Joint submission from New America's Open Technology Institute and Ranking Digital Rights

4. The RFoM Strategy Paper to Put a Spotlight on Artificial Intelligence and Freedom of Expression (#SAIFE) outlines various challenges to free speech when AI is deployed. What do you consider to be the biggest risk for freedom of expression when it comes to the use of AI? Please specify and, if possible, provide examples on what you, in your field of work, expertise or region, would consider to be the most important issue(s).

The deployment of artificial intelligence creates two key challenges to freedom of expression online:

1. **The Use of Algorithmic Decision-Making Can Result in Overbroad Censorship and Discriminatory Outcomes:** Internet platforms regularly tout artificial intelligence as a silver bullet solution to their content moderation problems. However, researchers have demonstrated time and time again that these tools are unable to accurately moderate content in situations that require subjective judgments, and contextual knowledge. For example, in June 2017, Google announced that it was using machine learning to enhance its detection of extremist content. This resulted in the erroneous flagging and removal of evidence of human rights atrocities, shared by groups such as the Syrian Archive, to raise awareness about extremist propaganda. In addition, as OTI outlined in its report *Everything in Moderation: An Analysis of How Internet Platforms Are Using Artificial Intelligence to Moderate User-Generated Content*, similar flaws can be seen in image recognition tools, which can detect nudity in an image such as a breast, but cannot determine whether the post is depicting pornography or permitted content like a work of art, or an image of breast feeding. Further, researchers have thus far been unable to develop comprehensive datasets that accurately account for the vast fluidity and variance in human language and expression. As a result, these automated tools cannot be effectively deployed across different cultures and contexts, as they are unable to reflect the various political, cultural, economic, social, and power dynamics that shape user expression. Research shows that the consequences of these limitations often disproportionately impact marginalized and vulnerable groups. In the United States, for example, a study at the University of Washington demonstrated that AI trained to detect hate speech could amplify racial biases, revealing that AI models used for detecting hate speech were 1.5 times more likely to flag tweets posted by African Americans and 2.2 times more likely to flag tweets written in African American English, commonly spoken by Black people, as offensive or hateful. What is deemed offensive depends on the social context, but algorithms are unable to judge context when making decisions, and they therefore amplify existing racial biases and societal inequities.

2. **Surveillance-based Business Models Threaten Human Rights:** As Ranking Digital Rights has outlined in its two-part report series *It's the Business Model: How Big Tech's Profit Machine is Distorting the Public Sphere and Threatening democracy* and as OTI

has outlined in its report *Special Delivery: How Internet Platforms Use Artificial Intelligence to Target and Deliver Ads*, the targeted advertising business model that most internet platforms rely on today is a root cause of the proliferation of discriminatory and harmful content such as disinformation, hate speech, and incitement to violence. Further, the ability to target ads and to personalize user-generated content on the basis of demographic and behavioral data causes discriminatory targeting of ads (whether wittingly or not) that is contrary to fundamental human rights principles. So far, platforms have failed to demonstrate that they are able to exercise effective oversight over such practices. While the OTI and RDR reports address the U.S. policy context in general, they are also relevant to ongoing discussions about the regulation of digital platforms in the EU.

In its "It's the Business Model" report series, RDR analyzed how reliance on revenue from targeted advertising incentivizes companies to collect massive amounts of data about their users and other individuals, and to design products that nudge users into spending more and more time on their platforms, thus divulging more and more data. As a former Facebook executive recently told the U.S. Congress, Facebook "sought to mine as much attention as humanly possible... We took a page from Big Tobacco's playbook, working to make our offering addictive at the outset." The report series concludes that content-layer interventions, such as improved content moderation and transparency standards like the ones we recommend, will not be sufficient to address the social harms that social media platforms cause or exacerbate. However, platforms must be much more transparent about their content rules (for both user content and paid advertising), enforcement processes, and the outcomes of these processes than they are currently. OTI, RDR and many other civil society organizations have long called on companies to disclose such information on a voluntary basis. If they continue to fail to do so, it may be appropriate for governments to mandate such disclosures. The OSCE could issue a normative framework or other "soft law" instrument to reinforce this expectation.

While necessary, transparency is not sufficient for platform accountability and governance. Legislative and regulatory interventions should target the business model itself, notably by strictly limiting data collection to that which is necessary to provide the service as expected by the end use (data minimization), enforcing purpose limitation safeguards, and generally enforcing the GDPR more strictly than has been the case to date. In Europe, this will likely require greater funding and capacity for member states' Data Protection Authorities. Additionally, governments should reform corporate governance to ensure that a company's Board of Directors and shareholders can exercise effective oversight over the company's environmental, social and governance (ESG) risks. Specific recommendations can be found in the second report in the series, "Getting to the Source of Infodemics: It's the Business Model."

