Ranking Digital Rights submission to the UNESCO consultation on a "model regulatory framework for the digital content platforms to secure information as a public good"

Submitted January 20, 2023

Introduction

Ranking Digital Rights (RDR) engages in research and advocacy to advance corporate accountability in the digital age. We are a non-profit program at the Washington, D.C.–based think tank New America and our mission is to hold tech companies to account for their obligations to promote and respect freedom of expression and privacy on the internet. We do this by establishing global standards for companies to respect and protect the human rights of internet users and their communities, and then we evaluate those companies on their realization of those standards in their terms of service, privacy policies, and other public policy disclosures and commitments. We publish the results of our evaluation in periodic rankings as part of the RDR Corporate Accountability Index, which comprises our Big Tech and Telco Giants scorecards. In addition to our research, we also advocate for laws and public policies that safeguard these fundamental rights and work with a growing array of investors seeking to manage risks associated with human rights harms that are enabled, exacerbated, and perpetuated by social media and other tech platforms.

RDR's set of 58 human rights standards provide specific guidance in three categories: corporate governance, freedom of expression, and privacy. We encourage UNESCO to review these standards—which have already provided policy direction to companies, intergovernmental bodies, ESG investor benchmarks, and industry associations—for incorporation into the guidelines for regulation. We maintain the premier data set covering the content moderation policies of major digital platforms; have given congressional testimony regarding the dangers of the targeted advertising business

model; and contributed to the second version of the Santa Clara Principles On Transparency and Accountability in Content Moderation.

RDR welcomes the opportunity to provide input on UNESCO's model regulatory framework for the digital content platforms to secure information as a public good. We have a long history of collaboration with UNESCO, including the co-publication in 2015 of *Fostering Freedom Online: The Role of Internet Intermediaries*.

This submission focuses primarily on our concerns surrounding the development of this guidance and the recommendations within, especially concerns about UNESCO's mandate to create the framework, the process used to develop it, and UNESCO's decision not to directly address the role of the targeted-advertising business model, which is responsible for the spread of so much undesirable speech.

Though our rankings reveal year-on-year progress, they also illustrate the need for regulation of digital platforms. None of the 14 major platforms we rank has ever achieved an acceptable score on our index, with many failing to disclose details of their content moderation process or offering limited opportunities for users to appeal mistaken decisions. Market forces are not sufficient to address these problems, especially since the business model of major platforms creates incentives that do not align with freedom of expression.

It may be possible to design a useful and generalizable framework for UN member states to implement regulation of digital platforms designed to address content moderation practices in line with human rights. We encourage UNESCO, however, to proceed only if it can guarantee an inclusive multistakeholder process.

Concerns about the mandate

It is not clear that UNESCO has a mandate to create this framework. As the civil society organization Article 19 stated in its comments on the proposed framework "[T]he elaboration of a model regulatory framework requires a decision by the General Conference in line with UNESCO's rules of procedure and that the general reference to UNESCO's global mandate to promote the free flow of ideas by word and image does not constitute a sufficient basis to create a mandate for this proposal."

Poorly scoped or vague regulation is a major threat to freedom of expression, so echoing comments by Article 19 and the Global Network Initiative, we believe that any future work by UNESCO on this framework should be done in partnership with subject matter experts at the Office of the High Commissioner on Human Rights.

Concerns about the process

The process for developing the framework was expedited, with only about a month provided between the first public draft and the deadline for public feedback. This is especially problematic because the draft guidelines invoke a multistakeholder approach, which requires input from a diverse range of organizations, including those who may not have the resources to reply quickly. Indeed, the hastiness of the process contradicts the advice UNESCO has offered to policymakers in the ROAM Principles.

The abbreviated comment period is likely to disproportionately stymie the participation of under-resourced organizations with limited capacity to respond quickly. These often include those representing marginalized groups, which are also excluded from national policy-making in their home countries. For the development of the second version of the Santa Clara Principles, to which the framework refers, multiple organizations undertook an extensive and global consultation process based on proactive and systematic outreach to civil society, even before the first draft was written. While far from perfect, this can serve as a baseline for the scope of consultation necessary to develop guidance applicable globally.

In addition, the framework lacks an evidence-based grounding clearly laying out the harms it attempts to remedy. UNESCO should not proceed with the process without conducting extensive empirical analysis of the current state of regulation of content moderation by platforms, situating it in diverse cultural and legal contexts.

Lack of attention to the targeted advertising business model

As we explained in our It's the Business Model report series, any regulatory approach to minimizing online misinformation, hate speech, or other potentially harmful expression must directly address the way that platforms make money. At RDR, we believe the targeted advertising business model is a chief driver of mis- and disinformation, hate speech, and other types of expression that lead to harm, such as ads for illegal trafficking operations or those containing false political or health information. Protecting people's data and regulating how companies and advertisers are able to use that data to target messages and ads, therefore, is the best way to avoid encroaching on the right to freedom of expression. So while we understand that UNESCO may consider privacy protections as out of scope for these guidelines, it is important to note for reasons outlined below that the speech harms of concern cannot be mitigated adequately through content moderation alone.

