Snap Inc.

SNAP INC RESPONSE TO OFCOM CONSULTATION: PROTECTING PEOPLE FROM ILLEGAL HARMS ONLINE

1 March 2024

CONTENTS

Introduction

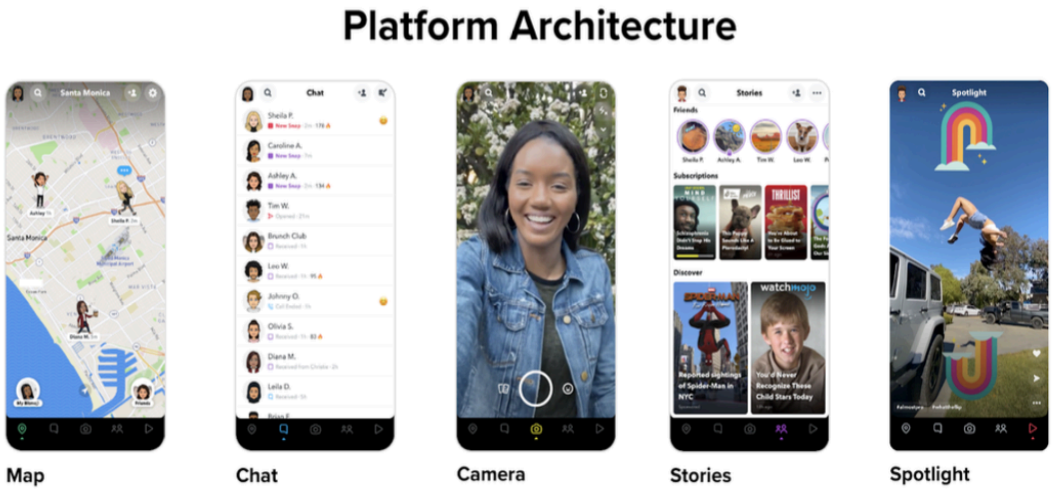
Thank you for the opportunity to respond to Ofcom's consultation on 'Protecting people from illegal harms online' as part of Ofcom's work to implement the UK's online safety regime.

To support our consultation response, it is first important to understand how Snapchat works and our efforts to serve the well-being of our community as Snap grows and faces new opportunities and challenges.

Snap Inc. is a technology company, which owns the well-known visual messaging platform, Snapchat. Snapchat is 13 years old and was built during the dawn of social media to offer people an alternative from the social media which cultivated a popularity contest between users, including chasing likes, comments and followers. Snapchat fosters relationships between real friends and family and it is primarily used as a private, visual messaging service without focussing on the pressure for social validation. The app opens into the camera to encourage creativity rather than passive consumption and scrolling through an endless public feed.

[REDACTED/CONFIDENTIAL]

While the heart of our platform remains as a visual messaging service, over the years it has evolved as our community and their interests have grown whereby we now have five tabs: Camera, Chat, Map, Stories and Spotlight:



Camera: Snapchat opens to your perspective. It's where Snapchatters can create Snaps, interact with Lenses AR experiences, and more.

Chat: Snapchatters can swipe left from the Camera to go to the Friends screen. Here, Snapchatters can talk with their friends and family. This is also where Snapchatters can select to interact with My AI (our opt-in chatbot powered by OpenAI's ChatGPT technology). The Chat screen will show when both friends are there at the same time, as well as indicators when a friend has opened and viewed a Snap or Chat. Snaps and Chats delete-by-default to mirror real life conversations, although Snapchatters can opt to save Chats by simply tapping on the ones they want to maintain, and we are launching new features in the near future that allow Snapchatters to select indefinite retention.

Snap Map: On Snap Map, Snapchatters can view Snaps from all across the world, including sporting events, celebrations, breaking news, and more. Snapchatters and their friends can also choose to share their locations with each other and see what is going on around them.

Stories: The Stories tab is where Snapchatters can find their friends' Stories, Publisher Stories, User Generated Content (UGC) Stories, and more. Stories are collections of Snaps, which typically last for 24 hours and play in the order they were taken. The 'Discover' section has content and recommended Stories from publishers, creators, Snap Stars, and the broader Snapchat community.

Spotlight: Right next to Stories is our dedicated entertainment platform, Spotlight. This is where Snapchatters can submit and watch short, fun, and creative videos for our community.

One of our highest priorities is **protecting the safety, privacy and wellbeing** of Snapchatters whilst ensuring they have a positive experience online. We have clear Terms of Service and Community Guidelines which prohibit the use of Snapchat for any illegal or harmful activity and we have made deliberate design choices from the outset (guided by safety- and privacy-by-design principles) to help prevent the spread of harmful and other violating content. For example, we do not have an open newsfeed or the ability to live stream; and the public areas of Snapchat – Spotlight and Discover – are largely curated or pre-moderated environments before content can reach a large audience.

We deploy strict privacy measures, including settings that are 'off-by-default'; safeguards to help stop strangers from finding and contacting younger Snapchatters (e.g. private friends lists, pop-up warning if they don't share any mutual friends); and parental tools - Family Centre - to give parents insights on who their teens are talking to and prompt conversations about healthy online habits. We also offer easy to use reporting tools and remove content that violates our policies or the law.

Summary

Overall, we welcome the guidance and many of the proposals set out in the consultation – these align with our policies and approach to safety, and have been implemented on Snapchat as standard practice for some time.

Our substantive response below provides detailed assessments of the consultation proposals or Ofcom's consideration. For the purposes of this summary, we have drawn out the key themes from our response where we either require further clarification and guidance or disagree with the Ofcom position:

- We welcome the consultation's recognition of risk profile and cost implications for companies in implementing some of the suggested proposals. However, the consultation only focuses on the largest companies and small or medium businesses impacted by the Online Safety Act (OSA). There is **no recognition for challenger companies**, like Snap, which do not quite fall within these definitions [REDACTED/CONFIDENTIAL]. We would welcome broader criteria beyond user-size when considering the size or definition of a company.
- Linked to this point, we note that in several parts of the consultation a number of fundamental safety design decisions or processes have been reserved for the largest services only on the premise of user-base size (over 7 million users in the UK, which we believe still allows for a sizeable user base as a smaller service). **We do not believe that lower user numbers should be a factor in whether or not the requirements apply - ultimately the question of whether a measure is necessary should be determined by the risk.** Failure to apply the proposals to all services would allow irresponsible design to be contemplated at an early stage in a service's lifecycle – the antithesis of the Act's aims on safety-by-design. The public is likely to see the most benefit to online safety when responsible practices are adopted across the industry.

