DDIA Public Comment to the Meta Oversight Board on Community-Driven Content Moderation

Submitted by:
Roberta Braga
Founder and Executive Director
Digital Democracy Institute of the Americas (DDIA)

December 9, 2025

Introduction and Context

My name is **Roberta Braga**, and I am the Founder and Executive Director of the **Digital Democracy Institute of the Americas (DDIA)**, a non-partisan, non-profit organization working to create a healthier Internet for Latino communities through research, capacity-building, and policy. DDIA conducts behavioral and narrative analysis across multiple digital platforms, languages, and countries in the Americas, with the goal of understanding how borderless information ecosystems shape public discourse and how they can be strengthened to empower communities.

I am submitting this public comment on behalf of DDIA and alongside two of our research consultants, **Cristina Tardaguila** and **Marcelo Soares**, who led the investigation underpinning our comment to the Oversight Board. This comment includes a summary of our research findings, implications of said findings for Meta's community-driven content moderation program, and overall recommendations. The findings below come from our analysis of over four years of data from **X's Community Notes**, a community-driven moderation model similar to the one Meta is now piloting. As Meta considers how to design, refine, and scale its own community-driven content moderation model, we believe the lessons from X's implementation offer important insights, including ways to improve systemic bottlenecks, lack of transparency, and disparities in multilingual participation.

Summary of Research

DDIA analyzed **1.7 million Community Notes** in English and Spanish submitted between **January 28, 2021, and March 12, 2025**. We examined:

- Program structure and limitations of the public dataset
- Participation trends across languages
- Publication dynamics and bottlenecks
- Behavioral patterns of top contributors

Our analysis relied on X's publicly available Community Notes dataset, which includes:

1. **Notes** (1,764,939 entries)
2. **NoteStatusHistory** documenting movement from submission to publication
3. **Ratings** (276+ million evaluations by contributors)
4. **UserEnrollment**, containing minimal contributor metadata

Because X revoked free API access, large-scale content analysis was not possible; manual content reviews were limited to the top contributors in English and Spanish.

Key Findings

1. Publication Rates Are Extremely Low, Fewer Than 10% of Notes Go Live

Across both English and Spanish, **under 10%** of submitted Community Notes are ever published.

- **English:** Publication dropped from **9.5% in 2023** to **4.9% in early 2025**.
- **Spanish:** Publication rose modestly from **3.6% in 2023** to **7.1% in 2025**, but still trails behind English overall.

This means the vast majority of contextual corrections never reach users, a critical gap when misinformation spreads rapidly.

2. Timeliness Has Improved, But 14 Days Is Still Too Slow for Viral Harms

Average time from submission to publication fell from **100+ days in 2022** to **14 days in 2025**. While this is a major improvement, two weeks is still too long for addressing viral misinformation, which often peaks and declines within hours. Without much faster turnaround, community-driven context tools will struggle to meaningfully reduce harm.

3. Opaque Criteria for "Consensus" and "Diverse Perspectives" Undermine Transparency

For a note to be published, contributors must agree on its helpfulness, and their consensus must come from "a diversity of perspectives." However, the dataset provides **no clarity** on:

- How "different viewpoints" are measured
- How contributors are grouped or weighted
- How many ratings are required
- What the consensus threshold is
- Whether ideological balance is considered or validated

This lack of transparency makes it impossible for researchers or the public to assess the fairness or accuracy of the process.

4. Participation Is Increasing, But Not Resulting in More Notes Published

Contributor enrollment and activity are rising in both English and Spanish:

- In **2024**, more than **126,000 English-language contributors** submitted a note — **double** the number from 2023.
- Nearly **25,000 Spanish-speaking contributors** participated — also more than double year-over-year.

More contributors, however, **have not translated into more notes being published**.

5. Bottlenecks Are Emerging, Especially in English

Even with increased participation, many notes remain **unseen or unrated**:

- **8%** of English notes submitted in 2024 received *no ratings at all*.
- In Spanish, the problem is far worse: **37%** of notes were never rated.

This indicates growing internal visibility bottlenecks, where the volume of submissions outpaces contributors' ability (or incentive) to rate them. These bottlenecks result in thousands of notes stuck in limbo, effectively invisible and unable to move toward publication.

6. Top Contributors Show Imbalances

Manual review of top contributors revealed:

- The *leading English contributor* appeared to be a **bot-like account** submitting over **43,000 notes**, primarily about crypto scams. Only **3.1%** were published.
- The *most active Spanish contributor* submitted **913 notes**, mostly about misinformation related to Venezuela. Its publication rate was **10.7%**.

This disparity reinforces concerns about:

- Overreliance on automated or high-volume accounts
- Skewed issue coverage
- Uneven publication standards by language and topic

It also demonstrates the risk of community-driven moderation being dominated by a handful of hyperactive users rather than a representative community.

Implications for Meta's Community-Driven Content Moderation Program

Meta, like X and other platforms, is embracing community-driven content moderation as a core component of its strategy to address harms on its platform. While this model has potential, our findings illustrate clear risks if deployed without adequate transparency, resourcing, and safeguards.

Below are lessons we urge the Oversight Board to consider:

Lesson 1: Community-driven moderation cannot be the only solution to online harms

The extremely low publication rates, slow turnaround times, and heavy reliance on user capacity show that this model **cannot scale fast enough** to address viral misinformation on its own. It must complement, not replace, more robust detection, enforcement, and editorial review processes.

Lesson 2: Transparency about algorithms, thresholds, and contributor behavior is essential

Without transparency regarding how "diverse perspectives" are calculated, how many ratings are needed, how consensus is determined, and how contributor identities or patterns factor into outcomes, these systems risk reinforcing bias, creating blind spots, or being gamed.

Lesson 3: Language equity must be a core design principle

The disparity in unrated notes (37% in Spanish vs. 8% in English) demonstrates that multilingual ecosystems require tailored design, recruitment, and moderation strategies. Platforms cannot assume that one universal system will work equitably across global contexts.

Lesson 4: Speed must be prioritized

A 14-day delay renders corrective context largely ineffective in combating fast-moving misinformation. Community-driven tools must move toward **near-real-time** functioning if they are to meaningfully reduce harm.

Recommendations

Based on DDIA's analysis, which was published in July of 2025, we recommend that Meta and the Oversight Board consider the following:

1. Do not rely on community-driven moderation as the primary intervention for misinformation and harms that violate Meta's terms of service.
2. Publish clear, auditable definitions of "consensus," "helpfulness," and "diversity of perspectives."
3. Ensure multilingual parity in visibility, rating volume, and publication thresholds.
4. Implement safeguards against bot-like overparticipation and disproportionate influence.
5. Improve turnaround speed to hours, not days.
6. Provide researchers with API access to enable independent evaluation.

Community-driven content moderation can be an effective tool in the toolbox of addressing online harms, but only when transparent, well-resourced, and coupled with complementary moderation strategies.

Thank you for your attention and for the critical work of the Meta Oversight Board. As we have done with TikTok and other actors in this space, DDIA stands ready to support efforts to build safer and healthier digital ecosystems for all communities.