Ofcom
making communications work for everyone

ACE Powered by vivace

ACE

Accelerated Capability Environment

User-Generated Content-Enabled Frauds and Scams

March 2023

A snapshot of platform responses against user-generated content-enabled frauds and scams

CONTENTS

1 FOREWORD 2

2 INTRODUCTION 3

3 USER-GENERATED CONTENT-ENABLED FRAUDS AND SCAMS 4

3.1 TYPES OF FRAUDS 4

3.2 FRAUD INDICATORS 5

3.3 RESPONDING TO BAD ACTORS 6

3.4 TECHNOLOGY AND COSTS 7

3.5 USE OF EXTERNAL SOURCES OF INFORMATION 9

3.6 INTELLIGENCE-SHARING INITIATIVES 10

4 CONCLUSION 11

5 ABOUT ACE 12

1 FOREWORD

Ofcom is the UK's converged communications regulator. We oversee sectors including telecommunications, post broadcast TV and radio. We regulate online video services established in the UK, including on-demand programme services (ODPS) and video-sharing platforms (VSPs). We will also take on new functions in respect of online safety under the Online Safety Bill.

In recent years a wide-ranging, global debate has emerged around the risks faced by internet users, with a specific focus on how users can be better protected from harmful content. Of this harmful content, one of the most common risks faced by internet users is online fraud. To help us broaden our understanding of this harm, Ofcom has commissioned ACE (Accelerated Capability Environment) to produce this report as a contribution to our evidence base.

2 INTRODUCTION

In the past few years, public concerns have grown about the prevalence of online frauds and scams being perpetrated on the social networking and marketplace platforms that we use daily. These online services themselves are under pressure to do more to mitigate against the threat of bad actors and suspicious content on their platforms, while the UK government is responding via the Online Safety Bill, which is expected to receive Royal Assent later this year.

The broader context for this short report is Ofcom's appointment as the Online Safety Regulator and its need to oversee the development of a new framework to ensure online platforms have appropriate systems and processes in place to improve the safety of their users.

This short paper aims to provide a snapshot of the ways in which platforms are seeking to mitigate the risks associated with user-generated content (UGC)-enabled frauds and scams. Such frauds and scams are enabled by user-to-user interactions on platforms via, for example, public posts that other users can see and respond to or direct messaging between users of a platform. This type of content is distinct from paid-for advertising where users or businesses pay for services to promote their content. Instead, UGC-enabled fraud occurs as users share or disseminate content on a platform without any payment being involved.

The research for the report was conducted through interviews with representatives from a targeted sample of platforms or people involved in anti-fraud mitigations in different fields, such as the online banking sector. This report summarises the findings in a way that preserves the anonymity of participating respondents. The views expressed in the report are the views of the interviewees and we do not therefore take them to be comprehensively representative of industry or wider stakeholder perspectives. They should also not be taken to represent Ofcom's views or any policy positions.

3 USER-GENERATED CONTENT-ENABLED FRAUDS AND SCAMS

3.1 Types of frauds

The UGC-enabled fraud that platform providers are working to mitigate falls into four broad categories, namely investment, pension or 'get rich quick' scams; romance or dating scams; purchase scams; and identity- and inauthentic activity-related fraud. The challenge with each is that it often blurs into other types of fraud and is therefore difficult to pin down with a single policy or approach.

The majority of platforms interviewed for this research have experience of scammers using investment, pension or get rich quick scams to target their users. This is where fraudsters present themselves as a trustworthy institution or advisor to pressure users to invest money, or by luring them with returns that are too good to be true. Some platforms refer to this type of fraud as "spam" because it usually involves high volumes of content being sent across the platform on topics such as cryptocurrency, loans or rapid weight loss.

Online marketplaces may also monitor seller behaviour for **purchase scams** (where products such as appliances, concert tickets, devices or pets are advertised and sold but are not received after the order or bank transfer has been made by the buyer), and for counterfeit goods scams such as fake designer goods, pirated copies of DVDs and computer games where buyers cannot check if the products are genuine until after the item has been delivered.

The main challenge for online platforms when it comes to these types of frauds is that, in many cases, the payment happens off platform. So while the online services may inadvertently enable investment, pension, get rich quick and purchase scams, platform providers said the point at which a fraudulent payment transaction occurs usually takes place outside their control.

Multiple industry respondents suggested that identity fraud presents a significant challenge on their platforms. This is where fraudsters pretend to be a user by accessing information about that user's identity (e.g. name, date of birth, current or previous addresses) and use it to obtain goods or services without permission.

