European Digital Media Observatory

Proposal for an Assessment of Risk Mitigations for Algorithmic Amplification of Disinformation, the Role of Platform Business Models & Demonetization

6 June 2024

Final version

Project number:	LC-01935415
Project Acronym:	EDMO
Project title:	European Digital Media Observatory
Start date of the project:	15/01/2023
Duration of the project:	30
Project website address:	https://edmo.eu/
Author of the report:	Jeff Allen, Abagail Lawson
Contributors:	Farhan Abrol, Spencer Gurley, Matt Motyl, Sarah Vieweg, Grady Ward
Submission date:	03/04/2024

This project has received funding from the European Union under Contract number: LC-01935

TABLE OF CONTENTS

This project has received funding from
the European Union under Contract
number: LC-01935

A. INTRODUCTION

This report provides details on indicators to measure the role of platform algorithms and business models in facilitating the spread of disinformation content, and the extent to which disinformation is monetized on platforms. These indicators could be used as part of an assessment of the impact of risk mitigation measures platforms have implemented to safeguard against these particular elements. This report seeks to build on past proposals to assess the integrity of, as well as the effectiveness of measures taken by online platforms, including the service-level and structural indicators for the EU Code of Practice on Disinformation and the impact assessment of the Digital Services Act. This report specifically builds on areas in the first and second EDMO proposal for structural indicators–expanding them to allow for a risk assessment by online platforms. The three indicators explored below reflect areas for further elaboration based on previous assessment proposals–that is, algorithmic amplification, the role of platform business models, and demonetization of disinformation.

What is laid out here will require data or access from platforms on which certain metrics can be calculated by external researchers, as well as self-reporting by platforms of particular metrics. We have identified places where the only realistic option is self-reporting by platforms, and where we believe it is feasible for researchers to conduct external assessment of the indicators below.

The exact scope of the content to be considered under the term "disinformation" needs to be defined by policymakers, as there is currently some misalignment in definitions. Many platforms do not use the term "disinformation" in their policies, but instead work with related concepts such as specific types of misinformation (e.g., "health misinformation") or coordinated inauthentic behavior. As noted by EDMO in the first proposal for a set of structural indicators, the Preamble within the 2022 Code of Practice considers disinformation to include misinformation (false or misleading content shared without harmful intent though the effects can be still harmful), disinformation (false or misleading content that is spread with an intention to deceive or secure economic or political gain and which may cause public harm), information influence operations (coordinated efforts by either domestic or foreign actors to influence a target audience using a range of deceptive means, including suppressing independent information sources in combination with disinformation), and foreign interference (coercive and deceptive efforts to disrupt the free formation and expression of individuals' political will by a foreign state actor or its agents).

In the context of this paper, we identify three possible sources of a definition of disinformation, which could be used for such an assessment, building on common platform policies and the Code definition in order to give an idea of what should be understood to be within the scope of each of the indicators below. The types of content that could be used to inform the public about disinformation on the platforms could include:

- **Content that violates platform misinformation policies** (including any policies around election, health, or other specific misinformation). Depending on platform policies, it is likely that this will not cover all "disinformation" as intended by the Code definition, but it would be a starting point. Platforms would be able to provide the information for a piece of content if it violated their misinformation policies or not. Eventually, a more rigorous practice would involve the inclusion of borderline content by using classifiers to evaluate the likelihood that a piece of content would violate a misinformation policy.
- **Content that violates platform policies around authenticity** (including content from accounts that violate authenticity policies). Platforms would be able to provide the information for a piece of content if it violated the relevant policies.
- **Content that is evaluated to fall under the Code of Practice on Disinformation definition of disinformation**: verifiably false or misleading content with the potential to cause harm, that is spread intentionally, for economic or political gain. As platforms typically do not currently have disinformation policies in place, this would require external fact-checkers or researchers to identify these pieces of content, or for platforms who are signatories to adopt the Code definition and label content provided for assessment in this way.

Whenever such an assessment is performed by platforms, it should include clear parameters for what was included in the definition of disinformation for the purposes of the assessment.

