Takshashila Discussion Document

Analysing Digital Platforms' Responses to COVID-19 Information Disorder

Discussion Document 2020-09

V 1.0 - 26 June 2020

By Prateek Waghre | Rohan Seth

This document can be cited as "Prateek Waghre and Rohan Seth, *Analysing Digital Platforms' Responses to COVID-19 Information Disorder*, Takshashila Discussion Document, June 2020"

This document is prepared for discussion and debate. It does not necessarily constitute Takshashila's policy recommendations. For any feedback, please email research@takshashila.org.in

Executive Summary

This Discussion Document analyses the responses of digital platforms to the information disorders around COVID-19:

1. During the pandemic, digital platforms may be able to exercise greater discretion in content moderation from increased calls for them to act as arbiters of truth.
2. Efforts to contain information disorders vary across the following areas:
    a. Changes in Information Flows: Platforms have made alterations to how information is spread.
    b. Funding: Platforms have invested significant sums of money into fact checking organisations, journalists, and media outlets.
    c. Changes in User Interface (UI): Platforms have made changes to their UIs to highlight credible information from public health authorities.
    d. Policy: Platforms have reacted by coming up with new policies, modifying existing policies, and deploying existing policies to deal with COVID-related information disorders.
3. Level and consistency of actions taken differ across geographies and types of actors.
4. Platforms’ response in controlling misinformation may lead to significant changes in the relationships between the government, society, and people.

Introduction

Information disorder is a collective term used to describe misinformation, disinformation and malinformation. These terms can be defined as -

- Misinformation: Information that is false but not intended to cause harm.
- Disinformation: Information that is false and is created/distributed with the intention or purpose to create harm.
- Malinformation: Information that is genuine/true which is shared to cause harm.

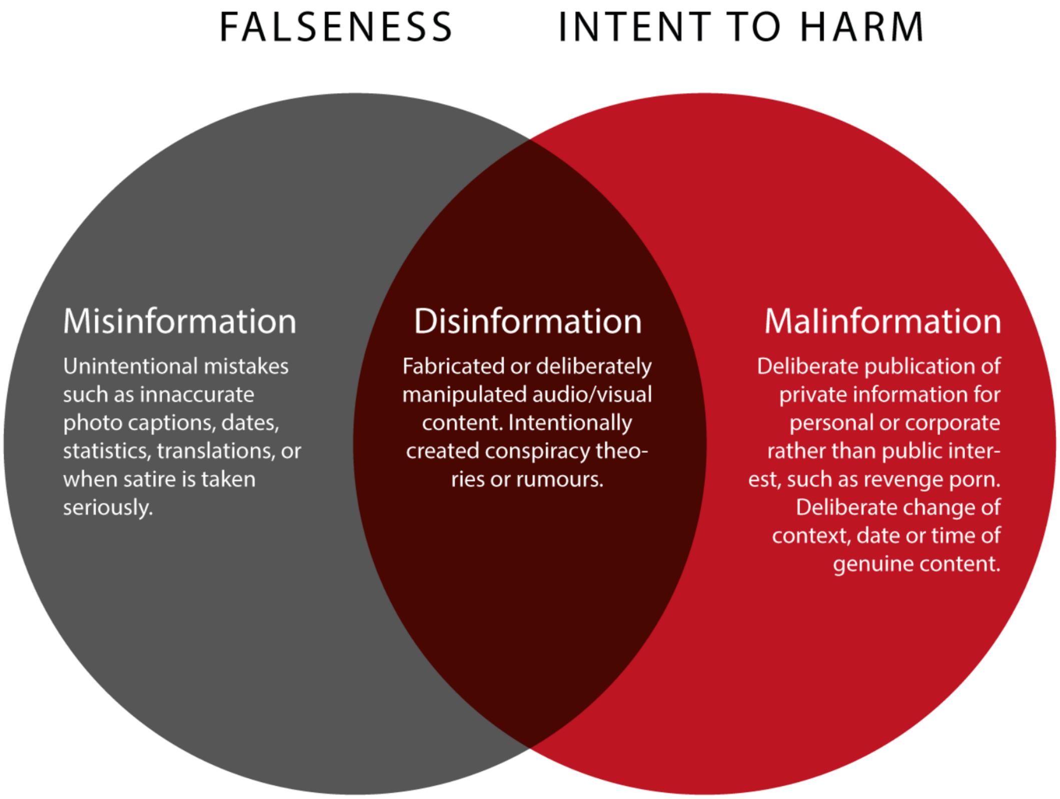
FALSENESS
INTENT TO HARM
Misinformation
Unintentional mistakes such as innaccurate photo captions, dates, statistics, translations, or when satire is taken seriously.
Disinformation
Fabricated or deliberately manipulated audio/visual content. Intentionally created conspiracy theories or rumours.
Malinformation
Deliberate publication of private information for personal or corporate rather than public interest, such as revenge porn. Deliberate change of context, date or time of genuine content.

Image 1: A visual description of misinformation, disinformation, and malinformation[1]

Due to the COVID-19 outbreak people are spending an increasing amount of time online, on social media and communication platforms. Surveys suggest that more news is being consumed[2] online and many[3] of these platforms have reported[4] significant[5] increases in usage[6],consumption[7], engagement[8], downloads.[9] This is in contrast with the state of the media and publishing industry where scaling back of advertising budgets have resulted in job losses.

Crisis informatics[10] researchers have pointed out[11] that events such as *'natural disasters, industrial accidents, terrorist attacks, and emergent pandemics are often times of high uncertainty'*. During such times, people resort to 'collective sense-making' as they try to obtain more information and take decisions in increasingly ambiguous situations. While this process has certain benefits it also results in a high degree of susceptibility to low quality information and information disorder. Motivated actors also look to exploit such situations whether it is to promote scams, influence operations, or just spread uncertainty. At the same time, the need for social distancing has meant that our interactions have moved onto internet platforms which are simultaneously increasing their reliance on automated content moderation since confidentiality agreements have limited remote working for such functions.

It is therefore important to analyse how various social media platforms have responded to information disorder. For this document, we have considered the following platforms listed in Table 1, which have a large user-base or wide reach in India. Some of them are also used to mobilise/motivate large numbers of people for various political and social movements.

Note that the scope of the analysis has been limited to COVID-19 and primarily considered from an Indian perspective. Therefore developments/events that are not directly related to COVID-19 have not been covered.

