Mapping platform child safety transparency

A Concept Note from the Center for Online Safety and Liberty

Executive summary

The Center for Online Safety and Liberty (COSL) proposes to publish a research report, the first of its kind, to shine a light into the child protection sector and to reveal how Internet companies, child abuse reporting hotlines, and software vendors make decisions to censor or restrict content on the stated grounds of child protection. This will enable those affected by these decisions to determine what policies have been have been applied, and what recourse may be available against their misapplication. Where transparency gaps exist, we will identify them and made recommendations for improvement.

What are hoped for goals or longer term effects of the project?

The elimination of child pornography, which the child protection sector nowadays refers to as child sexual abuse material (CSAM) and child sexual exploitation material (CSEM), is an essential and important objective for the sector as a whole, and for Prostasia Foundation (Malcolm, 2018a). But due to the importance of this work and its public interest character, it is essential that it be conducted in accordance with the highest standards of transparency and accountability, lest the machinery for eliminating unlawful images also be brought to bear on lawful content.

This, in fact, frequently happens. Just a few of the best known examples of the censorship of lawful content as child exploitation material include Facebook's censorship of the iconic Vietnam War image of Kim Phúc (Levin & London, 2016), the Internet Watch Foundation's blocking of Wikipedia over its refusal to censor an album cover (Metz, 2018), and Tumblr's censorship of all adult content shortly after being removed from the Apple Store for failing to address its illegal child pornography problem (Malcolm, 2018b).

The reliance on private companies as key actors with responsibility for the removal of illegal and harmful content means that rather than interpreting and enforcing the law, their flagging and removal of content often involves the interpretation and enforcement of their own private terms of service, which are frequently couched in terms of a broad and vague prohibition on "sexualizing minors."

However, the end result is the same; when an Internet company adopts a more expansive policy against "sexualizing minors," it is as if the government had imposed a new and broader legal definition of "child pornography:" books by art photographers and erotic

novelists disappear from Amazon, Japanese animation movies are removed from Netflix, fan artists have their content taken down from social networks, and search results disappear from major search indexes.

These decisions may well be substantively justified, and this project will not be rendering any judgments about where the line between lawful speech and harmful child exploitation should be drawn (although see "Who is this project for?" below for why this question is important). Rather, we simply want to uncover who is drawing these lines, and what are their stated justifications for doing so, so that these justifications can be held up to public scrutiny and interrogated.

The findings of this project will be published in the inaugural edition of an annual report on the practices of Internet platforms, software vendors, and content rating agencies, which will become an ongoing resource for those who are affected by the child protection practices of these bodies, and provide an aspirational standard for improvements in their accountability and transparency.

The end result of the project will be that the process of eliminating content on the grounds of child protection will become more open to public oversight, which will in turn enable the child protection sector to benefit from the input of a broader range of stakeholders, including those who are most impacted by the over-censorship of their own lawful speech, such as LGBTQ+ communities, sex educators, sex workers, artists, journalists, child sexual abuse professionals, and children themselves.

How will you do it?

The first edition of an independent annual report on the transparency and accountability practices of major platforms, consultants and agencies involved in online child protection will be prepared by Prostasia Foundation. The scope of the publication will extend to:

1. Major Internet platforms such as Facebook, Twitter, Medium, and Amazon, to determine whether their child protection policies are clearly stated, predictably applied, and whether decisions made under such policies are subject to the same mechanisms of review as decisions made about other types of content.
2. Vendors of software or services that are used by Internet platforms directly, or used by law enforcement officers in cooperation with Internet platforms, for censoring or moderating consent for purposes of child protection, including Microsoft, Google, Thorn, and the Child Rescue Coalition. For example the report will determine whether the software is documented, whether its source code is available for review or testing, and whether its use is audited.
3. Agencies such as the Internet Watch Foundation (IWF) and the National Center for Missing and Exploited Children (NCMEC) that supply "hash lists" (unique identifiers or known unlawful material) or "URL lists" (lists of Internet addresses pointing to unlawful

material) to be used by platforms in moderating content will also come under consideration for their own transparency and accountability practices.

