Your response

Please refer to the sub-questions or prompts in the annex to our call for evidence.

Question	Your response
Question 1: Please provide a description introducing your organisation, service or interest in Online Safety.	*Is this response confidential? – N* The Wikimedia Foundation (**Foundation**) submits these comments in response to Ofcom's Online Safety Call For Evidence (**CFE**). The Foundation

	appreciates the opportunity to offer input to some of the questions posed by the CFE. The Foundation hosts several free knowledge projects, the largest of which is Wikipedia. Wikipedia, the free encyclopedia, is a collaborative project created and maintained in over 300 languages by volunteers across the globe. The community of volunteers, who comprise the global Wikimedia Movement, collaboratively write and edit the content of the encyclopedia as well as create and enforce rules regarding both content and behavior on the platform. Since Wikipedia is organized around a singular goal, the construction and maintenance of an online encyclopedia, the types of potential harm on the platform are different than on social media platforms. The Wikimedia Movement's approach to addressing potentially harmful or illegal content has been tailored over years of community and organizational practice to promote fairness and minimize harm. This necessarily involves close collaboration between volunteer moderators and professional trust and safety staff.
Question 4: What are your governance, accountability and decision-making structures for user and platform safety?	*Is this response confidential? – N* Content moderation on Wikipedia, and other volunteer-run free knowledge projects that the Foundation hosts and supports, is largely conducted by a community of nearly 300,000 global volunteer contributors. In addition to editing Wikipedia, volunteers also collaborate to create and enforce policies as well as adjudicate disputes that arise under those policies. Many Wikimedia projects have boards dedicated to proposing new policies for the projects which are discussed and voted on by other volunteer community members until consensus is reached, not simply a majority vote. On English Wikipedia, for example, proposals to introduce or change policies must be announced on the "Village Pump" noticeboard and "require discussion and a high level of consensus from the entire community for promotion to guideline or policy." Wikipedia is collaboratively edited, which means that almost every change to articles, even small grammatical edits, are based on community-determined standards and could be considered an act of content moderation. Every article has a "history" section, which indicates what changes have been made and who has made those changes, and a "discussion" section, where users

can discuss changes they want to make before hitting “edit.” These basic safeguards build accountability into the editing process and put content moderation tools and processes in the hands of the entire community.

More experienced volunteers within the movement are given greater enforcement powers through a community selection process. These “administrators” and “bureaucrats” have the ability to block or unblock accounts, temporarily protect pages from being edited, and delete pages entirely. These volunteers have typically engaged extensively with the projects by contributing hundreds of edits when they are selected as administrators, and much of the proactive work to prevent vandalism and non-relevant content is done at this intermediary level of volunteer enforcement.

On English Wikipedia, our largest project, there is also an elected Arbitration Committee which handles disputes over content and conduct on the projects. These cases involve formal hearings, which can be private or public, as well as a formal appeals process. Once a dispute is settled, the Arbitration Committee will publicly publish its decision along with any consequences which have been taken.

While much of this dispute resolution is processed wholly within the volunteer community, the Foundation’s trust & safety and legal teams regularly engage in dialogue with users and community members, providing community members with opportunities to ask staff about policy decisions or other issues of concern. This close collaboration has led to initiatives like the Universal Code of Conduct, a policy developed with the community that offers new levels of protection for volunteers on Wikimedia projects when it comes to conduct disputes.

Finally, there are certain situations which cannot be handled by volunteers and are escalated to the Wikimedia Foundation trust & safety emergency response team to address. This includes situations where there is a threat of serious harm to someone’s physical safety as well as some higher level conduct issues which require a full, confidential investigation. This type of escalation is possible because of the trusted relationship between the Foundation and the volunteer administrators who maintain the Wikimedia