In its report *Special Delivery: How Internet Platforms Use Artificial Intelligence to Target and Deliver Ads*, (which is part of a broader series of reports that looks at how internet

platforms are using algorithmic content curation processes) OTI also outlined how the use of algorithmic targeted advertising systems can result in disproportionate harms to members of minority groups, such as through discriminatory targeting and delivery of ads for housing, employment, and credit. Algorithmic ad targeting and delivery systems enable advertisers to specify which categories of users they would like to target with their ads. This can result in users receiving relevant ads, but it can also result in the discriminatory exclusion of certain users, even when an advertiser sets non-discriminatory targeting parameters. This is because ad delivery algorithms make inferences based on engagement metrics and other data to identify users that are more likely to engage with an ad. Studies have shown that this can reinforce and exacerbate societal biases regarding race, gender, and socioeconomic status in housing, employment, and credit. For example, algorithms based on historical employment data may reflect patterns of discrimination and suggest that women and minorities would not be interested in certai jobs simply because they have been underrepresented in those fields in the past. Platforms have been slow to take action to redesign their ad algorithms to avoid perpetuating discrimination since digital advertising underpins their business models. Further, these platforms only offer users a limited set of controls over how their data is collected and how it is used to inform targeted advertising and algorithmic recommendations. Although these tools allow users to understand why certain ads or certain content has been delivered or recommended to them, they do not go far enough.

Moreover, the opacity of the online advertising ecosystem (which is largely powered by AI) makes it extremely difficult for regulators and civil society watchdog groups to verify that platforms and advertisers are following the law. For this reason, RDR joined the European Partnership for Democracy and 27 other European and international digital rights organizations in calling for mandatory, universal transparency for online advertising. Such a regime would enable oversight over the online advertising sector, thus better protecting citizens' fundamental rights.

5. How can the understanding of the implications of AI on free speech be increased among all stakeholders, including States, internet intermediaries and the general public? Please provide examples of good practices.

There are two specific approaches that we encourage when it comes to increasing understanding around the implications of AI on free speech can be increased among stakeholders:

- **Greater transparency and accountability from internet platforms:** Internet platforms should provide more granular and meaningful transparency around how their use of automated tools impacts free speech online. Such transparency can be achieved through mechanisms such as transparency reports, ad libraries, as well as through

human rights impact assessments. In its 2020 RDR Corporate Accountability Index methodology, RDR provides transparency and accountability benchmarks for content moderation and curation for online platforms. Similarly, OTI provides recommendations for transparency and accountability in its Transparency Reporting Toolkit on Content Takedowns, its report series on how internet platforms use algorithmic content curation practices, its report on how platforms are addressing COVID-19 misinformation and its latest report on how internet platforms are addressing election-related disinformation in the U.S. context.

- **Multi-stakeholder engagement forums:** Currently, engagements with internet platforms around issues of AI and free speech occur in a siloed manner in which companies engage with civil society, government representatives, and other groups separately. This prevents these groups from sharing insights and working collaboratively towards goals. As a result, there should be greater multi-stakeholder fora which bring together a broad range of actors. These fora should particularly focus on including voices of communities that are disproportionately impacted by these automated tools, such as communities of color and other vulnerable groups.

7. In the #SAIFE Paper, several underlying issues are identified that need to be taken into account when considering the impact of the use of AI on free speech. Internet intermediaries, especially social media platforms, act as gatekeepers by engaging in the selection of information that is published, in the ranking and editorial control over it, as well as in the removal of content. For these interventions, AI-powered tools are often deployed. The business model of most internet intermediaries, which is advertisement-based, builds on the collection and processing of massive amounts of data about their users, feeding into these AI-driven tools. Another underlying issue is that a few dominant internet intermediaries act as particularly powerful information gatekeepers in the online ecosystem. Are there additional underlying issues that need to be taken into account? Please explain.

The prompt adequately captures the central issues at play. Another related issue concerns linguistic diversity and equity, as the AI tools used to govern the placement, dissemination and moderation of both paid and unpaid content perform unevenly in different languages. Companies should be much more transparent about the strengths and shortcomings of their AI tools in various languages, conduct human rights due diligence on the potential impacts of this language-based stratification, and take steps to address these impacts.

8. The #SAIFE Paper addresses the role of "surveillance capitalism" business models in creating AI systems that can threaten freedom of expression and privacy. What

alternative business models, besides behaviourally targeted advertising and the monetization of users' data, can support hosting of user-generated content on a massive scale? Which alternative business model could better protect diversity of voices online, and how could such a model be designed for different kinds of content-hosting and other online services? Please provide examples of good practices.

