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:	
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:	
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: With any third party-risk assessment, there is a risk that a lack of depth of understanding of a business could influence any audit. This is a complex area, both due to the topic, as well as the range of services in scope. Even an independent third-party specialising in audits to mitigate and manage illegal content risks would need to develop an in-depth understanding of the working of the unique service in question. While services provide auditors with data and information for context, there can be significant	

complexity due to the business product and structure of online services, which can be unique to each service. This is especially so given the sheer diversity of online services in scope, and the nuances in how they operate.

Added to this, double-sided services (those servicing both businesses and consumers) can involve a further level of complexity as changes on one side of the service can have impacts on the other side and these are not always easy to foresee or predict.

[✂]

Ensuring consistency in this process is also critical. In other instances we have had experiences where consistency has been lacking with regard to what different stakeholders are looking for and expecting through such a process. It is therefore key that well established and not overly prescriptive requirements are set to deliver a consistent approach.

It is vital that third party audits deliver quality and meaningful outputs. However, in light of our experiences, we underline the risk of such audits not delivering this due to the complexity of online services, and the cost required to ensure that third-parties have a suitable understanding of services (which are often complex and unique).

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

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: We agree to a certain extent, but believe that refinements are required.
ii) Please provide the underlying arguments and evidence that support your views.
Response: Trustpilot recognises the importance of risk assessments and their use in this context. We also welcome Ofcom's commitments to ensuring its approach is proportionate and the steer that indicates that this approach is focussed on mitigating risks. The proposal to have a three month window from the guidance being published to compliance being required appears reasonable. However, we would note that Trustpilot is a company which already uses a similar risk assessment model, is already very proactive in tackling illegal harms content and which is closely following the application of the *Online Safety Act*. For services which do not currently adopt such an approach and/or who are not aware of this upcoming requirement, there may be challenges to comply within the three month window, particularly if they only become aware of the guidance a while after publication. A significant communications effort will be required to ensure that these new requirements reach all the companies in scope. As it stands, Ofcom is proposing that risk assessments would need to be reviewed in the event that the regulator updates its own guidance or if 'significant changes' are made to the service. Whilst these have a reasonable basis, we would stress a number of considerations in order to ensure that these requirements are workable and proportionate. <u>i) The frequency of Ofcom's updates to Risk Profiles</u> With respect to requiring services to review their risk assessments should Ofcom update a risk profile. Ofcom states "we expect to update our Risk Profiles to keep them up to date as our evidence base develops in the Register." While it is correct that the Risk Profiles should be kept up to date and relevant, we would stress the importance of ensuring that the risk assessment linked requirements are workable. In its proposals no details are set out with regard to how frequently Ofcom envisages updating the risk profiles. It is therefore difficult to judge how proportionate the linked requirement to review risk assessments will be in practice. At the very least, Ofcom should give an indication of how frequently they envisage making such updates so that adequate capacity can be held within services for conducting reviews of risk assessments. In order to be proportionate, we would suggest that a commitment should be made to group Risk Profile updates on an annual basis. This would seem a reasonable level of frequency at which to

release such updates and grouping them would prevent services having to review risk assessments more than twice a year on this basis. We argue that this would deliver a proportionate approach given how comprehensive the Risk Profiles are from the details provided at the outset.

Further to this, we would also ask for clarification as to the timeframe within which services need to respond to updates to the Risk Profile changes. Given the period of three months has been proposed for services complying with the guidance, we would suggest that it would be appropriate to mirror this timeframe in relation to this aspect as well.

ii) The level at which 'significant changes' are determined

Ofcom proposes to require services to "update their risk assessment whenever a 'significant change' to their service occurs". At 9.136 five factors are set out which the regulator notes are "the kind of changes we expect to amount to a significant change and trigger this duty". However, these are incredibly broad and lead to concern that small iterations to a service *could* be considered as a 'significant change' and thus require new risk assessments to be undertaken - particularly when b) and e) are considered.

This concern that unnecessary risk assessments are going to be triggered, is heightened by some of the questions provided in Table 13 of annex 5. A question for helping to determine whether a risk assessment *must* be carried out includes whether "changes to your business model in terms of how you generate revenue or your growth strategy".

