The spread of (mis)information: A social media experiment in Pakistan

Sarojini Hirshleifer, Mustafa Naseem, Agha Ali Raza, Arman Rezaee

PII: S0304-3878(26)00067-2
DOI: https://doi.org/10.1016/j.jdeveco.2026.103784
Reference: DEVEC 103784

To appear in: *Journal of Development Economics*

Received date : 7 July 2025
Revised date : 14 March 2026
Accepted date : 16 March 2026

Please cite this article as: S. Hirshleifer, M. Naseem, A.A. Raza et al., The spread of (mis)information: A social media experiment in Pakistan. *Journal of Development Economics* (2026), doi: https://doi.org/10.1016/j.jdeveco.2026.103784.

The spread of (mis)information: A social media experiment in Pakistan*

Sarojini Hirshleifer†, Mustafa Naseem‡, Agha Ali Raza§, and Arman Rezaee¶

March 14, 2026

Abstract

This study examines how controlling misinformation on a social media platform in Pakistan affects users' exposure to both accurate and false information. It combines an intervention to disseminate official information about the COVID-19 pandemic across the platform with a randomized experiment that measures the impact of fully controlling access to pandemic-related misinformation. The treatments rely on a higher-intensity, *ex ante* approach to moderating misinformation on the platform relative to the control, which relies on a typical *ex post* approach to moderation. Fully controlling misinformation, as in the treatments, reduces the number of daily users by 19%, indicating a distaste for moderation. Furthermore, the treatments reduce exposure to official information by 29% more than they reduce exposure to misinformation. A conceptual framework posits that these findings can be explained by the fact that, in this setting, official information is more trusted, and thus is more widely disseminated, relative to misinformation.

Keywords: social media, misinformation, information, digital economy, political economy, development economics, health, COVID-19, field experiment, randomized controlled trial
JEL Codes: L86, L82, D80, D83, O10, I10, I15, P00

*We gratefully acknowledge that this work was partially supported by the National Institutes of Health grant 5R21HD095696-02. We thank Fizzah Malik, Namoos Hayat Qasmi, Shan Randawa, Sacha St-Onge Ahmad, and Behzad Taimur for providing support for the implementation of the project and for extracting the data from the platform. We also thank Marcella Alsan, Natalie Bau, Roe'e Levy, Arun Chandrasekhar and seminar participants at the CEGA Research Retreat, University of Oregon, UC Irvine, UC San Diego Rady School of Business, BREAD, and NBER Summer Institute: Digital Economics and Artificial Intelligence for helpful feedback.

†University of California, Riverside and CEGA (email: sarojini.hirshleifer@ucr.edu)
‡University of Michigan (email: mnaseem@umich.edu)
§Lahore University of Management Sciences (email: agha.ali.raza@lums.edu.pk)
¶University of California, Davis and CEGA (email: abrezaee@ucdavis.edu)

1 Introduction

Social media is a powerful tool that can dramatically reduce the cost of information sharing and reach people who may not regularly engage with formal media. That any user can share content from any source, however, poses a risk: social media can allow both helpful, accurate information and harmful, inaccurate information to be disseminated much more widely than it otherwise would have. Both of these potential roles for social media are particularly relevant during times of crisis, such as political events, natural disasters, and the COVID-19 pandemic. On the one hand, sharing high-quality information on social media is widely recognized as an important tool for policymakers.[1] On the other hand, misinformation is especially likely to spread on social media, and is a critical risk factor in the harms caused by crises.[2] Although managing the dissemination of information on social media is a major policy challenge globally, it is particularly relevant to developing countries where high-quality information is more likely to be scarce.[3]

Using a randomized experiment, we study the impact of fully controlling access to misinformation across a social media platform on users' exposure to both official, high-quality information and misinformation itself.[4] Most platforms rely on *ex post* moderation which exposes users to misinformation before it is taken down. Given the potential harms of misinformation, however, a question arises as to the implications of a platform free of misinformation. In order to fully control access to misinformation, it must first be identified. This requires *ex ante* moderation of *all* information before it reaches the platform. Thus, a complete approach to controlling misinformation

[1] Government and official use of social media is widespread, and is recommended by researchers given high levels of engagement from users (Lin et al., 2016; Tursunbayeva, Franco and Pagliari, 2017). For example, the crisis communication plan of the CDC highlights the value of disseminating information through such platforms (CDC, 2018). In addition, during the early days of the pandemic, social media companies used their platforms to actively disseminate relevant information. Facebook had a coronavirus information center at the top of news feeds for a time (Dwoskin, 2020), while Twitter (which is now X) had a section on its Explore tab dedicated to news on COVID-19 (Twitter, 2020).

[2] Allcott and Gentzkow (2017) document the role of social media in spread of "fake news" during the 2016 U.S. presidential election. During the pandemic, the Director-General of the World Health Organization (WHO) noted in 2020, "...we're not just fighting an epidemic; we're fighting an infodemic. Fake news spreads faster and more easily than this virus, and is just as dangerous" (WHO, 2020). More generally, health misinformation on social media has been a concern since before the pandemic (Wang et al., 2019). Misinformation has also been a challenge in addressing natural disasters (Hsu, 2023).

[3] U.S. government agencies have frequently interacted directly with social media companies, which highlights the relevance of the dissemination of information on social media to first-order policy issues. The limits of this interaction are a topic of ongoing debate (Fung and Cole, 2023). Developing countries are widely perceived to be information scarce, particularly with regards to health information (Stiglitz, 2000; Kremer and Glennerster, 2011).

[4] We use the term "official information" throughout the paper when discussing evidence-based information or recommendations from policymakers. We recognize that the accuracy of this information may evolve over time along with the available evidence. In contrast, we use the term, "misinformation" to reflect information that is not grounded in an evidence-based understanding of the world. In the model that is put forth in Section 2, since both are intended to be a crystallized representation of the real world, and it is from the perspective of the policymaker, we simplify these concepts to "good" and "bad" information.

also has implications for the dissemination of official information. This study, in comparing *ex ante* to *ex post* moderation, provides insight into the trade-offs of implementing such a platform as well as the conditions under which it may be optimal. Conditional on fully controlling misinformation, we also examine two approaches to addressing it: never exposing misinformation or rebutting it directly.

This study took place in the context of a social media platform in Pakistan called Baang, early in the COVID-19 pandemic. The study had two main components. First, we used the platform to disseminate information about COVID-19 in the form of *official* posts from Baang. This content was available to all users of the platform throughout the study. Second, we implemented a novel, user-level randomized experiment that varied the approach to controlling user-generated misinformation about COVID-19 on the platform. In the control condition, users have access to a version of the platform that relies on lower-intensity, *ex post* moderation to address misinformation. The treatment versions of the platform, however, rely on higher-intensity, *ex ante* moderation: all user-generated content is reviewed by a moderator before being posted on the platform. In the *never post* treatment, content that includes misinformation is simply never posted to the platform, while in the *sunshine* treatment, misinformation is posted along with a rebuttal that debunks it. These rebuttals include high-quality information in line with official posts. A user's own condition assignment does not affect how their posts are distributed on the platform. Thus, aside from the posts that included misinformation, all content on the platform was the same in all of the conditions.

The Baang platform is a voice-based platform with a few thousand users calling in during the two months of the experiment. Its main features, however, are fundamental to social media and shared across all major social media platforms. Furthermore, it has direct implications for many other voice-based, non-profit platforms in developing countries (Raza et al., 2018). The content on Baang is generated by users in general, and this decentralization of content creation is an essential characteristic of social media platforms. Baang is typical with regards to the types of content posted by users, which focuses on the news, personal stories, and religious practice. Baang also has all of the standard mechanisms that allow users to engage directly with each other's content on social media: they can comment on, share, like and dislike each other's posts. All of the posts on Baang are also public and anonymous, and can be ordered by newest or most popular. Thus, the structure of Baang is analogous to the main page of reddit, one of the world's largest social media platforms (Curry, 2023). Furthermore, the users of Baang are in a demographic of young men with modest levels of education, which is of particular policy interest in developing countries, given their potential role in influencing a country's political and social stability (World Bank, 2006). This platform has broad external validity to other platforms and settings.[5]

[5]This is further discussed in Section 3.3.

A simple framework generates three main hypotheses that structure the results in this paper. The framework focuses primarily on the case of a social welfare maximizing social media operator, since misinformation on social media is increasingly the focus of regulation.[6] Determining the socially optimal approach to regulating misinformation is a first order question in this context. Since most platforms are for profit, however, we also consider a straightforward extension to a profit maximizing operator. In the framework, the social welfare maximizing operator chooses a high or low level of moderation to maximize net information exposure (i.e. good exposure net of bad). A user then chooses their levels of exposure to good and bad information, which are determined by their total exposure to the platform as well as their relative levels of trust in the sources of good and bad information.

The first prediction of our framework is that fully controlling access to information, as in both of the treatments, limits the overall usage of the platform if users have a distaste for moderation. Thus, we begin by documenting that the treatments reduce usage of the platform on both the extensive and intensive margins. On average, the treatments have 43.1 (19%) fewer daily users who spend 6.9 (26%) fewer daily minutes on the platform than in the control (both significant at the 1% level).

The second prediction is that if overall usage of the platform declines, then exposure to good information will also decline. This is confirmed by our first main result. We find that there is meaningfully *less* dissemination of official information in the treatments relative to the control condition. On average, users in the treatments listen to 0.33 (25%) fewer minutes of official posts (significant at the 5% level). We find qualitatively similar results for user-generated posts whose content is aligned with the official information.

The third prediction of our framework focuses on the impact of high moderation on users' *net* information exposure, that is, their exposure to official information and aligned user-generated information net of misinformation. It proposes that, given a distaste for moderation, high moderation will have a negative impact on net exposure if users' have a more favorable perception of the source of official information compared to misinformation. In the never post treatment, net information exposure *declines* by 21% relative to the control. Thus, even though we never post all of the misinformation to the platform in that treatment, the decline in exposure to official information is meaningfully larger than the decline in exposure to misinformation. In the sunshine treatment, including rebuttals as a source of official information, the decline in net information exposure is 29%. These effects are largely driven by exposure to official information.

[6] In 2024, the EU's Digital Services Act, which regulates content on large platform went fully into effect (EU, 2024; Chan, 2023). In 2023, UNESCO proposed a global framework for regulating social media (UNESCO, 2023).

Furthermore, consistent with the conditions in the framework for negative net information exposure, we find that users in this setting have relatively favorable perceptions of the official information. Specifically, 95% of users indicate that they trust official Baang posts more than user-generated posts. According to the framework, these perceptions matter because they determine users' relative exposure to good and bad information, for a given level of total exposure. That is also consistent with our findings in this experiment. In the control, the average official post is listened to 653.2 more times and shared 62.5 more times than the average misinformation post, which is listened to 11.0 times and shared 0.0 times.[7]

Finally, we examine the mechanisms that are driving the distaste for *ex ante* moderation in this setting. Ex ante moderation has two implications for how users experience the platform. The first implication is that it changes users' exposure to misinformation. One reason that users may have a distaste for moderation is if they have a preference for being exposed to misinformation. These preferences could be utility-maximizing or they could represent a behavioral response.[8] In that case, exposure to misinformation may increase future time spent on the platform. Thus, we examine how users respond to being exposed to misinformation posts compared to how they respond to being exposed to matched user-generated posts that do not contain misinformation. In the control, users spend relatively more time on the platform after being exposed to misinformation. Thus, we find evidence for the preference for misinformation mechanism.

Another reason that users may have a distaste for *ex ante* moderation is that it causes all content to be posted with a modest delay. This is the cost of a platform that is free of misinformation. In fact, all moderation leads to delays in identifying and addressing misinformation, even on major social media platforms.[9] The difference between *ex ante* and *ex post* moderation is how that delay is experienced by users. Most major platforms rely on *ex post* moderation, and thus delays leave misinformation on the platform for some period of time, which allows it to be disseminated. In the case of *ex ante* moderation, the user-generated content that is not misinformation will appear on the platform with some delay instead. The average delay from *ex ante* moderation in this experiment is relatively modest compared to *ex post* delays on major platforms. If users have a distaste for those delays here, however, the effects should be concentrated after posting as that is when users experience them directly. Using an event study approach, we demonstrate that for users who post,

[7] While our experiment was not designed to test it, we note that an interesting question is whether greater engagement with official information, including rebuttals, could have been generated through making it more entertaining. For example, research in Nigeria finds that watching edutainment TV with a sub-plot aimed at changing men's views around domestic violence is associated with large decreases in the belief that domestic violence can be justified (Banerjee, La Ferrara and Orozco, 2019).

[8] For example, as we discuss in Section 6, exposure to misinformation may lead to emotional disregulation which can increase addictive behaviors, and social media use may have some characteristics of an addictive behavior.

[9] See Section 6 for further details.

treatment effects are concentrated in the period after their initial post.

Our framework also highlights the settings in which the impact of high levels of moderation on net information exposure would be positive, instead of negative as we observe here. In particular, users' relative exposure to official information as opposed to false information under low moderation matters. In the setting of this experiment, relatively high levels of trust in the official information is likely instrumental in inducing users to seek out that type of information at high rates. This is in contrast to previous research on Twitter, for example, which finds that misinformation has a higher level of engagement relative to other types of information (Vosoughi, Roy and Aral, 2018). In that type of setting, which is characterized by relatively high levels of engagement with misinformation, the framework indicates that high intensity moderation would be optimal, even if users have a distaste for moderation. Furthermore, it clarifies that high intensity moderation will still not be optimal for a profit maximizing operator in such a setting. Therefore, the framework illustrates that the incentives of profit motivated firms and regulators are not aligned in settings with high levels of engagement with misinformation.

