Question 1: Please provide a description introducing your organisation, service or interests in Online Safety.

Thank you for an opportunity to provide our response to this evidence call. We are writing on behalf of REPHRAIN, the National Research Centre on Privacy, Harm Reduction and Adversarial Influence Online. REPHRAIN is the UK's world-leading interdisciplinary community focused on the protection of citizens online. As a UKRI-funded National Research Centre, we boast a critical mass of over 100 internationally leading experts at 13 UK institutions working across 37 different and diverse research projects and 23 founding industry, non-profit, government, law, regulation, and international research centre partners. As an interdisciplinary and engaged research group, we work collaboratively on addressing the three following missions:

· Delivering privacy at scale while mitigating its misuse to inflict harms

· Minimising harms while maximising benefits from a sharing-driven digital economy

· Balancing individual agency vs. social good.

We are addressing this call because REPHRAIN researchers are international experts on online safety across domains like privacy-enhancing technologies, misinformation, tech abuse, and many others. For years, we have been working to identify, evaluate and address online harms pertaining to the users of digital technologies across various demographics.

This is a submission from the REPHRAIN centre. Specifically, the following researchers contributed to the formulation of this response (in alphabetical order): Dr Ingolf Becker, Prof Madeline Carr, Dr Alicia Cork, Dr Andrés Domínguez, Chelsea Jarvie, Dr Ola Michalec, Dr Claudia Peersman, Prof Awais Rashid, Yvonne Rigby, Prof Paul van Schaik, Dr Pejman Saeghe, Dr Leonie Tanczer, Dr Honor Townshend, Dr Kami Vaniea, Dr Mark Wong.

As many of our comments and recommendations are based on researchers' work-in-progress, we are very happy to follow up with focused meetings and detailed evidence submission upon your request.

Question 2: Can you provide any evidence relating to the presence or quantity of illegal content on user-to-user and search services?

Several REPHRAIN projects are working to understand the presence of illegal content and user exposure to it:

- The Covid-19 Vaccine Safety project (COVSAF https://www.rephrain.ac.uk/covsaf/) identified the availability of COVID-19 related products (i.e., vaccines, falsified documentation etc.) across various dark web markets (DWMs). 118 DWMs were searched March 2020 - October 2021, showing forty-two listings of mainly unlicensed COVID-19 'cures' and vaccination certificates across 8 marketplaces, sold by 24 vendors. Significant variation in prices was observed, with certificates ranging between US$ 55 and 3,200. The listings were found to be geographically specific and followed the progression of the pandemic in terms of availability. Correlations between vendor portfolios of COVID-19 products and goods of other illicit nature were identified. A positive correlation was shown between the sale of vaccines and illegal weaponry, as

well as between claimed cures and medication/drugs of abuse. Our full findings also shed new light on the vendor motivations, as well as discussing appropriate regulatory measures and targeted interventions. We are happy to provide a paper draft upon request.
- The INTERACT project (https://www.rephrain.ac.uk/interact/) is currently collecting data for a study on how often individuals are exposed to harmful content online, who is exposed to harmful content, the types of harmful content that individuals see and the impact of the content. This study tracks individuals over a 7-day period and aims to provide a more granular view on what is being seen, by whom and the effects it is having. We expect this to be published in early 2023.

In our work, we often draw from internationally recognised organisations to access domain-wide data on illegal content. For example, the following reports provide data on the presence of child abuse material:

- The National Center for Missing and Exploited Children (NCMEC) and the Internet Watch Foundation (IWF) publish annual reports on the prevalence of child sexual abuse material found online (see https://www.missingkids.org/ourwork/ncmecdata and https://www.iwf.org.uk/about-us/who-we-are/annual-report-2021/)
- INTERPOL is investigating the scope and nature of online child sexual exploitation and abuse in their Disrupting Harm project (https://www.interpol.int/en/Crimes/Crimes-against-children/Projects-to-protect-children/Disrupting-Harm).