To the extent that we have been able to access that information, this report will answer the following questions in respect of each of these three groups of actors:

> What are the criteria related to child protection that respondents use for classifying content (including images, videos, and text) that is to be restricted?
>
> What technologies are used for classifying or restricting content, and how are these technologies publicly documented and reviewed?
>
> At what stages are the use of these technologies subject to human oversight, auditing, and impact assessment?
>
> If applicable, are these technologies used to scan private communications such as direct messages or emails?
>
> What form of notice, if any, is given by respondents to those whose content is classified or restricted?
>
> What statistics are published of the volume and nature of the content that the respondent has classified or restricted?
>
> What process exists for the respondent's determinations to be challenged or appealed, and by whom can this be done?

The report will be published under an open license and presented at relevant professional gatherings of the child protection, human rights, and Internet governance communities.

How long have you been thinking of working on this idea? What made you first think about it?

Since the first transparency report was issued by Google in 2010, transparency reporting by Internet companies has become accepted as an industry best practice, and the depth and scope of such reporting has been improving year on year (Bankston, Schulman, & Woolery, 2017).

But these improvements have not been uniform, with child protection policies often falling outside their scope, either explicitly or in practice. For example, Internet companies not only restrict content based on user reports and their own internal blocklists, but also based on lists of hash values of content that are derived from various third-party sources.

In addition to a list maintained by the National Council on Missing and Exploited Children (NCMEC), companies may also use a shared industry hash list, a list contributed by third-party nonprofits, and a list of content that falls short of the standard of "apparent child pornography" (Clark, 2017, pp. 6–8). The criteria for inclusion in each list are opaque, and

although it exercises statutory authority, NCMEC is not amenable to FOIA (*Lazaridis v. US Dept. of Justice*, 713 F. Supp. 2D 64).

There is no way for a user whose content is flagged by automated systems that use these lists to know which list caused their content to be flagged, the circumstances in which it was added to that list, whether their content has also been referred to law enforcement authorities as apparent child pornography, or who to contact to have corrections made.

Who is the project for?

Raising public confidence in the impartiality and bona fides of the various actors in the child protection sector benefits everyone who is affected by the decisions that they make. Greater transparency does not preempt a discussion of where the line between lawful speech and child exploitative content should be drawn. Rather, it is essential to illuminate that discussion, and to ensure that mechanisms developed for the narrow purpose of protecting children from exploitation are not used as cover for a broader regime of private censorship.

Increasing censorship of lawful sexual content under the guise of child protection has been observed following the passage of the law FOSTA, which narrows platforms' safe harbor protection from liability for users' content. As Prostasia Foundation has pointed out in our joint *amicus curiae* brief in a pending constitutional lawsuit, this has even extended to the censorship of materials dedicated towards the prevention of child sexual abuse (Freedom Network USA et al., 2019).

The line between child exploitation and lawful speech is especially likely to be misdrawn when those who are drawing the line exclude the perspectives of sexual minorities, most of whom are themselves stigmatized. Prostasia Foundation advocates for the rights of these groups to participate in policy discussions around child sexual abuse prevention (see the next section), but improved transparency of those discussions is an essential precondition of their ability to meaningfully participate in such discussions.

No matter where one chooses to draw the line between child exploitation and lawful speech, the openness, transparency, and inclusiveness of the process by which that line is drawn is a shared public good. This project will shed new light on existing processes of Internet content censorship by actors in the child protection sector, and thereby enable broader oversight and participation in that important work by currently excluded and marginalized groups.

What community currently exists around the project?

In May and June 2019 Prostasia Foundation held a Multi-Stakeholder Dialogue on Internet Content, Sexual Content, and Child Protection, in which we brought together representatives of Internet companies, experts from a range of disciplines, and stakeholders from affected communities, in a private expert-led seminar and open discussion about best practices to protect children without infringing the human rights of children or others.

By facilitating a dialogue with experts and stakeholders who are normally excluded from the development of child protection policies by Internet platforms, industry participants learned how to make these policies more evidence-informed, and more compliant with human rights standards. The objective was to improve their accuracy in the moderation of sexual content: removing more material that is harmful to children and has no protected expressive value, and less material such as lawful, accurate information on child sexual abuse prevention.

