ABSTRACT

Title of Dissertation:	PLATFORM DESIGN STRATEGIES AND IMPLICATIONS FOR USER BEHAVIORS
	Maya Mudambi
	Doctor of Philosophy in Business and Management
	2023
Dissertation directed by:	Siva Viswanathan Dean's Professor of Information Systems and Digital Innovation R.H. Smith School Decision, Operations, and Information Technologies Department

This work examines how the design, features, and moderation policies of online platforms impact user behavior in myriad ways and have significant externalities on society at large. The first two studies examine the effectiveness of different content moderation policies adopted by user-generated content platforms to address issues related to misinformation and verbal aggression, respectively. The third study examines how the design of financial incentive structures affects the behaviors of users on a crowdsourcing platform. The studies produce theoretical implications regarding human behavior on online platforms, from the spreading of misinformation to interpersonal verbal aggression, to the behavioral response to monetary rewards. I

additionally make recommendations for practitioners regarding optimal platform design and policies.

PLATFORM DESIGN STRATEGIES AND IMPLICATIONS FOR USER BEHAVIORS

by

Maya Mudambi

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2023

Advisory Committee:
Professor Siva Viswanathan, Chair
Professor Erich Battistin
Assistant Professor Jessica Clark
Professor Gordon Gao
Assistant Professor Lauren Rhue

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

The design, features, and moderation policies of online platforms impact user behavior in myriad ways and have significant externalities on society at large. Problems such as misinformation, verbal aggression, and harassment plague user-generated content (UGC) platforms. UGC platforms have adopted various strategies to combat these negative behaviors, two of which are examined within the body of work described in this dissertation. Additionally, online platforms can also provide an outlet for firms from all industries, researchers, and organizations to seek optimal solutions to problems using crowdsourcing. Crowdsourcing platforms' ability to provide solutions is a function of their design and incentive structure, which I examine in this work.

Beyond examining the effect of design and governance policies on general user behavior, this proposal examines the effect of said entities on particularly impactful groups of users. Both UGC and crowdsourcing platforms disproportionately rely on the contributions of a small group of high-activity users (Lakhani & von Hippel 2003; Nielsen 2006; Brandtzæg & Heim 2011; Hayashi et al. 2021). In Chapter 2 and 3 I focus on the impact of content moderation policies upon problematic users that produce misinformation and verbal aggression, respectively. In Chapter 4 I focus specifically on how platform features affect high-activity, high-performance users.

The work proceeds as follows: in Chapter 2, I examine the impact of a UGC platform's content moderation policy – prominence reduction – upon users' misinformation-spreading behavior. This is significant, as misinformation often has dire real-world consequences. Specifically, I study Reddit's quarantine policy, a prominence reduction policy that reduces the visibility of misinformation on the platform. I empirically assess the effectiveness, as well as spillover effects, of quarantine. I find that quarantine diminishes misinformation contributions in directly impacted groups, but causes a misinformation spillover in other topically related groups. This spillover comes solely from the most problematic users and dies out due to a lack of contagion to other users. Additionally, prominence reduction does not decrease non-problematic user contribution. My research sheds light on the unique nature of misinformation spread on online platforms, and shows that prominence reduction policies show promise for controlling misinformation.

In Chapter 3, I extend the previous essay to directly compare Reddit's prominence reduction policy with a banning policy when applied to a group high in verbal aggression. This study aims to directly compare the downstream effects of these policies. Additionally, I assess the differential effects of content moderation on multihoming versus non-multihoming users. I find that both prominence reduction and banning policies produce verbal aggression spillovers in outside groups. However, banning policies produce a wider spillover. This wider spillover is solely driven by multihoming users, as multihoming has a disinhibiting effect. Finally, I find evidence that these verbal aggression spillovers are contagious to other users. These findings have important implications, as they show that prominence reduction

policies produce harmful negative spillover effects, and that the efficacy of group banning policies is limited.

Finally, in Chapter 4 I study a specific design feature on crowdsourcing platforms: monetary rewards for high performance within contests. Specifically, the research focuses on crowdsourcing 'superstars' (high contribution and ability users), a small group of users with a disproportionately large influence on other users and solution outcomes (Tauchert et al. 2020; Hayashi et al. 2021). Utilizing a novel regression discontinuity design, I find that winning money significantly increases subsequent user performance. This performance premium arises from subsequent changes in team composition, rather than an increase in effort or contest selection.

This dissertation provides many contributions to the literature regarding the impact of platform design choices on human behavior. Regarding the spread of misinformation online, I show in Chapter Two that prominence reduction policies have the unintended effect of producing misinformation spillovers, but are ultimately successful due to a lack of contagion that produces a dampening effect. In Chapter Three, I find that both prominence reduction and banning policies applied to a group high in verbal aggression also produce spillovers. However, these are more harmful because they produce contagion. Multihoming users drive a wider spillover. Finally, I show in Chapter Four that monetary incentives are effective in increasing user performance.

This dissertation contributes to the theoretical understanding of negative behaviors such as misinformation spreading and verbal aggression online platforms. My research builds on prior work by examining the effectiveness of two specific content moderation policies: prominence reduction and banning. My work makes multiple novel theoretical contributions. I contribute to the literature on platform design, specifically content moderation policies, and policies to promote user contribution. I also contribute to the literature on informational cascades by examining how these policies impact the spread of problematic behavior.

Furthermore, this work builds upon theories of extrinsic motivators (Hannan et al. 2008; Murayama et al. 2010; Tran et al. 2012; Lourenco 2016). This is examined within the context of crowdsourcing. Specifically, the work in my third essay isolates the impact of providing a monetary incentive on subsequent user behavior. It is able to tease out the effect of financial incentives alone, even though other incentives such as recognition and learning are present.

The findings present within this body of work have important contributions for practitioners as well. The results presented in Chapter Two lead organically into specific recommendations for UGC platforms regarding controlling the spread of misinformation with prominence reduction. Chapter Three provides evidence for the relative efficacy of two content moderation strategies against verbal aggression that would be valuable for UGC platform designers. Finally, the work proposed in Chapter Four provides specific recommendations to crowdsourcing stakeholders on how to optimize incentive structures for their highest-performing and most valuable users.

Chapter 2: Fighting Misinformation on Social Media: An Empirical Investigation of the Impact of Prominence Reduction Policies

2.1 Introduction

Firms can incur serious financial losses and unwanted regulatory attention because of misinformation (Cusumano et al. 2021), as well as their policies towards it (Reuters 2022). As such, misinformation has attracted the attention of both researchers and practitioners in operations management (Cezar et al. 2020; Hossain et al. 2022; Konstantakis et al. 2022; Lee et al. 2018). Many recent studies focus on how misinformation can disrupt firm operations (Hossain et al. 2022; Konstantakis et al. 2022). Other studies assess how firms can best respond to it (Cezar et al. 2020), deploy it (Lee et al. 2018), or detect it (Chung Ng et al. 2023; Zhang et al. 2022a) to further their strategic goals.

