Ofcom
making communications work
for everyone

ACE | Powered by vivace

ACE

Accelerated Capability Environment

Illegal Harms

May 2023

Mitigating illegal harms: a snapshot

ACE | Powered by vivace

CONTENTS

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 they can be better protected from harmful content. This is reflected in the number of priority illegal harms added to the Online Safety Bill. To help us broaden our understanding of some of these harms, 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, illegal drugs and weapons sales; offers to facilitate illegal immigration; and the promotion of suicide and self-harm have increasingly developed an online component. This has led the UK government to respond via the Online Safety Bill, which is expected to receive Royal Assent later this year.

Ofcom's appointment as the Online Safety Regulator means that it needs 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 measures being taken to detect or prevent the sale or promotion of these illegal harms, as well as additional insights into the technologies used, the costs of implementation and the barriers to further mitigation efforts.

While there was some desk research involved, it was based mainly on primary research.
Most importantly, data was provided by platforms through a combination of interviews and written submissions. In addition, interviews were conducted with subject matter experts (SMEs) in illegal drugs and weapons, illegal immigration and suicide/self-harm.

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 ILLEGAL HARMS MITIGATIONS

3.1 Trading of goods and advertising: weapons/illegal drugs; illegal immigration

3.1.1 Detection and prevention measures

Platforms contacted for this study have put in place policies forbidding the sale of illegal drugs and weapons as well as the promotion of illegal immigration, backed up by a range of automatic and manual detection methods, including user reporting.

For the platforms interviewed, community guidelines form the basis of their detection and prevention measures, alongside national laws, and both the sale of illegal drugs and weapons and the promotion of illegal immigration are typically banned (although the latter may not always be specified). Meanwhile, efforts are also made to ensure that educational or artistic content, for example that related to small boats crossing the channel, does not come within scope of these measures.

Policies must also be enforced, and this is achieved through a combination of proactive detection and user reporting, respondents said. Machine learning (ML)-based automated technologies are typically used to moderate this type of content, supported by human reviewers. It is apparent that ensuring users are aware of these policies is also very important, and many platforms have educational measures in place to promote them. For some, user reporting is the most effective mitigation measure, supplemented by proactive detection that then goes to human moderators for review.

Platforms themselves regard the measures they have in place as broadly effective. However, others involved in tackling these harms sometimes see moderation as under-resourced and reactive, resulting in detection levels that are lower than they could be. Keyword-activated tools can be ineffective against the use of argot[1], emojis and images, for example, while the algorithms used to detect illegal content are only as good as those who train them and the data they use. There are other weaknesses too, including messaging platforms' inability to automatically detect violative content in one-to-one communication and private groups because of end-to-end encryption.

[1] The jargon or slang of a particular group or class.

3.1.2 Barriers to tackling harms

Barriers cited include content ephemerality, a widespread lack of awareness about age verification requirements and problems establishing context.

Among the barriers faced by e-commerce platforms, respondents said, are the fact that many sellers remain unaware of the age verification measures legally required in the case of knife sales, making enforcement significantly harder in this area than in others. It can also be challenging for platforms to determine whether content violates their policies because of problems of definition, again especially in the case of knives.

Platforms also told us that they were affected by the fact that UK laws differ from those elsewhere, an issue that is particularly acute when it comes to weapon sales. To take account of this, some platforms reported banning the sale of all knives (except cutlery) in the UK, while also hiding overseas listings of knives from UK buyers.

There are also problems with the high levels of nuance and ambiguity involved in making accurate determinations; it can be especially challenging to automatically identify weapons and narcotics content because there is a great deal of context dependency involved here. Keeping up with trends and changes can be difficult too, according to interviewees. Criminals may use argot, emojis and images to avoid detection, and moderation tools must be able to adapt to these identifiers and the way in which they change over time. Meanwhile, some respondents said the main barrier is simply the ephemerality of content on their platforms, which makes it much harder to detect.

3.1.3 Technologies

Platforms are using a combination of filter algorithms and ML-based technology to detect illegal harms, alongside human moderators who review flagged or potentially problematic posts.

Platforms generally find it easy to set up a filter algorithm based on keywords, and computer vision can be used to identify contraband and weapons when images are uploaded. However, algorithms still have problems differentiating between a kitchen knife and a zombie knife, for example, experts said, making it difficult to automatically flag harmful content without creating huge numbers of false positives.

Livestream video is also a problem because of a lack of technological solutions suited to this task. Accordingly, interviewees said technological- and human-based processes are used in combination, with automation and operators creating a kind of feedback loop where humans refresh the learning data.

