SafetyKit

Executive Brief · 2026

AI-Generated Media, Products, and Abuse

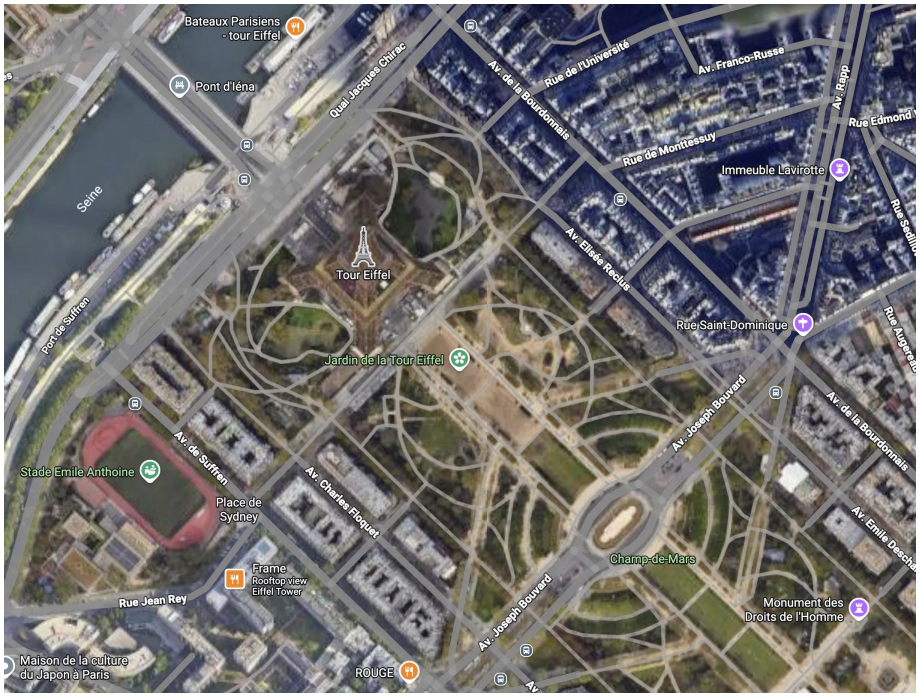
Core quality, fraud, and safety implications of AI generated content — and strategies to mitigate risks today and in the long term.

Seeing double?

Notice anything wrong with this photo of Paris?

Okay, you got me, AI. What about this one, though?

Maybe, though you could spend some time testing whether different street corners are plausible.



Google Street View: Street layout and tower orientation reveal spatial inconsistencies

The SafetyKit model was about nine minutes and fifty-nine seconds ahead of us, but it had an unfair advantage: analyzing the textures and pixel-to-pixel constitution of the image. AI image generation models necessarily leave artifacts and patterns at the pixel level, undetectable for humans but obvious for the SafetyKit model. In tricky cases, our model does check the Google Maps API and compare the street view, but that's rarely needed — for now.

Detecting AI media is a cat and mouse game. In the long run, the detector will lose when the media generation becomes identical to reality.

Marketplaces and **Payment Platforms** face new risks and operational challenges from AI: we outline SafetyKit's approach to the most salient ones here and how SafetyKit can protect your platform from fraud and low-quality content.

Table of Contents

Detection Strategy

While AI content remains detectable, SafetyKit will continue evolving its approach. We mix pixel-by-pixel analysis with offsite investigations, plus a rapid feedback loop with risk & policy partners at global marketplaces and payment processors.

1. A mixture of detection models

How are AI media generation models trained to be realistic? They make an image, then another "critic" model tries to guess whether the image is real or fake. A training round is completed when the critic hits 50% or lower accuracy: guessing. In other words, media generation models are *optimized* to avoid detection.

By using a mixture of models, you can stay ahead of the curve. Whenever a new generation model is released, SafetyKit runs comprehensive evals to ensure its models remain 99% accurate, then rolls out rapid updates.

2. Media extraction

Sometimes, an image is legitimate, but it features an AI generated product for sale (for example, taking a real picture of a "hand-painted" art print).