In today's media landscape, the vast majority of misinformation is circulated on online platforms. Indeed, online misinformation has been shown to spread via informational cascades, at a rate higher than the truth (Vosoughi et al. 2018). This has serious implications for societal welfare as well as platform success. Accordingly, platforms have experimented with various content moderation policies and design strategies aimed at halting its spread. The focus of this study is an empirical assessment of platform prominence reduction policies applied to misinformation. Prominence reduction is an approach in which the visibility of misinformation is decreased, the goal being to decrease its spread and impact on users.

Many operations management studies investigate how, in the face of informational cascades, online platforms can manipulate the visibility of information to increase metrics such as product adoptions, consumer purchases, or crowdfunding donations (Küçükgül et al. 2022; Maglaras et al. 2022; Xiao et al. 2021; Zhou & Chen 2016). However, there are fewer papers examining how platforms can utilize the same strategies to minimize unwanted user behaviors, which can also spread through a network. This paper complements the work of Candogan & Drakopoulos (2020), who model how platforms can optimally deploy informational labels to attenuate the informational cascade process in the case of misinformation. Instead of examining the policy of labeling misinformation, however, we study prominence reduction, which strategically hides misinformation rather than labeling it.

The efficacy and efficiency of platform content moderation policies have been studied across many fields. Despite the richness of this literature stream, there is a lack of consensus regarding the most effective approach to minimize misinformation on an online platform. Though deleting harmful content has been shown to mitigate its spread (Haaretz 2020), it must be done at a speed that may not be possible for large platforms to implement (Bak-Coleman et al. 2021; Papanastiasou 2020). Labeling misinformation has been shown to be detrimental to user engagement (Candogan & Drakopoulos 2020), as well as ineffective in halting its spread (Lees et al. 2021; Papakyriakopoulos & Goodmann 2022; Saltz et al. 2021; Zanettou et al. 2021). We complement and extend this previous work by assessing the efficacy of prominence reduction on misinformation, which is not well understood.

Prevailing revenue models and market forces, incentivize online platforms to maximize user contribution and utility. Indeed, there is a rich operations literature stream examining the economic value of user-generated content for firms (Chen et al. 2016; Duan et al. 2008; Lau et al. 2018). Many studies focus on what drives platform adoption and content contribution (Jin et al. 2022; Wang et al. 2019; Wei et al. 2021; Zeng et al. 2022; Zhou & Chen 2016). This is notable, as platform content moderation policies run the risk of decreasing user engagement and potentially alienating core users (Candogan & Drakopoulos 2020; Newell et al. 2016). We address this concern by assessing how prominence reduction not only impacts misinformation contribution but overall user contribution.

Motivated by these research gaps, as well as by managerial and societal import, we ask and answer the following questions: (1) How effective are prominence reduction policies in decreasing misinformation on an online platform? (2) How do prominence reduction policies heterogeneously impact different types of users? (3) How do prominence reduction policies affect overall platform contribution?

We answer these questions by empirically assessing the prominence reduction policy of 'quarantine' on the discussion forum platform Reddit. The goal of this policy is to reduce the visibility of harmful content and to prevent it from spreading to other users (Reddit Administration 2018). Reddit's quarantine policy reduces the visibility of problematic forums by removing links to them from site navigation and hiding them from search results. In essence, to access a quarantined forum, a user must have a direct link to it. Reddit has quarantined problematic forums in recent years for various reasons such as hate speech, illegal content, and misinformation.

We find that the prominence reduction policy employed by Reddit significantly diminishes misinformation in quarantined forums. However, after prominence reduction is applied, misinformation activity spills over to other topically related forums. This misinformation spillover dies down after a relatively short period, as misinformation is not found to be contagious within these spaces. We also find that this spillover is entirely produced by the most problematic users within the quarantined group. Finally, prominence reduction has no impact on users' non-misinformation contribution.

Our findings contribute to three literature streams addressing vital questions for online platforms. Many operations management studies assess how online platforms can best manipulate the visibility and accessibility of information to achieve a desired outcome (Küçükgül et al. 2022; Maglaras et al. 2022; Xiao et al. 2021; Zhou & Chen 2016). This work contributes to this by assessing how decreasing the visibility of misinformation via prominence reduction policies ultimately impacts its spread while factoring in the presence of informational cascades. Given the disruptive nature of misinformation, prior researchers study how firms can optimally respond to, detect, and even deploy misinformation (Cezar et al. 2020; Lee et al. 2018; Zhang et al. 2022a). This work contributes to this stream by empirically assessing how online platforms can best reduce the impact of misinformation. It builds upon previous work in detecting misinformation by utilizing a novel approach combining manual coding and automated artificial intelligence methods. Due to the economic value of user-generated content (UGC), many studies assess how platforms can boost user contribution (Jin et al. 2022; Wei et al. 2021; Zeng et al. 2022). We

contribute to this by assessing the impact of prominence reduction on non-problematic user contribution, as well as how it impacts the highest contributing users.

This work produces managerial insights for online platforms. These are significant as the presence of misinformation poses financial and regulatory risks for platforms (Cusumano et al. 2021). The results indicate that prominence reduction policies can be effective in containing misinformation on an online platform. However, they run the risk of producing misinformation spillovers. The duration and intensity of these spillovers are contingent upon newly exposed individuals' propensity to adopt misinformation. To address this concern, platforms should monitor users in problematic groups with a history of misinformation, and potentially apply additional sanctions. Additionally, these results indicate that these policies are not detrimental to overall user contribution. Overall, this study indicates that hiding misinformation on an online platform is an effective method that can be deployed with limited risk.

The remainder of this paper proceeds as follows. We summarize related literature, then describe the research context, data collection and variables, and the empirical approach. Finally, we provide results, discuss theoretical and operational implications, and provide limitations and potential extensions.

2.2 Related Literature & Research Gaps

This study is related to three streams of research: misinformation, informational cascades, and online platform design and policies. Work within these domains is highly interdisciplinary; we primarily draw from papers in operations management, management science, information systems, and economics, with additional insights from communications and sociology. Below we summarize significant prior studies and highlight the research gaps addressed.

2.2.1 Misinformation

The literature on the economic value of information is extensive and spans a variety of topics. Researchers have studied how to optimally provide information to consumers to increase product adoption (Zhou & Chen 2016), prevent customer churn (Allon et al. 2011; Jouini et al. 2011), and increase consumer surplus (Papanastiasou et al. 2018). Studies also show how consumers use different types of information to make purchase decisions (Davis et al. 2021; Ghosh et al. 2022).

