Ofcom

making communications work
for everyone

Your response

Please refer to the sub-questions or prompts in the Annex of our Call for Evidence.

Question	Your response
Question 1: Please provide a description introducing your organisation, service or interest in Online Safety.	Confidential? – N The Oversight Board is an independent model of online content moderation. Established in 2020, we have delivered assessments on some of the most significant challenges on social media today. Examples include hate speech in Myanmar, the suspension of former US President Trump, doxxing (the act of publicly revealing previously private personal information), COVID-19 misinformation, information shared during conflicts in places like Ukraine and Ethiopia and the treatment of content shared by journalists and news outlets. Our mission is to uphold freedom of expression, as well as other human rights by reviewing content moderation decisions taken on Facebook and Instagram. We are independent of Meta and funded by an irrevocable trust. The Board's work is two-fold: We issue binding decisions on content, but we also make recommendations to improve Meta's content moderation policies. We look for challenging cases and test if Meta's rules uphold human rights standards. Where they fall short, we propose ways to fix them. These changes can then be applied to all Meta's users globally to bring fairer, safer and more consistent standards on social media more broadly, while still taking into consideration local context and challenges. Our 23-member Board is comprised of global experts who are all specialists in their field, ranging from journalism and politics to activism and human rights. The Board began excepting cases from the public in late 2020. By this Fall, the Board would have decided 28 cases (more than half of which relate to countries in the Global South) reversing Meta's content moderation decisions in 20 of them. These reviews have resulted in the Board issuing well over 100 policy improvement recommendations; the majority of which Meta has committed to implementing or exploring the feasibility of implementing. These policies apply to all users on

Facebook and Instagram globally. They can also provide guidance or inspiration other social media and tech companies on best practices. Our mission has always been to inspire the wider industry to more seriously integrate human rights concerns into their decision-making processes.

The Board is firmly committed to transparency and our decisions are published with full, public explanations. To date, we have considered issues such as hate speech, incitement to violence and journalistic freedom. We have also been dedicated to fighting misinformation. Notable examples include the conflict in Ethiopia, where we have removed content relating to unverified rumours and made recommendations about stopping the spread of possible misinformation that was helping to incite violence. The fight against COVID-19 mis/disinformation has also been a key priority. We have taking two cases to date and are currently conducting a wide-ranging Policy Advisory Opinion on the future of content moderation around COVID-19 on Meta's platforms as the situation continues to evolve.

The Board always takes a human rights-based approach to analysing content moderation decisions and has public engagement firmly built into our deliberation processes. To date we have received more than 10,000 public comments that helped to shape all of our decisions. These also add aspect of public safety by allowing people to alert the Board and therefore Meta to issues / harms that are happening to users and non-users alike and to provide further context about these issues. People can do this anonymously if needed to ensure vulnerable groups and individuals are protected.

During the last two years, the Board has asked Meta hundreds of questions, opening a transparent space for dialogue with the company which did not exist before. In many more cases, the Board's work resulted in a voluntary decision by the company to reverse wrongful content moderation decisions.

The Board has a large base in London, with dozens of staff members and the Board's Director Thomas Hughes, all based in the UK. However, while the Online Safety Bill only operates in one legal jurisdiction, the Board's remit is global. At present, the Bill does not give the Government nor Ofcom the role of referee on difficult individual content moderation decisions, apart from imposing general principles about what causes harm or has the potential to cause harm. Our work to