B.INDICATORS & METRICS

1. INDICATOR 1: ALGORITHMIC AMPLIFICATION OF DISINFORMATION

An indicator measuring the role of platform algorithms in amplifying disinformation will have to start with a robust prevalence indicator to understand the scale of disinformation on the platform, which can serve as a baseline to understand what portion of its reach is being promoted and recommended by platform algorithmic systems. An indicator for prevalence was laid out in the EDMO proposals, and is adapted for the sake of a possible self-assessment by platforms.

1.1. Prevalence

Platforms should perform a comprehensive analysis of disinformation content on their platform, resulting in a prevalence measure. This is not just a measure of how many pieces of content violated their misinformation or authenticity policies compared to the total number of content on

This project has received funding from the European Union under Contract number: LC-01935

the platform, as this can be misleading. A comprehensive prevalence estimate should include the following:

1. **A random sample of public content weighted by views**[1] (at least 15,000 samples to identify prevalence of disinformation at 0.1% level with 95% confidence – see Footnote 2) in the monitored period, per member state and language (to estimate the prevalence of disinformation), which includes the following data for each entry in the dataset:
    a. reach (total unique views in the monitored period per member state)
    b. engagement (total number of interactions - depending on the service in question: i.e. comments, shares, and reactions with disinformation in the monitored period per member state).
    c. modality of content (audio, text, video, etc.)
    d. how each user came to view that content (see the options that should be detailed under point 8 of the "Causes of Exposures" section.)
    e. Information about monetization (was the impression on an ad, was the content in the impression monetized, etc.)

With this dataset, platforms should describe the criteria that they use to exclude pieces of content from this sampling, and describe the relative percentage attributable to each exclusion. For example, platforms should report on the volume of content that was excluded because it wasn't public.

Sample Size

The sample size and prevalence will impact the margin of error, and the statistical confidence in finding the type of content of interest. To estimate the margin of error at different sample sizes for different prevalence estimates, the following equation can be used:

$$M = z * \sqrt{\frac{P * (1 - P)}{n}}$$

Where,

- M = Margin of error
- z = z-score for a given confidence level

[1] Ideally, platforms would provide a random sample weighted by views, along with a truly random sample of content. There has been access to random content samples available for research before (e.g., the Twitter Sample Stream / Firehose, Bluesky has an open firehose now). This, alongside a view-weighted dataset would be the most robust examination of how much disinformation is on the platform.

This project has received funding from the European Union under Contract number: LC-01935

- P = Prevalence
- n = sample size

In some instances the margin of error may exceed prevalence, which indicates that the sample size is too small to have sufficient statistical confidence that the sample content provided would contain at least some examples of the content of interest.[2] We can use the following equation to determine the minimum sample size required to have sufficient statistical confidence that the sample would contain at least some examples of the content of interest:

$$S = z^2 \times P \times \frac{1 - P}{M^2}$$

Where,

- S = Sample size needed
- z = z-score for a given confidence level
- P = prevalence
- M = Margin of error

2. **Based on the sample, calculate:**
    a. Prevalence: How many exposures there were to disinformation content in the dataset
        - This will involve categorizing each piece of content according to the definition of disinformation above (which pieces of content violated the platform's misinformation policies, authenticity policies, or were evaluated to fall under the Code definition of disinformation).
    b. Use prevalence to estimate how many exposures there truly were to violating content (prevalence * total_number_of_views)
    c. Report the difference: Estimated views from prevalence - views on known violating content
3. **A total number of contents identified as disinformation.** This should also include information about the reach of disinformation:
    a. **The number of exposures to known disinformation occurred over X time window** (ideally this time period would be every 30 days, but for the initial reporting period, reporting over the 6 month time period would be acceptable)

[2] Different levels of confidence and margins of error may be tolerated depending on the available sample size. In the case of self-assessment by platforms where access to large samples is possible, a minimum size of 15,000 samples would be ideal to ensure with 95% confidence that content with .1% prevalence is present in the sample, while keeping the margin of error under half the size of prevalence.

