DISINFODEX

Gülsin Harman, Rhona Tarrant, Ashley Tolbert, Neal Ungerleider, Clement Wolf
Contact: disinfodex@googlegroups.com

EXECUTIVE SUMMARY

This paper outlines the creation of the Disinfodex project, an online database indexing public disclosures of disinformation campaigns issued by major online platforms, currently including Facebook, Instagram, Google, YouTube, Twitter, and Reddit. The project was created by participants of the Assembly:Disinformation Fellowship at Harvard's Berkman Klein Center.

By aggregating the resources in one searchable database, Disinfodex provides a way to analyze publicly available information on actions taken against disinformation networks from these companies. This paper explores considerations taken into account when building the database, which continues to be a work in progress at the time of this paper's release, and outlines current practices around public disclosure of disinformation campaigns by major online platforms.

Advancing the discussion further, the paper explores the current landscape of publicly available information for those researching and working to combat online disinformation, and examines how the Disinfodex project may be built upon to create a useful shared infrastructure within the space. The paper frames an approach to thinking about challenges and risks of building a shared infrastructure, and calls for future steps for collaboration.

DEFINITIONS

The study of online disinformation is fraught with definitional challenges, some of which we briefly touch base upon in this paper. We do not purport to resolve those questions, and as such, rely upon the following definitional framework:

- Claire Wardle and Hossein Derakhshan's 2017 definitions of **Disinformation, Misinformation, and Malinformation**[1]:
    - *Misinformation* is when false information is shared, but no harm is meant.
    - *Disinformation* is when false information is knowingly shared to cause harm.
    - *Malinformation* is when genuine information is shared to cause harm, often by moving information designed to stay private into the public sphere
- The Rand Corporation's stated definition of **Influence operations**[2]:
    - Information operations and warfare, also known as *influence operations*, includes the collection of tactical information about an adversary as well as the dissemination of propaganda in pursuit of a competitive advantage over an opponent
- For the sake of legibility, we will use the terms **"disinformation" or "disinformation campaign" as proxy** for the above, where it is the case that a given sentence can and should apply to multiple or all of these fields of study.

[1] Wardle, Derakhshan, 2017 – Information Disorder: Toward an interdisciplinary framework for research and policy making

[2] Rand Website, April 2020

In addition, we use “**platforms**”, “**social media companies**”, or “**companies”** interchangeably to refer to the broad ensemble of online services that malicious actors have used or could use to propagate disinformation. This includes social media companies *stricto sensu* but also search engines, online forums and bulletin board systems, photo or video sharing platforms, messaging services, and more.

Finally, when we refer to the “**actions**” that these platforms take, we mean deliberate interventions targeting specific accounts or content on the platform, usually in response to a violation of its policies or terms of use, as opposed to the normal course of product development. We call “**disclosures”** the reports that platforms publish about these actions, and **networks** the groups of coordinated accounts that platforms act against.

I. DISINFODEX

Over the past few years, there has been an increase in research, reporting, debate, and actions taken in response to the threat of disinformation. This multifaceted field involves numerous groups including researchers, policymakers, journalists, NGOs, and civil society groups, as well as platforms and companies.

As a result, there is a profusion of information available to those who seek to better understand the landscape of online disinformation. This ranges from open source investigations to disclosures from platforms, research papers, policy papers, intelligence reports, and more. However, this information is not as accessible or useful as it could be as it is scattered across multiple sites and properties.

In our investigation, we found that a number of projects had laid groundwork to address this accessibility challenge, such as www.io-archive.org, the Hamilton 2.0 dashboard, and more, but that many questions remained unaddressed. For instance, we could not find a comprehensive library of publicly disclosed platform actions or open source investigations.

As a result, it can be challenging for those working in the field to have a fully informed conversation based on currently available data, or to have a sense of how currently available information could be used in novel ways. It can also be challenging to form a clear opinion of what datasets may be missing or what new form of analysis could help advance the field.

The Disinfodex project was created over 10 weeks during the Assembly:Disinformation Fellowship at Harvard’s Berkman Klein Center. It represents an effort to help those working in the disinformation space to better aggregate, analyze and interpret some of the information that is already available in the open.

