# Fraudulent Advertising and Account Integrity: Expert Insights on Best Practice

Integrity Institute Report

July 2026

**Authors**: Spencer Gurley, Jeff Allen

**Experts**: Sofia Bonilla, Ariel Colon, Alexis Hochleutner, Henriette Cramer, Leah Ferentinos, Rob Leathern, Tara Sakhuja, Nick Shen

# About the Integrity Institute

The Integrity Institute is a non-profit membership-powered think tank that advances the theory and practice of protecting the social internet, made up of over 600 integrity professionals. With decades of combined experience mitigating harms to people and communities across more than 80 online platforms, we bring seasoned, practitioner-driven knowledge to those theorizing, building, and governing online spaces – helping them put integrity front and center.

Specifically, we:

- Build and empower a community of integrity professionals in the tech sector, giving them the tools and research they need to make online platforms safer and healthier for people and societies.
- Advise online platforms, policymakers, civil society, and academics to put integrity at the heart of company governance, compliance, and tech regulation.
- Educate the public about what an integrity-first future looks like for the online information ecosystem.

# Acknowledgements

This report was authored by **Spencer Gurley** and **Jeff Allen**, with input from members of the Integrity Institute community. It does not reflect the views of all Integrity Institute members. The authors are extremely grateful for the contributions from the experts listed below. Their inclusion here does not indicate their endorsement or support of all or any of the recommendations or statements in this report.

**Sofia Bonilla, Ariel Colon, Alexis Hochleutner, Henriette Cramer, Leah Ferentinos, Rob Leathern, Tara Sakhuja, Nick Shen**

# Ofcom Preface

Fraudulent advertising is a form of online harm that has affected millions of people in the UK.[1] It can undermine trust in digital services, cause direct financial and emotional harm to users, and distort the operation of online markets. As the independent regulator for communications in the UK, Ofcom is responsible for regulating compliance by service providers with the fraudulent advertising duties imposed by sections 38 and 39 of the Online Safety Act 2023 ('the Act').

The Act sets out the requirement for in-scope services to, among other things, take proportionate steps to prevent individuals encountering paid-for fraudulent advertising. While the Act sets out the duties that services must adhere to, it does not prescribe exactly how platforms should meet those obligations. Developing an evidence-based understanding of what effective mitigation looks like in practice is therefore a critical part of Ofcom's implementation of the Act, including the Fraudulent Advertising Code of Practice which Ofcom must prepare and issue in accordance with section 41(4) of the Act.

This report forms part of that evidence base. It was commissioned by Ofcom to explore how fraudulent advertising operates within modern online advertising systems, and to gather expert insight into the practices platforms use to detect, mitigate, and prevent fraud. The work was undertaken independently by the Integrity Institute, drawing on interviews with experienced Trust and Safety practitioners who have worked across a range of large online services. These perspectives provide useful insight into how advertising systems function in practice, and where current approaches may succeed or fall short. These insights have informed the consultation on the draft Code of Practice which has been published alongside this report.

**A central theme of this report is the complexity of the online advertising ecosystem**. Advertising ecosystems are often highly technical, involve multiple actors, and operate across organisational and national boundaries. Fraud can arise at many stages of the advertising lifecycle – from account creation and verification, through ad design and targeting, to ad delivery and off-platform user experiences. This complexity can make it difficult for platforms and other actors to monitor all online advertising, enforce standards consistently, and respond quickly when harm occurs. It also means that regulators need to remain abreast of the fast-moving nature of this space.

The report identifies a number of further insights, which are particularly valuable in the context of Ofcom's work. These include the following::

- **Interviewees described fraudulent advertising as dynamic and adversarial**, with fraudsters continually adapting their tactics. For example, fraudsters starting to use generative AI tools that are hard to detect but easy to use. This suggests that effective responses require

[1] *Ofcom and YouGov, 2026. Online paid-for advertisements research.*

monitoring the evolving picture of harm and using a combination of approaches that can be updated over time, rather than relying on static or single-point solutions.

- Interviewees also highlighted how **mitigation is further complicated by the structural differences between advertising models as well as variation in platform size and resources**. Platforms with internal advertising systems typically have greater access to data and more direct control over policy, moderation and enforcement. Larger platforms may also have dedicated teams to review ad impressions. By contrast, smaller platforms and those with external ad networks often have more limited access to data and fewer direct levers of control, usually depending on third-party vendors and intermediaries.

- Despite these differences, **interviewees identified a range of approaches that can be used across platforms to mitigate risks**. These include the use of multiple detection signals, such as advertiser behaviour, account characteristics, payment information and content indicators, alongside combinations of automated detection, human review, and user reporting. Interviewees also noted synergies with broader content moderation systems, where existing tools and processes for tackling harmful user-generated content can support the identification and enforcement of policy-violating advertisements. Interviewees noted that platforms can be lenient when deciding how to take action against users (e.g. strikes) and suggested that stronger penalties may be needed.

- Transparency was identified as another important, but challenging, area. Tools such as ad libraries can support researchers and civil society in identifying emerging fraud trends. However, **limitations in access to representative datasets and the lack of clear, standardised metrics remain barriers to developing mitigations**. Addressing these limitations could improve platform accountability.

- Finally, **while account integrity can be undermined by threats such as impersonation and account compromise, practical solutions already exist**. For example, account verification checks and multi-factor authentication can dramatically reduce the risk of unauthorised access to accounts, although it may introduce some friction for legitimate users.

**The perspectives set out in this report are solely those of the interviewees and authors, and are not endorsed by Ofcom, nor do they necessarily reflect the views of Ofcom. Interviewees' suggestions of what should be improved should not be considered a reflection of any policy position that Ofcom may adopt as part of our role as the online safety regulator.**

# Integrity Institute Foreword

The advertising systems used today by platforms are incredibly complex, can involve multiple companies, and present numerous opportunities for malicious actors to use the advertising system for fraudulent and other harmful purposes. Bad actors can use the modern advertising system to defraud consumers, other businesses, advertisers, and the platforms themselves. Based on interviews with Integrity Institute members who are Trust and Safety professionals experienced in working with online advertising systems and fraud, this report presents their perspectives on how to mitigate the risks of fraudulent advertising. Through this work, the Integrity Institute feels there is a range of implications for how to tackle fraudulent online advertising.

Mitigation of the problems that fraudulent advertising can cause people and companies begins with the platforms recognizing that it is in the best long-term interest of their business to prevent as much fraud on their platforms as possible. Interviewees noted that internally, company executives have highlighted that incidents of fraud on the platform, targeting users or businesses, hurt user and advertiser trust in the platform, and that a reduction in trust leads to a tangible reduction in engagement with the platform.

When platforms fail to properly combat fraud, there are examples of companies finding themselves in a situation where an extremely high fraction of their revenue[2] or ad spend[3] comes from fraudulent activity and creates a crisis for the company. Platforms, therefore, need to understand and be able to articulate exactly how they view fraudulent ads interacting with the business model.

Fraudulent ads can create tension between the short-term interests and long-term interests of the platform, so there is a genuine risk that platforms may choose to prioritize revenue growth in the short term by allowing additional fraudulent advertisements. Platform companies should have processes in place to prevent that and should be able to articulate how they manage that tension responsibly.

Once platforms have a strong incentive to reduce fraudulent ads, effectively combating fraud involves a coordinated, aligned effort that combines comprehensive and mature detection, moderation, platform design, policy, and governance. These efforts require expertise across many different roles, including engineering, analytics, operations, policy, threat intelligence, and leadership. Detection and measurement are core processes that enable platform employees to monitor trends and develop novel strategies for minimizing fraud.

Friction can be a key lever for platforms to use in their fight against fraud. Repeat offenders will

---

[2] Reuters, 2026. Meta Projected 10% of its 2024 Revenue Comes from Fraudulent Advertising

[3] Kumar, Saurabh, 2022. Uncovering Digital Ad Fraud: Lessons from Uber's $100 Million Ineffective Rider Ad Spend

always represent a significant portion of fraudulent activity on the platform. Introducing friction - such as identity checks or verification steps - directly disrupts these actors' abilities to operate at scale. The benchmark for effective fraud prevention is not perfection, but deterrence - platforms need to have enough protection to make committing fraud "not worth the effort" of the bad actors. Bad actors are constantly testing various platforms to see which ones offer the best opportunity for their tactics, so platforms are essentially competing against each other to frustrate bad actors.

However, increased friction in the process of running advertisements can also represent an opportunity cost for companies. Requiring advertisers to verify themselves to the platform, to use difficult-to-create external assets in the advertising process, enable multi-factor authentication (MFA) methods, and go through an account check process, are effective strategies to combat fraud and frustrate fraudster operations, but will cost the platform in the short term as opposed to allowing unrestricted advertising. Our findings from these series of interviews suggest that platforms could be doing more in this space, and there is an opportunity for regulators to create basic standards that keep users safe.

Regulation will not be without complications. Requiring greater public transparency from platforms risks creating some negative incentives and a desire to "move the goalpost" within the platforms. It is important that regulation includes comprehensive transparency requirements that make essential and key datasets publicly available, whilst also enabling regulators, auditors, and public interest researchers to ensure that platforms are fully honest in their reports.

Finally, there is a significant difference in the capabilities to combat fraud between platforms using internal ad systems and platforms using external ad systems. Platforms that use external ad systems are able to utilize only a small subset of the data and signals that are available across the full ad system. They are often reliant on, and at the mercy of, other companies and the data that they are willing to make available to publishers and platforms. There is an opportunity here for industry standards around transparency in the supply chain, similar to "buyers.json" type efforts, to improve, which will, in turn, empower publishers and platforms to set higher standards in the ads they display on their websites.

The problems surrounding fraudulent advertising, account integrity, and verification touch on all aspects of running a platform. These problems test the efficacy and trustworthiness of the design, operation, and governance of platforms. Ultimately, combatting fraud is in the long-term interest of individual platforms as well as the industry as a whole. There are many opportunities for industry-wide practices to be strengthened through regulation in a way that allows end-users and legitimate advertisers to build trust and safely engage with online platforms.

# Table of Contents


**1. Introduction** ..... **8**

1.1. Background and Context ..... 8

1.2. Methodology ..... 9

**2. How Does the Online Advertising Ecosystem Work?** ..... **11**

2.1. Types of Online Advertisements ..... 11

2.2. How Are Ads Shown to Users? ..... 15

2.3. External (Open Display) Ad Systems ..... 17

2.4. Internal (Walled Garden) Ad Systems ..... 18

**3. Platform Detection and Mitigation Strategies** ..... **21**

3.1. Platform Incentives to Combat Fraud ..... 21

3.2. How Do Platforms Detect Fraud? ..... 22

3.3. The Detection Process ..... 24

3.4. Threat Intelligence ..... 25

**4. Best Practice in Detection and Mitigation** ..... **26**

4.1. Best Practices in Detection ..... 26

4.2. Best Practices in Mitigation ..... 28

4.3. Best Practices in Technical Systems for Detection and Mitigation ..... 29

**5. Best Practices in Platform Design, Policy, and Governance** ..... **32**

5.1. Design ..... 32

5.2. Policy ..... 32

5.3. Governance ..... 33

**6. Example Case Studies** ..... **35**

**7. What Are the Key Challenges for Regulation and Platform Mitigation?** ..... **37**

7.1. Key Challenges for Regulation ..... 37

7.2. Key Challenges for Platform Mitigation of Fraudulent Advertising ..... 39

7.3. Costs ..... 41

**8. Policy Issue Deep Dive: Account Integrity** ..... **42**

8.1. Context: Who Buys Ads? ..... 42

8.2. Account Integrity Threats ..... 43

8.3. Signals of Inauthentic Accounts ..... 46
8.4. What Actions Can a Platform Take? ..... 48
8.5. Other Actions for Account Integrity ..... 50
8.6. What Detection Methods are Most Effective? ..... 50
8.7. What Should Proactive Detection Look Like? ..... 52
8.8. What Forms of Multifactor Authentication (MFA) are Most Effective? ..... 55
8.9. Metrics and Tracking Impact ..... 56
8.10. External and Internal Systems ..... 57
8.11. Cost Implications ..... 59
**9. Policy Issue Deep Dive: 'Account Checks' ..... 60**
9.1. How Do Platforms Prevent Repeat Offenders? ..... 60
9.2. How Do Platforms Verify Business Information? ..... 60
9.3. What Countries Enforce Account Checks ..... 62
9.4. Account Checks: External vs. Internal Systems ..... 63

# 1. Introduction

## 1.1. Background and Context

This report was funded by Ofcom, the United Kingdom's regulator for Online Safety and a range of communication services. The aim was to build understanding of mitigation practices relating to fraudulent advertising, in order to support Ofcom's work under the Online Safety Act 2023, which requires many of the largest online user-to-user and search services to, among other things, take proportionate steps to prevent individuals encountering fraudulent advertising.

As Ofcom worked to implement the Online Safety Act in relation to advertising and fraud, it identified a number of novel challenges that had not yet been adequately addressed by existing public research. This report reflects Ofcom's collaboration with the Integrity Institute to gather expert experiences and perspectives relevant to these key issues.

Under the United Kingdom's Online Safety Act 2023, "an advertisement is a 'paid-for advertisement"' in relation to an internet service if:

(a) the provider of the service receives any consideration (monetary or non-monetary) for the advertisement (whether directly from the advertiser or indirectly from another person), and

(b) the placement of the advertisement is determined by systems or processes that are agreed between the parties entering into the contract relating to the advertisement.[4]

For this particular report, this report designates **advertisements, "boosted" content, and sponsored content** as within scope, and generally excludes **organic brand content** as out of scope. We define these types of content in the next chapter.

### Types of Fraudulent Advertising

Online fraud can take many forms. The report focuses on user-sided fraudulent experiences rather than platform-sided fraud such as click fraud. The following types of fraud via advertising were considered as particularly relevant for the scope of this report:

[4] UK Parliament, 2023. Online Safety Act

1. **Bad user experiences**

Bad user experiences can cover a wide range of fraudulent experiences. This includes fraudulent, fake, or counterfeit products, products that are extremely low quality, products that do not match what is shown in the advertisement, or cases where no product or service is ultimately provided.

2. **Malvertising**

Malvertising is an advertisement that can directly compromise a user's device or account, or is an advertisement that leads to a website that will attempt to compromise a user's device or account. Advertisements that are themselves malware, exploits, or viruses, or lead to those harmful experiences, are covered in the report.

3. **Scams**

An advertisement that leads a user to be caught in a scam, which will typically be financial in nature. This can include fraudulent financial services or even advertisements that lead to scams involving blackmail.

4. **Data sales and retargeting**

Advertisements that exist solely or primarily for the purposes of harvesting data from a user. This data can be directly sold or used in additional scam operations.