This is the first publicly available experiment that fully controls access to misinformation across an entire social media platform.[10] This design uniquely allows us to consider the implications of fully controlling for misinformation on the dissemination of all types of information. It also avoids a potential external validity concern that is inherent in most experiments that rely on larger platforms, since they typically are conducted on a selected subset of users who respond to an ad and willingly install a plug-in that will affect how the platform appears to them. Furthermore, although these results are broadly relevant, this study is the first to experimentally examine questions of misinformation or moderation in a development setting. This is particularly important given the increasingly widespread use of social media in developing countries (We Are Social and Meltwater, 2024).

The few previous experiments on misinformation and social media have relied on selected samples as well as controlled environments, and have tested the impact of exposing people to an individual piece of misinformation as well as various approaches to addressing that misinformation (Barrera et al., 2020; Pennycook et al., 2020; Henry, Zhuravskaya and Guriev, 2022; Guriev et al., 2023).[11] In contrast, our experiment is designed to understand how naturalistic exposure to misinformation affects social media usage overall as well as exposure to official information. This further allows us to consider the broader policy implications of a platform free of misinformation.

[10] We know that social media companies experiment widely, but they do not always share the results of those experiments publicly.

[11] There is also a descriptive literature on the dissemination of misinformation on social media, see for example Vosoughi, Roy and Aral (2018) and González-Bailón et al. (2023).

In addition, this study is related to concurrent experiments on the moderation of toxic content. Jiménez-Durán (2022) finds minimal effects of moderation on those being moderated directly, while Beknazar-Yuzbashev et al. (2025) finds evidence for a distaste for moderation effect on Facebook and Twitter that is similar in magnitude as the one in this experiment. Both of these experiments rely on users downloading a plug-in to participate, and thus they are on a selected subset of platform users. Kalra (2025) reduces toxic content on a large platform in India by reducing the importance of a personalized algorithm. The experiment presented in this paper, however, is unique in examining the impact of moderation in the context of disseminating official information, which has important policy implications. The fact that our initial result on overall usage is reflected in other settings, however, is suggestive evidence of the relevance of our main results on official information exposure to other contexts. This paper also considers the unique implications of moderating misinformation, which is likely to be more costly to identify relative to toxic content.

This experiment is also related to a broader literature in economics on how social media can expose people to information and other types of persuasive content. Thus far, it has largely examined the impact of social media on political attitudes and outcomes.[12] This literature largely relies on natural variation; two recent exceptions are Levy (2021) and Nyhan et al. (2023) who conduct experiments that examine how varying the political polarization of content affects users' attitudes. Another thread has measured the impact of reducing exposure to social media on information acquisition (Allcott et al., 2020; Mosquera et al., 2020). A few recent experiments have examined the potential to use social media to disseminate useful health information through third-party advertising or influencers (Breza et al., 2021; Alatas et al., 2024). In this setting, however, the source of useful health information is the platform itself.

This paper proceeds as follows. Section 2 provides a conceptual framework that includes the perspective of a social welfare maximizing operator as well as for-profit operator. Section 3 explains the context, platform, and particularly, the external validity of the study. Section 4 describes the experiment. Section 5 presents the results informed by the three hypotheses generated by the framework. Section 6 contextualizes and examines evidence for two key mechanisms. Section 7 concludes.

[12] See, for example, recent work such as Bond et al. (2012); Enikolopov, Makarin and Petrova (2020); Bursztyn et al. (2020); Fujiwara, Müller and Schwarz (2024) as well as Zhuravskaya, Petrova and Enikolopov (2020) for a broader review.

2 Conceptual Framework

A simple two-period framework formalizes the hypotheses we test in this experiment. Unlike many settings, where agents are assumed to be passive recipients of a fixed amount of information, here, exposure to information is agents' key choice variable. This is a reflection of how users engage with social media platforms as well as specific types of information on those platforms. Like other researchers, we have found users' engagement on social media to be highly sensitive. In addition, note that we primarily develop this framework assuming a social welfare maximizing operator of a social media platform. Given the growing interest in regulating misinformation on social media, this is an inherently important question. Below, however, we also discuss a straightforward extension to a profit maximizing operator.

In the first period, a social welfare maximizing operator of a social media platform chooses either a high or low level of moderation m_k for $k \in h, l$ in order to maximize a user's positive net exposure to information $G = g - \theta b$, where g is the user's exposure to official, good information on the platform, and b is their exposure to user-generated misinformation. The operator could also assign a weight, θ, if they believe that the harm from bad information is greater than the benefit from good information, or vice versa.[13] In the second period, the user chooses their total exposure to the platform, $t(m_k)$, which is a function of the level of moderation. The user's exposure to both good $g(t(m_k), p_g)$ and bad $b(t(m_k), p_b)$ information is a function of t. It is also a function of p_j, the user's perceptions about the trustworthiness of a given source of information for $j \in g, b$, with relatively higher levels of perceived trustworthiness of an information source inducing relatively higher consumption.[14] Thus, net exposure is given by: $G_k = g_k(t(m_k), p_g) - \theta b_k(t(m_k), p_b)$. Using backward induction, the operator will choose a level of moderation by comparing G_l and G_h.

A functional form simplification helps to more clearly illustrate the hypotheses generated by this framework. Specifically, let $g_k = p_g * (t(m_k))$ and $b_k = p_b * (t(m_k))$, where the perceptions regarding the sources of good and bad content are simply the probabilities of seeking out the two types of content. Then, $G_l > G_h$ when $(p_g - \theta p_b)t'(m_k) < 0$. That is, low moderation will be optimal when $t'(m_k)$, which is the preference for moderation, and $p_g - \theta p_b$, which is the relative exposure to good as opposed to bad information, have the different signs. We consider potential mechanisms that

[13] Note, we abstract away from the quantity of good or bad posts on the platform and rather focus on the time users spend listening to those posts. Individual posts on social media can have dramatically varying levels of reach and thus the quantity of posts is likely to be second order to the amount of exposure.

[14] We refer to perceptions, rather than beliefs here, since evaluating whether users update is not in the scope of this study. It is intuitive that users will spend more time consuming information from sources perceived to be of higher quality. For example, someone who trusts the *New York Times* and has little trust in Fox News is likely to much more time consuming news from the former source, while someone with the opposite perceptions is likely to spend their time in a way that is reversed.

can explain the sign of $t'(m_k)$ in Section 6.

The framework generates the three main hypotheses that we focus on testing in this paper. First, we can identify the sign of $t'(m_k)$ by measuring the impact of treatment on overall usage of the platform. Specifically, we call the special case in which $t'(m_k)<0$ a distaste for moderation. Second, the direction of the change in g from control to treatment will be determined by the sign of $t'(m_k)$.[15] The third hypothesis is concerned with the conditions under which $G_l > G_h$. Specifically, if $t'(m_k) < 0$, then it must be the case that $p_g - \theta p_b > 0$.[16] That is, our third hypothesis is that, conditional on a distaste for moderation, low moderation will be optimal if the perception of, or trust in, the source of official information is greater than that of misinformation.

The predictions regarding the impact of sunshine as opposed to never post are ambiguous. The sunshine treatment will expose users to more misinformation relative to the never post treatment. It can also, however, expose users to more official information in the form of rebuttals. Unlike the official posts, which users must seek out, the users may come across the rebuttals in the course of listening to standard feeds.

This framework also has broader relevance. In particular, it highlights in what contexts high, rather than low, levels of moderation are optimal: $p_g - \theta p_b$ and $t'(m_k)$ should have the same sign. Thus, even when users have a distaste for moderation, if there is more bad information than good information on the platform, the social welfare maximizer will choose high moderation.[17] In addition, it is straightforward to extend the framework to platforms where users are not anonymous. Although in this setting, official information comes only from the platform, it could also come from official government accounts or other trusted sources.[18] In that case, p_g and p_b would be the weighted averages of the perceptions of the sources of good and bad information.

[15]Note we do not have a general hypothesis here concerning b since it varies across treatments, and it is zero by construction in the never post treatment. Furthermore, our data clearly supports a monotonic relationship for $g_k = p_g * (t(m_k))$. Whether $b_k = p_b * (t(m_k))$ is more difficult to test given the specifics of our study design. A more general model could allow p_b to vary according to the level of moderation. Instead, for this experiment, we simply allow the never post treatment to be a special case in which $b_k = 0$.

[16]Considering how policymakers should set θ is beyond the scope of this paper, and thus we will set it equal to one in our analysis. We also defer the question of cost-effectiveness to future work. It would be straightforward to include a cost parameter for G_h in the model, but it would not materially change our current hypotheses. Finally, note that $p_g - p_b > 0$ is sufficient for $G_l > G_h$, when $t'(m_k) < 0$. In the never post treatment, however, a further restriction is required for $G_l > G_h$, specifically, $p_g t'(m_k) + p_b * t(m_l) < 0$. That is, the absolute magnitude of $p_g t'(m_k)$, which is the decline in good information from treatment to control, must be larger than $p_b * t(m_l)$, which is total reduction in misinformation from control to treatment in that case.

[17]Note that if users have a preference for moderation, then high moderation will be optimal when $p_g - \theta p_b > 0$. In addition, in the special case of the never post treatment, if there is a preference for moderation, then high moderation is always optimal.

[18]On such platforms, users regularly see and make decisions about what content to engage with based on its source (Levy, 2021).

We primarily focus on the social planner case in this framework, not only because it is the first order question from a policy perspective, but also because the profit maximizing case is relatively straightforward. Since the objective of the profit maximizing social media platform operator is to maximize engagement, in the context of this framework, that implies maximizing total exposure: $T_k = g_k(t(m_k), p_g) + \theta b_k(t(m_k), p_b)$. Then, it is straightforward to see that a profit maximizer will choose the intensity of moderation based only on users' distaste for moderation. That is, if users have a distaste for moderation, the profit maximizing firm will not moderate misinformation. Thus, if it is also the case that $p_g - \theta p_b < 0$, then it will be socially optimal to moderate, but firms will not moderate. Therefore, regulation is likely to be appropriate in such contexts.

This framework focuses on user exposure to information, since that exposure represents important choices that users make with regards to their information-seeking behavior. Users' perceptions, however, could further lead them to weigh some sources of information more highly than others per unit of exposure. For example, higher levels of trust in official as opposed to user-generated information could lead users to not only increase their relative exposure to good information, but also give that information more weight in forming beliefs.[19]

3 Context

In the context of a public health crisis, disseminating health information widely and quickly is an important policy challenge in all types of countries. It is particularly relevant in the context of low-income countries such as Pakistan which are characterized by limited access to health information as well as health systems with limited capacity (Kremer and Glennerster, 2011; Dupas and Miguel, 2017). Holding other factors equal, the limited capacity of the health system increases the mortality risk from any outbreak. For example, Pakistan has 6.3 hospital beds per 10,000 people compared to the global average of 27.9 (WHO, 2021).

The study was implemented in the context of Baang, a non-profit, voice-based social media platform in Pakistan (Raza et al., 2018, 2022).[20] Voice-based social media platforms have relevance in many development contexts since they can be used by low-literate populations and those without an internet-connected phone or computer. Such platforms reach millions of users, especially in India, with Mobile Vaani being the most prominent example (Moitra et al., 2016). While these platforms often consist of primarily user-generated content, they also often aim to address particular information gaps. Baang followed this model during the time period of the experiment. It was a

[19] If prior beliefs about specific sources of misinformation are formulated outside the model, it may be more effective to directly rebut misinformation. That would make the sunshine treatment optimal *if* exposure to misinformation and official information is constant across the never post and sunshine treatments.

[20] Baang means rooster call in Urdu.

highly active general social media platform, but also had a few official posts about the COVID-19 pandemic.

Other platforms had a focus on promoting citizen journalism (Marathe et al., 2015), agricultural information exchange among farmers (Patel et al., 2010), connecting employers and employees in rural settings (White et al., 2012), and allowing people in rural areas to ask questions of community health workers (Sherwani et al., 2007).[21]

3.1 Baang platform

When users call into Baang, they are presented with a menu that gives them the option to: (1) listen to official Baang posts about COVID-19, (2) record their own posts, (3) listen to others' posts, or (4) listen to their own previously recorded posts.[22] After selecting option (3), users can then choose how they listen to others' posts, by: (a) newest, (b) trending (today's most liked), or (c) overall most liked. After each post plays, users are given the option to record an audio comment, listen to existing audio comments, forward (i.e. share), like, dislike or flag the post before moving on to the next post in the stream.[23] At any point while listening to posts, users can skip to the next post, and they frequently take advantage of this option. Option (1) is identical to (3) in that users are presented with a stream of posts and can comment on and engage with those posts, except (1) only includes the seven official Baang posts about COVID-19.[24] Finally, all of the users of Baang are anonymous, in so far as there are no public identifiers for each user. All of the posts are public.

Baang, like other voice-based social media platforms in developing countries, is a nonprofit platform. It has relied on grant funding to operate, and thus the operators have only been able to launch free deployments of Baang for limited periods of time. In this setting, ensuring the platform is free to use requires that the platform pays for the airtime of users while they are on the platform.[25] During initial free deployments over two years beginning in 2015 that totaled eight months, the platform reached more than 10,000 users through a combination of advertising and organic spread (Raza et al., 2018). These users actively engaged with the platform by calling in 293,657 times to participate through 35,677 posts and 155,352 comments. The posts were played 2.5 million times.

[21] See Raza et al. (2018) for further discussion of Baang and how it compares to other voice-based platforms.

[22] See Figure SA1 for a visual representation of the structure of the Baang platform.