	date shows that the Oversight Board is uniquely placed to participate in with online content moderation discussions in the UK and globally and has an extensive bank of expertise and knowledge to share as these conversations evolve. We would welcome collaboration on this going forward, particularly the establishment of a clear and formal consultation process that brings together the tech sector, civil society and organisations like the Oversight Board with Ofcom and the government that together can monitor, report and advise on the progress of the Bill.
Question 2: Can you provide any evidence relating to the presence or quantity of illegal content on user-to-user and search services?	Confidential? – N Our work does not cover illegal content.
Question 3: How do you currently assess the risk of harm to individuals in the UK from illegal content presented by your service?	Confidential? – N Over the last two years, the Board has identified a broad range of harms that resulted from poor content moderation on Meta's platforms. While the Board takes cases from all over the world, our policy recommendations and the standards we set are indented to be applied globally, including to users in the UK. For instance, cases relating to medications and plant-based substances that can be abused are extremely relevant to the UK where their use is also prevalent: • In February 2022, the Board determined that a US-based user's Facebook post, asking for medical advice about the prescription medication Adderall, should not have been removed by Meta and violated the user's freedom of expression. Meta had removed the content under its Restricted Goods and Services Community Standard and suspended the user for 30 days. • Similarly, a December 2021 case found that a Brazilian post concerning the plant-based brew ayahuasca, which made statements about the substance allowing one to "overcome fear" and "break free", did not violate Instagram's Community Guidelines. For both cases, the Board examined the possibility of direct or immediate connection between the content

and the possibility of harm and did not find one. Instead, it found that Meta’s response to non-medical drugs, which the company did not define publicly, was disproportionate. It then recommended that Meta change its rules to allow users to discuss the traditional or religious uses of non-medical drugs in a positive way. Although cases such as these do not originate from UK users, they reflect standards and issues that impact the UK.

In 2022, the Board also took its first UK specific case concerning UK Drill, a subgenre of rap music popular in the UK with a large number of Drill artists active in London. The case arose following a request by UK law enforcement to remove the artistic content from Instagram.

The post in question features a video clip of a UK Drill song. Shortly after the video was posted, Meta received a request from UK law enforcement to remove content that included this track. Meta says that it was informed by law enforcement that elements of it could contribute to a risk of offline harm. The company was also aware that the track referenced a past shooting in a way that raised concerns that it may provoke further violence. As a result, the post was escalated for internal review by experts at Meta.

Meta's experts determined that the content violated the *Violence and Incitement policy,* specifically the prohibition on "coded statements where the method of violence or harm is not clearly articulated, but the threat is veiled or implicit."

When Meta took the content down, two days after it was posted, it also removed copies of the video posted by other accounts. Based on the information that they received from UK law enforcement, Meta's Public Policy team believed that the track "might increase the risk of potential retaliatory gang violence", and "acted as a threatening call to action that could contribute to a risk of imminent violence or physical harm, including retaliatory gang violence."

Hours after the content was removed, the account owner appealed. A human reviewer assessed the content to be non-violating and restored it to Instagram. Eight days later, following a second request from UK law enforcement, Meta removed the content again and took down other instances of the video found on its platforms. Meta subsequently referred this matter

	to the Board, stating that this case is particularly difficult as it involves balancing the competing interests of artistic expression and public safety. The Board is currently grappling with these issues and deliberating the complexities of the case. It is expected to issue its decision and policy recommendations around the case before the end of the year. More information can be found here: https://oversightboard.com/news/385467560358270-oversight-board-announces-new-cases-and-review-of-meta-s-covid-19-misinformation-policies/
Question 4: What are your governance, accountability and decision-making structures for user and platform safety?	Confidential? – N The Board issues binding decisions on content and makes recommendations to improve Meta's content moderation policies. The Board looks for challenging cases and tests if the rules uphold human rights standards. Meta also requests input from the Oversight Board on some significant and difficult content decisions, policies, and enforcement issues. These decisions are advisory but, for accountability, Meta publicly responds to recommendations within 60 days. Human rights impact assessments are conducted on all our cases, to scope the rights holders who may be impacted whenever we select cases. We outline risks and mitigating actions, although this doesn't replace Meta's own corporate responsibility to conduct human rights due diligence. Our decisions are structured around the three-part test of legality, legitimacy, and necessity and proportionality enshrined in Article 19 of the International Covenant on Civil and Political Rights. But our decisions go further and identify other human rights implicated in a case. Several of the UNGP principles (18, 20, 21 and 31) are routinely applied to our work as we seek to create legitimate, accessible, predictable, equitable, transparent and rights-compatible pathways, while also showing a commitment to continuous learning. The effectiveness criteria run through all our work. We continuously push Meta for transparency and provide regular updates in our own transparency reports on the nature of appeals; the human rights standards referenced in each decision; and Meta's responsiveness to the Board's questions.