- In addition, this **will allow smaller services (with a notable user-size) to gain a competitive advantage** [REDACTED/CONFIDENTIAL] through reduced protection and safety cost. More broadly, we note that the impacts on competition have not featured in Ofcom's consideration when making its proposals. This is a worrying omission that forms part of Ofcom's wider regulatory remit and should be addressed
- We have found that in some parts of the consultation, **Ofcom's position has strayed from the OSA's focus on a systems and processes approach**; instead focusing on specific illegal content. For example, there are detailed requirements on tackling online fraud content (particularly online financial fraud) with the use of keyword detection and dedicated reporting channels/trusted flaggers. Arguably, these processes should underpin platform efforts to help tackle [REDACTED/CONFIDENTIAL] illegal harms and should be established as such (as Snap has done) rather than across one harm type where the risk may not be appropriate or proportionate to the ask [REDACTED/CONFIDENTIAL].
- We welcome Ofcom's attempts to ensure **no one-size-fits-all approach is taken and platform design and differentiation is accounted for**. However, there are still many proposals in the consultation which do not appropriately consider this and could lead to disproportionate outcomes for some services. For example virality as a prioritisation factor for content moderation; restoration of content following an appeal determination; providing users an ability to block all unconnected users accounts – these are all proposals which may not be necessary or applicable in the context of Snapchat's platform design.
- We also welcome alignment with other online safety regimes; namely the EU's Digital Services Act (DSA), which shares several similarities to the UK's OSA (e.g. in relation to risk assessments and reporting and complaints processes). However, **more can still be done to ensure closer harmonisation between regulations** (e.g. on Terms of Service) to reduce compliance burdens for services and achieve consistent online safety standards at a global level for the good of citizens. With Ofcom as the current chair of the Global Online Safety Regulatory Network, we note some of this work is already underway and would appreciate clarity on how this will be achieved.
- We understood that Ofcom would take learning from its experience as the regulator on the Video-Sharing Platforms (VSP) Regulation. However, we are dismayed to see that **key learnings from the VSP regime have not helped shape the proposals in the consultation.** As a VSP, we have highlighted notable examples in our response where these findings could offer better clarity and guidance to implement the online safety regime.
- We **disagree with the recommendation to exclude children from network expansion prompts and from receiving them if they have been designed responsibly and enable connections to people the child is likely to know in real-life** (as opposed to strangers). There is an abundance of research that demonstrates children use online platforms, like Snapchat, to connect to people they know. These types of features support their feelings of belonging and happiness.

Detailed response

To aid Ofcom's consideration of our response, we have consolidated our feedback according to each chapter set out in the main consultation document.

1. Governance, risk and accountability

Governance and accountability

Do you agree with our proposals in relation to governance and accountability measures in the illegal content Codes of Practice? Please provide underlying arguments and evidence of efficacy or risks to support your view.

We generally agree with Ofcom's proposals in relation to governance and accountability in Chapter 8, Volume 3; Sections 3A-3G. There are some areas, however, where we would like further clarification. In particular, it would be useful for Ofcom to clarify who it would envisage would be the person accountable to the governance body for compliance for the purposes of Section 3B and the staff who make decisions for the purposes of Section 3C. We believe the intention is that: (a) the 'person accountable to the governance body for compliance' is equivalent to the EU's Head of Compliance role (and could be performed by the same individual); and (b) the 'staff who make decisions' are the most senior Product/Engineering and Operations managers responsible for deciding on and implementing the risk mitigation measures. We would be grateful if Ofcom could clarify further whether our understanding of the intention for these Sections of the Code in this respect is correct.

Do you agree with the types of services that we propose the governance and accountability measures should apply to?

No, we do not consider it appropriate to limit governance and accountability measures to large and multi-risk services. We believe that all user-to-user services should face the same governance and accountability measures in accordance with their risk profile. The number of users, particularly such a large threshold of 7 million, should not be the qualifying criteria.

There's evidence for this in the General Risk Factors (Annex 5), which state that:

- *Low capacity or early-stage services may increase the likelihood of different illegal harms as they may have limited technical skills and financial resources to introduce effective risk management.*
- *A fast-growing user base may negatively affect effective risk management, given the increased scale and sophistication of the moderation technologies and processes required to keep track of a fast-growing user base (particularly since the sources of risk can change quickly as the user base develops).*

We agree with these general risk factors and they particularly arise with smaller services.

It is also important to recognise that the requirements of the OSA will inevitably have an impact on the speed of online service innovation and development, as companies will be concerned to ensure they have met all the accountability steps before they make changes or release new products. In the online world, where popularity of services can change very suddenly, this could be a competitive advantage where a company with a sizable user base is able to gain a significant foothold before having to put in place equivalent governance and accountability measures as a 'large service'. We have raised these concerns in respect of the DSA and its thresholds in the EU.

Are you aware of any additional evidence of the efficacy, costs and risks associated with a potential future measure requiring services to have measures to mitigate and manage illegal content risks audited by an independent third-party?

We and other 'Very Large Online Platforms' (VLOPs) are subject to external audit under the DSA. While it remains to be seen whether such a measure will have an impact on efficacy and risks as the audits are only just beginning, we are clear that such a measure is very expensive. [REDACTED/CONFIDENTIAL] Preparation for such an audit also requires considerable time and resources from frontline teams (such as

Trust & Safety Operations and Law Enforcement Operations), which is time that would arguably be better spent having them focus on their primary functions.

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?"

We do not have first-hand evidence regarding the efficacy, costs and risks associated with explicitly tieing remuneration for senior managers to positive online safety outcomes. We would note that digital services are not the same as financial products, and while a company may be able to produce data regarding complaints or violating content, it is harder to objectively determine online safety outcomes. In addition,

We are aware that this measure has been used in the financial and other sectors. Our understanding is that companies have found it particularly difficult to implement[1]. Compliance issues may not come to light for some time after remuneration has been paid, and clawing back compensation from responsible employees can be challenging and expensive (where lawful). We would be concerned that such drastic measures could have a chilling effect on collaboration, transparency and employee mental-health/well-being. For example, a recent documentary highlighted how UK water company executives are incentivised not to report sewage discharge issues for fear of damaging their remuneration[2].

We would ask that Ofcom also take into consideration the potential for abuse of such a measure. While the positive intent is clear, it might be used in a malicious manner. For example, when users complain about enforcement decisions, companies spend a lot of time reviewing the matter and checking that they have taken the right enforcement action. In some cases, even though it is clear the right enforcement action has been taken and this has been communicated, users can still feel very aggrieved despite being in the wrong. In a small number of these situations, this has led to users becoming bad actors who spend a lot of time looking for ways to 'punish' the company. In situations like this, a measure to tie remuneration could encourage abuse. It could create a perverse incentive for bad actors to try to generate negative online safety outcomes (for example, by flooding platforms with harmful content) because they know that this will directly impact the remuneration of the company managers responsible for correct decisions taken against them. In addition, we would also be concerned that a measure to directly tie remuneration to online safety outcomes could have a chilling effect on privacy, the freedom of information and expression, because platforms will be incentivized to only allow the most plain or blandest content on their platform to be certain of achieving subjective positive online safety outcomes.

Ultimately it is the company that is responsible for ensuring compliance and it is the company (and therefore its shareholders) that should be penalised for failures to comply. It is worth noting that many of the employees working on safety measures in U2U services are also shareholders in these companies and therefore already have an indirect incentive to keep the platform safe.

Risk assessment and review duties

Do you agree with our proposals? Please provide the underlying arguments and evidence that support your views.

[1] See for example: https://legalbriefs.deloitte.com/post/102ie00/compliance-sanctions-and-adjustments-to-executive-compensation
[2] See for example: https://www.thetimes.co.uk/article/joe-lycett-how-i-tackled-britains-poo-problem-with-toilet-humour-llfmc2tx8

We agree with Ofcom's proposals and we are delighted to see that they closely align with the risk assessment methodology that we have used for the risk assessment obligations under the EU's DSA.

