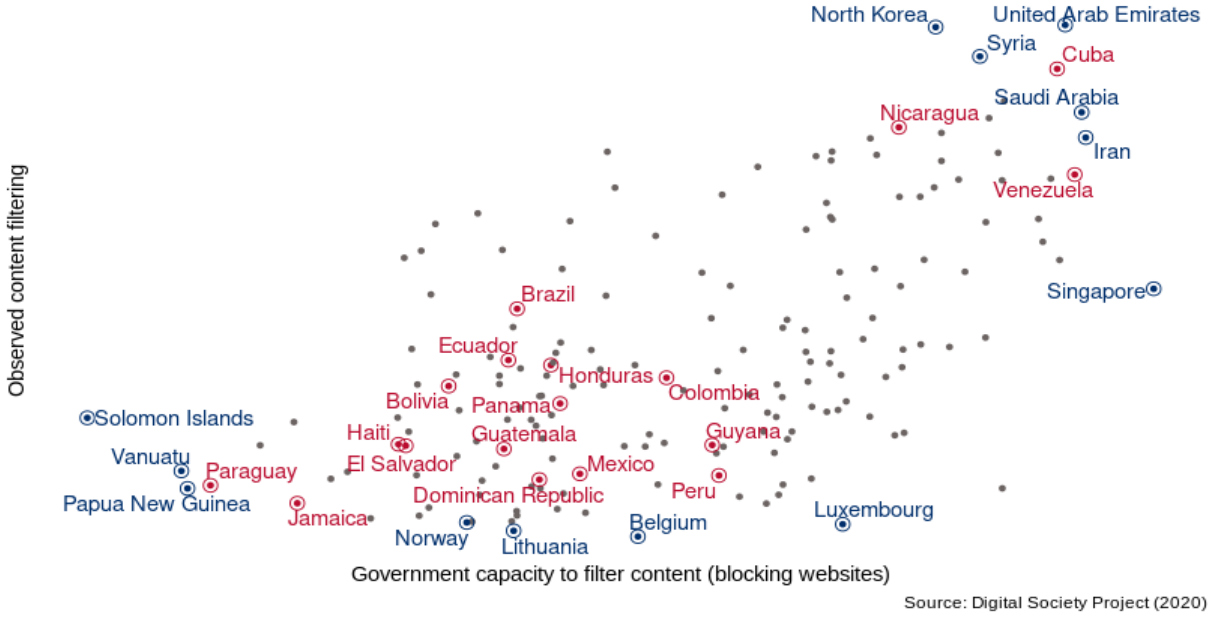
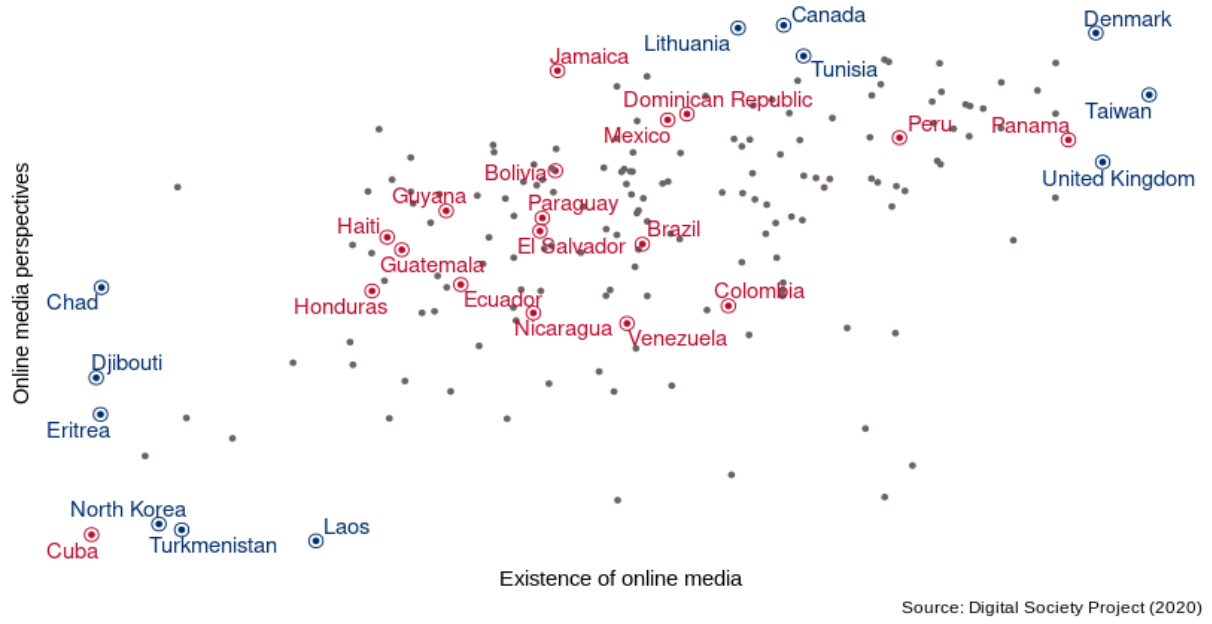
Responses to Systematic Harmful Content on Social Media