This project has received funding from the European Union under Contract number: LC-01935

   b. **The number of users that were exposed to known disinformation over X time window** (ideally this time period would be every 30 days, but for the initial reporting period, reporting over the 6 month time period would be acceptable)
   c. **The volume of the problem in the platform as extrapolated from the sampled rates:** Platforms should report the numbers from points a and b alongside their implied total volumes: the sampled numbers themselves have enough caveats that are subject to platform discretion that platforms should re-contextualize them through partially known numbers (like the percentage of the platform's corpus sampled) and unknown numbers (like the number of platform users). That would allow better media reporting on these figures and help regulators gain better insight into their relative magnitudes.

4. **A sample of TOP N (indicative number: 500) pieces of disinformation**[3] in a country, using the following metrics:
   a. reach (total unique views in the monitored period per member state)
   b. engagement (total number of interactions - dependent on the service in question: i.e. comments, shares, and reactions with disinformation in the monitored period per member state).

1.2. Algorithm Components

We can use the prevalence measure to study the role of algorithms by examining how prevalence of disinformation changes as a function of the various machine learning classifier scores used in platform ranking systems. Specifically, this should include the following measures:

5. **What are the most important classifiers used to rank and recommend content?**
   a. This will have to be provided by the platforms, and should be part of any systems description (a crucial component of an algorithmic risk assessment).
   b. The "most important" classifiers are those that have the most weight in impacting the final ranking score, for example by sharing the classifiers that have the highest coefficients in the Value Model.[4] However, platforms may have other ways of calculating importance of classifiers, so the broad requirement is that platforms provide transparency around what their method is for computing

[3] Reflects prevalence indicator from the EDMO proposal.

[4] The Value Model: This is the last layer of ranking which trades off between different things the platform cares about, for example (p=probability): p(click), p(share), p(view) among others. These are all weighted in different amounts to come up with a score for a piece of content. Similarly there are p(misinfo), p(harmful) which are used to downweight content too. Having a sense of which are the highest weighted levers will help understand which values are having the greatest impact on ranking content.

This project has received funding from the European Union under Contract number: LC-01935

importance of the classifiers, and are then transparent about what the most important classifiers are resulting from that method.

6. **How does the prevalence of disinformation change as a function of these classifier scores**?
    a. This would not require platforms to compromise proprietary information or share their code. They can report this information as **inputs** and **outputs** of their systems e.g.: When content is categorized by X classifier score, does the prevalence of disinformation increase as the X classifier score increases?
7. **What is the distribution of key classifier scores for disinformation content, compared to all content?**
    a. Similar to above, this gives another piece of information about how algorithm components and optimization are elevating disinformation.
    b. Distributions for disinformation content compared to other content should also be measured for key features used in ranking and recommendation algorithms, and final ranking scores. This could also be part of the risk assessments carried out by VLOPs and VLOSEs under the DSA.

While this information will give us a sense of what role the platform's algorithms are playing in amplifying disinformation, causal connections between changes in prevalence of disinformation and changes to the algorithmic systems would be more difficult to establish across assessment cycles. For one, there are external factors that will influence the prevalence of disinformation (e.g., around elections) that may have nothing to do with a change to the algorithm. While platforms will usually run A/B tests around algorithm changes which would show the impact of the change on the volume of harmful content, they may not run a full prevalence estimate on the A/B testing groups to be able to compare prevalence metrics directly.

1.3. Cause of Exposures

Understanding prevalence helps illustrate the volume of disinformation on a platform, how well the content moderation systems are catching it, and how many users are being exposed. The key to understanding the role that algorithms are playing is to understand the causes of *why* and *where* the users are being exposed to disinformation. This includes examining where users are encountering disinformation content and accounts that are deemed sources of disinformation[5], and whether the platform played a role in distributing the content to the user or if it was purely the result of present and prior user action. If the platform played a role, then it will be important to understand how that role functions in facilitating exposures to disinformation (e.g., how disinformation content performs in the key classifiers used to rank and recommend content).

[5] As defined in EDMO's proposal for SI-2: Sources of Disinformation.