This has extended to work studying the operational effects of false information, or misinformation. Significant work looks at the impact of consumers providing false information on firm strategy (Cezar et al. 2020), how fake news disrupts supply chains (Hossain et al. 2022; Konstantakis et al. 2022), and what incentivizes media firms to produce fake news (Hausken 2020). Regarding disinformation (misinformation released for the express purpose of misleading the public), researchers find evidence for its dissemination on online platforms to further business goals (Lee et al. 2018), as well as study how to detect it using machine learning (Chung Ng et al. 2023; Zhang et al. 2022a).

Online misinformation often arises from a small number of problematic users (Grinberg et al. 2019; Peres Nobre et al. 2022; Shao et al. 2018). Users are more likely to share misinformation if it has a high amount of engagement (Avram et al. 2020) or if they expect a positive response from others (Acemoglu et al. 2021). Misinformation is often closely associated with a given ideology, and individuals often prefer seeking out stances on misinformation that align with their own ideology (van der Linden et al. 2020; Zhou & Shen 2022). It is difficult to shift individuals' misinformation stances, as being presented with the opposing viewpoint often has the effect of making individuals argue more vigorously for their views (Lodge & Taber 2013). This phenomenon is known as a disconfirmation bias.

This study builds upon prior work within the misinformation research stream and addresses significant research gaps. Many studies establish the operational risks posed by misinformation (Cezar et al. 2020; Hossain et al. 2022; Konstantakis et al. 2022; Lee et al. 2018) and investigate how firms respond strategically. We address a research gap within this stream by studying how online platforms attempt to decrease the volume of misinformation itself. Online misinformation can often be sourced back to a small group of problematic users (Grinberg et al. 2019; Peres Nobre et al. 2022; Shao et al. 2018). We build upon this by examining how content moderation policies differentially impact problematic users and assess how misinformation spreads within a community where polarized views are present.

2.2.2 Informational Cascades

An informational cascade is a phenomenon in which individuals make decisions primarily based on the decisions of others. This concept has its origins in economics (Banerjee et al. 1992; Bikhchandani et al. 1992) and is also known as social learning or herding behavior. Many studies within operations management study the role informational cascades play in determining consumer behavior. Given consumers' tendency to consult others' reviews, which then influence their purchases, researchers model how firms can optimally design products and set prices (Crapis et al. 2017; Feldman et al. 2019).

Another research stream studies how firms can directly manipulate the information cascade process by optimally releasing information. Firms can selectively release information within a social network to increase product adoption (Zhou & Chen 2016). The content and visibility of information on an online platform can be manipulated to stimulate purchases (Küçükgül et al. 2022; Maglaras et al. 2022). The magnitude of the information cascade effect on online platforms can be moderated by both the frequency and topicality of informational messages (Xiao et al. 2021). Candogan & Drakopoulos (2020) model how user-generated content (UGC) platforms can attenuate the informational cascade process (in the case of misinformation spread) by selectively providing labels on misinformation. This study extends this work by empirically examining how manipulating misinformation visibility influences its propagation on a UGC platform (given the presence of informational cascades).

Informational cascades have been found to be subject to propagation via threshold models. Within this framework, individuals are influenced by others to adopt a behavior or information if a given number or proportion of others within their

social network do so. Thus, the likelihood of an unexposed individual adopting increases monotonically with their number of direct exposures to it (Granovetter 1978; Granovetter & Soong 1983; Hodas & Lerman 2014; Pastor-Satorras & Vespignani 2001). Informational cascades can halt if the requisite number of adopters is not met; the minimum number of adopters to sustain the process is referred to as the critical mass (Zuhair Al-Taie & Kadry 2017).

This study addresses multiple research gaps within the informational cascades literature stream. Many operations management studies investigate how, in the face of informational cascades, online platforms can manipulate the visibility of information to increase metrics such as product adoptions, consumer purchases, or crowdfunding donations (Küçükgül et al. 2022; Maglaras et al. 2022; Xiao et al. 2021; Zhou & Chen 2016). However, there are fewer papers examining how platforms can utilize the same strategies to minimize unwanted user behaviors, which can also spread through a network. This paper complements the work of Candogan & Drakopoulos (2020), who model how platforms can optimally deploy informational labels to attenuate the informational cascade process in the case of misinformation. Instead of examining the policy of labeling misinformation, however, we study prominence reduction, which strategically hides misinformation rather than annotating it. Furthermore, the empirical analyses provide a direct assessment of the efficacy of extant platform policies.

2.2.3 Online Platform Design & Policies

There has been significant debate within the operations management literature as to how to design platforms to address two competing goals that are important to managers. First, the presence of harmful user-generated content (UGC) on these platforms poses significant reputational and regulatory risks and incentivizes platforms to remove it or decrease its impact. Most popular platforms have adopted various content moderation policies with the goal of decreasing the prevalence of misinformation and other harmful UGC. Second, prevailing revenue models, as well as market forces, incentivize online platforms to maximize user contribution and utility. These two platform goals are potentially in opposition as content moderation policies run the risk of decreasing user contribution and potentially alienating core users, but letting harmful content remain can scare away advertisers and investors. Below we detail significant studies that address these goals and identify core research gaps.

2.2.3.1 Content Moderation Policies

When harmful content is widely circulated, platforms are penalized, be it from a loss of advertising revenues or from reduced user loyalty. To address this managerial concern, the operations literature has examined the efficacy of content moderation policies. The goal of these policies is to either remove or mitigate harmful content circulating on platforms. To effectively curb the spread of misinformation, some research suggests that it should be detected utilizing crowdsourcing methods and removed before it hits a critical number of shares (Papanastiasou 2020). This is supported by empirical studies: deleting misinformation from a platform can be effective (Haaretz 2020), but it needs to be removed from a platform as quickly as

within 30 minutes to halt its spread (Bak-Coleman et al. 2021). Nonetheless, using crowdsourcing methods to identify harmful content may be a flawed approach: such decentralized methods have been shown to be subject to in-group and out-group biases (Kwan et al. 2023). Large-scale algorithmic means of identifying misinformation show promise. For example, Chung Ng et al. (2023) find that fake news can be accurately identified by platforms using an augmented-AI approach.

The efficacy and efficiency of various content moderation policies aimed at curbing online misinformation have been studied across many fields, such as operations management, information systems, communications, and computer science. Researchers have examined the efficacy of utilizing fact-checked labels (Candogan & Drakopoulos 2020; Lees et al. 2021; Papakyriakopoulos & Goodmann 2022; Sharevski et al. 2022), deleting misinformation UGC (Bak-Coleman et al. 2021; Haaretz 2020), and banning problematic users (Chang & Danescu-Niculescu-Mizil 2019; Zhang et al. 2022b).