SafetyKit first extracts media in images, then scans each piece for AI-generated content.

3. Reality Checks

For platforms advertising physical spaces (e.g. AirBnB) or real-world products (e.g. Air Maxes on eBay), we search for the real thing online and run a side by side comparison, similar to the Eiffel tower example above.

This is more expensive than a basic detection check, and ideal for high-risk fraud and high-touch quality workflows.

Real printed photo of an AI-generated image

4. Robust evals and rapid user feedback

SafetyKit monitors product reviews and comments for core marketplace and UGC partners to ensure no fraudulent AI content slips through the cracks on detection. These reviews — plus close partnerships with policy experts and partner QA teams — provide immediate signal when detection is underperforming.

5. A fraud model

> **We will always be able to detect AI fraud**, no matter how good the generation models become.

When the image itself is indistinguishable from reality, its intent will still be possible to unmask. SafetyKit's AI detection model ingests signals from comments, reviews, transactions, metadata, and change history to unearth risk.

Original Image

Updated Image

We watched a seller update it several times, but the updates were highly unusual.

Typically, when a seller alters product images, they change layout, color, or orientation. Here, the underlying shapes of the image changed, indicating

AI image doctoring. Even if our pixel-by-pixel detection model misses this image on its own, we can catch the fraudulent product.

Marketplaces: Core Challenges in 2026

AI media blurs the line between fraud and quality. Platforms must determine:

- Is this product legitimate and high quality?
- Which internal team's responsibility is it to decide?

Is this an AI mockup for a legitimate product or a sign of fraud?
Early in 2025, many platforms took the approach that, if every image in your product listing is AI-generated, your product probably isn't real — or won't look anything like you promised.

> **This rule doesn't apply anymore**, as AI has become an incredible tool for mocking up new products, reducing the need for professional photography and staging.

A good bandaid fix for that rule is to allow custom products and services more leeway in AI use, since the actual product may not yet be ready to be photographed. A long term fix requires holistic review, where a model like SafetyKit's takes in 100s of signals from media, text, reviews, seller history, metadata, and offsite investigations to separate mockup from fraud.

Long term, this question of product legitimacy raises the importance of:

1. **High-quality, authentic reviews** — where we recommend platforms have zero tolerance for AI-generated media or text. Reviews prove that products are great.
2. **Great KYC and user monitoring**: as authentic, high-rated accounts become more valuable, preventing them from being hacked or flipped becomes even more critical.
3. **Granular actions for suspicious accounts**: how many platforms can say that their human moderators have a button they can click which

messages a seller, asking them to add a non-AI image to their listing, then automatically checks in a week later? AI media enforcement is a product and tooling problem as much as a quality problem.

Is AI a quality problem, fraud problem, or content policy problem?
There are three teams that could own risks relating to AI media. For any platforms interested in allowing *some* AI media, we recommend a four-step waterfall.

You'll run four checks: AI detection, quality, fraud, and content policy.

We almost *always* recommend starting with AI detection. This content often requires new workflows and case management approaches, so filtering down via AI detection limits the scope of the problem.

In some cases, a platform will only want to review certain products for AI-generated media, e.g. specific categories, products over a price point, or products from new merchants — treat the waterfall as starting *after* that filter step.

1. **AI Detection**
   Filter AI-generated content

2. **Quality**
   Do we want this on our marketplace?

3. **Fraud**
   Is the product real? Is the merchant legitimate?

4. **Content** Policy
   Is this allowed on our platform?

Order these steps based on pass rate and cost. If your content policy review is 10x cheaper than quality and catches more violations, you may want to put it ahead of quality. If fraud is a long investigation with a low pass rate, you may want to put it last.

SafetyKit integrates via API at every step of your waterfall, providing human-level accuracy at scale, and human case management tooling for the hardest cases.