However, we have a few areas in where we disagree or seek further clarification:

- An area in which Ofcom's proposals differ from the DSA is the extent to which a significant change to a risk profile can trigger a need for companies to update their risk assessments. It would be helpful to explore with Ofcom further the situations in which it envisages this would occur and in particular, the timelines which Ofcom would expect organisations to conduct an ad-hoc review in response to a significant risk profile change.
- Ofcom identifies a number of examples of the kinds of changes which it expects to amount to a significant change to a service and trigger the duty to review or carry out a new risk assessment. We do not agree with several of these factors and would recommend their removal from the guidance, as follows:
    - Most changes to a service will apply to a substantial proportion of a services' user base. As a result, we do not believe this is a significant factor in determining whether a change is significant. For example, a change to a font colour could apply to a substantial proportion, if not all, of a service's user base. However, unless that font colour will have a material impact on access to illegal content, it is unlikely to be significant for the purposes of an illegal content risk assessment.
    - Similarly, most changes to a service that is accessible to children will inevitably impact children. For the same reasons as above, it should not be the case that just because a proposed change impacts children, this should be considered a significant change. To continue the above example, unless a change in font colour will have a material impact on a child or other vulnerable user's access to illegal content, it is unlikely to be significant for the purposes of an illegal content risk assessment.
    - While there are fewer changes that impact a service's revenue model, its growth strategy and/or its ownership that affects its design, we would again query why these factors alone are sufficient to consider a change to be significant. It is possible, for example, that a change in ownership could lead to a change that affects service design without impacting the risk of illegal content appearing on a service. For example, a new owner might remove certain functionality from a service or redesign the user interface with different colours. As above, we believe the key factor in determining whether such changes are significant is whether they impact the risk to users from illegal content on the service.
- It would be helpful if Ofcom could provide a mapping of the OSA illegal content categories against the illegal content categories in other similar legislations (for example, the categories set out in articles and recitals of the DSA), so that international companies can have confidence that a single risk assessment could be used for many common illegal content categories, and minimise the aspects that are required for a single jurisdiction. We recognise this may be more appropriate for the Global Online Safety Regulators Network but Ofcom has the opportunity to advance this alongside the OSA implementation as the current chair of the forum.
- With respect to the risk assessment evidence base, we are concerned by the expectation to consult the enhanced input list A5.11: *Expectation to consult a list of enhanced inputs even if clear evidence that no/limited harm is occurring*. Some enhanced inputs are challenging and create legal privilege issues and if we can reasonably conclude from the core inputs that no or limited harm is occuring, then this should be sufficient.

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

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?

Yes, this four-step approach is the approach we have taken for our DSA risk assessment and report. We query though the:

- The circumstances in which we would need to update risk assessments in the event of a change to Ofcom's Risk Profiles and the timelines for this (as above);
- Risk level table on page 23 (A5.73) - confusion between likelihood and severity of harm e.g. "broad scope" (see below our responses on Content Moderation Prioritisation); and
- [REDACTED/CONFIDENTIAL]

Are the Risk Profiles sufficiently clear, and do you think the information provided on risk factors will help you understand the risks facing your service?

Yes. We would primarily use the risk factors as a starting point for likelihood consideration but the reality is that 'social media' platforms are very different so aside from noting the risk factors - we would immediately switch to looking at what the evidence on our platform shows. [REDACTED/CONFIDENTIAL]

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

We are aligned with Ofcom's proposals in its draft record keeping and review guidance. We expect the most significant burden will come from updating our holistic digital and data risk assessments (which include DSA risk assessments) to demonstrate that we have assessed Ofcom's specific risk factors and address its mitigation/alternative mitigation requirements. In other words, our risk assessments and reports for the areas we have already assessed for the DSA will still need to be revised to ensure we are in a position to demonstrate compliance with the OSA, even if the substantive content remains the same.

We also also have some queries:

- We use Google documents to record our risk assessments. These automatically record changes and version dates. We assume this will be sufficient to meet the requirement that "the written record should be dated when it is made and on each occasion it is updated".
- The retention period for the OSA and the DSA are not aligned, being 5 years and 3 years respectively. As we will maintain holistic risk assessments incorporating all compliance requirements, this will need to set retention at the longest period for all of our documentation. It would be helpful to understand the justification for a longer retention period than the DSA.
- Where we are meeting the requirements of the relevant code of practice but have additional mitigation measures, would Ofcom consider these as 'alternative measures' under the guidance?

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?

Yes, and if Ofcom has made this decision, why are you not adopting a broader approach to governance and accountability for all services?

2. **Content moderation**

Do you agree with our proposals? Please provide the underlying arguments and evidence that support your views.

[REDACTED/CONFIDENTIAL]

We welcome the broad approach Ofcom has set out to content moderation processes in Chapter 12, Volume 4, which recognises that a one-size fits all approach would not be appropriate and flexibility should be given to services to establish effective systems in line with their platform design and functionalities.

Content moderation systems

We agree with Ofcom's recommendations that the systems or processes that services choose to put in place must be designed in a way that ensures illegal content of which the service is aware is removed swiftly.

We also agree that 'swift' removal should not be tightly defined, as this will depend on the circumstances. At Snap, we operate internal performance targets, which include the timeliness in which reports are enforced or taken down (see below) and our Transparency Report provides median turnaround times per harm type. However, we recognise that there will be circumstances in which a 'swift' removal can vary (e.g. complexity of the content involved/edge cases; law enforcement cooperation) and this is baked into our content moderation systems and processes.

Ofcom specifically states that such systems should:

- *'make an illegal content judgement in relation to the content and, if it determines that the content is illegal content, swiftly take the content down; or*
- *where the provider is satisfied that its terms and conditions for the service prohibit the types of illegal content defined in the Act which it has reason to suspect exist, consider whether the content is in breach of those terms of service and, if it is, swiftly take the content down.'*

We have concerns about the first option and would not recommend this as an appropriate course of action: the implication that each piece of content would require legal analysis and judgement by services raises human rights concerns. These concerns were previously recognised by the courts and struck out in the context of Germany's NetzDG (Network Enforcement Act) and France's Social Media law. It was found that judicial authorities (with the consent of governments) should be the arbiters of whether content or online speech is legal or illegal. We suggest Ofcom applies the same principle. Not only would this practice be cost prohibitive for services, but judgments made by services would necessarily be speculative. In order to comply with this requirement, online service providers would err on the side of over-reporting, thereby chilling freedom of speech and infringing on their users' privacy rights. Even with Ofcom's proposed mitigations (setting performance/accuracy targets, training, and appeals), it is likely that there will be over-enforcement and over-removal of content that is not illegal.

Snap's position is aligned to the second option but we would welcome some operational clarification on the level of granularity that would be required as part of the service's terms and conditions. For example, would it suffice that the terms and conditions cover the relevant illegal harms by subject with a recognition that users must follow local laws, including UK laws? Please refer to our comments on Terms of Service

where we have asked that they remain high-level to support user understanding on the proviso that these are supplemented with more detailed materials that users can easily access, including on specific illegal harms. Furthermore, Snap's Terms of Service apply globally – there is a version that applies to those based in the U.S. (Snap is an American company) and a version for those based outside of the U.S. (i.e. rest of the world). It would be inappropriate to cover each illegal harm in granular detail under the UK legislation given that these may not apply to every jurisdiction in which we operate outside of the U.S.

We recommend that in the context of content moderation, Terms of Service also remain sufficiently high-level regarding illegal content prohibited on the service to maintain overall consistency of approach. This should be supplemented with detailed, internal content policies and guidance on what is and what is not allowed on Snapchat (as noted below).

Internal content policies

We agree that service providers should have internal content policies specifying what content is or is not allowed on the platform, as well as how they should be operationalised and enforced [REDACTED/CONFIDENTIAL]. We also agree that service providers should not be obligated to publish these internal policies due to the risk of essentially training potential violators on how to evade enforcement and game our systems.

However, linked to performance targets below, there should be stronger recognition by Ofcom in its consultation of the need to also assess the effectiveness of services' internal policies and guidance for moderators. Ofcom's VSP report 'What we've learnt about VSPs' user policies' reviewed this area as part of the regulation and highlighted Snapchat's analysis of moderators' performances to test the effectiveness of its internal policies and guidance as an example of good practice.

Performance targets

We generally agree with Ofcom's recommendation of setting time-based and quality/accuracy-based targets for content moderation teams. The key will be in striking the correct balance between timeliness and accuracy of decision-making, which can be difficult and somewhat of a moving target for services as they constantly adjust to perpetually dynamic external factors (e.g. elections, wars) and new threat profiles (e.g. financial sextortion) as well as the evolution of their own products and user base. It is also important to note that bad actor behaviour tends to evolve rapidly. Therefore, quality/accuracy-based targets also need to account for content moderation teams evolving their own responses to this behaviour. Despite Ofcom's suggestion that performance targets could mitigate the risk of stifling free speech, there is still a risk that services may over-enforce and remove legal content to hit time-based targets while erring on the side of caution, especially during times of acute pressure.