Facebook / Instagram	11% of Facebook's total user base (241 million users)[12] is from India. In addition, in Jan 2019, Instagram had ~155 million users in India.[13] Both numbers are expected to grow over time, making Facebook (and Instagram) the biggest players in the Indian social media ecosystem.
WhatsApp	WhatsApp is estimated to have ~400 million users in India, giving it the largest footprint. It is an outlier in this group given that it is a closed and encrypted instant messaging platform. Nevertheless, it is an important source of content for many Indians and a significant player in the information ecosystem.
TikTok	TikTok reportedly has 200 million users[14] in India and expects that number to grow to around 300 million by the end of 2020. It is therefore among the top 3 platforms in the country and as such has an important role to play when it comes to curbing misinformation and disinformation.
Google	For most people in India, Google is ubiquitous with internet usage. Google is a dominant part of how people access the internet, starting from using Android and Google Assistant platforms, Google Search as yellow pages, and Google Maps for directions. It is the go-to choice for people in India and the world to sort and access information.
YouTube	YouTube has 265 million[15] monthly active users. This makes India YouTube's biggest audience and one of the fastest growing audiences in the world. YouTube's consumption on mobile has increased [16]to 85%, with 60% of the watch time coming from outside of the six largest metros in the country.
Twitter	Twitter does not break out numbers by country and estimates of the number of users in India vary between 13M and 35M.[17] While its footprint is among the smallest in this group, it is extensively used to mobilise political activity and therefore its response to information disorder is of interest.
Sharechat	Sharechat identifies itself as a regional language social media platform. Reportedly, it has around 60 million monthly active users (MAU)[18] giving it a larger user-base than Twitter. With its focus on regional language content, how it handles information disorder is crucial.

Table 1: Platforms Analysed

Efforts to contain Information Disorder

The pandemic has presented a new set of challenges for platforms in managing information disorder surrounding the virus. Each platform is different in terms of its size, the kind of content they host and sort, and their user interfaces. For instance, while TikTok may find it easier to restrict or remove ads from the platform, WhatsApp does not face that challenge given the user interface does not support ads yet. Similarly, while Google or Facebook may be in a position to make substantial grants to combat information disorder, ShareChat may not be able to do so as it has had to contend with job losses during the pandemic.

A combination of these factors results in significant differences in terms of the responses they have to the challenge. The table below shows, with some degree of granularity, the response each platform has taken. We have identified eleven different kinds of efforts. At the time of writing, it is too early to say how effective each of these initiatives is going to be.

	f		TikTok	G			ShareChat
Misinformation research / journalism grants	✓	✓	✓	✓		✓	
Promoting official/reliable sources	✓	✓	✓	✓	✓	✓	✓
Creating dedicated covid-19 resources		✓	✓	✓		✓	
Content removal/flagging	✓				✓	✓	
Downranking/restricting content	✓	✓				✓	✓
Chatbots		✓					✓
Restrict/remove ads	✓			✓	✓	✓	
Dedicated Covid-19 misinformation category			✓				
Priority moderation queues			✓			✓	
Post interaction warnings	✓						
Verifying credible sources/advertisers		✓		✓		✓	
Created new policies				✓	✓		
Modified existing policies	✓					✓	
No change to policies			✓				✓

Table 2: Summary of efforts for each platform considered in this study

Based on information summarised in the table, we can classify responses by platforms to information disorder in the following four categories:

1. Allocating funds: Platforms that host or sort user generated content, such as TikTok and Facebook, have allocated grants to various institutions to support their efforts to manage information disorder. This involves grants to fact-checking organisations, media outlets, and sources of credible journalism.
2. Changes to User Interface: Platforms that have a 'feed' as part of their UI have taken this step by inserting and highlighting information from international organisations to make it easier for users to access information from credible sources. This also involves modifying how explore sections function or how search results related to COVID-19 are structured.
3. Changes to information flows: This category has the most nuanced responses from platforms. It involves downranking content on news feeds, restricting or removing ads, providing public health authorities with verified chatbots, installing priority content moderation queues, and notifying users about their interactions with misinformation.
4. Policy Changes: In this category, we observed that platform responses varied across a spectrum. Most of them either created new COVID-19 specific policies or modified their existing policies in response to COVID-19. Some of them chose not to make any specific changes but continued to apply existing policies as they were.

Facebook

Facebook is providing the World Health Organization (WHO) with "as many free ads as they need"[19] as well as "ad credits" to other organisations. This was meant to give organisations a mechanism to put out credible and accurate information to users without worrying about the ad spend.

Facebook and Instagram also made UI changes to direct users to reliable and official sources. On Instagram, when a user taps on a hashtag related to COVID, the app shows resources[20] from the WHO, and other local health authorities. Facebook began inserting[21] a box in users' news feeds directing to the Centre for Disease Control's (CDC) page (or alternative relevant local authorities) about COVID-19.

Facebook claims that it will remove[22] false claims and conspiracy theories that have been flagged by leading global health organisations, as well as block people from running ads that try to exploit the situation. In addition, Facebook has also promised to down-rank and remove[23] pages and groups which spread misinformation about vaccinations from recommendations. Similarly, Instagram's announcement says that it will downrank content[24] that has been rated false by third-party fact-checkers and remove[25] false claims and conspiracy theories highlighted as false by leading global health organisations. Also, Instagram has indicated that, 'COVID-19 accounts' and content will be removed from recommendations and the Explore tab respectively, unless posted by a health authority.

In terms of advertising, Facebook claims that Ads that contain misinformation about COVID-19[26], will be reviewed and rejected on Facebook and Instagram. Targeting options that may be harmful, such as, based on vaccine controversies will not be allowed. Instagram also claims to prohibit[27] misleading ads that refer to COVID and has removed the ability to search for COVID-19 related AR effects unless they were developed in partnership with a recognised health organisation.

According to an update on Facebook's Newsroom,[28] Facebook has applied existing misinformation policies to misinformation about COVID-19 to 'remove posts that make false claims about cures, treatments, the availability of essential services or the location and severity of the outbreak'.

The following Facebook Advertising Policies and Community standards are relevant:

- Advertising Policy- Misinformation[29]: Facebook pledges to prohibit ads that have been debunked by third-party fact checkers (in case of COVID, this will likely be substituted for the WHO and other local authorities). Repeated offenses of this policy may result in Facebook placing restrictions on future advertisements.