Question 3: What can providers of online services do to enhance the clarity and accessibility of terms of service and public policy statements?

A number of recent REPHRAIN projects addresses the need for clarity and accessibility of Terms of Service and Public Policy Statements:

- A team of researchers at REPHRAIN (Daniel Kirkman, Daniel Woods and Kami Vaniea) is conducting work on automatically detecting dark patterns in cookie dialogs and the impact large cookie dialog providers have on the patterns as well as on cookie setting behaviours. Our results detail how such patterns hide the actual behaviour of sites and nudge users into providing more information than they might if given a free choice. We are happy to provide a paper draft on request.
- In the recently launched WELL-CONSENT project (https://www.rephrain.ac.uk/well-consent/), we will identify user needs for online consent and ways to design consent forms to those needs.
- The PRIME (https://www.primecommunities.online/) project team is addressing the needs of Minority Ethnic communities in online environments. This will be achieved by making the terms of service and public policy statements available in multiple languages explaining technical terms in language legible to non-technical audience.
- The ongoing work by the PriXR project (https://www.rephrain.ac.uk/prixr/) goes beyond the current online services and draws attention to deceptive designs in augmented and virtual reality applications. We have been running co-design workshops with XR experts and deceptive design experts to investigate the ways in

which deceptive designs may manifest in the context of XR. We are currently drafting this work for publication and happy to provide more details on request.

Based on the work of REPHRAIN experts, we outline the following recommendations:

- Online service providers should adopt a co-design and participatory approach when developing (or revising) their terms of service/policy statements (Wong, 2022), ensuring the needs and priorities of the end-users, especially those who are most vulnerable, are adequately addressed. This should not merely be a consultation (Nesta, 2021) where end-users do not have meaningful opportunities (Alonso Curbelo and Wong, 2020) to influence the decision-making process.
- Policies should be standardised, and put user protection at the front. Users should not have to read every policy document, their rights should be protected by default. Variations should not be allowed or should not require the user to give consent, since that consent is rarely meaningfully given in an online context. We are happy to provide with a draft of our work to date that contains further information.
- Deceptive designs are typically used to reduce the transparency of, and/or accessibility to terms of service and public policy statements. (Sources: "Ease" and "Default Settings" (Forbruker Radet, 2018), "Obstruction" (Gray et al., 2018), and "Hidden Legalese Stipulations" (Bösch et al, 2016)). Deceptive designs are intentionally used because, in some shape or form, their usage benefits the service providers. In this context, a more relevant question may be: *How can service providers be encouraged to enhance the clarity and accessibility of terms of service and public policy statements, especially given that doing this may go against their interests?*

Question 4: What can providers of online services do to enhance the transparency, accessibility, ease of use and users' awareness of their reporting and complaints mechanisms?

We outline the following recommendations:

- Online service providers should adopt a co-design approach when developing their reporting and complaints mechanism, where people who are most vulnerable to harms are meaningfully included in the design process (Szostek and Wong, 2022). Evidence (Eubanks, 2019) shows that historically marginalised groups, such as Minority Ethnic people, are more likely to under-report complaints and harms or have perceived fear of the consequences of reporting. A co-design approach (Wong, 2022) helps anticipate such barriers and to co-design solutions that are inclusive, equitable, fair. In this case, equal opportunity does not mean fair.
- Adding to the point of a co-design approach, different reporting and complaints mechanisms should be available for different user groups. Especially for vulnerable groups, using such mechanisms requires establishing trust first.
- The processes for complaints should be prescribed, alongside deadlines and escalation procedures. There should be APIs for complaints processes so that trusted parties can submit complaints on behalf of individuals automatically.

- We highlight the need for parental insight/oversight technologies in the context of adolescent use of social XR. Our findings suggest that parents need guidance with regards to which intervention methods should be employed to deal with child safety. We are currently drafting this work for publication and are happy to provide more details on request.