5. **Other forms of illegal Content**

An advertisement that includes content that is itself illegal or advertises the sale of illegal goods or goods that are illegal to advertise. This could include the sale of illegal narcotics or weapons where prohibited.

## 1.2. Methodology

To produce this report, the Integrity Institute identified seven members of our community with experience working on integrity and Trust & Safety issues related to advertising and who were no longer working for a platform at the time of study. The interviewees have a combined 20+ years of experience working on integrity and Trust & Safety issues in advertising. Their expertise spans seven online platforms, each with millions of annual UK users, all either user-to-user or search services. Each platform features paid-for online advertising and has functionalities that are relevant to services that have additional duties under the Online Safety Act 2023 (such

as recommender systems). Each participant was chosen for their knowledge, expertise, and awareness of industry best practices in fraud mitigation and platform operations. Across the seven interviewees, we conducted ten one-hour-long interviews in 2025 to gather the findings and insights detailed in this report.

This methodology means that the report is not an exhaustive account of mitigations related to fraudulent advertising, and there are inherent limitations given interviewees no longer work at these platforms. We have, however, only included evidence that reflects wider industry knowledge about Trust & Safety practices in this area.

It should be noted that all recommendations and opinions expressed in the main sections of this report are the opinions of interviewees, and not directly of Ofcom or the Integrity Institute.

# 2. How Does the Online Advertising Ecosystem Work?

Understanding how fraudulent advertising operates requires first understanding the digital advertising ecosystem itself. This section includes baseline information on the types of online advertising and the systems that display them.

## 2.1. Types of Online Advertisements

There is a range of different types of advertising within the platform ecosystem.

*Table 1. Types of Online Advertisements*

| Category | Definition | Financial Relationship |
|---|---|---|
| 1. Advertisement | Paid promotional content, falling in either the feed or advertising section, where advertisers pay for placement. | Advertiser pays the platform/publisher directly for ad space or impressions. |
| 2. Boosted Content | A piece of content that a user pays to promote to a network beyond their existing followers. | Advertiser pays the platform/publisher directly for impressions or reach. |
| 3. Sponsored Content | Paid content created specifically for promotional purposes, often in partnership with publishers or influencers. It's designed to blend with the platform's regular content while often being disclosed as paid promotion. | Advertiser pays a user (often an influencer) to produce and distribute content.<br><br>The platform may also take a cut (like YouTube's ad revenue share or Instagram's branded content fees) |
| 4. Organic Brand Content | Non-paid content that brands create and share naturally through their own channels. | No advertising spend. |

1. **Advertisements** - *Advertiser pays the platform, via a bid, to advertise in a section of the platform*

This happens in both internal ("walled garden") and external ("open display") ad platforms. Typically, these ads appear either in the feed as a post, before a video plays, or to the side in a designated advertisement space. These are by far the most common types of advertisements. Also included in this scope are 'dark advertisements', adverts that target specific users and only appear in their feed, often simulating "organic" content. Since advertisements represent the dominant revenue model for platforms, interviewees noted it as a significant pathway for fraud within the system.

2. **"Boosted" content** - *User pays the platform to increase distribution for "organic" content*

Boosted content starts as normal organic content. After posting, an advertiser pays the platform to boost content impressions on a post for a certain number of impressions or a period of time. This typically extends the "reach" (total number of people exposed to the ad) of a post beyond the traditional audience by targeting users that wouldn't otherwise have the content recommended to them.

From an outsider's perspective, there may seem to be little difference between "boosted" content and advertisements. Boosted content does include a monetary transaction between the user/advertiser and the platform. However, interviewees noted that organizationally boosted content can exist outside of a company's traditional ad integrity team's remit, and have separate moderation systems. Boosted content can fall into a less governed space between advertisements and traditional content generated by users, since boosted content is typically simpler to set up than paid advertisements; can be less expensive; may not be investigated as heavily as traditional ads; won't get checked by the same content classifiers as traditional ads; and is excluded from ad transparency products such as ad libraries and APIs.

**Case Study:** Boosted content was a tactic used by Russia to interfere with the Moldovan elections. The Integrity Institute's analysis[5] on boosted content in Moldova showed how boosted posts were less moderated, less researched, and had the potential to reach and influence a large number of people. Though this example focuses on election interference, fraudsters use similar strategies to avoid detection.

3. **Sponsored content** - *An advertiser pays a platform user to advertise a product to their follower base*

A sponsor pays a user, often an account with a large following or an influencer, to advertise a product to their follower base. As a result, these types of advertisements can present a challenge to the platform, since the financial relationship between the advertiser and the account may or may not be

[5] Integrity Institute, 2025. Boosted Content and Electoral Risk in Moldova

disclosed to the platform and may or may not be disclosed to the audience, depending on local regulation and whether the advertiser and account are following local regulation.

Interviewees shared that this style of advertising presents numerous surfaces for fraud. The sponsor may be paying influencers to spread the fraudulent content or products to get access to account data or financial information, or the influencer may have even more direct involvement with the fraud. The sponsor may be defrauding the influencer themselves, offering free products or shipping, withholding payment, offering fictitious sponsorship programs, or looking for ways to hack users or influencers.

The platform is not involved in the financial relationship, which only exists between the content creator and the advertiser. However, since the platform is offering the medium in which sponsored content operates, interviewees suggested that platforms should hold some accountability since a large reach can be easily achieved through sponsored content on the platform. In addition, they noted platforms have existing mechanisms for sponsorship disclosure, and may have a financial incentive to facilitate, rather than restrict, sponsored content where they take a share of the creator's earnings.

4. **Organic Brand Content** - *An organization's account posts about its product*

Many businesses have accounts on a platform where they post about their company, products, deals, or news in an organic fashion. Additionally, these are the kinds of content that organizations often pay to boost. Interviewees stated that fraudulent links in these posts won't trigger the advertising integrity team review because they fall under organic content, and will be looked at under other content-moderation processes.

However, many users might consider these posts 'ads' despite organizations not paying the platforms (unless content is also boosted). Interviewees noted that users have made complaints about the number of 'advertisements' from organizations that appear in their content feed, even though these posts are organic. Anti-spam measures can sometimes cover this behavior when fraud is involved; however, since there is no financial relationship between the content creator and the platform, these posts don't have the same level of safety as more traditional ads.

**Case Study:** Platform "Shops" and Marketplaces

Some platforms have created marketplaces within the platform, so that vendors can sell their products to users of the platform, all while staying on the platform itself. This can be beneficial to both the platform and the vendor, because the platform can take a share of the sale price of the goods as a fee from the vendor, and the vendor can use payment services offered by the platform in addition to accessing the large user base as potential customers.

However, it also brings the platform directly into the operations of the sale of goods. In the case of fraudulent goods, this can create direct monetary benefit to the company that operates the platform from the sale of fraudulent goods or services. In some cases, content creators can create content that highlights goods for sale in the platform's "shop" section, and then for any sale of that product based upon a user seeing that content, a share of the sale will go to the content creator and the platform company. This creates a three-way financial relationship between the platform, creator, and vendor. A post can exist as organic content, sponsored content, boosted content, or an advertisement, as a creator organically posts sponsored content and then boosts it to get more customers cuts across all types of ads. Suddenly, any platform with a shop will be combined with a marketplace platform, which creates an ambiguous, less-regulated ad market.

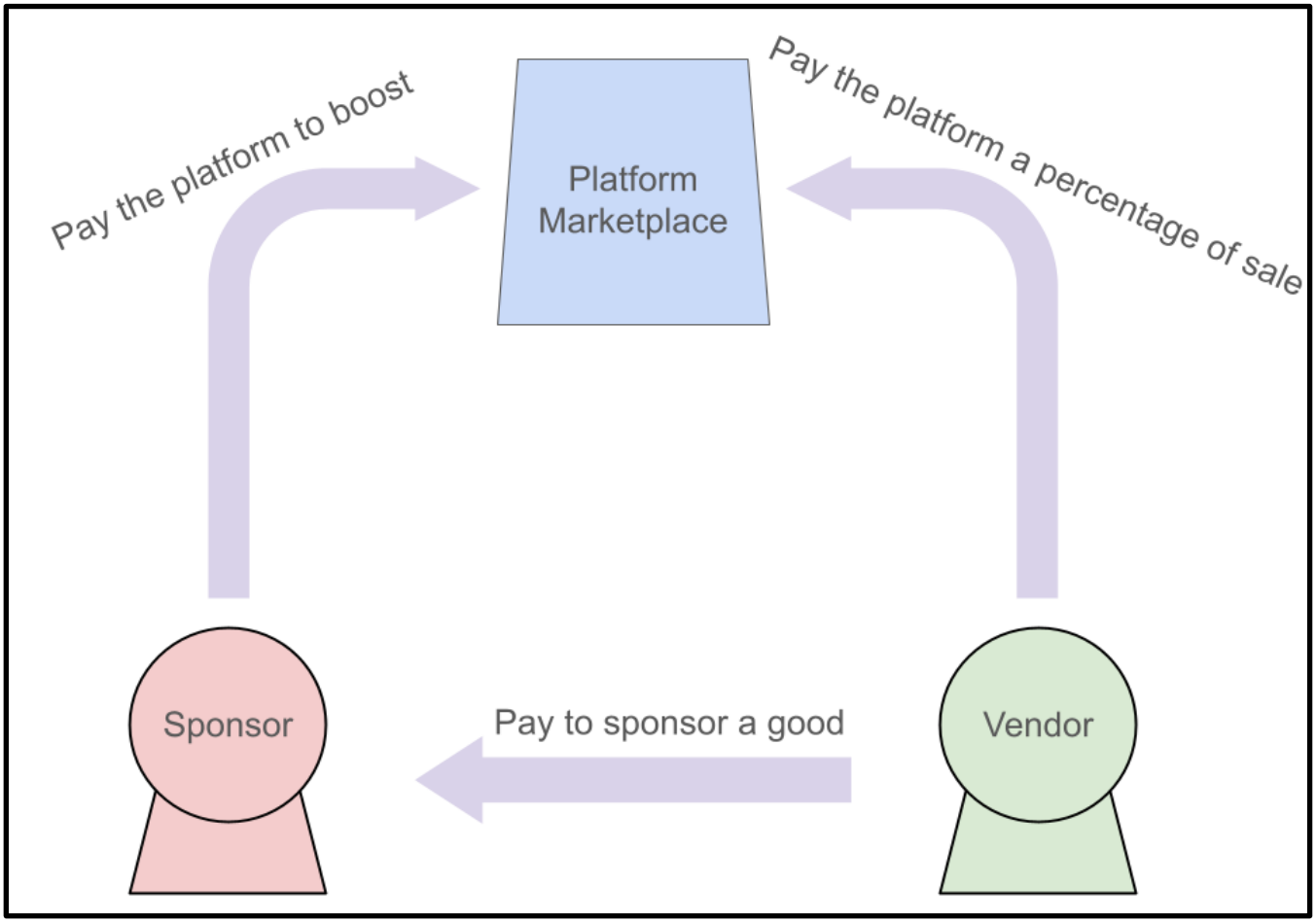


*Figure 1. Relationship Between Vendors, Sponsors, and Marketplaces*

## 2.2. How Are Ads Shown to Users?

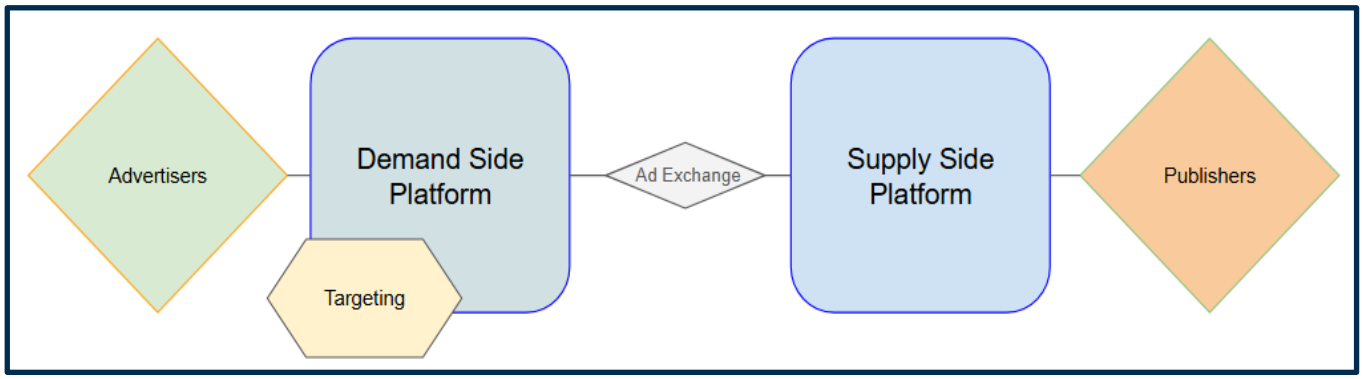


*Figure 2. Mapping of the Ad Ecosystem*

*Table 2. Advertisement System Components*

| Component | Description |
|---|---|
| 1. Advertisers | Brands create and manage advertising campaigns through dedicated platforms. |
| 2. Demand Side Platform | Advertisers purchase ad space and define their target audiences using various criteria and optimization strategies. |
| 3. Ad Exchange | Ad placements are determined through competitive bidding systems. |
| 4. Supply Side Platform | Marketplaces connect advertisers and publishers to buy and sell ad inventory in real-time. |
| 5. Publishers (Platform/Provider) | Publishers manage and sell their available advertising space to generate revenue. |

Based on our conversations with interviewees and wider industry knowledge and evidence on this topic, the main components of online ad systems can be described as follows:

1. **Advertisers**

Advertisers are brands, companies, or individuals who pay to promote their products, services, or content. Advertisers drive demand for ad impressions by purchasing advertising space to increase audience reach or sales.

- *Advertiser Platform*
  A platform for advertisers to upload content and create ad campaigns.

- *Auction-Based Ad Display*
  A system where advertisers bid for ad impressions. The final ad selection can be based on bid value or a mixture of bid value and engagement.

**2. Demand Side Platform**

A demand-side platform (DSP) is a service used by advertisers to buy digital ad inventory. Ad targeting systems are used on these platforms so that advertisers can set criteria used to select the audience. There are several different frameworks used:

- *Targeting Criteria*
  The advertisers can specify characteristics of the desired audience, such as age, gender, and location.

- *Optimization Goals*
  Advertisers can optimize for views, clicks, or purchases

- *Interest/Topic-Based*
  Uses platform-tracked interests and keywords to recommend ads

- *Manual Lists*
  Advertisers can upload user lists (email-based targeting)

- *Lookalike Audiences*
  Expands target groups to similar groups

**3. Ad Exchange**

An ad exchange is a digital marketplace where advertisers (via DSPs) and publishers (via SSPs) buy and sell ad inventory in real time, usually through automated auctions.