RDR's "It's the Business Model" reports series links business models predicated on targeted or behavioral advertising (aka surveillance capitalism) to human rights harms notably affecting the rights to privacy, freedom of expression and information, and non-discrimination. While neither OTI nor RDR's research to date has focused on the economic viability of alternative business models that would eliminate or at least reduce the risk, severity and prevalence of such harms, examples include contextual advertising, fee-for-service models, "freemium" subscription models (where basic features are free of charge but more advanced features require payment), and broadly targeted advertising based on data that is freely provided by the user rather collected or inferred by the platform.

9. Do you consider the dominance of a few internet intermediaries to be a challenge for freedom of expression online? If so, which specific concerns do you see?

The intermediary role that large online platforms exercise can threaten freedom of expression, privacy, human rights, and civil rights. Because there are so few platforms dominating so much of the internet airspace, these companies also act as gatekeepers, and so their content policies and practices effectively determine who has a voice online and defines what speech is permissible. Further, large online platforms' reliance on algorithmic tools for content moderation can result in disproportionate harms to members of minority groups, such as through discriminatory targeting and delivery of ads for housing, employment, and credit. In addition, market consolidation can have negative impacts on innovation and create significant barriers for new entrants.

10. Is there a need to create a policy and normative environment that is conducive to a diverse, pluralistic information environment in the AI domain, or to ensure competition to prevent the concentration of AI expertise? Should the "network effect" and limited interoperability of online services be addressed? If so, how?

When platforms acquire smaller companies in adjacent markets, they often acquire user data that can be consolidated to give the platforms a unique competitive advantage, especially in the AI domain. Usage data, information about how individuals use a product, is unique and cannot be easily replicated by competitors. OTI recently wrote to the European Commission to express our views that the data advantage Google would gain by acquiring Fitbit's data would not be remedied by creating a silo for the health and wellness data that would prevent that data from being used for targeting advertising. In-depth merger reviews and competition enforcement are necessary to create a diverse and pluralistic information environment in the AI domain.

OTI also explained that the Google/Fitbit merger would create an incentive for Google to limit the interoperability of competing wearables. Interoperability decreases barriers to entry and facilitates greater competition by enabling new players to offer access to the users on, and at least some of the features of, the entrenched platforms. It also expands the overall market for a particular service or type of service by letting third parties fill in the gaps around the platform's feature set, as many games and other apps have done with Facebook's platform. Interoperability is a promising lever for regulators to use in their efforts to oversee and correct monopolistic abuses amongst the dominant online platforms. It has a unique ability to promote and incentivize competition—especially competition between platforms—and can also offer users greater privacy and better control over their personal data generally.

11. Is there a need for different approaches or different free speech safeguards on AI depending on the specific internet intermediary, their size, capability, extent of risks of human rights impact, and services offered?

Yes, any effort to institute safeguards for free speech related to the use of artificial intelligence must account for the fact that different companies offer different kinds of services. As a result, any transparency and accountability safeguards should be relevant to the type of service a company provides. In addition, it is important to ensure that the institution of safeguards does not place undue financial or resource-intensive requirements on smaller platforms to the extent that it creates an anti-competitive effect. It is much easier for larger platforms to comply with such requirements, but placing such burdensome requirements on smaller platforms can undermine their ability to remain competitive.

12. The #SAIFE Paper addresses how the surveillance of individuals' activities through AI technologies, by States (often relying on data collected through, and shared by, private companies) and by the private sector resulting from their business model, can seriously impede freedom of expression. What are the main risks stemming from the use of AI for surveillance techniques?

Surveillance tools that rely upon algorithms are dangerous for numerous reasons. First, like all algorithms, algorithms used in the surveillance context are trained upon historical data, and therefore perpetuate historical problems, such as racialized policing. Second, these AI-driven tools are often flawed and opaque, and any decisions they make (which are often high stakes in the context of state surveillance) may be not only inaccurate, but difficult to challenge in court. These issues have come to the forefront lately within the context of facial recognition technology in the U.S., where two stories emerged this summer of wrongful arrests and imprisonments based on facial recognition mismatches, both on Black men in Detroit, Michigan. But long before, key studies had demonstrated that facial recognition technology presents alarming inaccuracies, particularly with regard to women and darker-skinned individuals. Further, and most relevant to free expression, such surveillance tech tools drastically increase the scope and scale of state surveillance, and therefore can have a dramatic chilling effect upon free speech and expression. As OTI has recently written, the use of these technologies at recent Black Lives

Matter protests has again highlighted these free expression issues within surveillance tech tools, as well as the need to rein them in.