- Community Standards- False News[30]: Facebook does not remove false news from the platform, instead, lowers its distribution in the Feed. Given the Newsroom update, this does not apply to misinformation around COVID-19. Facebook will remove posts that make false claims about cures, treatments, the availability of essential services or the location and severity of the outbreak.

- Community Standards- Manipulated Media[31]: Facebook aims to remove media where the manipulation isn't apparent and could mislead, particularly in the case of video content. This specifically applies to videos, standards for which are outlined in the policy.

WhatsApp

While WhatsApp is part of the Facebook platform, it has announced USD 1M[32] grant to Poynter Institute's International Fact-Checking Network's (IFCN) #CoronaVirusFacts[33] alliance.

As a closed, encrypted messaging platform, no changes were made to its user-interface. However, interactive chatbots in collaboration with organisations such as WHO[34], IFCN[35] were developed and promoted.

To potentially reduce the spread of information disorder, frequently forwarded messages[36] (indicated visually by a double arrow icon ») were restricted such that they could only be forwarded to one chat at a time compared the usual limit of 5. For reference, messages are defined as frequently forwarded if they are forwarded more than 5 times.

Since WhatsApp does not have visibility into any of the content being shared over its platform, it was not considered for the policy change classification.

TikTok

In a short span of time, TikTok has evolved into one of the most frequently used platforms in India. The frictionless sharing with low dependency on text also makes it a target for misinformation. The company has declared a misinformation research grant up to USD 50K (INR 3.5 million) with the objective of understanding the misinformation ecosystem on social media better.

A new category for In-app reporting of COVID-19 related misinformation[37] has been created and content reported in this section is sent to a priority moderation queue manned by an internal task force and escalated to third-party fact checkers. It is not specified whether this is specific to health/medical information or not. An in-app informational page[38] has also been pinned to the top of the 'Discover' tab containing information from authoritative sources like WHO and MOHFW. And, to build awareness around the spread of COVID-19 an in-app quiz[39] available in 11 Indian languages was created. Hashtags related to COVID-19 were intended to be accompanied by a public service announcement.

Content shared[40] by credible partners such as MyGov, PIB, WHO, UNDP India and UNICEF India on the platform will also be elevated. Simultaneously, it also claims to have increased its moderation effortsibid and removed 'thousands' of videos.[41]

TikTok has not updated its Community Guidelines[42], since January 2020. It has not instituted new COVID specific policies either. Its existing policy already covered misleading information which included misinformation 'meant to incite fear, hate or prejudice' as well as content which 'may cause harm to an individual's health'.

On its Safety Resources page for COVID-19[43], it does state its intention to err on the side of caution for misinformation that could cause harm to the TikTok community or the public in general, resulting in the removal of otherwise 'borderline' content.

It also mentions a partnership with Vishvas News - the fact-checking arm of the Jagran group.

Google and YouTube

Because both Google and YouTube have the same parent (Alphabet), their initiatives to deal with information disorder have been collaborative. As a result, for the purposes of simplification, we have mentioned them in the same section.

Google is providing[44] USD 6.5 million in funding to fact-checkers and nonprofits fighting misinformation around the world. According to the press release, this is also supplemented by increased access to data through funding bodies such as SciLine and Australian Science Media Centre.

According to Google's India Blog, Google will show[45] the latest updates and health advice from the Ministry of Health and Family Welfare (MoHFW) and international health authorities across Google Search, Google Maps, and YouTube.

In addition, YouTube has begun to highlight and add information from the WHO in the middle of search results related to COVID. The platform has also begun to add a link to the MoHFW's COVID page to the top of search results related to COVID.

On the advertisement front, Google will expand its program[46] of verification of advertisers to weed out fraud and "bad actors." This is part of a longer process and may not be completed immediately. According to The Verge, Google has also blocked[47] tens of thousands of ads "capitalising" on the virus. It has also pulled ads from YouTube videos that discuss COVID-19.

Google searches related to the virus now trigger[48] an "SOS Alert," which presents the user with a dashboard on information related to the virus, with news from mainstream publications displayed prominently.

Besides, YouTube claims that it does not allow[49] content that spreads medical misinformation that contradicts the World Health Organization (WHO) or local health authorities' medical information about COVID-19. According to their new policy on misinformation related to COVID-19, this is limited to content that contradicts WHO or local health authorities' guidance on: Treatment, Prevention, Diagnostic, Transmission.

The terms are better explained in the policy, with a list of examples of content that will be in violation of the update.

In case content violates the policy, it will be taken down and an email notification will be sent to the creator. Should the channel be a first-time offender, it will be let off with a warning and no penalty.

Should the channel not be a first-time offender, YouTube will issue a strike against them, after three of which, the channel will be terminated.

Twitter

Credits under Ads for Good[50] will enable nonprofit organisations to build fact-checking campaigns. Organisations suchs @factchecknet (@maldita_es and @malditobulo) and @taiwantfc are believed to be making use of such credits already. The company has pledged 1M USD[51] to the Committee to Protect Journalists and International Women's Media Foundation with the aim to support journalists amid economic uncertainty.

In terms of changes to the user interface, a dedicated search prompt that will prioritise credible, authoritative information has been created. A number of other dedicated resources have also included an 'Events' feature- pinned to the top of the timeline for users in 30+ countries including India and a dedicated Twitter 'Moment' linking to external content about COVID-19 globally.[52]

A new system of labels and warnings[53] has been rolled out to flag misleading information, disputed claims and unverified claims and inform people that they are engaging with content that conflicts guidance from public health experts. Notably, these tweets will be identified through 'proactive monitoring' by the company's own admission. Amplification of tweets with these labels will also be prevented. Content that may lead to 'increased exposure or transmission' will be prioritised for review.

The #KnowTheFacts[54] function has been expanded so that auto-suggest results will avoid directing users to non-credible content. This feature was originally designed to provide credible information on immunisation and vaccination. Twitter accounts which provide credible COVID-19 updates will be verified. Accounts associated with an 'authoritative organisation or institution will be prioritised.[55] Twitter claims to have official partnerships with national public health agencies in over 70 countries[56], including India**.** To facilitate research, it also released[57] a new endpoint into Twitter Developer Labs to enable approved researchers to study public conversations around COVID-19 in real-time. The dataset from this end-point/API is provided at no-cost.

On the policy front, Twitter has taken incremental steps towards what it describes as protecting the public conversation around COVID-19.

