Your response

Question	Your response
Question 1: How do you measure the number of users on your service?	Confidential? – Y / N
Question 2: If your service comprises a part on which user-generated content is present and a part on which such content is not present, are you able to distinguish between users of these different parts of the service? If so, how do you make that distinction (including over a given period of time)?	Confidential? – Y / N
Question 3: Do you measure different segments of users on your service? • **Do you segment user measurement by different parts of your service? For example, by website vs app, by product, business unit.** • **Do you segment user measurement into different types of users? For example: creators, accounts holders, active users.** • **How much flexibility does your user measurement system have to define new or custom segments?**	Confidential? – Y / N
Question 4: Do you publish any information about the number of users on your service?	Confidential? – Y / N

Question	Your response
Question 5: Do you contribute any user number data to external sources/databases, or help industry measurements systems by tagging or sharing user measurement data? If not, what prevents you from doing so?	Confidential? – Y / N
Question 6: Do you have evidence of functionalities that may affect how easily, quickly and widely content is disseminated on U2U services? • **Are there particular functionalities that enable content to be disseminated easily on U2U services?** • **Are there particular functionalities that enable content to be disseminated quickly on U2U services?** • **Are there particular functionalities that enable content to be disseminated widely on U2U services?** • **Are there particular functionalities that prevent content from being easily, quickly and widely disseminated on U2U services?**	Confidential? – NO There are a variety of functionalities that are likely to affect the dissemination of content on social media. A recent review paper by Johansson, Enock and colleagues (2022) identified and evaluated several such functionalities in the context of the dissemination of misinformation, though many of these apply similarly to other kinds of harmful online content. On the platform side, content moderation may involve removing content or creators of content that is considered to be harmful, with the aim of preventing exposure to and thus further dissemination of harmful content entirely. *Early stage moderation* One such kind of content removal is known as **early stage moderation**, which involves blocking content at the point of upload to prevent certain content from ever appearing on the platform, preventing users from further exposure. As early stage content removal interventions sit with the platforms, the evidence base to evaluate them remains thin. However, several researchers have attempted to explore how platforms could intervene at the point of upload. For example, some research examines how analysing the contents of a post could help to stop the spread of misinformation. Zhou and colleagues (2020) found that by mining news content for certain attributes, they could predict misinformation with 88% accuracy. In many instances, the spread of harmful content online can be traced back to activity by algorithmically driven social media accounts, also known as bots. Research shows that the curbing of bots could be an effective strategy for mitigating the

Question	Your response
	spread of certain kinds of content in the early stages, particularly as a large proportion of the total traffic that carries misinformation can be traced back to relatively few accounts (Shao et al., 2018; Vosoughi et al., 2018). In their analysis of 14 million messages, Shao and colleagues (2018) find that roughly one-third of "low-credibility" content is spread by only 6% of accounts. While any automated solution risks miscategorising accounts, the identification of bot accounts is a relatively low-risk strategy to adapt. However, if bot accounts or networks become increasingly sophisticated they may manage to go undetected, making this strategy difficult to maintain long term. Predicting the probability that content has the potential to be harmful before it starts to propagate on social media, either by scanning content, links, or propagation networks, typically relies on AI based solutions. While such solutions do scale effectively to the volume of content on social media, they also create further risks, such as a potential difficulty for users in terms of understanding why content has been deleted and the potential for incorrectly deleting acceptable content. Hence, caution is needed alongside robust monitoring and appeals mechanisms for any such automated systems. Sometimes platforms may take a hybrid approach, making content invisible to users until it has been reviewed by a human moderator (a so called 'shadow ban'). *Deplatforming* Another form of content removal is known as **deplatforming**, which is the process of removing a user, channel or forum from a platform when they post content classed considered to be harmful (or violating other rules of the platform). The aim is to prevent generation of further harmful content from the same sources. While platforms routinely delete large volumes of user accounts, the intervention is also often associated with high-profile cases and is typically used only as a last resort, and in many cases as a reaction to public pressure. It is important to consider how platforms go about the deplatforming process. Facebook was criticised in 2019 when it