16. The #SAIFE Paper outlines how the assessment of the (il)legality of content is a complex task, and depends on local context, local languages, and other societal, political, historical and cultural nuances. AI-driven decisions for content removal can fail to understand nuances underpinning the pieces of content, resulting in the filtering and taking down of legitimate content. Is there a need for a "human in the loop" in AI applications? If so, what level of human review or genuine human involvement should be ensured?

Internet platforms often tout automated tools as silver bullet solutions to their content moderation and curation problems. However, as OTI has outlined in its report *Everything in Moderation: An Analysis of How Internet Platforms Are Using Artificial Intelligence to Moderate User-Generated Content*, automated tools are unable to accurately and effectively moderate content when a category of content has fluid definitions. This is often seen with categories of content such as hate speech and extremist propaganda. In addition, as previously outlined, researchers have thus far been unable to develop comprehensive datasets that accurately account for the vast fluidity and variance in human language and expression. As a result, these automated tools cannot be effectively deployed across different cultures and contexts, as they are unable to reflect the various political, cultural, economic, social, and power dynamics that shape user expression. Further, automated tools lack subjective decision-making and are unable to assess context. As a result, yes, companies should only use these tools to augment human review efforts, and ensure that humans are kept in the loop in order to adequately safeguard user speech and rights.

21. What measures, if any, need to be implemented to ensure effective remedies for AI-powered tools? How can it be ensured that those impacted by partially or fully automated decisions enjoy protection against erroneous or discriminatory outcomes? What internal complaint and redress mechanisms need to be installed for users in relation to AI?

OTI is one of the authors of the Santa Clara Principles, which outline minimum standards that companies must meet in order to provide adequate transparency and accountability around its content moderation practices. Two of the Principles emphasize that companies must provide impacted users with adequate notice and with a timely and meaningful appeals process. By providing users with adequate notice, companies can demonstrate transparency around their content moderation process and how it impacts users. In addition, by providing users with a meaningful appeals process, companies give users access to vital remedy and redress mechanisms.

In particular, the Principles outline that companies should provide notice to each user whose content is taken down or account is suspended about the reason for the removal or suspension. At a minimum, this notice should include: 1) the URL, content excerpt, and/or other information

sufficient to allow identification of the content removed, 2) the specific clause of the guidelines that the content was found to violate, 3) privacy compliant information on how the content was detected and removed (flagged by other users, governments, trusted flaggers, automated tools, etc.), and 4) an explanation of the process through which the user can appeal the decision. Notices should be available in a durable form that is accessible even if a user's account is suspended or terminated. In addition, users who flag content should have access to a content log where they can keep track of content they have flagged and the outcome of the moderation processes.

Further, the Principles outline that users should have access to a robust and timely appeals process for any content removal or account suspension. At a minimum, this appeals process should include: 1) human review by a person or panel of person not involved in making the initial moderation decision, 2) an opportunity for the impacted user to present additional information that will be considered in the appeal review process, and 3) notification of the results of the review, and a statement of the reasoning sufficient to allow the user to understand the decision. In the long term, independent external review processes may also be an important component for users to be able to seek redress.

OTI has underscored the need for these notice and appeal mechanisms when platforms deploy algorithmic content curation processes more broadly. In its report series, which looks at how internet platforms use algorithmic decision-making for content curation purposes, OTI outlines the need for platforms to offer meaningful notice and timely appeals processes to users whose speech has been impacted by curation processes such as downranking and recommendation systems. OTI also emphasizes the need for users who flag content to also have access to a notice and appeal function. Further, OTI has also pressed internet platforms to provide greater transparency and accountability around how many appeals they are receiving, how many appeals were successful, and what impact this has had on user speech. Such data is also important as alternative appeals models and structures, such as Facebook's Content Oversight Board, are being established.

In its 2020 RDR Corporate Accountability Index methodology Ranking Digital Rights emphasizes the need for companies to 1) clearly disclose it has a grievance mechanism that enables users to submit complaints if they feel their freedom of expression or privacy has been adversely affected by the company's policies or practices, 2) clearly disclose the company's procedures for providing remedy for freedom of expression or privacy related grievances, 3) clearly disclose relevant timeframe for its grievance and remedy procedures, 4) clearly disclose the number of complaints received related to freedom of expression and privacy, and 5) clearly disclose evidence that the company is providing remedy for freedom of expression and privacy. Further, in the context of content moderation appeals, RDR outlines that companies should 1)

clearly disclose that it offers affected users the ability to appeal content moderation decisions, 2) clearly disclose that the company notifies the users who are affected by a content moderation action, 3) clearly disclose a timeframe for notifying affected users when it takes a content moderation action, 4), clearly disclose when appeals not permitted, 5), clearly disclose its timeframe for reviewing appeals, 6), clearly disclose that such appeals are reviewed by at least one human not involved in the original content moderation action, 7) clearly disclose what role automation plays in reviewing appeals, 8) clearly disclose that the affected users have an opportunity to present additional information that will be considered in the review, 9) clearly disclose that it provides the affected users with a statement outlining the reason for its decision, and 10) clearly disclose evidence that it is addressing content moderation appeals.