Industry respondents said they are incentivised to focus on identity fraud for two main reasons. Firstly, because they can leverage established tools and third-party services, which can help assist them with mitigating this problem effectively. Secondly, because tackling identity fraud helps reduce the number of bad actors on their platforms, which in turn limits the scope for these same actors to perpetrate other types of fraudulent activity.

Romance and dating scams, which were noted as being prevalent on dating and social networking sites, happen when fraudsters pretend to be someone else or lie to gain a user's affection and trust, eventually asking for the user's money or financial information to purchase goods and services. Measures to remove bad actors are a high priority for dating platforms to safeguard their brand reputation.

The **authenticity of the content** itself is increasingly an area of importance in its own right. For platforms where the community value lies primarily with the integrity and quality of the user experience (based on the quality of advice, reviews or answers, for example), inauthentic content generated by fake accounts, software bots or state-sponsored bad actors strikes at the heart of their business models. While inauthentic account activity may not reach the threshold required for law enforcement intervention, it is taken very seriously indeed by platforms given its potential to negatively impact and degrade user experience.

3.2 Fraud indicators

Several platforms operate a two-tier approach to capturing fraud indicators. There is a bottom-up system of notification for suspicious content that begins with users themselves reporting bad content, along with teams of moderators monitoring content for inappropriate activity. There is also a top-down approach, which relies on the use of software to monitor the platform for unusual patterns of behaviour as well as banned actors and content.

Bottom-up fraud indicators. This primarily relies on self-reporting from users and moderators, and is reactive in nature. For UGC, most platforms expect one-off or individual incident content transgressions to be flagged by users. In this way, the platform's user community operates as the early-warning system for spotting suspicious content, an approach supported via platform-provided guidelines and education.

Above the users themselves, the next level of bottom-up fraud indication comes from content moderators who may have access to platform-provided tools to enable them to spot banned actors and content. Moderators also play a large role in refining the machine learning (ML) algorithms used to automatically monitor the platform from the top down. They do this by helping to build up categories of classified and labelled images and content, which are then used to update the rules-based monitoring systems or to train the ML algorithms.

Top-down fraud indicators. This approach tends to be centrally managed and proactive in manner as the platform fraud specialists use rules-based systems, ML models and neural networks to monitor keywords, images and unusual user behaviour. In the case of keywords, for example, platforms look for requests for money or services, surveys being sent out to get identity information, or accounts posing as financial institutions.

Because of the scale and volume of organic content on many platforms, one of the most effective indicators of fraud is to scan the platform for certain behaviours and patterns of activity associated with fraudulent activity, rather than focusing on the content itself.

While human moderation alone may struggle to spot the indicators of fraud at scale, it is often central to ensuring that monitoring by technology is relevant and accurate. Consequently, platforms have reported that they feel they need a robust level of human moderation as well as artificial intelligence (AI)-based tools. It is the human moderation component that is often the most expensive for platforms to provide, and for this reason some see human moderation as an outsourcing opportunity.

3.3 Responding to bad actors

Most platforms use centralised software tools to monitor for the presence of bad actors. This may:

- **Generate automatic warnings to users that something is not right and that they should consider stopping the interaction**
- **Generate nudges to potential bad actors that their actions are not appropriate**

Human-led detection via reporting channels is an important aspect of responding to bad actors. For this reason, many platforms proactively issue alerts and warnings, reminding users about their conduct on the platform and their duty to report inappropriate behaviour. In addition, platforms often produce online guidelines intended to establish the boundaries of behaviour while using their services.

Some supplement this with an internal community-based, user-generated scoring system that rates individuals based on their past interactions. If users dip below a given rating threshold, they are removed.

Ultimately, however, it is usually up to human agents to make the decision as to what happens next – whether an account is to be blocked, content taken down, or external authorities alerted. This tends to rely on the human resources available to the platform because it requires manual as opposed to automated review processes.

The response towards a bad actor typically depends on each platform's own scoring system. This provides a weighting of the risk factors for suspicious behaviour and determines how the team acts towards bad actors. Each platform's model for scoring bad actors is considered proprietary and is rarely publicly divulged.

In cases where the value proposition lies with the community interaction or networking aspect rather than with commercial transactions, many platforms err on the side of giving potential bad actors the benefit of the doubt before acting decisively to close them down. This is primarily driven by a concern that trust will be broken among users if the platform is seen to be acting in an unwarrantedly punitive manner. Consequently, it is common for the behaviour of a flagged bad actor to be monitored for a while and for a warning to be issued before any further action is taken. However, where fraudulent activity is confirmed, accounts are usually blocked swiftly.