	The most recent example can be found here: https://oversightboard.com/news/572895201133203-oversight-board-publishes-transparency-report-for-first-quarter-of-2022/
Question 5: What can providers of online services do to enhance the clarity and accessibility of terms of service and public policy statements?	Confidential? – N When Facebook and Instagram users feel that they have been treated unfairly, or that violating content has remained on the platforms in error despite appeals to Meta, they are then able to appeal their case to the Oversight Board. To ensure this process is easy and clear, our website, where appeals are submitted, is available in a wide variety of languages. This includes the appeals process, FAQs and case information. Almost two million cases have been submitted to date from all around the world, including from the UK. Many of the Board's recommendations, born from the case review process, have identified serious issues around a lack of clarity and transparency in rules and standards on Meta's platforms. The Board has time and again called for these to be simplified and presented in a way that users can easily understand. The Board has also called for more cohesiveness between standards on Instagram and Facebook, which should ease user experiences across both platforms. In addition, we have called for greater clarity for users to receive information about why their content was removed. This has been an integral part for our drive for greater transparency, aimed at impowering users. Key examples of our work toward enhancing accessibility include: • Calling for Meta to publish and translate community standards and policies into more languages. This will lead to 100s of millions of people finally being able to understand the rules governing content moderation on the platforms. However, more work remains to be done and our decisions to date highlight that Meta's rules for Facebook and Instagram are still not available in all user languages. We also continue to raise concerns about whether Meta was investing sufficient resources in moderating content in languages other than English.

	• Calling for the creation and publication of the company's Crisis Policy Protocol, which is used to codify Meta's content policy response to crises and assess situations that require a new or unique policy response. On the Board's recommendation, in August 2022 Meta published this in part, but the Board is continuing to push for greater clarity for users in this regard.
Question 6: How do your terms of service or public policy statements treat illegal content? How are these terms of service maintained and how much resource is dedicated to this?	Confidential? – Y / N N/A
Question 7: What can providers of online services do to enhance the transparency, accessibility, ease of use and users' awareness of their reporting and complaints mechanisms?	Confidential? – N Transparency is clearly an area where Meta must improve, and the Oversight Board's mission is to try and be part of the solution. In October 2021, we began publishing our first quarterly transparency reports. We are continuously pushing Meta to be more transparent and treat users better. Our recommendations have repeatedly urged Meta to follow some basic tenets of transparency: make your rules easily accessible in your users' languages. Tell people as clearly as possible how you make and enforce your decisions. And, where people break your rules, tell them exactly what they've done wrong. We've already seen some early wins for user transparency based on the recommendations issued so far: • Health misinformation policies are now consolidated, with clearer guidance on the harms Meta is seeking to reduce. Our current policy review of COVID-19 mis/disinformation will hopefully only add further clarity and transparency to how these issues are handled.) • Meta will update its Transparency Centre on content removed for violating its Community Standards following a formal report by a government, including the number of requests it receives. • Meta added information to its Community Standards on when content is eligible for fact-

	checking, including whether public institutions are eligible. This is just the start but, as the Board continues to better understand Meta's processes and policies, we will only further increase transparency that will benefit users. A key example of this will be our Policy Advisory Opinion on the cross-check system, due this autumn. It will shed some light on Meta's system of moderation protocols for high-profile users. In terms of the Board's own processes there are three ways cases for content to be reached for review: appeals by people, case referrals by Meta, and requests for Policy Advisory Opinions (PAOs). Users can submit an appeal in 10 minutes on the Board's website – and significantly more than 1.5 million have done so to date. The Board then prioritises cases that are challenging, globally relevant and can inform future policies which impact the almost three billion Instagram and Facebook users, including those in the UK. Members of the public can submit evidence in cases and a written explanation of the final decision – as well as provide public comments – which are always made available publicly. The Board is also always looking for further avenues to enhance engagement and collaboration. The Board's first annual report, which explains our work on transparency in more detail, can be found here: https://www.oversight-board.com/news/322324590080612-oversight-board-publishes-first-annual-report/
Question 8: If your service has reporting or flagging mechanisms in place for illegal content, or users who post illegal content, how are these processes designed and maintained?	Confidential? – Y / N N/A
Question 9: If your service has a complaints mechanism in place, how are these processes designed and maintained?	Confidential? – Y / N N/A
Question 10: What action does your service take in response to reports or complaints?	Confidential? – Y / N N/A