23. What minimum standards of transparency should be introduced for the use of AI? What elements should these standards contain?

As outlined in the 2020 RDR Corporate Accountability Index methodology that provides transparency and accountability benchmarks for content moderation and curation for online platforms, the Santa Clara Principles, and OTI's Transparency Reporting Toolkit on Content Takedowns, and report series focusing on algorithmic content curation tools, as well as COVID-19 misinformation and U.S. election-related disinformation, minimum standards for transparency related to the use of artificial intelligence should include:

1. **Transparency standards for user content moderation:** Digital platforms can and should set rules prohibiting certain content or activities, such as toxic speech or malicious behavior. However, when companies develop and enforce rules about what people can say and do on the internet—or whether they can access a service at all—they must do so in a way that is transparent and accountable in order to ensure that freedom of expression and information rights are being respected. As a result, online platforms should therefore:

- Clearly disclose policies describing what types of content and activities are not permitted on their platforms and services and how they enforce these rules (RDR Index, Indicator F3a, OTI's content curation report series). Companies should also disclose their processes for identifying breaches to targeting rules.
- Ensure terms of service are easy to find and understand (RDR Index, Indicator F1a, OTI's content curation report series), and provide prior notice of changes to these terms (RDR Index, Indicator F2a, OTI content curation report series). In addition, to the extent that companies create new and specific policies related to ongoing events such as presidential elections or the COVID-19 pandemic, they should ensure all policies related

to these events are available in one central location and are not inaccessible because they are spread out on disparate webpages (OTI's COVID-19 and election disinformation reports)

- Provide evidence of enforcement of their terms of service (RDR Index, Indicators F4a and Fb) by publishing data on content and accounts removed as a result of breaches to platform rules, and which rule(s) were violated. Further, companies should publish data on how violating content was detected (e.g. through user flags, automated detection tools etc.) which of their products these removals occurred on, how many appeals were received for moderation decisions, and how much content was restored as a result of appeals or proactively based on platform recognition of errors (OTI's content curation report series, OTI's Transparency Reporting Toolkit on Content Takedowns). Companies should publish this data regularly, in a structured data format and ensure that links are consistent and static.

- Disclose evidence of conducting regular, comprehensive, and credible due diligence, such as through robust human rights impact assessments, to identify how their processes for policy enforcement affect users' fundamental rights to freedom of expression and information, to privacy, and to non-discrimination, and to mitigate any risks posed by those impacts (RDR Index, Indicator G4b). Companies should also proactively conduct or welcome audits of their algorithmic content moderation systems to identify any threats to users' fundamental rights such as freedom of expression (OTI's *Everything in Moderation: An Analysis of How Internet Platforms Are Using Artificial Intelligence to Moderate User-Generated Content*).

2. **Transparency regarding the development and use of algorithmic systems for content moderation and governance:** Using algorithmic systems to moderate and govern the dissemination of user content can have adverse effects on fundamental human rights, specifically, the rights to free expression, access to information, privacy, and non-discrimination. Algorithmic content curation, recommendation, and ranking systems play a critical role in shaping what types of content and information users can see and access online. In addition, systems that are optimized for user engagement -- i.e. designed to keep users on the platform viewing more and more content -- can have the effect of prioritizing controversial and inflammatory content, including content that is not protected under international human rights law. Over time, reliance on algorithmic curation and recommendation systems that are optimized for engagement can alter the news and information ecosystems of entire communities or countries. These systems can be manipulated to spread disinformation and otherwise distort the information ecosystem, which can in turn fuel human rights abuses. The development and testing of algorithmic systems can also pose significant risks to privacy, particularly when companies then use the information collected about users to develop, train, and test these systems without the data subject's informed consent. Online platforms that

develop and deploy algorithms should therefore:

- Disclose a clear commitment to uphold international human rights standards in their development and deployment of algorithmic systems (RDR Index, Indicator G1, Element 3), in line with the Council of Europe's Recommendation CM/Rec(2020)1 of the Committee of Ministers to member States on the human rights impacts of algorithmic systems.