- *Open Ad Exchange*
  Virtual marketplace that offers open auctions

- *Private Ad Exchanges*
  Private marketplace offered to premium publishers

- *Preferred Ad Exchanges*
  Marketplace for ads at preferred or specific price

**4. Supply Side Platform**

A supply side platform (SSP) is used by platforms, publishers, and other websites to manage, sell, and optimize their available ad "inventory," which is potential impressions of ads to be shown to users.

5. **Publishers (Platform/Provider)**

Publishers are the entities that have ad space available for sale.

The terminology around DSPs, SSPs, and ad exchanges is primarily used for external ad systems, where each element can be its own external vendor. However, interviewees noted that platforms with internal ad systems will still have an equivalent system with components that correspond roughly to the above.

## 2.3. External (Open Display) Ad Systems

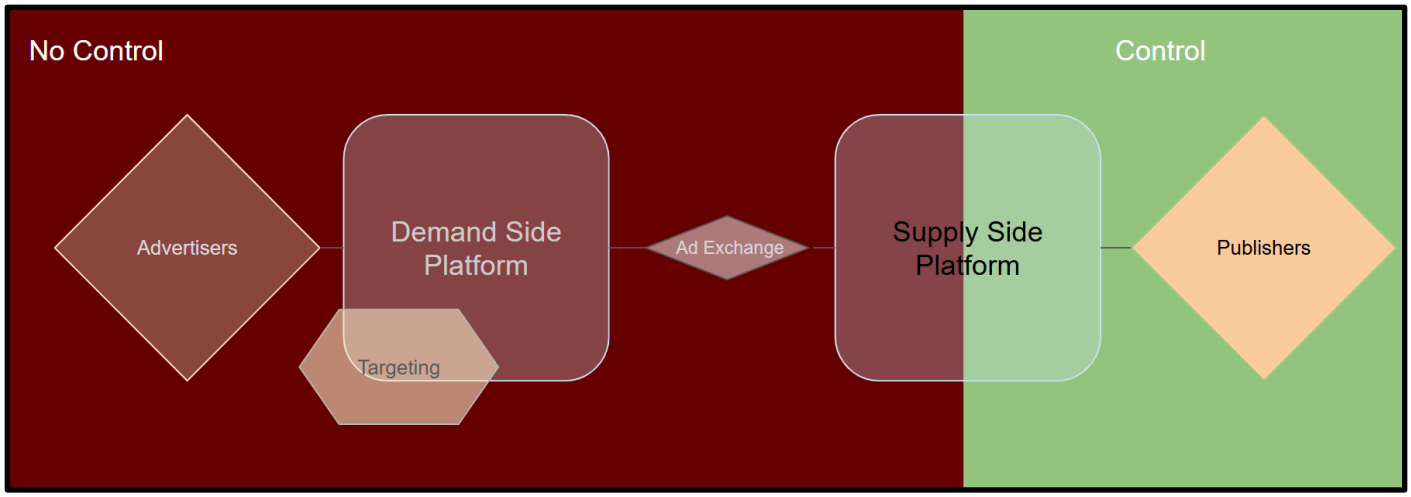


*Figure 3. External Ad System Control*

A platform that uses an External Ad System only has direct control over the publisher (themselves) and what supply-side platform they use. The ad exchange, demand side platform, and which advertisers get selected are up to their external vendor and system.

Interviewees stated this creates challenges relative to platforms that run their own internal ad system. Using an external ad network, publishers and platforms must rely on third-party services and vendors, potentially with lower ad standards than the platform would like to set, or a lack of granularity in options that a platform might need.

There are third-party vendors that offer services that can help monitor and maintain quality standards for publishers, but the level of control will generally be much lower for platforms using external ad systems compared to platforms with internal ad systems. There are generally strong incentives for publishers and platforms to rely on third-party vendors for protecting against fraudulent ads. Third-party vendors enable platforms to begin displaying ads more rapidly than would otherwise be possible, while also providing access to established DSP and advertiser partnerships.

## 2.4. Internal (Walled Garden) Ad Systems

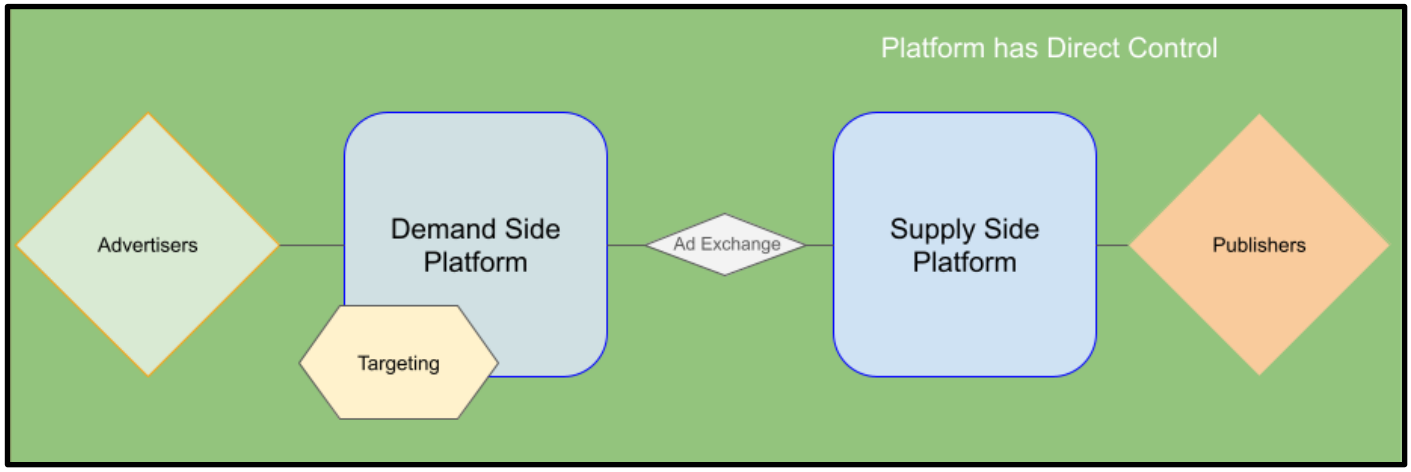


*Figure 4. Internal Ad System Control*

Many of the larger platforms use internal ad networks. The top platforms (*Google, Facebook, and Amazon*) use walled gardens and accounted for 78%[6] of global digital ad revenue in 2022, encompassing a massive user base. Internal ad networks give the platform much more control over the user experience, policy standards, detection of harmful advertisements, and enforcement against them. Interviewees noted that more control also means more powerful ad targeting through in-house algorithms and first-party data.

Interviewees characterized Walled Garden ad systems as giving platforms **comprehensive control over all aspects of the stack**. They control the criteria to become an advertiser, what they advertise, how users are targeted, and how users are reached.

Verification processes for ads and advertisers range from basic algorithmic detection to full manual review. Some platforms opt for a hybrid approach, which combines internal teams as well as third-party vendors for reviewing certain content areas or threat intelligence, enabling subject matter experts to handle certain content that is highly researched or tailored.

Platforms own and control how they collect and use user data, which means that all data around the advertising system, coming from advertisers, the advertisement itself, or users viewing the ads, can be used for detection, monitoring, and enforcement against harmful ads. These systems require robust in-house tooling and resources to be able to perform all data collection, labeling, and ad functionalities for their users and advertisers – but there is a strong incentive for companies to build them since they are essential to enabling revenue. Additionally, since the system is entirely in-house, it makes it more difficult for auditors or the public to gain transparency.[7]

---

[6] J. G. Navarro, 2026. Share of walled gardens versus the open internet in digital advertising revenue worldwide from 2017 to 2027

[7] US House Science, Space and Technology Committee, 2021. Testimony of Laura Edelson, NYU

## Internal Ad Systems Ecosystem: Operational Steps

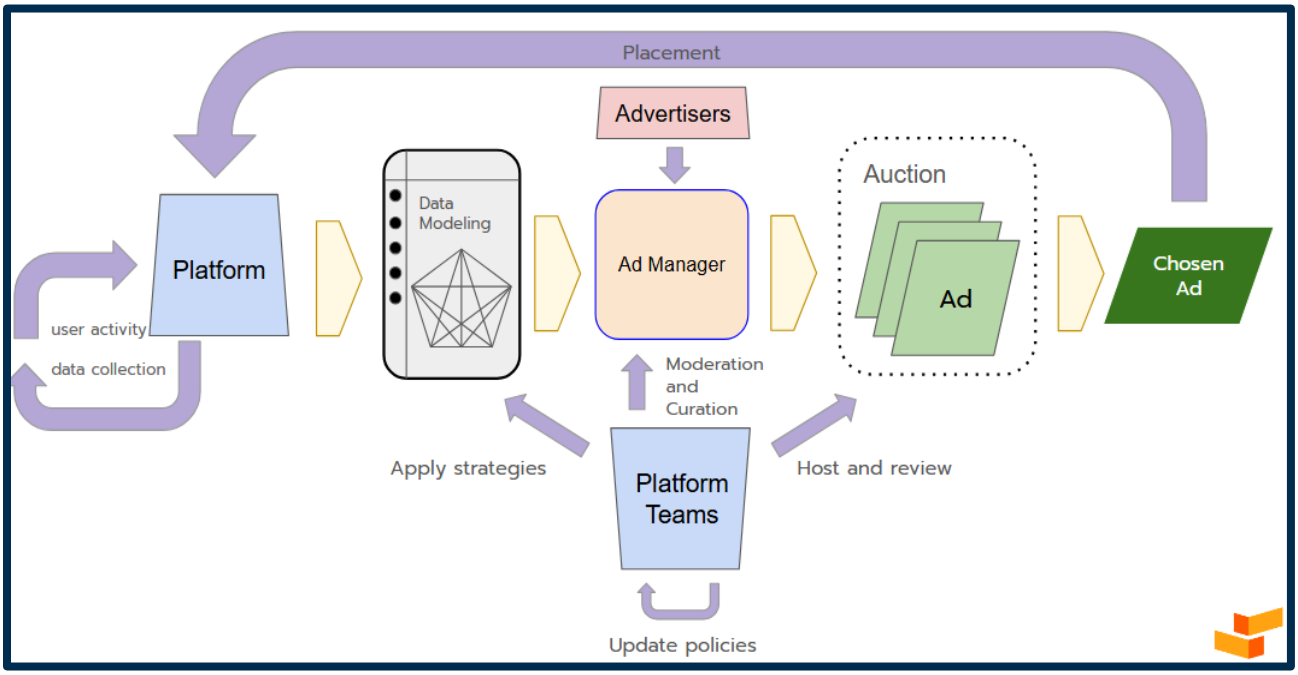


*Figure 5. Overview of Operational Steps in an Internal Ad Ecosystem*

Internal ad systems can grow to be extremely complex, since they incorporate all the elements of the open display ecosystem, which is typically spread across many companies, into one single system. A basic list of the key operational elements, identified by interviewees, of an internal ad system includes:

*Table 3. Operational Steps in Internal Ad Systems*

| **Operational steps** | **Description** |
|---|---|
| In-platform user activity and data collection | Assorted behavioral, interest, and technical signals like watch time, likes, comments, follows, skips, device, geo-location, language, and on-platform purchases are collected from users' activity on the platform. |
| User data organization and modeling | The collected data is centralized and structured to be used in the advertising system. The data is used to build interests graphs and predict demographic and geographic properties of users. It is also used to allocate users into targeting buckets to predict what actions individual users will take in response to seeing an ad. |

[Cybersecurity for Democracy]

| | |
|---|---|
| Ad access and operations via the platform ad manager | Tools for advertisers to use for running advertisements. This will include creating and managing accounts to run ads, as well as ad manager tools to create ads, set objectives, choose ad formats, and define audiences. |
| Ad review and moderation | Systems will be put into place to review advertisements. This includes both automated content moderation systems run proactively as well as manual review, done by human moderators, of advertisers and ads that can be initiated after user reports of policy violations. The review of material does not need to be limited to on-platform content and actions.<br>Practices include regular review of off-platform content, in particular, the landing pages and domains of any links in the advertisement. Once an assessment is made, moderation actions can be taken, which can include ad approval, rejection, pause, and escalation for further consideration. Advertisements and ad campaigns can also be reviewed in a continual process, if users report it or new signals of abuse or fraud arrive, either from internal or external sources. |
| Ad auction and delivery | Advertisements will "compete" with each other to be shown to users. The competition can incorporate an assessed quality of the ad and the likelihood of the user to take an action on it, as well as the price the advertiser paid to have it shown. |
| Ad placement | Ads will be placed in relevant locations on the platform, which can include feeds, before or after videos, in direct messages to the user, or notifications to the user. |
| Ad performance tracking and reporting | Engagement and performance metrics for the advertisement, including views, clicks, comments or replies, off-platform conversions, or direct purchases, will be tracked and reported back to the advertiser. |

# 3. Platform Detection and Mitigation Strategies

## 3.1. Platform Incentives to Combat Fraud

Interviewees talked about how platforms have incentives to lower the amount of fraudulent advertising on the site, and felt they should be able to communicate how they balance any tensions

between revenue from fraudulent advertisements, user safety, and long-term user trust in the platform. Large companies understand the impact fraud has on the use of their platform and should want to minimize fraudulent ads without external influence. From a purely business perspective, if users are defrauded, platforms have evidence that users will reduce their engagement with the platform.

Interviewees therefore felt platforms have an interest in minimizing the fraudulent experiences on the platform to ensure engagement does not decrease.

Interviewees discussed how the relationship between platforms, advertisers, and users is built on trust. The advertiser trusts the platform to promote their content safely and truthfully. The user trusts the platform to promote advertisements that lead to trustworthy experiences. As soon as a platform loses this trust, it is difficult for the platform to get it back, leading to users leaving the platform or no longer engaging with advertisements, and thus advertisers pulling out. Advertisers want to know they are getting what they paid for via accurate tracking data from the platform. Additionally, the threat of lawsuits from users, government agencies, and/or advertisers encourages platforms to combat fraud.

> *"A lack of data leads to lack of trust from advertisers.*
> *A lack of trust leads advertisers to pull ads from platforms."*
>
> - Interviewee with 5 Years Experience at a Large Social Platform

Regulators, platforms, users, and advertisers all want (and benefit from) fewer fraudulent experiences on platforms' advertising systems. However, interviewees described how platforms manage tensions between short-term and long-term incentives. Allowing less regulated ads may generate short-term revenue, but it will hurt long-term trust and engagement. Implementing extensive safety checks may hurt initial revenue, but it keeps users and advertisers happy, setting up the platform for long-term sustainable income.