This project has received funding from the European Union under Contract number: LC-01935

For all exposures on identified disinformation content, the following metrics should be measured:

8. **Discretion in Display: To what degree did the platform cause each exposure to disinformation content?** This information should be detailed in the Prevalence datasets. For each exposure or piece of content, it should be listed where the exposure occurred, defined along these four categories:
    a. The content was shown in direct response to user action (ex: search, navigated directly by URL)
    b. The content appeared via a subscription (ex: podcast feed, followed shared source, groups, messages)
    c. The content appeared via a subscription that the user made based on a platform recommendation
    d. The content was proactively recommended by the platform (ex: ads, proactive content feeds)

If multiple sources apply, the platform should select the least harmful option (a, b, c, d in ascending order of harm) so as not to overestimate the role of the platform's systems in exposing the user to the content.

There are additional categories that could provide more nuance to this understanding and insight into dynamics at the edge of what qualifies as a "proactive recommendation". For example, a user seeing a post in their feed because someone they follow commented on that post is one degree removed from a post created by an account the user has explicitly chosen to follow. Eventually, platforms should include these edge categories in the range of discretion in display.

Using the above information, percentages should be calculated to demonstrate the weight of the platform's discretion. These should include:

9. **What percentage of exposures to disinformation came from product surfaces that are algorithmically ranked?**
    a. This can be calculated as a percentage of total disinformation (provided under Prevalence), and include a definition of which surfaces are included in the category of "algorithmically ranked".
10. **What percentage of exposures to known disinformation content were recommended to the user, meaning the user did not follow the account sharing disinformation?**
11. **For the remaining exposures where the user did follow the account sharing disinformation, what percentage of the time was that follow the result of an algorithmic recommendation of the platform recommending the account?**

This project has received funding from the European Union under Contract number: LC-01935

a. This is important to demonstrate the extent to which algorithms are not just amplifying content but sources of disinformation.
b. By way of example, it was reported that on Facebook, 64% of joins to groups that share violent extremist content were the result of the platform recommending those groups to users. This demonstrates the significant role that the platform's recommender system way playing in getting violative content in front of users, even when the user technically subscribed to the accounts or groups.
c. While some platforms may not have this detailed logging in place, many VLOPs likely will, and all signatories should be encouraged to put this into place.

12. **What is the "amplification factor" for disinformation posts?**
    a. This can be calculated per user that posted disinformation: how much engagement and views did each user get for their disinformation content vs. the engagement/views of non-disinformation content they posted. This should be expressed as the difference in the average log of the engagement for disinformation posts and the average log of the engagement for non-disinformation content posted by the account.
    b. Companies should report the amplification factor for each source of disinformation, along with the mean, median and standard error of that ratio for each week within a 30 day period.
13. **What percentage of exposures were the result of a content sharing mechanism on the platform, such as a reshare on Facebook or a repost on TikTok?**
    a. Resharing capabilities have been shown to increase the spread of harmful content on platforms, and the probability that a piece of content will be reshared can contribute to its final ranking score in the platform algorithm.

2. INDICATOR 2: BUSINESS MODEL

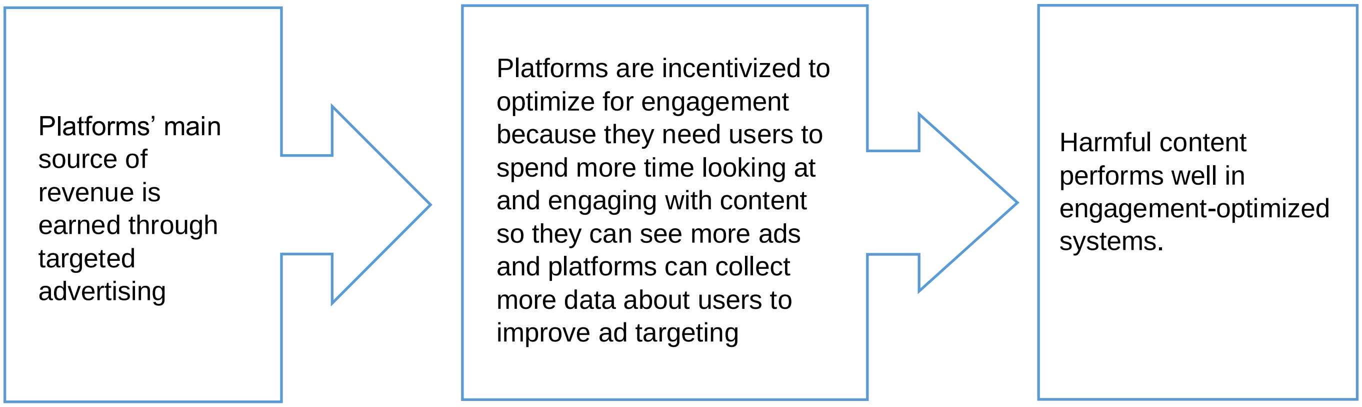
The metrics above can give us a good understanding of the role that platform systems (and thus, design decisions) are playing in spreading disinformation. But it is also important to understand the incentives that are driving these decisions and design of algorithms, and the extent to which these incentives contribute to the spread of disinformation. This requires examination of platform business models.