As an output of this series of meetings, we have developed a set of draft *Best Practice Principles for Sexual Content Moderation and Child Protection* (Prostasia Foundation, 2019) emphasizing that preventing harm to children should be the touchstone for child protection policies, and that any restrictions on content that does not directly harm children should be evidence-based and take account of the human rights impacts of such restrictions. An online community that includes participants from our May and June meetings is currently finalizing these guidelines in an open and transparent process.

This project is envisioned as a step towards making those draft recommendations actionable. We cannot assess content moderation and censorship decisions against a best practice standard unless those whose content is being evaluated know where, when, and by whom this is being done.

Our work is also situated within a broader community of digital and human rights activism. We will draw upon and supplement other reports on the transparency and accountability of Internet platforms, such as the Ranking Digital Rights project (Ranking Digital Rights, 2019) and the Who Has Your Back report (Gebhart, Crocker, Opsahl, & Mackey, 2019).

However as explained above, due to the fact that content governance in the child protection sector draws on a radically different set of stakeholder inputs, technologies, and guiding principles, there are no existing documents that cover the same ground as our planned report. The 2019 Who Has Your Back report, for example, explicitly excludes child sexual exploitation content from its coverage for this very reason.

Why is this project needed?

Navigating the grey areas of content removal related to child exploitation and its prevention is a thankless and difficult job for Internet content moderators. Due to the extreme stigma that surrounds this topic, and the enormous pressure that governments place on Internet companies to provide solutions to the problem of child sexual abuse, choosing to remove content in cases of doubt has generally been a pretty safe decision.

But as previously explained, taking a precautionary approach to the removal of content is not without its own risks and costs. The over-removal of lawful sexual content, both visual and written, has had documented adverse impacts on LGBTQ+ communities, sex educators, sex workers, artists, journalists, child sexual abuse professionals, and children themselves.

It's important for the public to be able to know where the line between lawful sexual content and child exploitation is being drawn, by whom, and on what basis. This is made more difficult by the fact that in the child safety sector specifically, there is a marked lack of transparency and accountability in comparison with other areas of content governance. Policies are not spelled out as clearly, partnerships with agencies such as trusted flaggers and software vendors are not disclosed, and explanations for takedowns are not given.

To give just one example, if you conduct a Google search that would otherwise return the Wikipedia page for the topic "Lolicon," you will find that page conspicuously omitted from the search results. Unlike in the case of content removal on copyright grounds under the Digital Millennium Copyright Act (DMCA), there is no indication in the search results that any result has been removed.

It may be that this page is among a list of that United Kingdom authorities requested Google and Microsoft to remove from their search indexes in November, 2013. But there is no public record of what pages were removed from the search indexes, any subsequent such requests, the geographical scope of the removal, or the impact of this censorship on the behavior of people assumed to be searching for child sexual abuse material (Jütte, 2016, p. 7).

One reason that companies give for refusing to part with such information is to prevent the information being used to support offending, or for the development of countermeasures against censorship. However, this justification cannot apply to the restriction of legal material (such as Wikipedia pages censored from search results), nor to the distribution of image hashes used for filtering (as actual images cannot be derived from these hashes), nor to the criteria for classifying such images, nor to the documentation of software used for their detection and elimination, or much of the other information that we are seeking to uncover.

The adage that security through obscurity is no security at all, applies equally to the domain of child protection. If the child protection sector's response to the proliferation of illegal sexual images of minors depends on keeping certain keywords, hashes, or assessment criteria secret, then that regime is overdue for review. By identifying how processes for censoring or restricting content are being deliberately obscured from the public, our report will raise the bar of transparency for the child protection sector as a whole.

About COSL

The Center for Online Safety and Liberty (COSL) is a nonprofit dedicated to empowering individuals and communities to thrive online by building safer digital spaces, fostering creativity, combating harm, and championing digital rights. COSL serves as an incubator for independent projects that tackle pressing issues such as age verification mandates, Section 230 rollbacks, encryption battles, and content-scanning overreach, while also developing open source trust-and-safety tools and nurturing inclusive online communities.