When examining policies to optimally detect and label misinformation (rather than deleting it), Candogan & Drakopoulos (2020) conclude that all posts above a set certainty threshold should be labeled. They find that this design is at odds with a design aimed at maximizing user engagement. Nevertheless, studies in other fields suggest that there is little evidence to suggest that labels significantly reduce misinformation spread. Research has shown that labels have a limited effect on engagement or spread (Lees et al. 2021; Papakyriakopoulos & Goodmann 2022). Furthermore, some researchers suggest that these labels have the unintended consequence of increasing engagement with harmful content (Zannettou 2021), and that some users view them as biased and punitive (Saltz et al. 2021).

Despite the richness of this literature stream, there is a lack of consensus regarding the most effective content moderation policies. Prior research has underlined the need for effective content moderation but has not identified the most effective policy. This motivates a systematic examination of the potential of a prominence reduction policy. With prominence reduction, since content is not labeled or deleted, but rather hidden, this can be seen as a compromise approach. Yet, the efficacy of prominence reduction for minimizing the harmful effects of misinformation has not been directly addressed in the literature. We address this core research gap empirically in this study.

2.2.3.2 User Contribution

There is a rich operations literature stream examining the economic value of user-generated content (UGC) for firms (Chen et al. 2016; Duan et al. 2008; Lau et al. 2018). Consequently, many studies focus on what drives platform adoption and UGC contribution. Platform adoption can be maximized by targeting and selecting the ideal members within a social network (Zhou & Chen 2016). The size of an online social network is positively correlated with contribution frequency (Burke et al. 2009; Zhang & Zhu 2011), quality (Wang et al. 2019), and the presence of social nudges (Zeng et al. 2022). Social network structure also drives contribution frequency and quality (Jin et al. 2022; Wei et al. 2021).

Online platforms disproportionately rely on the contributions of a small number of zealous individuals (Chen et al. 2016), referred to in this study as

'superusers'. This has been verified on a variety of platforms (Adar et al. 2000; Qiu & Kumar 2017). This is generalized by the so-called 90-9-1 rule: on online platforms, 90% of users do not contribute, 9% contribute rarely, and 1% contribute the majority (Carron-Arthur et al. 2014; Nielsen Norman Group 2006). High-contributing users have also been observed to be highly influential (Hayashi et al. 2021).

While much of this literature focuses on how to boost user contribution, there is a lack of consensus as to how content moderation policies affect this key metric. We address this by assessing how prominence reduction impacts overall non-misinformation contribution. As the contribution of superusers is critical to maintaining a successful platform, we address an additional research gap by assessing how content moderation policies impact the behavior of superusers.

2.3 Research Context

2.3.1 Anti-Vaccine Misinformation

To dive deeper into misinformation contribution behaviors, it is helpful to focus on a particular misinformation context. Our research context is anti-vaccine misinformation. Online platforms have been highly polarized regarding vaccines, and research has examined multiple aspects of the vaccination debate (Schmidt et al. 2018). The core position of the anti-vaccine ('anti-vax') movement alleges that large pharmaceutical companies and governments are covering up the dangers of vaccines for sinister reasons (Jolley et al. 2014; Smith & Graham 2019). Examples of myths associated with this movement include that there is a link between vaccines and conditions such as autism, sudden infant death syndrome (SIDS), and cancer; that toxic elements are present in vaccines; and that diseases such as measles are harmless.

The proliferation of the anti-vax movement in recent decades has been tied to changes in vaccination rates, as well as outbreaks of preventable diseases (Yiannakoulias et al. 2019). The MMR vaccination rate in the UK dropped from 92 percent to 84 percent from 1996 to 2002; correspondingly there were measles outbreaks in the country (Hussain et al. 2018). Vaccine-related media has been shown to have a significant effect on vaccination rates in many studies (Kata 2012).

Though online misinformation spans many different topics, we chose the context of anti-vaccine discourse because the ground truth is well-established. Most anti-vaccine myths are definitively contradicted by scientific and medical scholarship. The scientific consensus is that vaccines are typically well-tolerated, contribute dramatically to declines in morbidity and mortality from preventable diseases, and provide benefits that outweigh their risks (Kata 2012). This makes misinformation within our dataset easier to identify as compared to misinformation surrounding political events. The ground truth in those cases often falls along political lines, and there is less evidence explicitly debunking false claims (Corner 2017). Furthermore, many online platforms have taken action against anti-vaccine misinformation (New York Times 2021), so these results are relevant beyond the specific study context.

2.3.2 Reddit

Reddit.com is a massive social news aggregator platform with more than 57 million daily active users (New York Times 2023). The platform is comprised of thousands

of topic-specific discussion forums, which its users create. Reddit has a primarily de-centralized moderation structure: each forum is governed by its own rules, which are enforced by volunteer user moderators. Reddit is also governed by employee administrators, who manage site operations and can override decisions made by these user moderators.

Reddit forums have a nested replies functionality, in which users can directly respond to comments. Figure 1 shows the structure of a typical Reddit conversation thread and defines terms utilized in the paper. The first post that initiates a conversation thread is a 'conversation thread starter'. As replies are iteratively replied to, this forms a 'conversation branch' within the thread. Each thread can contain multiple conversation branches. Replies are denoted by their level number (i.e., distance from the conversation thread starter within the branch).

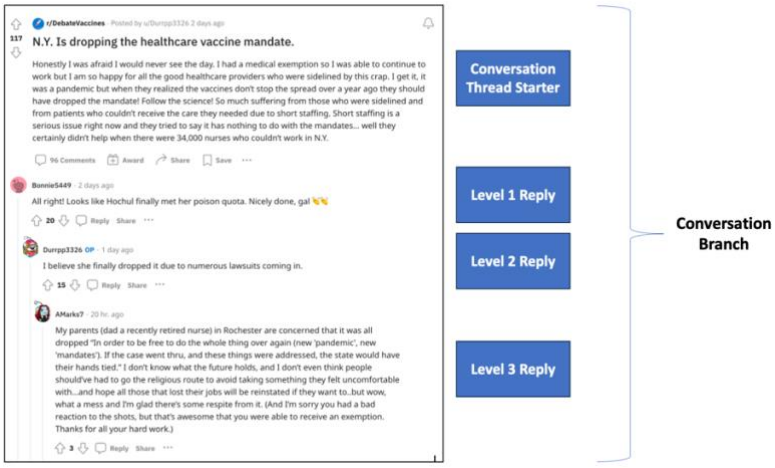


Figure 1. Reddit Conversation Thread Structure

2.3.2.1 Reddit Quarantine

In September 2018, Reddit ramped up its existing prominence reduction policy: forum quarantine (Newsweek 2018). When a forum has been observed to contain a high proportion of problematic content, Reddit intervenes with a quarantine. Quarantined forums do not show up in Reddit or Google searches, do not generate ad revenue, and require users to see a warning and explicitly opt-in to view them. Quarantined forums' content will not appear in a user's newsfeed unless they were a previous subscriber. Examples of problematic forums quarantined include forums dedicated to violence, hate speech, illegal activities, and misinformation. Reddit implemented this policy with the goal of reducing the visibility of harmful content and preventing it from spreading to other users (Reddit Administration 2018).

