Your response

Volume 2: The causes and impacts of online harm

Ofcom's Register of Risks

Question 1:	
i)	Do you have any comments on Ofcom's assessment of the causes and impacts of online harms?

Response:

Our community has identified some key missing functionalities that may cause harms, they are as following:

- Monetization tools
  - Access to monetization tools for children
  - User-sold subscriptions and ad sales provide a major risk vector bc they introduce a direct financial benefit to scammers or hackers
- Private vs. public content
  - Outside of strictly E2EE

Our community has identified some key missing significant harms, they are as following:

- Sale of illegal substances
- Bullying (distinct from harassment)
- Surveillance
  - Difference between adult user and child user
- Extreme graphic imagery
- Most platforms do work to proactively identify harms and harmful content, which are generally tracked in the content policies in their terms of service. It is reasonable to ask for platforms to perform risk assessments for "all content that violates the policies of the platforms" or "any content platforms deem violating of the platforms policies"

ii)	Do you think we have missed anything important in our analysis? Please provide evidence to support your answer.

Response:

iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)

Response:

Question 2:	
i)	Do you have any views about our interpretation of the links between risk factors and different kinds of illegal harm? Please provide evidence to support your answer.
Response:	
ii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Volume 3: How should services assess the risk of online harms?

Governance and accountability

Question 3:
i) Do you agree with our proposals in relation to governance and accountability measures in the illegal content Codes of Practice?

Response:

There are some key issues around the way these proposals frame the issue of governance, namely:

- The framing of the responsibilities of someone tasked with managing safety risks as tracking increases in content puts the focus on finding and destroying content, not systemic approaches that could have an impact on reducing the production of, and/or incentives to create, harmful content.
- Content tracking also misses huge swaths of harms like fraud and some ephemeral interactions like in gaming or VR/AR
- This also codifies unweighted content prevalence as the metric of success, which isn't advisable especially as perspectives change on the most effective way to deal with harms, and prevalence of harmful content on platforms is not indicative of the platform's impact on dealing with root causes of harms or the scale of harms caused, such as total number of users exposed to harms.

The focus for strengthening platform governance should be instead on platform design, and building transparency and accountability for the ways in which it contributes to the dissemination of illegal content, and incentivizing better design choices. Platform governance goes beyond just who is designated in a particular role, but also includes what metrics the company uses to measure success, how it tests its systems, and how decisions are made about how different values (safety, growth, etc) are weighed.

On metrics

Metrics are a key way in which integrity and trust and safety can be better embedded into platform governance processes. When deciding whether or not to release a new feature or an update to a ranking and/or algorithmic system, metrics will often be core to the decision making process.

If the decision making process is dominated by platform growth and business interest metrics, and integrity and trust and safety metrics are not integrated until later, it will be difficult for risk mitigation measures to be properly incorporated to the launch of new features, products or changes.

Incorporating integrity and trust and safety metrics into platform decisions and governance can be a challenge. Typically, platforms first build out growth and business interest metrics before integrity and trust and safety. Growth and business metrics will typically cover platform usage, engagement, ad impressions and revenue. So typical metrics here would be daily/weekly/monthly

active users, user retention, time spent on platform, user engagement actions on the platform (likes, faves, shares, comments, views), number of ad impressions, total revenue from ads. All these metrics will largely be correlated with each other. The more total time spent on the platform, the more ads that will be viewed for example.

Integrity metrics will naturally be in tension with growth and engagement metrics. Harmful, violating content will generate engagement, and often will generate engagement more efficiently and effectively than non-harmful content. This means that any effort to reduce harmful content will in general decrease overall engagement on the platform.

Platforms are typically highly optimized to accomplish the company goals around growth, engagement, and business interests. This means that there are few options to further increase the total time spent on the platform, or the total number of ad impressions. When platforms are in a state of high engagement optimization, then any change to the platform will inevitably reduce engagement, because the platform is at an engagement maximum. To break out of this, platforms should not use engagement metrics to measure success. Options for metrics beyond engagement include quality focused metrics, user surveys of how much value they are getting from the platform, and/or only counting clearly good forms of engagement. Google Search and their quality metrics provide an alternative and demonstrate that trade-offs can be navigated. For example, for every A/B test Google Search runs on real people, Google Search eliminates 190 potential new features through their quality, non-engagement based assessments.

On AB tests

While under current industry practice, these tests are often evaluated against metrics that align with the business interests of the platform, in order to truly contribute to an assessment of risk, they should include integrity and safety focused metrics, such as the prevalence of harmful content, exposures to predicted harmful content, or user reports of harmful experiences.