- Platform Manipulation: In January[58], it announced its intention to monitor any coordinated campaigns to spread disinformation related to COVID-19. In March, it went further to state that it is taking a 'zero tolerance approach to platform manipulation'. In both updates, it mentioned that no significant coordinated manipulation efforts had been detected yet.

- Content Moderation: COVID-19 also required platforms to rely more on machine learning and automation for content moderation. Twitter acknowledged that this has its limitations and announced that no accounts would be permanently suspended based 'solely' on automated systems. It also stated its intent to rely less on reporting and more on close coordination with 'trusted partners, including public health authorities and governments'.

- It announced the creation of a content severity triage system in order prioritise action against rule violations that presented greater risks of harm.

- Broadened definition of harm: In March and April, Twitter also broadened its definition of harm to include tweets that contradict local health authorities (e.g. effectiveness of social distancing), promote ineffective cures - both harmful and not immediately harmful, denial of scientific facts, conspiracy theories, impersonation of authorities, and inciting people to action that could lead to the damage of critical infrastructure, etc.

In the event that world leaders violated COVID-19 guidelines, it would leverage the 'Public Interest Notice'[59] approach that was updated in October 2019[55] which was already applicable in cases such as promotion of terrorism, clear and direct threats of violence, releasing private information, promotion of self-harm, child sexual exploitation and posting/sharing intimate photos without consent.

In May[60], as part of an updated approach to misleading information it announced that content that is misleading, contested, or unverified would be labeled accordingly so that users interacted with the labels first, before seeing the actual content itself.

- Ads: Clients and partners that have dedicated Account Managers can post ads containing implicit or explicit references to COVID-19 with regard to changes to business practices/models and support for customers/employees. Restrictions applicable include inflation of prices of products related to COVID-19, any promotion of products such as facemasks, alcohol hand sanitisers, and sensational content likely to incite panic.

Sharechat

There were no reports in the media, nor any updates on Sharechat's websites to indicate whether the company had donated orpledged any financial support for fact-checking operations or media outlines. On the contrary, there were reports that the company needed to eliminate[61] jobs during this period.

While there were no official updates to its website/blog listed COVID-19 specific feature updates, the Chief Business Officer, in a press release[62] stated that trending tag will be displayed at application launch showing users the latest and official updates from the Ministry of Health and Family Welfare. A chatbot, in partnership with All_IN_Call to address queries and provide COVID-19 related factual information from government and global health sources was also announced.

The same press release also mentioned that algorithmic changes to the trending feed section have been made so that content that has been debunked by third party fact checkers will not be displayed. Since an important aspect of the app is regional language support, content review is supposed to take place across 13 languages and will also cover conspiracy theories, fake news, and misinformation campaigns.

Sharechat has not officially updated its Content and Community Guidelines[63] since October 2019. It has not announced any new policies since the outbreak of the pandemic either.
The existing guidelines which do cover 'Fake News' state that content which 'spreads deliberate misinformation, hoaxes or fake propaganda with the intent to mislead, is not permitted'.

Content meant to create 'a sensation by straight-up' fabrication or 'damage someone's reputation, hurt their financial or political standing on the basis of false information' will not be allowed to 'spread' on the platform.

The guidelines explicitly state that satire and parodies, which are meant to amuse 'not mislead' will not be 'confused' with 'fake news'.

According to the press release mentioned above, the company's CBO was quoted as saying that 'fighting fake news is a focus area' and also stated that third party fact-checkers were reviewing content across 13 languages.

Understanding the Operating Environment

In theory, a public health crisis offers a suitable situation for interventionist behaviour by platforms to act on. During the current pandemic, social media platforms have witnessed increased engagement and content creation. In general, more information about COVID-19 is being created. A framework[64] proposed by Stratechery posits that with the increase in content creation, there will be a corresponding rise in low quality information as well as misinformation and disinformation. Additionally, research on fact-checked items[65] in India indicates that the nature of information disorder has also shifted with time. It is not limited to health-related topics but has extended into existing political themes. Thus, addressing information disorder related to COVID-19 is not a straightforward task for platforms and requires the consideration of various tradeoffs.

Twitter states[66] that it has taken action against more than 2600 tweets containing misleading and potentially harmful content and over 4.3 million accounts which were targeting COVID-19 related conversations. Facebook claims[67] to have labelled around 50 million pieces of content based on ~7500 fact-checked articles. WhatsApp estimates[68] that sharing of frequently forwarded messages had reduced by as much as 70%. TikTok says[69] it has removed 'thousands of videos' that violated its Community Guidelines in India. Various academic studies[70] demonstrated that misleading content continued to stay up. The spread of the #plandemic documentary[71] also revealed how challenging it is to completely remove content even when there is a strong willingness to do so. This highlights the difficulties that platforms face to moderate even public health related content. Moreover, politicisation of subjects such as effectiveness of lockdowns, protests against stay-at-home orders further increase the challenges faced by platforms. As a result, they are faced with multiple different kind of trade-offs that need to be incorporated in their decision-making processes that deal with taking action against misinformation and disinformation.

International v/s Local Sources of Information

Platforms such as Google and Facebook manage information on a global scale. During the pandemic, a key part of making credible information readily available to people is deciding where that information should come from. With that decision, comes an inherent tradeoff between promoting American healthcare institutions e.g. CDC, International healthcare institutions e.g. The WHO or locally relevant bodies, e.g. The Ministry of Health and Family Welfare in India, or the Ministry of AYUSH.

When it comes to choosing between whether the source of information should be American or International, the companies examined in this document have always tilted towards the latter, along with information from sources relevant to the country the platforms are operating in.

While a situation of the sort has not occurred yet, settling the trade-off this way can present conflicts when there is a divergence in messaging by local and international bodies. For instance, earlier this year, because of the controversy around hydroxychloroquine, the WHO suspended trials of the drug[72], claiming that it led to an increase in mortality rate. However, at the same time, The Indian Council of Medical Research (ICMR) expanded the usage of HCQ[73] as a preventative measure against COVID-19.

While the conflict in messaging did not play out disproportionately on social media, conflicting messaging between authorities is a challenge inherent in this trade-off.

Managing different classes of actors

Platforms broadly have four different types of actors:

- Influencers: Users with large numbers of followers.
- Organisations
- Ordinary users
- Bots/trolls/accounts participating in coordinated information campaigns/influence operations

Each of these actors have different levels of influence with their ability to spread misinformation.