We agree that to mitigate these risks, it is important to quality assure the targets and ensure they remain effective. The level of content reinstations may also be used by services to assess effectiveness (although as noted below, this may be less applicable to Snapchat given the ephemeral nature of content) [REDACTED/CONFIDENTIAL].

Prioritisation

We agree with Ofcom's recommendation that service providers should operate a content moderation system and (where relevant) set out certain factors to which they should have regard when designing

their content moderation systems and processes. The factors articulated by Ofcom, however, (virality of content, severity of content, and likelihood that content is illegal) require further clarification.

Whether virality should be a factor in content moderation prioritisation will depend on the nature and design of the platform and this should be reflected in Ofcom's final guidance. Virality is a much lower risk on Snapchat than some other platforms because Snapchat does not offer tools that allow for quick and unvetted 'reposting' or 'resharing' – this is a conscious design decision that vastly reduces the risk of harmful and other violating content spreading rapidly across the platform. Not offering these tools reduces engagement, but we believe this is the right trade-off to improve the safety of our platform. Offering 'resharing' or 'reposting' functionalities is one of the key factors that allows information to spread rapidly across a network, and these features reward sharing polarising content that create outreach, division and harm. The features are often overlooked or taken for granted, but have major safety implications, and we urge Ofcom to take those design decisions into consideration. Also, Snapchat was not designed to broadcast user-generated content – we have no open newsfeed or option to live-stream content (see Introduction for more details). Snapchat is primarily used as a messaging app between friends and content on our public surfaces, such as Spotlight and Discover, are subject to strict content moderation processes before they are able to reach a large audience. This further limits the virality of harmful or illegal content on our platform.

Severity of the content is likely to vary from service to service, depending on platform design, product offerings and user demographic, among other factors. While we agree that severity, generally, should be a factor in prioritisation, that severity may not align with Ofcom's priority illegal harms. For example, a harm may be considered "severe" or "priority" by Ofcom, but have almost no prevalence on the service. We recognise that this consideration alongside *'likelihood that content is illegal'* will be best informed by the risk assessments of services to ensure proportionate prioritisation as part of a content moderation system. This includes harms outside the list of priority illegal offences which may have a high-severity or high prevalence on the platform.

We agree with Ofcom that signals, such as reporting or complaints, including from Trusted Flaggers, can be a clear indication on the likelihood that content is illegal and should be prioritised for removal. At Snap, we indeed use such signals to help inform our content moderation processes alongside other signals, including user and/or account data (depending on the harm). However, we are mindful of the weight Ofcom places on Trusted Flagger complaints. While we agree that information from Trusted Flaggers tends to be high-quality, it does not always signify that the harm is widespread on the platform and should be prioritised. [REDACTED/CONFIDENTIAL] We suggest that Ofcom considers broader criteria in making this assessment.

Resourcing

We agree that it is sensible to consider what languages are common in the regions in which we operate and to ensure language expertise as part of our moderation efforts. We also agree that for certain external events, such as elections, it may be possible for service providers to plan ahead and anticipate an increase in demand for content moderation. More often than not, however, external events such as wars or pandemics, are largely unpredictable. In these cases, we can rely on playbooks and past learning to inform our critical response, including whether additional resources are required, but each case needs to be assessed on its own merits and the situation can be fluid. [REDACTED/CONFIDENTIAL]

It should be noted that the pull on resources usually not only impacts content moderation but business-wide resourcing, as a cross-functional approach is often vital to ensure the response is appropriate, robust and consistent. This can mean a re-prioritisation of resources rather than new resourcing. [REDACTED/CONFIDENTIAL]

Ofcom should also be mindful of quantity versus quality (as noted in the proposals for performance targets) and how different platforms will require different moderation and resourcing levels (as set out in 'Prioritisation' above).

Provision of training and materials to moderators

We appreciate Ofcom's observation that there is no universal best practice for training content moderation teams. What will be effective for each service depends on the nature of the platform, products offered and user demographic, among other factors. We agree that it is sensible to consider a service's own risk assessment and any information divulged pertaining to signals of emerging harm. We also agree that it is sensible to develop a feedback loop between content moderators and policy developers and subject matter experts to remedy any gaps in moderation staff' understanding of specific harms. [REDACTED/CONFIDENTIAL]

We also note that providing real-time guidance, advice and support to content moderators is something that service providers already do, including Snap, and will continue to do. Training is not static, but is a constant, iterative, calibration. The estimated costs provided by Ofcom would seem to omit these costs and efforts, which are significant.

3. Automated content moderation

Do you agree with our proposals? Do you have any views on our three proposals, i.e. CSAM hash matching, CSAM URL detection and fraud keyword detection? Ofcom would like the underlying arguments and evidence that support your views.

On the whole, we agree with the reasoning and proposals that Ofcom sets out in Chapter 14 of Volume 4. We believe these are proportionate and afford reasonable flexibility to platforms in applying the safeguards proposed. To a large extent these reflect the measures that Snap already employs on Snapchat through CSAM image scanning and hash matching with PhotoDNA and Google CSAI, and proactive detection methods that include URL reputation service flags. [REDACTED/CONFIDENTIAL]

However, we have concerns with the proposal on fraud keyword detection. While our processes align with those proposed by Ofcom, we believe the use of keyword detection or an 'Abusive Language Detection' (ALD) tool (as in Snap's case) should be used more broadly to support service efforts to detect and moderate illegal content – not only fraud, and in particular, articles for use in frauds (i.e. stolen credentials). Please refer to our comments in the *Summary* and *Dedicated fraud reporting channel* parts of this response on ensuring Ofcom's measures meet the aims of the OSA – it promotes a systems and processes approach rather than focusing on specific pieces of illegal content (with the noted exception of CSAM). While Ofcom's research notes that fraud is the most commonly experienced illegal harm, this will not be true for every service and should not be the justification for requiring the use of keyword detection on a specific harm type. Ofcom also states that the effectiveness and accuracy of standard keyword detection is another reason it is not being recommended as a broader process on other illegal content.

However, we believe this assessment should be at the discretion of the service (Ofcom acknowledges that many services already use keyword detection) and is proportional to the risk(s) on the platform.

Notwithstanding this, fraud is a complex issue and can relate to many types. A service may consider itself to be at high or medium risk of certain types of fraud (e.g. romance scams) in its risk assessment but this risk may not be associated with articles of fraud (e.g. stolen credentials). It would therefore be unreasonable and disproportionate for services to adopt the keyword detection measure. It is not clear how Ofcom would distinguish these nuances in risks. Further guidance and clarification from Ofcom would be welcomed on this point. [REDACTED/CONFIDENTIAL]

Do you have any comments on the draft guidance set out in Annex 9 regarding whether content is communicated 'publicly' or 'privately'?

We believe this recommendation is aligned to Snap's approach and is appropriate in striking a balance between advancing the safety of users while honouring their privacy rights with regard to their direct communications with others.

4. Reporting and complaints

Do you agree with our proposals? Please provide the underlying arguments and evidence that support your views.

Complaints

We agree with the majority of Ofcom's recommendations set out in Chapter 16 Volume 4 in relation to complaints handling:

- all user-to-user services have complaints processes which enable UK users and affected persons to make certain types of relevant complaints in such a way that the service will take appropriate action;
- complaints processes for all types of relevant complaints must be easy to access, easy to use (including by children), and transparent; and
- there are processes for content prioritisation and performance targets to handle complaints about suspected illegal content.

