Written evidence submitted by Wikimedia Foundation

The Wikimedia Foundation welcomes the opportunity to comment on the draft Online Safety Bill ("Bill"). As the nonprofit that supports Wikipedia and other free knowledge projects, we are concerned the Bill may result in a less useful, more dangerous internet for everyone. The requirements and costs associated with complying with the Bill may force companies to limit access to reliable information that risks qualifying as harmful content under the draft Bill. Furthermore, the tracking and identification requirements in the Bill seriously threaten the safety of our users online, particularly in countries where freedom of expression and access to information is highly restricted.

The Bill lacks clarity on key enforcement points, and should look to already existing international and human rights standards to fill those gaps.

Our first concern with the draft Bill is the overbroad definition of "harmful content," which could lead platforms to restrict access to critical information, including health information. Wikipedia, for example, includes encyclopedic information on a variety of potentially "harmful" topics, including sexual activity, drug use, suicide, and self-harm. Wikipedia's content standards—including requirements of neutral point of view, verifiability, and educational content—ensure readers have access to reliable information about a subject, and minimize the likelihood that the content is triggering or otherwise harmful. Under a broad definition of harmful content, however, information that is encyclopedic and reliable may have to be suppressed, leading users to seek information on less reputable websites, including websites which cannot be compelled to answer in the UK.

We are also concerned about the broad discretion granted to the Secretary of State for Digital, Culture, Media and Sport ("the Secretary") and OFCOM, which creates uncertainty on how the Bill will be enforced. The Secretary has the discretion to expand the meaning of "illegal content," "content that is harmful to children," and "content that is harmful to adults." Additionally, the Secretary can decide who is eligible to make "super-complaints" under section 106. An eligible entity could have political motives or institutional biases against particular platforms; for example, an eligible entity might dislike their article on Wikipedia because it contains unflattering information about a scandal related to the organization. In such instances, an eligible entity can argue that content on Wikipedia must be investigated since, by definition, encyclopedic information about that scandal qualifies as harmful content under the broad definitions in the draft Bill.

The requirement that Parliament must approve most regulations under section 132 is reassuring, particularly since it provides a needed limitation on the Secretary's and OFCOM's discretion. However, this does not offset the potential harm that could result from the broad discretion still afforded to the Secretary to grant and repeal exemptions. MPs should prioritize establishing a clear framework for when exemptions can be granted and withdrawn by the Secretary to limit the Secretary's discretion and provide stability in this matter. Additionally, the Secretary's degree of influence over OFCOM's strategic priorities and autonomy provide further

areas of potential overreach, limiting OFCOM's neutrality and independence [sections 57, 109, and 33(1)(a)].

We appreciate the procedural safeguards put into place by the draft Bill. In particular, the requirement in section 29(5)(f) to consult with human rights experts with expertise in Articles 8 and 10 of the European Convention on Human Rights is welcome in addition to other comments made in the Explanatory Notes on this matter. We would like to see this approach expanded to include other international agreements and standards that already address harms both online and offline. For example, The World Health Organization ("WHO") develops guidelines for critical global health issues, including mental health and suicide prevention. The WHO recently published guidance incorporating best practices for suicide prevention from a diverse group of countries. This guidance includes media reporting best practices, which have been shown to reduce the risk of suicide. The WHO also published information on the risk of suicide among victims of domestic violence, which most likely qualify as a "certain group of people" under sections 45 and 46 of the Bill.

Since international institutions are already working to address online harms based on their specific mandates, legislators should defer to their expertise and global reach rather than that of UK government officials. To minimize the risks of politically-motivated regulations that may result in human rights violations, the definition of "psychological harm" and "harmful content" should be based on international human rights standards and guidance from the WHO. Additionally, the draft Bill should qualify compliance with rigorous standards like the UN Guiding Principles on Business and Human Rights as automatic compliance with the Bill.

The vague definition of harmful content in the Bill will ultimately incentivize the suppression of legitimate content

We likewise appreciate the flexibility that the draft Bill tries to give platforms to conduct content moderation their own way. However, by instituting a duty of care, the draft Bill eschews the traditional intermediary liability protections which allow for effective moderation and proactive removal of content. This burden shift, along with the vague definitions of harmful content in the Bill, will incentivize the use of automated tools to find, filter, and remove content, much of which may not qualify as "harmful" under the Bill.

Examples of such content which is likely to be suppressed under such a law are activist content, documentation of government crimes and abuses, art, and criticism. At scale, there is little time to consider each piece of content in its full ambiguity and context, and thus broad lines will likely be drawn in order to avoid the stiff penalties that come with failing a "duty of care." Similarly, the threat of jail time for senior leadership in sections 72 and 73 may provoke some platforms to engage in over-censorship.