Question	Your response
	announced the upcoming removal of two far-right influencers prior to their actual ban, allowing them to usher followers to other platforms (Martineau, 2019). Aside from deplatforming single individuals, platforms can also do sweeps of de-platforming users based on a common characteristic. X (at the time called Twitter), for example, removed 70,000 QAnon accounts after the storming of the US Capitol (Conger, 2021). Finally, deplatforming can also encompass the shutting down of whole forums. Key concerns when considering the use of deplatforming are the progression of topics and the reach of the de-platformed (Rauchfleisch & Kaiser, 2021), the activity of their supporters (Jhaver et al., 2021), any potential backlash or counter-reactions (Innes & Innes, 2021) and migration of follower bases (Bryanov et al., 2022; Ribeiro et al., 2021; Rogers, 2020). A study by Jhaver and colleagues (2021), based on 49m Tweets collected six months before and after the deplatforming of Alex Jones, Milo Yiannopoulos and Owen Benjamin, concluded that de-platforming significantly reduced the popularity of many of the anti-social ideas associated with the influencers. A small group of supporters did however increase their activity in reaction to deplatforming—consistent with other findings which show that removal might have negative counter reactions both on the platform in question, and across the wider ecosystem (Ali et al., 2021). However, other research on de-platforming shows the opposite effect. A study by Innes and Innes (2021) collected data mentioning two Covid-19 conspiracy influencers, QAnon affiliated David Icke and Kate Shemirani. Icke was de-plaformed from Facebook in April 2020 because of repeatedly spreading Covid-19 misinformation. Their research showed that the removal attracted additional attention to the influencer, and in the 7 days following his account removal his mentions on Facebook increased by 84%. Informed by empirical analysis, their study proceeds to conceptualise two possible reactions to de-platforming, the

Question	Your response
	creation of so-called “minion accounts” and general efforts to “re-platform”. ‘Minion accounts’ are accounts that surface after deplatforming and which have a clear association with the removed ‘leader’. The accounts continue to post in promotion of the mission or message or the leader, although not under any direction. The emergence of ‘minion accounts’ is one of many ‘re-platforming’ responses, showing persistent diversification of strategies to disseminate their ideas such as diversifying their cross-platform presence. Despite much talk of migration of alt-platforms, few studies quantify these movements. An analysis by Rauchfleisch and Kaiser (2021) found that of the 516 far-right YouTube channels analysed in their 2018–2019 study, 111 had been de-platformed of which 20 could be found on BitChute. They concluded that deplatforming was effective in minimising the reach of misinformation on YouTube, and that despite some users flocking to alternative platforms, these do not make up for the loss in viewership, which is consistent with other quantitative findings on that their audiences on new platforms ultimately ‘thin’ (Rogers, 2020). However, despite having smaller audiences, others point to how deplatforming can have the tendency to harden the views of followers and those engaging with the misinformation (Dwoskin & Timberg, 2021) or how the act of deplatforming can, at least temporarily, bring more attention to them—the ‘Streisand effect’ (Innes & Innes, 2021). *Demonetisation* Aside from removing potentially harmful content or the creation of such content, platforms can also slow the spread of such content through alternative means. **Demonetising content** lessens the incentive for its creation by ensuring that creators and publishers cannot make money from it, for example by generating an ad-revenue. For example, Google updated their ad policy to reflect a ban on misinformation concerning Covid-19 (Dang, 2020), as well as climate misinformation (Hiar, 2021). Ahead of elections held in the United States (2020), Germany (2021) and France (2022), Meta announced a number of policies focusing specifically on securing

Question	Your response
	the integrity of elections, such as banning ads that delegitimise details of the vote or undermine voter safety. Beyond platforms, there are calls for ad networks to do more to reduce the monetary incentives to spread misinformation. One study found that of the top-10 credible ad-servers, those that liaise between retailers and websites selling ad-space account for 66.7% of fake ad traffic (Bozarth & Budak, 2021). Regardless of who acts on it, demonetisation presents itself as an appealing solution for a range of actors, as it sidesteps the arguments often concerned about freedom of speech, instead targeting the incentives that have allowed the misinformation industry to flourish. However, one of the key difficulties of its implementation revolves around effectively identifying information at scale, especially in an industry where many adverts are bought and sold automatically. *Algorithmic downranking* An additional method of slowing the dissemination of content without outright removing it is through **algorithmic downranking.** By changing algorithms such that particular types of content appears less frequently, is shown to fewer users, or appears further down a ‘feed’ or list of recommendations, platforms can limit a piece of content’s amplification on the platform or service. This intervention makes a distinction between the right to have certain content published and its amplification; the right to publish content remains, but there is no ‘right to reach’. This is especially important in the case of some kinds of content such as viral misinformation, which, it has been suggested, is often more successful at achieving high position in many social media ranking algorithms (Shin & Valente, 2020), or is likely to be recommended to users even if they have not shown a prior interest in such content (O’Callaghan et al., 2015). Algorithmic downranking is currently one of the most commonly used interventions by platforms. For example, YouTube downranks unauthoritative content (Courchesne et al., 2021) and Facebook