Frequently, there will be tension between the direct business interests of the company, total engagement on the platform, and the integrity and safety of the platform, minimizing exposures to harmful content. Updates to the platform intended to increase the safety of the platform by reducing exposures to harmful content will typically have a negative impact on engagement metrics. Therefore, it is extremely important to understand how they make trade-offs between the business interests and driving up user engagement, and the integrity and safety of the platform.

One additional comment - to mimic Ofcom's stance on bookkeeping, there this section should include:

- Keeping record of when responsibilities for senior members is revised
- Keeping record of when a code of conduct is revised
- If the staffing and service's approach on compliance is revised

ii)	Do you think we have missed anything important in our analysis? Please provide evidence to support your answer.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)

Response:

Question 4:	
i)	Do you agree with the types of services that we propose the governance and accountability measures should apply to?
Response:	
ii)	Please explain your answer.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Question 5:	
i)	Are you aware of any additional evidence of the efficacy, costs and risks associated with a potential future measure to requiring services to have measures to mitigate and manage illegal content risks audited by an independent third-party?
Response:	
ii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Question 6:	
i)	Are you aware of any additional evidence of the efficacy, costs and risks associated with a potential future measure to tie remuneration for senior managers to positive online safety outcomes?
Response:	
ii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Service's risk assessment

Question 7:	
i)	Do you agree with our proposals?
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.

Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Specifically, we would also appreciate evidence from regulated services on the following:

Question 8:	
i)	Do you think the four-step risk assessment process and the Risk Profiles are useful models to help services navigate and comply with their wider obligations under the Act?

Response:

It is useful to provide platforms with some structure to indicate what the regulator is looking for in risk assessments, and in this regard the risk profiles and four steps are a useful starting point. However, there's not a 1 to 1 relationship between risk and solution, which makes it very tricky to report, so it must be clear that there is room for the platforms to discuss areas of overlapping risks and mitigations. Additionally, in this proposed process, there could be a significant lag between identifying the risk/implementing the fix and reporting it.

In our view, a comprehensive platform risk assessment, in particular one that focuses on how platforms are contributing to the spread of illegal or harmful content, should include the following pieces:

1. An overview of the risk area:

The overview of the risk area should demonstrate that the platform has a comprehensive understanding of the problem, bad actors or perpetrators which amplify the problem, potential victims of the problem, and how the problem impacts people and societies.

The overview of the risk area should include: description of the risk area, description of the harms that it can cause people and societies, external organizations and research that the company uses to understand the risk area, a description of any perpetrators and victims of the harms

2. An overview of the platform's systems and associated risks

Starting with a list and overview of key systems, similar to a systems description found in other industries and security system audits (SOC 2 technology audits in the USA), platforms should produce a systems description that outlines how their major technology systems contribute to key categories of risk. This will then lead to an assessment of how the risks manifest on the platform, including how the risks interact with every component of the platform's algorithmic systems. This is where Ofcom's risk profiles can be useful to help platforms understand the types of risks associated with certain features. We'd also point to projects like Focus on Features that explore the ways platform design contribute to harm and what different interventions can look like). Features described in the risk profiles are helpful, but there are also choices platforms can make in designing elements that may exacerbate or reduce risk. For "recommender systems" for example, there are design choices that can be made at every level of the system that will have an impact on risk and should be considered.

This assessment includes transparency around how harmful content associated with the risk area performs in the platform's systems, including on the scale (how much harmful content related to

a risk area is on the platform), nature (who is seeing the harmful content), and causes (why are they seeing it).

3. Steps the platform will take to mitigate risks and make each component resilient against the risk.

Risk mitigation plans should correspond to the pathways for risk identified in the earlier sections of the risk assessment, and should involve taking steps to ensure that the platform's algorithmic systems are not amplifying content that contributes to the risks. Each component of the algorithmic system should be strengthened and improved to mitigate risks.

ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Question 9:	
i)	Are the Risk Profiles sufficiently clear?
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Do you think the information provided on risk factors will help you understand the risks on your service?
Response:	
iv)	Please provide the underlying arguments and evidence that support your views.
Response:	
v)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Record keeping and review guidance

Question 10:	
i)	Do you have any comments on our draft record keeping and review guidance?
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Question 11:	
i)	Do you agree with our proposal not to exercise our power to exempt specified descriptions of services from the record keeping and review duty for the moment?
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Volume 4: What should services do to mitigate the risk of online harms

Our approach to the Illegal content Codes of Practice

Question 12:	
i)	Do you have any comments on our overarching approach to developing our illegal content Codes of Practice?
Response:	
ii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Question 13:	
i)	Do you agree that in general we should apply the most onerous measures in our Codes only to services which are large and/or medium or high risk?
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Question 14:	
i)	Do you agree with our definition of large services?
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Question 15:	
i)	Do you agree with our definition of multi-risk services?
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Question 16:	
i)	Do you have any comments on the draft Codes of Practice themselves?
Response:	
ii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Question 17:	
i)	Do you have any comments on the costs assumptions set out in Annex 14, which we used for calculating the costs of various measures?
Response:	
ii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Content moderation (User to User)

Question 18:	
i)	Do you agree with our proposals?
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Content moderation (Search)

Question 19:	
i)	Do you agree with our proposals?
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Automated content moderation (User to User)

Question 20:	
i)	Do you agree with our proposals?
Response:	

On the use of keyword detection for fraud: Keyword detection is less effective than a more tailored approach.

To improve the robustness of the keyword approach, other signals such as the account history including content toxicity, coordinated link-sharing behavior, sharing unreliable media, engagement patterns, "disbelief" responses to their posts, account reporting history, creation date, posting frequency, etc. should all be considered in the review of a flagged post or account. This would also serve a meta-analysis of trends in future flagging efforts.

This pre-supposes the availability of keywords in multiple languages that the platform might serve content in. If there is not a reliable, prescribed way to source these keywords, it is likely that this will lead to unmeasured biases in the sourcing of actionable content, with downstream impacts on certain user groups.

ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Question 21:	
i)	Do you have any comments on the draft guidance set out in Annex 9 regarding whether content is communicated ‘publicly’ or ‘privately’?
Response:	
ii)	Is this response confidential? (if yes, please specify which part(s) are confidential)

Response:

Do you have any relevant evidence on:

Question 22:	
i)	Accuracy of perceptual hash matching and the costs of applying CSAM hash matching to smaller services;
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Question 23:	
i)	Ability of services in scope of the CSAM hash matching measure to access hash databases/services, with respect to access criteria or requirements set by database and/or hash matching service providers;
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Question 24:	
i)	Costs of applying our CSAM URL detection measure to smaller services, and the effectiveness of fuzzy matching for CSAM URL detection;;

Response:

Many of our members believe that while hash matching is a current best practice, is accessible for newer organizations and should be recommended for large or multirisk platforms, there are numerous issues with hash matching including:

- Hash matching only works for still images
- Images must be an exact match. Cropping the image even slightly will not trigger the hash and create additional hashes for the same image.
- False positive matches sneak in that triggers against 1000s of images, floods even a very large team quickly
- While checking images against an API is not costly, smaller services will struggle when dealing with reporting and legal aspects.

- Applying to other harms still has the same issues.

An additional potential positive of hash matching is that platforms have the challenge of coming up with definitions of what constitutes various harms such as suicide and self injury or eating disorder content. Hash matching could be a useful tool for these harms as it would encourage platform alignment of these issues. We would additionally encourage companies to do more cross platform hash sharing, as more scrutiny and oversight towards these harms can lead to more cross industry alignment while additionally making the platform ecosystem safer.

ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Question 25:	
i)	Costs of applying our articles for use in frauds (standard keyword detection) measure, including for smaller services;
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Question 26:	
i)	An effective application of hash matching and/or URL detection for terrorism content, including how such measures could address concerns around 'context' and freedom of expression, and any information you have on the costs and efficacy of applying hash matching and URL detection for terrorism content to a range of services.
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Automated content moderation (Search)

Question 27:	
i)	Do you agree with our proposals?
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

User reporting and complaints (U2U and search)

Question 28:	
i)	Do you agree with our proposals?
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Terms of service and Publicly Available Statements

Question 29:	
i)	Do you agree with our proposals?
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Question 30:	
i)	Do you have any evidence, in particular on the use of prompts, to guide further work in this area?
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Default settings and user support for child users (U2U)

Question 31:	
i)	Do you agree with our proposals?

Response:

Defaults are useful especially for maintaining child safety on platforms. We note that while some level of age verification could be beneficial in creating safer spaces, however we strongly caution against ID-based age verification requirements, as there are major trade offs presented with age verification and privacy. There are other approaches to ensuring that children accounts receive the safest settings, including setting all accounts to the safest defaults, or using device-level verification.