At the time of this paper’s release, Disinfodex has aggregated and indexed reports of actions taken by several platforms, sourced directly from the websites and social media sites of these companies. In total, the database indexes 53 individual disclosures from Facebook, Instagram, Twitter, Google, YouTube, and Reddit, starting in September 2017. Together, these disclosures allow for a cross-platform analysis of the threats the platforms have reported on and the actions they say they took in response. They also allow for the comparison of network attributes as disclosed by platforms; for instance, how many of these networks have ties to or targeted a given country.

In order to develop this project we had to consider the existing taxonomies for categorizing disinformation and consider how they may impact our work [section II]; to examine the pros and cons of such disclosures by platforms and others and how our work might affect that [section III]; and to explore broader challenges and risks related to this project [section IV].

To conclude, we outline how aggregating more publicly available information could help to develop a shared infrastructure for practitioners, scholars, and others interested in the field, and share thoughts about next steps for this project [section V].

II. TAXONOMY AND CATEGORIZATION

One of the most urgent tasks facing scholars, practitioners, and the public alike is identifying shared definitions of disinformation and related topics across academic disciplines (Spies, 2019). Indeed, while there exists an abundance of relevant and valuable scholarly work on this issue, multiple overlapping definitional frameworks continue to coexist – with diverse perspectives on the relevance of falsehood, intentionality, means, or strategic objectives to establish the contours of disinformation (Wanless, Pamment, 2019).

It is worth noting that real-life covert influence operations may not stick to misinformation or disinformation; as documented by open source investigators over previous years, threat actors have been known to generate or amplify information that aggravates existing tensions in societies, including content that is identical to that posted by genuine, good-faith platform users. For instance, a 2019 Graphika report on "the IRACopyPasta campaign"[3] notes that "*Many of [the threat actor's] posts consisted only of memes, without accompanying text. Posts that included text were usually sourced from viral posts made by American accounts.*"

As such, when responding to disinformation, platforms have tended to focus on addressing malicious actors by developing rules and policies tackling nefarious behavioral patterns rather than focusing on the content that they propagate[4]. While there is no joint taxonomy across platforms, each of the services whose disclosures we considered for Disinfodex had their own terms to characterize such problematic behaviors, from "Influence Operations" for Google and YouTube, to "Coordinated Inauthentic Behavior" at Facebook and Instagram, and "Platform Manipulation" at Twitter.

We observed that platforms provide diverse levels of substantiation with regards the nature of the behaviors they consider a violation of their terms of use or policies, but these behaviors seemed similar enough for the platforms to share tips with each other on multiple occasions, leading to action across multiple companies.

To help users grasp these nuances, Disinfodex includes the descriptions used by each individual platform about the actions they took, and outlines the platform policy violations that grounded these actions, when specified.

[3] *The IRACopyPasta Campaign*, Graphika, 2019

[4] For more on the Actors, Behaviors, Content taxonomy – see: *Actors, Behaviors, Content: A Disinformation ABC*, Camille Francois, 2019

III. DISCLOSURES

Public disclosure of disinformation campaigns, including when and how much should be disclosed, has become a source of debate among practitioners and experts. The process of deciding when to disclose actions taken against networked actors or campaigns must take into account important and complex considerations including the public's right to know, the risk of amplifying the work of the threat actor, or the risk of lowering public trust by unintentionally giving the impression that disinformation is more widespread or impactful than it actually is (Tufekci, 2017).

As of April 2020, there are no industry-wide standards or guidelines on whether, how, and when companies should disclose actions taken against disinformation campaigns. In the absence of such standards or legal requirements, the ethical and practical considerations relating to the disclosure of actions taken against disinformation campaigns lie primarily with the platforms.

Over the course of our work, we observed that platforms have taken diverse approaches since 2017 – which, from the information we surveyed across platforms, was the year public disinformation-related disclosures first took place. While all companies tend to announce actions against disinformation campaigns through their websites or platforms after an investigation has been completed, the modalities of these disclosures vary. For instance, Facebook shares information with third party companies who investigate disinformation campaigns to help with their own reporting, while Twitter and Reddit make databases of accounts and content that they have removed or banned publicly available.