Question 11: Could improvements be made to content moderation to deliver greater protection for users, without unduly restricting user activity? If so, what?	Confidential? – N The Oversight Board's system allows for minimal restriction of user activity whilst ensuring that content moderation takes place. The moderation is in fact driven by users themselves, who submit cases for consideration. The case-based approach means that, once a difficult issue has been considered by the Board, this content moderation can be taken forward immediately by Meta. As with traditional media, after the Online Safety Bill comes into force, self-regulation will still form a huge part of day-to-day decision-making on content policy. It is right that those rules should be assessed on an ongoing basis through an independent entity outside of a social media company.
Question 12: What automated moderation systems do you have in place around illegal content?	Confidential? – Y / N N/A
Question 13: How do you use human moderators to identify and assess illegal content?	Confidential? – Y / N N/A
Question 14: How are sanctions or restrictions around access (including to both the service and to particular content) applied by providers of online services?	Confidential? – Y / N N/A
Question 15: In what instances is illegal content removed from your service?	Confidential? – Y / N N/A
Question 16: Do you use other tools to reduce the visibility and impact of illegal content?	Confidential? – Y / N N/A
Question 17: What other sanctions or disincentives do you employ against users who post illegal content?	Confidential? – Y / N N/A
Question 18: Are there any functionalities or design features which evidence suggests can effectively prevent harm, and could or should be deployed more widely by industry?	Confidential? – N The Board was created to test the model for independent oversight and transparency that impacted how social media companies should operate and improve global moderation practices. We believe, two years on from taking our first public case, the Board is

	clearly generating results and shown the benefit of independent oversight more broadly. This model works to complement government regulation and is a system that would bring significant benefits to a broad range of social media and tech companies which should be encouraged to ensure independent oversight is better built into their processes. In more granular terms, the Board is increasingly exploring the use of new tools in content moderation such as placing warning screens on graphic and misleading content, as well as using external fact-checking services. The Board's current review of Meta's COVID-19 misinformation policy will see us issue recommendations on issues like: • The effectiveness of social media interventions to address COVID-19 misinformation; • The use of algorithmic or recommender systems to detect and apply misinformation interventions, and ways of improving the accuracy and transparency of those systems; • The fair treatment of users whose expression is affected by social media interventions to address health misinformation, including the user's ability to contest the application of labels, warning screens or demotion of their content. These guidelines will shape best practices not only at Meta, but we hope across the social media ecosystem.
Question 19: To what extent does your service encompass functionalities or features designed to mitigate the risk or impact of harm from illegal content?	Confidential? – Y / N N/A
Question 20: How do you support the safety and wellbeing of your users as regards illegal content?	Confidential? – Y / N N/A

Question 21: How do you mitigate any risks posed by the design of algorithms that support the function of your service (e.g. search engines, or social and content recommender systems), with reference to illegal content specifically?	Confidential? – Y / N N/A
Question 22: What age assurance and age verification technologies are available to platforms, and what is the impact and cost of using them?	Confidential? – Y / N N/A
Question 23: Can you identify factors which might indicate that a service is likely to attract child users?	Confidential? – Y / N N/A
Question 24: Does your service use any age assurance or age verification tools or related technologies to verify or estimate the age of users?	Confidential? – Y / N N/A
Question 25: If it is not possible for children to access your service, or a part of it, how do you ensure this?	Confidential? – Y / N N/A
Question 26: What information do you have about the age of your users?	Confidential? – Y / N N/A
Question 27: For purposes of transparency, what type of information is useful/not useful? Why?	Confidential? – Y / N N/A
Question 28: Other than those in this document, are you aware of other measures available for mitigating risk and harm from illegal content?	Confidential? – Y / N N/A