Question	Your response
	downranks exaggerated or sensationalist health claims (Perez, 2019) Content may even be ranked to zero, meaning it has no ranking and will therefore not be algorithmically delivered to other users in the feed, but it will remain on the platform (Saltz & Leibowicz, 2021). In 2018, Facebook claimed that its downranking efforts cut future views by more than 80% on average for posts that had been labelled as ‘false’ by third-party fact-checkers (Lyons, 2018). Despite being a strategy commonly deployed by platforms, algorithmic downranking remains understudied. Data access for researchers is limited meaning that a true understanding of the workings of algorithms is usually difficult to establish. The proprietary algorithms underpinning platforms is also an important intellectual property, meaning there are strong limitations on whom it can be shared with. *Delisting* Another option available to platforms to keep unwanted content at bay is **de-listing** it. Removing content from any search results provided by the platform means that users who are not specifically looking for misinformation or already involved in communities that spread conspiracies are less likely to find it. Such actions can also include removing hashtags which are another common mechanism of content discovery. For example, Pinterest blocked search results for anti-vaccine terms even before the COVID-19 pandemic (Telford, 2019). There have been multiple instances of platforms banning hashtags that are associated with specific misinformation campaigns, such as those related to the conspiracy that the 2020 US election was “stolen” (Perez & Hatmaker, 2020) and hashtags related to coronavirus misinformation (Jin, 2020). Similar to other interventions discussed in this work, de-listing content is an option that benefits from upholding freedom of expression whilst limiting users access to misinformation and therefore limiting its reach. *User-facing prompts* On the user experience side, platforms have also experimented with adding ‘friction’ to the process of redistributing content, curbing its dissemination. For example, some platforms use **prompts** to encourage people to pause before liking or sharing content, for example by asking people if they would like to read a

Question	Your response
	full article before sharing a headline. The shift in focus intends to induce extra caution or to make users think twice prior to sharing. The intervention targets the act of sharing content, shown to substantially reduce its reach (Andı & Akesson, 2021; Pennycook et al., 2021). The appeal of posting prompts are that they are proactive, as well as that they allow full freedom for users to decide for themselves what content to share, and remove technology companies from the position of having to decide what is true or false. Pennycook and Rand's (2022) recent meta-analysis showed that asking users to consider the accuracy of content prior to sharing reduced the intention to share false headlines by 10% relative to the control, thereby increasing the quality of news people shared (Pennycook & Rand, 2022). Research on posting prompts has inspired multiple campaigns such as the United Nations misinformation initiative, "Pause" ('Pause before Sharing, to Help Stop Viral Spread of COVID-19 Misinformation', 2020), which encourages users to pause before sharing content relating to Covid-19. Platforms have also been seen to have been using this intervention in various contexts. For example, X (formerly Twitter) prompts users to read articles if they haven't opened the link prior to retweeting. The platform later reported that this initiative had resulted in users opening articles 40% more often when having received this message (Hutchinson, 2020). The following year, Facebook followed suit and added a similar prompt (Spangler, 2021). WhatsApp has added friction by limiting the number of times a message can be forwarded as well as the number of people that can be in one group, limiting groups to a maximum of 512 people and allowing messages to be forwarded to a maximum of five groups at once. *User-facing fact check labels* Another intervention on the user experience side are **fact check labels**. These are full or partial overlays which typically warn users that claims made in particular pieces of content have been disputed by fact-checkers, often offering links to more information about the topic. Fact-check labels are

Question	Your response
	usually based on the judgements of expert human fact-checkers. These labels are most commonly implemented by social media platforms and search engines. There is overall support for the efficacy of fact-check labels in reducing susceptibility to misinformation and reducing misinformation sharing intentions (for a review, see Nieminen & Rapeli, 2019). One research study showed that adding a ‘disputed’ or ‘rated false’ tag to a headline significantly lowered its perceived accuracy relative to a control condition and that these overlays were more effective at reducing susceptibility to misinformation than general warnings (Clayton et al., 2020). Additionally, in the same study, fact-check labels did not affect the perceived accuracy of unlabelled false or true headlines (unlike general warnings mentioned above). Other studies similarly find that exposure to a fact-check tag improves accuracy judgments about the specific content (Ecker et al., 2010; Nyhan et al., 2020). However, one issue often noted with fact-check labels is that they rely on the judgements of professional fact-checkers. The process is laborious and human checkers struggle to keep up with the enormous amount of content posted on social media each day. Additionally, individual fact-checkers may be (or may be perceived as being) biased or politically motivated in their assessments. Addressing this issue, recent work suggests that fact-checking may be reliably crowdsourced without impairing the quality of judgments. One recent paper showed that average veracity ratings of politically balanced groups of laypeople correlate highly with judgments of professional fact-checkers, suggesting that employing a 'wisdom of the crowds' approach is a promising way to enhance scalability and reduce perceived bias in fact-checking interventions (Allen et al., 2021). Twitter’s (now X)‘Birdwatch’ (Coleman, 2021) implemented this crowd-sourced approach. Here, members of the programme write notes contextualising posts or provide related information about certain pieces of content. According to the