Reddit has multiple forums related to vaccines that promote various differing ideologies. These ideologies are made explicit and can be determined by consulting forum descriptions as well as forum rules. Reddit "anti-vax" forums are devoted to discussing the danger of vaccines and circulating conspiracy theories. In addition, Reddit has opposing forums dedicated to ridiculing the anti-vax movement and promoting the usage of vaccines, which we characterize as "pro-vax" forums. An example of a pro-vax forum's description is below in Figure 2; it is made clear that

the purpose of the forum is to mock anti-vaxxers. Finally, some forums are devoted to the discussion of vaccines from a neutral standpoint and allow both pro- and anti-vax content, denoted here as "neutral-vax" forums.

Figure 2. Description of Pro-Vax Forum, r/vaxxhappened

In the spring of 2019, Reddit quarantined five anti-vax forums, specifically with the goal of preventing the further spread of misinformation. We examine the impact of this quarantine policy on users in anti-, pro-, and neutral-vax forums, collectively called the "vaccine network".

2.4 Data

2.4.1 Data Collection

Reddit implemented a quarantine of anti-vax forums on May 23, 2019. As in previous literature (Habib et al. 2022; Habib & Nithyanand 2022; Phadke et al. 2022), we obtain the identity of these five quarantined forums, as well as the date of quarantine, utilizing the Reddit forum r/reclassified[1]. We identify all other forums dedicated to vaccines by searching the Reddit platform for forums using a set of vaccine-related keywords. Forums with zero subscribers are discarded, as well as those whose primary focus was not vaccines. We utilize the forum descriptions as well as forum rules to confirm that each forum's primary focus was vaccine discourse, and to classify each forum as anti-vax, neutral-vax, or pro-vax. The final list of forums includes 5 quarantined anti-vax forums, 1 non-quarantined anti-vax forum, 4 pro-vax forums, and 2 neutral-vax forums. For the primary analyses, we scrape all Reddit posts within these forums ranging from February 28, 2019, to August 15, 2019[2]. A survey of Reddit policy changes during the period of interest showed no other changes relevant to the study context. Posts produced by known bots are discarded from the dataset. The restricted dataset contains 307,976 posts.

To analyze the effect of quarantine upon posting behavior, we identify all users (44,746) who posted in the vaccine network during the baseline period prior to

[1] This forum lists a crowd-sourced subset of all quarantine events, their dates, and the reasoning.

[2] To conduct long-term neutral-vax analyses, we collected posts from those forums until 3/1/2020. Data collection was halted at this date due to the COVID-19 pandemic. This totaled 30,023 additional posts.

quarantine. We collect their posts on the entire platform during the study period (2/28/19 - 8/15/19), as well as their account features.

2.4.2 Variable Construction

User-level variables based on posting metrics, as well as user account features, are constructed to analyze the effect of quarantine on user behavior. All variables and their definitions are available in Table 1. We first identify users who posted in the quarantined forums during the baseline period pre-quarantine (2/28/19 - 5/22/19). We create a time-invariant dummy variable *Quarantine Exposed*$_i$. This takes the value of 1 the user posted in a quarantined anti-vax forum during the baseline period, and 0 if not. This yields 545 users with *Quarantine Exposed*$_i$ equal to 1; we refer to them as "quarantine-exposed users". The quarantine-exposed group is the treatment group within this study. We refer to other users with *Quarantine Exposed*$_i$ equal to zero (44,201) as 'control' users.

The control variables are a set of user account features and baseline activity. The dependent variables are each user's number of misinformation posts in the follow-up period, as well as the number of non-misinformation posts on the entire platform. Three variables that describe types of users are described as well: *Superuser*$_i$ (the highest activity users), *Problematic User*$_i$ (users most likely to spread misinformation), and *Neutral-Vax Native*$_i$ (control users who posted in neutral-vax forums at baseline).

Table 1. User-Level Variables

Variable Type	**Variable**	**Definition**
Dependent Variables	Misinformation Posts$_i$	Number of misinformation posts (posts classified as 'agrees with misinformation') in the vaccine network
	Non-Misinformation Platform Posts$_i$	Number of posts on the Reddit platform minus misinformation posts in the vaccine network
Independent Variable	Quarantine Exposed$_i$	Takes the value of 1 if user *i* posted in one of the 5 quarantined anti-vax forums during the pre-quarantine baseline period (2/28-5/22/19); 0 otherwise
User Account Features	Is Moderator$_i$	Takes the value of 1 if user *i* is a moderator for a forum on Reddit; 0 otherwise
	Has Premium Account$_i$	Takes the value of 1 if user *i* pays for a premium account on Reddit; 0 otherwise
	Verified Email$_i$	Takes the value of 1 if user *i* has verified the email associated with their Reddit account; 0 otherwise
	Tenure$_i$	Number of days between user *i*'s first post on Reddit and May 23, 2019

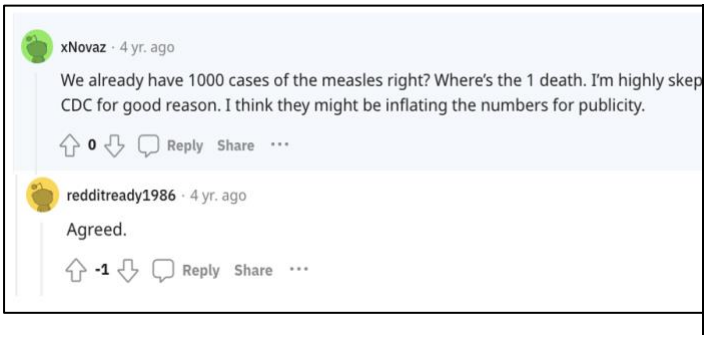
Baseline Activity	Vaccine Network Misinformation Posts$_i$	Number of misinformation posts user *i* made within vaccine-related forums during the pre-quarantine baseline period
	Vaccine Network Posts$_i$	Number of posts user *i* made within vaccine-related forums during the pre-quarantine baseline period
	Vaccine Network Forums$_i$	Number of vaccine-related forums user *i* posted in during the pre-quarantine baseline period
	Outside Network Posts$_i$	Number of posts user *i* made within non-vaccine-related forums during the pre-quarantine baseline period
	Outside Network Forums$_i$	Number of non-vaccine-related forums user *i* posted in during the pre-quarantine baseline period
Dependent Variables	Vaccine Network Misinformation Posts$_i$	Number of misinformation posts user *i* made within vaccine-related forums during the follow-up period
User Types	Superuser$_i$	Takes the value of 1 if user *i* was in the top 1% of contributors (38 posts) in vaccine network forums during the pre-quarantine baseline period; 0 otherwise
	Problematic User$_i$	Takes the value of 1 if user *i* had 1 or more misinformation posts in vaccine network forums during the pre-quarantine baseline period; 0 otherwise
	Neutral-Vax Native$_i$	Takes the value of 1 if user *i* had 1 or more posts in neutral-vax forums AND has *QuarantineExposed*$_i$ equal to 0; 0 otherwise