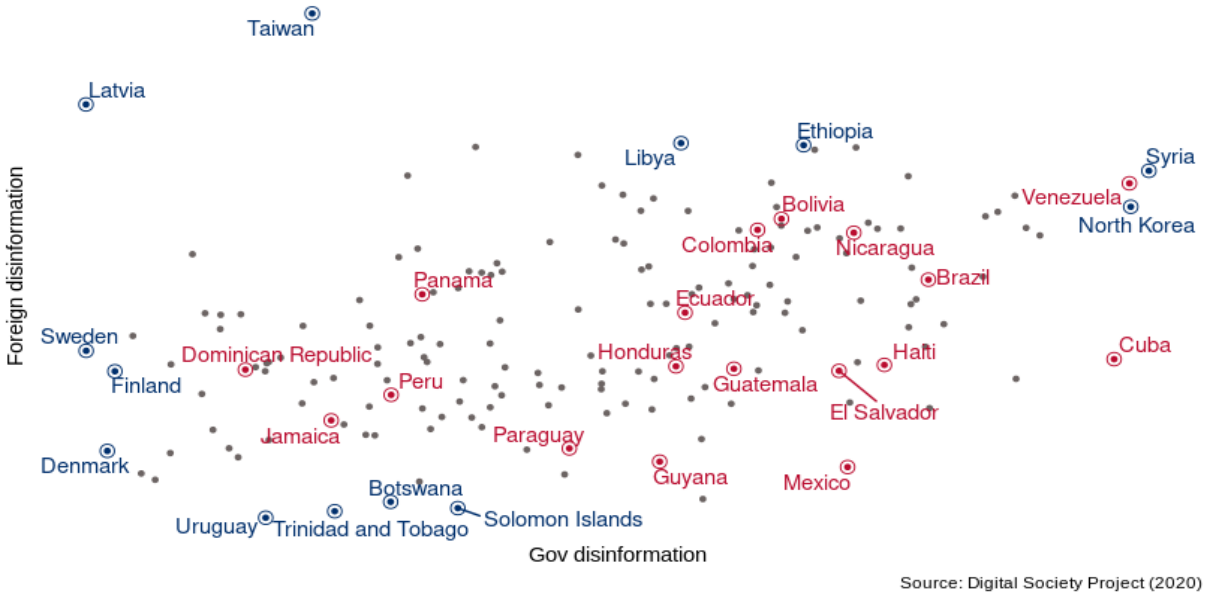
USAID has strong media and civil society programs that track the online conversation and point out inappropriate use. It should be noted that most of these programs are in Asia and E&E. The graph below shows that these regions have been more active in censoring the internet. Still, the LAC region shows worrying trends, especially with Cuba, Nicaragua and Venezuela as clear outliers on the top right.



LAC countries vary widely in both the quantity and diversity of online media. Panama and Peru (top right) are doing as well as many rich countries, whereas Haiti, Guatemala, and Honduras are more restrictive. Cuba (bottom left) is a clear outlier, while Nicaragua and Venezuela aren't too far from the rest of the region.



Regarding government disinformation, Venezuela and Cuba (top right) lean heavily, while others (Brazil, Haiti, Nicaragua, and El Salvador) aren't far behind.



Reporting social media content

Platforms like Facebook/Meta or Twitter are receptive to taking down content when it falls within three categories: 1) harmful content, 2) deceptive behavior, and 3) manipulative actors. However, reported content may not be removed or may only be deprioritized if platform adjudicators determine that it does not violate the terms of service.

- Harmful Content: The content needs to demonstrate violations of the platform's community standards. Hate speech, harassment, and calls for violence usually qualify. Facebook/Meta has set up a "court" that they call "oversight board" to adjudicate incidents, meaning that it's not necessarily a straightforward process.
- Deceptive Behavior: If evidence is provided, most platforms will remove content / accounts demonstrating what Facebook calls "coordinated inauthentic behavior" (e.g., troll farms or bots) violating their terms of service.
- Manipulative Actors: In some of USAID's more developed programs, partners, specifically investigative journalists or think tanks doing online investigations are able to map out the bad actors on social media platforms. They use the evidence to petition the platforms to get 'inauthentic" accounts taken down. Cases are not necessarily straightforward. For example, public figures receive more leniency to remain online in spite of violations of terms of service.

For example, AID/Georgia's Information Integrity Program described the following actions to limit harmful content:

1. Analyze online narratives and identify actors engaging in "coordinated inauthentic behavior". The partner will share a list of accounts with Facebook/Meta who in turn will dig deeper and ultimately take them down or de-promote content. While this work begins "manually" there is technology that can identify patterns that reveal the source of harmful content. It should be noted that online content may fall under different categories (Fake/Manipulation/Out of Context) with different levels of response. Facebook/Meta also works with trusted organizations such as the International Network of Fact-checking Organizations (IFCN),Graphika or the Atlantic Council's digital forensics research lab to do the leg work and flag activities through Facebook/Meta's user interface.
2. Support local fact-checking organizations to apply adequate standards and eventually become part of IFCN and Facebook/Meta's fact checking program. Facebook/Meta's

Trusted Partners program allows vetted partners (e.g. local CSOs) to weigh in on content takedown requests, and for example, make determinations on whether a takedown request is legitimate.