Platforms are mainly using in-house technology solutions to tackle illegal harms, some of which have come from acquisition. While details of the costs associated with implementing these technologies are hard to come by, interviews indicate that expenditure is high in this area and often runs into many millions of dollars for each platform.

Yet although this technology can be quite effective, more accurate artificial intelligence systems are needed, interviewees said. In addition, there is a need for tools that can be applied to livestream video-focused technology and end-to-end encrypted content without compromising user privacy.

3.1.4 External information and information sharing

Several external information sources are used, particularly those provided by government agencies, but sharing between platforms remains a relative weak point, respondents said.

Some platforms make most use of data provided by government agencies to tackle illegal harms. When it comes to illegal drugs, for example, they may rely on advice from the UK's Medicines and Healthcare products Regulatory Agency[2] or the United States Food and Drug Administration to provide information about unsafe products. Reports from other competent authorities and trusted flaggers may also be used to improve platforms' knowledge of illegal harms. However, interviews indicate that the level of sharing and collaboration is still quite minimal, perhaps because of the potential for negative publicity that such external engagement could cause for services seen to be hosting illegal content.

Platforms may also share intelligence and collaborate with law enforcement organisations such as the National Crime Agency (NCA) to help prevent illegal content appearing on their platforms, but this too is somewhat limited, and intelligence held by European law enforcement agencies on people smugglers, for example, is currently underused, experts said. More broadly, they believe there is scope for much greater levels of collaboration between platforms and European law enforcement agencies such as Europol, Interpol and the NCA.

Knowledge sharing between platforms themselves is particularly low, respondents said, with no drugs and weapons equivalent of the Global Internet Forum to Counter Terrorism's (GIFCT) hash-sharing database, for example[3]. Labelling of content could be improved too if there was more sharing between platforms, while exchanging information about the latest use of argot and emojis would surely help flag more illegal content.

[2] https://www.gov.uk/government/organisations/medicines-and-healthcare-products-regulatory-agency
[3] The GIFCT was established by Facebook, Microsoft, Twitter and YouTube to help prevent terrorists and violent extremists from exploiting digital platforms.

3.2 Harmful activities: illegal suicide/self-harm

3.2.1 Detection and prevention measures

Policy-based technology and human-based detection systems are being used to prevent the promotion of self-harm and suicide, according to interviewees, bolstered by tools such as automated messages and signposts.

Platforms typically operate policies that prohibit the promotion or encouragement of self-harm and suicide, banning any content that may endanger a user's life or encourage negative physical behaviour such as eating disorders. They may also prohibit users from asking how they can harm themselves. However, as such content is often created by people who themselves are experiencing mental health problems, platforms also aim to provide help and resources where appropriate.

As with all other forms of illegal harms, the starting point is a coherent policy for the given harm, with content moderation conducted via a combination of technology and paid and volunteer community moderators. Among the latter may be trained crisis counsellors who are highly familiar with the subject matter. Policies are then enforced via proactive activities as well as reactive reviews of user reports. Reviews can be particularly useful, respondents said, because self-harm-related content is so context-dependent and people may use language in deliberately deceptive ways. Despite these enforcement measures, discussions that focus on research, advocacy and educational issues related to self-harm or suicide are usually permitted.

Platforms may put in place other mitigation measures too, including signposts that appear on screen before users gain access to search results and automated messages that interact with users who appear to be at risk. These efforts can be quite effective, the platforms believe, leading in some cases to a meaningful reduction in self-harm content.

3.2.2 Distinguishing between violative and non-violative content

Platforms reported distinguishing between violative and non-violative content based primarily on context, with human decision making often prioritised over automatic responses.

There is often a fine line between expressing self-harm or suicidal intent and encouraging others to engage in such activity, respondents said. Accordingly, reports on self-harm- and suicide-related topics are typically reviewed by moderators who have been given clear criteria on which to base their decisions. They might make a distinction between violative and non-violative content by utilising contextual information and internal policies, for example. Specific language or keywords can also be used to help make decisions, respondents said, with ongoing training of reviewers based on actual examples to help distinguish between categories.

However, there are a great many grey areas still. Given the highly context-dependent nature of differentiating between violative and non-violative content, platforms often struggle, and this is exacerbated by the fact that while discussion of suicide may be permitted in some contexts (such as within specialist channels set up to prevent it), it remains strictly forbidden in others. Although platforms understand the need for users to discuss self-harm- and suicide-related topics, all platforms interviewed draw the line at any content that glorifies or promotes these activities. With this in mind, platforms prohibit users from sharing information about specific strategies or methods related to self-harm or suicide, even when they may be sharing their own experiences.