This is an incredibly low bar given that most changes can in some way be linked to either revenue or growth - or indeed both - given growing and increasing revenue is a goal of most businesses and online services.

Overall, we are concerned that this approach is very burdensome and fails to take account of the realities of how online businesses operate and evolve. There is a significant risk that innovation will be stifled, that costs of disproportionate risk assessment *prior* to enacting product changes will outweigh the benefits.

Avoiding a drag on innovation

We are concerned that - should the balance not be struck correctly - there is a risk that the 'significant change' requirement for further risk assessment could result in a drag on innovation. For businesses that are interacting and updating their product, this could conceivably result in conducting full risk assessments multiple times a quarter. While risk assessment is an inherent and natural part of good product development, the danger is that for lower risk services, it could yield little practical gain via this approach.

To retain the value to the exercise, it is imperative that this approach to risk assessment becomes a useful tool which prompts necessary discussions, rather than a bureaucratic exercise. As it is developed, the process must remain workable and not become too convoluted. Documentation is important, but such requirements should stay proportionate so that they are workable and successful, delivering good outcomes, without stifling innovation.

Further to this, whilst split testing and other testing is touched on in the guidance, we believe there would be value in clarifying that testing *does not* constitute a 'significant change'. Rather, that testing is encouraged as a useful part of mitigating risk and not inadvertently discouraged by ambiguity or statements to the contrary.

Testing plays a significant role in iterative product development and enables assessments to be tested. Assumptions about how changes will impact user flow can be wrong, and thus testing is needed to bring these issues to the fore. Given the critical role of testing, it's important that businesses are not penalised for attempts to assess how changes impact content on their systems and for exploring mitigation solutions, even if this does not always have the intended effect. The knowledge garnered from this stage of product development is so valuable and can be a real asset in ensuring that risks are identified and mitigated before product changes are fully rolled out.

We request that Ofcom enhances its guidance to clarify this important aspect.

<u>Ensuring parity across services</u>
The final consideration to raise regards the potential risk - which is present across this Act and in this approach to risk assessment - that the requirements are more proportionate to meet as a big tech firm, and less proportionate for services which are not leviathans of the tech world.

It is well acknowledged that big tech firms, including the social media giants, have significant resources and staff numbers, often with large legal teams which can be deployed to minimise the impact of such proposals.

Whereas more modestly sized and resourced firms (including some of those who will be deemed 'large' under the threshold proposed by Ofcom, but are vastly smaller than big tech) will see a far higher proportion of their resource needing to be directed on compliance because they do not have the large resources to make intricate legal arguments based on this guidance. This risks creating an unlevel playing field, reinforcing bias towards the leviathans of the tech industry - which sits in tension with the *Digital Markets, Competitions and Consumers Bill* which is passing through Parliament currently.

We urge Ofcom to think carefully about how this risk of a built-in bias towards big tech firms can be balanced, particularly when considering the proportionality of this approach.

<u>Solutions</u>
In order to address these concerns we recommend that Ofcom:

- Defines 'significant change'

We propose that a suitable and workable definition would be "A 'significant change' is substantial modification of or transformation to a service's underlying design or in its operation. It does not apply to routine changes and/or developments for the service's design or operation".

This conveys key - and currently missing - detail. A change is classified as a "modification" and significance is determined as being "substantial".

- Provides clarity regarding testing and split testing

We believe there would be value in clarifying that testing *does not* constitute a 'significant change'.

- Refines its examples

As highlighted the current examples to help assess a 'significant change' are currently too broad and therefore have a high risk of placing burdens for unnecessary additional risk assessments. These examples should be refined in light of feedback gathered to ensure they are proportionate, targeted and workable, whilst - most importantly - effectively targeting risk.

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

Response: No

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:

Yes

ii)	Please provide the underlying arguments and evidence that support your views.

Response:

The four-step risk assessment process broadly aligns with what is considered good practice in this area. The Risk Profiles are also helpful.

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

Response: No

Question 9:

i) Are the Risk Profiles sufficiently clear?