[23] Sharing users input the phone numbers of anyone that they wish to receive a post. Each receiving phone number is then sent an SMS message, which includes the sharing user's phone number, inviting them to call into Baang to listen to the shared post. If the receiving user calls in following the SMS invitation, they are taken straight to the shared message before being sent to the main menu.

[24] See Section 4 for more details on the official Baang COVID-19 posts.

[25] As in most developing countries, in Pakistan, people typically pay for cell phone airtime by the minute, and purchase it in relatively small amounts at a time.

In the months prior to the RCT, however, the platform was not free and users had to pay for their minutes when they called in. Thus, it had a smaller number of committed users, with 392 calling in on a typical day before treatment began. Although we made the platform free to use again at the beginning of the RCT, we did not advertise the platform during that time period.

3.2 Survey and user characteristics

Several months after the experiment, we conducted a phone survey on a sub-sample of 259 Baang users. This allowed us to learn users' demographics as well as their perceptions of different sources of information, including official Baang posts.[26]

The user base of Baang is largely younger males with modest education levels.[27] The average user is 30 years old, and around half (47%) have less than 10 years of education (Table 1). One-fifth have less than 8 years have education, and thus never reached upper secondary. Ninety-nine percent of users are male. Given that voice-based platforms are designed to be accessible to people without smartphones, a higher than expected percentage of users have a smartphone (91%). In addition, almost all of those with a smartphone (96%) regularly use WhatsApp, a common form of social media in this setting.[28] This suggests the potential broader relevance of voice-based platforms, since most users could spend their time on other higher profile social media platforms but use Baang anyway.

3.3 External validity

This experiment relies on a smaller-scale platform, but this allows for a unique advantage relative to experimenting on larger platforms with regards to external validity. In our experiment, everyone on the platform participates, and they are unaware of the experiment. In contrast, many experiments on larger social media platforms rely on specific and likely meaningfully selected subsamples of users who respond to an ad, agree to a survey, or opt-in to downloading a plug-in that controls their access to the platform (Beknazar-Yuzbashev et al., 2025; Levy, 2021; Allcott et al., 2020; Mosquera et al., 2020). Those may be users that already are interested in changing their relationship to social media. Furthermore, those experiments have generally targeted English-speakers in wealthy

[26]The survey took place in April 2021 and included three samples: 94 randomly sampled users, 87 of the most active users, and 86 of the users most exposed to misinformation. Since these three groups of Baang users all have similar characteristics in practice, we focus on the randomly sampled users in the analysis discussed in the paper. This survey was never intended to collect outcomes as is self-evident from the small sample size. See Section 5.2.1 for a discussion of the data on perceptions collected in the survey.

[27]That women mostly do not participate is perhaps unsurprising given the barriers to female participation in some aspects of public life in Pakistan (Schwab et al., 2016).

[28]Although WhatsApp is a messaging application, in many countries it is also used as social media as users join large groups where they do not know the other members.

countries. Thus, neither approach has an obvious advantage in external validity. Nonetheless, we assess the external validity of this study in depth with regards to the four 'transparency conditions' (selection, attrition, naturalness, and scaling) proposed by List (2020). Given its ability to meet those conditions, we find a compelling argument for the external validity of this study.

We begin by considering the selection condition, which considers how the study population is selected. Baang is in Pakistan, which is a large low-income country of 235 million people (World Bank, 2024). Men in Pakistan have a median age of 23 (Central Intelligence Agency, 2024) and 29% of them have completed secondary school (World Bank, 2024). A number of developing countries have similarly a large demographic of young, relatively uneducated men. This group is of particular interest in many developing country contexts, as they are seen as a high risk group to foster social and political instability and conflict (World Bank, 2006; Blattman and Annan, 2016).[29]

Baang users are a selected subset of this highly relevant demographic of younger males with modest levels of education. As discussed above, Baang users are 29.6 years old on average and 53% have completed secondary school. Thus, Baang users are only modestly older and more educated than men in Pakistan on average. Baang users also use other social networks such as WhatsApp (91%), so they are likely indicative of social media users more generally. The fact that Baang users are almost entirely male is also representative of social media users given Pakistan's large gender gap in mobile phone usage (GSMA, 2021).

This study is particularly well positioned on the next two transparency conditions. With regards to attrition, we observe the universe of Baang users and all of their actions on the platform across the entire study. In particular, the outcome variables for our main specifications, such as time spent on the platform, are observed once for each person who is part of the study, and thus, there is zero attrition on these measures. The study also performs well on naturalness. Our data capture Baang users' normal day-to-day usage of the social network except for what is affected by our treatment.

3.3.1 Relevance to other platforms

Finally, we consider scaling. To do so, we consider the relevance of this platform to other platforms. Of course, this study is highly relevant to the millions of people who use voice-based social media platforms in South Asia, which are discussed above. This alone gives the study broad relevance.

[29]For discussion of the so-called "youth bulge" in Pakistan, see, e.g., Hafeez and Fasih (2018); Idrees et al. (2023); Urdal and Hoelscher (2009).

In addition, Baang has the fundamental characteristics of major social media platforms, such as Facebook or X (Boyd and Ellison, 2007). Users on Baang can generate and publicly share content, and then they can engage with other's content in a number of ways. Furthermore, posts on Baang do receive substantial engagement; the average post receives 4.4 comments, 6.4 shares, and 7.2 likes.[30] Of the major social media platforms, Baang is most analogous to the main page of reddit, which has been frequently called 'the front page of the internet' (Singer et al., 2014). reddit is a popular social media platform with 5 (2) billion monthly visits from across the globe (U.S.), and is the 7th (4th) most visited website in the world (U.S.), with a similar popularity to Instagram (Semrush, 2023). The format of Baang is also similar to browsing the "trending topics" page on X. Both the main page of reddit and the trending topics page on X feature posts that are simply the most popular topics on their respective platforms at the time and, like Baang, do not rely on a tailored feed of posts. That said, recent research on platforms that do rely on tailored feeds, such as Facebook, have found that showing people a generic set of posts has relatively little impact along many measures compared to a tailored feed (Nyhan et al., 2023).

Furthermore, much of the content on Baang is typical of other social networks. People mostly share and comment on the news, or tell personal stories. Many users also leverage the audio nature of the platform to recite or sing religious poetry. In this sense, it also aligned with large social media platforms, which have become less text-based with the advent of TikTok.

Research also suggests that social network use is consistent with fundamental human behaviors such as the psychological need for social rewards (Lindström et al., 2021). Thus, there may be behavioral responses to social media functionality that in many cases will be consistent across settings. Therefore, we are not surprised to find that the impact of moderation on platform usage for Baang users is consistent with a related intervention for Facebook and Twitter users (Beknazar-Yuzbashev et al., 2025).

Finally, our conceptual framework further extends the relevance of this study, including to settings with relatively high levels of trust in misinformation (see Section 2). One aspect of our experimental setting that may be less prevalent is that trust in misinformation is relatively low. Of course, this may be explained by our intervention itself, particularly the official information, influencing attitudes towards misinformation rather than innate aspects of the setting. Regardless, our framework addresses settings with high engagement with misinformation by illustrating that the optimal policy in those cases would be high intensity moderation.

[30] See Section 4.2 for more on the usage of the platform.

4 Experiment Design

4.1 Timeline

In April 2020, we made available official COVID-19 posts to all users on the platform. We then conducted the content moderation experiment for two months, from June 27th to August 26th, 2020.[31] In addition, the day the randomized experiment began, we made Baang free to use in order to encourage the user base to grow. The platform was only completely free, however, until July 25th. After that, due to funding limitations, users only had 30 free minutes a day to use the platform, with the potential to gain some additional minutes by forwarding the platform.[32] These adjustments to the cost of the using the platform were the same across all conditions, and thus were orthogonal to treatment.

4.2 Platform usage

During the randomized experiment, the platform generated meaningful engagement, from a total of 3698 users.[33] In total, users called into the platform 116,124 times. In addition to listening to content, the users recorded 13,315 posts and 69,768 comments. This implies that one post was recorded every 6.5 minutes on average, though in busy parts of the day this was much higher. Users further engaged with the platform through 109,844 likes and 96,693 shares. Over this time period, the platform had 583 average daily users, with the mean (median) user spending 23 (6) minutes on the platform per day.

User-generated COVID-19 content was a relatively small part of the total platform, which is perhaps not surprising given the demographics of the Baang user base. During the experiment, users generated 389 COVID-19 related posts and 532 COVID-19 related comments. This is approximately 1.1% of the total content on the platform (2.9% of posts and 0.7% of comments). Still, the total engagement with this content was substantive. Users spent 3886 minutes listening to user-generated COVID-19 posts, which generated 1380 comments, 294 shares, and 2034 likes.

[31] This study was not pre-registered because of the tight timeline of implementation after the beginning of an unprecedented global pandemic. Our first priority was to get useful COVID-19 information on the Baang platform while COVID-19 was rapidly taking hold in Pakistan. The broad trade-off we test in this paper as well as the treatments and outcomes were laid out in funding submissions prior to the start of the experiment, however. Also note we did not select outcome variables for this study but rather we examine the universe of exposure outcomes in our data: time spent on the platform, likes, dislikes, shares, and comments. We did post register the paper: AEA RCT Registry AEARCTR-0009954.

[32] Free minutes accrued across days. In addition, the option to gain additional minutes by forwarding the platform began on July 30th, and the number of minutes was increased on August 13th.

[33] Of those, 43% called in during the pre-experiment period that started in April, and the remainder called in for the first time during the experiment itself.

4.3 Interventions

Before the experiment started, we added the official posts about COVID-19 to the platform. They were introduced with a clarification that they were official posts from Baang and were based on recommendations from local official sources such as the NIH, Pakistan. These seven posts stayed the same for the duration of the experiment and totaled approximately 6.5 minutes of content. The first sentence or two of each post contained its main message so that critical information in the post would reach users who did not listen to the entire post.

Given the very limited number and length of the official posts on the platform, the engagement they generated was substantial. During the experiment, the full experimental sample of 3698 users spent 2717 minutes listening to those seven posts. They also engaged with the official posts through 162 comments, 978 shares, and 489 likes. Overall, 34% of users listened to official posts during the experiment. Taken together, these statistics suggest that had more official posts been made available during the study, the findings reported in this paper could have even been more pronounced.

In contrast to the content in the official posts, which relied on local official sources, the content in the rebuttals largely relied on content from international sources, such as the WHO. This was because a logical starting point for the rebuttals on Baang were existing published rebuttals to COVID-19 related myths at that time, and these were largely put together only by international organizations.[34] Both the official posts and rebuttals were recorded by a single professional voice artist to further help users identify the official content as such.

4.3.1 Treatments

This experiment tested three approaches to addressing misinformation on the platform using two treatments and a control. The two treatments relied on *ex ante* moderation, which means that all user-generated content on the platform was reviewed by a moderator before being made publicly available. Much of the misinformation on the platform could be identified by relying on pre-existing lists of myths created by international public health authorities.[35] In the *never post* treatment, we never posted the content identified as misinformation related to COVID-19. In the *sunshine* treatment, we posted all of the identified misinformation content, but we included a specific rebuttal with each piece of content.[36] These rebuttals played automatically immediately after the misinformation content, and were identified as official responses from the platform. Users did

[34] See Section SA1.1 for further details.

[35] Most examples of misinformation in this setting were largely unambiguous and included folk cures for COVID-19 as well as various conspiracies about it. See Section SA1.1 for additional details about the moderation and the rebuttals.

[36] A meta-analysis of lab experiments in psychology finds that specific rebuttals are more effective than simply denying misinformation in causing people to update their beliefs (Chan et al., 2017).

meaningfully engage with the rebuttals.[37]

These two treatments are compared against a control condition that relied on *ex post* community-based moderation. This approach to moderation is similar to that of many social media platforms, and it was the standard on Baang before the study began. In the control, all user-created content was available immediately as it was posted, but users could tag messages as potential COVID-19 misinformation. These tagged posts were then sent to moderators to remove from the platform if found to be misinformation.[38]

It is important to note that users in all three conditions were exposed to the same content in general. The only two exceptions were that users in the never post treatment were not exposed to COVID-19 misinformation posts, and users in the sunshine treatment were exposed to the official rebuttals. Otherwise, whenever a user posted to the platform, regardless of that user's own condition assignment (never post, sunshine, or control), that post was available immediately to everyone in the control condition. The same post would only become available to users in the two treatment conditions, however, once it was moderated. In addition, we did not announce to treatment users that the content they were exposed to had been *ex ante* moderated. If some users became aware that other users did not receive the announcement, it might have induced them to shift across conditions, threatening the internal validity of the study. Furthermore, both potential mechanisms we propose as explanations for the distaste for moderation observed in this study (see Section 6) do not require users to be aware that they are being *ex ante* moderated.

4.4 Random assignment

We designed our randomization to account for networks of users. Specifically, treatment assignment depended on how a user reached the platform for the first time. *Original* users, who called in directly, were randomly assigned to one of the three conditions when they called into the platform for the first time during the study period. *Referral* users, who called in because they were forwarded content from the platform by another user, were assigned to the same condition as the user who forwarded them content. Regardless of how a user came to the platform initially, once a user was assigned to a condition, they remained in that condition every time they called into the platform thereafter. Users were identified by phone number.[39]

[37] While posts on Baang were skipped in the first 3 seconds on average, rebuttals were skipped after nearly 15 seconds on average.

[38] During the experiment, these posts were already being reviewed, but they were only taken down in the control if identified by the community in order to ensure that typical moderation protocols were maintained. Users flagged 459 posts as misinformation during the study, but none of them were deemed misinformation by the moderators, suggesting the limitations of relying on community moderators in this setting.

[39] For more on this see Section SA1.1.4.

