Annex to illegal content public consultation response

Mozilla welcomes the opportunity to contribute to the European Commission's ongoing reflections around illegal content online. As a mission-driven not-for-profit technology company, we are uniquely placed to provide thoughtful input in the ongoing discussions in Brussels and beyond on how to manage the harms of illegal content online within a rights-protective framework. Ultimately, illegal content on the web – and substandard policy and industry responses to it – undermine the overall health of the internet and as such, are a core concern for Mozilla.

Regrettably, the chosen format for the public consultation makes it extremely difficult for respondents to capture the many nuances at play with respect to illegal content on the web and policy responses to it. Moreover, many of the questions are framed in such a way as to infer a certain outcome, making it more difficult for respondents to provide objective and reasoned feedback. We fear that the result will be an opaque and partial view of stakeholder perspectives and policy suggestions, to the determinant of the Commission's desire for substantive input. In an attempt to address this, Mozilla has opted to submit additional comment in the form of an annex, to provide context and nuance to our responses in the consultation.

In brief, our annex holds that:

- There is no one-size-fits-all approach to content regulation. While some solutions can be generalised, each category of content has nuances which must be appreciated when crafting policy and operational solutions.
- In the effort to enhance trust and safety, automated content control solutions are no panacea. Such solutions, especially those that involve content filtering are of little value when context is required to assess the illegality and harm of a given piece of content.
- Trusted flaggers offer *some* promise as a mechanism for enhancing notice & takedown procedures. However, such entities can never replace judicial authorities as assessors of the legality of content and as such, their role should be limited to fast-track notice procedures.

- Fundamental rights safeguards should be included in illegal content removal frameworks *by design*, and should not be patched on at the end. Transparency and due process should be at the heart of such mechanisms.

There is no one-size-fits-all approach to content regulation

As Mozilla has argued for many years, it is difficult to craft generalised and overarching policy solutions to optimise the fight against illegal content. Indeed, in some crucial cases, nuance and narrow-tailoring are essential. For instance and as we explain in more detail below, automated content recognition may be useful in helping with the detection of child sexual abuse material (CSAM), but is completely inappropriate for content where *context* is a determining factor to illegality or harm (e.g. copyright infringement, 'hate speech'). It is thus critical to craft a regulatory approach that does not conflate policy responses and technical solutions across different types of illegal content.

This is not to say that *certain* general principles cannot be deployed across different types of illegal and harmful content. For instance, notice & action regimes have applicability across the board, and rightly remain the preeminent mechanism for ensuring trust and safety online. Ultimately, it is crucial for policymakers to avoid the assumption that simply because a certain measure works with respect to a given content form, it will necessarily work with respect to other content forms. Reflection, nuance, and tailoring are essential when considering policy and operational responses, to avoid inapt and excessive interferences with speech and other forms of online content.

Automated solutions are not a panacea

Linked to this, we note with concern that the consultation questions lead towards the position that automated content control tools, such as preemptive upload filters, are a panacea in the fight against illegal and harmful content on the web. Mozilla has argued for many years that automated tools - especially those that aim at automatically detecting and filtering illegal and harmful content - are a crude control instrument, and are of limited use for assessing the legality of content where *context* is essential. This includes elements such as the sampling of small snippets of copyright-protected content as part of original user-generated content, or strongly-

worded commentary around political issues of the day, satire, culturally-motivated depictions and more.

This assessment challenge is exacerbated with respect to 'harmful' content online, given the even greater context needed to assess whether such content qualifies for control actions (e.g. potential audience, relation to broader cultural norms, etc) and the high-risk implications for freedom of speech. In an era where artificial intelligence and related technological advances are excessively revered, a sober reflection on the limits of such solutions for tackling content-related challenges is necessary.

Ultimately then, we strongly believe that filtering tools should only ever be considered in a few discrete cases, given that they are effective only with respect to *certain* categories of content - most notably, CSAM, spam, etc - and even then, only with respect to a *small subset* of that given category of illegal content. Moreover, filtering tools can often lead to the inadvertent suppression of legal speech, engendering a free expression chilling effect. While the Commission's consultation focuses on possible safeguards to mitigate against the free expression harms, we note that such safeguards cannot offset the corrosive harms inherent in the use of such technologies for content control.

Trusted flaggers offer potential but caution is required

As we note above, notice & action is and should remain the cornerstone of our efforts to tackle illegal content online. In context of the Mozilla mission and the recent addendum to our Manifesto, Mozilla strongly supports the exploration of mechanisms that aim at making notice & action more efficient and scalable.

In that context, so-called 'trusted flagger' mechanisms hold some promise, given the possibility they offer for a greater volume of high quality and actionable notices to service providers. Consequently, it is not without reason that 'trusted flagger' mechanisms (both ad hoc and formal) have come to play a role of increasing importance within many service providers' trust & safety strategies.

Worryingly however, trusted flagger mechanisms take content control further outside of a rule of law framework, given that the opinion of a trusted flagger

can never replace the assertion of a judicial authority. Indeed, while they may have authority to assess whether certain content is *harmful* and *likely* to infringe on legal standards, trusted flaggers themselves cannot make the decision as to whether a certain piece of content is in fact illegal. For that reason, it is inappropriate to suggest - as some of the consultation questions seem to imply - that trusted flaggers should be empowered to mandate automatic takedowns of content. This same reasoning applies to law enforcement authorities and the so-called 'Internet referral units' that exist within Europol and several EU Member State law enforcement authorities.

On the basis of the above we believe that an entity with trusted flagger status should simply benefit from a 'fast track' notification mechanism, such that the service provider prioritises the processing of notifications submitted by the trusted flagger. And in that context, it is of paramount importance that both standards and safeguards are incorporated into any future 'trusted flagger' certification schemes and also the codes that govern their interaction with service providers. Furthermore, such certification schemes should be subject to regular review and auditing.

Ultimately, it is difficult to speak substantively about trusted flaggers when the concept is so vague and fluid. Consequently, much elaboration and thoughtful consideration will be required before it is appropriate to position trusted flaggers within a substantive policy framework for tackling illegal content online.

Safeguards need to be built into the system

On the point of safeguards more broadly, the consultation's dedicated section on that issue constitutes a welcome acknowledgment that the envisaged content control mechanisms pose a serious risk of erroneous removal of *legal* content and the fact that greater transparency is needed from both OSPs and competent authorities to ensure adequate reporting of *what* content is removed and *why*.

However, while transparency is valuable in this context, it is just one of the many safeguards that need to be present in future regulatory interventions around illegal content. Due process is also essential and mechanisms such as appeal processes and independent dispute resolution should be incorporated into the framework *by design*. In that vein, policymakers must appreciate the

fact that the removal of illegal content and the protection of fundamental rights cannot be traded against each other.

Conclusion

In conclusion, we would like to reiterate the recommendation we offered in our feedback to the recent Inception Impact Assessment, namely that the Commission should focus on continued monitoring of the progress of the various ongoing initiatives at national and EU level that aim at fighting illegal content. A multi-stakeholder discussion on how to create an adaptive, rights-based framework for notice & action and content responsibility is warranted, but must *precede* any new legislative intervention in the area. The next European Commission mandate, with reinvigorated political impetus and a five-year span, will be the best place to begin and culminate this ambitious project to set a global standard in how to address illegal content within a rights-protective framework.

In any case, Mozilla looks forward to constructively engaging with the Commission in this space, to ensure the internet remains an empowering and integral part of modern life for all.