- Disclose evidence that they conduct regular, comprehensive, and credible due diligence, such as through robust human rights impact assessments, to identify how all aspects of its policies and practices related to the development and use of algorithmic systems affect users' fundamental rights to freedom of expression and information, to privacy, and to non-discrimination, and to mitigate any risks posed by those impacts (RDR Index, Indicator G4d). Companies should also proactively conduct or welcome audits of their algorithmic content moderation systems to identify any threats to users' fundamental rights such as freedom of expression (OTI's *Everything in Moderation: An Analysis of How Internet Platforms Are Using Artificial Intelligence to Moderate User-Generated Content*).

- Publish policies that clearly describe the terms for how they use algorithmic systems across their services and platforms (RDR Index, Indicator F1d) and to what extent humans are kept in the loop when using these systems (OTI's *Everything in Moderation: An Analysis of How Internet Platforms Are Using Artificial Intelligence to Moderate User-Generated Content*). Companies that use algorithmic systems with the potential to cause human rights harms should publish a clear and accessible policy stating the nature and functions of these systems. This policy should be easy to find, presented in plain language, and contain options for users to manage settings.

- Publish information about whether they use algorithmic systems to curate, recommend, and rank content (RDR Index, Indicator F12, OTI's content curation report series). They should disclose how these systems work, what options users have to control how their information is used by these systems, and whether such systems are automatically on by default or users can opt-in to have their content automatically curated by the algorithmic system.

- Clearly disclose algorithmic system development policies in a way that users can easily access and understand, so that users can make informed decisions about whether to use a company's products and services (RDR Index, Indicator P1b, OTI's content curation report series).
- Clearly disclose that they provide users with options to control how their data is used for the development of algorithmic systems (RDR Index, Indicator P7, Element 7, OTI's content curation report series).

- Clearly disclose whether they use user data to develop algorithmic systems by default, or if users must affirmatively consent to such use of their data (RDR Index, Indicator P7, Element 8; OTI's content curation report series).

- Provide greater transparency and accountability around their use of algorithmic content moderation tools by disclosing what kinds of information datasets contain (e.g. how regionally, linguistically, and demographically diverse the data are), what kind of outputs models generate, and how they are working to ensure tools are not being misused or abused in unethical ways. This should also include data on accuracy rates for human and automated detection and removal, including the false positive, true positive, and false negative rates, as well as the precision and recall metric (OTI's *Everything in Moderation: An Analysis of How Internet Platforms Are Using Artificial Intelligence to Moderate User-Generated Content*).

3. **Transparency standards for ad content and ad targeting**: The ability for advertisers or other third parties to target users with personalized content—based on their browsing behaviors, location information, and other data and characteristics that have been inferred about them—can significantly shape (and in some cases, distort) a user's online experience and information diet. Personalization, which can affect both paid and unpaid content, can amplify offline social inequities and can be overtly discriminatory. It can also result in so-called "filter bubbles" as well as amplify problematic content, including content intended to mislead or to spread falsehoods. Therefore, online platforms that enable advertisers and other third parties to target their users with personalized ads or content should:

    - Disclose evidence that they conduct regular, comprehensive, and credible due diligence, such as through robust human rights impact assessments, to identify how all aspects of their targeted advertising policies and practices affect users' fundamental rights to freedom of expression and information, to privacy, and to non-discrimination, and to mitigate any risks posed by those impacts (RDR Index, Indicator G4c). Companies should also proactively conduct or welcome audits of their ad targeting and ad delivery and optimization algorithms to identify any threats to users' fundamental rights such as freedom of expression and should ensure that sensitive categories of ads such as political ads and housing, employment, and credit ads are subject to human review before they can run (OTI's *Special Delivery: How Internet Platforms Use Artificial Intelligence to Target and Deliver Ads*).

    - Ensure ad content and ad targeting policies are easy to find and understand (RDR Index, Indicators F1b and F1c, OTI's *Special Delivery: How Internet Platforms Use Artificial Intelligence to Target and Deliver Ads*), and provide prior

notice of changes to these terms (RDR Index, Indicators F2b and F2c, OTI's *Special Delivery: How Internet Platforms Use Artificial Intelligence to Target and Deliver Ads*).

- Clearly disclose ad content and ad targeting policies (RDR Index, Indicators F3b and F3c, OTI's *Special Delivery: How Internet Platforms Use Artificial Intelligence to Target and Deliver Ads*).), including which types of targeting parameters—such as audience categories, age, location, or other characteristics—are prohibited. Companies should also disclose their processes for identifying breaches to these rules.