Randomizing referral users into the same treatment as their original user allows us to account for potential spillovers across conditions. Thus, each of the original users and their referral users, if any, form a cluster. Although the clusters only partially captured sharing networks, any spillovers across condition assignment that we were unable to fully capture with this randomization design would generally work against us finding effects.[40] Thus, we do not expect that any cross-treatment sharing is driving our results. Furthermore, we are able to measure any cross-condition spillovers that take place on the platform. Our results are robust to accounting for these spillovers directly in the analysis.[41]

We assigned a latent treatment status to each user as they called in starting in April. Of course, until the experiment began in late June, all users were effectively in the control. Thus, it is useful to differentiate two types of referral users. *Pre-treatment referrals* first used the platform before the experiment began, and thus the referral could not have been endogenous to treatment. *Post-treatment referrals* could conceivably have been selected into a given condition, since their condition was assigned after the study began. Thus, we consider our results on two samples. The sample of original users and pre-treatment referral users had treatment assigned exogenously. The results for the full experiment sample are also an object of interest, however. If one condition is attracting more people or people who listen to more content, that is relevant to understanding the implications of a that condition.

The total experimental sample includes 3698 users. There are 2077 original and pre-treatment referral users. Although is just 56% of the full experimental sample, these users account for most of the platform usage. These users made 91% of the calls, recorded 94% of the posts and made 95% of the comments.[42] Since the condition assignment of an original user determines the assignment of their referral users, this study relies on cluster-level random assignment. The average cluster includes just a few users, and thus there are 1408 clusters in the full sample and 1259 in the sample of original and pre-treatment referral users. As designed, the original users are almost exactly split across treatment conditions, with 367, 366, and 371 users assigned to the never post, sunshine and control conditions respectively.[43]

[40]For example, if control users forward the official COVID-19 posts to users who are in one of the treatment conditions, that would reduce the impact of being assigned to the treatments on exposure to official COVID-19 misinformation.

[41]See Section SA2.2 for more details on spillovers.

[42]They account for 484 average daily users out of a total of 583, with the mean (median) user spending 25 (9) minutes on the platform per day. Forty-six percent of these users listened to official posts during the experiment and an additional 18% before the experiment began.

[43]Due to natural sampling variation, in the full sample, there are 1153, 1258, and 1287 users in the never post, sunshine and control conditions respectively. In the sample of original and pre-treatment referral users, there are 681, 672, and 724 users in the never post, sunshine, and control conditions respectively. In addition, note that total number of original users is 1104, which does not equal the number of clusters (1408) in the study. This is because in some

Finally, we consider the validity of the randomization. First, we note that the randomization was hardcoded into the platform and occurred automatically as users called in for the first time, thus the randomization procedure was unlikely to be vulnerable to implementer-driven threats.[44] Furthermore, the sample size is sufficiently large that it is unlikely to face imbalances due to small sample considerations. Nonetheless, we confirm the validity of our randomization using three approaches (see Section SA2.1 for the details of this analysis).[45] First, we do not find evidence of imbalance for our experimental sample on the date users first joined Baang. Second, we do not find evidence of imbalance on the outcomes of interest for the subsample of experimental users who also used the platform during the pre-treatment period. Furthermore, we confirm that our main results are robust in this subsample. Third, we find that there are no meaningful differences in pre-treatment usage trends, but there are large and significant differences post-treatment.

4.5 Outcome data

The main analysis in this study, and the outcomes in particular, rely on data that is automatically collected in the platform log files as users interact with the platform. The main outcomes in the experiment examine exposure and engagement at the user-level for three sources of information. We particularly focus on the impact of the treatments on the *official* information posts, since these posts were designed to provide high-quality information about COVID-19.

We also examine two sources of user-generated information: useful and misinformation posts. *Useful* posts contained information about COVID-19 that is aligned with the content in the official posts. In some cases, that included personal experiences with COVID-19 that emphasize that it is real. This type of content is aligned with one of the goals of the official content, which was to confirm that COVID-19 was not a hoax. *Misinformation* posts contained false information about COVID-19. User-generated COVID-19 posts were twice as likely to be useful (21%) as opposed to misinformation (8%). Useful information was identified and categorized after the experiment, while misinformation was identified during the experiment through moderation. This categorization was double-checked after the experiment.[46] Most user-generated posts about COVID-19 (71%), however, were neither useful nor misinformation and thus were categorized as *neutral*.

cases an original user called in before the experiment began, referred the platform to someone else, and then never called in during the experiment itself. Thus, there are 304 clusters that do not include an original user.

[44]Of course, coding errors are possible. In addition, the referral users entered the platform through a different process, although it should also be random.

[45]Note that traditional balance tables are not possible in this setting since only a fraction of the experimental sample was exposed to the platform before the study began.

[46]For more on content categorization, see Section SA1.1.

For each of the three types of information, we conduct our analysis for three exposure and engagement measures. The focus of our analysis is on exposure to information, which is measured through minutes spent listening to a given type of post. Since users have a great deal of control on how they spend their time on the platform, this is the key measure of information-seeking behavior that is of interest here.

In addition, we consider two measures of engagement. Increased engagement can induce additional exposure of other users directly, through sharing, or indirectly through increasing the popularity of posts. It can also potentially characterize the intensity of users' exposure. We separately examine one measure of engagement, the number of shares, since it is the primary outcome of interest for researchers focusing on the determinants of the spread of misinformation. We also examine a standardized index of the other measures of engagement: comments, likes, and dislikes. For more on how engagement works on the platform, see Section 3.1.

4.6 Estimation

Our main results are intention-to-treat (ITT) estimates of the impact of the treatment at the user-level for the full study period. We initially present our results graphically, which exploits the time series nature of our data. We focus on the user-level analysis for the main results, however, because it eliminates considerations about attrition or selection over the course of the study. Everyone who is in the experiment, regardless of condition assignment, appears in the dataset for the main results once. Thus, when official information exposure is an outcome variable in our main results, for example, it includes the total amount of exposure across all days that the user called in during the RCT.

Thus, the main estimating equation is given by:

$$Y_i = \beta_1 NeverPost_i + \beta_2 Sunshine_i + \epsilon_c,$$

where $NeverPost_i$ is an indicator for having been assigned to the never post treatment and $Sunshine_i$ is an indicator for having been assigned to the sunshine treatment. Our other main estimates consider the effect of being assigned to either the sunshine treatment or never post treatment using the indicator $Treated_i$. In both cases, we cluster the standard errors at the level of an original user and their referral users, since they are jointly randomized into a given treatment status (see Section 4.4). The mechanism results rely on a non-parametric event study approach, which is discussed in that section.

As noted above, we present our main results for both the sample of original and pre-treatment

referral users and the full experimental sample. For other results, however, we focus on the former sample.

5 Results

The first hypothesis of our framework is that if users have a distaste for moderation, then usage will decline under the higher level of moderation in the two treatments. Thus, we document that both treatments reduce the overall usage of the platform (Figure 1). We begin by examining the extensive margin measure of usage: total number of users per day. On average, the treatments attract 43.1 (19%) fewer daily users relative to the control. Turning to the intensive margin, we find that conditional on calling into the platform, treatment users spend an average of 5.6 (26%) fewer daily minutes on the platform. Thus, not only do fewer users in the treatments call in on any given day than in the control, but conditional on calling in, those users spend less time on the platform. The combined effect on the extensive and intensive margins is that treatment users spend a total of 2120 (42%) fewer minutes on the platform.

We confirm that these results are highly significant using user-level regressions relying on our main specification (Table 2). Focusing on our main experimental sample (Panel A), users in the treatments spend 144 minutes less on the platform over the two months relative to control users who spend an average of 386 minutes on the platform (significant at the 1% level). Similarly, users in the treatments share posts 19 fewer times compared to users in the control who share posts an average of 55 times during the study period (significant at the 10% level). The engagement index is not significant, however, and is close to zero.[47] These results are similar across the two treatments when analyzed separately.

These declines in platform usage as a result of moderation are aligned with estimates from Beknazar-Yuzbashev et al. (2025) and Kalra (2025), the only two papers of which we are aware that have related estimates. In Beknazar-Yuzbashev et al. (2025), the authors find that removing toxic content on Facebook and Twitter led to a 23% decrease in intensive-margin content consumption relative to the mean. This is remarkably similar to our 26% estimate, though in a very different context. There are key differences in our results, however. For example, Beknazar-Yuzbashev et al. (2025) find that users consume less content but spend the same amount of time on the platform. In our context, users are spending less time on the platform as a result of moderation. Kalra (2025) finds that reducing toxic content causes users to view 35% fewer posts.

Finally, note that Figure 1 illustrates that the impacts of being assigned to the treatments are con-

[47] This is because although comments are negative and significant, neither likes nor dislikes are significant. Furthermore, although not significant, dislikes are positive.

centrated in the first half of the experiment. The two reasons for that concentration of effects both suggest that our results are a lower bound on the potential impact of treatment. First, the red line on the figures indicates the point at which we had to increase the cost of spending time on the platform, irrespective of treatment (see Section 4). The large resulting decline in overall usage makes detecting the impact of treatment more difficult. In addition, the official posts remained the same throughout the study, so once study users sampled them, there would likely have been diminishing marginal returns to repeat listens. Adding more official posts throughout the study might have increased the observed impacts of treatment.

5.1 Main results

Our main results are motivated by the second hypothesis of the framework, which posits if users have a distaste for moderation, they will decrease their exposure to official information. In addition to examining exposure to official information, we also measure the impact of treatment on our measures of engagement, since they are potentially important in both influencing and characterizing exposure. The estimates reported below are for the sample of original and pre-treatment referral users, but the results are qualitatively similar for the full experimental sample, as is evident in the referenced tables.

Table 3 provides estimates of the impact of the two treatments on exposure to and engagement with official information posts. Since the results are similar for each of the two treatments, our focus is on their average effect. Users assigned to the treatments listen to 0.33 (25%) fewer minutes of the official posts relative to users assigned the control, a result which is significant at the 5% level. In the control condition, users listen to an average of 1.33 minutes of the official posts, which is somewhat more than one such post. The results on the engagement measures are the expected sign but are marginally insignificant for the main specifications, in which we average the effects across the two treatments. This is unsurprising given that users who engage with posts are typically a small subset of the users who consume posts. Since levels in the control group are much lower for the engagement measures than for exposure, it is more difficult to detect decreases in engagement from a low level on these measures that have a lower bound of zero. They are significant, however, for some individual treatment effects. Thus, these findings are suggestive, but not definitive with regards to the impact of treatment on engagement.[48]

Next, we examine the impact of being assigned to either of the two treatments on exposure to useful

[48] We also consider how treatment affects the relationship between total time spent on the platform and time spent listening to official posts. We find that users spend a somewhat larger percentage of their time listening to official information in the treatments (approximately 1.6%) compared to the control (approximately 0.8%). This does suggest that p_g varies by treatment status, but the difference between p_g under high and low levels of moderation in this case is not large, and thus our empirical results still align with the predictions of the model.

user-generated information.[49] The treatment effects on useful posts are smaller in absolute magnitude but larger in relative magnitude than the effects on official posts (Table 4). Users assigned to either of the two treatments spend 0.14 (38%) less minutes listening to useful posts than in the control, which is significant at the 5% level. In the control, the exposure to useful user-generated content is 0.35 minutes, however, which is lower than for official posts. This highlights the reach of official content on the platform, which is further examined in Section 6 below. As is the case for the results on official posts, the average treatment effects on the engagement measures are at most marginally significant.

Finally, we examine the impact of treatment on exposure to misinformation (Table 5). We conduct this analysis separately by each treatment, since the two treatments had different objectives with regards to addressing misinformation. In the control condition, users are exposed to 0.126 minutes of misinformation on average. As intended, users in the never post treatment are exposed to effectively zero misinformation. Thus, the treatment effect on being assigned to the never post condition is negative 0.122, an 97% decrease relative to the control.[50] In the sunshine treatment, however, we do not see statistically significant differences in exposure to misinformation relative to the control.

In Section SA2, we document that these results are robust to accounting for outliers and spillovers.

5.2 Net information exposure

Motivated by our third hypothesis, we now test whether high moderation has a negative impact on net information exposure, or good minus bad information exposure. For the purposes of this analysis, we weight exposure to good and bad information equally, but we recognize policymakers may choose to weight differently depending on the context.

We begin examining this hypothesis by measuring the average effect of being assigned to one of the treatments on net exposure. Comparing official information to user-generated misinformation, we find that treatment decreases net exposure relative to the control by 25%. This approach to assessing net exposure is directly aligned with our model, which assumes that official information and misinformation comes from different sources. We also measure net exposure including useful posts in the accounting of good information. Using that measure, we find that treatment decreases net exposure relative to the control by 29%. These findings capture that the absolute magnitude of the treatment effect is larger for good information than for bad information. Furthermore, they

[49] Note that useful information is aligned with the good information in our framework, and as discussed in that section it is possible to have multiple sources of good or bad information.

[50] This exposure to misinformation is not exactly zero since a re-examination of all COVID-19 posts on the platform after the experiment was complete identified a small number of misinformation posts that were not identified initially.

confirm that it is largely official information that matters in this setting.