It is difficult to come up with a metric that clearly lays out "X impressions on disinformation led to $X for the platform", although we can examine the ways in which the structure of the platform's business model may contribute to the amplification of disinformation. We can also, to some extent, measure how platform monetization schemes (including advertising and revenue sharing programs with users), which generate revenue for the platforms, are playing a role in the distribution of disinformation. The goal of this indicator should be to get platforms to think through how they might be financially benefiting from disinformation in all their products.

This project has received funding from the European Union under Contract number: LC-01935

While we address advertising here, we address revenue sharing from monetization programs with users in the Demonetization indicator.

Plenty has been written that connects platform business models to societal harms, and the oversimplified equation looks something like this:



Platforms' main source of revenue is earned through targeted advertising

Platforms are incentivized to optimize for engagement because they need users to spend more time looking at and engaging with content so they can see more ads and platforms can collect more data about users to improve ad targeting

Harmful content performs well in engagement-optimized systems.

This assumes that the third block in the above graphic is true: that disinformation generates more engagement than other types of content, performs well in engagement based systems, and will therefore be profitable for platforms. There is sufficient evidence that this is the case.[6] The indicators in the previous section are intended in part to measure this. However, to understand the role the platform's business model is playing, we have to examine the first part of the equation: the links between profit/business interests and systems optimization. This can be done through understanding how platforms are measuring their own success and how they are measuring safety, and how those measures are weighed in decision making processes. Examining the business model will also require metrics for understanding the extent to which targeted advertising is amplifying disinformation and generating revenue for the platform.

2.1. Metrics and Decision Making Processes

The following metrics are intended to evaluate the ways platforms weigh user engagement, or other business considerations, against integrity concerns relevant to disinformation, and how platform business models may be incentivizing certain considerations. The following will require

[6] See for example, *What We Know About Using Non-Engagement Signals in Content Ranking*, Tom Cunningham and Sana Pandey and Leif Sigerson and Jonathan Stray and Jeff Allen and Bonnie Barrilleaux and Ravi Iyer and Smitha Milli and Mohit Kothari and Behnam Rezaei (2024), arXiv:2402.06831]; Katarzyna Szymielewicz (March 2024) "Safe by Default," Panoptykon Brief, https://panoptykon.org/sites/default/files/2024-03/panoptykon_peoplevsbigtech_safe-by-default_briefing_03032024.pdf; *On Risk Assessment and Mitigation for Algorithmic Systems*, Jeff Allen and Abagail Lawson, Integrity Institute Report (Feb. 2024), *https://integrityinstitute.org/news/institute-news/risk-assessment;* and "Why Is Instagram Search More Harmful Than Google Search?" Jeff Allen, Integrity Institute Blog (Feb. 13, 2024), https://integrityinstitute.org/blog/why-is-instagram-search-more-harmful-than-google-search.

This project has received funding from the European Union under Contract number: LC-01935

self-reporting by platforms, and for VLOPs and VLOSEs, these are also part of our risk assessment framework.