- Provide evidence of enforcement of ad content and ad targeting rules (RDR Index, Indicator F4d; OTI's *Special Delivery: How Internet Platforms Use Artificial Intelligence to Target and Deliver Ads*) by publishing data on the number and type of ads removed as a result of breaches to ad content policies and by disclosing the rule(s) violated, as well as the specific policy that the ad violated, the product it was submitted to run on, how the ad was detected, whether an appeal was received for the removal and what the result of this appeal was. Companies should also publish this data at least once a year and in a structured data file with static links.

- Explain to users why the platform collects, infers, and shares user data. This information should outline the purposes and scope of each of these practices. It should also include an explanation of the risks associated with such data collection, inference, and sharing practices. This information should be easy to access and understand. Users should also have access to controls that allow them to easily manage whether and how data is collected, inferred, and shared, how this data is used, and how it influences the content that they see. This should include the option to delete this data entirely. (OTI's *Special Delivery: How Internet Platforms Use Artificial Intelligence to Target and Deliver Ads*).

- Create a publicly available online database of all of the ads that a company has run on its platform. This can help explain a platform's ad operations in a comprehensive manner and can also enable meaningful trend analysis and research. This database should include ads from all categories of ads on the platform, including categories of ads that could have significant real-life consequences such as political ads, housing ads, employment ads, and credit ads. It should also be user-friendly. In particular, this database should include search functionality. In order to protect user privacy, the information in this database should not enable the identification of users that received the ad. (OTI's *Special Delivery: How Internet Platforms Use Artificial Intelligence to Target and Deliver Ads*). At a minimum, this database should disclose the following information about each of the ads in the database:

- The format of the ad (e.g. text, video, etc.)
- The name of the advertiser
- What region the ad was run in
- How much the ad spend for the ad was
- The time period during which an ad was active
- Granular engagement and interaction information, such as how many users saw the ad, and the number of likes, shares, and views that an ad received
- What targeting parameters the advertiser specified to the advertising platform
- What categories of users the ad was eventually delivered to (i.e. what targeting parameters did the ad delivery system eventually select and optimize for)
- Whether the ad was delivered to a custom set of users or one generated by an automated system (e.g. Lookalike users)
- Clearly disclose that targeted advertising is off by default: users should only be shown personalized ads if they explicitly opt in (RDR Index, Indicator P7, Element 6).

25. How often should transparency reports be made publicly available? Should there be any other criteria for publication?

Internet platforms and governments should publish transparency reports as regularly as possible. Generally, we recommend that they be published at least on a quarterly basis. As outlined in OTI's Transparency Reporting Toolkit on Content Takedowns and in the RDR Corporate Accountability Index, entities publishing transparency reports should also consider the following criteria:

- Publish data in a structured data format: Companies should make all report data available in a CSV (comma separated values) format, rather than or in addition to a flat PDF file. The CSV format is most helpful to researchers, journalists, and others who want to make use of the report data, as it simplifies the data extraction process and makes reports more accessible.

- Link relevant reports to one another: All of an entity's past and present transparency reports should be accessible in a single convenient location. Parent companies that own subsidiaries of products and government agencies whose departments publish independent transparency reports should also similarly collect all of their available transparency reports in one central location.

- Publish reports at static and functioning URLs: In order for transparency reports to be meaningful, they have to be accessible. Maintaining static and functioning URLs is especially important for older versions of a company's transparency report. If a company is acquired or re-branded and the links to its transparency reports subsequently change, they should clearly note this and direct users to updated links.

- Publish reports using a non-restrictive Creative Commons license: Companies should use a non-restrictive Creative Commons license for their reports, which enable others to build on their work even for commercial purposes.

26. How can transparency norms and expectations be harmonized to ensure that disclosures are comparable, accurate, and useful to a broad range of stakeholders?

OTI's Transparency Reporting Toolkit on Content Takedowns and the RDR Corporate Accountability Index both provide detailed and mutually compatible roadmaps for how companies should structure their transparency reports. We strongly recommend that the OSCE and any other norm-setting bodies working on this topic ground their recommendations in a careful review of our work, which is based on years and research and stakeholder consultations. As we have outlined in our past work, a lack of standardization in transparency reporting often makes it difficult to compare company efforts and understand what the larger content moderation or content curation landscape looks like. However, variations in transparency reporting, especially related to the metrics that companies are reporting in, can also reflect the unique contexts, roles, and services that companies play and offer in the digital sphere. These variations also allow innovative and creative thinking around the concept of meaningful transparency. As a result, the OSCE and other norm-setting bodies should ensure that there is a balance between the need for a degree of standardization among platforms' transparency reports and the different contexts and services that companies offer. Any efforts to harmonize transparency tools should not overly restrict companies' efforts to innovate around the metrics they produce and exploring different components and definitions of meaningful transparency.