Question	Your response
	platform, once a Birdwatch note is attached to a tweet, users are 15 to 35 percent less likely to engage with it compared to users who aren't shown the note (Kelly,2022). Recent work shows promise in using Natural Language Processing (NLP) to match social media content with already fact-checked content (Kazemi et al., 2021, 2022; Shaar et al., 2020). This method allows for scalable cross-referencing, as a more efficient way to fact-check than doing so post-by-post. In Kazemi et al. (2022) the classification and retrieval experiments were conducted in monolingual, multilingual, and cross-lingual settings, achieving 86% average accuracy for match classification. References All cited work is discussed in: Johansson, P., Enock, F., Hale, S., Vidgen, B., Bereskin, C., Margetts, H., & Bright, J. (2022). How can we combat online misinformation? A systematic overview of current interventions and their efficacy. *arXiv preprint arXiv:2212.11864*.
Question 7: Do you have evidence relating to the relationship between user numbers, functionalities and how easily, quickly and widely content is disseminated on U2U services?	Confidential? – Y / N
Question 8: Do you have evidence of other objective and measurable factors or characteristics that may be relevant to category 1 threshold conditions?	Confidential? – Y / N
Question 9: Do you have evidence of factors that may affect how content that is illegal or harmful to children is disseminated on U2U services? • **Are there particular functionalities that play a key**	Confidential? – Y / N

Question	Your response
role in enabling content that is illegal or harmful to children to be disseminated on U2U services? - Do you have evidence relating to the relationship between user numbers, functionalities and how content that is illegal or harmful to children is disseminated on U2U services?	
Question 10: Do you have evidence of other objective and measurable characteristics that may be relevant to category 2B threshold conditions?	Confidential? – **NO** In response to Question 10 we wish to highlight to two characteristics that may be pertinent to category 2B threshold conditions: U2U platforms that are frequently redirected to from larger U2U services, as well as U2U services that have lacking content moderation policies or are limited in their user safety choice architecture. As large U2U platforms become increasingly regulated users have migrated to other platforms for illicit content (UNESCO, 2022). Researchers at the Alan Turing Institute have studied this phenomenon, also known as 'signposting', for holocaust denial groups. Their study points to how users on large, mainstream, services have advertised telegram channels posting far more explicit material on the same topic (ibid). The mainstream sites function to direct users to more radical forums, and having data on what services users are being redirected could be an indicator of risk and a characteristic relevant to category 2B threshold conditions. The research shows that users would post content that complies with platform policy, but embedded within them would be links to other platforms, many leading to Discord and Telegram channels. More of this content might emerge as the Online Safety Bill passes, and significant redirection to alternative U2U services might act as a qualifier for category 2B. Similarly, other research points to how these cross-platform operations are a key element to extending the reach of disinformation campaigns that span alternative and mainstream social media (Lazerson, 2023). Including this

Question	Your response
	‘referral’ characteristic might lead to capturing broader cross-platform operations of extremists and other threat actors. Secondly, there are several characteristics one would expect to see on a user-to-user service, and a lack of these could be an indicator of risk, therefore warranting that these characteristics serve as category 2B threshold conditions. The first of these would be the existence of a content moderation policy, and proof of that this is enforced in a timely and consistent manner. However, several U2U services also offer safety features (also known as user controls). These are features users themselves can opt-in to that affect content dissemination by, for example, locking their accounts (so that only approved people can see their content) or disabling comments on their content. In this regard, it is worth highlighting that forthcoming research from the Alan Turing Institute shows that roughly 90% of social media users are aware of safety controls, such as blocking and reporting, and roughly 50% are aware of controls that impact their feed settings, for example hiding comments and likes, or choosing to have a chronologically presented feed (Johansson and colleagues, forthcoming). Our research shows that these tools are used at least by one in three users, with many indicating that they use them to protect themself from harmful content, or to protect their wellbeing. The absence of this segment of functionalities or safety by design, could therefore be seen to be limiting users' choice architecture regarding their online safety, and might come to be helpful in the categorisation of U2U services meeting category 2b threshold conditions.
Question 11: Do you have evidence of matters that affect the prevalence of content that (once the Bill takes effect) will count as search content that is illegal or harmful to children on particular search services or types of search service? For example, prevalence could refer to the proportion of content surfaced against each search term 16 that is illegal or harmful to children, but we	Confidential? – Y / N

Question	Your response
welcome suggestions on additional definitions. - **Do you have evidence relating to the measurement of the prevalence of content that is illegal or harmful to children on search services?**	
Question 12: Do you have evidence relating to the number of users on search services and the level of risk of harm to individuals from search content that is illegal or harmful to children? - **Do you have evidence regarding the relationship between user numbers on search services and the prevalence of search content that is illegal or harmful to children?**	Confidential? – Y / N
Question 13: Do you have evidence of other objective and measurable characteristics that may be relevant to category 2A threshold conditions?	Confidential? – Y / N

Please complete this form in full and return to os-cfe@ofcom.org.uk.