Question 5: Could improvements be made to content moderation to deliver greater protection for users, without unduly restricting user activity? If so, what?

First, we outline the following remarks on the topic of automated content moderation:

- Caution must be exercised when considering the use of automated or AI systems for content moderation, which are being proposed by some social media platforms. ML-based content moderation is highly context, culturally and temporally sensitive (Gillespie et al. 2020; Gorwa, Binns, and Katzenbach 2020). Such systems are especially susceptible to discriminatory bias and algorithmic harm (Noble, 2018), e.g. in the training data, prediction models, and risk/individual profiling, and they may amplify existing inequalities, especially racialised and gender inequalities, and disproportionately restrict or incur over-surveillance of racialised communities, for example (Eubanks, 2019). Depending on the specific benchmarks used to train these models, content can be misclassified as harmful/toxic or purposely crafted by adversaries to dodge detection. The degree of automation and whether or not to deploy automation should be assessed not only with regard to standard undifferentiated metrics of performance but with attention to culture-specific and situated community norms, topicality, language, etc.
- Any automated or ML-based content moderation tools should be evaluated before being deployed. Within REPHRAIN, we have a team working on a case study in which we are evaluating five Proof of Concept tools built to automatically detect child sexual abuse material in E2EE environments. The study involved collecting feedback from the community regarding the evaluation criteria. The document describing the final evaluation criteria can be found here.

Second, we have the following comments on the role of people in content moderation:

- The MITIGATE (https://www.rephrain.ac.uk/mitigate/) project is currently exploring the impacts of visible and non-visible moderation in different types of social groups (e.g., the effects of telling individuals that something has been deemed inappropriate versus simply removing the content). We are also assessing the role of the moderator's identity in communicating a message that something is not acceptable. We expect this research to be published by mid-2023.
- As a part of the MITIGATE project, we are also working on the topic of personalisation in content moderation and content reduction. Our upcoming article explores the idea that individuals have different thresholds and sensitivities to different types of legal but harmful content (e.g., nudity, profanity, aggression). Whilst at the moment it is down to platforms to determine where the appropriate threshold lies when protecting users from harmful content, our analysis explores under what circumstances individuals should be

able to set their own thresholds which may differ across different types of harm. For example, users could declare idiosyncratic topics that may cause harm to them, even if they would not cause harm to 'most people' (e.g., an individual who has just suffered a miscarriage may not want to view any parenting related content or advertising). This aligns with the Online Safety Bill's aim to encourage greater user empowerment. Psychologically, this allows users to gain greater agency over not only what they see online, but also what they are able to avoid. The role of the user moves from a vulnerable individual reliant on platforms/government to protect them, into being more equipped to ensure their own safety online. Further, it does not unduly restrict the activities of other users; this approach turns the debate from one of freedom of expression into one about the freedom to be heard by others. The pre-print will be available in late September 2022.

Question 6: Are there any functionalities or design features which evidence suggests can effectively prevent harm, and could or should be deployed more widely by industry?

We recommend deployment of the following:

- Consistency, Feedback, Help/Documents (See our paper: Usability analysis of shared device ecosystem security: informing support for survivors of IoT-facilitated tech-abuse, Parkin et al, 2019)
- Threat Modelling methodology could be used to investigate potential cybersecurity attacks, to shift the conventional technical focus from the risks to systems toward risks to people (See our paper: Threat Modeling Intimate Partner Violence: Tech Abuse as a Cybersecurity Challenge in the Internet of Things, Slupska and Tanczer, 2021).

Question 7: What age assurance and age verification technologies are available to platforms, and what is the impact and cost of using them?