We welcome the flexibility afforded by Ofcom to establish the most appropriate and proportionate approach in line with Snapchat's platform design and risk profile rather than mandating specific requirements. We believe this is the right approach to ensure our users are given appropriate tooling to support their user experience and interface.

At Snap, we have already assessed and updated our reporting and complaints system on Snapchat to ensure users can exercise their rights to raise an issue or appeal a decision by virtue of the requirements under the EU's DSA. We provide users and non-users with reporting options in-app and on our website. This includes facilitating the 'relevant' types of complaints listed by Ofcom in the consultation and meeting the high-level requirements around accessibility and findability. We confirm receipt and communicate the result of our decisions. When we take action against content or an account, we provide a statement of reasons and instructions on how users can appeal the decision if they believe it is incorrect. We have

dedicated teams in place to review reports and handle appeals, and have an internal prioritisation framework and performance targets for handling both reports and appeals. Users can further escalate via the dedicated contact point on our website to our Head of DSA compliance, whose team may conduct a further review. This area is a good example of where Ofcom has sought to align the UK's legal requirements with other online safety regimes to ensure the burden on companies is reduced.

Indicative timelines for complaints

We agree with Ofcom's conclusion that complainants should receive an acknowledgement that their complaint or report has been received as opposed to a detailed approach, which would include proactive status updates. However, we have concerns on the requirement that an indicative timeline is provided to the complainant. Snap takes user complaints and reports very seriously and has internal goals and metrics regarding turnaround times but it would be very difficult to adopt a one-size-fits-all approach on a timeline for user complaints or reports. This is because some complaints are resolved very quickly, while others are more complex and take longer to resolve correctly. Operationally, this requirement would likely result in the message being sent to the user with the longest possible time frame, as Snap cannot assess ahead of time how complex of an issue the user is reporting.
To ensure that services remain accountable for responding to user complaints and reports in a timely manner, we suggest including turnaround time metrics in transparency reporting obligations (which Snap already does). Alternatively, we suggest that if a complaint has not been resolved within a reasonable time frame (as decided by the service provider), we provide the complainant with an update, including if we need more time to handle their case.

Appeals

We agree with a number of Ofcom's recommendation in relation to appeals processes of content moderation and enforcement, as measures that Snap has already adopted on Snapchat (see above):

- Service providers deal with appeals of content moderation and takedown decisions promptly.
- Service providers should set their own internal metrics and performance targets for appeal timeline and accuracy decisions and adhere to them.
- Service providers notify users of their rights to bring proceedings for breach of contract if the use of proactive technologies has been used to takedown content or restrict access to content in a way inconsistent or not contemplated by our terms of service.
- Service providers establish a triage process for relevant complaints, with a responsible team leading this process and ensuring that complaints reach the most relevant function or team.

Prioritisation of appeals

Ofcom puts forward criteria which may be used by service providers to prioritise appeals. We agree with the proposal that service providers should prioritise certain appeals based on the severity of the action taken against the user. For example, at Snap this could mean prioritising an appeal from a user whose account has been locked over one who simply received a warning. We also consider that other baseline criteria should be used to inform prioritisation, such as the date/chronology of when an appeal was submitted to avoid backlogs.

Turning to the other proposed criteria, we have some concerns:

- *Whether the decision that the content was illegal content was made by proactive technology:* this involves a moving target, as proactive technology in detecting illegal harm content is constantly improving and changing. Therefore, developing a prioritisation system for appeals based on the accuracy of proactive technology would be extremely difficult and require significant resources to constantly update it to real-time.
- *The service's past error rate in making illegal content judgements of the type concerned:* as above, this creates similar operational challenges and would require constant updating; diverting vital resources from working on the technology and/or moderation and appeals [REDACTED/CONFIDENTIAL].

Action following appeal determination

We largely agree with Ofcom's recommendation that if a service provider reverses a content moderation decision on the grounds that the content was not, in fact, illegal, that the provider should, if necessary, adjust the relevant content moderation guidance and its automated content moderation technology to prevent such errors in the future. However, the unique nature of Snapchat and the fact that content is available for only a short period of time before it deletes by default makes it difficult for us to "restore" content that was erroneously taken down. We urge Ofcom to consider platform differentiation as part of their proposals to ensure the follow-up action is appropriate.

Dedicated fraud reporting channel

As set out in the *Summary* above, we have concerns with the proposals recommended by Ofcom in this section of the consultation. The premise of the OSA is for service providers to demonstrate a duty of care to mitigate the risk of harm to their users by adopting a systems and processes approach, rather than focusing on specific types of content. Yet this part of the consultation places full weight on one specific harm type: online fraud (outside of the codes of conduct stipulated by the Act). While we note Ofcom's research on the increasing risk of online fraud in the UK and the engagement challenges encountered between external organisations and service providers, we do not believe this justifies a diversion from the legal position and the prescriptive requirements to tackle online fraud - compelled further with the specific focus on financial fraud when online fraud is a complex issue and can encompass many different types.

We urge Ofcom to take a more proportionate approach to maintain the integrity of the legislation. A Trusted Flagger programme or Dedicated Reporting Channel should be deployed as an effective system to help service providers tackle illegal content and understand any emerging threats on their platforms – not one that prioritises fraud, especially when it may be deemed as medium risk against other much higher-risk illegal harms. Ofcom rightly highlights in its consultation that Snap already operates a Trusted Flagger programme. Our Trusted Flagger programme offers a dedicated reporting route to our Trust & Safety Team and is not limited to one type of harm but spans across and prioritises the most egregious illegal harms to ensure these are handled appropriately.

[REDACTED/CONFIDENTIAL]

Turning to the proposed Trusted Flaggers on fraud, we can confirm that many of these are already Trusted Flaggers at Snap (some of which have been highlighted in Ofcom's consultation). However, we disagree that these should all be onboarded to ensure parity on accessibility and engagement with platforms; rather this should be left at the discretion of the service provider. At Snap, we have a clear policy and process for onboarding and reviewing the status of our Trusted Flaggers. This is assessed on

a case-by-case basis and on criteria including the organisation and their relative expertise; types of harm(s) they wish to report; volumes and risk/relevance on Snapchat. Where we have not onboarded organisations as Trusted Flaggers, it will be because they do not meet this criteria (or they are content with the existing engagement channels and have not requested or see a need for Trusted Flagger status). It should be noted that even when organisations do not have Trusted Flagger status, they are still offered appropriate engagement channels at Snap to discuss matters. For example, some organisations do not wish to raise specific reports but discuss broader trends and/or policies with policy experts in the company.

If we were expected to onboard all these organisations as Trusted Flaggers, this would also raise capacity concerns, as mentioned above, as well as increased operational costs. [REDACTED/CONFIDENTIAL] It would also be helpful to understand whether Ofcom's proposed list of Trusted Flaggers is exhaustive or will be reviewed and added to.

5. Terms of Service

Do you agree with our proposals? Please provide the underlying arguments and evidence that support your views

Substance of Terms of Service

On the whole, we support the guidance set out in Chapter 17, Volume 4. We agree that service providers should clearly and publicly state how they protect individuals from illegal content, including information about the technology used and how it works, as well as the policies and processes that govern the handling and resolution of relevant complaints.

Indeed, Ofcom's VSP report 'What we've learnt about VSPs' user policies' found that Snap's terms and conditions cover a broad range of different types of content that may cause harm to children as an example of best practice. The report also recognised more generally that Terms of Service are necessarily lengthy and detailed, and can often be difficult for users to read and understand. As a result, we would welcome additional guidance from Ofcom on how to balance the level of detail proposed with ensuring that the average Snapchat user understands our terms. As part of this, we believe that disclosures around a service's safety strategy and methodology – including how we protect users and how the technology works – should be kept at a high-level for various reasons:

- To avoid overwhelming the user;
- To ensure the content is at an age-appropriate reading level [REDACTED/CONFIDENTIAL];
- To prevent potential violators from gaming our systems and evading our technologies; and
- To remain flexible and agile to changing factors, including real world events and user needs, so that adjustments can be made quickly as needed.

