Response to Ofcom Call for evidence: First phase of online safety regulation

Introduction

Samaritans is the UK and Ireland's largest suicide prevention charity. We respond to a call for help every ten seconds and, in 2021, Samaritans volunteers spent over one million hours supporting people who called us for help.

Over the last three years we have developed a hub of excellence in suicide prevention and the online environment with the aim of minimising access to harmful content and maximising opportunities for support. Our Online Excellence Programme includes industry guidelines for responding to self-harm and suicide content, an advisory service for sites and platforms offering advice on responding to selfharm and suicide content, a research programme exploring what makes self-harm and suicide content harmful and for whom, and a hub of resources helping people to stay safe online.

We warmly welcome the opportunity to respond to this consultation. The new online safety regulatory regime is a once-in-a-generation opportunity to create a suicide-safer internet and it is hugely significant that 'assisting suicide' has been defined as priority illegal content on the face of the Bill.

In order to be able to respond meaningfully to this consultation we have taken a broad interpretation of 'illegal suicide content' as there is not yet clarity on the practical parameters of this for the purposes of the online safety regime. We believe that some of the material we are concerned about may fall into the 'legal but harmful' category and are continuing to push for the strongest possible protections for people of all ages from all harmful suicide and self-harm content as the Online Safety Bill progresses through Parliament.

Can you provide any evidence relating to the presence or quantity of illegal content on user-touser and search services? We are particularly interested in evidence about how this might vary across different services or types of service, or across services with particular users, features or functionalities

CONFIDENTIAL RESPONSE

What can providers of online services do to enhance the clarity and accessibility of terms of service and public policy statements? Please submit evidence about what features make terms or policies clear and accessible.

Our industry guidelines recommendations around accessibility suggest that users should be provided with clear and accessible community guidelines about what content is allowed on the site. They should also be given step-by-step information including how to make a report and what action may be taken. This information should be clearly displayed to new users, and existing users should be regularly reminded, empowering them to report any content that concerns them.

Our recent research with people with lived experience of self-harm and suicidal thoughts indicates that whilst there it a good understanding of the purpose of community guidelines "to keep users safe" very few online users have seen or read them. Many participants said that they would only check out the community guidelines in response to having their own content removed by the platform. Overall, there were very low levels of awareness that community guidelines specifically relating to suicide and self-harm content existed. It is therefore of vital importance that platforms take additional steps to o make community guidelines more visible and accessible for users, so it is clear what content is and is not allowed on their site.

In our research, what resonated most with online users were messages that:

- adopted a human and friendly tone, that avoids authoritative or triggering language
- used simple and directive language without too much text and jargon -guidelines should be easy to navigate and for all users to understand.
- included specific examples of things that are allowed and prohibited. This could be a list of things that users can and can't do
- included clear guidelines on appeals and contact information for a person to speak to about the appeal, rather than an automated help system.
- included clear and straightforward information about what happens if users breach the guidelines.

What can providers of online services do to enhance the transparency, accessibility, ease of use and users' awareness of their reporting and complaints mechanisms? Please submit evidence about what features make user reporting and complaints systems effective, considering: • Reporting or complaints routes for registered users; • Reporting or complaints routes for nonregistered users; and • Reporting routes for children and adults

Our industry guidelines include suggestions for accessible reporting processes. For small sites this may be a dedicated email or reporting form. Larger sites may implement more sophisticated reporting functions, such as self-harm and suicide content specific reporting categories and trusted flagger functions, whereby credible organisations and users with a track record of making responsible and accurate reports are able to have their reports fast-tracked. Our research with people with lived experience showed that reporting categories used by platforms should also be ordered by priority, so self-harm and suicide should come higher in the list of options compared to categories like spam.

Language around reporting functions for self-harm and suicide should also be treated with caution and sensitivity. Quite often if someone is posting about suicide or self-harm, while it still might be harmful to other users, it is not done with malicious intent. Users may therefore avoid reporting concerning content because they worry about getting the poster in to trouble. The lack of self-harm and suicide specific reporting categories can also deter users from reporting content. For example, if the categories don't feel as relevant or are more associated with other types of online harms, such as 'abusive/offensive content'.

Reporting and complaint requirements in new online safety regulations should also uphold impartial appeals process or independent ombudsman provision. Platforms should be required to have user redress and victim support measures, policies, and systems in place, that can be easily located and utilised by users.

Could improvements be made to content moderation to deliver greater protection for users, without unduly restricting user activity? If so, what? Please provide relevant evidence explaining your response to this question. Please consider improvements in terms of user safety and user rights, as well as any relevant considerations around potential costs or cost drivers.

Our industry guidelines call on all sites and platforms to moderate user-generated content, ensuring that self-harm and suicide content policies are successfully implemented and that users are protected from harm and directed to support. Sites with low volumes of user-generated content may be able to rely on human moderation alone. This can be an effective way of detecting and responding to self-harm and suicide content as moderators can understand the nuance around

selfharm and suicide language, provide users with personalised responses, and quickly identify and react to emerging trends. But consideration should be given to the speed at which content can be identified and times of day and night when it is most likely to be posted. All sites implementing human moderation should ensure moderators are provided with high quality training and support.

Platforms hosting higher volumes of user-generated content should complement human moderation with artificial intelligence (AI) to prioritise user reports, flag potentially harmful content to be reviewed, and prevent harmful content from being uploaded. AI allows for the assessment and identification of harmful content at scale which can enable early detection and can prevent content from being widely shared. An example of this is self-harm scars, as while graphic images and open wounds should be censored, it is not always appropriate to censor or remove content where selfharm scars are visible.