Next, we examine average treatment effects on net exposure separately for the never post and sunshine treatments. In the never post treatment, users are not exposed to any misinformation, which may increase net exposure. In the sunshine treatment, however, users are exposed to the rebuttals, an additional source of good information not available to never post users. When we include only official sources of good information in our measure (including rebuttals), we find that the never post treatment decreased net exposure relative to the control by 15% where the sunshine treatment does so by 25%. This pattern persists, with declines in net exposure of 21% and 29% respectively, if we include useful posts in our measure of good information. It is important to note that net exposure declines in the never post treatment even though we have prohibited all of the misinformation from the platform. That is, the decline in misinformation in the never post treatment is constrained by the total amount of user exposure to misinformation on the platform in the control. In this setting, the absolute decline in exposure to good information is greater than all of the misinformation exposure in the control. A final consideration in weighing the tradeoff across sunshine and never post in a given context is whether the rebuttals are an opportunity to refute misinformation that is circulating outside the platform. We do find evidence that the misinformation on the Baang platform was disseminated through other sources in Pakistan (see Section SA3).[51]

5.2.1 Conditions for negative net exposure

Given we find negative net information exposure from high moderation, we now examine the conditions of the third hypothesis. This hypothesis states that given a distaste for moderation, net exposure will be negative under high moderation if users perceive the sources of good information to be of higher quality than those of the bad information. In our survey, we measure perceptions by asking users about their trust in various information sources (Table 6). A random sample of users almost universally (95%) trust the official Baang posts over users' posts on COVID-19. This is also reflected in their responses to a separate set of questions, in which they are asked to rank their trust in COVID-19 information from different sources on a five-point scale. Their trust in official posts was 3.1 while their trust in users' COVID-19 Baangs was 2.2, a difference that is significant at the 1% level. This range of 1.1 points on the scale also covers a large percentage of total observed range of trust levels in different sources of information. The least trusted source of information is users of other types of social media aside from Baang (1.8), while the most trusted sources are

[51]Note our analysis only considers net exposure on Baang. While we cannot estimate whether users substitute exposure to information on Baang with exposure to information elsewhere, past work suggests an exogenous reduction in social network usage leads to declines in overall knowledge and thus substitution effects are minimal (Allcott et al., 2020).

government announcements (3.8) and doctors (3.8).[52]

Perceptions of different information sources matter, according to the framework, since if users have greater levels of trust in official posts relative to user-generated misinformation posts, then they may be more likely to seek them out and engage with them. This would then lead to negative net information exposure under moderation. So, to complement the survey finding on trust, we directly examine user exposure to and engagement with different types of COVID-19 posts in the control group. Note that this analysis focuses on the intensity of exposure to or engagement with each post, and thus abstracts away from the quantity of different types of posts on the platform. Compared to misinformation posts, which are listened to 11 times on average, official posts are listened to an 653 additional times.[53] The official posts are also shared 62.5 additional times and their engagement index is approximately 13σ greater than misinformation posts. The engagement index is normalized to zero for misinformation posts. Exposure to and engagement with useful information posts is modestly but significantly greater than that of misinformation posts, with users listening to such posts an additional 3.2 times more than the misinformation posts. This pattern is in contrast to studies in other settings where misinformation received more engagement than other types of posts (Vosoughi, Roy and Aral, 2018).[54]

6 Mechanisms

As outlined in the conceptual framework, whether users have a preference for or distaste for moderation is a key determinant of the optimal approach to moderation from the perspective of a social planner. Using the platform more under high moderation indicates a preference for moderation, while using it less indicates a distaste for moderation. To better understand users' preferences over moderation, we consider two implications of high or *ex ante* moderation in this setting.[55] Note that

[52]The Baang posts were aligned with government announcements, but it is perhaps unsurprising that they are less trusted since they are not a direct source. Trust in the official posts were in a similar range with trust in local imams (3.2).

[53]See Table SA1. A listen is defined as ever beginning to listen to a post. There are structural reasons why listens might be higher for user-generated posts. Specifically, they play automatically in a user's feed, while a user would have to actively seek out official posts. In addition, official posts are also a small percentage of the total platform, with just seven total posts. At the same time, however, the official posts are directly accessible throughout the study.

[54]One possible explanation for the low levels of engagement with misinformation in this setting is that the official posts have an inoculation effect that increases skepticism towards misinformation, which could lead users to skip such posts. Roozenbeek et al. (2022) finds that exposure to information on rhetorical techniques has an inoculation effect against misinformation.

[55]Note although we will find evidence for distaste for moderation according to revealed preference through both of mechanisms proposed in this section, in surveys, Baang users do overwhelmingly indicate a preference for moderation. Specifically, 99% prefer that Baangs are moderated and, in a separate question, 100% prefer that they are moderated by the Baang team. The survey, however, does not indicate the details of moderation, thus users are likely considering *ex post* moderation as opposed to no moderation. This is particularly likely given that WhatsApp is a common alternative form of social media in this setting, and it doesn't have any moderation.

these implications do not require that the user is aware that they are being moderated, and they are not in this setting.

First, *ex ante* moderation has implications for whether and how users are exposed to misinformation in this experiment. In the never post treatment, users are not exposed to misinformation, and in the sunshine treatment that exposure is mitigated by rebuttals. On the one hand, users may prefer moderation since they may not enjoy being exposed to misinformation. If that is the case, we may expect users to spend less time on the platform after being exposed to misinformation. On the other hand, users may have a distaste for moderation, in that they enjoy being exposed to misinformation, and they use the platform more after being exposed to it.[56] Alternatively, there may be a behavioral explanation for a distaste for moderation type effect. In particular, exposure to misinformation may induce an emotional reaction (Brady et al., 2017; Lewandowsky and Van Der Linden, 2021; Vosoughi, Roy and Aral, 2018). This could lead to increased usage of the platform after exposure to misinformation if social media use is addictive or correlated with emotional disregulation (Sun and Zhang, 2021; Liu and Ma, 2019).

Second, *ex ante* moderation is required in order for a platform free of misinformation, and this study highlights the cost of that approach for users. Specifically, *ex ante* moderation requires that all user-generated content is posted to the platform with some delay in order to allow time for moderation. In this experiment, the 99% of content that is not about COVID-19 is the same across the three conditions, and the average delay with which that content was posted in the treatments was a relatively modest 67 minutes.[57] It is not surprising that even modest delays may affect usage. Social media usage patterns are consistent with understanding it as a reward system based on likes or engagement with one's own content (Lindström et al., 2021). Furthermore, social media use is correlated with delay discounting, and thus people who prefer instant rewards are more likely to be users (van Endert and Mohr, 2020).[58]

These delays are relevant to moderation in other social media platforms. In general, major platforms rely on *ex post* moderation that only requires reviewing a fraction of the content on plat-

[56]There have been recent high profile examples of both of these types preferences. On the one hand, when Twitter management explicitly stated they would be doing less moderation of misinformation the platform in late 2022, many users left the platform (Sweney, 2023). On the other hand, social networks such as Parler and Truth Social have made a stated lack of moderation their selling point and have become a haven for prominent figures that had posts removed from Twitter due to misinformation (Lima, 2021).

[57]COVID-19-related posts not deemed misinformation had, on average, 270 minute delays to posting. COVID-19-related posts deemed misinformation had, on average, 554 minute delays. These increased delays were caused by our moderation process requiring a supervisor and in many cases a PI to sign off on such decisions.

[58]In practice, all posts go up the control immediately, and thus are equally likely to receive engagement. Users in the treatments, however, will not necessarily be aware of that engagement. In particular, we confirm that 44% users of who experience a delay check the main feed between the time they post and the time the post goes onto the platform. These users could expect to find their own post and be disappointed.

form. In contrast, *ex ante* moderation requires moderating all content. Despite this more limited approach, *ex post* moderation is still significantly delayed on such platforms.[59] This is perhaps due to the costs of moderation, which include thousands of human moderators.[60] Under *ex post* moderation, however, the implication of those delays is that people are exposed to misinformation.

Furthermore, moderating misinformation has unique challenges, and platforms have until recently largely focused on moderating more traditionally regulated toxic speech. New types of misinformation are always arising, and identifying misinformation can require significant additional time, since it typically requires expertise beyond that of standard moderators.[61] AI is likely to have a growing role in moderation, but is also expected to increase the dissemination of misinformation, as well as the challenges in identifying it.[62] Furthermore the challenges of identifying misinformation and relying on AI moderation are even greater in development settings and in languages other than English.[63] Determining what is misinformation will take time and human judgment for the foreseeable future.

We focus on an event study approach to test these potential mechanisms.[64] This approach is appropriate here given the timing of the events are variable, and they can take place throughout the experiment. We also rely on local polynomial regressions given the nonlinear nature of our data. In this analysis, we rely on a conditional parallel trends assumption for identification.[65]

[59] There has been little attempt to comprehensively quantify overall moderation delays on major platforms, but Goldstein et al. (2023) find that average time to post removal on Facebook is approximately 21 hours in late 2020.

[60] Although figures are not reported publicly, according to some reports, Facebook relies on 10,000 to 15,000 human moderators (Barrett, 2020).

[61] Goldstein et al. (2023) finds content removal was significantly delayed after the U.S. capital riot on January 6th, 2021, as Facebook changed its policies to address new types of misinformation content. More generally, Facebook relies on an independent council to make determinations about what types of posted content is misinformation (Oversight Board, 2023).

[62] For a summary of the challenge of moderating misinformation on social media platforms and the role of AI in this process, see Gallo and Cho (2021). For a specific overview of the challenges of scaling content moderation through AI, see Gillespie (2020).

[63] According to documents released by a whistle-blower, the accuracy of a Facebook algorithm in detecting hate speech in the Afghan context was 0.2%. Furthermore, in Arabic, an algorithm falsely flagged innocuous content 77% of the time (Scott, 2021).

[64] In addition, as a robustness check to this mechanisms analysis, we examine whether treatment impact differs by whether users who appear in the pre-treatment data also post during that pre-treatment period (Table SA2). We find that the treatment effects are concentrated among users who post in the pre-treatment period. This approach is a useful robustness check since it relies on experimental identification, although for the subset of users who use the platform before the study begins. Unlike the event studies, however, it does not allow us to isolate the separate impacts of the two mechanisms since users who post before the study are more likely to be exposed to delays in posting after the study begins, but they are also more likely to be exposed to misinformation in the control.

[65] We plot standard error bands to allow for visual examination of parallel trends.

6.1 Exposure to misinformation mechanism

In order to understand users' preferences for misinformation exposure, we examine user behavior after being exposed to misinformation. To do so, we take an event study approach in which the treatment event is exposure to a user-generated *misinformation* post for the first time and the counterfactual or control event is exposure to a comparable *useful or neutral* post. Given there are more non-misinformation posts than misinformation posts related COVID-19 during our study, we selected a matched subsample of non-misinformation posts to serve as counterfactuals identified through a propensity score approach.[66] The matching exercise simply selects the counterfactual group, however. For identification, in this event study framework, we confirm that usage before exposure to either type of COVID-19 post follows the same trend. We conduct this analysis separately for the two conditions in which users are exposed to misinformation, the control and the sunshine treatments, since the rebuttals may impact user behavior.

We find that, in the control, users spend more time on the platform after exposure to misinformation posts relative to similar non-misinformation posts (Figure 2). This indicates a distaste for moderation on the part of users. Notably, we do not find that misinformation has the same effect on usage in the sunshine treatment. This suggests that rebuttals may have a mitigating effect on post-misinformation usage. If users like misinformation, they may dislike the rebuttals and be driven off the platform by them. Alternatively, if the misinformation induces an emotional response, the rebuttals may mitigate that response.[67]

6.2 Delay mechanism

We hypothesize that the delay mechanism is most likely to be observable for users who post, since they have directly experienced their own content being delayed as it is posted to the platform.[68] Thus, in order to test whether this is a mechanism for the observed distaste for moderation, we examine whether the treatment effects are concentrated during the time period after users post for the first time. Specifically, using a non-parametric event study, we examine both overall engage-

[66] This exercise compares subsequent engagement with the platform for two sets of users who came across a COVID-19 related user-generated post for the first time while listening to their feed. The two sets of users are those for which that first post contains misinformation and those for which it contained useful or neutral information. Since there was not a user-specific algorithm ordering feeds in our context, whether the first user-generated COVID-19 post that users come across is a misinformation post or a useful/neutral post is likely to be effectively random. We matched on two post characteristics: date and number of total listens. These are not intended to capture post quality, which we expect to vary across misinformation posts as opposed to other posts. Instead, these are intended to capture behavior on the platform around the time of exposure.

[67] We also compare the difference in the impact of misinformation exposure across the sunshine and control arms and confirm that it is statistically significant (Figure SA2).

[68] This is in contrast to users who spend time on the platform simply listening to content, and who are likely to be less aware of the fact that the content they are listening to is being posted with a modest delay.

ment with the platform and specific engagement with official posts as outcomes of interest. An alternative approach to isolating the delay mechanism could have been to create artificial delays in the control group. We chose not to do so since, as discussed above, *ex ante* moderation and delays cannot be separated in practice, and thus our experiment captures the most policy-relevant trade-off here.

Furthermore, we are able to exploit variation in whether users directly experience moderation by posting to isolate the effect of delays using an event study. Thus, this analysis is limited to the subsample of 722 users who ever post from the sample of original and pre-treatment referral users.[69] We verify that users in both the treatment and control have similar engagement with Baang until their first post.[70]

In Figure 3 Panel A, we find that, after a user's first post, treatment users' exposure to the platform is significantly lower than control users. Usage increases in both the treatment and control immediately after posting, which is consistent with evidence from other platforms (Grinberg et al., 2016). For users who have been assigned to the treatment, however, their usage declines more quickly that for those in the control. Although the confidence intervals are wide immediately after posting, the treatment effect persists from four days after posting to the end of our event study. In Figure 3 Panel B, we find a similar pattern for exposure to official posts specifically, further confirming our result. Finally, we do not find evidence that these results are particularly sensitive to the length of the delay, which is not surprising given users are likely to notice even short delays in their posts reaching the platform.[71]

7 Conclusion

We conduct the first randomized controlled trial with publicly available results in which the two treatments aim to fully control (mis)information across an entire social media platform. In this case, we focus on information specific to COVID-19. We combine this experiment with an intervention that provides access to official posts that contain high-quality information regardless of treatment assignment. We document that a substantial percentage of users seek out these high-

[69] Focusing on this subsample allows us to examine whether pre-trends are parallel, which is a test of the identifying assumption for this approach. The pre-period is not defined for those who never post, since the timing of posting for the first time is variable and defined at the individual level for those who post. Thus, it is less straightforward to test an identifying assumption for the subsample who never post, and we exclude these users from this analysis. We expect that users who post are a selected subsample, and thus, this analysis does not allow us to determine whether those who never post are affected by being assigned to treatment or not.