We note that Community Guidelines provide a more digestible and user-friendly way to explain rules to users using the service. However, these can still be under-utilised and we believe further provisions are needed to support the user experience. This is why Snap provides policy explainers that sit beneath its Community Guidelines and we have a dedicated Safety Centre to ensure clear signposting to support pages or articles and blog posts for users to access the information that is most pertinent to them regarding our rules, products, measures and operations (as opposed to having to scroll through long

documents/pages to surface the information they need). This is also complimented by in-app notifications and resources for users' awareness and understanding. For example, we have a Safety Snapshot series which educates users in an interactive way (video) about illegal harms related to sexual content and how they can report. Some of these provisions are in direct response to the DSA (e.g. policy explainers) and it would be helpful if Ofcom can consider alignment in this respect to ensure a consistent approach to Terms of Service that best serves the needs of users and manages the compliance burden for services. In other words, where the DSA and OSA have similar requirements, we would ideally meet them with the same substance for both jurisdictions.

Clarity & accessibility

We agree with Ofcom's recommendations that services provide clear and accessible provisions by considering the following factors:

i. Easy to find: i) clearly signposted for the general public, regardless of whether they have signed up to or are using the service; and ii) locatable within the Terms of Service;
ii. Laid out and formatted in a way that helps users read and understand them;
iii. Written to a reading age comprehensible for the youngest person permitted to agree to them; and
iv. Designed for the purposes of ensuring usability for those dependent on assistive technologies.

Snap already considers these factors in our Terms of Service, Community Guidelines and other associated policies for our users (and those that are not users, such as parents). This also includes feedback from relevant bodies or experts, such as the areas of improvement recommended in Ofcom's VSP report 'What we've learnt about VSPs' user policies'.

More broadly, it would be helpful to understand why Ofcom took a more prescriptive approach on its recommendations regarding clarity and accessibility under this regime than the one proposed under the transposed EU's Audiovisual Media Services Directive (AVSMD). We had also understood that VSP learnings would be used to help inform the online safety regime but we cannot see how these learnings have been considered in the consultation (the Content Moderation chapter is another prime example where the VSP regime offered key findings that have not been reflected here).

We also believe that Ofcom's estimated costs for achieving this level of clarity and accessibility (£16,500 at the high end) are far too low [REDACTED/CONFIDENTIAL].

6. Default settings and support for child users

Do you agree with our proposals? Please provide the underlying arguments and evidence that support your views.

Default settings for child accounts proposal

To note: A user must declare that they are aged 13 or above to join Snapchat. We determine children on Snapchat as those aged between 13-17 and refer to them as 'teenagers'; 'teens'.

We agree that default settings for children should protect them from harm, especially grooming and child sexual exploitation and abuse (CSEA). We support a number of the measures proposed by Ofcom in Chapter 18 of Volume 4 of the consultation:

- We agree that child users should not be exposed in connection lists (e.g. friend lists) of other users, nor should connection lists of child users be displayed to other users. Snapchat by design keeps user friend lists private (for all users, not limited to child users) and only visible to the users themselves.
- We agree that children should not receive direct messages from people they are not connected to ("likely to be friends with in real life"). As recognised in Ofcom's consultation, *'Snapchat already limits discoverability of teen accounts on their platform to people that users are "likely [to] know", such as where there is a mutual connection.'* For all users on Snapchat, there must be mutual acceptance of a friend request or an existing phonebook presence before any communication can begin.
- We agree that user location information should be OFF by default. This is consistent with Snapchat's location-based feature design, Snap Map, where all users must opt-in to in order to activate, and then proactively select which bidirectional friends to share their location information with. We do not allow users to share their location with people they are not friends with.

These measures speak to Snapchat's ethos as the platform for 'real friends' (please see introduction for more information) and how we can support these connections rather than allow strangers to target our users.

As such, we do not agree with the Ofcom proposal that children should be removed from network expansion prompts and receiving network expansion prompts. Our research shows that:

- **Two-thirds of young people** say direct messaging with family and close friends makes them feel extremely or very happy[3].
- Snapchat users report **higher satisfaction with the quality of friendships** and relationships with family than those who do not use Snapchat[4].
- **Over 90% of Snapchat users** say they feel comfortable, happy and connected when using our service[5]

Furthermore, Internet Matters' recent third annual Digital Wellbeing Index report for the UK found that:

- There has been a rise in the positive developmental, emotional, and social experiences of children online from 2022 to 2023. **Two-thirds (65%) of children say spending time online makes them feel at least mostly happy.**
- **A large majority of children continue to agree that digital technology is key for keeping in touch with friends (82%).** It is also clear that digital devices and online platforms are not just about games and videos; **they are often about community, friendship, and support. This year, 60% of children say that being online makes them feel like they're part of a group.**

We believe the evidence supports the need to make available responsible network expansion opportunities to teenage users by default - 'Quick Add', in the case of Snapchat. These are key to the

[3] 2024, Source
[4] 2024, Source
[5] 2022 Alter Agents study commissioned by Snap Inc.

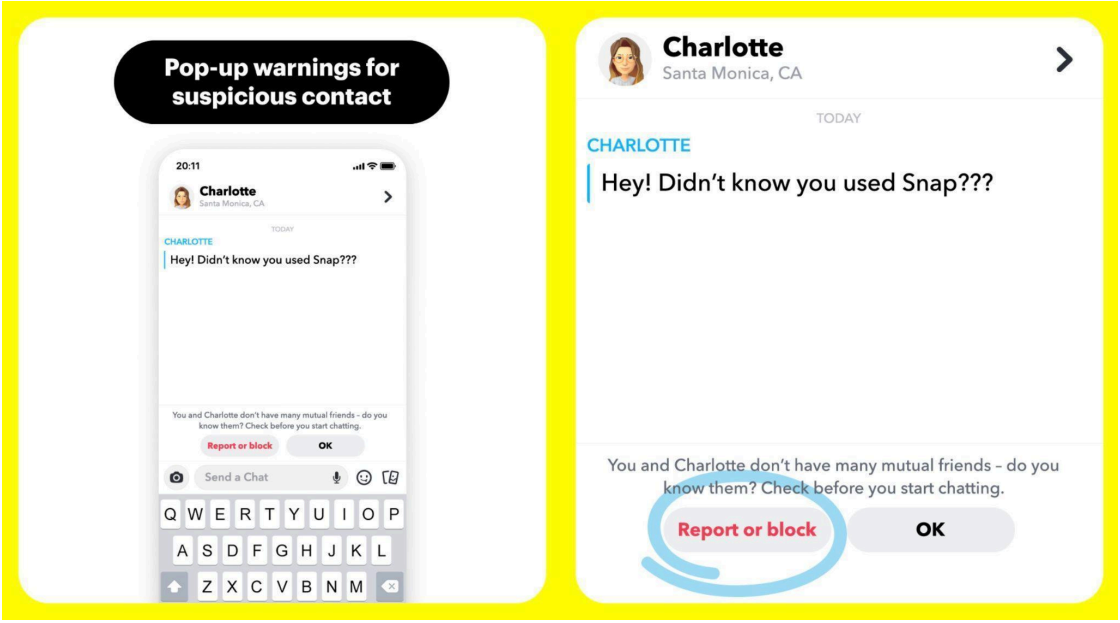
feelings and sense of connection, and children should not be excluded from them. It helps to easily connect them to people they already know (for example, classmates, team mates etc). This is all the more important in light of the 'loneliness epidemic' which recognises that the loss of social connections can have an acute impact on an individual's mental and/or physical health.