Project coordinator

The coordinator for this project will be our Executive Director, Jeremy Malcolm. Dr Malcolm has significant experience of managing multiple complex, international and multi-stakeholder projects. He has raised and managed six-figure project budgets, working with donors such as Ford Foundation, Open Society Foundations, the International Development Research Centre (IDRC), and Google. While employed at Consumers International as Senior Policy Officer (2008-2014), he coordinated its global program Consumers in the Digital Age and was responsible for spearheading proposed revisions to the United Nations Guidelines for Consumer Protection. While he was Senior Global Policy Analyst at the Electronic Frontier Foundation (2014-2018), he led the development of the Manila Principles on Intermediary Liability, which have become an aspirational global standard on that topic. Dr Malcolm graduated with degrees in Law (with Honours) and Commerce in 1995 from Murdoch University, and completed his PhD thesis at the same University in 2008 on the topic of Internet governance. Dr Malcolm's background is as an information technology and intellectual property lawyer and IT consultant. He is admitted to the bars of the Supreme Court of Western Australia (1995), High Court of Australia (1996) and Appellate Division of New York (2009). He is a member of the Multistakeholder Advisory Group of the United Nations Internet Governance Forum.

References

Bankston, K., Schulman, R., & Woolery, L. (2017, February 9). Case Study #3: Transparency Reporting. Retrieved July 9, 2019, from New America website: https://www.newamerica.org/in-depth/getting-internet-companies-do-right-thing/case-study-3-transparency-reporting/

Clark, J. (2017, April 25). Testimony of John F. Clark of the National Center for Missing and Exploited Children for the European Parliament Committee on Civil Liberties, Justice and Home Affairs. Retrieved July 11, 2019, from http://www.europarl.europa.eu/cmsdata/117902/john-clark.pdf

Freedom Network USA, Sex Workers Project, New York Transgender Advocacy Group, Sharmus Outlaw Advocacy and Rights Institute, Decriminalize Sex Work, National Coalition for Sexual Freedom, … St. James Infirmary. (2019, February 20). Amicus Brief of Prostasia Foundation et al in support of Appellants in Woodhull Freedom Foundation v United States.

Gebhart, G., Crocker, A., Opsahl, K., & Mackey, A. (2019, June 12). Who Has Your Back? Censorship Edition 2019. Retrieved July 10, 2019, from Electronic Frontier Foundation website: https://www.eff.org/wp/who-has-your-back-2019

Jütte, S. (2016). *Online child sexual abuse images: Doing more to tackle demand and supply*. Retrieved from NSPCC website: https://learning.nspcc.org.uk/media/1194/online-child-sexual-abuse-images.pdf

Levin, S., & London, J. C. W. L. H. in. (2016, September 9). Facebook backs down from "napalm girl" censorship and reinstates photo. *The Guardian*. Retrieved from https://www.theguardian.com/technology/2016/sep/09/facebook-reinstates-napalm-girl-photo

Malcolm, J. (2018a, July 4). Child protection and the limits of censorship. Retrieved July 11, 2019, from Prostasia Foundation website: https://prostasia.org/blog/child-protection-and-the-limits-of-censorship/

Malcolm, J. (2018b, December 4). Tumblr's adult content ban is an admission of defeat. Retrieved July 10, 2019, from Prostasia Foundation website: https://prostasia.org/blog/tumblrs-adult-content-ban-admission-defeat/

Metz, C. (2018, December 7). Brit ISPs censor Wikipedia over "child porn" album cover. *The Register*. Retrieved from https://www.theregister.co.uk/2008/12/07/brit_isps_censor_wikipedia/

Prostasia Foundation. (2019, July 3). Draft best practice principles for sexual content moderation and child protection. Retrieved July 11, 2019, from Prostasia Forum website: https://forum.prostasia.org/t/draft-best-practice-principles-for-sexual-content-moderation-and-child-protection-master-thread/21

Ranking Digital Rights. (2019). *2019 Ranking Digital Rights Corporate Accountability Index*. Retrieved from https://rankingdigitalrights.org/index2019/assets/static/download/RDRindex2019report.pdf