MEEDAN

1355 Market St., Suite 488
San Francisco, CA 94103
meedan.com

To: Ms. Irene Khan, Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, United Nations
From: Meedan
Date: 15 February 2021
Re: UN OHCHR Report on Disinformation

At this critical juncture in the history of the internet, it's becoming increasingly clear that how we approach online content and disinformation — namely, content moderation — needs a set of standards expressed as a code of ethics. This code of ethics needs to function globally and across platforms, with international human rights norms as the critical foundation.

There is no authoritative definition of "content moderation," but the term broadly refers to the process of monitoring, judging, and acting on user-generated content and the user-generators themselves to enforce policies, rules, or guidelines (often called community standards) determined by a governing body. As a practice, content moderation should be thought of as encompassing a range of *content mediations* from take-downs to fact-checks to content warning flags to account suspensions. To this end, Meedan has begun some of this work with a Content Moderation Toolkit for Civil Society, which we hope can serve as a meaningful contribution to a larger discourse around standards and codes of ethics in content moderation.

Public engagement with content moderation most prominently centers the role of social media companies in moderating content on their platforms. Therefore, social media companies are most commonly the assumed governing bodies. In reality, there are many intersecting layers of governance. As has been highlighted through the COVID-19 pandemic, we must prioritize an infrastructure to support collaboration across stakeholders to respond to disinformation across languages as they inform fact-checking and content moderation efforts for platforms.

At scale, good content moderation has much to learn from good journalism. Fundamentally, Meedan believes that context-aware systems must support but not replace human judgment, and that human judgment should be guided by fair policies and procedures. Indeed, the enormous complexity of responsible and effective content moderation requires many of the same tools as good journalism: a deep understanding of context and power, a commitment to the public trust, and practices of transparency and

accountability. And just as codes of ethics and working standards guide journalists, so must a code of ethics help guide the work of content moderation.

One of the biggest challenges of responsible content moderation that promotes free expression and effectively addresses disinformation and hate speech is the sheer scale of the challenge. With more than half the world's population now online amid a context of hundreds of social media platforms and hundreds of governing interests in how the internet operates, content moderation can look like a wicked problem, creating paralysis in and of itself. To respect rights at this scale, content moderation must be understood as a process, not a solution in and of itself.

To that end, Meedan believes we must break down the issues into a series of five core challenges raised by disinformation that are worthy of investigation:

1. **Stakeholders:** Content moderation decisions are often adjudicated by companies in major technology locations, like Silicon Valley and the Pearl River Delta, and outsourced to workers in places like India and the Philippines, but the consequences are felt globally. These platforms depend on local jurisdictional support to operate and so cannot be considered impartial in matters of political dissent and calls for take-downs from authorities. The Global South must be elevated in this conversation, as must those from marginalized communities such as women, people with disabilities, Black, indigenous and LGBTQ communities, and children and marginalized castes.
2. **Platform affordances and business models:** Technology platforms have different features, designs and functionalities and therefore require different approaches, from public feeds like Twitter and TikTok to private messaging apps like WhatsApp and Signal. In its current state, most internet platform business models seek to shape attention at the expense of any secondary consideration, whether that is the health of a society or the accuracy of a consumer decision. The current business models of large technology companies often prioritize profit over the public interest, and shareholders over civics. Good content moderation must take the public benefit into account, embracing an ethos of "do no harm" that is adaptive for encrypted and unencrypted spaces.
3. **Language:** Language inequities on the internet are a full stack problem, exacerbating the challenges of information. While tech companies and social media platforms want to bring more users to the internet, many parts of the internet have little to offer to users in the primary languages they speak. People who don't speak digitally dominant languages have a smaller pool of information they can access, inferior algorithms that have not been trained for their languages, and often no senior decision-makers at the platforms who speak the language or understand the local context.