Therefore, the Code should allow for older children (i.e. teenager users (13-17)) to be included in network expansion prompts for other users and be presented with network expansion prompts by default – but only where the service has limited these to cases of there being multiple mutual contacts or the contact is already in the users' device (or OS account) address book. This ensures that there are clear parameters in place so that the teen is very likely to know the other user in real life (whether another teen or an adult). [REDACTED/CONFIDENTIAL]

[REDACTED/CONFIDENTIAL] We have a zero tolerance for this type of illegal activity on Snapchat. However, given the strong benefits of reasonable network expansion opportunities to the vast majority of teenage users explained above, we do not agree that a total prohibition on network expansion opportunities is the right measure to set in the Code at this time.

Instead, we believe that the Code should require services that provide reasonable network expansion opportunities to mitigate this risk via an appropriate mix of proactive and reactive measures identified elsewhere in the Code. These should include:

- Ensuring teenage user profiles do not have any personal information associated with them and are not set to public (at least by default). We believe this is important because this information makes it easy for perpetrators to identify teen accounts (described by Ofcom as the 'scattergun approach'). On Snapchat, unlike other platforms, we do not have direct personal information associated with user profiles. For example, the username does not have to be associated with their actual name. It could be anything. Teenage user profiles (13-17) also cannot be set to public on Snapchat (with the exception of a few, well known users).
- Ensuring group chats in small, private social spaces used by teenagers are not public and users cannot search for them in light of Ofcom's research that perpetrators of grooming and CSEA tend to infiltrate such group chats as a way of connecting with younger users and manipulating the algorithm for expansion prompts. It should only be possible to join such spaces or groups if invited by the group chat creator (or other members of the group that have been granted that permission) to make infiltration by strangers difficult. We believe this is important because once part of such group chats, users will usually be able to access the usernames of other members - reducing the impact of responsible limits to network expansion prompts (like Snapchat's Quick Add) to prevent strangers connecting with children. Larger public groups intended for communities, such as colleges and universities, should have eligibility criteria and checks to ensure they are limited to members of that community. To note: Snapchat's small, private social groups are not public and users cannot search for them, and our larger, public social groups have eligibility checks.
- Ensuring services have confidential chat and group reporting tools to raise any issues, and services are able to respond to such reports promptly. Snapchat has such tools and our transparency reports show that we take quick enforcement actions.
- Ensuring that when a teenager is contacted by someone they do not share a mutual friend with (i.e. they may not know), they are automatically sent a pop up warning, asking if they know the person and giving them the option to report or block the individual. [REDACTED/CONFIDENTIAL]



- Ensuring services use proactive signal-based detection to identify and take appropriate action against bad actors to prevent them having the opportunity to target and victimise others in the first place.

We also believe parental tools have a role to play in helping parents to further mitigate this risk. Our website - parents.snapchat.com - provides resources and advice for parents about Snapchat and how teens are using it. Our parental tool - Family Centre - offers valuable insights to parents on who their children are talking to on Snapchat to trigger conversations with their teens on any suspicious activity; or they can easily and confidentially report accounts that may be of concern directly to our Trust and Safety teams, who work around the clock to investigate reports and take quick enforcement actions.

In addition, we understand that grooming (along with other sexual offences, such as financial sextorion) often starts via a seemingly innocent contact offline or on more broadcast-style platforms that still connect (and even recommend) strangers to young people. When those services are required to apply the Code and help prevent contact with strangers, we believe this will have a significant impact on the industry as a whole and reduce the migration of grooming onto responsible, private messaging services like Snapchat.

When you consider the totality of these safeguards, we believe the risks presented by responsible network expansion prompts (such as Quick Add on Snapchat) are relatively low and this measure should not be disabled for younger users. We believe that at Snap we have struck the right balance between offering a positive experience that allows our younger users to connect and deepen their relationships with their friends while helping to protect them from unwanted contact. The significant impact that we believe will result from requiring all services to limit their network expansion to similar responsible levels should be studied before a total prohibition is considered.

Support for child users

We agree that child users should be provided with clear and easy to understand information regarding their choices so that they can make knowing decisions as they engage with different features in the app,

and take any actions that could impact their safety. On Snapchat, we provide users with access to information when they first access certain features and in our support pages. We also provide educational content if, for example, users search for certain terms related to harmful content, such as our Here for You episodes (which would appear if users search for terms like suicide or self-harm and focus on mental health) and Heads Up content (which would appear if users search for terms associated with drugs). These encourage and empower users to report and get support.

We think that further consideration should be given in the Code to striking a balance between providing important information to users at the moment each choice is being made versus part of the user onboarding and education process when their account is created:

- At the risk of creating an overly burdensome user experience with detailed information presented at the point of every triggering decision or event, there should be:
    i. an option for the information provision to occur the first time a user takes the triggering action/experiences the event or there has been a material change/update to the feature;
    ii. an option for users to dismiss future repeated displays due to redundancy, i.e. a “Don’t Show Me Again” button; and
    iii. An option for an infrequent reminder, particularly if a user has not used the service for a while.
- To ensure users have access to important information at all times, platforms should also present easy-to-access resources that summarise important information and provide further details for users' recourse where appropriate. On Snapchat these can be found via our Support Hub.
- There is a high likelihood that providing the amount of recommended information at the point of any of the listed scenarios in the consultation may result in visual (and even cognitive) overload for child users. Therefore, design flexibility and discretion should be explicitly afforded to services with regard to how such information is presented. This is particularly acute when services like Snapchat are predominantly mobile-app based and there are practical implications to display information on a small screen.

Are there functionalities outside of the ones listed in our proposals, that should explicitly inform users around changing default settings?

Location sharing. A child user who elects to change their default setting to share their precise location with another user should be provided with information regarding the potential risk of sharing that information with the selected user(s), as well as reminders:

i. to only share location with friends that they know and trust; and
ii. how to easily disable location sharing with users.

We provide this information to our users of Snap Map.

7. Recommender systems

Do you agree with our proposals? Please provide the underlying arguments and evidence that support your views.

We agree that services whose risk assessment indicates that they are medium or high risk for illegal content relating to priority offences should test for the prevalence of such content and log these safety

metrics. However, we disagree with Ofcom's conclusions as to who the measures would apply to. We agree that there is significant cost in developing such testing services, and it would essentially penalise responsible platforms if this OSA requirement only applies to those that already employ on-platform testing. An irresponsible platform, or new platform, would have a significant competitive advantage over an existing platform if they can avoid the cost of deploying on platform testing.

In our view, neither Option A, B nor C in paragraph 19.47 should be adopted by Ofcom. Instead, the key factor in determining whether user-to-user services should deploy illegal content testing should be based on whether the service has assessed that they are high or medium risk for at least one type of illegal harm identified in paragraph 19.53 of volume 4, or the extent that the risk is unknown. We believe this is proportionate as the costs and complexity of such testing are expected to be proportional to the size of the service and the volume of content being published, as well as the design of the service (i.e. the complexity of a recommender system and whether its design inherently reduces the volume of content needed to be tested). We agree that this measure should not apply to services whose risk assessment indicates that users face a low risk of encountering harms relevant to recommender systems.

We are aware of design features and parameters that can be used in recommender systems to minimise the distribution of illegal content, e.g. ensuring content/network balance and low/neutral weightings on content labelled as sensitive. Are you aware of any other design parameters and choices that are proven to improve user safety?