[70] This is not surprising since users are randomized into treatment groups initially, so if they are largely not treated before they post for the first time, then the determinants of the outcome would be similar in the pre-period for people who post.

[71] See Section SA2.5 for analysis of this question.

quality official posts on the platform (44%). We find that fully controlling access to misinformation through high or *ex ante* moderation reduces usage of the platform. This leads to meaningfully less exposure to the official posts. We also find that the treatments substantially reduce net exposure to information (good minus bad).

The never post treatment in particular is designed to understand the implications of a social media platform that is free of misinformation. Specifically, in the never post treatment, there is no COVID-19 misinformation on the platform. The decline in exposure to official information is greater than the decline in misinformation under treatment, however, even though we have never posted misinformation to the platform.

A conceptual framework helps to contextualize these results. It identifies that users' preference for moderation and their relative trust levels in good as opposed to bad information will determine the optimality of low or high moderation. In this setting, almost all users trust the official posts more than user-generated COVID-19 posts. This may explain why users are much more likely to listen to and engage with official posts compared to user-generated posts on the same topic. Furthermore, higher levels of trust in official posts may lead users to give the official information more weight than that from the misinformation posts. In contrast to this setting, in settings with lower levels of trust in official information and more dissemination of misinformation, *ex ante* moderation is likely to be optimal.

In this experiment, users exhibit a distaste for moderation. Thus, we examine two potential mechanisms for that distaste, and we find evidence for both. First, users engage with the platform more after being exposed to misinformation. So, it is not surprising that they use the platform less after that type of information is prohibited. Second, this experiment highlights that a cost of *ex ante* moderation is modest delays in user-generated content being made available on the platform, and that users dislike delays. These delays are uniquely relevant to moderating misinformation in particular. Given that there are significant delays in the *ex post* moderation of content on major platforms in the status quo, these delays are also relevant to considering the implications of *ex ante* moderation on large platforms.

Our conceptual framework suggests that *ex ante* moderation would be socially optimal for large platforms where misinformation is more prevalent and trusted than official information. The framework also indicates that it is unlikely that it would be individually optimal, however, for profit-maximizing platforms to *ex ante* moderate under those conditions since it would lead to a decline in overall usage. Still, to minimize the potential harms of misinformation, social media companies can take a two-pronged approach. First, platforms can step up efforts to actively disseminate high-quality information from trusted sources, and work to increase trust in reliable information

sources. Second, they can continue to limit the spread of misinformation even if they fall short of ensuring the platform is free of misinformation via *ex ante* moderation.

References

Acharya, Avidit, Matthew Blackwell, and Maya Sen. 2016. "Explaining causal findings without bias: Detecting and assessing direct effects." *American Political Science Review*, 110(3): 512–529.

Alatas, Vivi, Arun G Chandrasekhar, Markus Mobius, Benjamin A Olken, and Cindy Paladines. 2024. "Do Celebrity Endorsements Matter? A Twitter Experiment Promoting Vaccination in Indonesia." *Economic Journal*, 134(659): 913–933.

Allcott, Hunt, and Matthew Gentzkow. 2017. "Social media and fake news in the 2016 election." *Journal of Economic Perspectives*, 31(2): 211–36.

Allcott, Hunt, Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow. 2020. "The welfare effects of social media." *American Economic Review*, 110(3): 629–76.

Banerjee, Abhijit, Eliana La Ferrara, and Victor Orozco. 2019. "Entertainment, Education, and Attitudes Toward Domestic Violence." *AEA Papers and Proceedings*, 109: 133–137.

Barrera, Oscar, Sergei Guriev, Emeric Henry, and Ekaterina Zhuravskaya. 2020. "Facts, alternative facts, and fact checking in times of post-truth politics." *Journal of Public Economics*, 182: 104123.

Barrett, Paul M. 2020. "Who moderates the social media giants." *Center for Business*.

Beknazar-Yuzbashev, George, Rafael Jiménez-Durán, Jesse McCrosky, and Mateusz Stalinski. 2025. "Toxic content and user engagement on social media: Evidence from a field experiment." CESifo Working Paper 11644.

Blattman, Christopher, and Jeannie Annan. 2016. "Can employment reduce lawlessness and rebellion? A field experiment with high-risk men in a fragile state." *American Political Science Review*, 110(1): 1–17.

Bond, Robert M, Christopher J Fariss, Jason J Jones, Adam DI Kramer, Cameron Marlow, Jaime E Settle, and James H Fowler. 2012. "A 61-million-person experiment in social influence and political mobilization." *Nature*, 489(7415): 295–298.

Boyd, Danah M, and Nicole B Ellison. 2007. "Social network sites: Definition, history, and scholarship." *Journal of Computer-Mediated Communication*, 13(1): 210–230.

Brady, William J, Julian A Wills, John T Jost, Joshua A Tucker, and Jay J Van Bavel. 2017. "Emotion shapes the diffusion of moralized content in social networks." *Proceedings of the National Academy of Sciences*, 114(28): 7313–7318.

Breza, Emily, Fatima Cody Stanford, Marcella Alsan, Burak Alsan, Abhijit Banerjee, Arun G Chandrasekhar, Sarah Eichmeyer, Traci Glushko, Paul Goldsmith-Pinkham, Kelly Holland, et al. 2021. "Effects of a large-scale social media advertising campaign on holiday travel and COVID-19 infections: a cluster randomized controlled trial." *Nature Medicine*, 27(9): 1622–1628.

Bursztyn, Leonardo, Georgy Egorov, Ruben Enikolopov, and Maria Petrova. 2020. "Social media and xenophobia: Evidence from Russia." National Bureau of Economic Research Working Paper 26567.

CDC. 2018. *Crisis Emergency Risk Communication Manual.*

Central Intelligence Agency. 2024. CIA World Factbook, accessed May 2024.

Chan, Kelvin. 2023. "EU investigates X over potential violations of social media law." *Associated Press*. Dec 18, 2023.

Chan, Man-pui Sally, Christopher R Jones, Kathleen Hall Jamieson, and Dolores Albarracín.

2017. "Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation." *Psychological Science*, 28(11): 1531–1546.

Curry, David. 2023. "Reddit Revenue and Usage Statistics (2023)." *Business of Apps*. January 9, 2023.

Dupas, Pascaline, and Edward Miguel. 2017. "Impacts and determinants of health levels in low-income countries." In *Handbook of Economic Field Experiments*. Vol. 2, 3–93.

Dwoskin, Elizabeth. 2020. "Facebook launches one-stop shop portal for coronavirus information." *Washington Post*. March, 18th, 2020.

Enikolopov, Ruben, Alexey Makarin, and Maria Petrova. 2020. "Social media and protest participation: Evidence from Russia." *Econometrica*, 88(4): 1479–1514.

EU. 2024. "The Digital Services Act." https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en.

Fujiwara, Thomas, Karsten Müller, and Carlo Schwarz. 2024. "The Effect of Social Media on Elections: Evidence from the United States." *Journal of the European Economic Association*, 22(3): 1495–1539.

Fung, Brian, and Devin Cole. 2023. "Biden administration defends communications with social media companies in high-stakes court fight." *CNN*. August, 10th, 2023.

Gallo, Jason A, and Clare Y Cho. 2021. "Social Media: Misinformation and content moderation issues for Congress." *Congressional Research Service Report*, 46662.

Gillespie, Tarleton. 2020. "Content moderation, AI, and the question of scale." *Big Data & Society*, 7(2): 2053951720943234.

Goldstein, Ian, Laura Edelson, Damon McCoy, and Tobias Lauinger. 2023. "Understanding the (in) effectiveness of content moderation: A case study of facebook in the context of the us capitol riot." *arXiv preprint arXiv:2301.02737*.

González-Bailón, Sandra, David Lazer, Pablo Barberá, Meiqing Zhang, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Deen Freelon, Matthew Gentzkow, Andrew M Guess, et al. 2023. "Asymmetric ideological segregation in exposure to political news on Facebook." *Science*, 381(6656): 392–398.

Grinberg, Nir, P Alex Dow, Lada A Adamic, and Mor Naaman. 2016. "Changes in engagement before and after posting to facebook." *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 564–574.

GSMA. 2021. "Addressing the Mobile Gender Gap in Pakistan."

Guriev, Sergei, Emeric Henry, Théo Marquis, and Ekaterina Zhuravskaya. 2023. "Curtailing false news, amplifying truth." Working paper available at SSRN.

Hafeez, Erum, and Tayyaba Fasih. 2018. "Growing Population of Pakistani Youth: A Ticking Time Bomb or a Demographic Dividend." *Journal of Education and Educational Development*, 5(2): 211–226.

Henry, Emeric, Ekaterina Zhuravskaya, and Sergei Guriev. 2022. "Checking and Sharing Alt-Facts." *American Economic Journal: Economic Policy*, 14(3): 55–86.

Hsu, Tiffany. 2023. "Falsehoods Follow Close Behind This Summer's Natural Disasters." *New York Times*. August, 30th, 2023.

Idrees, Muhammad, Ayesha Ch, Babak Mahmood, Rao Sabir Sattar, and Faiza Anjum. 2023. "Youth Bulge and Social Unrest: A Pessimistic Image of Demographics Dividend." *Social Evolution and History*.

Jiménez-Durán, Rafael. 2022. "The economics of content moderation: Theory and experimental

evidence from hate speech on Twitter." George J. Stigler Center, University of Chicago Booth School of Business Working Paper 324.

Kalra, Aarushi. 2025. "Hate in the Time of Algorithms: Evidence on Online Behavior from a Large-Scale Experiment." *arXiv preprint arXiv:2503.06244*.

Kremer, Michael, and Rachel Glennerster. 2011. "Improving health in developing countries: evidence from randomized evaluations." In *Handbook of Health Economics*. Vol. 2, 201–315.

Levy, Ro'ee. 2021. "Social media, news consumption, and polarization: Evidence from a field experiment." *American Economic Review*, 111(3): 831–70.

Lewandowsky, Stephan, and Sander Van Der Linden. 2021. "Countering misinformation and fake news through inoculation and prebunking." *European Review of Social Psychology*, 32(2): 348–384.

Lima, Cristiano. 2021. "Gettr, Parler, Gab find a fanbase with Brazil's far-right." *Washington Post*. November 9, 2021.

Lindström, Björn, Martin Bellander, David T Schultner, Allen Chang, Philippe N Tobler, and David M Amodio. 2021. "A computational reward learning account of social media engagement." *Nature Communications*, 12(1): 1311.

Lin, Xialing, Patric R. Spence, Timothy L. Sellnow, and Kenneth A. Lachlan. 2016. "Crisis communication, learning and responding: Best practices in social media." *Computers in Human Behavior*, 65: 601–605.

List, John A. 2020. "Non est Disputandum de Generalizability? A Glimpse into The External Validity Trial." National Bureau of Economic Research Working Paper 27535.

Liu, Chang, and Jian-Ling Ma. 2019. "Adult attachment style, emotion regulation, and social networking sites addiction." *Frontiers in Psychology*, 10: 2352.

Marathe, Megh, Jacki O'Neill, Paromita Pain, and William Thies. 2015. "Revisiting CGNet Swara and its impact in rural India." *Proceedings of the Seventh International Conference on Information and Communication Technologies and Development*, 1–10.

Moitra, Aparna, Vishnupriya Das, Gram Vaani, Archna Kumar, and Aaditeshwar Seth. 2016. "Design lessons from creating a mobile-based community media platform in Rural India." *Proceedings of the Eighth International Conference on Information and Communication Technologies and Development*, 1–11.

Mosquera, Roberto, Mofioluwasademi Odunowo, Trent McNamara, Xiongfei Guo, and Ragan Petrie. 2020. "The economic effects of Facebook." *Experimental Economics*, 23(2): 575–602.

Nyhan, Brendan, Jaime Settle, Emily Thorson, Magdalena Wojcieszak, Pablo Barberá, Annie Y Chen, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, et al. 2023. "Like-minded sources on Facebook are prevalent but not polarizing." *Nature*, 620(7972): 137–144.

Oversight Board. 2023. "Oversight Board." http://www.oversightboard.com, accessed August 2023.

Patel, Neil, Deepti Chittamuru, Anupam Jain, Paresh Dave, and Tapan S Parikh. 2010. "Avaaj otalo: a field study of an interactive voice forum for small farmers in rural india." *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 733–742.

Pennycook, Gordon, Adam Bear, Evan T Collins, and David G Rand. 2020. "The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings." *Management Science*, 66(11): 4944–4957.

Raza, Agha Ali, Bilal Saleem, Shan Randhawa, Zain Tariq, Awais Athar, Umar Saif, and Roni Rosenfeld. 2018. "Baang: A viral speech-based social platform for under-connected populations." *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–12.

Raza, Agha Ali, Mustafa Naseem, Namoos Hayat Qasmi, Shan Randhawa, Fizzah Malik, Behzad Taimur, Sacha St-Onge Ahmad, Sarojini Hirshleifer, Arman Rezaee, and Aditya Vashistha. 2022. "Fostering Engagement of Underserved Communities with Credible Health Information on Social Media." *Proceedings of the ACM Web Conference 2022*, 3718–3727.