While large social media companies may tout the effectiveness of their automated content detection systems, these systems are still relatively limited in what they can detect accurately. As we see in the case of copyright enforcement, when automated processes are used to detect and remove content online, the consequence is often the over-removal of entirely legal, legitimate content. These tools have been shown to be subjective and to not capture nuances and contextual variations of human speech. Additionally, the fundamental

flaws in the underlying datasets and tools cannot be fully mitigated with human oversight. If companies are expected or forced to replicate these models on content which requires additional context or tone to interpret, these broad over-removals and perpetuating bias will increase.

In fact, by incentivizing the use of automated filters, the Bill threatens to interrupt effective content moderation systems developed and implemented by individual communities. At Wikimedia, much of our content moderation is done by our editing community; these community content moderation processes have been shown to work efficiently and to provide users with complaint mechanisms to challenge removal. The volunteer editors on Wikipedia and other Wikimedia projects currently use a machine learning tool called Objective Revision Evaluation Service ("ORES") to flag vandalism on the projects and to predict article quality, but ORES itself does not make final decisions. This relationship acknowledges the limitations of machine learning while harnessing its strengths—these tools help our community of volunteer editors to do their jobs more efficiently, but are no substitute for human review.

Since Wikipedia is written for a general audience, limiting access to all information that meets the definitions of "harmful content" for children or particular groups of adults under the Bill would be impossible. New information is continuously being added to Wikimedia projects, and some encyclopedic content may have an adverse psychological impact on some readers. As such, organizations like the Wikimedia Foundation must have the flexibility and resources to invest in processes and tools that empower users to participate in content moderation processes by, for example, having access to machine learning tools and having the ability to improve them, and to quickly remove harmful content.

The Bill grants too much power to large tech companies, while disrupting effective content moderation on smaller platforms

Preparing to comply with the vague requirements of the Bill will require implementing new tools. Smaller companies investing in compliance with one system may not be able to comply in the highly efficient manner large tech companies are able to, creating a greater imbalance within the industry than already exists. As such, legislators should amend the Bill to limit government officials' discretion to ensure organizations, for-profit and nonprofit alike, have the flexibility to direct their resources to the processes that are most effective for making their platforms safe.

There are several sections throughout the draft Bill that require service providers to perform their content duties based on their capacity, size, type of service, type of content included on their platform, and to use systems and processes that are proportionate to their activities as long as they have regard for online safety objective and users' human rights [Sections 9, 10, 21, 22, 31, 49, 56, 115, and 36]. Service providers can appeal OFCOM's decisions to categorize them as a particular type of service provider and appeal technology notices under sections 104-106. However, appealing OFCOM's decisions, or an organization's decision to remove content, is costly and time intensive. While larger companies have the resources to appeal to the Upper Tribunal, smaller companies and most individuals do not. As such, these safeguards do not, in practice, provide smaller parties with equal opportunities to appeal content moderation decisions. These sections are welcome safeguards, but do not mitigate the

discretion OFCOM and elected officials have to issue guidance and regulations that circumvent or otherwise undercut them.

By instituting a broad duty of care for platforms, the Bill forces quick, top-down directives from the platforms, disrupting existing community processes and creating distrust between platforms and their communities. Research has shown Wikimedia's community content moderation processes work efficiently and provide users with recourse for edits in the form of talk pages and other complaint mechanisms to challenge removals. The majority of content moderation that takes place on the Wikimedia projects is handled by the Wikimedia community before any issues are raised to the Wikimedia Foundation. While these processes are effective in removing content, they are slow, deliberative, and user-initiated. Requiring decisions to be made by Wikimedia Foundation staff will compromise these processes by removing a significant amount of human oversight, making it even harder to achieve the aim of a safe online community.

Regulatory efforts that presuppose all platforms handle content moderation in-house or with subcontractors fail to account for the myriad other viable forms of content moderation that power the modern internet. Wikipedia is not alone in this format: message boards have relied on volunteer community moderators for decades, and volunteers continue to help maintain content standards in Facebook groups, Reddit subreddits, and more. Harsh liability requirements, demands for explanations behind individual takedowns, and brief mandatory takedown windows quickly become unworkable in such systems. Imposing such requirements on Wikipedia and other organizations will fundamentally break the structure that has allowed this historically unique public reference work to thrive.

The Bill institutes national regulations on global platforms, while leaving some of the UK's most vulnerable residents behind

One of our final concerns with the current draft Bill is that it would require global companies, nonprofit organizations, and other institutions to treat users around the world differently. The United Kingdom, as a member of the United Nations and a party to several human rights treaties, including the International Convention on Civil and Political Rights, should strive to make the internet safe for individuals around the world and not simply within its own borders. We recognize the bill is not happening in isolation, but is part of a global trend that includes the Digital Services Act in the European Union, the Network Enforcement Act in Germany, the Canadian online harms proposal, and the Australian Online Safety Act. This trend has resulted in a proliferation of similar yet competing standards. Indeed, this increased interest by governments around the world to see who can address problematic content online first will result in a fractured set of rules that no global platform can comply with completely.