Industry respondents claimed the perception platforms are slow to react to bad actors is often driven by:

- **A hesitancy to be seen as heavy-handed as this may create a reputational problem among users**
- **The fact that a response relies on human intervention, which is constrained by cost and resourcing**

3.4 Technology and costs

While most platforms use external third-party products or services in areas such as account verification, many believe that internally developed automated detection tools are the best way of mitigating against the risk of frauds and scams. This is primarily because frauds and scams are often tailored to specific online environments, meaning that each platform must tackle fraud in a unique way.

Technology for verifying account creation

Requesting two-factor authentication (email and phone number) in order to set up an account is standard practice for online marketplace platforms as well as recruitment platforms. However, this is not the case for many dating and social networking sites – often, just an email address is required to set up an account, with IP scanning used to check if an email address has previously been banned.

Dating and social networking platforms are among those making use of technology to verify a person's age and ensure that the face being scanned is live and not a photo.

Technology for monitoring posted content

For all platforms, the monitoring of posted content is addressed by a combination of AI tools (that may either be rules-based or use ML models designed to identify patterns of suspicious behaviour) as well as human moderation.

Many industry respondents suggested that rules-based approaches are often much less flexible than ML techniques, adding that they can become very complex over time as the hundreds of rules accumulate, slowing down platform responses to suspicious content.

Meanwhile, ML anti-fraud technologies can adapt more quickly as new data becomes available because they are better at identifying patterns of behaviour based on multiple signals.

Technology for user-to-user communication

For direct user-to-user messaging, many platforms use in-house AI software that picks up keywords and phrases being monitored. However, respondents from online marketplace and recruitment platforms suggested the main challenge they face is that direct user-to-user communication occurs on other social networking platforms beyond the reach of their trust and safety measures.

Costs

It is difficult to establish costs for user-to-user anti-fraud operations. This is not simply a question of commercial sensitivity, but rather that many platform teams often cannot conceptualise how to answer the question. This is because fraud is diverse, with many different platform teams involved in tackling it, and the technologies involved may be used across multiple areas of harm. Another issue is that security technology will have been built into the platform from the beginning and is not a cost component that is broken out.

In addition, platforms often view spend to combat fraud/cybercrime as simply the cost of doing business, meaning fewer concerns are raised than might be expected, even among secondary platforms. **However, while platforms said they regard tackling fraud as an important priority because of the negative impact it has on user experience and perceptions of platform safety/integrity, cost will always be a restraining factor to some degree**.

Protecting vulnerable individuals

Beyond age assurance/estimation and verification systems to identify child users, industry respondents said they do not monitor a wider set of user vulnerability characteristics. Many believe that to identify users in this way would be intrusive as well as unmerited because the platforms have ‘safety by design’ built in that offers protection to all users, even to the extent that it blocks all content from certain countries. The premise for safety by design is that everyone is treated as potentially vulnerable. Typically, it is executed by inbuilt smart filtering software on the platforms, which automatically prompts alerts that are pushed to users when scams are suspected.

Many platforms cite data privacy concerns as the reason why they do not seek data about more vulnerable users on their platforms. However, others are less hesitant about asking for additional information about users via opt-in data collection if it provides revenue stream potential. An example of this is job recruitment platforms, where establishing the (anonymised) diversity of the recruitment pool offers useful value to recruiters.

Some platforms have considered offering everyone the added protection of online security checks but decided against it given the danger of giving users a false sense of security. It also adds costs as well as more friction to the account creation process.

Despite not having internal policies for tracking potentially vulnerable users (beyond age-related child identification systems), many platforms recognise that some cohorts of users may be more vulnerable than others. Older users, for example, especially new retirees, are perceived as being less digitally literate and therefore as a more vulnerable online group.

Several US-headquartered platforms specifically monitor for incoming investment scams targeting US veterans, deploying a similar tactic to UK banks that puts extra flags on accounts held by individuals in their sixties as they are likely to draw down pension funds and be targeted for investment scams.

3.5 Use of external sources of information

All platforms participating in the study reported working with external law enforcement agencies when requested. However, some find this relationship frustrating because it is always reactive – the platform will only be looped in after the fraud has been committed. Platforms said they proactively share data and collaborate with law enforcement agencies in the case of significant threats such as terrorist alerts.

Use of external sources of information is far more prevalent among larger platforms than it is among their smaller counterparts. Dating and adult content platforms appear to use more paid-for external sources of information than others because of the need to keep known sex offenders off their platforms. For marketplace and online review platforms, application programming interfaces (APIs) connecting these platforms to information from financial regulators such as the Financial Conduct Authority and trading standards associations are important.