1. **How does the platform measure success?**
    a. What are the topline metrics the company uses to measure success? How many of them are engagement or growth focused, and how many of them are quality, harm-reduction, or disinformation focused?
    b. From this information, it can be explored how disinformation content can contribute to these metrics. For example, if the company measures total engagement on the platform as a metric of success, then engagement that goes to posts that contain disinformation will contribute to the apparent "success" of the company.
2. **Assessment of the role that integrity metrics play in A/B tests or launch processes used to evaluate products and changes, specifically:**
    a. What are key metrics used in A/B tests to evaluate product changes? (E.g., companies should provide the 25 most common metrics included in their A/B tests)
    b. How many (what percentage) of those are metrics that measure the impact on harms, and how many of those are specifically disinformation focused?
    c. How does disinformation content perform on key metrics used in A/B testing?
    d. What metrics are included in the company's A/B testing process to protect against an increase in disinformation?

Insight into this indicator will require platforms to share information about decision-making processes and criteria about how they weigh business interests against the spread of disinformation. It may be the case that a platform uses "conversion factors" to translate engagement to integrity. For example, a company could say that "reductions in the prevalence of violating content are as valuable as engagement". This would then enable any integrity features to launch, as long as the reduction in impression on violating content was equal to or greater than overall reductions in engagement. For example, if a new feature reduced violating impressions by 1% and overall engagement by 0.5%, then it could launch, but if it reduced overall engagement by 2%, then it would be held back. Companies may establish different thresholds for this through guardrail agreements between teams working on integrity and those working on engagement/growth that set thresholds for these conversions. Ideally, the company would have a measure of success that does not have frequent tension with reduction in disinformation, but understanding what measures companies have in place will demonstrate the literal value that companies place on reducing disinformation against their growth. The following data would illustrate this:

3. **How are considerations about potential spread of disinformation weighed against business interests?**

This project has received funding from the European Union under Contract number: LC-01935

   a. What are the conversion factors or guardrail agreements used to evaluate product changes that impact growth and integrity metrics?
   b. How are safety and teams tasked with studying disinformation integrated into product teams, and what level of influence they have. Have there been products blocked from launching because of concerns about disinformation?
   c. What teams have influence over content moderation policies and decisions? Do sales or marketing staff have influence? Do PR or lobbying staff?
   d. How do product launches get approved, and who has the power to delay or deny a launch?
   e. Does the company have a measure of success that does not frequently correlate with an increase in prevalence or views on disinformation?

2.2. Prevalence of Disinformation in Ad Content

4. **For all exposures on identified disinformation content, calculate the percentage of exposures to known disinformation that occurred on paid advertising posts.**
5. **A random sample of advertising content on the platform (weighted by views) that is evaluated against disinformation policies.** This should include the following measures:
   a. What is the overall percentage of ads that were identified as disinformation in the sample.
   b. What percentage of the ads identified as disinformation in the prevalence process escaped moderation or identification in their standard content moderation and ad selection processes. Meaning, what disinformation is uncovered for the first time in the prevalence measure, rather than disinfo that was identified naturally by the platforms systems and processes.
6. **How many ad dollars did the platform make from disinformation?** This includes the following specific measures:
   a. How much revenue was generated from ads purchased by actors that were later found to be violating platform disinformation policies?
   b. How much revenue was generated from ads that were later found to violate platform disinformation policies?

3. INDICATOR 3: DEMONETIZATION OF DISINFORMATION

To consider monetization, we have to look at two flows of money. Sources of disinformation can make money on posts, as well as the platforms, from their involvement in platform monetization programs (including through ads), as well as from the traffic to a domain owned by sources of disinformation generated by these accounts and content. To the extent it is

This project has received funding from the European Union under Contract number: LC-01935

possible to measure the revenue generated for platforms by disinformation content, we have included these metrics under SI-2 on the platform business model.

This indicator will focus on how sources of disinformation are generating revenue for themselves, by monetizing their content. This is hard to measure in a comprehensive way, because monetization can happen through third party avenues that fall outside official platform monetization programs.

3.1. Official Monetization Programs

The first metric will cover the level to which a platform's official monetization programs are supporting sources of disinformation. This will not give a complete picture, but a place to start to understand the strategies of disinformation sources and gaps in platform policies. Information should be analyzed at both the account and content level.