Additionally, we argue that defaults are a good tool for improving safety for all users, not just child users. Therefore it is our suggestion that these defaults should be recommended for **all** users as improving user awareness of tools that platforms provide to make informed decisions around content and U2U interactions is always beneficial. Allowing these controls for all users and defaulting to the safest ones is a way to encourage safety for all users. Any reasoning as to why adult users should be given less informative default settings should be researched and thoroughly justified.

We note that there should be thresholds when discussing child users as there is a stark difference between child users age 3-5 and any child user under 13. Therefore we argue that there needs to be parameters around these thresholds. Additionally, we note that the wording of these proposals seems overly prescriptive and fails to be future proof. “Receiving a direct message” fails to encapsulate first time interactions in VR or interactions in a public game chat for example. This assumes that direct messaging will be the sole communication style in x years. We argue that instead of prescribing to the type of interaction, to prescribe towards the **principle** of ‘this is the first time a child user is being contacted’ instead. Finally, we urge platforms to consider alternative approaches, such as implementing penalties for harmful actors rather than implementing more controls on child users. Reducing child freedoms on any platform will not be a permanent solution. Stricter thresholds around those that are promoting the harms are encouraged to target the source of the risk.
ii) Please provide the underlying arguments and evidence that support your views.
Response:
iii) Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

Question 32:
i) Are there functionalities outside of the ones listed in our proposals, that should explicitly inform users around changing default settings?
Response:
ii) Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

Question 33:
i) Are there other points within the user journey where under 18s should be informed of the risk of illegal content?
Response:
ii) Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

Recommender system testing (U2U)

Question 34:
i) Do you agree with our proposals?
Response:

There are good practices when testing recommender systems to see if they are lifting harmful content, and evaluating changes.

Best practices in evaluating changes to algorithmic systems would begin with "offline" evaluation of the change. An "offline" evaluation means that the change is not evaluated on live users, but is instead evaluated using various simulation scenarios. Google Search provides a good example of how this is done.

The first stage of evaluations could be strictly simulated. For the same query or same user, two sets of ranked results could be produced for test and control, and these sets of ranked results could be compared to each other based on existing data on the results. For example, if all the results already have quality scores assigned to them, which is the case for Google Search due to the Search Quality Evaluator Guidelines process, then the ranking of high quality results relative to low quality results could be evaluated directly.

The second stage could involve human evaluations. The two sets of ranked results could be presented to evaluators in "side-by-side" tools, where the evaluator is given both lists of results, and they pick which set of results is better according to various criteria, which could be objective or subjective.

If a change looks good according to both these stages, then a live A/B test on users can be run. This would involve assigning a small subset of users, 1% for example, into a test or control group. Users will remain in the test or control groups as long as the test runs, which could be anywhere from a few hours to a few weeks. At the end, metrics are computed for the test and control groups and compared to assess the impact the change would have.

Platforms have choices about how to design and run their A/B tests. Ideally, platforms would have lots of integrity and trust and safety metrics in their tests. This should include:

- Metrics that track negative experiences
    - Exposures to content that are likely violating: This means exposures to content that score above X% in various content violation classifiers
    - Prevalence of violating content: This involves running a prevalence measurement for both the test and control buckets, and can be very expensive to run for every test, so is likely only run on a limited set of tests.
    - Number of user report/flags of violating content
    - Number of exposures to low quality content or accounts
- Metrics that track good experiences
    - Impressions on high quality content
    - User responses on surveys about experience ("Net Promoter Score" as an example)
    - Clearly good engagement signals (for example, instead of just measuring all impressions, measure impressions on high quality content as a positive engagement signal)

Bad Practices

However, the primary metrics used by platforms are typically motivated by business interests, and will include things like user engagement, time spent on the platform, retention, and impressions on advertisements and revenue. This is poor practice that does not account for any risks. Such poor practice typically begins with goal metrics the company has for some defined period of time, for example "increase time spent on the platform by 5% over the next 6 months". The platform will then develop changes that try to increase the goal metrics to hit the target. Evaluating primarily on this type of metric does little to help understand what risks the algorithmic systems pose to users and society.

Another poor practice would be to immediately begin experimenting on users in live A/B tests. This means running experiments on users with little more than a hunch that it will lead to an acceptable experience (and often it does not). After the A/B test concludes (hours to weeks depending on how much data needs to be collected), then it will be assessed. A poor assessment may be as simple as looking at a few engagement metrics, such as overall time spent on the platform and daily active users, and then launching. These types of narrowly focused decision making processes based on bad testing practices do not account for risk, and may pose risks themselves.