	projects.
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?	*Is this response confidential? – N* The terms of use [ToU] for the Wikimedia projects prohibit a broad range of harmful activities, and explicitly prohibit the misuse of the service for illegal purposes or activities. Our ToU are officially translated into 29 different languages, and we maintain a "Governance Wiki" where we maintain documentation related to policies and governance of the projects. The Wikimedia volunteer community also enforce project-specific policies which address illegal content, like these from English Wikipedia. We also engage in ongoing legal education for the Wikimedia volunteer community. Since the Wikimedia projects are organized by language, rather than geography, there can sometimes be questions from the community about how to respond to potentially conflicting laws. When particular issues surface consistently, we publish educational content analyzing the legal issues at stake to help users better understand their own risks when posting content. We also occasionally publish blog posts about litigation the Foundation has been involved in, explaining our legal theories, the decisions that have been made, and how they will impact our volunteer community's efforts on Wikipedia. We also regularly interact with the community through organized conferences and community conversation hours dedicated to specific legal topics. The close coordination and continuous dialogue with the large volunteer community provides us with early insights when issues do arise, while allowing us to take the community's input into account while we are making decisions about content and governance.
Question 9: If your service has a *complaints* mechanism in place, how are these processes designed and maintained?	*Is this response confidential? – N* While most complaints about content on the projects can be remedied by directly changing the content in dispute, the Foundation does occasionally receive legal requests to remove or change content on the projects. In these cases, teams at the Foundation consider several elements when determining whether to act on a complaint. First, for issues emerging internationally, we apply a multi-factor analysis on a case-by-case basis to determine whether non-U.S. law applies. Since the Foundation is headquartered and incorporated in

	the U.S., non-U.S. laws do not always apply. In addition to evaluating whether a law applies to an issue and the risks presented to the Foundation and/or the Movement, this analysis examines whether compliance with a law would be in line with international human rights norms, including rights related to privacy, free expression, and dignity. We consider all complaints through a human rights lens, in alignment with the Foundation's human rights policy. Our processes are designed and maintained with human rights principles in mind, and subject to human rights impact assessment. At the moment, conduct disputes are typically opened through an email ticketing system. Once the trust & safety team decides to open a case, they investigate the facts of the case and can take punitive actions such as global project editing bans or event bans. The Foundation has committed to making these workflows even more seamless in the future, and a dedicated anti-harassment team is working to develop and implement user reporting systems, which will be native to the Wikimedia platforms instead of relying on email.
Question 11: Could improvements be made to content moderation to deliver greater protection for users, without unduly restricting user activity? If so, what?	*Is this response confidential? – N* The Wikimedia community is already highly effective at removing illegal and harmful content on the projects. Researchers at the Berkman Klein Center for Internet and Society at Harvard University found that the median amount of time harmful content remained on English language Wikipedia was 61 seconds. They found that Wikipedia's system of identifying and removing harmful content is largely effective, despite Wikipedia's large scale, the variety of content, and different interpretations of the Wikimedia Foundation's guidelines and policies. However, improvements can always be made to further protect the human rights of readers and editors. Recognizing that many histories and perspectives have been excluded by structures of power and privilege, the Wikimedia Foundation envisions a key role that free knowledge projects can play in achieving inclusive, equitable, quality education, and in realizing the human right to non-discrimination. We are committed to improving knowledge equity for women, LGBTQ+ communities, historically underrepresented racial and ethnic groups, people

	with disabilities, and communities in underserved regions, and in people's native languages. Furthermore, we recognize that not all content on our platforms may be appropriate for all audiences, including children. We strive to support our volunteers with training to ensure that content is handled sensitively and appropriately where it may be deeply disturbing or result in harm. The Foundation has commissioned a Child Rights Impact Assessment, scheduled to conclude in 2023, which will inform further steps that the Foundation and volunteer community might take. The Wikimedia Foundation is committed to making these improvements in a manner that is consistent with international human rights standards. As the hosts of Wikimedia projects, the Foundation operationalizes our commitment to human rights by conducting ongoing human rights due diligence, including periodic human rights impact assessments, in addition to regular and robust engagement with rights holders and their legitimate representatives. Staff also track and publicly report on our efforts to meet our human rights commitments as well as provide access to redress, proportionate to the type and manner of harm. Through our Movement Strategy process, which set out goals for the Wikimedia Movement and which was conducted with extensive input from the community, we have committed to promoting equity by working with partners across the private sector and government to advance the realization of rights we are uniquely positioned to support. There will always be improvements that can be taken to make online spaces safer and more welcoming for marginalized users, and the Wikimedia Foundation's greatest innovations in this field have come after years of listening to community concerns and collaborating to address specific problems raised by the community.
Question 12: What automated moderation systems do you have in place around illegal content?	*Is this response confidential? – N* Editors on Wikipedia employ a multi-layered approach to discovering and removing harmful speech on the projects. The Foundation seeks to empower users to participate in content moderation processes by, for example, providing them access to machine learning tools which they can use to improve or quickly remove content. While the Foundation may assist developers with building tools, they are used and maintained by community members.