4. **Access:** While more than 4.5 billion people are online, their experiences of the internet vary significantly, from zero rating services to high-speed broadband. Online information makes it possible for people to exercise their rights as citizens and seek accountability from the authorities. Yet we've witnessed internet shutdowns and slowing of bandwidth during elections and other public events. When governments shut down the internet during elections and in politically volatile circumstances, the measure is alleged as a way to curb misinformation. In reality, fact-checkers, newsrooms and human rights activists experience storms of misinformation when the internet is either slowed down, or mobile internet and broadband services are completely shut down.
5. **Public health:** Information inequity is a public health issue. Today's world has more internet users than people with access to essential health services such as primary care, dental care, or surgery. Through the COVID-19 pandemic, online information searches serve as a supplement to reduced availability and accessibility of in-person healthcare. The availability of accurate, accessible and relevant information online that can reduce the impacts of disinformation on public health outcomes is a necessary component of realizing the right to health. Content moderation infrastructure must prioritize sustainable and reliable connections between health experts, with an understanding of how local contexts, customs and traditions shape perceptions of health information, and the decision-makers involved in public health content moderation policies.

Given this, we recommend five areas of address:

1. **Map policy and government actions.** While the EU's GDPR and Brazil's LGPD represent the most comprehensive responses so far to content moderation, countries such as Egypt and the Philippines have implemented "fake news" laws which have been used to crackdown on free speech and independent journalism under the guise of preventing the spread of actual misinformation. **Conduct research on the impact and potential of key initiatives.** There is no single, universal solution to disinformation. We need a broader evidence base for the impact of disinformation response efforts in different contexts. This involves ensuring that case studies and bodies of evidence are developed across languages and in collaboration with communities around the world. Such research requires greater academic–practitioner collaboration and data sharing, which will highlight gaps in current disinformation response efforts and opportunities for impact in underserved settings.
2. **Build capacity for civil society, media organizations and fact-checkers and tech workers to work at the scale of the internet.** This includes the development of pathways and data standards to enable subject matter experts to directly support communicators and disinformation responders at scale, and it must involve the support of technology worker activism and collective organizing. Speed and

accuracy are essential for the containment of false and potentially damaging information; in encrypted spaces, protecting user privacy should be complemented with an ability for users to easily and securely query a third party fact-checking or information service with any content they encounter. It is necessary to amplify authoritative voices and credible data in order to ensure that the integrity of the information ecosystem is maintained, clear channels are provided to report bad actors, and support is provided for moderation of content published in local languages.

3. **Localize content moderation approaches:** In addition to strengthening transparency, there needs to be infrastructural shifts in content moderation approaches that allow for unique context layers for different languages, countries and communities, instead of simply applying the same models across all. Content moderation approaches need to take into account local contexts, languages, politics and power dynamics.
4. **Develop technological approaches that augment human efforts.** We should be embracing human-led fact-checking and journalism efforts in the fight against disinformation and not seek to replace them. Machine learning can identify repeated instances of content that has already been fact-checked, spot potential emerging misinformation, and assist fact-checkers, journalists, and general users in putting a piece of content in the wider context. But the priority for combatting all disinformation should be led by local, knowledgeable people and organizations.

In the 18th and 19th centuries, as news consumption transitioned to newspapers and broadcast media, new codes of ethics emerged to deal with the series of harms that emerged in the new media of its day, from yellow journalism to conflicts of interest. We stand at a similar juncture today, as the internet brings forth new flavors of old challenges around free expression, disinformation and rights to health and safety. Fortunately, we can learn from others standards-based efforts, like international human rights law and journalistic codes of ethics, to serve as a foundation for digital content frameworks.

Much of the work of content moderation should have a clear aim in mind: establishing a code of ethics that is grounded in principles of transparency, accountability, impartiality, integrity and minimization of harm. With the increasing complexity of content on the internet and norms around governance, new standards are needed to help bridge the gap between industry and those seeking to engage with industry.

Written with contributions from Meedan team members: Azza el Masri, Eric Mugendi, Scott Hale, Shalini Joshi, Sneha Alexander, Christin Gilmer, Isabella Barroso, Nat Gyenes, Ed Bice and An Xiao Mina.

Meedan's Board of Directors: Maria Ressa, Zeynep Tufekci, Tim Hwang, Hanan Heakal, Jon Corshen, Ed Bice