While testing for safety metrics is a positive look on recommender testing we note:

- There is no accountability enforced for platforms to use the results of recommender system testing to actually make changes.
    - Is this requirement simply that the service document its understanding of the impact of proposed changes on safety metrics, not guidance on whether or not the service can ship those changes even if the tests show negative impacts?
- It is also worth noting that testing for prevalence of illegal content has its limitations, and expectations should be realistic. Prevalence of illegal content on platforms (compared to total content) is often so low that it may be difficult for even the largest platforms to find statistically significant metrics in test results.
    - Will be hard to see effectiveness of results.
    - Will rarely see illegal content go down X% where is X looks like a meaningful number

To combat the accountability piece, we recommend the results of these tests and definitions of the metrics used be available to third party auditors. Platforms should also be transparent about the most important features and machine learning models used in their recommender systems, and test how harmful content performs in these models. This can be tested internally and released publicly without violating any proprietary information: platforms can report inputs and outputs of their systems. When X classifier is weighted heavily in recommendation scoring systems [input], does level of [X type of harmful content] go up or down [output].

Due to the general rarity of illegal content, and the difficulty in seeing changes in live A/B tests, arning classifiers used in the ranking and recommendation systems. If the average classifier scor platforms should also study how their recommendation systems respond to illegal content. For example, for each machine learning classifier that plays a significant role in recommendations, the platform should keep track of how illegal content performs in it, meaning, what is the distribution of classifier scores that illegal content got. We assume here that the platform will log scores given

to all content, and then when content is found to be illegal, the platform will look through their historical records of scores and record the scores the particular piece of illegal content received. Then, the platform can compare the distribution of scores that the illegal content received to the overall distribution of scores, and see if illegal content systematically scores higher in the platforms machine le e for illegal content is higher than the overall average score, or the prevalence of illegal content becomes higher at higher classifier scores, then that particular classifier can be said to amplify the illegal content and thus play a significant role in the risk of spread of illegal content.
ii) Please provide the underlying arguments and evidence that support your views.
Response:
iii) Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

Question 35:
i) What evaluation methods might be suitable for smaller services that do not have the capacity to perform on-platform testing?
Response:
ii) Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

We are aware of design features and parameters that can be used in recommender system to minimise the distribution of illegal content, e.g. ensuring content/network balance and low/neutral weightings on content labelled as sensitive.

Question 36:
i) Are you aware of any other design parameters and choices that are proven to improve user safety?
Response: Platforms have choices in designing their recommendation and ranking systems that can be made at every level of the system. Broadly speaking, choices that favor engagement based ranking will be risky. **Inventory** (what Ofcom refers to as "content pool"): The more opportunities there are for users to be shown lower quality content from accounts they don't follow, the larger the risk is that violating and risky content will be broadly distributed. Evidence: Facebook 2020 series of research. In one experiment, researchers removed all reshared content from test users News Feeds. The result was that users saw a 30% reduction in the amount of content they saw from untrustworthy news sources, which was defined as news sources that posted two or more pieces of misinformation. Additionally, leaked research from companies have shown instances where recommendation systems were responsible for 64% of users joining hate groups.

Integrity Institute research into the amplification of misinformation by platforms: In our study, we saw the highest levels of amplification of misinformation on Twitter, TikTok, and Facebook Video. These are the platforms that rely heavily on engagement based ranking and where most views on content come from algorithmically recommended and reshared content.

Features (what Ofcom refers to as "content signals"): Features based on the historical engagement of users will pose more risk than other types of features that machine learning models can use, such as those related to the quality of the content or the safety of historical content from the creator.

Machine learning models and final ranking score ("tuning prediction weights"): Machine learning models that predict the likelihood of a user engaging with content will pose more risk than other types of machine learning models, such as those predicting the safety of the content or the quality of the content. The more the ranking score is dependent upon engagement based factors, the higher risk it poses.

What is missing from Ofcom's proposal is a discussion of other features of platform design that can improve user safety. (Again, we refer to the Focus on Features project for more detailed discussions of specific features and what harms they can intervene on.)

Companies can implement platform features that guard against gaming the systems. Examples of these types of features include:

- Friction: e.g. requiring more steps before resharing a post, to discourage low-information reshares.
- Slowing the spread of highly viral content.
- Aging in: requiring accounts to exist for a certain period of time before they gain access to certain features with higher risk potential.
- Limiting the size of chat rooms, groups, and account followers.