Response:

The Risk Profiles are relatively clear, but in some areas further clarity is required. These are with regard to recommender systems and pseudonymity.

ii) Please provide the underlying arguments and evidence that support your views.

Response:

Recommender systems

When considering recommender systems, Volume 2 of the guidance points to two specific recommender systems which are considered as potentially risky. These are:

- Content recommender systems
- Network recommender systems

As described in Volume 2, these systems - and effects - are most commonly found on social media services.

While honing in on these two types of recommender systems is reflected somewhat in the Risk Profiles, it is important to retain the specificity and this could be clearer than it currently is (the title in the Risk Profiles suggested a far broader scope "8 Services with recommender systems").

The *Online Safety Act* is estimated to cover over 100,000 online services. Within these, a number of services would not have *content recommender systems* or *network recommender systems* as defined in Volume 2, however would have systems which could fall under the much broader umbrella of general "recommender systems" which is denoted in the Risk Profile title ("8 Services with recommender systems"). The ambiguity of this title is therefore unhelpful and could result in services with more banal recommender systems having to give attention to this area when they are not the intended targets, arguably diverting critical resources away from tackling relevant areas of risk.

As Ofcom itself recognises, not all recommender systems carry the same level of risk and it is the specific types of recommender systems which are raised in the guidance which are relevant. It is therefore important that this delineation is also recognised in the Risk Profiles too.

We therefore recommend that the types of recommender systems which are being targeted should be better reflected in the U2U Risk Profiles. For example, in Table 14 of Annex 5, "8 Services with recommender systems" could be refined to be "8 Services with *specific types of* recommender systems".

This alignment would be a clearer signal for U2U services that have *other* types of lower-risk recommender systems. Enabling them to more efficiently address this part of the risk assessment and dedicate their resources to areas of relevant risk for them.

iii)	Do you think the information provided on risk factors will help you understand the risks on your service?

Response:

We broadly agree with the types of services that the Risk Profiles focus on. It is understandable that not all types of service can, or indeed should, be considered at an in-depth level, given the diversity of services in scope.

From analysis of our reviews and reviews flagged by consumers and businesses, we believe that many types of illegal harm in the *Online Safety Act* are highly unlikely to be prevalent on our service, due to the nature of online reviews. As such, this justifies not addressing online review services directly as a sub-type of service within the guidance.

Given this context, much of the insight that a service like Trustpilot gains from the risk factors is indirect. The approach taken is useful to understand which other types of services are likely to have such risks, and to understand how those risks manifest themselves in such examples.

We are then able to utilise our own understanding of the *differences* between online review services and those other types of services listed in the Risk Profiles to compare and contrast this information.

That said, we do think an improvement can be made with regard to the area of pseudonymity. This is set out below.

iv)	Please provide the underlying arguments and evidence that support your views.

Response:

<u>Pseudonymity</u>

To further enhance the clarity provided by the Risk profiles we suggest further improvements are made regarding pseudonymity.

In the context of the Risk Profiles, we note that anonymous profiles are stated as having the potential to be risky “in certain contexts”. This implies that not all instances carry the same level of risk, which we agree with.

However, within the definition of “anonymous user profiles”, pseudonymity is also included. This is set out in annex 5, p57, footnote 39 which says “instances where a user's identity (an individual’s formal or officially recognised identity) is unknown to other users, for example through the use of aliases ('pseudonymity').”

We would argue that pseudonymity should be distinguished from anonymity, and that pseudonymity attached to a registered account should not be considered to generate the same level of risk as anonymous profiles.

This is especially the case where platforms require users to create and register an account (attached to a valid email address) in order to post user-generated content. Even though other

users of the service may not know the true identity of the poster, this layer allows the service in question to have a fixed way of contacting the user behind the account, including for redress if necessary.

In the context of online reviews, pseudonymity can be particularly beneficial for users writing reviews because it allows freedom to write critical reviews. It is vital that the benefits this brings are recognised and that, from a risk perspective, pseudonymity is not seen as being akin to anonymity which is quite different, as the above example demonstrated.

On this basis, we recommend that further clarity is provided by Ofcom to distinguish pseudonymity from anonymity, and that pseudonymity attached to a registered account *should not* be considered to generate the same level of risk as anonymous profiles.