The targeted-advertising business model accounts for the vast majority of revenue at all major social media platforms. It entails the delivery of advertisements personalized with user information. That same information is used to create design recommendation engines that algorithmically sort user-generated content in a way that is addictive and that prioritizes virality over accuracy or any other social good. These issues are inherent to the fundamental incentive structure of a targeted advertising-based business model, which, to grow, must maximize the amount of attention and information it can gather about individual users.

The framework does not address the targeted advertising business model at all. The only mention of recommendation engines is in Paragraph 31.2, which states that users should have some control over whether and how they receive algorithmic recommendations.

Keeping the conversation at the level of content, rather than addressing the business model, limits regulation to partial and temporary fixes which can not keep up with the scale and speed with which content harmful to human rights is generated and disseminated. No human enterprise can provide enough content moderation for a major global platform to be free of objectionable content, and no artificial intelligence-based solution has the necessary accuracy to fill the gap without significantly risking the erosion of the right to freedom of expression.

Reducing or eliminating platforms' ability to microtarget ads and user-generated content based on algorithmically inferred attributes would do more to improve the online ecosystem than any approach to content moderation. It would also cause less of a risk to freedom of expression than popular regulatory approaches, which generally use the threat of fines or liability to push platforms to moderate aggressively. Such aggressive moderation leads to mistakes that threaten freedom of expression, especially for marginalized communities, including linguistic minorities, for whom the moderation apparatus is not well equipped.

Instead of focusing on content moderation, regulation should place strict legal limits on platforms' ability to collect user data and employ it for content and ad targeting. This represents a more holistic approach. Through a shift in financial incentives for platforms, these businesses will in turn be incentivized to create true enabling environments for civic discourse, rather than addictive, algorithmically intensified echo chambers powered by privacy-violating data collection. A variety of alternative business models exist, but platforms need not reinvent themselves entirely. Simply basing ads on the content of the viewed page (contextual ads), rather than users'

personal information, would greatly ameliorate the threats to civic discourse and privacy.

Even if governments do not ban or aggressively limit ad targeting, any regulation covering online content should still reduce its threat to human rights by setting basic standards for a more ethical targeted advertising paradigm. As we argued in our essay *We can't govern the internet without governing online advertising*, this should require of targeted advertising systems:

- regular human rights due diligence;
- a highly transparent system of ad moderation with transparency reporting
- the ability for advertisers to appeal ad moderation mistakes; and
- independent audits to ensure appropriate enforcement of privacy and moderation standards.

Specific areas for improvement

Despite our foundational concerns about the drafting process and emphasis of the framework, we will also take this opportunity to highlight some specific ways it could be improved. This section draws on our detailed human rights and transparency standards for digital platforms. These standards were developed through two rigorous consultation processes that sought explicit input from civil society from its earliest stages.

Accountability for all targeted ads (Paragraph 35)

Transparency recommendations for political advertising should be fully extended to cover all advertising. Non-political advertising can be extremely harmful through provision of medical misinformation, incitement to violence, and discriminatory targeting. Further, accurately determining at scale what counts as political advertising is a significant technical challenge that no platform has ever fully solved.

Guidance on multistakeholderism

The framework mentions multistakeholderism repeatedly but does not provide clearly scoped guidance about how states should effectively convene stakeholders or balance their interests.

Policies available in all major languages (Paragraph 37)

Major platforms should have their full terms of service available in the major languages of every country where they operate. The framework currently suggests that using the six UN languages or the 10 most spoken languages in the world could be sufficient. It contradicts freedom of expression to subject people to speech rules that they are not capable of understanding. We submit that it is negligent for a company to operate a platform in a given country without properly accommodating its languages.

Clearly scoping commercial confidentiality (Paragraphs 22.2 and 38)

The framework is laudable in its recommendation for greater transparency by platforms regarding their content moderation. We caution, however, that since the draft mentions commercial confidentiality as a justification for limiting transparency, it should illuminate the boundaries of this concept. For example, commercial confidentiality could reasonably justify a company's decision to keep certain details of its content moderation apparatus secret to prevent spammers from gaming the system. However, it could not justify keeping that apparatus secret to hide preferential treatment of certain users (as in Meta's XCheck program) in the interest of avoiding embarrassment.

Companies must have clear rules explaining the boundaries of commercial confidentiality and a requirement to justify any opacity in their content moderation apparatuses. Without such accountability, the concept of commercial confidentiality can serve as a smokescreen to justify any sort of opacity with even a remote possibility of affecting a company's competitive position.

Banning shadowbanning (Paragraph 33.4)

All users whose posted content is subjected to content moderation actions, including downranking in algorithmic recommendation systems, should be immediately informed and given a chance to appeal the decision. This requirement rests on the human rights principle of due process and is necessary for moderation to be legitimate. The practice of shadowbanning—reducing or terminating the circulation of a poster's content without notifying them—should be disallowed in all cases.

Final remarks

We appreciate that the framework was created with good intentions, but we agree with Article 19 and the Global Network Initiative that its development should not proceed

unless the issues with UNESCO's mandate can be resolved and a true multistakeholder process can be undertaken with sufficient time for a diverse range of groups to participate. Further, we believe that freedom of expression online is fundamentally dependent on reigning in or eliminating the targeted advertising business model, and any approach to regulation must address this. Other areas of improvement should focus on ensuring that the framework is not vague, that it is comprehensive in addressing moderation of all advertisements as well as user-generated content, and that it does not inadvertently harm the right to freedom of expression.