2.4.3 Operationalizing Misinformation

To get a sense of the types of misinformation present in the dataset, we take a random sample of 1,083 posts and cross-reference them to multiple academic reviews (Ayers et al. 2021; Guidry et al. 2020; Kata 2010; Kata 2012; Offitt & Hackett 2003) of anti-vaccine online discourse. Posts that directly mention a myth or trope present in literature are flagged. We then assess which myths are most common in the dataset. We then determine the scientific consensus regarding those myths by performing a literature review. The myths that are most common in the dataset, are mentioned in the anti-vaccine discourse literature, and are clearly in opposition to the scientific consensus are selected to be representative of misinformation within the data. This yields twelve separate myths. Appendix 1 details the misinformation typology, literature sources, and which myths were ultimately selected.

The goal is to capture the full scope of harmful misinformation discourse on Reddit. Due to Reddit's nested replies functionality, discourse can extend well beyond an initial post affirming a myth. The stance of these replies towards

misinformation can be impossible to accurately assess in isolation. Below is an example in Figure 3. To fully capture harmful misinformation, it is important to collect not only posts that can be judged to clearly agree with misinformation in isolation, but also other posts that explicitly affirm misinformation through agreement or amplification of others' posts. This is a complex task that differs from and builds on prior literature. Previously, misinformation was classified by looking at the spread of false news articles (Chung Ng et al. 2023; Grinberg et al. 2019; Shao et al. 2018; Vosoughi et al. 2018) or through text analysis of content in isolation (Peres Nobre et al. 2022). Past approaches provide a solid foundation but are incomplete.



Post explicitly affirming the myth "Measles is a harmless childhood disease."

Reply: explicit agreement with myth, but could not be assessed in isolation

Figure 3. Misinformation Assessed Through Agreement

To capture a broader scope of misinformation, we first ascertain where misinformation discourse begins within Reddit threads. We start by compiling a set of relevant keywords associated with each of the twelve vaccination myths. These keywords, as well as their scientific basis, are available in Appendix 2. We take all conversation thread starter posts and level one replies - referred to henceforth as 'parent posts' - with a misinformation keyword present in them (43,484), and manually code them as 'agrees with misinformation', 'neutral to misinformation', 'disagrees with misinformation', or 'unrelated to misinformation'. Posts with a label of 'agrees with misinformation' are counted as misinformation. This step is performed manually by human coders due to the conditional nature of agreement in Reddit conversations. It is critical to have a high degree of confidence in the stance and relevance of the initial posts in a conversation. We validate these labels utilizing the ChatGPT[3] API and obtained an accuracy of 90.7%. The validation process for these manual labels is detailed in Appendix 3.

We discard all parent posts that were classified as unrelated to a misinformation myth; this yielded 16,796 total parent posts. We then select all replies to these parent posts (constructing conversation branches). With these parent posts, our goal is to assess which replies within each branch agree with the misinformation myth. We utilized ChatGPT to parse these conversation branches and assess which replies agree with the relevant misinformation myth. All replies that are classified 'agrees with misinformation' are also counted as misinformation in our dataset. We manually verify these labels using human coders and achieved an accuracy of 90.0%. Full performance statistics, as well as the verification procedure, are available in

[3] ChatGPT is a state-of-the-art artificial intelligence large language model trained using both supervised and reinforcement learning released by the research laboratory OpenAI.

Appendix 3. All parent posts, as well as all subsequent replies within conversation branches that are coded as 'agrees with misinformation' are counted as misinformation within this study. This yielded 4,788 posts.

2.5 Empirical Analysis

2.5.1 Semiparametric Difference-in-Differences Estimation

This section describes the empirical methodology. Because we are analyzing an event – quarantine –difference-in-differences (DID) seems warranted. However, difference-in-differences estimation requires that "in the absence of the treatment, the average outcomes for treated and control groups would have followed parallel paths over time" (Abadie 2005). This assumption may not hold if pretreatment characteristics that affect outcomes differ significantly between treatment and control groups.

Semiparametric DID estimation provides estimates for the average treatment effect on the treated (ATT) when the deterministic pretreatment characteristics that affect outcomes differ significantly between treatment and control groups. The assumption that must hold in this case is the *conditional* parallel trends assumption- that if treatment and control groups had identical pretreatment characteristics, and no treatment occurred, outcomes would follow parallel trends. (Abadie 2005).

The estimation process is as follows: propensity scores are semi-parametrically estimated for each unit of analysis utilizing a polynomial series of control variables. These scores explicitly account for any observed cofounders that affect the likelihood of treatment as well as subsequent outcome trends. We utilize a linear probability model to estimate propensity scores. For an observation x_b with *i* binary control pretreatment characteristics and *k* continuous control pretreatment characteristics, the propensity score $\hat{\pi}(x_b)$ is estimated as follows below in Equation 1:

$$\hat{\pi}(x_b) = \widehat{\gamma_0} + \hat{\gamma}_1 \times x_1 + \sum_{i=1}^{k} \widehat{\gamma_{2i}} \times x_2^i \quad (1)$$

where x_1 is a binary variable, $x_2^i = \prod_{j=1}^{i} x_2$, and the coefficients $\widehat{\gamma_0}, \hat{\gamma}, \ldots\ldots, \widehat{\gamma_{2i}}, \ldots\ldots, \widehat{\gamma_{2k}}$ are estimated using an ordinary least squares estimator. Observations for which the propensity score is less than 0 or greater than 1 are discarded from the analysis.