Roozenbeek, Jon, Sander van der Linden, Beth Goldberg, Steve Rathje, and Stephan Lewandowsky. 2022. "Psychological inoculation improves resilience against misinformation on social media." *Science Advances*, 8(34): eabo6254.

Schwab, K, R Samans, S Zahidi, et al. 2016. "The Global Gender Gap Report 2016: World Economic Forum."

Scott, Mark. 2021. "Facebook did little to moderate posts in the world's most violent countries." *Politico*. October, 25th, 2021.

Semrush. 2023. "Top Websites." http://www.semrush.com, accessed August 2023.

Sherwani, Jahanzeb, Nosheen Ali, Sarwat Mirza, Anjum Fatma, Yousuf Memon, Mehtab Karim, Rahul Tongia, and Roni Rosenfeld. 2007. "Healthline: Speech-based access to health information by low-literate users." 1–9, IEEE.

Singer, Philipp, Fabian Flöck, Clemens Meinhart, Elias Zeitfogel, and Markus Strohmaier. 2014. "Evolution of reddit: from the front page of the internet to a self-referential community?" *Proceedings of the 23rd international conference on world wide web*, 517–522.

Stiglitz, Joseph E. 2000. "The contributions of the economics of information to twentieth century economics." *Quarterly Journal of Economics*, 115(4): 1441–1478.

Sun, Yalin, and Yan Zhang. 2021. "A review of theories and models applied in studies of social media addiction and implications for future research." *Addictive Behaviors*, 114: 106699.

Sweney, Mark. 2023. "Twitter 'to lose 32m users in two years after Elon Musk takeover'." *The Guardian*. December 13, 2022.

Tursunbayeva, Aizhan, Massimo Franco, and Claudia Pagliari. 2017. "Use of social media for e-Government in the public health sector: A systematic review of published studies." *Government Information Quarterly*, 34(2): 270–282.

Twitter. 2020. "COVID-19 tab in Explore." https://blog.twitter.com/en_us/topics/company/2020/covid-19#explore.

UNESCO. 2023. "Online disinformation: UNESCO unveils action plan to regulate social media platforms." https://www.unesco.org/en/articles/online-disinformation-unesco-unveils-action-plan-regulate-social-media-platforms.

Urdal, Henrik, and Kristian Hoelscher. 2009. "Urban youth bulges and social disorder: An empirical study of Asian and sub-Saharan African cities." *World Bank Policy Research Working Paper*, , (5110).

van Endert, Tim Schulz, and Peter NC Mohr. 2020. "Likes and impulsivity: Investigating the relationship between actual smartphone use and delay discounting." *PloS One*, 15(11): e0241383.

Vosoughi, Soroush, Deb Roy, and Sinan Aral. 2018. "The spread of true and false news online." *Science*, 359(6380): 1146–1151.

Wang, Yuxi, Martin McKee, Aleksandra Torbica, and David Stuckler. 2019. "Systematic literature review on the spread of health-related misinformation on social media." *Social Science*

& Medicine, 240: 112552.

We Are Social, and Meltwater. 2024. "Digital 2024 Global Overview Report." https://datareportal.com/reports/digital-2024-global-overview-report, accessed September 2024.

White, Jerome, Mayuri Duggirala, Krishna Kummamuru, and Saurabh Srivastava. 2012. "Designing a voice-based employment exchange for rural India." 367–373.

WHO. 2020. "Speeches of the Director General: Munich Security Conference." February, 15th.

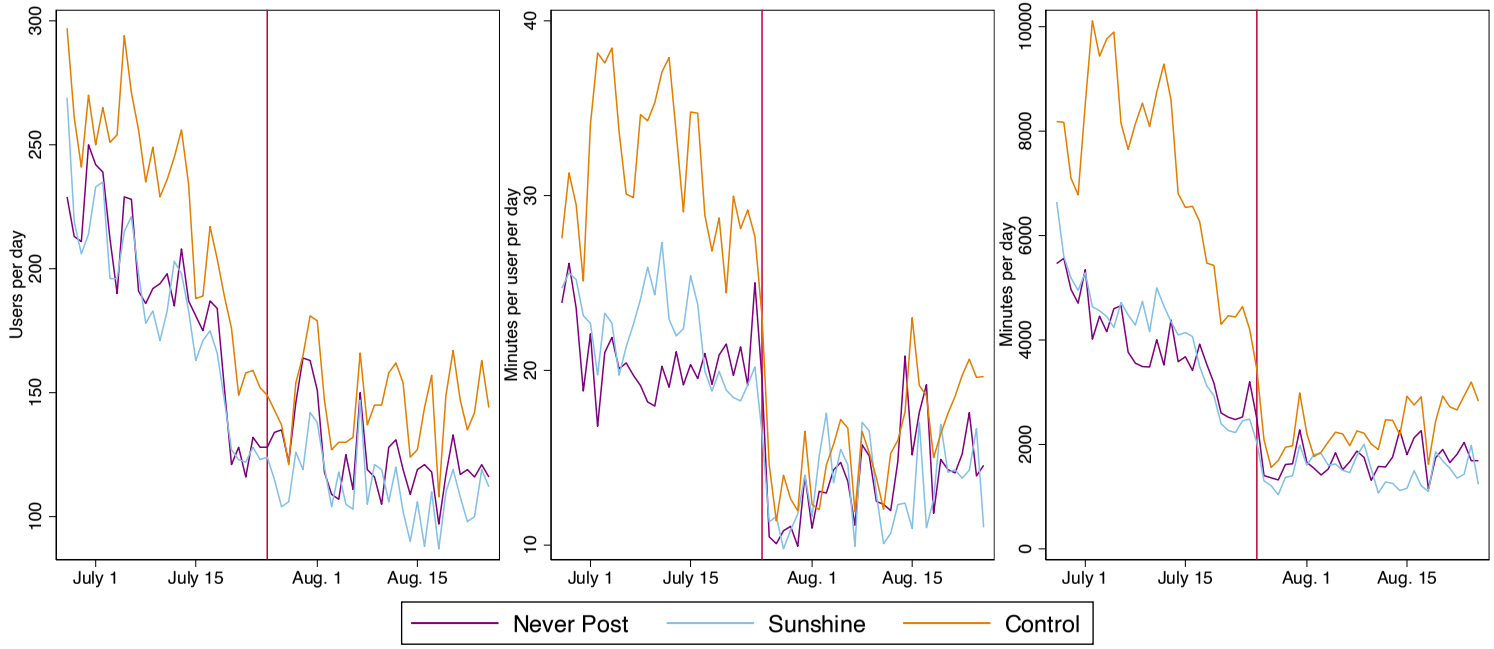
WHO. 2021. "*Global Health Observatory*." Dataset accessed January 2021. https://www.who.int/data/gho/data/indicators/indicator-details/GHO/hospital-beds-(per-10-000-population).

World Bank. 2006. *World Development Report 2007: Development and the Next Generation.* Washington, D.C.:World Bank.

World Bank. 2024. World Bank Open Data Portal, accessed May 2024.

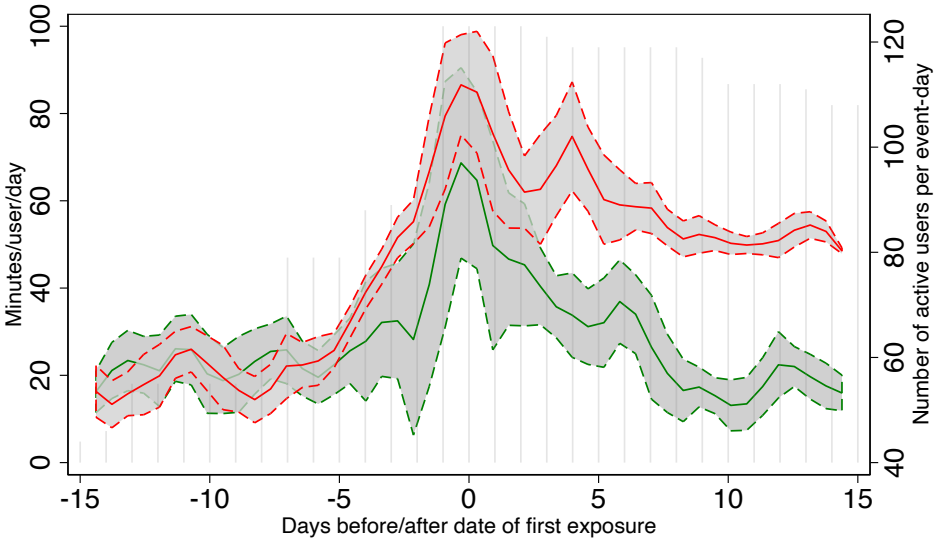
Zhuravskaya, Ekaterina, Maria Petrova, and Ruben Enikolopov. 2020. "Political effects of the internet and social media." *Annual Review of Economics*, 12: 415–438.

Figure 1: User exposure to the platform during the experiment

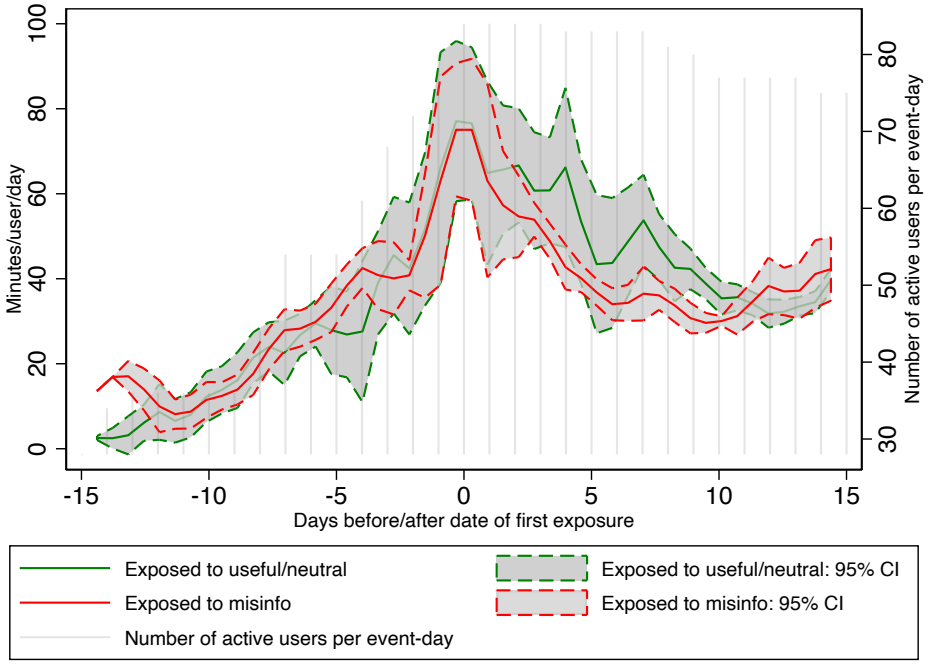


Notes: The red vertical line represents the date that free access to the platform was limited for each user. In panel A: *Users per day* captures whether a user called in on that day and listened to at least one second of a post or comment. *Minutes per user per day* is the average minutes spent on the platform by users per day. *Minutes per day* is the total minutes spent on the platform across all users per day. In panel B: *Users/day exposed to official posts* captures whether a user listened at all to an official information post on that day. *Mins/user/day listened to official posts* is the average minutes that a user spent listening to official posts on that day conditional on listening at all. *Mins/day listened to official posts* is the total minutes spent listening to official posts across all users per day. Data includes 29,053 user-day observations.

Figure 2: First misinformation exposure event study



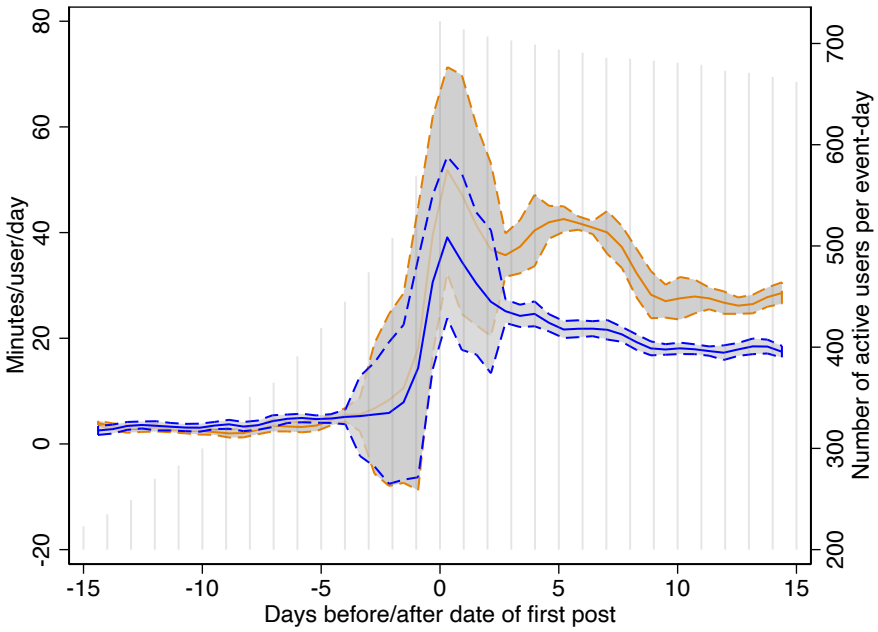
(a) Control users only



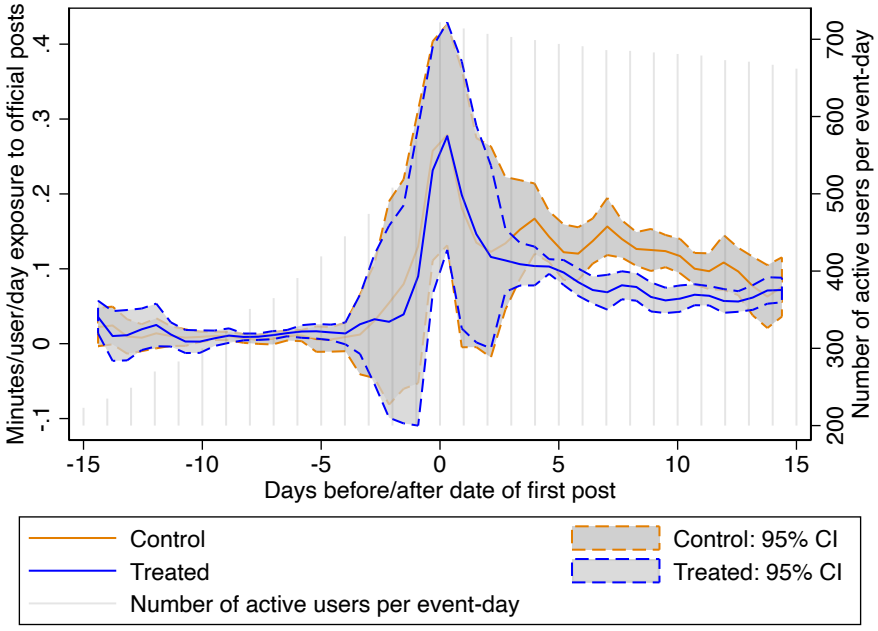
(b) Sunshine users only

Notes: This figure demonstrates an event study in which the treatment event at day zero is a user being exposed to a user-generated misinformation post for the first time. The matched counterfactual event is being exposed to a COVID-19-related but not misinformation post, selected to equal the number of misinformation posts via propensity score matching. The outcome measure in both panels is total minutes spent on the platform by user-day. Panel A considers only control users and Panel B considers only sunshine users. Lines are local polynomial regressions with an epanechnikov kernel with bandwidth 1. Data includes 12,420 user-day observations.