The Wikimedia Foundation, as operator of the Wikimedia projects, works on a global scale, distinguishing our projects by language rather than country so people from around the world can collaborate and build a reliable knowledge resource together. As a website that provides vital, but occasionally sensitive information to the public, we also strongly defend the rights of our users and readers to privacy on Wikimedia projects. As such, laws which fail to

consider globally accepted standards for human rights or data protection are antithetical to Wikipedia's global perspective and reach.

In fact, laws which have attempted to address harms broadly online have resulted in the further marginalization of those they attempted to help. Online harms are real harms which require real solutions, but they are reflective of systemic, offline issues that cannot be resolved through internet regulation alone. The United Nations High Commissioner on Human Rights has highlighted that algorithmic tools reflect and sometimes amplify historic racial and ethnic biases that exist offline. Laws like the draft Bill can also result in important health, safety, and contextual information about illegal behavior being removed alongside strictly "illegal" content, driving users to rely on information from less reliable sources and from platforms with less effective content moderation processes. As such, while the Bill is well-intended, the issue of online harms will not disappear until systemic offline harms are addressed.

Recent legislation in the United States provides a helpful example regarding the potential harms caused by ill-considered content moderation bills. In 2018, the US passed the Fight Online Sex Trafficking Act (FOSTA), which carved out a subject-matter exception to the intermediary liability protections provided by Section 230 of the Communications Decency Act, despite objections from digital rights organizers and impacted sex workers. Overzealous content moderation and fears of liability prompted many websites to prohibit not only sex trafficking-related content, but also content pertaining to consensual sex work, educational information about sexuality, and even fully legal businesses such as non-sexual massage therapy. Just three years later, several of the bill's advocates and co-sponsors already recommend re-examining and repealing the bill. We encourage the Secretary to learn from this, as well as other rightfully rejected proposals in the UK, and listen closely to marginalized groups who express concern about this Bill.

What qualifies as harm differs across cultures and languages, as well as across lived experiences. Since the guidance and regulations specific to this bill will be drafted by UK regulators, there may be significant blind spots in the types of harm recognized and addressed. Often, marginalized communities know better than regulators what content is truly harmful to those communities, and broad rules enforced by platforms can actually disempower communities from telling their stories and lived experiences. The German-language Wikipedia, for example, identified Nazi symbols as a possible type of harmful content after Germany passed a broad law declaring the symbols unconstitutional, and developed specific rules concerning those symbols. The community acknowledged the historical importance of preserving reliable information about the Holocaust, while recognizing the adverse psychological impact this content may have on Holocaust survivors and their families, and developed language-specific policies to address possible harms. This type of nuanced policy, instituted by the community it impacts, is one of the great strengths of allowing community content moderation to flourish.

We are also concerned that requirements in the draft Bill that platforms track their users discourage free communications, especially in countries where online censorship is prevalent and users are most vulnerable. Wikipedia does not track its readers. This type of fracturing national law, and the draft BBill in particular, are not complementary and are often contradictory with international law, including agreements on human rights and data

protection. This is important for protecting individuals' right to privacy and reader and contributor autonomy alike.

Limiting the information we collect on users is crucial for the safety of Wikipedia contributors who contribute or moderate content on sensitive topics, or who contribute from regions where their personal safety could be at risk for editing Wikipedia. In order to comply with the draft bill as written, Wikimedia would have to collect information about users' ages and protected characteristics, even where collecting such information could present a personal risk to that user.

These requirements could also violate international human rights standards by undermining protections like encryption. Under section 70 of the draft Bill, OFCOM can request essentially any information, which could include encrypted information. The Special Rapporteur on the promotion and protection of freedom of opinion and expression has stated that encryption is needed to protect the most vulnerable individuals, who rely on encryption to protect themselves online, from interference with their right to freedom of opinion and expression. Restrictions on encryption must be necessary and proportionate to achieve a legitimate objective. The Secretary has yet to issue regulations on OFCOM's information powers, organizations' online safety duties, and the definition of harmful content. While protecting children and vulnerable groups is a legitimate and important aim, at this time, it is unclear whether the restrictions on encryption and demands for information will be necessary and proportionate restrictions on individuals' right to freedom of opinion and expression, and privacy.

We urge the Secretary to reconsider the existing provisions of the Bill with a broader set of stakeholders in mind. This includes considering alternative models of content moderation, including community governance, and listening to marginalized groups about the potential consequences of certain language in the Bill. The Bill should be grounded in international human rights principles, which help to keep the internet free, open, and safe for everyone. More specifically, we hope to see more precision in the definitions in the Bill and more information about how discretion is being granted to government officials. By making the costs of compliance more clear, the UK government can avoid a regulation which would compromise the promise of community-driven projects like Wikipedia and threaten to fragment the internet as a whole.

17 September 2021