## 3.2. How Do Platforms Detect Fraud?

In general, interviewees described significant overlap between methods used for the detection of fraudulent and harmful adverts and those used for the detection of all types of policy-violating content on platforms. Since platforms should have policies in place against fraud, there will be similar processes and operations used to identify policy-violating content that will be used for identifying fraudulent advertisements. This will include automated detection, user flagging, and off-platform identification and detection. Human review and moderation operations are also used to detect and take action against fraud.

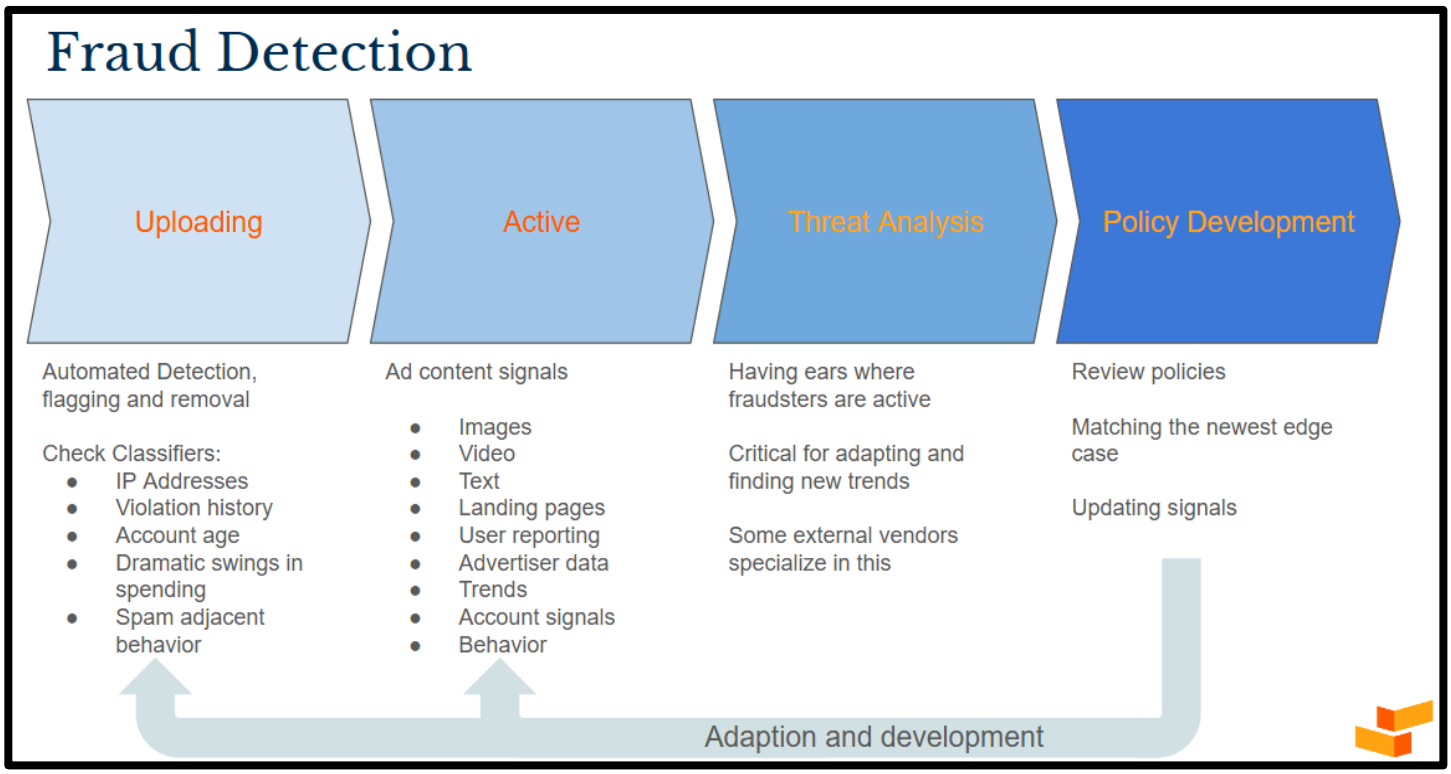


*Figure 6. Overview of different stages of fraud detection*

*Table 4. Summary of Different Fraud Detection Approaches*

| | |
|---|---|
| **I. Automated Detection** | Platforms deploy machine learning classifiers to identify advertisements and advertisers that likely violate ad policies using signals such as IP addresses, account age, spending patterns, and violation history. |
| **II. User flagging** | Users can report individual ads or accounts they suspect violate policies, which helps platforms prioritize content for automated or human review. |
| **III. Off-Platform Identification and Detection** | Platforms address fraud occurring outside their ecosystems by hiring specialized experts or partnering with third-party vendors to investigate activity they cannot directly monitor. |

### I. Automated Detection

Interviewees noted that platforms, especially large platforms, will have machine learning classifiers that are used to predict if content violates any of their policies. This will typically be discussed around classifiers that identify adult content or graphic violence. However, the same processes are used on advertisements to identify ads and advertisers that are likely to violate ad policies.

Some common fraud detection technical signals include:

- Actor-based signals
- IP Addresses
- Violation history
- Account age
- Dramatic swings in spending
- Spam adjacent behavior

### II. User Flagging

Platforms also have processes for users to flag or report content that they believe violates either content or advertisement policies. Interviewees said this will include user reporting flows that enable users to report individual ads or entire accounts as possibly violating. The reports will be used to inform how content is enqueued and prioritized for additional automated or human review.

### III. Off-Platform Identification and Detection

Off-platform experiences and highly motivated bad actors create special challenges. Interviewees described how this inherently involves digital spaces where the company will not have full access to the data that tracks all activity in the space. Special skills will be required for the platform to gain insights into whether off-platform experiences represent fraud. For this, platforms can either hire experts, rely on third-party vendors, or both.

## 3.3. The Detection Process

If a platform has prior data, classifiers, or trends on its backend, interviewees noted, detection begins at the uploading process. Machine learning classifiers or other technical signals and systems are frequently used to assess if the content and account are safe under the platform's policies and guidelines. If there is a new trend or a new account from a fraudster, sometimes the signals don't exist. This is where further analysis may be necessary as more data is collected (user reporting, further actions by account, collection of external signals).

The process of detecting and evaluating fraudulent ads depends on the size and resources of the platform. For larger platforms, internal teams will regularly review randomly sampled ad impressions and the most viewed ads. For smaller platforms or publishers, external vendors can review ads. These evaluations should result in an estimate of the prevalence of policy-violating ads, which should include a breakout for fraudulent ads, which should take the form of both the prevalence of violating impressions and the fraction of revenue generated from violating ads.

Interviewees shared several key questions that platforms should consider when evaluating potential signs of fraud:

- *Where does the money come from?*
- *Why is this person paying money to advertise?*
- *How is their money showing up on the platform systems?*
- *What is the advertiser funding?*
- *What is the full customer experience?*
    - Interviewees stressed this should include looking externally into ad assets and landing pages in the ad to evaluate end-to-end user experience, from clicking an ad to purchasing, to determine fraudulent activity.
    - However, they noted the amount of experience a single platform takes responsibility for is not consistent across the industry. They reported platforms tend to include "one click" in their fraud evaluations, meaning only the landing page of links in the ad is included in the evaluation.

Interviewees emphasized that thorough evaluations involving these topics, including the evaluation of customer experiences, can be done on a regular basis, especially for larger platforms. Ideally, this would enable not only the continuous measurements of fraudulent activity but also the development of new policies around fraud detection and identification. Interviewees stressed that effective detection practices should lead to policy development and change based on findings, leading the way to the development of new techniques for dealing with existing policy violations or ways of identifying new types of fraud.

**Case Study:** Well-intentioned businesses get flagged

Interviewees brought up that sometimes, well-intentioned businesses get flagged as fraudulent due to a single problematic admin. Advertisers will typically have multiple users tied to their ad accounts as admins. Large advertisers will use many admins. If one of these admins from the large advertiser is caught spreading fraudulent ads from a separate ad account, the advertiser can still appear fine and non-violating, since it was one admin out of many.

However, this is challenging for a small advertiser with one or two admins. They can be negatively impacted by working with a potentially problematic consultant with prior or recent fraudulent activity. It is extremely punishing when well-intended businesses get flagged as engaging in fraudulent activity, especially smaller businesses with only one or two admins.

## 3.4. Threat Intelligence

Interviewees noted how it is important for companies to pay attention to digital spaces where fraudsters are active. They described how the threat intelligence process, which can include monitoring fraudsters in spaces where they share tactics and sell their services, is a vital step for countering fraud comprehensively, since fraud can happen during off-platform experiences. Investigating the "dark web," tracking new tactics, monitoring user data sales, identifying malware and redirects, examining hidden code, and other similar behaviors was said to be crucial for stopping fraud. However, it was also recognised that it requires a high level of expertise both to infiltrate spaces and understand those discussions. This type of threat intelligence can be a specialty of some external vendors, since detection of off-platform issues requires engineering expertise (and familiarity with malware, cloaking, data harvesting, and other specialist areas).

Deceptive behavior online is a constantly evolving space because bad actors are highly motivated, highly adversarial, and always developing new techniques. Genuine expertise is needed, on an ongoing basis, to combat these threats.

# 4. Best Practice in Detection and Mitigation

This section examines strategies for detecting and mitigating fraudulent advertising, drawing on insights from the expert interviews.

## 4.1. Best Practices in Detection

**Strong Governance**

Interviewees emphasized that detection of fraud begins with a clear articulation of the business incentive to remove fraud from the platform and build trust with users and advertisers, and to incorporate this understanding into the platform's governance and operation. Without a clear articulation of the harm that fraudulent activity causes the platform, there will be an increased risk that a platform will allow fraud to occur on the platform unnecessarily or fail to resource anti-fraud efforts.

Interviewees noted the inherent tensions between a platform's short-term and long-term business interests. For example, if platforms begin to allow fraud, they will gain more revenue tomorrow but lose user and advertiser trust over time. Interviewees stressed that platforms should have processes in place to manage the tradeoff since strong systems today can be eroded over time by bad governance.

**Internal Processes Alignment**

An additional best practice noted by interviewees is the significant operational overlap for dealing with any type of violating content. Fraudulent ads will be just one type of violating content for any platform, and the detection and mitigation of it can rely on the typically large operations companies built around all violating content. Research, policy development, measurement, and monitoring are all steps of any significant effort to curb violating impressions universally, and so all of these can be brought to bear on fraud.

Below are examples of best practices highlighted during the interviews:

*Table 5. Summary of best practice mitigations raised by interviewees*

| **Category** | **Identified Best Practices** |
|---|---|
| Regular measuring and monitoring in line with other policy violations | • Measurement and monitoring of prevalence should be continuous and done at regular intervals.<br>• Measurement and monitoring can also be supported by vendors, which can help add capacity to smaller companies. |
| Measurement of standard content moderation metrics for ad fraud | Employing the following metrics:<br>• The prevalence of violating content, which can include prevalence at the level of impression, as well as revenue.<br>• The total number of impressions on violating content and the total amount of revenue earned from policy-violating ads.<br>• The total number of user reports on content, including ads, and the tracking report reason, which should include fraudulent activity. |
| Operational Effectiveness Metrics | Employing the following metrics:<br>• Time to removal<br>• Views before removal<br>• Precision and recall of the operation. |

It was recommended that platforms implement and standardize these practices across their operations to ensure consistency and long-term effectiveness in combating fraudulent advertising.

**Unique Challenges and Enhanced Transparency**

Interviewees identified a number of distinct features where fraudulent advertising presents unique challenges, which stem from: the off-platform experiences, the wide variety and high motivation of actors, the continuous innovation in bad actor tactics, and important regional contexts which can lead to large variations in region and language. Additionally, the variety of types of advertisements, products, and types of fraud adds complexity to addressing fraud comprehensively. While interviewees noted that there was some overlap with other areas of violating content, such as foreign information, manipulation, and influence (FIMI) operations, some of these features will be unique to fraudulent advertisements.

This is where external consultation was noted as a method of importance and value. Interviewees noted how transparency tools, such as ad libraries, can help external groups flag issues. Ad libraries have proven to be a useful tool that enables external researchers to monitor and successfully identify and track scams, fraudulent ads[8], and foreign influence operations[9]. Ad libraries have been made available by several large platforms.

Interviewees noted that the current offerings of platforms could be greatly improved to better empower researchers to identify new trends in fraud. To support platform efforts against fraudulent advertising, researchers have stated the need for access to broad, representative ad datasets rather than being limited to searches for specific actors.

Interviewees indicated that improving these capabilities would significantly enhance the ability of researchers and civil society organizations to identify emerging fraud trends and assist platforms in addressing them. The following features were identified as essential components of an effective ad library:

**Comprehensive data**: spend estimates, targeting parameters, and impression counts.

**Advanced search functionality**: beyond keyword-based queries, including image/video detection.

**Representative sampling:** methods to obtain random samples weighted by impressions or spend.

These enhancements would strengthen transparency and collaboration, enabling external stakeholders to monitor and mitigate fraudulent advertising more effectively.

## 4.2. Best Practices in Mitigation

Interviewees noted that there is a similar set of practices that are commonly used by platforms to mitigate fraudulent advertising. Mainly, they noted mitigation begins with monitoring, and that if a platform isn't monitoring for fraudulent ads, then it won't understand there is a problem to mitigate. Similar to fraud detection, fraud mitigation measures are able to take advantage of a platform's large systems, processes, and governance structures, borrowed from other policy-violating areas affecting the platform.

[8] ProPublica, 2024. Exploiting Meta's Weaknesses, Deceptive Political Ads Thrived on Facebook and Instagram in Run-Up to Election

[9] AI Forensics, 2024. No Embargo In Sight: Meta lets Pro-Russia Propaganda Ads Flood the EU

#### Comprehensive Policies

A throughline best practice from the interviewees was that platforms should put comprehensive policies in place that capture all forms of fraud and are adaptable to new tactics developed by fraudsters. These policies help operational teams handle ambiguity and include enabling the operational teams to take a variety of actions against violating or likely violating ads. Interviewees stated that having a spectrum of actions, such as the ability to remove ads, remove ad accounts, pause ads or ad accounts, or otherwise limit ads that potentially violate policies, will help ensure that fraudsters are not able to exploit loopholes in platform policy.

#### Governance and Operational Effectiveness

Additionally, interviewees indicated that effective mitigation requires coordination across multiple roles and teams, including ad reviewers, product teams, and policy teams. Each of these groups plays a key part in reducing the prevalence of advertisements that are fraudulent. Prevalence on a platform will be different depending on region, language, or user base. Ensuring that these roles and teams are sufficiently staffed and also have the capability to share insights and coordinate their efforts across the platform was seen as crucial.