36. Do you want to add any specific observations in the context of the COVID-19 pandemic, and the tendency, as observed in the #SAIFE Paper, to revert to technocratic solutions, including AI-powered tools, which may lack adequate societal debate or democratic scrutiny?

The unprecedented spread of COVID-19 around the world forced internet platforms to adjust their content moderation operations, as due to privacy and security concerns, many of their content moderators were unable to work remotely. As a result, these platforms announced that

they would be relying more on artificial intelligence. For some platforms, this meant that they would rely more on automated tools to detect certain categories of content, while for others it meant that they would be relying more on automated tools for content review and potentially removal. Some platforms, such as Facebook, also suspended their appeals process amidst the outbreak of the pandemic, leaving users with no mechanism for remedy or redress in the face of increased automated content moderation procedures.

This shift to relying more on automated tools has raised significant concerns. As outlined in OTI's report *Everything in Moderation: An Analysis of How Internet Platforms Are Using Artificial Intelligence to Moderate User-Generated Content*, these tools have proven to be inaccurate when moderating content across cultures and languages, content that has fluid definitions, and content that requires subjective decision-making and contextual understanding. In response to this shift, many platforms have warned their users to expect more content moderation mistakes. This raised further concerns given the rapid spread of COVID-19 misinformation and disinformation online, as well as the rise of election-related misinformation and disinformation in the United States.

As OTI outlined in a blog post with the Electronic Frontier Foundation, during emergency periods such as the COVID-19 pandemic, companies must maintain their commitment to digital rights. Frameworks such as those proposed by the Santa Clara Principles can act as a guiding mechanism during these times. In particular, companies must ensure that given their increased reliance on automated tools, they are providing adequate transparency around their content moderation efforts, are notifying impacted users, and are offering impacted users access to a timely and robust appeals process.

Links to resources:

1. Holding Platforms Accountable: Online Speech in the Age of Algorithms Report Series: https://www.newamerica.org/oti/reports/report-series-content-shaping-modern-era/
2. Everything in Moderation: An Analysis of How Internet Platforms Are Using Artificial Intelligence to Moderate User-Generated Content Report:

https://www.newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-user-generated-content/

3. Maréchal, Nathalie, and Ellery Roberts Biddle. 2020. *It's Not Just the Content, It's the Business Model: Democracy's Online Speech Challenge - A Report from Ranking Digital Rights*. Washington D.C.: New America. http://newamerica.org/oti/reports/its-not-just-content-its-business-model/.
4. Maréchal, Nathalie, Rebecca MacKinnon, and Jessica Dheere. 2020. *Getting to the Source of Infodemics: It's the Business Model*. Washington D.C.: New America. https://www.newamerica.org/oti/reports/getting-to-the-source-of-infodemics-its-the-business-model/.
5. Special Delivery: How Internet Platforms Use Artificial Intelligence to Target and Deliver Ads Report: https://www.newamerica.org/oti/reports/special-delivery/
6. 2020 RDR Corporate Accountability Index methodology: https://rankingdigitalrights.org/2020-indicators/
7. How Internet Platforms Are Combating Disinformation and Misinformation in the Age of COVID-19 Report: https://www.newamerica.org/oti/reports/how-internet-platforms-are-combating-disinformation-and-misinformation-age-covid-19/
8. Protecting the Vote: How Internet Platforms Are Addressing Election and Voter Suppression-Related Misinformation and Disinformation Report: https://www.newamerica.org/oti/reports/protecting-vote/
9. Transparency Reporting Toolkit on Content Takedowns: https://www.newamerica.org/oti/reports/transparency-reporting-toolkit-content-takedown-reporting/
10. Santa Clara Principles on Transparency and Accountability in Content Moderation: https://santaclaraprinciples.org/
11. Facebook's Final Charter for Its Content Oversight Board Takes Encouraging Steps, But Fails to Prioritize Transparency and Accountability: https://www.newamerica.org/oti/press-releases/facebooks-final-charter-for-its-content-oversight-board-takes-encouraging-steps-but-fails-to-prioritize-transparency-and-accountability/
12. The Santa Clara Principles During COVID-19: More Important Than Ever: https://www.newamerica.org/oti/blog/santa-clara-principles-during-covid-19-more-important-ever/
13. AI Proves It's a Poor Substitute for Human Content Checkers During Lockdown: https://venturebeat.com/2020/05/23/ai-proves-its-a-poor-substitute-for-human-content-checkers-during-lockdown/