Accounts with large levels of interaction and high follower counts are important to platforms. These include influencers such as content creators, politicians/public figures, celebrities etc. However, studies by NewsGuard have highlighted the roles of super spreaders on Facebook (in Europe)[74] and Twitter[75] respectively. Reports[76] in the media also covered the role of celebrities in the amplification of 5G related COVID-19 conspiracy theories. Twitter has deleted tweets[77] containing misleading content posted by Brazilian President Jair Bolsanoro, Venezuelan President Nicolas Maduro and in India, Rajnikanth[78] with regard to COVID-19. Facebook has taken action against[79] posts by Jair Bolsanoro for COVID-19 related information on Facebook and Instagram.

Meanwhile, neither company has flagged or removed content posted by American President Donald Trump for COVID-19 related misinformation in the context of his advocating the use of hydroxychloroquine as a prophylactic or injecting disinfectant as a possible cure. Twitter did restrict activity using #InjectDisinfectant and #InjectingDisinfectant. This prevented users that can be classified as 'ordinary' from participating in conversations using these hashtags.

Twitter claims that it has not yet observed significant coordinated activity related to coronavirus misinformation, but a study[80] by CMU estimated that nearly half of COVID-19 related information was posted by bots. It should be noted that several academics/researchers[81] have expressed skepticism about its findings in the absence of information regarding the methodology and classification of 'bots'. Facebook banned a publisher[82] named 'Natural News' after discovering that troll farms in North Macedonia and Philippines frequently posted content from it. In India, none of the platforms have reportedly acted against or even flagged posts claiming that coordinated "clapping will kill the coronavirus"[83], which was debunked by several fact-checkers.

Level and consistency of actions taken vary significantly across different geographies and sets of actors indicating that different internal trade-offs inform the final course of action followed.

***Note**: Twitter did ultimately flag tweets by Donald Trump under different policies (civic integrity and glorification of violence) for non-COVID-19 content. Therefore, those instances are not explicitly covered in this analysis.*

Ad revenue/engagement and curbing misinformation

The crisis induced by the pandemic also presents a revenue opportunity for platforms. Due to the increasing relevance, misinformation posted by users around the virus would lead to more clicks and better engagement rates. At the same time, it would also present an uptick in advertising for COVID-related goods and services, such as mask suppliers. This would also apply to users classified under labels like 'vaccine controversies'.

This presents an inherent tradeoff between maximising revenue and limiting misinformation about the virus. During the research for this document, we repeatedly came across instances of firms sacrificing ad revenue to control misinformation around the virus.

For instance, Facebook blocked advertisers from running ads that would aim to exploit the situation, notably by trying to sell a cure. Such instances have been rife on social media websites, particularly when trying to advocate for alternative medicine as potential cures. This involves cures based on kadha, turmeric, cumin, and garlic.[84]

While advertisements around these 'cures' would have been opportunities to rake in advertisement revenue, changes in policies have tilted in the favour of controlling misinformation instead.

Broader Policy Implications

Information Disorder and platform responses to it will shape interactions between States, Platforms, and Societies. In this section we make some predictions regarding how information age politics will play out on platforms.

Target ▶️ ***Actor*** 🔽	***States***	***Platforms***	***Society***
States	Compete over effectiveness of pandemic responses Engage in covert and/or overt influence operations targeting other states.	Seek to impose more control & accountability on platforms States may be extra cautious in their messaging, to keep it in line with evolving content policies.	Seek to impose greater control on information by nudges or coercion or monitor it. Stricter punishment for individuals for sharing 'perceived' misleading content on platforms.
Platforms	Actively moderate speech from political leaders/states. Attempt to work more closely with governments.	Signal intent for closer collaboration to address information disorder. Compete with other platforms for favourable public perception.	Signal 'good faith' actions against information disorder Take on a more interventionist approach to moderating content.
Society	Expect states' social media channels to provide official/reliable information. Calls for government regulation of speech on platforms.	Expect active content moderation from platforms. Self-censorship. Consequently, cede more control of public discourse to platforms.	Community efforts to combat information disorder. Increased polarisation of inter-group interactions on social media.

Table 3 Summary of Interaction between States, Platforms, and Society

States

As per the United Nations' Department of Economic and Social Affairs approximately 86%[85] of member states are using online channels to disseminate COVID-19 related information. Based on statements by various social media platforms, public health authorities have also partnered with them to provide official COVID-19 information and engage with individuals. Governments are using social media as one of many channels to demonstrate the effectiveness of their responses to the pandemic or gain political points with their respective bases. In doing so, they may compare themselves favourably with other states and even engage in propagating information disorder. The primary target of such campaigns is likely to be domestic, however, such operations may also be carried out with the intention of managing[86] perceptions among the international community. The social media platforms analysed in this document have users across multiple countries and are being leveraged for both covert/overt information operations against rival states. The aim[87] of such actions is to erode trust in public authorities among domestic constituents. This was an ongoing trend which has been accelerated by the pandemic. As a result, competition among states for narrative dominance on social media platforms will intensify.

With regard to platforms, there are two ways that states could proceed. In the first scenario, in the Indian context, there is a high probability that the Union Government will advocate for stricter moderation of content by platforms and seek to impose greater control over them. This would help deal with misinformation regarding COVID and would also fit a broader pattern gauging by the proposed changes[88] to the intermediary guidelines calling for intermediaries not to *"host display, upload, modify, publish, transmit, update or share any information that… threatens public health or safety"*. It also advocated for "*tech based automated tools"* for proactively identifying and removing access to unlawful information or content. The Ministry of Home Affairs has also instructed[89] social media companies to delete misinformation from their platforms.

The other slightly unlikely scenario would be for the Government to tread lightly when it comes to dealing with platforms. Platforms' responses to misinformation are still shaping up, and Governments, at the Union and State level, do rely on them to reach out to people and mobilise support.

A recurring[90] trend of the pandemic has been governments and states seeking to acquire more power to monitor and control people and their interactions in a bid to contain their spread. In India, this is reflected in the approach towards information disorder too as different levels of the state machinery appear to have taken a paternalistic approach, issuing strict warnings[91], guidelines[92], and notices[93] to prevent the circulation of 'fake news' and 'mischief' on social media platforms. In some instances[94], authorities have filed[95] FIRs[96] and even made arrests[97] under the various sections of the Indian Penal Code, the Disaster Management Act or the Epidemic Diseases Act for content shared on various social media platforms. There will also be efforts to monitor speech on social media proactively with the stated intent of managing information disorder. In addition, attempts to spread awareness about the dangers of information disorder should be expected.