Companies can impose limits around risky behaviors (usually these are discussed in the context of protecting children). This includes limits such as:

- No recommended content (only content from people a user has elected to follow)
- No recommended follows of unknown accounts
- No contact (DMs, comments) allowed from unknown or unconnected accounts

ii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Enhanced user control (U2U)

Question 37:	
i)	Do you agree with our proposals?
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.

Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Question 38:	
i)	Do you think the first two proposed measures should include requirements for how these controls are made known to users?
Response:	
ii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Question 39:	
i)	Do you think there are situations where the labelling of accounts through voluntary verification schemes has particular value or risks?
Response:	
ii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

User access to services (U2U)

Question 40:	
i)	Do you agree with our proposals?
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Do you have any supporting information and evidence to inform any recommendations we may make on blocking sharers of CSAM content? Specifically:

Question 41:	
i)	What are the options available to block and prevent a user from returning to a service (e.g. blocking by username, email or IP address, or a combination of factors)?
Response:	
While Ip addresses can be useful in following up on accounts, it isn't a useful tool in handling CSAM recidivist accounts as changing an IP address or using a VPN are easy ways to combat IP	

blocks. That being said, IP blocks and username checks are easy enough for any size platform to incorporate as an early step.

We encourage platforms (especially ones with high risk of CSAM content) to track social graphs. Users that recidivate typically join similar social graphs and groups immediately after returning to the platform.

Regarding punishment for sharing CSAM content:

- If someone posts something that needs to be reported to NCMEC, their account should be locked no matter the level of offense
- Some instances of CSAM sharing are less malicious. We argue that malicious or high impact CSAM sharing should be on a zero strike policy.

While IP and username blocking along with tracking social graphs are best practice measures for addressing CSAM content, as it stands, human review is by far the most applicable to date. Considering that human review is necessary for legal considerations when reporting to NCMEC, having a human in the loop is an essential step for addressing CSAM.

ii)	What are the advantages and disadvantages of the different options, including any potential impact on other users?
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Question 42:	
i)	How long should a user be blocked for sharing known CSAM, and should the period vary depending on the nature of the offence committed?
Response:	
ii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

There is a risk that lawful content is erroneously classified as CSAM by automated systems, which may impact on the rights of law-abiding users.

Question 43:	
i)	What steps can services take to manage this risk? For example, are there alternative options to immediate blocking (such as a strikes system) that might help mitigate some of the risks and impacts on user rights?
Response:	
ii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Service design and user support (Search)

Question 44:	
i)	Do you agree with our proposals?
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Cumulative Assessment

Question 45:	
i)	Do you agree that the overall burden of our measures on low risk small and micro businesses is proportionate?
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Question 46:	
i)	Do you agree that the overall burden is proportionate for those small and micro businesses that find they have significant risks of illegal content and for whom we propose to recommend more measures?
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Question 47:	
i)	We are applying more measures to large services. Do you agree that the overall burden on large services proportionate?
Response:	

ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Statutory Tests

Question 48:	
i)	Do you agree that Ofcom's proposed recommendations for the Codes are appropriate in the light of the matters to which Ofcom must have regard?
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Volume 5: How to judge whether content is illegal or not?

The Illegal Content Judgements Guidance (ICJG)

Question 49:	
i)	Do you agree with our proposals, including the detail of the drafting?
Response:	
ii)	What are the underlying arguments and evidence that inform your view?
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Question 50:	
i)	Do you consider the guidance to be sufficiently accessible, particularly for services with limited access to legal expertise?
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Question 51:	
i)	What do you think of our assessment of what information is reasonably available and relevant to illegal content judgements?
Response:	
ii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Volume 6: Information gathering and enforcement powers, and approach to supervision.

Information powers

Question 52:	
i)	Do you have any comments on our proposed approach to information gathering powers under the Online Safety Act?
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Enforcement powers

Question 53:	
i)	Do you have any comments on our draft Online Safety Enforcement Guidance?
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Annex 13: Impact Assessments

Question 54:	
i)	Do you agree that our proposals as set out in Chapter 16 (reporting and complaints), and Chapter 10 and Annex 6 (record keeping) are likely to have positive, or more positive impacts on opportunities to use Welsh and treating Welsh no less favourably than English?
Response:	
ii)	If you disagree, please explain why, including how you consider these proposals could be revised to have positive effects or more positive effects, or no adverse effects or fewer adverse effects on opportunities to use Welsh and treating Welsh no less favourably than English.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	