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

Response: No

Record keeping and review guidance

Question 10:

i)	Do you have any comments on our draft record keeping and review guidance?

Response:

We agree with the need - and benefit of - services holding clear and well-maintained records which document compliance.

We would note, however, the importance of ensuring that this can be delivered in a manner which is proportionate, workable and useful to all involved.

ii)	Please provide the underlying arguments and evidence that support your views.

Response:

As it stands, the priority offences are already quite extensive and so assessing them in relation to potential harm could be quite a considerable exercise.

Were Ofcom to develop online tools to assist assessments we would recommend that a framework approach is taken which services can draw on whilst allowing a level of flexibility to apply them as best fits their circumstance and service.

[✂]

To this end, we would recommend that Ofcom takes a more flexible approach, ensuring that the record keeping and review guidance gives a clear steer on this area, but enables flexibility to meet the requirements in the most appropriate way for each service.

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

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: Trustpilot commends the approach being taken in considering size *and* functionality when determining what measures apply to services. This is a significant step up from the approach being taken via the *Digital Services Act* (DSA) which focuses only on the *size* of services. This has been an area of concern with the approach being taken under the DSA, and raises questions as to whether focussing on size alone tackles risk as effectively as it could. Combining functionality with size takes a significant step forward in targeting risk.That said, we do have some concerns with regard to both how size is being established and how multi-risk services are defined. These concerns, and our suggestions for refinement, are set out in response to Q14 and Q15 respectively.	
ii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response: No	

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: We challenge how 'large' services are being determined and also how low the bar is being set to determine whether a company is 'multi-risk'. We would propose that these metrics are both refined. Our views on this are set out in response to Q14 and Q15.	
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: No	

Question 14:
i) Do you agree with our definition of large services?
Response: We do not agree with the definition in the proposal. This is on the basis of a number of considerations which are set out in the following response.
ii) Please provide the underlying arguments and evidence that support your views.
Response: In Ofcom's consultation documents, the approach to 'users' is defined in a number of key places, these include at: ● 6.10 - which sets out the questions used to identify the risk factors relevant to each group of characteristic for the "user base" ● 9.59 - where the rationale is set that "In general, all else being equal, the more users a service has, the more users can be affected by illegal content and the greater the impact of any illegal content. We have therefore proposed that services which reach certain user numbers should consider the potential impact of harm to be medium or high. ● 9.62 - where it is noted that the Service Risk Assessment Guidance noted that "in some instances the number of users may be a weak indicator of risk level." And as such user numbers need to be considered alongside other risk factors. ● 11.58 - where the proposed definition for 'large services' is set. Throughout, it is clear that the aim of this guidance is to ensure that the impact of illegal content on individual users who are active users, engaging with content on services and thus at risk of encountering illegal content. <u>Concerns regarding the definition of 'user'</u> However, the definition of user (presented at 6.5) does not match this purpose. A user is defined in the following terms - "User base refers to the users of a service. A user does not need to be registered with a service to be considered a user of that service." Registered users are a very important metric for services - helping services to track activity and take targeted approaches to developing their services, including in areas such as delivering product improvements and fixes. Expanding the definition of 'user' beyond registered users could lead to a very broad and onerous approach for services and platforms, including those which are lower risk. Overall, this very high level definition is incredibly broad meaning that anyone who clicks on a service would be counted as a user regardless of whether they *interacted* with any content on the service. For example, in the instance of Trustpilot, a user could accidentally click through to our service when scrolling on a different website page which displays one of our widgets, and then leave

straight away. The user content we host is text based, as such a 'click on, click off' user would not be able to engage and process written content in order to be exposed to harm.

However, despite the user not having engaged with the written user review content and thus not having had the potential to be exposed to any illegal content, they would count as a 'user' under this proposal. As this demonstrates, applying this broad approach across all services goes beyond the purpose of what the Act intends to tackle.

Proposed new definition for 'user'

To resolve this issue, we propose refining the definition of 'user' to be 'active user'. This language is utilised in both the EU and Australia - jurisdictions which Ofcom has noted it's alignment with in setting the bar for what is considered a "large service".