These propensity scores $\hat{\pi}(x_b)$ are utilized to weight each unit of analysis's change in the outcome variable over time (Δy_t). **d** is equal to 1 if the observation is for a treated participant; 0 if they are in the control group. The unbiased estimate of the ATT is obtained by comparing the weighted change of dependent variable *y* over time as below in Equation 2. Other variables can also be added into the estimation process to see how the ATT varies with other characteristics.

$$E\left\{\frac{\Delta y_t}{P(d=1)} \; x \; \frac{d - \hat{\pi}(x_b)}{1 - \hat{\pi}(x_b)}\right\} \quad (2)$$

2.5.1.1 Justification for Usage

To justify the usage of semiparametric DID, it is necessary to demonstrate statistically significant differences in baseline characteristics between control and treatment groups that impact the likelihood of treatment (Houngbedjie 2016). Table 2 provides the group means, as well as statistically significant differences assessed by t-test for all baseline continuous variables.

A comparison of quarantine-exposed and control group sample means of baseline activity shows statistically significant differences between the groups. The average quarantine-exposed user posts at a much higher volume than the average control group user (both misinformation and non-misinformation posts), within a wider breadth of forums both inside and outside the vaccine network. Additionally, the average quarantine-exposed user has a significantly shorter tenure on the platform.

Table 2. Continuous Baseline Variables

	Quarantine-Exposed Group Mean	**Control Group Mean**	**Difference**
Vaccine network misinformation posts	1.712	0.022	-1.690**
Vaccine network posts	40.703	3.532	-37.170**
Vaccine network forums	2.189	1.039	-1.150**
Outside network posts	401.123	212.662	-188.460**
Outside network forums	143.939	72.672	-71.268**
User Tenure	626.634	753.850	127.216**
*** p<0.05; ** p<0.01**			

Table 3 provides the group proportions for baseline dummy variables, as well as statistically significant differences assessed by t-test. A comparison of quarantine-exposed and control group proportions shows statistically significant differences between those groups. Quarantine-exposed users are significantly more likely to serve as moderators somewhere on the platform, and less likely to have a verified email account. However, there is no statistically significant difference in the proportion of users with premium accounts between groups. Taken with the results above comparing baseline activity between quarantine-exposed and control groups, this is strong evidence of a difference in pretreatment characteristics that could affect both a

user's likelihood of being exposed to quarantine, as well as their response to it, and justifies the usage of semiparametric DID estimation[4].

Table 3. Dummy Baseline Variables

	Quarantine-Exposed Group Proportion	**Control Group Proportion**	**Difference**
Is Moderator	0.305	0.226	-0.079**
Has Premium Account	0.017	0.013	-0.003
Verified Email	0.635	0.704	0.069**
*** p<0.05; ** p<0.01**			

2.5.2 Empirical Approach

The intervention of interest is exposure to quarantine. Quarantine-exposed users comprise the treatment group, and the control group is made up of users who were not exposed to the quarantine. The data comprise user-level posting data both before and after quarantine. The baseline period is the period before quarantine (2/28/19 - 5/22/19), and the follow-up period is the period after quarantine (5/23/19 – 8/15/19).

We run semiparametric difference-in-differences analyses to empirically assess the effect of quarantine. We estimate the average treatment effect on the treated (ATT) by comparing the change in the outcome of interest between follow-up and baseline periods across treatment groups while adjusting for differences in observable characteristics at baseline. The appropriate specification is below:

$$\text{ATT} = \left\{\frac{Outcome_{follow-up} - Outcome_{baseline}}{P(QuarantineExposed=1)}\ x\ \frac{QuarantineExposed - \hat{\pi}(x_b)}{1-\hat{\pi}(x_b)}\right\} \quad (3)$$

where x_b refers to the set of user account features and baseline activity variables (available in Table 1).

[4] For subsequent semiparametric analyses comparing quarantine-exposed users to neutral-vax native users, the same comparison between baseline variables was performed. Pervasive statistically significant differences between these variables were found between the two groups to justify the empirical approach.

2.6 Results & Discussion

The following section describes the main results. After the presentation of each set of results (First Order Effects, Effect of Quarantine on Other Vaccine-Related Forums, Heterogeneity Analyses, Effect of Quarantine on Platform Contribution), we discuss each result, highlight its relationship to prior literature and established research gaps, and provide managerial insights.

2.6.1. First-Order Effects

An initial analysis provides evidence that users significantly decrease activity within the quarantined anti-vax forums. As shown in Figure 4, the number of unique contributors per week to quarantined anti-vax forums plummets after quarantine, suggesting that new users are failing to find these forums and an insufficient number of users are returning. Between the baseline and follow-up period (+/- 12 weeks around quarantine), the average number of unique contributors per week drops 70.6%, indicating a significant drop in group size. This is accompanied by a 61.27% drop in misinformation posts within these forums. Quarantine-exposed users decrease misinformation contribution between baseline and follow-up periods by -0.664 posts on average.

This result is consistent with previous findings in literature across multiple streams. It has been shown that the size of an online social network is positively correlated with contribution frequency (Burke et al. 2009; Wang et al. 2019; Zhang & Zhu 2011). Thus, when group size drastically drops in these forums due to a decline in their visibility and accessibility, it follows that contribution would drop as well. Many operations management studies investigate how platforms can manipulate information visibility to achieve a desired outcome such as product adoptions, consumer purchases, or crowdfunding donations (Küçükgül et al. 2022; Maglaras et al. 2022; Xiao et al. 2021; Zhou & Chen 2016). This result complements these studies – it shows that decreasing the visibility of misinformation via prominence reduction precipitates a significant decline in misinformation within problematic communities.

This primary result suggests that managers should apply prominence reduction policies to problematic communities, as they result in a dramatic decline in misinformation. As the drop in group size is the driving force behind the misinformation decline, this suggests that this policy would achieve the same result across other types of misinformation topics (beyond anti-vaccine myths). We elaborate on the managerial insights in greater detail after subsequent analyses below.