Disinfodex evidentiates that these platforms differ in the frequency, nature, and amount of information they release. There may be many legitimate or illegitimate reasons for these inconsistencies, ranging from different appetites for legal risks to different levels of exposure to malicious actors, or different scales of the problem across platforms. The purpose of the Disinfodex project is not to speculate on the reasons behind inconsistent disclosure practices or to determine what constitutes good or bad disclosure behavior, but rather to provide a dataset and tool to help researchers, journalists, and investigators to explore companies' disclosures and make their own determinations.

IV. CHALLENGES AND RISKS

While we hope Disinfodex will prove useful to scholars, journalists, policymakers, and practitioners alike, this project has both challenges and limitations, which we outline below:

Our scope is limited to the information that platforms disclose: In indexing only public releases from several major social media companies at this point, Disinfodex comprises information the companies have elected to disclose publicly. As such, the dataset cannot and should not be construed as representing the sum total of disinformation or influence operations that have taken place online or that have been tackled by platforms.

For a more comprehensive aggregation of relevant activities, one should also consider publications from researchers, journalists, civil society groups, for-profit companies, non-governmental organizations, etc., which Disinfodex does not index at this point. The database may expand over time to include some or all of these categories, although this expansion would come with a new set of challenges including the development of robust inclusion criteria.

It is also worth noting that some platforms (e.g. Facebook) work closely with third parties of their disclosures. In such cases, third party reports represent an important complement to the information directly released by the platform. We intend to add these reports to Disinfodex in the future, but in the interim, we urge users to consult these reports when available.

The numbers we index may lead to misinterpretation or misleading analysis: While we believe it is important to provide simpler means to analyze platform disclosures over time, these numbers may be misinterpreted. For instance:

- We do not seek to compare platform A's high number of removed accounts related to a disinformation campaign to platform B's comparatively lower number. That inquiry does not suggest that platform B may have missed details. Without knowing the actual exposure of platform B to that campaign, a firm conclusion cannot be drawn from such a comparison.
- Additionally, the numbers provided by platforms are not necessarily reliable to assess the effectiveness or reach of a given disinformation campaign. Research is still ongoing to understand how online disinformation affects the people it reaches; it is not the case that two campaigns that reached the same scale in terms of number of users exposed to the content will have the same impact.

All disclosures risk amplifying the work of malicious actors: The Disinfodex project considered the risk of unintentionally amplifying the disinformation campaigns that platforms have taken action against by giving them more exposure. We believe this risk is mitigated by the fact that the platforms' disclosures do not repeat the narratives of these disinformation campaigns. However, we also considered whether this project's focus on disinformation campaigns may give the impression that disinformation campaigns are more widespread on platforms than they actually are – thus furthering the sense of distrust that may be the ultimate goal of a number of threat actors[5]. This risk is challenging to mitigate, as the discussion of disinformation contributes to magnifying this threat. We simply urge the reader to remember that even the highest numbers reported by platforms or third parties represent a fraction of the activity taking place on platforms.

We do not take a position on the actions or disclosure practices of platforms: The Disinfodex project does not take a position on the disclosure practices of platforms, nor does it address salient questions of whether platforms should disclose more information about their services or actions and to whom, an issue elevated by multiple researchers and authorities around the world (e.g. Walker et al, 2019) and a matter of ongoing deliberations for platforms.

We were deliberate about this approach for two reasons:

- Participants in the project group have direct ties to one or more of the platforms at stake; their involvement in an analysis or position about the platforms' work on disinformation would have the potential to create a conflict of interest or the impression thereof, thus jeopardizing the credibility of this project.
- We believe that to function as a trusted hub for data releases, Disinfodex needs to be an independent aggregator that people can use to form their own assessments. In

[5] See for instance: Yardena Schwartz, CJR, 2017: Putin's throwback propaganda playbook

demonstrating and itemising the information platforms have publicly released, Disinfodex may be used to better conduct cross-platform analysis, to determine commonalities and patterns, or to provide historical context for research and reporting on platform responses. Similarly, it could be used to determine what gaps still exist in the information that is released, to help establish that more helpful conclusions can be derived from publicly available information than previously thought, or to drive further discussion on how and when information should be released.

