC measures. Who's reviewing others' work, and how often? Is there a policy that guards against moderators snapping photos of sensitive info on their work screens using their personal phones? (Skim our Guide to Selecting a UGC Moderation Partner for more.)

Special Situations to Watch

Curious about common online dating scenarios that give automated moderation trouble? Check out these examples. If these situations apply to your app, plan for some human review:

Virtual reality and Metaverse features

The Metaverse is a recent frontier, and it adds aptitude for intimacy. However, 3D objects are difficult for AI models to train on, review, and flag, and virtual reality adds a behavioral dimension only humans can oversee. AI doesn't understand inappropriate pantomime, gestures, or invasions of personal space. This is why WebPurify is investing in dedicated human VR moderation teams.

Drawing and captioning

Many apps allow users to modify images with doodle tools or to caption photos and videos. An uploaded image, even one previously reviewed by AI and found safe, must be re-reviewed if it's been edited on your app. The edit creates a "new" image that could have something inappropriate added to it. True, AI optical character recognition (OCR), coupled with a profanity filter, catches bad words. But a harmless word on top of a harmless image, together, can create a harmful meaning or double entendre. Many clients escalate images or videos with any words added, so human teams can ensure nothing is suddenly mean-spirited (or worse) in the new context.

In addition, any art, from stick figures to paintings, usually needs human help. Most AI models are trained on images, and art is often far from photo-real, yet it can still be upsetting or hateful.

Livestreams

As a rule, we recommend having AI check livestreams first to ensure scalability. It doesn't take long for concurrent live video to overwhelm a human team. Also, be sure there's an option to skip video frames. Spot-checking every fifth or tenth frame is a great way to reduce costs without much compromise on safety. Barring clear hits on highly sensitive things like CSAM, AI should escalate potential offenders to humans for a closer look. Did a single frame trigger a false positive, or did the AI catch a stream of bad behavior? Looping in a human to watch more before pulling the plug may be the difference between warning an account and closing it.

Your platform should also track and alert a human moderator to any livestream seeing unprecedented views or a sudden, sharp rise in viewership. These indicate that a stream has just taken an interesting turn. Many times, this is not a bad thing. Sometimes, though, it signals something macabre, illegal, or explicit is being shown. Thus, the need for a closer look.

Subtle or custom concerns

If there isn't an AI model for it, you'll need to train one and either use humans to review temporarily or indefinitely in the meantime. Common custom criteria often include images that foretell something inappropriate happening. For example, images that foreshadow sex or violence without showing events in frame. Consider this upsetting image: A profile photo of someone harshly bringing down a brick over a dog's head without contact. This will—unfortunately—fool almost any AI.

You may prefer that users upload no logos in any content. Many AI models are trained to detect major company logos, such as those in the Fortune 500, but it's impossible to detect them all. Humans are required.

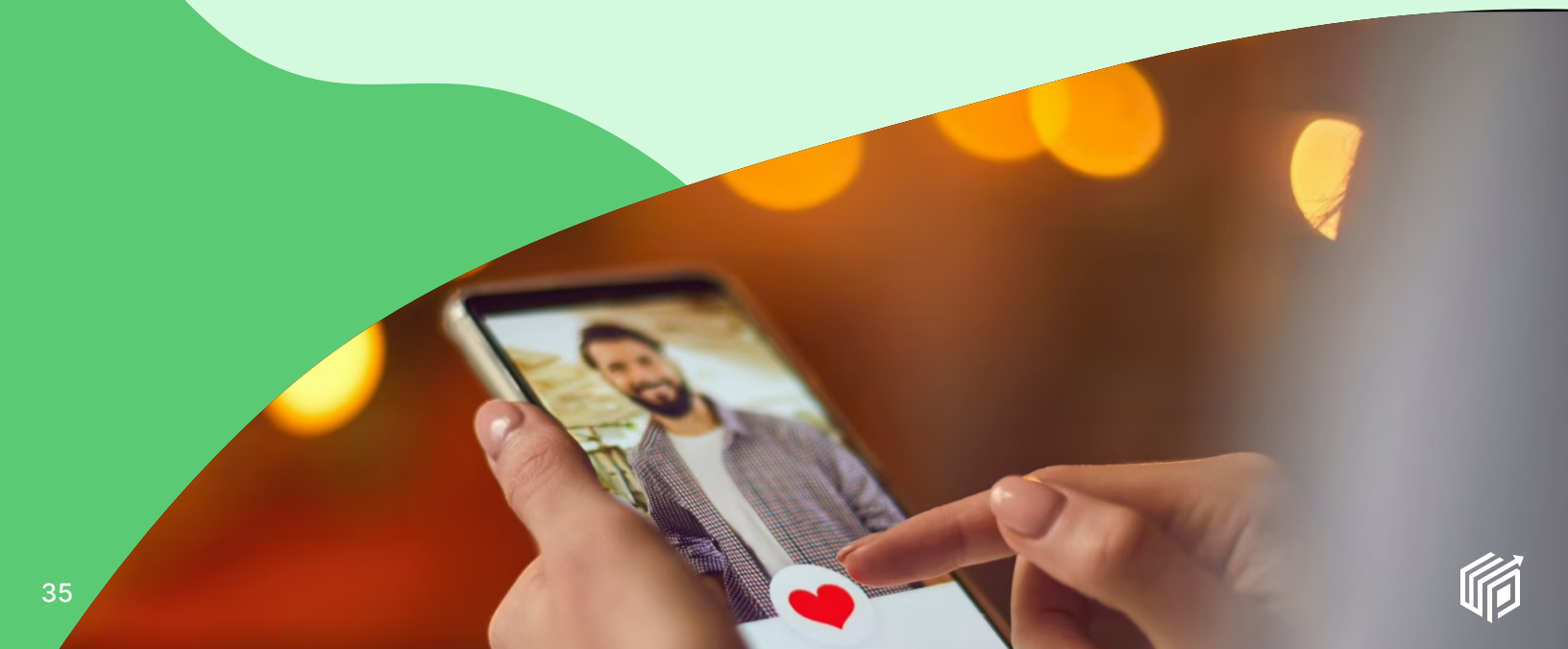
A final example is imagery that implies unwanted attitudes, such as racial stereotypes. AI doesn't detect ethnic caricatures. They are too numerous, often hand drawn, and highly contextualized.

If budget is going to force your hand, this is where we advise compromise. Sometimes this means paring down criteria and relying on a human team to review content reported by the community. AI's limitations are technical, while human limits are time and budget.

Wrapping Up

It's difficult to ascribe to the internet any one main objective or effect on society, but the idea of connectivity is at the top of the list. A subset of this connectivity has always been romantic and, in the last 15 or so years, has evolved into an extensive ecosystem of mobile apps and websites that help people of all stripes meet, flirt, commune, and forge companionship. Sometimes it's for the short term, but increasingly what starts with online banter and emojis gives way to lifelong relationships.

Online dating isn't just here to stay: At present, it's the most common way new couples meet in many countries across the globe. As the number of people turning to dating apps continues to climb, it is incumbent on everyone who builds these products to safeguard their users, both because it's the right thing to do and for the sake of your company's reputation and success. In this spirit, we hope this guide is helpful and something your team can reference from time to time. The landscape of trust and safety and the world of online dating are ever-evolving. This primer is a very solid start to addressing any content challenge that comes your way.



Since 2006, we've been providing elegant, robust solutions that protect kids, online communities, and our clients' brands. Every day, we moderate 3.5 million text submissions, half a million images, and thousands of videos.

webpurify.com

@webpurify

877-751-4046

sales@webpurify.com