Finally, interviewees highlighted how employing teams dedicated to and rewarded for driving the prevalence rate of fraudulent ads down can encourage employees to find new mitigation strategies. Due to the nature of fraudulent advertising, teams need continuous monitoring to mitigate these problems. For effective mitigation, interviews indicated that a platform should look at being more effective than its competitors, since bad actors are looking for the easiest platforms to take advantage of, and any platform that puts enough friction to deter fraudulent actors will minimize fraudulent experiences on their platform.

## 4.3. Best Practices in Technical Systems for Detection and Mitigation

Technology plays an essential role in effective detection and mitigation. Interviewees referenced the importance of detection, scraping, reporting, computer vision, machine learning classifiers, and other technical systems. It was determined that the technical systems to support the detection and mitigation of fraudulent ads need to be comprehensive and continually developed and evaluated. There are a few key areas where it was felt that technical systems need to be developed:

*Table 6. Summary of interviewee recommendations for best practices in technical systems*

| Technical System | Description |
|---|---|
| 1. Content Analysis | Platforms need comprehensive systems that analyze text, images, and videos to detect both known and novel fraudulent content, with regular accuracy measurements and updates to identify new fraud patterns. |
| 2. Behavioral Analysis | Platforms track advertiser signals such as IP addresses, device IDs, payment methods, and spending patterns to identify accounts linked to previous violations or exhibiting suspicious activity, with systems requiring routine updates and evaluation. |
| 3. Landing Page Analysis | Platforms extract signals from advertisement landing pages, root domains, and linked pages to detect malware and threats, often relying on third-party vendors since bad actors actively obfuscate harmful content. |
| 4. Ad Design | Platforms with internal ad networks leverage insights into campaign design, targeting, and optimization to develop predictive signals for identifying fraudulent advertisements. |
| 5. User Reporting | Platforms collect user engagement data and violation reports on running ads, though delayed fraud discovery (such as counterfeit products) may require ad transparency tools that let users review and report previously seen advertisements. |

**1. Content Analysis**

Interviewees emphasised that platforms need comprehensive systems that can evaluate content and estimate the likelihood that it could be fraudulent. Interviewees determined the baseline of these systems to include: analyzing the text, image, and video content; matching to content that is known to have been fraudulent, as well as assessing novel content; and having ongoing accuracy measurements, such as precision and recall for identifying violating content. They felt these processes should be refreshed and updated for new fraudulent content on a regular basis.

**2. Behavioral Analysis**

Interviewees indicated that platforms track behavioral and identity signals from advertisers to make predictions as to whether or not they have or will post fraudulent ads. They noted that these signals should be based on trying to match the advertiser's current account with accounts that have previously posted violating content, such as IP addresses, device IDs, geolocation, and payment

methods, as well as behavioral signals that could indicate suspicious activity, such as wild changes in advertising spend in a long-dormant account. Similar to the content analysis systems, these should be routinely updated and evaluated.

### 3. Landing Page Analysis

Interviewees stated that signals should be extracted from the landing page of any links in an advertisement. These could be combined with additional signals from the root domain or additional pages on the domain or linked to from the landing page. These signals would likely need to be evaluated against known malware and threats. It was noted that this process presents challenges for platforms, as the gathering of these signals is adversarial. Bad actors will obfuscate the harmful content or make it appear benign to systems run by the platforms. One workaround interviewees suggested was for platforms to rely on third-party vendors to support internal systems and analysis.

### 4. Ad Design

Platforms with internal ad networks will have full insights into how the ads are designed, targeted, and optimized. Interviewees stated these can be used to develop signals and systems for predicting if an ad campaign is fraudulent.

### 5. User Reporting

As an advertisement is running, the platform can collect signals from users, including how users engage with the ad or users reporting the ad as violating. Interviewees noted that, from experience, some forms of fraud will not be captured by user reporting. For example, if a user has a bad customer experience because the product was counterfeit, they may not know that until days or weeks after they have seen the ad, and may not be easily able to track down the advertisement that led them to the product. A potential mitigation noted by interviewees was encouraging the platforms to counteract this by allowing users to see advertisements they have engaged with in a user transparency type product, allowing them to easily report potential fraudulent advertisements.

# 5. Best Practices in Platform Design, Policy, and Governance

This section examines strategies for operationalizing design, policy, and governance processes in addressing fraudulent advertising, drawing on insights from the expert interviews.

## 5.1. Design

Interviewees identified friction as a key tool, enabling platforms to slow down fraudsters and encourage them to find other platforms to move to, especially considering that, in their experience, repeat offenders are responsible for the majority of fraudulent adverts. Better detection and operational frustration for fraudsters can cause bad actors to move on once they realize the cost or effort of attempting to commit fraud on the platform is not worth their time.

Effective ways of introducing friction that were suggested include: phone number and physical address verification, device verification (through two-factor authentication methods), geolocation verification, and creating delays between creating an account and reaching a large audience through ads.

Another area of best practice identified was to enable a wide range of feedback from users, make it as easy as possible for users to provide feedback, and ensure that reporting is easy for all forms of advertisements on the platform. This should include being able to report ads as violating, but also gathering softer signals from users, like enabling them to remove the ad from their feed, block the advertiser, or even report the advertisement as low quality or untrustworthy to encourage more trustworthy advertising.

## 5.2. Policy

Interviewees noted how platform policy changes were identified as the major catalyst for tackling fraudulent advertising. Their experience concluded that constantly making sure that certain edge cases are accounted for is critical for tackling fraud comprehensively. Typically, platform policies go through a regular baseline review every 6-12 months, but that duration can change depending on data received from various teams (i.e., detection teams may report a spike[10] in novel forms of fraud or operational teams may report a spike in edge cases that are not being caught). Empowering teams to create effective policy solutions was a key best practice identified by all the interviewees.

[10] 'Spike' refers to a noticeable increase of a particular metric, such as a large increase in reports on a particular advertisement.,

Interviewees identified that internal platform policies often focus on explicitly prohibited content, such as weapon sales, and that edge cases, such as certain cases of hate speech or misinformation, are often extremely difficult to target and, thus, are often exploited. It was recommended as best practice that platforms should take into account the following considerations:

- Is there a policy on a given borderline situation? Are those policies executable?
- How are teams able to create new policies or adapt prior policies?
- What is the methodology for fine-tuning ad fraud policies?

## 5.3. Governance

Interviewees emphasized the spectrum of organizational responses and structures that companies could use to effectively counter fraudulent advertisements. Some platforms may leave fraud to policy teams, engineering teams, or Trust & Safety teams. The core responsibilities of reducing fraudulent advertising are interspersed across teams, with each platform dispersing the responsibility uniquely. However, interviewees agreed that a clear structure for accountability is important, as situations where accountability is ambiguous can allow problems to slip through the organizational cracks.

> *"Who is responsible for executing? There should be an accountable owner [of all the processes] within the company."*
>
> - Interviewee on Policy Teams, 6+ Years Experience

Interviewees noted that, from experience, teams that combat fraud also run the risk of interfering with other teams within the company, such as teams focused on revenue growth or advertiser account growth. This follows, since removing fraudulent accounts has inverse success metrics, such as total account growth. As efforts to combat fraud become more effective, there could be a short-term reduction in revenue since fewer advertisers are being allowed on the platform.

**Case Study:** Safety tradeoffs and decision-making

Governance teams need a strong understanding of the tensions and tradeoffs that exist when adopting processes that prioritize safety. In the worst case scenario, the largest advertisers on a platform are fraudsters, and the platform can't tackle the issue without serious financial loss. As a result, governance teams need levers (processes, policies, and mitigation strategies) and a deep understanding of the tradeoffs of allowing fraudulent advertisers to exist on the platform to determine the most effective and practical mitigation plan of action.

Interviewees emphasised how important it is for companies to have clear governance structures in place that make clear how the company ensures that the desire for short-term revenue growth will never outweigh fighting fraud and increasing user safety on the platform. Potential best practices identified include a clear understanding and articulation of how fraudulent advertisements hurt the long-term health of the company, and clear processes and accountability for any situations where relevant tensions arise.

*Table 7. Summary of Best Practices Across Design, Policy, and Governance*

| Category | Identified Best Practices |
|---|---|
| **Design** | Platforms should introduce friction through verification methods (phone numbers, addresses, devices, geolocation) and account delays to deter repeat offenders, while making it easy for users to report, block, or provide feedback on advertisements. |
| **Policy** | Platforms to regularly review and update ad policies every 6-12 months (or more frequently based on emerging fraud patterns), empowering teams to address edge cases and create executable policies that close loopholes exploited by fraudsters. |
| **Governance** | Platforms need clear organizational structures that assign accountability for combating fraudulent advertising across policy, engineering, and Trust & Safety teams to prevent problems from falling through the cracks. |

# 6. Example Case Studies

In the interviews, we asked participants to talk us through how platforms might want to respond to different scenarios they could face relating to fraudulent ads.

**Case Study:** Open Display (External) vs Walled Garden (Internal) - Response to Increase in Reporting

**Scenario:** Many users are reporting fraudulent experiences on the ads you serve, resulting in a large increase of reported fraud.

Open Display (External)

**To investigate this further, platforms should ask:** what data is available to the platform?

- Platform: User data, user reports
- Ad Exchange: Advertiser information of the bid winner, ad placement information, and post impressions
    - However, interviewees noted that advertiser information can be extremely limited or obfuscated by resellers
- SSP: Performance and analytic data
- *Sometimes* ad creative (swapping possible) and sometimes off-site landing page (swapping possible)[11]

**After investigating the available data, what actions are possible for the platform to take?**

- Block advertiser or reseller
- Reach out to the Ad Exchange or SSP
    - It may or may not be possible for the platform to change its settings in the SSP to limit the impact.

Walled Garden (Internal)

**To investigate this further, platforms should ask:** what data is available to the platform?

- The answer: *everything*
    - Advertiser signals (Device IDs, financial signals, full history on platform)
    - Ad signals (Full content, full targeting information, true links to offsite locations)
    - User signals (Full user reporting)

**After investigating the available data, what actions are feasible for the platform to take?**

- Block advertiser or reseller
- Policies can change
- Platform design can change
- Teams can be (re)organized
- Specific advertisers can be banned

[11] 'Swapping' refers to a technique where an advertiser submits a policy-compliant advertisement for platform review and approval, then subsequently replaces the approved creative or landing page.

**Case Study:** Challenges with the user's ability to retrospectively flag a fraudulent ad

**Scenario**: A user is defrauded after viewing an advertisement on the platform. It takes two weeks for the product to ship and arrive, and only once it arrives does the user realize they have been defrauded. The user has no way to report the advertisement that led them to purchase the product.

**Potential Solution**: The more a platform has control over its internal advertising, the more it is able to mitigate the risk of users not being able to report ads retrospectively. Participants discussed whether platforms could be more responsible for off-platform activities or areas with limited control. Platforms maintain records of some of the ads that users have seen, such as certain political ads. Maintaining these records could enable users to report ads retrospectively. A potential best practice could be to require large platforms to register and track a much higher proportion of ads, to then enable users to report advertisements that they saw in the past, once the ad has been identified as fraudulent.

**Case Study:** Limited control using external ad networks

When using external ad networks, the platforms are sometimes entirely dependent upon other companies.

**Scenario**: A spike in fraudulent ads is seen coming from a large ad reseller. The platform doesn't get advertiser information because the data that they get in the SSP only indicates the last link in the supply chain, which, in this case, is a reseller. The platform only knows that ads are coming in via a large, established Reseller.

**Potential solution**: Currently, the platform can flag ads to the reseller and can request more information, but blocking a large reseller (typically the only available mitigation) can be challenging, because it is possible that a significant amount of revenue goes through the reseller. This is especially true if the reseller controls a large portion of ads and only a small percentage are truly fraudulent. Interviewees identified that more transparency and controls for publishers in open display networks could help give more control to the external ad platforms and provide them with more knowledge to understand the nature of fraud on their platform. This could include "buyers.json" or other forms of comprehensive supply chain transparency, which would empower platforms and publishers to take more targeted actions to reduce fraud.

# 7. What Are the Key Challenges for Regulation and Platform Mitigation?

This section examines interviewees' views on key challenges facing regulators and platforms in addressing fraudulent advertising as regulation and mandatory public transparency are increased.

## 7.1. Key Challenges for Regulation

Interviewees explained that increased transparency requirements can create negative incentives for companies, and that public-facing metrics risk being manipulated if the transparency provided by platforms, or the metrics required by regulation, are not sufficiently comprehensive. Interviewees felt that providing more public transparency around the prevalence of violating content on a platform and the number of exposures to violating content creates a potential incentive for the companies to modify their policies in order to narrow the scope of content that ends up violating the policies. For example, a platform may have a broad and comprehensive definition of negative ad experiences, which will result in a particular prevalence level. If they narrow the definition of negative ad experiences, that will inherently reduce the prevalence, since fewer advertisements will fall in scope. This could, in theory, make the platform's metrics seem like they are trending in a positive direction, when in reality the user experience is worsening.

This can lead to a situation where platforms "move the goalposts," changing policies to show a decline in identified fraud. In order to combat this, interviewees discussed how comprehensive transparency is required. For fraudulent ads, interviewees said this should include both public components and a transparency requirement to regulators, auditors, and public interest researchers to ensure that platforms are not changing policies to report fewer cases of fraud.

Interviewees stressed the importance of transparency requirements that allow external researchers and regulators to accurately track platform policies, classifiers, and overall fraudulent advertising mitigation efforts. Interviewees identified several essential public transparency metrics regarding fraudulent ads that they thought should be part of comprehensive platform requirements:

- The number of exposures to advertisements that were deemed to be violative by the platform, and the revenue generated by the platform from them
- The prevalence of violative ads, measured by both impressions and revenue
- An estimate of the true number of exposures to violative ads based upon the prevalence measurements
- The total number of advertisements that were deemed violative
- A breakdown of the above metrics by

    - Demographic properties of the audience of the advertisements (age, location)
    - Country of origin of the fraudulent advertisements, based on the ad account

- Metrics around the operational performance of the ad moderation system
    - Time to moderation, views before moderation, accuracy of moderation decision, accuracy of automated systems
- The metrics the company uses to measure the overall success and safety of their advertising systems, including topline growth metrics, revenue metrics, Trust & Safety metrics, and how frequently there are tradeoffs between those metrics, and how the company manages tradeoffs between them
- Public datasets of
    - The most viewed ads on the platform
    - A random sample of ads weighted by impressions
    - A random sample of ads weighted by revenue generated

However, interviewees argued that making these metrics public could create some incentive for the platforms to redefine negative ad experiences in a way to make total exposures and prevalence drop, simply due to the definition. To counteract this, it was suggested that regulators, auditors, and public interest researchers could be granted access to the full history of the policy definitions for violating advertisements, the full history of the human moderator guidelines used to evaluate if ads violate, and datasets of advertisements that went through the moderation operation and the decision that was made based upon them. This would enable experts and stakeholders to verify that platforms maintain consistent definitions of violating content rather than making arbitrary changes.