Beyond these sources of external information, platforms also use services from technology vendors. These include Google Safe Browsing APIs, Microsoft PhotoDNA, telecoms intelligence from Twilio, domain tools to provide information on suspicious domains, Mailgun for email validation, and MaxMind for geolocation fraud detection of IP addresses[1].

Effectiveness is measured by the number of false negatives and positives these technologies generate on the platform. A false positive is where an anti-fraud system incorrectly flags transactions as fraudulent, while a false negative is where an anti-fraud system misses fraudulent activity. Although platforms want to minimise false negatives, false positives create a lot of friction for users, and for online marketplaces this may lead to lost revenue while a transaction is halted because of an investigation.

In order to improve anti-fraud systems, there is an appetite for access to more external information among many of the platforms, as long as the information is timely, easy to access via an open API, and free or very low cost to use. In particular, there appears to be general agreement among platforms participating in this study that **they would like to be able to see information held by other online platforms about bad actors**.

One suggestion is that email providers could do more to provide information about compromised accounts. Another suggestion is for Apple and Google, as the primary app store providers, to play a bigger role, using their own internal data to block activities by bad actors.

[1] This involves scammers manipulating or falsifying GPS data and IP address data to conceal their actual location.

3.6 Intelligence-sharing initiatives

Most platforms say they are involved in intelligence-sharing initiatives. The more established of these are typically not initiatives with each other, despite this being cited as being required to stop bad actors. Rather, they are with non-governmental organisations, law enforcement agencies and banks. However, there is a burgeoning number of initiatives between platforms to share information. In the UK, among larger primary platforms, these include the Online Fraud Steering Group, co-chaired by Tech UK, as well as membership of Stop Scams UK. In the USA, the Federal Trade Commission is working with some platforms around data sharing to safeguard the authenticity of user reviews. Dating and relationship platforms discuss cross-platform issues via the UK-headquartered Online Dating Association. Beyond this, secondary platforms headquartered in the USA say they informally and confidentially share information about scams.

Meanwhile, data privacy concerns and General Data Protection Regulation (GDPR) obligations are frequently cited as the main barrier to sharing intelligence between platforms. Commercial risk, in the context of ceding competitive advantage, is mentioned less frequently. More significantly, smaller platforms struggle to reconcile the effort of intelligence sharing with the business value it generates.

From the platforms' perspective, the challenge GDPR poses is that platforms are not confident about what data they can share, nor how to go about sharing it, and they are looking for regulatory guidance in this area.

Platforms, especially smaller ones, may also not see the relevance of sharing information based on a belief that they are too specialised to gain much value from knowledge of frauds and scams on other platforms. These platforms would require persuasive cost–benefit analysis before participating in intelligence-sharing initiatives with their peers.

4 CONCLUSION

Despite reservations about the approach, platforms consulted as part of this study typically understand that they could do more to mitigate the risks associated with UGC-enabled fraud, and many are interested in exploring new approaches.

However, there are still several barriers. For example, fraud involving financial transactions is problematic for online platforms to deal with because the fraudulent act often happens off platform, and platforms may only be looped in after the fact.

Platforms also vary in their response to account verification designed to address identity fraud. Some, such as online marketplaces, mandate two-factor authentication, others suggest it, and a few offer the option to take part in a Know Your Customer process. Some simply ask for an email address. Typically, however, there is a belief that platform users would find more account verification measures intrusive.

Platforms contributing to this study also highlighted the importance of greater cross platform and cross-industry collaboration to combat fraud. **Meanwhile, when it comes to UGC-enabled fraud, there was agreement among participants that cross-platform identification of bad actors and data about suspicious patterns of behaviour is the best way forward, rather than focusing on the content itself**.

In order to overcome their fears about breaching GDPR, many platforms would welcome regulatory guidance and best practices about sharing data to enable this collaboration to happen. A framework to provide guidance on what information can and cannot be shared, and how that data should be shared, would be gratefully received. Interviews with smaller UK platforms suggest they rarely use external sources of information because of the cost this would involve and because they are less likely to be included in the formal and informal partnerships that larger platforms have set up with banks and financial service sector players. These smaller platforms said they would be interested in a simple, free or low-cost way to access shared information about fraudulent patterns of behaviour and bad actors.

5 ABOUT ACE

This report was produced on behalf of Ofcom by the Futures & Insight team at the Accelerated Capability Environment (ACE), a Home Office unit that takes a highly innovative and disruptive approach to solving technology and data problems facing public sector agencies.

ACE's Futures & Insight service covers a broad range of market intelligence, horizon scanning and foresight, and currently serves several government customers including Ofcom. Please contact ace@homeoffice.gov.uk for more information.