1. **Platforms that provide monetization programs for users[7] should share what percentage of exposures on known disinformation content come from sources that are in their monetization programs.** This can be calculated from the dataset in SI-1, if the platforms provide enough metadata around each sample to know which sample includes content that was monetized.
    a. Additionally, what percentage of exposures on content in their monetization programs was on disinformation content?
2. **Platforms should release a sample of all content (weighted by views) that was eligible for monetization through all its different programs, and calculate a prevalence metric:**
    a. How much of monetization-eligible content is disinformation?
    b. How much of that identified disinformation was only found through the prevalence calculation?
3. **How much money was paid out by the platform for disinformation content eligible for monetization?** This should also specify how much revenue was generated for the platform through any revenue-sharing program, which would contribute to SI-2 above.
4. **How much money was paid out to accounts that:**
    a. **Have ever posted disinformation?**
    b. **Were since removed from the platform for violating disinformation policies?**
    c. From this, a revenue sharing breakdown could be calculated to understand how much money the platform generated from disinformation purveyors, and could contribute to SI-2.

[7] Such as the YouTube Partner program, the TikTok and Instagram Creator funds, Instant Articles and Ad Breaks on Facebook, subscriptions on X and Substack, X "Ads Revenue Sharing," etc.

This project has received funding from the European Union under Contract number: LC-01935

3.2. Off-Platform Monetization

Accounting for revenue generated outside of official platform programs will be more difficult to measure in a comprehensive way. A possible starting place is to measure the prevalence of disinformation content that was not in an ad or monetization program but had other indicators, such as the use of #ad usage in captions or content.

5. **For known disinformation content, what percentage of impressions occurred on content that was not in official monetization programs, but included markers such as #ad, #sponsored, #paid or affiliate link?**

Another measure of the success purveyors of disinformation are having in promoting their content could be taken by looking at the top accounts by popularity in a given time period (popularity defined as total views, total engagements over a time period).

6. **For the top N accounts by popularity in a time period, measure how many:**
    a. are officially "monetized" through platform programs
    b. are unofficially monetized (e.g., by often driving people to a different domain full of ads or selling things)
    c. Are officially monetized and have spread more than 2 pieces of disinformation in the last 90 days.
    d. Are unofficially monetized and have spread more than 2 pieces of disinformation in the last 90 days.

C. CONCLUSION

While we have presented this as a self-assessment method by platforms, questions about how these metrics can be measured in a consistent way by outside researchers remain open. In their latest proposal on Structural Indicators for the EU Code of Practice, EDMO considers “API access as the most feasible approach in accessing platform data for testing and implementing the structural indicators.” More comprehensive researcher APIs would enable measurement of many of these metrics, although there are some exceptions (particularly under Indicator 2 Metrics & Decision Making, such as the internal metrics that platforms use to measure success). Another approach may be to ask platforms to provide data and datasets they have in their own schema, along with explanations of their own definitions and metrics, and then a third party researcher can engage in the work of cleaning and calculating the specific metrics. But even in this instance, there are still some metrics that could only be realistically calculated or shared by the platforms themselves through specific, comprehensive data-sharing with researchers.

There are additional data points that platforms could provide to increase the confidence in the data they are sharing. First, if all the datasets with the relevant components outlined under the first section should be shared, researchers should be able to validate many of the other

This project has received funding from the European Union under Contract number: LC-01935

measures requested. Aside from this and importantly for all the metrics, platforms should provide transparency around how the metrics are calculated by providing the query or MapReduce spec used to generate the statistics concretely.

Acknowledging the reality that all of these metrics cannot be calculated without significant platform cooperation, we suggest that the most robust approach would take advantage of the opportunity provided by the DSA as an instrument for risk mitigation. It is our position that all of these metrics could be reported by platforms (or calculated by researchers based on datasets provided by platforms) as part of their risk assessment obligations under the DSA, and should certainly be a part of platform audits. This would only apply to those platforms that are VLOPs/VLOSEs, but could set a “best practice” standard for the broader industry. It is also true that if platforms provide sufficient data access along the lines outlined above, then a large portion of this methodology could also be used in an expanded third party assessment of structural indicators under the Code of Practice.

This project has received funding from the European Union under Contract number: LC-01935