**Case Study:** "Ad Farms" Policy

As an example of a "moving the goalposts" situation, we can consider a common policy that many platforms have against "ad farms", which are websites with an excessively high number of intrusive ads on them, which are made simply for generating ad revenue for the owners. Such sites are strategically designed with no other purpose than displaying ads.

An interviewee discussed how a platform might have an external-facing policy that states they do not allow advertisements on the platform that lead to ad farms, and an internal-facing policy in which an ad farm is an external site where, for example, 30% of pixels on the page are ads means it is considered an ad farm. If new transparency requirements call on the platform to demonstrate how their strategy lowers the number of ad farm impressions, they might change the figure on their internal policy from 30% to 80%. This will provide them with one potential way to claim they have lowered ad farm impressions without having meaningfully done so.

## 7.2. Key Challenges for Platform Mitigation of Fraudulent Advertising

Platforms also face consistent challenges as they combat fraudulent advertising, and there are areas identified where there is some genuine uncertainty in the industry. These were some of the identified challenges from the interviewees:

*Table 8. Summary of Interviewee Views on Challenges with Mitigation for Platforms*

| **Topic area** | **Challenges faced** |
|---|---|
| **Off-Platform experiences**<br><br>What happens when an ad is policy-appropriate, but the landing page isn't? | • Most platforms claim responsibility for "one click away"<br>• User reporting is difficult for many types of fraud, especially when fraud occurs off the platform<br>• Users who were defrauded days ago can't locate the original ad<br>• Users who get malware won't be immediately aware<br>• Users whose data is sold won't be aware |
| **Adversarial nature of fraudsters on the platform**<br><br>The adversarial nature of fraudsters on the platform and the tactics they use make auditing fraudulent advertising difficult for multiple reasons: | • Users aren't able to retroactively report fraudulent ads due to time passing since being defrauded<br>• Some scams operate over long timelines<br>• Images and content can be swapped out<br>• Some scams look for small exposure to the "right" audience.<br>• Fraudsters adapt quickly<br>• Platforms that use external ad networks (and those that don't manage the entire stack) will have much less control, often being unable to target ban or accurately respond to fraud as it appears |

| | |
|---|---|
| **Novel forms of advertising**<br><br>Platform advertising has evolved far beyond traditional display ads. | • Normal display ads<br>• Boosted content<br>• Sponsored content<br>• Platform shops<br>• Platform teams are consistently trying to develop new forms of advertising, which may create new vulnerabilities to fraud. |
| **Global scale** | • Moderators and external checking processes may lack sufficient cultural context to make an informed decision about whether or not some adverts are violative. The prevalence is typically much higher in some countries than in others, depending on language and other factors. |
| **Policies and definitions can change** | • The technical definitions of policy violations cannot be public, since bad actors will exploit them, but the definitions could be made available to auditors or regulators to ensure compliance and good practice. |
| **Public scrutiny is limited** | • Most ad tools only enable access to public ad data by brand or keywords, not by engagement, date, or any other features.<br>• Additionally, researchers need to be able to search for fraudulent behaviors that matter, which means allowing researchers the ability to search for platform metrics and measurements, for example, prevalence, engagement, or takedowns. |

*"Illegal sellers don't want 10M views, they want 1000 highly targeted views in 10 minutes [and] then [to] disappear."*

- Interviewee with 6 Years of Experience Combating Fraud

## 7.3. Costs

The cost of internal ad safety staffing varies widely because roles differ significantly in expertise, location, and operational scale. Ad reviewers are relatively low-cost but fluctuate based on geography and required volumes, while senior product roles command much higher salaries due to their specialized skills and strategic responsibilities. As a result, total annual costs span a broad range depending on how mature and globally distributed the ad platform's Trust & Safety function is.

**Cost Considerations (Internal Ad Platform)**

*Note: There is no standardization amongst the Trust & Safety ecosystem; these are estimates and based upon historical evidence and interviewee experience.*

*Table 9. Cost Considerations of Internal Platforms*

| Role | Cost per Head | Headcount Range | Annual Cost Range |
|---|---|---|---|
| Ad Reviewers | ~£17,250 (average)<br>• £11,500 (India)<br>• £23,000 (Dublin) | 10-50 *minimum* | £115K - £1.15M |
| Senior Product Team Members | £190K - £380K | 2-10 | £380K - £3.8M |
| **Total Internal Ad Staffing** | | | **~£500K - £5M** |

**Cost Considerations (External Ad Platform)**

For platforms that are using an external ad system, the costs they incur will largely be determined by the third-party vendors they use for combating fraudulent ads. Using third-party vendors will be a preferred solution for smaller and medium-sized platforms that have limited visibility into all the data inside ad exchanges and the speed at which vendors can turn on basic functionality for advertisement safety.

It is not impossible to reach **£1M** in vendor and operational costs of running a full external ad system. This cost will largely track the typical fees from companies that provide ad suitability and safety services to publishers.

# 8. Policy Issue Deep Dive: Account Integrity and Inauthentic Accounts

The section explores account takeover scenarios and the mitigation strategies platforms employ to combat them, highlighting the distinct approaches used in internal versus external advertising systems.

## 8.1. Context: Who Buys Ads?

Interviewees noted that the purchase of advertisements as an action on a platform can be taken by a variety of different platform assets, entities, or accounts. The people and companies purchasing ads can represent themselves in many different ways when doing so. From the interviews, we identified three primary groups:

**Individuals:** A person buying an ad

**Companies:** Typically, a team of people representing a company buying ads

**Agencies:** Typically, a team of people that may represent many different companies

Each group represents itself using many different "assets" or "entities" on platforms, as noted in the diagram below.

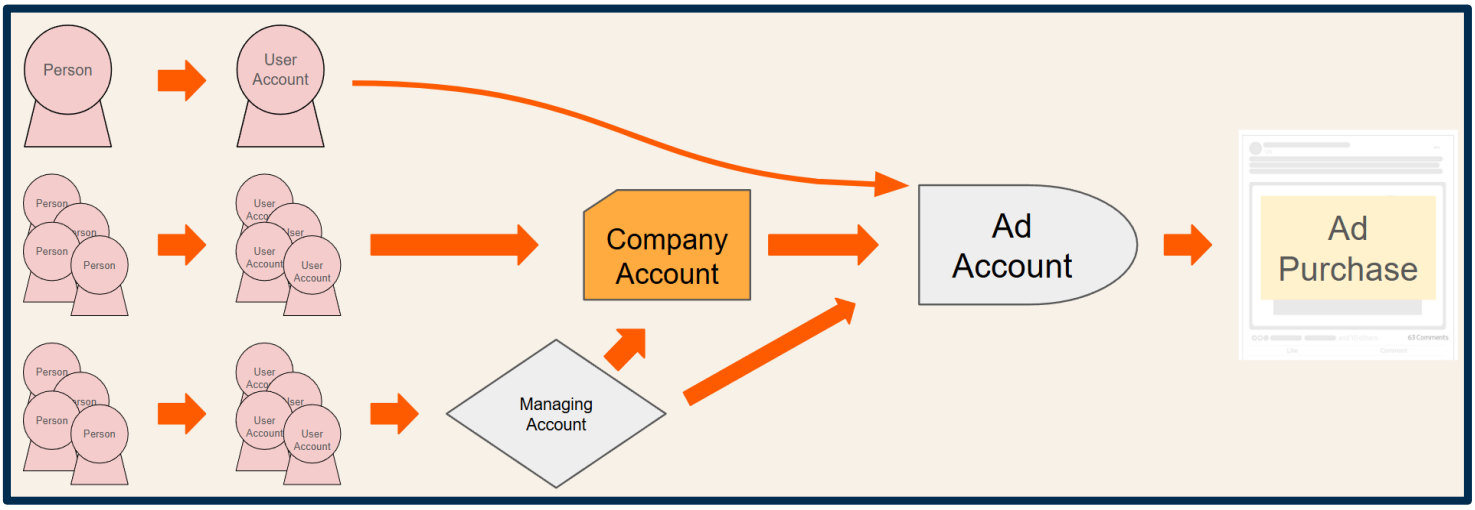


*Figure 7. Variety of Account Types that Buy Advertisements*

These assets and entities include

**User Accounts**

**Company Accounts**

**Ad Accounts**

Some accounts have multiple people behind them, tied to either numerous individual accounts or managed via a shared login. Business, company, and ad accounts will commonly have multiple users. Ads can also link on-platform entities to off-platform entities, such as webpages, domains, or accounts on other platforms.

Interviewees noted there are opportunities for inauthentic activity in each of these different on and off-platform entities, and in every step of this process of mapping people to companies to ads, which can lead to users, companies, and the platform itself being taken advantage of.

To track the impact of inauthentic activity within the ad systems, interviewees described how platforms can track metrics such as:

Ad payments and benefits

- What % of (ads, ad impressions, ad revenue) stem from accounts that have authenticity issues?
- What % of (ads, ad impressions, ad revenue) benefit domains/companies that have authenticity issues?
- What % of (ads, ad impressions, ad revenue) lead to scams, fraud, counterfeit products?

Geographic and demographic breakdowns

- Where is the origin of the fraud?
- Who are the targets of the fraud?

Interviewees noted that large platforms should be, and typically are, measuring all of these metrics.

## 8.2. Account Integrity Threats

Several common problems reduce the integrity of accounts that can purchase ads on platforms, which impact all on-platform accounts, but also off-platform entities as well. Interviewees reported that platforms typically have relevant policies in place that enable mitigation efforts.

These were the three most common threats identified and described by interviewees, who have professionally handled these threats at platforms internally.

*Table 10. Summary of Account Integrity Threats*

| Account Integrity Issue | Description |
|---|---|
| Impersonation | Bad actors create fake accounts or domains that mimic trusted public figures or companies, using stolen or AI-generated content to exploit established credibility and trust. |
| Fraudulent Businesses | Real companies with actual employees operate to sell counterfeit products or fraudulent services, either misrepresenting legitimate goods or deliberately imitating established brands through deceptive advertising. |
| Account Compromise | Malicious actors gain full or limited control of user, business, or ad accounts through takeovers, incorrect permissions, or off-platform domain hijacking, leveraging existing credibility and verification status to conduct fraud or resell access. |

**Impersonation**

Impersonation is when a bad actor creates assets, such as user accounts, company accounts, or domains, that impersonate another person or company, typically a well-known public figure or company, which allows the bad actor to exploit the broad trust people have towards them.[12] There have been many high-profile impersonation attacks that have caused real financial harm to companies and people.[13]

Impersonation is typically accomplished by stealing or copying images, videos, and text from the real people or companies being impersonated. This is a fairly easy and effective tactic, but it is also relatively easy to identify on the platform, since there is often an official brand page or celebrity account the fraudster is attempting to impersonate. Interviewees expected that generative AI tools will enable bad actors to create convincing accounts that are not as trivial to detect as stolen content, which would lower the effort scammers and fraudsters have to exert.[14]

[12] The Times, 2025. Warning over scammers impersonating doctors on social media

[13] The Guardian, 2023. Blue-tick scammers target consumers who complain on X

[14] The Guardian, 2025. Scammers using deepfakes to steal 26M Pounds

**Case Study:** A musician's account that had bought ads years ago is taken over and begins spamming ads.

Fraudsters will be more successful at tricking the fans with a legitimate account. If payment accounts are still linked, the fraudster can advertise for free until the owner or platform notices. In general, the fraudster targets these accounts with two distinct objectives:

1. Directly messaging the followers or posting content in the feed with the intention of using the musician's credibility to evade detection and utilize established trust.
2. Spamming ads using the musician's financial details until the account receives disciplinary actions, using the musician's verified status to make moderation actions difficult.

Fraudsters are less likely to target larger corporations since larger companies monitor their online presence more heavily.

**Fraudulent businesses**

A fraudulent business is a real company, with real employees, but the products or services they sell are fraudulent. This is commonly the case behind counterfeit products sold online. The companies are often selling products that are as advertised, with the intention of misleading buyers via the advertisement, or selling products that are designed to look like common brands and advertising as that brand.

**Account compromise**

Account compromise occurs when a user account is taken over by a malicious actor. In some cases, this will give the attacker full access to the user account, but there are also exploits that only enable limited account access. The compromise of company, business, and ad accounts can also occur when malicious actors are added to a Business or Ad account incorrectly. This again will give the bad actors potentially full or limited control over the business or ad account.

Interviewees noted that compromise can also happen off-platform. A malicious actor could gain control of a domain. This would enable a bad actor to, for example, change the webpage that an existing ad links to contain malware. Inappropriate access can also happen due to poor practices at the company itself. For example, a former employee may still be left with access to the business account. The goals of account compromise can be wide. Some operations will use the accounts to impersonate the real person or company[15]. Compromised accounts may also be sold[16], used for any linked payment methods to the account, or used to gain access to additional, higher-value accounts. Taking over accounts is useful for fraudsters because they gain check marks, verification, and

[15] BBC, 2020. Twitter hack: 130 accounts targeted in attack

[16] Integrity Institute, 2022. The Hidden Economy of Spam

credibility. This makes it much more difficult for the platform to take action as compared to a fresh account. However, interviewees noted that it is not only high-profile accounts that are at risk of takeover. Fraudsters target lower visibility accounts, ideally accounts that have advertised in the past and accounts with additional permissions or status, such as verified statuses or the ability to post political advertisements.

**Case Study:** Account Takeover - Taking over a political ads account

Business and personal accounts, as well as major corporations, must get special approval or reapproval to post political ads. An interviewee identified a threat scenario where a local politician applied to run political ads and received approval to post political ads. After the campaign, they stopped using the account but didn't delete it.

It was noted that these accounts can get taken over easily because they aren't meant to be long-term accounts and are often no longer monitored. A historical account can be an easy target where fraudsters can post a lot of misinformation, spam, and political ads. The historically authentic nature of these accounts will help the bad actors evade classifiers that are evaluating for spammy activity from new accounts. This is an actual threat identified by an interviewee.