	One of the tools editors can use is ClueBot NG, an automated tool which uses a combination of different machine learning detection methods and requires a high confidence level to automatically remove vandalism on the projects. Another tool is a machine learning tool called Objective Revision Evaluation Service (ORES) which assigns scores to edits and articles in order to help human editors improve articles. Additionally, users with special privileges have access to the AbuseFilter extensions, which allows them to set specific controls and create automated reactions for certain behaviors. While automated tools are used to support existing community moderation processes, the bulk of the work is still done manually. Wikimedia uses select automated tools to scan for child sexual abuse material and works closely with law enforcement to report content that violates applicable law.
Question 13: How do you use human moderators to identify and assess illegal content?	*Is this response confidential? – N* The Foundation's ToU describe the rights and responsibilities of users and the Foundation, but each Wikimedia project also has its own set of policies and guidelines. These include speedy deletion policies, which allow administrators to immediately delete pages or media without going through the formal deletion procedures. Criteria that make articles or pages subject to speedy deletion include pure vandalism and blatant hoaxes, as well as attack pages that "disparage, threaten, intimidate, or harass their subject or some other entity, and serve no other purpose." In addition to quickly removing content on the basis of speedy deletion policies, some automated tools developed by the community automatically remove content and assign scores to the quality and reliability of revisions. Administrators and users with advanced and specialized technical permissions are then able to limit the visibility of harmful content by limiting the discoverability of flagged edits in an article's revision history, which is visible to all readers. They are also able to block users, IP addresses, and investigate cases of disruptive editing and sock puppetry. Additionally, there are editors and administrators who specialize in moderating particular categories of content and are therefore able to quickly identify and remove harmful content. Indeed, some editors have set up automatic alerts for any changes made to

	Wikipedia articles and pages they personally commit to monitoring, and are therefore able to quickly escalate harmful content issues to administrators or address the issues themselves. It is important to note that Wikipedia's guidelines and policies do not specifically define "harmful content" as a category. However, researchers identified that harmful content on Wikipedia generally falls into five broad categories: (1) harassment, (2) threats, (3) defamation, (4) identity-based attacks, and (5) posting personal information of another without consent. These same researchers also found the median amount of time harmful content remained on English language Wikipedia was 61 seconds due to the volunteer community's robust content moderation practices.
Question 15: In what instances is illegal content removed from your service?	*Is this response confidential? – N* The Wikimedia Foundation removes content in accordance with the applicable law determination policy. Content will only be removed in cases where it is clear that the law applies to the facts at issue and the Foundation is within the local court's jurisdiction. Additionally, the case must be one that presents safety, technical, and/or monetary risk. Finally, the case must be one where compliance is in line with international human rights standards and meets the Foundation's commitment to human rights under the Foundation's Human Rights Policy.
Question 16: Do you use other tools to reduce the visibility and impact of illegal content?	*Is this response confidential? – N* Independent of community moderation processes and automated tools maintained by community members, the Foundation does not use other tools to reduce the visibility and impact of illegal content on the projects. This is because algorithmic highlighting or amplification are not deployed on the projects. Wikimedia's non-profit, public interest projects are devoted to a single mission: enabling people everywhere to freely share factually verified knowledge. Unlike some other commercial platforms, the Wikimedia projects do not amplify or target content to maximize reader engagement or attention. To the contrary, the projects are structured in a way that does not allow illegal content to spread virally on the projects, limiting the threat of illegal content being widely viewed.
Question 19: To what extent does your service encompass functionalities or features designed	*Is this response confidential? – N* The fundamental principles of Wikipedia are

to mitigate the risk or impact of harm from illegal content?	summarized in five "pillars," which set expectations for all community members, visitors, and other users. The encyclopedic focus and related policies mitigate significant risks related to illegal content—Wikipedia is written from a neutral point of view, all articles must have verifiable sources, and topics are limited to notable topics. The ethos of being a free culture project also often disincentivizes posting illegal content. For example, Wikimedia Commons, our free image repository, often removes copyrighted content even if there may be another legal exceptions or justification for hosting the content. On a more granular level, all edits, contributions, and other actions taken on Wikipedia are documented and publicly displayed. There are edit histories for articles and contribution lists for users—including anonymous users that are identified by their IP address. This policy is a safeguard and means that no one can upload, add, or edit content without leaving a footprint that is attached to an identifier, be it a user name or an IP address. Any articles that are being vandalized or are otherwise controversial will be locked, so that no further edits, contributions, or comments can be made. By giving community-elected administrators a wide variety of tools to use when investigating or preventing vandalism, they can choose the tool that is most appropriate for the given circumstance.
Question 20: How do you support the safety and wellbeing of your users as regards illegal content?	*Is this response confidential? – N* In addition to the policies and practices mentioned elsewhere, we continue to evolve our practices around supporting our volunteer community of editors as they work to improve the Wikimedia projects. One of the goals identified by the Wikimedia Movement during the Movement Strategy process was to "Provide for Safety and Inclusion." Under this recommendation, the Foundation and the volunteer community will work together to develop tools, resources, and practices which support safety on the projects. As a part of the implementation of this strategy thus far, the Foundation has collaborated with the volunteer community to develop a Universal Code of Conduct governing behavior on projects as well as a pilot program for peer support aimed at supporting the mental health of volunteers,