3.2.3 Barriers to tackling harms

Barriers include difficulties interpreting nuance, hybrid content that may be both soliciting a methodology and calling for help, gaps in livestream detection technology, and the scale of the problem.

For some, the main barrier is the lack of livestream-focused moderation technology (able to tackle, for example, illegal content within live videos), while for others it is the high level of nuance involved and the fact that automating or standardising decision making of this kind is far from easy. There may also be difficulties associated with hybrid content that is soliciting a methodology for self-harm or suicide while also calling for help.

In addition, although removing self-harm- and suicide-related content is clearly very important to the platforms, from their point of view it is just one of the many very high priorities they must deal with, some of which – child sexual abuse material (CSAM) and terrorism in particular – carry both greater reputational risk and the prospect of fines.

3.2.4 Technologies

ML systems and keyword and language detection systems are among the technologies being used to detect and mitigate violative content.

Platforms said they typically use ML reporting systems to surface violative content. These run constantly in the background to automatically identify banned content, whether or not a human has reported it. Some of the platforms interviewed told us that anything considered to fall within a grey area is then reported to a human moderator for review. As with other forms of illegal harm, language processing and keyword detection technologies are also used, including those focusing on potential suicidal hashtags and curated keywords. Meanwhile, moderators also perform manual platform sweeps. Given the sensitivity and nuance in this area, some have rejected algorithmic approaches entirely and instead rely on human review to understand context.

Because of the specific requirements of each platform and the need to understand nuance, most of this technology in use has either been developed in-house or acquired, interviewees said. Cost details are not easy to obtain, but platforms typically describe expenditure levels in this area as high. However, as these systems are not restricted solely to self-harm and suicide, costs can be spread across several mitigation areas.

While platforms mainly regard their technology as effective, the current high rate of false positives is a matter of concern to the platforms, along with the technology's substantial cost and difficulties gaining inputs to train ML models because of a lack of suicide and self-harm content. Problems distinguishing between harmful and supportive content also continue to impede progress somewhat.

3.2.5 External information and information sharing

Platforms are selectively sharing information with charities, other specialist organisations, SMEs and industry peers, although these practices are far from widespread.

Some platforms engage with charities like Crisis Text Line and Samaritans to share information and ensure their policies are as effective as possible. They may also work with law enforcement agencies where appropriate. Platforms may also collaborate with SMEs to ensure that those engaged in self-harm or experiencing suicidal thoughts are able to share their personal experiences appropriately.

Industry conversations with platform peers are another means of sharing information, although some may only do this on an ad hoc basis – for example, at technology conferences – and, even in the best cases, these practices are far from widespread, according to interviewees. This may be due to each platform's unique characteristics, or at least their own perception of such, but the effect is to limit the sharing of valuable information, reducing the overall effectiveness of platforms' mitigation measures.

4 CONCLUSION

When it comes to the illegal harms covered in this report, platforms believe they have put in place comprehensive policies backed up by a range of automatic and manual detection methods. These are then enforced through a combination of proactive detection and user reporting.

However, platforms still face many barriers, including problems making determinations and a high rate of false positives. ML-based solutions can also be slow and reactive, as well as unsuited to sparse and weak signal recognition of the kind necessary in areas like illegal immigration, according to some respondents. Meanwhile, ephemeral content remains difficult to moderate, and there are other blind spots too as a result of end-to-end encryption on messaging platforms.

These interviews suggest that problems are being exacerbated by the current low level of information sharing both by and between platforms. More cross-platform collaborations like GIFCT would be highly beneficial for all concerned, according to experts consulted as part of this study, as would better engagement with some of the more innovative law enforcement initiatives[4].

And while the platforms have made great strides in recent years, some experts believe that the level of importance assigned to tackling illegal drugs, weapons and self-harm is still below that of CSAM, for example. The same level of effort made to tackle CSAM, these experts said, should now be applied to mitigating a wider range of illegal harms, ensuring the safety and security of everyone online.

[4] An example being COPKIT[4], a European Union-funded project by the Austrian Institute of Technology in Vienna, which applies natural language processing to posts on darknet forums.

ABOUT ACE

This report was produced on behalf of Ofcom by the Futures & Insight team at 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's service covers a broad range of market intelligence, horizon scanning and foresight, and currently serves several government customers as well as Ofcom. Please contact ace@homeoffice.gov.uk for more information.