That said, while we believe that echoing the language on 'active user' is useful, we would call on Ofcom to use this in a more tailored manner - setting a reasonable and proportionate bar for what an active user is in this context.

We would propose to determine that an 'active user' is "engaging at least once in [a set period of time] and being exposed to information for [a certain period of time - for example three or five seconds]".

The rationale behind this definition is that, firstly a user is considered to be an 'active user' and thus in the target of this Act - a user engaging with content and at risk of being exposed to illegal content.

Secondly, the topic of whether said 'user' actually engaged with content and was indeed 'active' is also addressed in the form of having an opportunity to meaningfully and actively engage with content and thus be at risk of harm. With respect to what the time frame is set for how long a user must be on a service to digest the content, we suggest that this is linked to the *type of content* and the time frames are established via user testing. Our hypothesis being that a user accidentally clicking through to a service and clicking straight off could engage with picture content far faster than written content.

Indeed, we understand that in Australia a framework based on a format of materials is being used in their approach to social media content - separating audio and text from other formats.

This Act will cover over 100,000 vastly different services - the limitations of user numbers alone is already recognised by Ofcom. Given this, a more nuanced approach is required to target the risk under consideration and this approach presents a nuanced and evidenced based way to do this.

The bar for determining 'large services'

We are concerned that the threshold for 'large services' is being set at 10% of UK users and that this is too low. This concern is heightened when the definition is coupled with the broad approach to measuring 'users'. Our views on the definition of 'users' being set out above.

If the definition of 'user' is kept as is, we strongly advocate that the 10% threshold is raised to take account of this.

More broadly, we are concerned how broad the categorisation of 'large services' will be. While there is rightly a wish to ensure differentiation of services, classing a service with 10.1% of UK users in the same grouping as a service with over 50% of UK users seems out of proportion.

If no further nuance is given to how services are categorised by size we would request that commitments are made to ensure that this definition is not copied for use in other policy and legislation. This request comes from experience in the EU, where the blunt 'Very Large Online Platform' definition is now being drawn on in relation to a range of policy proposals including in financial services which is completely inappropriate given the sectoral differences.

Ofcom has recognised the limitation of purely looking at user numbers by coupling the size with functionality to determine risk, but we strongly caution against allowing this definition of size to be used as a basis of other policies and laws.

<u>Using resources efficiently</u>
A further point to raise is in relation to the mechanics of counting and reporting user numbers.

From the perspective of a platform which is in the scope of both the *Online Safety Act* and the *Digital Services Act* we highlight an opportunity to avoid unnecessary duplication, if the broad definition of "users" is retained.

Platforms, like Trustpilot, which are not classed as 'Very Large Online Platforms' under the DSA have to measure users via a dashboard which uses a six-month rolling average. With the DSA now in force, this dashboard has been created and is maintained by our Data Team.

In Ofcom's proposed approach, services will need to use a 12-month average of UK users per month. This will require additional resources across services to create an updated dashboard to capture and record the data in this manner.

Given the 6-month rolling average of EU users vs the 12-month average of UK users has little material difference, we would suggest that Ofcom aligns with the DSA's approach. This will enable services to use their resources as efficiently as possible - simply requiring a UK data set to be added to an existing dashboard, rather than requiring the creation of a new dashboard altogether. This would lessen the compliance burden for the majority of services without impacting the quality of the data which is being gathered.

We fully recognise that Ofcom is enforcing the UK regime and this will differ to the EU's regime. However, where there are opportunities to remove unnecessary duplications of workload for services, this should be taken, as in this case.

<u>"Large" services have varying levels of resource</u>
Further to this, we also draw Ofcom's attention to the assertion made through the guidance regarding the resourcing of 'large' services.

While large services have significantly more resources than small or micro services, when a comparison is made with those at the opposite end of the scale (namely, big tech), some large services are more akin to the former group of services.

Given the bar for 'large' services is currently set at those with 10% of UK users (based on very basic definition), there could be a huge variation in both the nature and size of services which are labelled 'large'. For example, a service with 'user' numbers equating to 10.1% of the UK population looks very different to a popular social media service, for example, which has 'user' numbers equating to over 50% of the UK population.