	especially those who have experienced harassment. The Foundation places editor safety in high regard, and is committed to continue evolving and learning how best to support the volunteers who create and maintain the Wikimedia projects.
Question 26: What information do you have about the age of your users?	*Is this response confidential? – N* Wikipedia collects no mandatory demographic information on its editors and readers in accordance with our Privacy Policy. Because we collect so little information, requiring birthdate or age information for all editors, readers, and others who may access the site would run counter to our commitments to data minimization principles and to upholding our readers' right to privacy. We deliberately do not collect any information about the age of our users because of the restricted, educational subject matter of the Wikimedia projects. While we are aware that some content on Wikipedia may be objectionable, as on all open platforms, we believe that access to knowledge is an important right for everyone of any age. Wikipedia is first and foremost, an encyclopedia, and encyclopedias in the physical world are not age-restricted or censored based on the age of the person holding the volume, though they may contain material that could be considered disturbing for younger readers. Many of the threats that younger users face on social media platforms are also less prevalent on Wikipedia due to the nature of the platform. Any open conversation on the platform tends to revolve around the building of that encyclopedia and the existing content therein. There are no private messaging capabilities for users. Any in-thread or other communication between users is publicly posted and visible, meaning that private predatory behaviors cannot take place on the Wikimedia platform itself.
Question 27: For purposes of transparency, what type of information is useful/not useful? Why?	*Is this response confidential? – N* At the Foundation, we produce bi-annual transparency reports which report on requests for user information along with requests to remove or alter content. The numbers in the report highlight the efficacy of the volunteer community's content moderation efforts, but also some of the unique challenges of transparency reporting on smaller platforms. Platforms which receive relatively few reports have less data to report on, meaning that the more granular the categorization of complaints, the

	greater the chance that those complaints will become at least partially identifiable. Additionally, platforms which practice data minimization practices, collecting as little data as possible about users, may actually end up having to collect more data about individuals if certain reporting categories are required (i.e. identifying the number of minor users of a service requires collecting age data about a platform's users). It is important that when developing any transparency requirements, Ofcom intentionally balance the need for transparency with the need to protect the privacy of individual users and complaints. Further, for transparency reports to be effective in comparing actions across platforms or on one platform over time, it is especially important that any transparency reporting requirements are clear. If platforms are left to individually determine what constitutes "harmful" or "illegal" content when reporting, the resulting reports are likely to have significantly different interpretations of those categories. Thus, while the Foundation recommends against including overly specific categories of reporting, which may compromise user privacy as discussed above, there should be clear definitions and examples provided for broader categories of content which may be left up to platform interpretation otherwise.
Question 28: Other than those in this document, are you aware of other measures available for mitigating risk and harm from illegal content?	*Is this response confidential? – N* One additional step the Wikimedia Foundation has taken to mitigate risks and harms from content on our projects is to conduct a Human Rights Impact Assessment (HRIA) to identify human rights risks related to the Wikimedia projects as well as opportunities to address and mitigate those risks. Our inaugural HRIA report identified several steps which could be taken to reduce harmful content and mitigate risks to child rights specifically. As discussed in question 11, we are now conducting a Child Rights Impact Assessment, which will allow us to gain greater insight to the risks to child rights on the platform and additional opportunities for mitigation. Additionally, peer-to-peer education and training can help the volunteer community to be better prepared to address harmful content if they encounter it, and can even improve digital literacy skills overall. Wikimedia UK, a local chapter of the Wikimedia Movement, regularly works with schools and universities to put on classroom

	education activities, teaching students how to contribute to Wikipedia and educating them about how information is shared and spread online. These programs were designed with digital literacy skills development in mind, and help students to better exercise their writing, research, and critical thinking skills while navigating content online.