This trend[98] is also visible in other countries, especially in Asia. As per resources compiled by Poynter/IFCN[99] government actions against information disorder include: laws, media literacy campaigns, internet shutdowns/restrictions, threats, setting up investigative committees/task forces, etc. In the medium/long term this may result in greater control over speech and consequently, dissent.

Platforms

Platforms are already changing their privacy policies and working to manage COVID-related information disorder. This approach broadly fits along the line of states seeking to impose control over them and pushing them to increase content moderation and act on information disorder along standards set by a government within its sovereign territory. However, since these platforms are global in nature different states will seek to impose different standards. They will, thus, be under pressure to intervene directly in cases of speech by political leaders andpublic figures.

Since the responses to information disorder are varied, there is likely to be some competition between them when messaging how successful their efforts have been. Perception management combined with a genuine intention to address the problem, they are also likely to collaborate[100] to address coordinated information disorder operations across platforms. There could also be a convergence of approaches over time, though not all platforms are likely to be willing participants in this.

They have been facing an increasing 'techlash'[101] from society over the last few years. The pandemic has presented platforms with an opportunity to signal that they are 'good faith' actors when it comes to dealing with information disorder. This may even result in slowing down[102] the 'techlash'. However, this will depend largely on how effectively and consistently they are able to balance their actions across a range of political ideologies. Nevertheless, increased responsibility thrust on them by states coupled with a willingness to accept stricter moderation by society, platforms are likely to emerge stronger from the pandemic resulting in greater concentration of power in their hands.

Society

Early surveys[103] by the Reuters Institute indicate a favourable response to government communications around the pandemic. In the face of ambiguity, there is also a demand[104] from academics and civil society groups for governments to engage in social media discourse to limit the spread of information disorder.

Calls for regulating speech and/or democratic oversight of platforms to make them accountable for content posted on them have grown louder. With the COVID-19 pandemic and the immediate harm posed by health-related information disorder, societies placed even more pressure on platforms to act, giving them space to assume the role of 'arbiters of truth'. This has happened even as some platforms publicly stated that they would be relying more on algorithmic content moderation.

Faced with a high degree of uncertainty, societies are looking for technological solutions to address problems in the short/medium term and this has been reflected in the case of information disorder as well. It appears that societies are conceding more control over public discourse to social media platforms in a tradeoff to contain information disorder.

With societies witnessing an increase in tribalism[105] fueled by social media platforms, information disorder and any inconsistencies in platform responses will further cleave them along existing fault lines. This will result in greater polarisation of inter-group engagement on social media. There is also a growing awareness of the importance of fact-checking and realisation that information disorder is both a supply and demand problem. This could result in community-led efforts to increase fact-checking as well as awareness of the harms posed by even passive propagation of information disorder.

Conclusion

Based on this assessment, we conclude that the increasing amount of engagement online will contribute to giving platforms more power and discretion regarding how content is to be handled. In essence, they have been given authority to act as arbiters of truth.

Responses to COVID-19 related information disorder can be classified into two buckets - action and policy. The former involves providing funding to relevant organisations, changes in UI, and changes in information flows. The latter addresses creating new COVID-19 specific policies, modifying existing policies and applying existing policies as they were.

As a public health crisis, COVID-19 has presented an opportunity for platforms to exercise a higher degree of good faith interventionism. However, now that they have taken such steps, it will result in more strident calls for action in more politically sensitive and ambiguous areas. We have already seen this manifest in the form of Twitter flagging tweets by Donald Trump and The White House under citing multiple policies, which has increased pressure on Facebook to follow suit.

We are still in the midst of the pandemic. The evolving nature of information disorder suggests that platforms will have to continue to reactively make changes to their policies and how they are implemented across users and geographies. This resulted in greater subjectivity and complications in enforcement of both existing and COVID-19 specific policies.

Changes to policies and their enforcement may impact the relationships between governments, society, and platforms. The government's relationship with platforms may change with the former advocating for stricter content moderation standards. Alternatively, it may lead to the Government being more cautious about its messaging keeping in mind the updating rules of moderating misinformation.

Society's relationship with platforms may change with the society conceding more control over public discourse to social media platforms in a tradeoff to contain information disorder.

Between government and society, the latter may call for greater moderation of speech on platforms resulting in a tradeoff between free expression and the harms posed by information disorder while governments will seek to control/monitor what people post on platforms.