It is critical to recognise that services which are not classed as 'Category 1' services in the UK but are considered 'large' are likely to have less resources than their larger counterparts.

In this regard, it is important to note that multiple factors need to be recognised when considering resourcing:

i) Firstly, services that have to comply with multiple regulatory regimes need to deploy their limited resources efficiently and effectively in order to maintain their ability to innovate and compete with larger incumbents.

ii) Secondly, even if lower-risk services are able to justify *differentiation* to fulfil their duties under the *Online Safety Act* and explain any deviation from the recommended guidance measures, they must still follow the overall process, including via documentation and regular reassessments, all of which requires resourcing.

iii) A further layer of complexity is added where rules consider some services more directly than others, for example social media. Whilst direct application has perhaps been focussed on one type of service due to risk level or prevalence, for other services - such as review services - direct application may not be as straightforward as envisaged, requiring more resources to be dedicated to interpreting and applying it.

Against this backdrop, for many services there will be a trade off on a sliding scale between regulatory compliance and innovation. It is therefore important that all required changes are carefully weighed, as even small changes can impact innovation.

Overall, Trustpilot supports efforts to improve online safety, and the implementation of this Act is very important. To maximise the benefits and minimise unintended consequences, it is important to ensure that the application of this part of the Act does not become an academic exercise of justification and differentiation. Otherwise, it risks favouring the largest and most well-resourced services - those with access to the most extensive legal and data resources to deploy. Such an outcome would be of detriment to competition and to services which fall at the lower end of the 'large' category. Likewise, it is important that unnecessary impacts on innovation are also avoided.

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

Response: No

Question 15:
i) Do you agree with our definition of multi-risk services?
Response: Overall, Trustpilot agrees with the approach being taken and the importance of incorporating functionalities into assessment of risks. As stated in response to Q12, we think that overlooking the role of functionalities is a weakness of the EU's approach via the DSA, so it is welcome to see that recognised here. That said, we do have a number of concerns with the current approach which are set out in response to part ii of this question.
ii) Please provide the underlying arguments and evidence that support your views.
Response: Our concerns regard: <u>i) The lack of rationale for setting the bar for 'multi-risk' services</u> In setting out its approach to categorising firms as 'multi-risk' there is a lack of evidence or rationale provided by Ofcom as to why the threshold is being set at two 'medium' or 'high' risks designations from the 15 types of priority offences which constitute a service being deemed 'multi-risk' service. As set out below, we believe that this should be refined, however if it is not and the threshold remains at this level, Ofcom should set out the rationale for this level being appropriate. <u>ii) The level of the bar for being a 'multi-risk' service</u> Our view is that setting the bar for a 'multi-risk' service at two medium or high risk categories being identified from the 15 types of priority offences, is too broad and too low. For example, a review service could have in place systems and processes to reduce the harm of hate speech by filtering out this content before it is published. Yet, in the event that the system doesn't catch every single instance (no system is perfect) and those cases are flagged and reported by users, this would be counted as 'medium risk' under the process being proposed. This is on the basis that one of the thresholds to assist services in determining their risk level (as set out in Table 6 of Annex 5), sets out that a service would be considered "medium risk" on the basis that "You assess that there is a moderate likelihood that this illegal harm could occur on your service" and a condition that meets this is if "There is evidence that harm is likely to occur on your service" or "There is some evidence of harm taking place on your service". The drafting in Table 6 of Annex 5 is very loose and leaves pretty much all services open to meeting these very low bars. A single piece of evidence would count as evidence of harm occurring and, by virtue of a user flagging it, an argument could be made that harm was

experienced. We strongly suggest that tighter wording is required to create a meaningful threshold.

This approach also underlines the issue with how 'multi risk' is being defined with the same weighting being given to medium and high risk categories when defining whether a service is 'multi risk'.

That a review service which has a handful of instances of written reviews containing hate speech being flagged - having slipped through their systems - and swiftly taken down, is given the same weighting under this framework as a site which say, regularly exposes users to CSAM content which can go viral and be pushed through recommender systems seems out of balance and disproportionate in the context of the goals of this Act.