V. CONCLUSION AND FUTURE WORK

While disinformation is not a new or strictly online problem, the study of online disinformation is a relatively novel, complex, fast-moving field where many fundamental questions remain open (notably the effect of exposure to online disinformation and the usefulness of mitigations).

In order to give those working on this issue the best chance of countering disinformation, we need tools that can foster a better understanding of the landscape. Such tools, relying upon publicly available information, could help better identify cross-platform disinformation campaigns, identify trends over time, and build solutions. Disinfodex is one step towards that goal; it is designed to be built upon and improved.

While we hope that others will develop similar or complementary projects to advance the field, we plan to iterate upon Disinfodex. Provided we can secure resources and partnerships to move the project forward, we aim to expand the work in the following ways:

- **Database content:** The initial release of Disinfodex represents only a subset of the publicly available information that could be useful to scholars and practitioners. As such, we hope to iterate it by:
    - Keeping the database up to date**:** We will continue adding platform disclosures as they are released, and hope to improve aggregation methods.
    - Automatic scraping: We would like to write a script to crawl existing platforms for notification of disclosure releases, which would auto populate the Disinfodex data table with newly available information.
    - Adding open source investigation reports: The reports of reputable open source investigators would provide a helpful complement to the disclosures of platforms (for example, FireEye or DFRLab releases, etc).

- **Database design:** At time of launch (May 2020), Disinfodex operates as a simple search and discovery interface. Moving forward, we would consider:
    - Visualizations and dashboards: We would like to make the information more accessible by allowing users to generate visualizations and dashboards.
    - Exporting or downloading: As information in Disinfodex is neither proprietary or confidential, we want to make it easy for users to export and analyze it.
    - Direct input: We would like to integrate solutions such as an API to make it possible for select third parties (for example, platforms or open source investigators) to add content pertaining to their work directly into the database.

- **User interface/User experience:** We are exploring ways to make the content of Disinfodex more easily available to a broader set of users. This could include:
    - Improving UI Design: Disinfodex's interface is a simple front-end tool that allows search and query. We would like to iterate upon it to include tailored data views to serve each audience, and present data in a way that addresses specific cases for each audience - for example, writing an article, writing a thesis, completing a threat intelligence report.
    - New query methods: We would like to develop ways for users to query the database directly from other platforms to either (1) add to, or (2) check against our existing data. For example, we could create a bot for social media to automatically pull and publish answers to simple user questions, adding real-time assistance to users.
    - Mobile application: As Disinfodex matures, we would like to create a mobile application to broaden usage.

Our ability to engage in these workstreams is limited by the team's time constraints and our technical and financial resources. However, we are committed to furthering the work of Disinfodex, and welcome feedback from users, thoughts on how we might build upon it, and ideas about groups or organizations we could partner with to help us expand this project. Our team can be contacted via email at disinfodex@googlegroups.com, and we intend to communicate updates about the project on the Disinfodex website.

BIBLIOGRAPHY

Francois, Camille. *Actors, Behaviors, Content: A Disinformation ABC* (September 20, 2019), Available at https://science.house.gov/imo/media/doc/Francois%20Addendum%20to%20Testimony%20-%20ABC_Framework_2019_Sept_2019.pdf

Schwartz, Yardena. *Putin's throwback propaganda playbook* (January 28, 2017), Available at https://www.cjr.org/special_report/putin_russia_propaganda_trump.php

Spies, Samuel. *Defining "Disinformation"* (October 22, 2019), Available at https://mediawell.ssrc.org/literature-reviews/defining-disinformation/versions/1-0/

Walker, Shawn, et al. *The disinformation landscape and the lockdown of social platforms, Information, Communication & Society* (July 19, 2019), Available at https://www.tandfonline.com/doi/full/10.1080/1369118X.2019.1648536

Wanless, Alicia, and Pamment, James. *How Do You Define a Problem Like Influence?* (December 30, 2019), Available at https://carnegieendowment.org/files/2020-How_do_you_define_a_problem_like_influence.pdf

Wardle, Claire, and Derakhshan, Hossein. *Information Disorder Toward an interdisciplinary framework for research and policymaking* (September 27, 2017), Available at https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c

Zeynep Tufekci. *Twitter and Tear Gas: The Power and Fragility of Networked Protest* (2017), Yale University Press