## 8.3. Signals of Inauthentic Accounts

Interviewees reported that many of the signals for identifying fraudulent ads overall would also be useful in identifying inauthentic accounts. This includes content, behavioral, and identity, landing page, and ad design signals. Basic signals identified by interviewees included:

*Table 11. Summary of Signals Used to Recognize Account Integrity Threat*

| Account Integrity Issue | Signals |
|---|---|
| Impersonation | • IP address from one country advertises heavily in another<br>• Unverified account impersonating a celebrity who has an official account<br>• Emails and links from an uncommon or commonly fraudulent domain (xyz.com) |

| | |
|---|---|
| Fraudulent Businesses | • IP address from one country advertises heavily in another<br>• Users report the business |
| Account Compromise | • Change in identity signals (IP address, device IDs, location)<br>• Change in behavioral signals (high-risk actions, change in advertising behavior)<br>• An organization or company account begins posting irrelevant or scam content |

*"It's not so often that an account takeover will lead to legitimate ads posted by a [...] fraudster. They will often take over these accounts to post really scammy organic content."*

- Interviewee with 7 years experience

**Basic Signals: Account Information**

Interviewees noted that regular user and business user accounts can advertise, be independently inauthentic, and, depending on the platform, independently accrue strikes for this inauthentic activity. Signs of inauthenticity identified by interviewees included:

- IP address change or bounces back and forth
- Email or phone number changes
- Sudden changes in account behavior (ad spend increases by an excessive amount)
- Multiple accounts tied to the same IP's
- One account tied to multiple accounts IP's
- Change in language (less common now that AI can translate on the page)
- Credit Card changes
- Low-quality content, products, etc.
- Reports

Interviewees also shared the following examples of content signals relating to inauthenticity, often coming from fraudsters impersonating known brands:

*Generic sounding name* - "[Brand] Services" | "[Brand] Solutions"

*Product service or category* - "[Brand] Ads Branding" | "[Brand] Customer Support Services"

*Affiliate domain* - "[Brand] Partnerships Program" | "[Brand] Partner Network"

*Common Fraudulent Topics* - Spammy content | Scams, Crypto, etc.

## 8.4. What Actions Can a Platform Take to Tackle Inauthentic Accounts?

Platforms have a range of actions available when responding to suspicious or potentially fraudulent activity, varying in both effort and severity. One of the lightest-touch interventions is to remind users of platform policies. Often described by interviewees as a "nudge" or "prompt", this approach requires minimal effort and can nonetheless be effective in discouraging behaviour that may breach platform rules, with a positive association with safer user engagement.

At the other end of the spectrum, platforms may seek to restore the account to its rightful owner. This is typically considered the most desirable outcome, as it resolves harm while preserving legitimate access. However, it is also one of the most resource-intensive responses, requiring careful checks to ensure that access is not mistakenly returned to a fraudulent actor rather than the genuine account holder.

Where risks persist or evidence of abuse strengthens, platforms may suspend or disable the account entirely. This is often treated as a last resort, as it carries notable downsides: it can reduce platform revenue and, in cases of account takeover, may unfairly penalize the legitimate user instead of the individual responsible for the fraudulent activity.

Between these two extremes, platforms may apply more targeted restrictions. Limiting posting, for example, allows platforms to reduce the reach or volume of potentially fraudulent advertising without fully suspending the account, helping to contain harm while further review takes place. Similarly, freezing payments can be an effective way to minimise financial harm, though platforms noted that this typically requires a higher evidentiary threshold, as it can be difficult to justify without clear proof of fraudulent intent.

Platforms may also request that a user re-verify their identity, providing an additional check on account ownership. While this can strengthen security, platforms must balance its benefits against the risk of frustrating legitimate users who are asked to re-verify without clear cause. Finally, accounts may be flagged for ongoing monitoring. This is relatively easy to implement once suspicious activity is detected, though its effectiveness depends on whether sufficient behavioral signals can be gathered over time to conclusively establish fraudulent activity.

**Case Study:** A Valid Account Gets Compromised – How do platforms identify when this happens?

Platforms may identify that a previously valid account has been compromised through a combination of account-level and content-based signals. At the account level, unusual IP or device changes, such as logins from new countries or rapid switching between locations, can indicate suspicious activity. Changes to geolocation or device information may also raise flags. In some cases, compromised accounts show changes to payment details, or a change in the account email address if an attacker replaces it with their own. Account age, however, is not typically a useful signal, as compromised accounts may be long-standing.

In addition, platforms may detect compromise through changes in content behaviour. This can include a sudden or dramatic shift in the topics an account posts about, or the account beginning to post fraudulent or harmful content that is inconsistent with its previous activity.

**Case Study:** A Fraudulent User Whose Account Was Removed Returns – How do platforms identify when this happens?

Platforms may identify repeat fraudulent actors by comparing signals from new accounts with those associated with previously removed accounts. At the account level, this often includes matching technical indicators such as IP addresses, device IDs, and payment fingerprints that are the same as, or closely aligned with, those used by a removed account. Email patterns can also be a strong signal: a new account may use the same email address as a removed account, or an address that follows a similar naming structure (for example, small variations on the same username or domain). In contrast, the age of the account is typically not a reliable signal, as these users often create brand-new accounts that begin abusive activity shortly after registration.

Content behaviour provides additional signals. Repeat actors may resume posting identical or highly similar content to that shared before the original account was removed, allowing platforms to link current activity to prior enforcement actions.

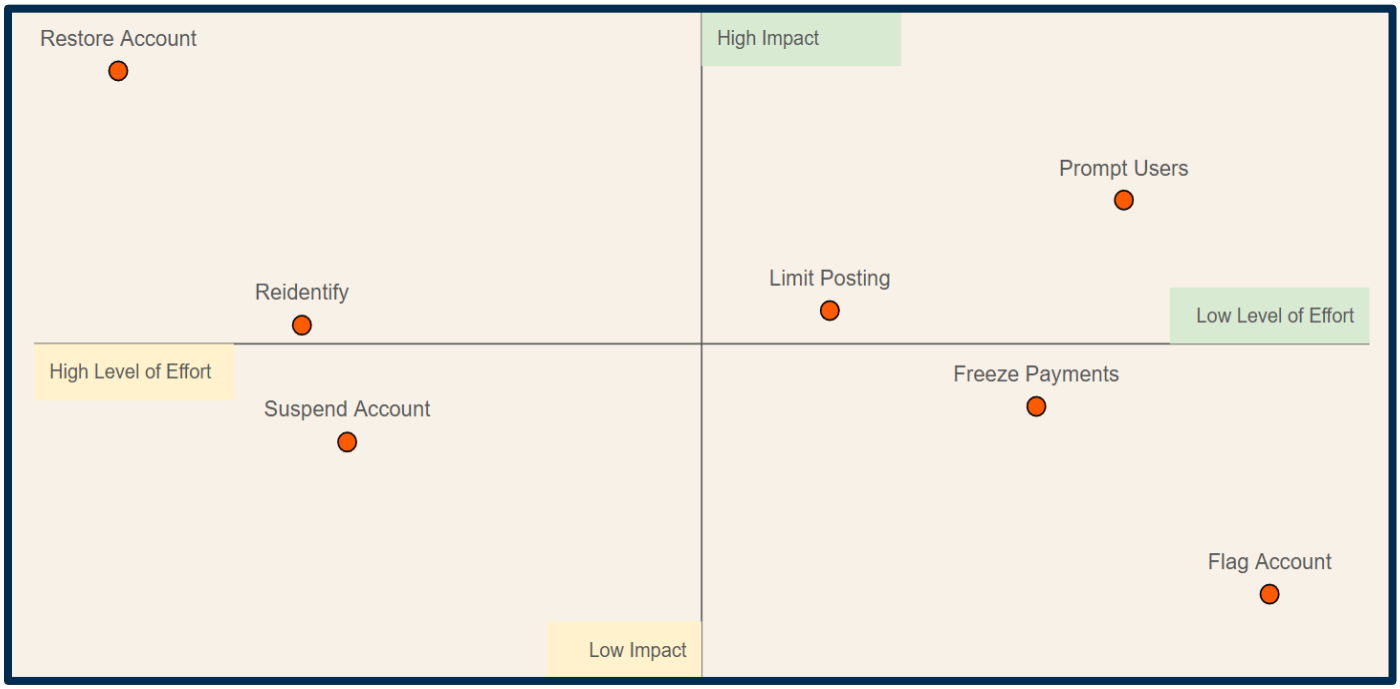


*Figure 11. Impact/effort matrix for account integrity interventions. Based on interviewees' views on what platform mitigations are high or low impact and high or low effort.*

## 8.5. Other Actions for Account Integrity

Platforms have policies in place that cover a wide variety of account authenticity issues. Interviewees identified that impersonating accounts and accounts posting fraudulent ads will normally be policy-violating, and the typical content moderation process for the company will have processes in place to moderate those issues.

Additionally, interviewees determined that account takeovers are more complicated than other fraud tactics, as the solution isn't simply a case of removing the account or not. The ultimate goal in an account takeover should, of course, be the restoration of the account to the rightful owner. However, while the platform attempts to validate who the rightful owner is, validation can take time as it must meet the necessary restoration threshold. During this process, interviewees mentioned that the platform may impose limits that prevent the highest-risk behaviors. This can serve as an effective interim safeguard. If it is impossible to validate who the rightful owner is, the account may ultimately be permanently deleted to prevent misuse.

## 8.6. What Detection Methods are Most Effective?

At a high level, interviewees noted that the detection of account authenticity issues will mirror the detection of fraudulent advertisement issues. There are historic cases of inauthentic accounts that

will be used to train automated systems that can predict the likelihood of an account being inauthentic, which can be used as a signal in measurement and mitigation operations. Interviewees determined that the detection needs to combine internal and external signals, as well as wider intelligence, and will include historical cases of account takeovers. Interviewees identified a number of signals:

**Internal Signals**

- IP monitoring
- Tracking logins
- VPN detection
- Credential monitoring
- Two-Factor Authentication (2FA) / Multi-Factor Authentication (MFA)
- Login alerts
- Behavioral monitoring
    - Ad spend
    - Language changes
- Content monitoring
    - Spam, scams

**Threat Detection and External Signals**

- Honeypots
- Fake sites designed to attract fraudsters
- Dark web monitoring

Detection systems estimate a likelihood that an account is inauthentic, which can be used in the mitigation. Interviewees reported that the strongest confidence will be in cases where there are hard links to previous bad actors, such as device IDs, IP addresses, and payment information that matches previously removed accounts.

If the system is highly confident that the account is inauthentic, the account can potentially be removed promptly after creation. However, many interviewees noted that platforms are typically judging accounts at the time they are posting content rather than at creation, unless it is extremely obvious upon creation that the account is fraudulent. If the system has lower confidence, then the account can be placed in a queue for human review. Nevertheless, interviewees brought up that some platforms are reluctant to remove accounts, and will wait for user reports of the account to come in before taking any action.

User reports could cover authenticity issues, such as:

"My account has been taken over."
"This account is fraudulent."
"This account is impersonating me/my business."

Interviewees emphasised that one signal is never enough, because any of these signals **can** relate to authentic user activity:

A user can move, changing their IP address.
A business can share an account within a family, meaning multiple IPs use the same account.
A business can authentically add a new credit card.

Overall, interviewees explained that platforms are reluctant to take action on new accounts, because there are legitimate causes for new accounts rapidly spending on advertising. Platforms want to enable new small businesses to use their services and advertise, and don't want to make it burdensome for businesses that have legitimately set up their new website recently. Platforms will also be reluctant to take action on accounts unless there is a clear policy violation, because there are legitimate explanations for many actions and seemingly suspicious signals. Bringing in multiple signals, while essential, is challenging to combine, track down, and 'make the case' for why any given account is potentially linked to a fraudster.

> *"[As someone working on mitigating fraud within a platform] You're not generally requiring companies to prove that they control a given asset...*
>
> *you're also not making connections between on platform assets like an ad account and an off-platform asset like a domain."*
>
> \- Interviewee with 6 Years Experience

## 8.7. What Should Proactive Detection Look Like?

Despite platforms analyzing at the time of content posting, ideally, interviewees stated that larger platforms should have proactive detection of inauthentic activity and account takeovers in place. Interviewees state that the best practices of these systems should include:

- Identity signals
  - Suspicious IP/device/location signals
  - Location signals pointing to distant geographic areas
  - Strong MFA signals
- Content signals
  - Spammy content
  - Content previously associated with inauthenticity
  - Content similar to that previously associated with inauthenticity
- Behavior signals

    - Rapid change in ad spend
    - Shift in content topics
    - Change in language or time zone

It was noted that classifiers can be trained to detect when an account may be taken over, and these signals can be used in proactive prevention or remediation.

**Where can platforms increase effort?**

There are major challenges within the anti-fraud space. Interviewees noted that platforms can struggle with country-to-country variation in identification and business registration, multiple people legitimately having authority to act on behalf of the advertiser, and false positives that can incur a tangible cost for users.

Interviewees noted that despite these challenges, there are still several practical best practice steps and areas for improvement, such as:

- Increased investment globally, ensuring that there are no "easy" countries for fraudsters to enter
- Increase in partial measures, such as ID and address verification
- Increased effort around verifying authority and permissions
    - Interviewees noted that one of the biggest challenges is in verifying someone's authority to act on behalf of an entity. They suggested a need to attempt to create tools to empower companies to have stricter authentication processes.
- Verifying email addresses and connections to digital assets, including whether:
    - Emails link to established domains
    - Platform accounts link to established domains
    - Domains have existed for a long period of time with the same owner
    - Domains match the domain of other workers
- Increased penalties for repeat offenses. Interviewees noted that:
    - Actor and behavioral strikes do not often lead to deactivation or a ban for advertising accounts
    - Historically, fraudulent activity has not been treated as the highest priority violation, compared to user-generated posting of more manifestly illegal content, meaning numerous repeat offenses are sometimes allowed
        - *Specifically, interviewees noted that users are banned after posting **one** piece of illegal content, and may stay on even after posting **15 pieces** of fraudulent content.*
    - Platforms are sometimes proactive in strikes given out, but lenient in the consequences
- Increasing the cost of fraudster operations
    - Interviewees suggested requiring assets that are difficult to recreate in the purchase

of ads. For example, platforms accept many "use once" payment methods (prepaid cards, PayPal accounts), which makes it easy for fraudsters to recreate assets for subsequent operations. Disallowing these would frustrate bad actors.

While this list is not exhaustive, interviewees determined it to be a good foundation for platforms of any size to focus on in an effort to reduce fraudulent experiences.

**Case Study:** Proactive Detection - Multiple layers of accounts

Some platforms have a business account infrastructure on top of their personal account infrastructure. An interviewee suggested a scenario where business accounts spanning multiple personal accounts are posting ads, meaning somewhere in the pipeline, users end up with permissions they're not meant to have:

- Someone leaves a business and still has access to an account
- Someone gains access they aren't meant to have
- An admin goes rogue or is compromised

Account permissions end up in uncertain circumstances. This makes it challenging to determine who has access at what time and whether those who have access are meant to have it or not.

Entanglements between ad and personal accounts also lengthen the amount of time to take down a compromised account, since the account is now in a compromised review queue. These queues can mean it takes much longer to review, which allows fraudulent spending to continue longer than it might otherwise.