We strongly believe that all services that offer public content spaces should design them such that they limit the ability for unmoderated content to be shared widely (please see our comments under Content Moderation - Prioritisation). This is the best way to improve user safety. On Snapchat, we offer two main public content platforms – Stories and Spotlight — where Snapchatters can find public Stories and videos published by vetted media organisations, verified creators, and Snapchatters. In these sections of our app, we limit the ability for unmoderated content to be shared widely. We use proactive detection tools prior to publication and additional review processes to make sure this public content complies with our guidelines before it can get broadcast to a large audience. We have proven that such proactive measures, combined with rapid reaction to reports of illegal content that evades proactive moderation, improves users safety. There is no reason why other services should not make similar proactive content moderation choices.

8. **Enhanced user control**

Do you agree with our proposals? Please provide the underlying arguments and evidence that support your views.

We agree with Ofcom's proposals that users of services that have assessed themselves as being at medium or high risk of relevant harm should be offered protection from other individuals (whether or not they are connected to that service), and we support the outcomes that Ofcom is looking to achieve through blocking controls. However, we believe the Code should have greater flexibility in how this protection is achieved because not all services are designed in the same way, and we do not believe this measure should be limited to large services, as follows:

- While we broadly agree that services should provide the ability to block and mute individual accounts, we have concerns in respect of the proposal to provide users an ability to block all

unconnected users accounts. In the case of Snapchat, the service is already designed to prevent unconnected users (i.e. users who are not mutually accepted or in their contact book) from exchanging chats and snaps, and viewing / replying to user stories; it is not possible for users to enable this (please see Default Settings above for more information). We believe this important safeguard is already providing the intended protection and we do not agree that our service should be required to go further as proposed by Ofcom. The Code should recognise that the design of a service may render some measures unnecessary, where the service is able to show that it achieves the intended protection as an inherent part of its design.

- As expressed in various parts of this response, we do not agree that these requirements should only apply to large services. Blocking and muting functionality is a fundamental design decision for user-to-user services and we find it hard to believe that existing and new services do not consider this design choice when developing their service. Failure to apply this proposal to all services would allow irresponsible design to be contemplated at an early stage in a service's lifecycle – the antithesis of the Act's aims on safety-by-design – and will allow those services that do not provide appropriate protections to gain a competitive advantage through reduced protection and safety cost.

We agree with Ofcom's proposal that services should allow users to disable comments on content, and that users be given an easy way to find and use a functionality to disable comments. However, similar to our comment above, we believe the Code should provide greater flexibility in how the desired outcome is achieved. In the case of Snapchat, by default, the user must manually approve any inbound comments before they can appear publicly. We believe this default achieves a stronger protection than that proposed in the Code and effectively prevents any users from commenting on content posted by that user. Also, as above, we believe the obligation to achieve this protection should apply to all services. Again, a decision on whether to provide an ability to pre-approve or disable comments is a fundamental design decision that all services should consider. We do not believe lower user numbers should be a factor in whether or not the requirement applies - ultimately the question of whether a measure is necessary should be determined by the risk, and we do not believe the Code should be inadvertently incentivising irresponsible 'smaller' services (which can still be quite large in practice) with reduced costs and competitive advantage.

We agree with Ofcom's proposal for requiring services to have clear internal policies and better public transparency around notable user verification, including paid-for user verification schemes. As above, we believe this proposal should also apply to smaller services. We recognise that smaller services may be less likely to have a notable user verification or paid-for user verification scheme. However, where they do, the key question should be whether a service is subject to a relevant high or medium risk. While they may have lower reach, 'smaller services' can still be quite large in practice. Bad actors removed from larger services are likely to seek smaller services instead and the public is likely to see the most benefit when responsible practices are adopted across the industry.

Do you think the first two proposed measures on disabling comments should include requirements for how these controls are made known to users?

We think that transparency requirements around how users can disable or block comments would be appropriate. However, we recommend that specific methods of transparency should be at the discretion of the services and not left to prescription within the Codes. Services understand how their users typically engage with their platform, and therefore ought to be best positioned to achieve user transparency in the most effective way.

Do you think there are situations where the labelling of accounts through voluntary verification schemes has particular value or risks?

Risk of misuse and exploitation of verification schemes to impersonate or deceive users and thereby expose them to illegal content can arise whether the scheme is voluntary or paid-for. In each case, the risk can best be mitigated by a thorough and rigorous set of processes that services apply when operating a verification scheme to achieve satisfaction that the user is credibly the notable person that they claim to be.

9. User access

Do you agree with our proposals? Please provide the underlying arguments and evidence that support your views.

We mainly agree with Ofcom's recommendations in Chapter 21 of Volume 4 of the consultation; namely that service providers remove user accounts when service providers become aware that the account is operated by or on behalf of a proscribed organisation or where a significant proportion of a reasonably sized sample of the content recently posted by the user account is proscribed organisation content.

[REDACTED/CONFIDENTIAL]

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

[REDACTED/CONFIDENTIAL]

What are the advantages and disadvantages of the different options, including any potential impact on other users?

[REDACTED/CONFIDENTIAL]

How long should a user be blocked for sharing known CSAM, and should the period vary depending on the nature of the offence committed?

Notwithstanding an appeal being successful and overturning the decision that led to the block, we generally believe that any user who shares known CSAM should be blocked permanently. However, as discussed below, we believe that the ultimate decision to block a user can sometimes depend on context, and that service providers are best positioned to make those determinations. We recommend that any legislation ultimately promulgated by Ofcom recognise that there are some edge cases where some discretion may be appropriate.

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. 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?

As discussed in Section 3 (Automated Content Moderation), Snap and other platforms utilise hash matching technology, specifically PhotoDNA and Google CSAI, to proactively scan for CSAM. These two tools are only able to scan for known CSAM (as opposed to first generation/novel CSAM), via the Internet Watch Foundation's database and the hashes are made available to various companies for this particular purpose. Therefore, when a user posts or shares an image that is picked up by our proactive automated screening system and flagged as CSAM, this means that the user's image contains hashes (which are like a photo or video's unique DNA or fingerprints) that are identical to those in the known CSAM database. The technology is well-established and continuously improving - we are confident that the risk of error here is low. [REDACTED/CONFIDENTIAL]

However, in order to keep the errors low and to avoid situations where lawful content is incorrectly flagged as CSAM by proactive technology, we recommend that any further guidelines by Ofcom on this subject are limited to scanning for known CSAM at this time (as defined as CSAM that resides in hash databases), as opposed to first-generation or novel CSAM. Due to strict U.S. federal regulations on sharing and distributing CSAM materials, service providers are still exploring viable tech solutions to engage in accurate and proactive screening for first-generation CSAM which also do not infringe the privacy rights of the user.

Generally speaking, the vast majority of users who share CSAM on Snapchat are permanently blocked and reported to the National Center for Missing and Exploited Children (NCMEC) in line with US law. However, we agree with Ofcom that there may be times where a user is blocked for sharing something that is flagged as CSAM, but where a permanent block is not the most appropriate option. In these cases, context often matters and an understanding that CSAM can take various forms. For example, there are various "meme" content items that contain elements of CSAM, but that are often shared in jest, without sexual context or intent, often by ignorant and gullible children who do not understand the seriousness of the content they are viewing and sharing. In these cases, we feel that it would not be appropriate to simply lock the user out. Instead, we believe the more productive approach is to delete the content, warn the user, and focus on educating the user as to why such content is harmful, illegal, and cannot be shared. To this end, we agree with Ofcom that a strikes system is a good measure to deploy, and indeed, we utilise a strikes system with our users. However, since some cases are context-dependent, we believe that the discretion to either issue a strike or move to a permanent block against the user in the case of CSAM sharing should be left to the service providers.

We would welcome a reflection by Ofcom on edge cases and providing platforms with some discretion to work within as part of the final guidance.

Conclusion

We hope this response is helpful to Ofcom's consideration when finalising the codes and guidance on illegal harms under the OSA. Please let us know if you have any questions or require additional information and we would be happy to discuss in further detail if required. [REDACTED/CONFIDENTIAL]