Figure 3: First post event study



(a) Overall exposure to the platform



(b) Exposure to official information posts

Notes: This figure demonstrates an event study in which the event at day zero is a user posting for the first time. The outcome measure in Panel A is total minutes spent on the platform by user-event-day. Panel B is minutes listened to an official information post on that user-day. The sample is limited to original users. Lines are local polynomial regressions with an epanechnikov kernel with bandwidth 1. Data includes 43,320 user-day observations.

Table 1: User characteristics

	Sample Means		
	Random	Most active	Exposed to misinformation
User characteristics			
Age	29.61	29.97	30.77
	(6.30)	(4.85)	(4.72)
Female (=1)	0.01	0.03	0.03
	(0.10)	(0.18)	(0.18)
Less than 8 years of education (=1)	0.19	0.10	0.17
	(0.40)	(0.31)	(0.38)
Less than 10 years of education (=1)	0.47	0.54	0.62
	(0.50)	(0.50)	(0.49)
Has a smartphone (=1)	0.91	0.83	0.80
	(0.28)	(0.38)	(0.40)
Uses WhatsApp at least once a day or more (=1)	0.91	0.78	0.76
	(0.28)	(0.42)	(0.43)
Observations	94	87	86

Notes: *Random* is a random sample of users who ever called into the platform. The *Most active* sample are the users that comment the most times. The *Exposed to misinformation* sample are the users most exposed to misinformation. Standard deviations are reported in parentheses.

Table 2: Exposure and engagement outcomes for overall platform usage

	Minutes listened	Number of shares	Engagement index
Panel A: Original and pre-treatment referral users only			
Treatments combined			
Treated (=1)	-144.20***	-19.26*	-0.02
	(49.42)	(9.96)	(0.04)
Treatments separated			
Remove (=1)	-145.40***	-18.97*	-0.01
	(52.09)	(10.14)	(0.04)
Sunshine (=1)	-142.98***	-19.55*	-0.04
	(53.40)	(11.72)	(0.04)
Remove = Sunshine?	0.95	0.95	0.48
Control mean	386.31	55.80	-0.00
# Clusters	1259	1259	1259
# Users	2077	2077	2077
Panel B: All post users			
Treatments combined			
Treated (=1)	-89.63***	-12.73**	-0.02
	(28.54)	(5.68)	(0.03)
Treatments separated			
Remove (=1)	-83.44***	-11.12*	0.00
	(29.85)	(5.86)	(0.03)
Sunshine (=1)	-95.31***	-14.20**	-0.04
	(32.27)	(6.69)	(0.03)
Remove = Sunshine?	0.64	0.58	0.26
Control mean	234.72	34.44	-0.00
# Clusters	1408	1408	1408
# Users	3698	3698	3698

Notes: *$p < 0.1$, **$p < 0.05$, ***$p < 0.01$. The unit of observation is the user. All outcome measures sum usage across the entire platform (not only for COVID-19-relevant content). *Treated* is an indicator for being assigned to either the sunshine or remove treatments. *Total engagements/user* is a z-score average of comments, likes, and dislikes, with each engagement normalized relative to the mean and standard deviation of control users' posts. Reports OLS regressions with standard errors clustered at the level of an original user and their referral users in parentheses. P-values are reported for the test that rejects Remove = Sunshine.

Table 3: Main exposure and engagement outcomes for *official* information posts

	Minutes listened	Number of shares	Engagement index
Panel A: Original and pre-treatment referral users only			
Treatments combined			
Treated (=1)	-0.331**	-0.246	-0.043
	(0.168)	(0.156)	(0.028)
Treatments separated			
Never Post (=1)	-0.305*	-0.292*	-0.025
	(0.181)	(0.167)	(0.032)
Sunshine (=1)	-0.357*	-0.199	-0.062**
	(0.188)	(0.177)	(0.031)
Never Post = Sunshine?	0.731	0.518	0.177
Control mean	1.335	0.605	-0.000
# Clusters	1259	1259	1259
# Users	2077	2077	2077
Panel B: All post users			
Treatments combined			
Treated (=1)	-0.228**	-0.140	-0.053**
	(0.097)	(0.089)	(0.022)
Treatments separated			
Never Post (=1)	-0.224**	-0.164*	-0.034
	(0.109)	(0.095)	(0.025)
Sunshine (=1)	-0.231**	-0.118	-0.071***
	(0.105)	(0.102)	(0.024)
Never Post = Sunshine?	0.942	0.587	0.081
Control mean	0.883	0.356	-0.000
# Clusters	1408	1408	1408
# Users	3698	3698	3698

Notes: $*p < 0.1$, $**p < 0.05$, $***p < 0.01$. The unit of observation is the user. All outcome measures focus on official information posts about COVID-19. *Treated* is an indicator for being assigned to either the sunshine or never post treatments. *Engagement index* is a z-score average of comments, likes, and dislikes, with each engagement normalized relative to the mean and standard deviation of control users' posts. Reports OLS regressions with standard errors clustered at the level of an original user and their referral users in parentheses. P-values are reported for the test that rejects Never Post = Sunshine.

Table 4: Main exposure and engagement outcomes for *useful* information posts

	Minutes listened	Number of shares	Engagement index
Panel A: Original and pre-treatment referral users only			
Treatments combined			
Treated (=1)	-0.143**	-0.010	-0.022
	(0.064)	(0.007)	(0.031)
Treatments separated			
Never Post (=1)	-0.139**	-0.006	-0.006
	(0.067)	(0.008)	(0.036)
Sunshine (=1)	-0.146**	-0.015**	-0.038
	(0.068)	(0.007)	(0.035)
Never Post = Sunshine?	0.859	0.186	0.328
Control mean	0.350	0.022	-0.000
# Clusters	1259	1259	1259
# Users	2077	2077	2077
Panel B: All post users			
Treatments combined			
Treated (=1)	-0.084**	-0.009*	-0.023
	(0.037)	(0.005)	(0.023)
Treatments separated			
Never Post (=1)	-0.073*	-0.004	-0.002
	(0.039)	(0.006)	(0.028)
Sunshine (=1)	-0.093**	-0.013***	-0.043*
	(0.040)	(0.005)	(0.025)
Never Post = Sunshine?	0.446	0.045	0.115
Control mean	0.208	0.017	-0.000
# Clusters	1408	1408	1408
# Users	3698	3698	3698

Notes: *$p < 0.1$, **$p < 0.05$, ***$p < 0.01$. The unit of observation is the user. Useful information posts are user-generated. All outcome measures focus on posts about COVID-19. *Treated* is an indicator for being assigned to either the sunshine or never post treatments. *Engagement index* is a z-score average of comments, likes, and dislikes, with each engagement normalized relative to the mean and standard deviation of control users' posts. Reports OLS regressions with standard errors clustered at the level of an original user and their referral users in parentheses. P-values are reported for the test that rejects Never Post = Sunshine.

Table 5: Main exposure and engagement outcomes for *misinformation* posts

	Minutes listened	Number of shares	Engagement index
Panel A: Original and pre-treatment referral users only			
Treatments combined			
Treated (=1)	-0.027	0.001	-0.086***
	(0.027)	(0.002)	(0.028)
Treatments separated			
Never Post (=1)	-0.122***	-0.001	-0.115***
	(0.020)	(0.001)	(0.027)
Sunshine (=1)	0.068	0.003	-0.057*
	(0.043)	(0.003)	(0.034)
Never Post = Sunshine?	0.000	0.082	0.007
Control mean	0.126	0.001	-0.000
# Clusters	1259	1259	1259
# Users	2077	2077	2077
Panel B: All post users			
Treatments combined			
Treated (=1)	-0.012	0.000	-0.068***
	(0.016)	(0.001)	(0.021)
Treatments separated			
Never Post (=1)	-0.069***	-0.001	-0.089***
	(0.011)	(0.001)	(0.020)
Sunshine (=1)	0.040	0.002	-0.049*
	(0.026)	(0.002)	(0.025)
Never Post = Sunshine?	0.000	0.088	0.013
Control mean	0.071	0.001	0.000
# Clusters	1408	1408	1408
# Users	3698	3698	3698

Notes: *$p < 0.1$, **$p < 0.05$, ***$p < 0.01$. The unit of observation is the user. Misinformation posts are user-generated. All outcome measures focus on posts about COVID-19. *Treated* is an indicator for being assigned to either the sunshine or never post treatments. *Engagement index* is a z-score average of comments, likes, and dislikes, with each engagement normalized relative to the mean and standard deviation of control users' posts. Reports OLS regressions with standard errors clustered at the level of an original user and their referral users in parentheses. P-values are reported for the test that rejects Never Post = Sunshine.

Table 6: User attitudes

	Sample Means		
	Random	Most active	Exposed to misinformation
Perceptions of Baang content			
Trusts official more than users' COVID-19 posts (=1)	0.95	0.85	0.94
	(0.23)	(0.36)	(0.25)
Trust in official COVID-19 posts (1-5)	3.08	2.96	3.24
	(0.81)	(0.81)	(0.77)
Trust in users' COVID-19 posts (1-5)	2.23	1.92	2.01
	(0.91)	(0.80)	(0.78)
Prefers Baangs are moderated (=1)	0.99	1.00	1.00
	(0.10)	(0.00)	(0.00)
Prefers Baang team moderates (as opposed to users)	1.00	1.00	1.00
	(0.00)	(0.00)	(0.00)
Trust in sources of COVID-19 information (1-5)			
Government announcements	3.83	3.57	3.65
	(0.82)	(0.64)	(0.78)
Doctor	3.83	4.02	3.93
	(0.81)	(0.71)	(0.72)
Friends & family	3.64	3.75	3.87
	(0.65)	(0.55)	(0.66)
Local imam	3.09	3.16	3.23
	(0.80)	(0.73)	(0.81)
Social media	1.79	1.69	1.73
	(0.71)	(0.70)	(0.64)
Observations	94	87	86

Notes: *Random* is a random sample of users who ever called into the platform. The *Most active* sample are the users that comment the most times. The *Exposed to misinformation* sample are the users most exposed to misinformation. Standard deviations are reported in parentheses.

Highlights for

"The spread of (mis)information: A social media experiment in Pakistan"

by Sarojini Hirshleifer, Mustafa Naseem, Agha Ali Raza, Arman Rezaee

- First RCT to control misinformation across a social media platform, focusing on COVID-19
- Higher moderation substantially reduces overall platform usage, indicating a distaste for moderation
- Also reduces exposure to official COVID-19 information on the platform by more than it reduces exposure to misinformation
- A framework shows this is driven by official information being more trusted than misinformation in this setting
- In contrast, higher moderation is likely to be optimal where misinformation is more trusted than official information

Disclosure Statements for

"The spread of (mis)information: A social media experiment in Pakistan"

by Sarojini Hirshleifer, Mustafa Naseem, Agha Ali Raza, Arman Rezaee

This paper benefited from National Institutes of Health allowing some funds to be reallocated to this project from NIH grant 5R21HD095696-02, which was largely intended for a different project. All co-authors received direct support from that grant.

This project received IRB approval from Lahore University of Management Sciences (03032021AAA-FWA-00019408) and University of California, Davis (1717776-1).

Sarojini Hirshleifer declares that they have no additional relevant or material financial interests that relate to the research described in this paper.

Mustafa Naseem declares that they have no additional relevant or material financial interests that relate to the research described in this paper.

Agha Ali Raza is the co-founder of Baang. This is an unpaid role in a non-registered entity. As a co-founder of Baang, he has unlimited access to Baang data. Aside from this role, he has no additional relevant or material financial interests that relate to the research described in this paper.

Arman Rezaee declares that they have no additional relevant or material financial interests that relate to the research described in this paper.