We summarise our contributions in the area of age verification as follows:

- The most common method utilised online for age verification is a tick box, whereby the user ticks a box to confirm they are an adult, or they have to enter a date of birth. Neither method verifies whether the user is in fact an adult, they only act as a way of transferring responsibility from the provider to the consumer. Commercial age verification products utilise a variety of methods to verify a user's age. The predominant methods are database checks, government ID checks or AI to determine the user is underage or not.
- The tick-box or date of birth methods both only require web-development time and do not have a per-verification fee. It can be assumed this is why they are the most popular methods today. The commercial age verification products tend to charge per verification, and these range from 25-45p. This initially doesn't seem like a lot of money per transaction, but for businesses with small profit margins, this could prove unaffordable. It's unclear what set up or onboarding costs are associated with the commercial products.

- Very few of the commercial mechanisms preserve users' privacy. Most of these mechanisms use third party identification mechanisms as a proxy for age verification. This is an overkill solution but works very well for the vendors in terms of covering themselves from a legal perspective. Yet the user has to sacrifice their own privacy to use the service. The AI services have the potential to be more privacy preserving, whereby the user doesn't need to identify themselves because the AI model will determine their age. While exploring the potential of AI age verification, we ought to consider the already mentioned caveats of context-dependency, potential for bias and increasing inequalities of already marginalised communities. Additionally, we report that solutions currently being developed achieve accuracy 2 years mean error rate (Yoti, 2021). However, by using a third party to identify a user this does mean that websites aren't required to store and process personal information which is beneficial from both a security and privacy perspective.
- We believe age verification technologies need to possess three key qualities; they are effective, privacy preserving and affordable. This means the technology should be able to determine an adult from a child without the need for the user to be identified. The technology must also be affordable so that all businesses, large or small can implement an effective solution which will prevent children from accessing adult content, products or services online. We are happy to attach our pre-print reviewing age verification mechanisms and recommendations on request.

Question 8: For purposes of transparency, what type of information is useful/not useful? Why?

We would like to share the following comments regarding transparency:

- Transparency of algorithms and automated decision-making systems are particularly useful to examine the impact or potential harms (Crawford, 2021) of such systems on racialised and minority ethnic communities. This information needs to be not only transparent but also accountable (AI Now, 2019). There needs to be robust mechanisms to make service providers accountable (Wong, 2022) if such harms, such as algorithmic harm, could be reasonably prevented before implementation.
- Transparency itself is also often neither sufficient nor fit-for-purpose, when considering algorithmic design, as it doesn't necessarily evaluate the impact, harms, or unfairness in the outcomes/outputs of the algorithms. There needs to be more transparent and robust (Crawford, 2021) impact assessments by service providers relating to the use of algorithmic and data-driven systems. We are happy to provide a chapter from our upcoming book on request.
- Moderation algorithms can have harmful impacts on users when these fail to detect problematic content or erroneously flag/censor harmless content. More transparency is needed around algorithms' tendency to create echo chambers where harmful content is less likely to be flagged and allowed to circulate. Similarly, recent works recommend more transparency around the quality, relevancy and authoritativeness of datasets training automated moderation systems (Gebru et al. 2021; Gilbert and Mintz 2019). We conducted a study looking at the harms of content classification algorithms and suggest

transparency reports should allow for different levels of auditing including the criteria used for curating training datasets, anticipated limitations and harms and reflexive disclaimers about developers' methods, problem statements, institutional affiliations and sources of funding which could bias data collection and model construction. We also recommend measuring and report the uncertainty of moderation algorithms. This can be crucial in aiding human intervention and help to avoid overreliance on classification algorithms. We are happy to share a draft of our paper.

- Given the potential harmful impact on users' lives, especially in the context of online child protection, unambiguous justifications for decisions produced by any automated prevention or detection system should be available to help users, developers, law enforcement and regulators understand the decision-making process of such tools. This should include reasonable disclosure regarding how and when an automated prevention or detection system is engaging with the user, without enabling offenders to circumvent the system. Additionally, transparency should be provided on different levels, e.g. transparency about design, implementation, prior evaluations, training data, matching data, the processes triggered upon illegal content detection, matching results during deployment, false positive rate, etc.

We are happy to provide drafts of our work-in-progress with further information.