Added to this, we note that whilst there will be an obligation on services to comply with this Act, there is also a requirement to uphold freedom of expression under Article 10 of the *Human Rights Act*. It is therefore critical that the approach to implementing and enforcing the *Online Safety Act* does not result in users' speech being over censored by services.

Our understanding is that Ofcom expects a low number of services to be categorised as multi-risk. However, with the definition set so broad and the criteria arguably very low, we would envisage that many more companies would actually categorise themselves as being 'multi-risk' under the current proposals and we are not convinced that this will necessarily result in risk being tackled most effectively.

<u>iii) Unintended negative impacts of the proposed approach</u>

Added to this, in setting the threshold at the low level of two medium risks or above, there is a risk that services prematurely divert their resources from mitigating measures to target the two medium risks in question to reduce the risk level, to putting in place broader, multi-risk obligations.

Where a company has four or five medium or high different risks there is a clear benefit in adhering to cross-cutting multi-risk obligations. Yet, for just two medium risks, the advantages seem less clear and, indeed, the ability to target those risks is arguably watered down - a counter intuitive outcome.

Ultimately, to impactfully identify and target risk through this framework, the ability to effectively differentiate risk and services is key. We are concerned that the proposed application framework does not deliver this.

<u>iv) Potential bias to the largest and best resourced firms</u>

As it stands, a company deemed 'large' which is at the smaller end of the "large" scale, could conduct a conscientious and subjective assessment and be placed in the same space as a much larger service which has a vastly different risk profile.

What is more, the factor of resource arguably then becomes a key factor (a risk factor in other parts of this approach too). Services with extensive resources (often the largest firms) can harness these to minimise how much stringent requirements impact them.

Ofcom's approach should deliver meaningful action, but there is a clear risk that firms with more resources could target these on justifying and documenting their assessments in a manager which gives them more conscience to downplay their risk as 'low'.

Proposed solution

Ultimately, we want an approach where risk is tackled most effectively, where services and risk are meaningfully differentiated and where services with abundant resources can't take advantage of the system in place and minimise their compliance simply due to their resources, in a way that smaller companies cannot.

To address these concerns, we propose that the threshold for multi-risk is raised to require *either*:

- The threshold for being categorised as 'multi-risk' is raised to *four* medium or high risk categorisations, with *at least two* required to be high; or
- A grading or points system is introduced to differentiate between levels of harm. This would enable the process to take into account levels of harm and have services categorised accordingly. The differentiation could be based on *either* evidence of what categories cause the most severe online harm or the level of sentencing in relation to the sentences for culprits of these harm using the hierarchy established by the application of the law to date. Both options would provide an evidence based hierarchy from which to refine Ofcom's approach. Using this framework, a suitable level can then be set for what accumulation accords to a service being 'high risk'.

The refinement of this approach is especially key if Ofcom retains its approach to designating the size of a platform based on a blunt definition of 'users' (our concerns with regards to this are raised in response to Q15). If this approach is retained, then the role of functionality becomes even more important and the benefits for creating clearer differentiation are heightened in order to accurately target the largest risks. Regardless, there is a strong case for this to be refined.

iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response: No	

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:	
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:	
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: With respect to standard keyword detection as discussed in Volume 4, Chapter 14 - we would note that our experience determines that keyword detection is only partly effective in this context and would require constant moderation and human input. Standard keyword detection has a number of positives as well as negatives, which include the generation of false positives, limited ability to understand the context in which the keywords which were used, a need for keywords to constantly be updated and the ability of scammers to identify the use of key words and deploy attempts to circumvent them and.	

In considering fraud and financial offences, we would agree that keyword detection would not be sufficient to tackle offences. Keywords may be a good start if services have nothing in place, or limited technician resources, but more developed solutions are required to avoid a "whack-a-mole" approach. At Trustpilot, we use more developed solutions such as AI and machine learning which are arguably far more effective and specialised than keyword detection. We therefore agree with Ofcom's approach to not proposing keyword detection technology as a panacea to the issue identified.
iii) Is this response confidential? (if yes, please specify which part(s) are confidential)
Response: No

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:	
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:	
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:	
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:	
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:	