How are sanctions or restrictions around access (including to both the service and to particular content) applied by providers of online services? Please provide evidence around the application and accuracy of sanctions/restrictions, and safeguards you consider should be in place to protect users' privacy and prevent unwarranted sanction

The nuanced area of online suicide content means that platforms and sites may need to first try to establish the intention behind posting such content. Even where content illegally encourages or assists suicide, this may not have been posted maliciously and could also have been posted by someone who is in vulnerable circumstances themselves.

Where sites and platforms need to remove illegal suicide content as far as possible this should be done using safe and empathetic approaches. Care should be taken to minimise any distress caused to the user, by ensuring the tone of the communication is sensitive and avoids negative language, and explains why the content has been removed, how to re-post safely and where to find support.

If a user repeatedly posts content that breaks community guidelines, companies may decide to pause their membership or close their account to protect other users. Companies should be mindful that this could withdraw a user's vital, and in some cases only, source of support. If pausing a membership, the user should be provided with an explanation of why their membership is being paused, signposts to support and information about how to appeal the decision. It is worth considering if it is possible to suspend parts of a user's account, i.e., the ability to post publicly,

whilst still allowing them access to old content they have posted and to directly message users they have existing relationships with.

Are there any functionalities or design features which evidence suggests can effectively prevent harm, and could or should be deployed more widely by industry? Please provide relevant evidence explaining your response to this question.

Our industry guidelines on managing user-generated suicide and self-harm content online as well as our research with people with lived experience highlights a number of suggestions around functionalities and design. These include:

- Ensuring site algorithms don't push harmful self-harm and suicide content towards users. For example, platforms that make suggestions based on previous browsing should disable this functionality for self-harm and suicide content.
- Ensuring that censored content does not appear as suggested content for users
- Blocking harmful site searches, such as those relating to methods of suicide, online suicide challenges and hoaxes, or searches for websites that are known to host harmful content.
- Autocomplete searches turned off for harmful searches such as those relating to methods of harm and associated equipment.
- Using age and sensitivity content warnings, to warn users that content may be distressing as it mentions self-harm or suicide.
- Embedding safety functions, allowing users to have more control over the content that they see. For example, by having more functions to block content by muting words, phrases and hashtags.

What age assurance and age verification technologies are available to platforms, and what is the impact and cost of using them? In particular, please provide evidence explaining: • how these technologies can be assessed for effectiveness or impact on users' safety; • how accurate these tools are in verifying the age of users, and effective in preventing children from accessing harmful content; • steps that can be taken to mitigate any risk of bias or exclusion that may result from age assurance and age verification tools; • the costs involved in implementing such technologies;

and • the safeguards necessary to ensure users' privacy and access to information is protected, and over restriction is avoided.

Can you identify factors which might indicate that a service is likely to attract child users?

In research that we commissioned from Swansea University (which has not yet been published), we found that whilst users generally support age verification and restrictions across social media and online platforms, these are easily bypassed by children. In a sample of over 5200 participants, over three quarters saw self-harm content online for the first time at age 14 years or younger, with nearly a fifth saying that they were 10 years or younger. It was highlighted by participants that date of birth alone is not sufficient as age verification as using a fake birthday was a simple way to get around this.

Recommendations from the research are that age verification tools, parental controls and censoring/ filtering of content are used alongside further steps of increased education around safer internet use in schools and for professionals and carers.

For purposes of transparency, what type of information is useful/not useful? Why? In particular, please consider: • Any evidence of public information positively or negatively affecting online user safety or behaviours, how this information is used, and by whom; • What information platforms should make available, considering frequency, format and intended audiences; • What information Ofcom should make available through its transparency report, considering frequency, format, intended audiences and potential use cases by external stakeholders; • The benefits and/ or drawbacks of standardised information and metrics; and • Any negative impacts or potential unintended consequences of publishing certain types of information, and how these may be mitigated

We believe that annual transparency reporting is vital for best practice and to hold platforms accountable for the action taken against harmful content on their platform.

Transparency reporting should be proportionate to the size of the platform but all platforms hosting user generated content should be subject to this requirement.

Transparency reports should include key information, such as:

- **Prevalence of self-harm and suicide content on the platform**
- **Average number of views of content that breaks community guidelines**
- **Mechanisms in place to detect and respond to self-harm and suicide content**
- **Action taken to content that breaks community guidelines** – including the percentage of reported content that is removed and average time taken to respond to user reports.
- **Resource allocated to responding to self-harm and suicide content.**

Publicly reporting high prevalence rates of self-harm and suicide content may inadvertently encourage vulnerable users to access specific platforms to seek out potentially harmful content. Platforms should consider how they can mitigate these risks (e.g., by reporting figures as rates per 10,000 views) and including signposting to support services within the transparency reports. Ofcom could also provide useful guidance on how to strike the right balance between meaningful transparency and reducing the risk of inadvertently drawing attention to harmful content.

Other than those in this document, are you aware of other measures available for mitigating risk and harm from illegal content? We would be interested in any evidence you can provide on their efficacy, in terms of reducing harm to users, cost and impact on user rights and user experience.

It is essential that companies work more collaboratively and establish ways to safely share insights on illegal content with one another, with Ofcom, and with professionals with relevant interests in order to promote excellence across the industry and better protect online users. This could include

alerting other platforms to new online trends or discussion of emerging methods of suicide and sharing insights on how to manage and respond to these effectively. There is a precedent for a multiagency alert system that includes Samaritans Online Harms Advisory Service to ensure safe approaches to emerging developments or incidents.

In cases where illegal self-harm and suicide related content has been shared, companies should also explore how they can promote helpful content to their users that encourages help seeking and directs users to appropriate support.