**Case Study:** Celebrities' likeness in scams

Online scams commonly rely on using images or videos of celebrities, in an unauthorized fashion, for ads with language such as "Celebrity X just endorsed this investment/product." However, interviewees noted that training a machine learning classifier to detect a celebrity's image can require their permission. Platforms can end up waiting for approval from the public figure to automate identifying scams, and the response of public figures can be varied. This pipeline dramatically stretches out the amount of time necessary to begin targeting 'celebrity-bait' scams.

## 8.8. What Forms of Multifactor Authentication (MFA) are Most Effective?

MFA is highly effective at preventing account takeover, particularly for smaller, lower-profile accounts. MFA methods that do have vulnerabilities, such as text messaging, still provide strong coverage for people who aren't the target of more sophisticated operations. Interviewees noted that physical hardware keys provide very robust protection, even for the most at-risk users, but are unlikely to be adopted by the general public due to costs and difficulty of use. Interviewees determined a ranked list of MFA methods as follows, stating that hardware keys are the most secure while SMS is the least:

*Table 12. Summary of Types of MFA*

| Type of MFA | Characteristics |
|---|---|
| Hardware security keys, YubiKeys | Physical devices that generate authentication codes offline offer the strongest protection because they cannot be remotely intercepted or phished. |
| Authenticator Apps | Software applications that generate time-based codes on a user's device, providing strong security since codes are generated locally rather than transmitted over networks. |
| Device-based biometrics (Fingerprint/Face ID) | Authentication methods using unique physical characteristics stored on the device offer convenient security, though potentially vulnerable to sophisticated spoofing techniques. |
| SMS codes | One-time passwords sent via text message. The least secure option because they can be intercepted through SIM swapping, network vulnerabilities, or phishing attacks. However, even SMS codes provide a much higher level of security compared to having no MFA. |

**Costs and Limitations**

Platforms that implement MFA can end up paying a small fee, such as a per-text fee, while security vendors that offer MFA start cheap, such as a "free tier", but can get expensive quickly as the platform scales up.

Additionally, interviewees noted that there are a number of hidden costs in implementing MFA. Some users won't sign up due to friction. Some won't want to share a phone number (for authentic and inauthentic reasons). Some don't understand how MFA works. Problems with MFA may increase customer support needs and costs.

## 8.9. Metrics and Tracking Impact in relation to Inauthentic Accounts

Platforms are tracking a vast number of internal data signals to deal with fraud. Interviewees noted that ideally, platforms should have standardized metrics for tracing the prevalence and total amount of inauthentic activity contributing to ads. A number of potential metrics were identified that would be best practices to measure in analyzing fraud:

**Prevalence - *Scale***

Interviewees noted that prevalence is the industry standard when analyzing fraud, as it is with most policy violations. Estimating the prevalence of inauthentic accounts can largely follow the same operational processes as the prevalence of any violating content. However, they also noted that the effort to evaluate account authenticity, in a vacuum with far fewer user reports, is significantly higher than typical policy violations. Interviewees recommend evaluating what percentage of ad impressions are on ads that violate policies. They noted, however, that evaluation for inauthentic activity, and in particular account takeover, is more labor-intensive than for content policies.

**Impact of detected incidents - *Severity***

Interviewees highlighted the importance, for account authenticity specifically, to use the impact of detected incidents in cases where full prevalence estimates are too difficult or not comprehensive enough. They said it is not only important to understand the number of fraudulent ads, but also the severity of harm that each one causes:

- What percentage of ad impressions, or revenue, were found to be part of an inauthentic operation?
- How many ad impressions, or revenue, were found to be part of an inauthentic operation?
- How does this breakdown by country of fraud origin and demographics of fraud targets?

**Proactive Detection**

Finally, interviewees mentioned that proactive detection methods are important to consider when measuring fraud – in particular, what percentage of ad impressions have strong, moderate, or weak signs of inauthenticity, according to inauthentic activity classifiers.

Most platforms provide no public metrics around ads or ad violations in their transparency reports, despite interviewees noting that most platforms measure signals of authenticity, in addition to signals of inauthenticity. Authenticity signals include accounts that have gone through verification processes and use strong MFA methods. In addition to reporting out metrics detailing the scale of inauthentic accounts, interviewees highlighted that platforms that report out metrics detailing the scale of authentic accounts would be supporting independent research and analysis into the scale of fraudulent advertising. This would largely create a positive incentive for the companies to scale up their practices around authenticating accounts and create a more authentic advertising experience for users.

## 8.10. Account Integrity: External and Internal Systems

The nature of using an external ad network means the platform's ability to verify an advertiser is restricted. Interviewees noted that the data that a platform has access to regarding the identity of the advertiser is limited to what the ad networks provide, which can be minimal or, at the very least, determined by the ad network and not by the platform.

Internal platforms, on the other hand, have unrestrained access to:

- Account information
    - IP
    - Device ID
    - Payment fingerprints (billing zip code)
    - Email
    - Account Age
- Content information
    - Content
    - Creative
    - Landing Page

This creates challenges for platforms using external ad networks to set and enforce their own standards, and enforce them using potentially limited data. On the other hand, internal systems have access to all user and content data needed to analyze fraud on the platform. Interviewees determined that, as with overall fraudulent advertisements, supply chain transparency efforts would empower platforms to be more proactive and secure.

Interviewees identified three key system-level challenges.

1. **Supply chain issues and entry points**

Interviewees noted that external ad systems have some supply chain issues that lead to challenges with ad integrity. Ad websites hosted externally have the potential to cause damage, considering code can be running on ad servers, potentially leading to integrity faults. Platforms using external ad networks rely solely on their vendors to do the integrity work on the ads they are hosting, as they do not have direct access to the advertisement data. Interviewees brought up that these external vendors may or may not do integrity work to the level of rigor of the platform or to the granular standards that the platform desires.

> *"There is no possible way to guarantee that you won't have malicious ads in there [when using external ad systems] unless you are explicitly certain about all the entities and all the people working for those entities."*
>
> \- Interviewee with 10 Years of Expertise

Fraud can occur at any stage of the advertising pipeline, with interviewees noting that violations happen even within platforms' internal systems. However, interviewees reported that external ad networks — which involve numerous independent vendors and companies — create significantly more entry points where a compromised entity can introduce fraud into the pipeline.

2. **Limits of third-party Integration**

Interviewees brought up that platforms, within both internal and external ad systems, frequently partner with vendors who offer services to detect bad actors in ads. They noted that that many vendors provide very close integration, where vendor and platform employees will be in shared communication channels. Vendors can work on the same systems as employees, operating in the same codebase and systems.

However, interviewees noted that vendors typically have hard limitations on data and metrics, and the platform cannot access any metrics that the vendors don't already track. In general, the external vendor may offer useful reports, information, and audits, but the platform won't be able to get specified information or data outside of what's already tracked. The platform may have limited ability to encourage vendors to provide one-off solutions, studies, or data.

3. **"Ecosystem Enforcement"**

Many interviewees brought up how it can be challenging for platforms to take strong actions against bad actors, especially when bad actors come through a large reseller, since often a platform will only be able to stop the entire reseller, resulting in authentic ads being included in the ban.

This creates challenges if a significant amount of platform revenue is coming from a large reseller that includes fraudulent activity. Platforms find themselves in challenging situations, and interviewees described how each may have different thresholds at which they would take major action. In 2014, App Nexus fired a third of its customers, over suspected fraud.[17]

Ideally, interviewees said, if you are a reseller, a platform will hold you to the same standards as any other entity, whether you are just one person or made up of 10,000 entities. One potential best practice identified was to increase data sharing and transparency to publishers and platforms about who is purchasing ad impressions from them, and who has been identified as a malicious actor.

## 8.11. Cost Implications Related to Account Integrity

Interviewees noted that platforms are able to avoid penalties that incentivize other businesses to avoid fraud. For example, chargebacks are a common fee that payment processors will apply to a business for any disputed customer purchases. This incentivizes most businesses to ensure that they are providing a level of service and product quality that matches their customers' expectations. But for platforms that are selling advertising, the chargebacks would only apply to them if a customer disputed the cost they paid for advertising. By this nature, interviewees noted how advertising-based platforms avoid the penalty that typically comes from connecting customers to fraudulent experiences.

Interviewees additionally noted that platforms are incentivized to move from short-message service (SMS) MFA to authenticator apps because they don't want to pay for the SMS messages, since they are charged per message. SMS two-factor authentication costs platforms about half a cent US per SMS authentication, noted by one interviewee, which can become very costly at scale. Additionally, platforms are incentivized to avoid ID verification since ID verification costs about a dollar US per check, noted by one interviewee, and may dissuade users from signing up due to privacy concerns.

The primary costs platforms face here are in opportunity costs. Interviewees noted that platforms are incentivized to make the process of posting an ad as frictionless as possible, because any amount of friction lowers the amount of potential advertisers (even legitimate advertisers). Slowing things down results in lower revenue, and advertisers may leave the platform if they have to verify identity. This may not always be a signal of fraudulent behavior, as many individuals have mistrust of data practices and privacy, and will not begin advertising due to the burden of joining.

[17] Ad Exchanger, 2014. App Nexus Fired a Third of Their Customers over Suspected Fraud

*"The way that big tech thinks about this is not in terms of financial cost; they think about it in terms of productivity loss."*

- Interviewee with 6 Years of Experience

# 9. Policy Issue Deep Dive: 'Account Checks'

Interviewees reported that platforms have processes to make sure that the advertiser is who they say they are, has legitimate payments, and follows the ad standards of the platform. Platforms conduct **identity** verification, **business** verification, and **payment** verification.

## 9.1. How Do Platforms Prevent Repeat Offenders?

Despite these account checks, in general, interviewees noted that platform practices are not extensive or comprehensively applied, as buying ads is often designed to be as "low friction" as possible.

Platforms rely on operational and technical signals rather than identification methodologies. Platforms may rely on matching IP, device ID, geolocation, and other signals, but must combine multiple indicators rather than verifying identification.

Unless it is extremely obvious that a new account is a malicious prior actor, a platform will most likely not proactively remove the account.

*"When it comes to ads... It's [not usually] the case that an ad is approved to run, and more [so] that an ad is not disapproved."*

- Interviewee with 6 Years of Ad Fraud Experience

## 9.2. How Do Platforms Verify Business Information?

Interviewees mentioned a number of account checks that platforms will pursue. Platforms will request and verify an accurate business identification number, as it corresponds to the country of origin for the business. Platforms may also purchase databases of known businesses and match that way. Finally, platforms make sure payment information is accurate and matches what was given.

In general, most platforms are split between **content/policy-based compliance** and **account/behavioral-based compliance**. Platforms typically prioritize content-based compliance over account-based compliance.

*Table 13. When Do Platforms Require Account Checks*

| | |
|---|---|
| Often Requires Verification | • Insurance products<br>• Mortgages<br>• Loans (long and short-term)<br>• Investment products and opportunities<br>• Credit card applications |
| Sometimes Requires Verification | • Brand ads for banks and insurance companies<br>• News articles related to financial products or services<br>• Information about financial education<br>• Education, training, or skill-building to apply for loans or manage them<br>• Educational advertisements about a financial service or product that does not provide the ability to obtain or connect with that product or service |
| Sometimes Requires Verification *After Violating Policies* | • Accounts post violating content in ads<br>• Accounts post ads in high-risk industries (financial)<br>• Accounts use features that are involved in the misuse of the platform<br>• Accounts post sensitive ads (Involving brand names) |

Information collected in verification includes: email, legal business name, phone number, website, business license or local regulator license and articles of incorporation.

## 9.3. What Countries Enforce Account Checks

Interviewees noted that the following countries implement account checks for financial advertising:

**Australia**

Financial advertisers must hold an Australian Financial Services Licence (AFSL). These checks have been strengthened in recent years, largely in response to significant consumer losses caused by *"celeb-bait" deepfake financial ads* and other scam activity.[18]

**European Union**

Under the Digital Services Act (DSA), platforms are required to collect and verify information from financial advertisers, including two key fields:

- Beneficiary: the full name of the individual or organization on whose behalf the advert is presented
- Payor: the full legal name of the entity that paid for the advert

**India**

Financial advertising verification is overseen by the Securities and Exchange Board of India (SEBI).[19]

**Taiwan**

Account Checks are mandated under the Fraud Crime Hazard Prevention Act.[20] Taiwan has fined platforms in the past for not removing harmful ads.[21]

**United Kingdom**

In the UK, financial advertisements may generally only be issued by firms authorized by the Financial Conduct Authority (FCA), or by unauthorized firms where the content has been approved by an FCA authorized firm with the appropriate permission, or where a statutory exemption applies. Regulatory and verification mechanisms are used to restrict the promotion of regulated financial products and services to authorized, approved, or exempt communications.

---

[18] The Guardian, 2024. More than 9,000 scam Facebook pages deleted after Australians lose $43.4m to celebrity deepfakes

[19] Securities and Exchange Board of India, SEBI

[20] Taiwanese Government, 2024. Fraud Crime Hazard Prevention Act

[21] Taipei Times, 2025. Meta fined NT$2.5m over delayed fake ad removal

### Additional Verification Steps and Challenges

Interviewees noted how verification between different digital assets may be useful or necessary. For example, a platform could require verification of domain ownership through HTML files or tags, or could also require that email addresses for company or agency accounts match the domain. Additionally, it could require that domain registration contain public information of ownership. More basic personal ID verification is also an option, as validating people's identities can be more straightforward than validating businesses. These steps would all increase the frustration of bad actors while, in some cases, also increasing the security of ad accounts. However, interviewees noted that while verification can frustrate repeat bad actors, it is not a perfect form of deterrence.

Verification is also challenging, it was noted, because advertising is international, and not every country has the same standards around business registration. Bad actors can seek to use countries with less comprehensive standards to create a legitimate-looking business and begin posting ads. Some companies even exist to act as intermediaries between businesses and the platforms.

## 9.4. Account Checks: External vs. Internal Systems

### External

Interviewees indicated that if the data needed for an account check isn't in a report or audit provided by an SSP or vendor, platforms using external ad systems aren't guaranteed to get it. When a platform lacks access to behavioral signals, such as the business email associated with an account, it has fewer indicators to verify customer authenticity. Some vendors offer a full account check solution. Some offer standards around transparency of the buyer in the "Demand Chain," including features like "buyers.json," a list of all the advertisement buyers on the platform.

Interviewees also indicated that it is challenging for external ad systems to gather enough information about advertiser identity, increasing the likelihood that platforms with external ad systems will hire an account check specialist vendor solution. The vendor can verify business information, but the external vendor is also susceptible to fraudulent actors. Sometimes fraudsters target platforms that use external vendors, as there may be more opportunities to avoid suspension and action.

### Internal

For platforms using internal ad systems, interviewees noted that all data is held in-house and the internal teams are able to make reports, audits, and control the entire internal flow, from policy to operations to engineering and back. This centralized control allows platforms to implement more comprehensive authorization processes tailored to their desired standards or regulatory obligations.