References

1. Wardle, C. Information Disorder, Part 3: Useful Graphics. *firstdraft* https://medium.com/1st-draft/information-disorder-part-3-useful-graphics-2446c7dbb485 (2018).
2. Neilsen, R. K., Fletcher, R., Brennen, J. S., Newman, M. & Howard, P. Navigating the 'infodemic': how people in six countries access and rate news and information about coronavirus. *Reuters Institute* https://reutersinstitute.politics.ox.ac.uk/infodemic-how-people-six-countries-access-and-rate-news-and-information-about-coronavirus#sum3 (2020).
3. Perez, S. Report: WhatsApp has seen a 40% increase in usage due to COVID-19 pandemic. *TechCrunch* https://techcrunch.com/2020/03/26/report-whatsapp-has-seen-a-40-increase-in-usage-due-to-covid-19-pandemic/ (2020).
4. Roy, T. L. How ShareChat has been monetising stay-at-home traffic. *exchange4media* (2020).
5. Agarwal, S. Youtube, Facebook, and Netflix drive rural India's internet usage. *The Economic Times* (2020).
6. Perez, S. Report: WhatsApp has seen a 40 percent increase in usage due to COVID-19 pandemic. https://techcrunch.com/2020/03/26/report-whatsapp-has-seen-a-40-increase-in-usage-due-to-covid-19-pandemic/?guccounter=1.
7. Schultz, A. & Parikh, J. Keeping Our Services Stable and Reliable During the COVID-19 Outbreak. *Facebook Newsroom* https://about.fb.com/news/2020/03/keeping-our-apps-stable-during-covid-19/ (2020).
8. Twitter Inc. Coronavirus: Staying safe and informed on Twitter. *Twitter* https://blog.twitter.com/en_us/topics/company/2020/covid-19.html#metrics (2020).
9. Business Standard. TikTok hits 2 bn downloads; becomes most installed app amid Covid-19 crisis. *Business Standard* (2020).
10. Anderson, K. & Palen, L. Crisis informatics—New data for extraordinary times. *Science (80-.).* **353**, 224–225 (2016).
11. Starbird, K. How a Crisis Researcher Makes Sense of Covid-19 Misinformation. *medium* https://onezero.medium.com/reflecting-on-the-covid-19-infodemic-as-a-crisis-informatics-researcher-ce0656fa4d0a (2020).
12. Fuscaldo, D. Facebook Now Has More Users in India Than in Any Other Country. *Investopedia* https://www.investopedia.com/news/facebook-now-has-more-users-india-any-other-country/ (2019).
13. NapoleonCat. Instagram users in India. *NapoleonCat* https://napoleoncat.com/stats/instagram-users-in-india/2019/01 (2019).
14. Dredge, S. TikTok expects to have 300m users in India by end of 2020. *Musically* https://musically.com/2020/03/05/tiktok-expects-to-have-300m-users-in-india-by-end-of-2020/ (2020).
15. Laghate, G. Youtube in India has over 265 mn monthly active users 1200+ channels with 1mn+ subs. *Economic Times* (2019).
16. LiveMint. YouTube hits 265 million monthly active users in India. *LiveMint* (2019).
17. Sehl, K. Top Twitter Demographics That Matter to Social Media Marketers. *hootsuite* https://blog.hootsuite.com/twitter-demographics/ (2020).
18. Shrivastava, A. & Sanghamitra, K. Sharechat to stay focused on users, unique content. *Economic Times* (2020).
19. Zuckerberg, M. An Update by Mark Zuckerberg. *Facebook* https://www.facebook.com/4/posts/10111615249124441/?d=n (2020).
20. Comms, I. Instagram Comms on Twitter. *Twitter*

https://twitter.com/InstagramComms/status/1235984308994703360 (2020).
21. Newton, C. How the coronavirus is changing Big Tech. *The Interface* https://www.getrevue.co/profile/caseynewton/issues/how-the-coronavirus-is-changing-big-tech-231110?utm_campaign=Issue&utm_content=view_in_browser&utm_medium=email&utm_source=The+Interface (2020).
22. Newton, C. How coronavirus is changing Big Tech. *The Interface* (2020).
23. Facebook. Combatting Vaccine Misinformation. *Facebook Newsroom* (2019).
24. Instagram. Keeping People Informed, Safe, and Supported on Instagram. *about.Instagram.com* (2020).
25. Instagram. Keeping People Informed, Safe & Supported on Instagram. (2020).
26. Facebook. Combatting Misinformation around the Vaccine. *Facebook Newsroom* https://about.fb.com/news/2019/03/combatting-vaccine-misinformation/?utm_campaign=The Interface&utm_medium=email&utm_source=Revue newsletter (2020).
27. Instagram. Keeping People Informed, Safe, and Supported. *Instagram* (2020).
28. Clegg, N. Combating Coronavirus Misinformation Across Our Apps. *Facebook Newsroom* (2020).
29. Facebook. Facebook Advertising Policies. *Facebook* https://www.facebook.com/policies/ads/prohibited_content/misinformation.
30. Facebook. Facebook Community Standards: False News. *Facebook*.
31. Facebook. Facebook Community Standards: Manipulated Media. *Facebook*.
32. Chaturvedi, A. WhatsApp donates $1 Million to international fact-checking network for coronavirus facts alliance. *Economic Times* (2020).
33. Poynter. Fighting the Infodemic: The #CoronaVirusFacts Alliance. *Poynter*.
34. WhatsApp. The World Health Organization launches WHO Health Alert on WhatsApp. *whatsapp.com* (2020).
35. Grau, M. New WhatsApp chatbot unleashes power of worldwide fact-checking organizations to fight COVID-19 misinformation on the platform. *Poynter* (2020).
36. WhatsApp. About forwarding limits. *whatsapp.com*.
37. Narayan, A. Our efforts towards fighting misinformation in times of COVID-19. *TikTok Newsroom* (2020).
38. TikTok. Keeping our community safe during the COVID-19 outbreak. *TikTok Newsroom* https://newsroom.tiktok.com/en-in/keeping-our-community-safe-during-the-coronavirus-outbreak (2020).
39. TikTok. TikTok mobilizes user community to raise awareness and strengthen fight against COVID-19. *TikTok Newsroom* https://newsroom.tiktok.com/en-in/tiktok-mobilizes-user-community-to-raise-awareness-and-strengthen-fight-against-covid.
40. TikTok mobilises user community to raise awareness and strengthen fight against COVID-19.
41. TikTok. Keeping our community safe during the Coronavirus outbreak. *TikTok Newsroom* (2020).
42. TikTok. Community Guidelines. *TikTok* (2020).
43. TikTok. Safety Center. *TikTok*.
44. Mantzarlis, A. COVID-19: $6.5 million to help fight coronavirus misinformation. *Google Blog* (2020).
45. Gupta, S. & Senupta, C. Google India Blog. *Google India Blog* (2020).
46. AFP. Google Announces measures to fight COVID-19 misinformatio, will verify the location of all advertisers. *Economic Times* (2020).