Payments: Emerging Fraud Risks, Powered by AI Media

Today's leading financial technology platforms have cut the friction out of payments — reducing onboarding times from days to minutes. Fast onboarding increases fraud exposure, but platforms have stayed ahead with machine learning algorithms. AI media poses a new challenge on two fronts:

- **Circumventing onboarding**: Creating fake documents, face verification, and websites has never been easier. Complex fraud operations abound.
- **Increasing access to illegal products**: AI makes building illegal businesses faster and cheaper. Today, this manifests primarily with deepfakes and malware services, but could expand to more industries as AI media improves.

Complex Fraud Operations

Platforms are exposed to sophisticated fraud at a never-before-seen scale. People always faked receipts, but AI media generation enables faking entire companies, leaving a paper trail the size of an Amazon warehouse.

Want to fake real-time face verification? Hold up a tablet with an AI video or, better yet, run a real-time face swap.

Coordinating safety measures at the model provider level is near impossible, with dozens of high-quality providers and open-source models.

AI detection works here, but only with media extraction, because often the whole document is not AI generated. This can mean that detection is prohibitively expensive.

Solution: Index heavily on user feedback to interactions with customers. As a payment platform, look offsite (with SafetyKit) for reviews and adverse media. If there is no user feedback, look for web presence and be suspicious of the absence of web presence. If this is an in-person business, analyze their transaction patterns for traditional fraud signals, but expect that in the future, it may be quite cheap for you to "audit" an in person business anywhere in the world with drones.

AI Deepfake Platforms

Unlike fraud platforms, deepfake platforms are easy to detect. They're selling a clear-cut, malicious service and advertising it to find buyers.

These platforms often violate the TAKE IT DOWN ACT legislation in the US and UK, as deepfakes constitute non-consensual intimate imagery.

To avoid detection, these sites often practice transaction laundering: registering with a payment processor under another name and funnelling payments through their "clean" entity. Access to legitimate, well-known processors and cards is critical for deepfake platforms, as their users are often wary of using crypto or other suspicious checkout options.

> **Finding and removing these platforms requires flipping the script on detection.** Instead of searching for violators in your own data, SafetyKit runs millions of internet, dark web, and telegram searches across the globe to find malicious websites, run test transactions, and identify the upstream "clean" site.

SafetyKit's risk platform detects and mitigates risks upstream and down-

stream of AI media. We action millions of fraud, transaction laundering, and high-risk cases for the world's largest payment platforms. Book a demo to learn more.

Recommended Policy Lines: Preserving quality, mitigating risk

These are battle-tested strategies across quality, risk, fraud, and trust & safety.

Quality SOP — For Marketplaces and Creator Platforms

1. Allow AI use, but require disclosure

There are two possible disclosure integrations:

- **Seller input**: require the seller to add the disclosure and assign a strike if they do not
- **Product integration**: Add a disclosure based on AI detection result

For a platform that puts high trust in sellers (e.g. rental platforms, crowd-funding), leveraging seller input should be the highest priority. A direct product integration works best when AI use has become a rampant, urgent issue, or when rolling out seller disclosure would require a prohibitively large backlog review and comms effort.

2. Require "proof of work" for listings over $100

Sellers must upload a video demonstrating either their manufacturing/creation process or a functional use of the product. Strictly prohibit AI use in these videos with a high-severity penalty.

3. Prevent new and inactive sellers from "starting" with majority AI-generated media

Block these products without assigning a strike, ideally as a step in the product creation flow. Encourage users to add "proof of work" or non-AI images to their product. We define new sellers as users with zero sales and inactive sellers as ones with no sales in the past three months.

The measure against inactive sellers is designed to prevent flipped and hacked accounts from selling low quality, AI-generated products.

4. Block AI use in comments and reviews

Reviews are an incredible counterbalance to AI content, but the second they stop being trustworthy, platform quality degrades.

5. Escalate AI-generated products priced over $500 to fraud review

Beyond the risk of a low-quality product, AI-generated products could signal a myriad of other concerns: a $500 AI generated painting, for example, could be a sign of money laundering.