3. Promote benchmarks on social media advertising policies. Facebook/Meta has increased its standards in the U.S. especially after the recent U.S. elections' claims of the site's vulnerabilities. The program is lobbying with Facebook/Meta to introduce the same standards overseas, especially during elections.

The above description has focused on Facebook/Meta. Working with Google, YouTube, TikTok or Twitter has been more difficult as they are behind the curve on controlling harmful content. Nonetheless, addressing takedown requests is incredibly complex as company decisions are informed by their business models and local and international laws on internet governance, among other factors. For example:

- Reporting harmful content can be abused by malign entities seeking to discredit their adversaries.
- Facebook/Meta has decentralized and very spread out "digital threats" teams making it difficult to react quickly. Facebook/Meta, for example, has different enforcement capabilities depending on the market/jurisdiction -- in places where they have a strong market presence or interest in expanding, it employs a much more robust network of content moderators, who are fluent in local languages, meaning the policy enforcement is extremely uneven.
- In some instances, Facebook/Meta chooses not to react for fear of being exposed for housing malicious content.
- The rules are also subject to change. Recently, for example, Facebook/Meta changed its policy to allow users from certain Eastern European countries to call for violence against members of the Russian government and military. This is unprecedented and raises serious concerns.
- The machine learning models and algorithms Facebook/Meta and other social media companies use to verify harmful content is incredibly imperfect, to the point of oftentimes being harmful. For example, it's not uncommon for posts about atrocities or hate speech to get flagged by an algorithm as a violation of the terms of service and get taken down, when they are in fact posted by a journalist or human rights organization to raise the issue.
- Even in places where Facebook/Meta has hired robust and well-trained content moderators, these moderators frequently find themselves subject to abuse, harassment, and threats when content moderation decisions contradict government requests. Additionally, there are new laws being proposed and enacted by various governments that require social media platforms to hire local staff - what some civil society groups are terming hostage-taking laws - that could further put local staff at risk.

Additional Resources

This section does not intend to offer an exhaustive list of all resources at USAID and State. It simply offers a sample to begin the conversation with the experts.

Resource	Description	Links
Innovation, Technology and Research (ITR) Hub *USAID Bureau for*	The ITR Hub's Technology Division works to find innovative ways to address gaps in digital access, affordability, and **advance the use of technology and advanced data analysis** in development.	USAID Digital Strategy Digital Ecosystem Country Assessments (DECAs)

Development, Democracy, and Innovation (DDI)	**Coordinates** on Bureau and Mission strategies.	
Center for Democracy, Human Rights and Governance (DRG) *USAID Bureau for Development, Democracy, and Innovation* (DDI)	As the **hub of global DRG programming**, the Center focuses on the application of **policy, best practices, and research** to enable innovative foreign assistance.	Center for DRG Intranet Site Intro to Disinformation Presentation USAID Disinformation Primer Greater Internet Freedom (GIF) Program Factsheet
USAID MDM Initiative *Led by DRG Center*	Community of practice a) developing a **USAID approach to** misinformation, disinformation, and malinformation (**MDM**); b) working with the interagency and partners to **advance collective learning** and program design on MDM; c) developing evidence-based learning around MDM approaches and best practices to **advance thought leadership.**	
Cybersecurity Team, Technology Division *USAID Bureau for Development, Democracy, and Innovation* (DDI)	DDI Team that a) Provides **technical resources** for all Missions, Sectors, and Bureaus in cybersecurity; b) manages an agency-wide **cybersecurity working group**; c) represents USAID in **interagency cybersecurity discussions** that focus on development programming.	Cybersecurity Team Intranet Site Cybersecurity Primer Digital Security Resource Guide (under review)
USAID MDM Programs in LAC *DRG Center and LAC Bureau*	USAID bilateral, regional and global **counter MDM,** media development, safety, and digital literacy **programs across the LAC Region**.	MDM Programs in LAC
Global Engagement Center (GEC) *U.S. Department of State*	Direct and coordinate Federal Government efforts to recognize, understand, expose, and counter foreign state and foreign non state **propaganda and disinformation** efforts aimed at undermining or influencing the policies, security, or stability of the United States, its allies, and partner nations.	GEC Website Selected GEC Resources Disinformation briefing in LAC
Bureau of Global Public Affairs	**Communicates U.S. foreign policy priorities** to American audiences and engages foreign publics to enhance their	Bureau of Global Public Affairs

U.S. Department of State	support for the values and policies of the United States.	

Guidance for reporting social media accounts and content

1. Facebook
   - Reporting a Message: https://www.facebook.com/help/1709360766019559/?helpref=related_articles
   - Reporting an Abusive Group: https://www.facebook.com/help/266814220000812/?helpref=popular_articles
2. Twitter
   - Reporting Harassment and Threats of Violence or Physical Harm: https://help.twitter.com/en/forms/safety-and-sensitive-content/violent-threats/legal-rep
3. Instagram
   - Reporting Harassment or Bullying https://help.instagram.com/547601325292351