47. Newton, C. Google has been unusually proactive in fighting COVID-19 misinformation. *The Verge* (2020).
48. Bergen, M. & Vynck, G. Google Scrubs Coronavirus Misinformation on Search, YouTube. (2020).
49. Google. COVID-19 Medical Misinformation Policy. *YouTube Policies* (2020).
50. Team, T. P. P. Stepping up our work to protect the public conversation around Covid-19. *Twitter Blog* (2020).
51. Gadde, V. Protecting and supporting journalists during COVID-19. *Twitter Blog* (2020).
52. Team, T. P. P. Stepping up our work to protect the public conversation around the novel Coronavirus. *Twitter Blog* (2020).
53. Roth, Y. & Pickles, N. Updating our Approach to Misleading Information. *Twitter Blog* (2020).
54. Harvey, D. Helping you find reliable public health information on Twitter. *Twitter Blog*.
55. Team, T. S. Defining public interest on Twitter. *Twitter Blog* (2019).
56. Team, T. P. P. Stepping up our work to protect the public conversation around the coronavirus. *Twitter Blog* (2020).
57. Twitter Inc. COVID-19: Staying safe and informed on Twitter. *Twitter Blog* (2020).
58. Chu, J. & McDonald, J. Helping the world find credible information about novel #coronavirus. *Twitter Blog* (2020).
59. Twitter Inc. World Leaders on Twitter: principles & approach. *Twitter Blog* (2020).
60. Roth, Y. & Pickles, N. Updating our Approach to Misleading Information on Twitter. *Twitter Blog* (2020).
61. Mandavia, M. ShareChat lays of 101 employees as advertising market tanks. *Economic Times* (2020).
62. PTI. ShareChat sees 15% rise in daily average users during lockdown. *The Hindu Businessline* (2020).
63. ShareChat. ShareChat Content and Community Guidelines. *privacy.sharechat.com* https://privacy.sharechat.com/content-policy.html.
64. Statechery. Defining Information. *statechery.com* https://stratechery.com/2020/defining-information/ (2020).
65. Akbar, S. Z., Kukreti, D., Somya, S. & Pal, J. Temporal Patterns in COVID-19 misinformation in India. *michigan.edu* (2020).
66. Twitter Inc. Novel Coronavirus: Staying safe and informed on Twitter. *Twitter Blog* (2020).
67. Rosen, G. An Update on Our Work to Keep People Informed and Limit Misinformation About COVID-19. *Facebook Newsroom* https://about.fb.com/news/2020/04/covid-19-misinfo-update/.
68. Singh, M. WhatsApp's new limit cuts virality of 'highly forwarded' messages by 70%. *TechCrunch* (2020).
69. Gandhi, N. Prioritising privacy and user safety on TikTok, curbing misinformation together. *TikTok Newsroom* (2020).
70. Avaaz.org. How Facebook can Flatten the Curve of the Coronavirus Infodemic. *avaaz.org*.
71. Newton, C. How the 'Plandemic' video hoax went viral. *The Verge* (2020).
72. Staff, S. Covid-19: WHO suspends trial of hydroxychloroquine, after study shows it increases mortality rate. *Scroll.in* (2020).
73. Saikia, A. Covid-19: India is relying on flimsy evidence to expand use of HCQ despite concerns about dangers. *Scroll.in* (2020).
74. Newsguard. Tracking Facebook's 'Super Spreaders' in Europe. *Newsguard.com* (2020).
75. Newsguard. Tracking Twitter's super spreaders. *Newsguard.com* (2020).
76. Adams, T. 5G, coronavirus and contagious superstition. *The Guardian* (2020).

77. Lyons, K. Twitter removes tweets by Brazil, Venezuela presidents for violating COVID-19 content rules. *The Verge* (2020).
78. Stalin, S. Why Rajinikanth's Post On 'Janata Curfew' Was Removed By Twitter. *NDTV* (2020).
79. Hannon, E. Twitter, Facebook Delete World Leaders' Misleading Coronavirus Posts. Could Trump Be Next? *Slate.com* (2020).
80. Allyn, B. Researchers: Nearly Half Of Accounts Tweeting About Coronavirus Are Likely Bots. *NPR* (2020).
81. Douek, E. Evelyn Douek on Twitter. *twitter* (2020).
82. Collins, B. & Zadrozny, B. Troll farms from North Macedonia and the Philippines pushed coronavirus disinformation on Facebook. *NBC* (2020).
83. Kundu, C. No Title. *India Today* https://www.indiatoday.in/fact-check/story/fact-check-no-clapping-together-at-5-pm-during-janta-curfew-will-not-kill-coronavirus-1658438-2020-03-22 (2020).
84. Sharma, P. Fake coronavirus cures on social media include ginger, honey & hot air. *WION* (2020).
85. United Nations Division for Public Institutions and Digital Government. UN/DESA Policy Brief #61: COVID-19: Embracing digital government during the pandemic and beyond. *un.org* (2020).
86. Rosenberger, L. China's Coronavirus Information Offensive. *Foreign Affairs* (2020).
87. Singer, P. W. & Brooking, E. T. *Likewar: The Weaponization of Social Media*. (Eamon Dolan, 2018).
88. MEITY. The Information Technology [Intermediaries Guidelines (Amendment) Rules] 2018. *meity.gov.in* (2018).
89. Government, E. Government tells social media platforms to remove videos citing misinformation. *Economic Times* (2020).
90. Roth, K. How Authoritarians Are Exploiting the COVID-19 Crisis to Grab Power. *Human Rights Watch* (2020).
91. IANS. DC bans unverified news on social media in Raj district. *newsd* https://newsd.in/dc-bans-unverified-news-on-social-media-in-raj-district/ (2020).
92. Arunachal24.in. WhatsApp admins to be held responsible for 'fake news, rumours' on groups-Arunachal Police. *Arunachal24.in* (2020).
93. India, D. Order aimed at curbing social media misinformation, not govt criticism: Mumbai Police. *DNA India* (2020).
94. Dore, B. Fake News, Real Arrests. *Foreign Affairs* (2020).
95. Navhindtimes.in. Unknown person booked for spreading rumour on COVID-19. *navhindtimes.in* (2020).
96. PTI. Web portal owner booked for spreading fake news in Punjab's Kapurthala. *theweek.in* (2020).
97. NE Now News. Mizoram: Two arrested for a fake Facebook post. *North East Now News* (2020).
98. Sochua, M. Coronavirus 'Fake News' Arrests Are Quieting Critics. *Foreign Policy* (2020).
99. Funke, D. & Flamini, D. A guide to anti-misinformation actions around the world. *Poynter*.
100. Shu, C. & Shieber, J. Facebook, Reddit, Google, LinkedIn, Microsoft, Twitter and YouTube issue joint statement on misinformation. *TechCrunch* (2020).
101. Smith, E. The techlash against Amazon, Facebook and Google—and what they can do. *The Economist* (2020).
102. Times, F. So much for the 'techlash'. *Financial Times* (2020).
103. Neilsen, R. K., Fletcher, R., Newman, M., Brennen, J. S. & Howard, P. Navigating the

'infodemic': how people in six countries access and rate news and information about COVID-19. *Reuters Institute* (2020).
104. Limaye, R. J. *et al.* Building trust while influencing online COVID-19 content in the social media world. *Lancet* **2**, 277–278 (2020).
105. Wheeler, T. Technology, tribalism, and truth. *Brookings* (2020).