6. Block high-risk services

Block services with explicit mention or innuendo of:

- Deepfake/face swap content
- AI generated illegal sexual content (e.g. non-consent, bestiality)
- AI generated minor content with sexually suggestive language

Where possible, employ red-teaming services like SafetyKit's to check if the product/service for sale will generate illegal/underage content.

Modifying Existing Minor Safety & Adult Content Policies for AI use

1. Treat realistic AI content the same as photographic content

AI-generated content can still be considered CSAM and must follow the same rules. Treat illustrative content the same way you would human-made illustration, sculpture, and other craft media.

2. Draw the line between illustrated and photographic content with a custom Golden Set

The line between what is photographic and illustrative blurs quickly and varies heavily from platform to platform. We recommend putting together a set of 20-50 real product listings and aligning internally on what should be treated as photographic (and therefore actioned).

This alignment process will help define core rules for enforcement. There's

Monet's painting (left) shows delicate brush-strokes applied with rhythm and direction. The AI-generated image (right) uses vague, uniform strokes layered over flat color blocks.

no easy definition here, but a Golden Set of expert-labeled examples is an ideal starting point.

Fraud SOP — For Marketplaces and Creator Platforms

Investigations often start here for big-ticket items generated with AI. Keep in mind that AI use is a risk factor, not a risk on its own, so this may be a chance to trigger your standard merchant review workflow rather than a fully be-spoke one.

1. **Analyze seller history for rating, refunds, and past AI use**

Immediately flag sellers with past suspicious activity. Good standing does not exempt a seller from further review because their account could have been flipped or stolen. If you have robust access logs, checking them here is invaluable.

2. **Check for known bad actor patterns**

Look for patterns common to coordinated fraud operations.

3. **Analyze product complexity and "proof of work"**

Are the new products more sophisticated than previous ones? Is there proof in the product media or description that the seller can deliver?

4. Review online presence and adverse media

Identify the legal entity behind the seller and run web searches to see if they have presence on other marketplaces or creator platforms — or bad public reviews. SafetyKit performs this step automatically.

Fraud SOP for Merchant Onboarding/Monitoring

1. Block any AI use for critical documentation

There is no legitimate reason for company documents, receipts, or tax forms to be AI generated. Extracting objects from inside images is critical here, as bad actors can generate media, then print it out and take a real picture of it to evade detection.

2. When initial AI detection did not catch AI, but the merchant's authenticity is suspect, run a full AI usage investigation

This is where the SafetyKit model searches the location on Google Maps, runs reverse image searches, and orchestrates a far-reaching online investigation.

3. Analyze change history

A merchant's original documents may have been legitimate, but their account could be taken over or abused.

Does the lighting, background, font, and signature in new documents differ from the originals? Do parts of the new document appear identical to old documents in a way that suggests doctoring rather than a fresh photograph?

Summary

AI-generated media is not going away. The tools will get better, the content will become indistinguishable from reality, and the fraud operations will grow more sophisticated.

But the platforms that treat this as a layered problem, not a single detection challenge, will come out ahead. The most resilient strategies combine multiple detection models, robust user feedback loops, holistic fraud signals, and granular enforcement tooling.

Three principles should guide your approach:

1. Detection is necessary but not sufficient.
Pixel-level analysis works today. It will not work forever. Build systems that incorporate seller history, transaction patterns, reviews, and offsite signals so you are not dependent on any single layer.

2. Policy clarity prevents internal confusion.
Decide now whether AI media is a quality problem, fraud problem, or content policy problem for your platform. In most cases, it is all three, and you need clear ownership and escalation paths for each.

3. Invest in tooling, not just detection.
The platforms that win will be the ones where moderators can take precise, automated actions: request additional verification, flag for fraud review, or message sellers directly. Enforcement is a product problem as much as a policy problem.

SafetyKit partners with the world's largest marketplaces and payment platforms to detect AI-generated content, investigate merchant fraud, and enforce nuanced policies at scale. Our platform handles detection, case management, and automated enforcement in a single API.

Get in touch at safetykit.com to learn how we can help protect your platform.