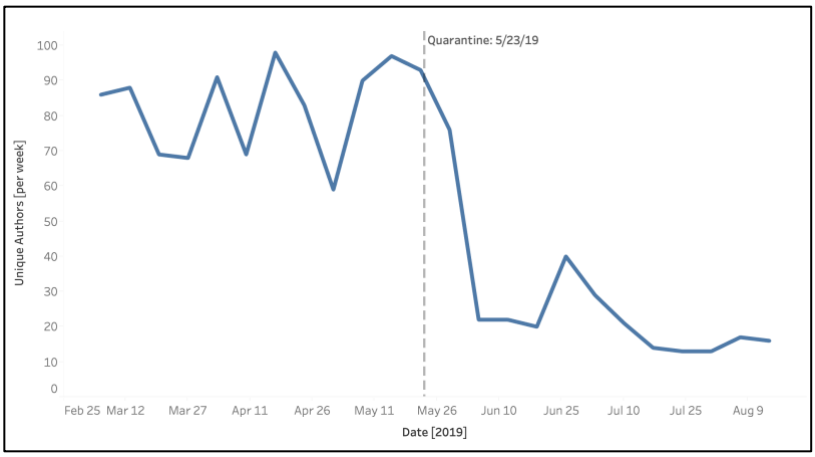


Figure 4. Unique Users/Week in Quarantined Anti-Vax Forums

2.6.2. Effect of Quarantine on Other Vaccine-Related Forums

To assess the effect of quarantine on users' misinformation spreading within other vaccine-related forums, we utilize a semiparametric DID analysis. The results are shown in Table 4. The number of misinformation posts is the dependent variable. The estimate of the average treatment effect on the treated (ATT) is positive and significant (ATT= 0.803; $p<0.026$) for misinformation within neutral-vax forums. Notably, the magnitude of this increase (0.803 misinformation posts) exceeds the corresponding average decrease (-0.664 misinformation posts) within quarantined anti-vax forums. This indicates that quarantine-exposed users spill over from quarantined anti-vax to neutral-vax forums, where they fully substitute (and then exceed) their misinformation contribution. This spillover is restricted to neutral-vax forums: quarantine-exposed users do not significantly change their misinformation behavior within pro-vax forums (ATT = -0.057; $p<0.211$).[5]As a robustness check, we run a difference-in-differences analysis on a matched sample which produces consistent results; this is available in Appendix 5.

Table 4. Effect of Quarantine on Misinformation Posts in Vaccine-Related Forums

Forum Type	**ATT**	**Standard Error**	**z**	**P>z**	**Confidence Interval**	
Neutral-Vax Forums	0.803	0.361	2.23	0.026	0.096	1.511
Pro-Vax Forums	-0.057	0.046	-1.25	0.211	-0.147	0.033
N: 43,670						

[5] Analysis for the single non-quarantined anti-vax forum was excluded due to a cessation of activity prior to quarantine, as well as core differences in forum discourse. Full analysis and details are in Appendix 4.

We assess the impact of this spillover of misinformation to neutral-vax forums by performing an additional set of analyses. These analyses directly compare quarantine-exposed users with neutral-vax natives, to assess if the observed change in quarantine-exposed user behavior has an impact on other users. As described in Table 1, neutral-vax natives are control users who posted in neutral-vax forums during the baseline period. Figure 5 shows a descriptive comparison between the two groups. In the short term (until 8/15/19), quarantine-exposed users significantly increase misinformation contribution, while neutral-vax natives remain constant. However, it shows that in the long-term (until 3/1/20[6]), this increase does not persist. Beginning in August 2019, quarantine-exposed users steadily decrease misinformation contribution to close to pre-quarantine levels. This indicates that the misinformation spillover to these forums is significant but relatively short-lived.

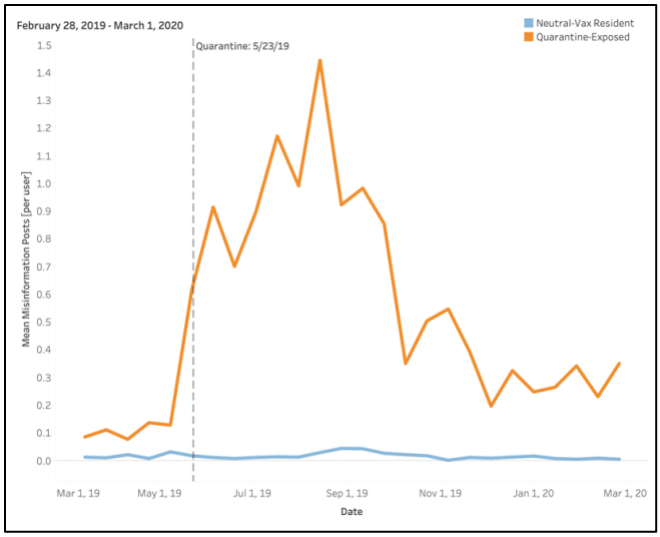


Figure 5. Misinformation in Neutral-Vax Forums: Quarantine-Exposed Users vs. Neutral-Vax Natives

These descriptive results are confirmed by semiparametric DID. Table 5 shows that quarantine-exposed users significantly increase misinformation contribution compared to neutral-vax natives in the short-term (ATT= 0.835 misinformation posts; $p<0.033$). However, in the long term, the difference between quarantine-exposed users and neutral-vax natives becomes marginally significant ($p<0.091$).

Table 5. Comparison between Quarantine-Exposed Users and Neutral-Vax Natives, Misinformation

Forum Type	**ATT**	**Standard Error**	**z**	**P>z**	**Confidence Interval**

[6] Analysis was cut off here due to potential disruption from the COVID-19 pandemic.

Neutral-Vax Forums, Short Term	0.835	0.939	2.13	0.033	0.066	1.605
Neutral-Vax Forums, Long Term	1.293	0.764	1.69	0.091	-0.204	2.791
N: 1,305						

To dig deeper into this result and ascertain why the misinformation spillover does not persist, we run analyses assessing if misinformation becomes contagious to neutral-vax natives. These are run on all posts from natives made during the long-term follow-up period (May 23, 2019 – March 1, 2020). The goal is to assess if direct interaction with misinformation on a conversation thread makes these control users more likely to post misinformation themselves subsequently. For more details regarding the identification strategy for this analysis, as well as a table with full regression results, please refer to Appendix 6. These analyses show that neutral-vax natives are not significantly more or less likely to spread misinformation ($p<0.291$) after directly interacting with it on a conversation thread. Thus, there is not strong evidence for misinformation contagion to neutral-vax natives. This may be a factor that contributes to the short-lived nature of the misinformation spillover in these forums.

These results provide strong evidence that quarantining anti-vax forums initiates a misinformation spillover in topically related neutral forums. We find evidence for a substitution effect: though users decrease misinformation within anti-vax forums, they increase it (at a higher magnitude) in neutral-vax forums. Neutral-vax forums are also centered around the focal topic of the quarantined forums (vaccines) but are ideologically mixed, allowing both pro- and anti-vax views. However, as the only non-quarantined anti-vax forum had become inactive by the time of quarantine (see Appendix 4), these forums were the closest ideologically for quarantine-exposed users.

Prior work within the operations management literature has established the presence of homophily on online platforms: individuals tend to cluster by homogenous characteristics (Hwang & Krackhardt 2020). This applies to misinformation as well: it has been shown that individuals prefer seeking out stances that align with their own beliefs (Zhou & Shen 2022). It stands to reason that quarantine-exposed users from anti-vax forums would spill over to these neutral spaces (rather than pro-vax forums) to engage in vaccine-related discourse.

Both descriptive and empirical evidence indicate that this misinformation spillover is relatively short-lived, beginning to decline after roughly three months (September 2019). However, supplemental analyses indicate that quarantine-exposed users maintain a high level of non-misinformation contribution. These analyses are available in Appendix 7. This indicates that the attenuation of the misinformation spillover is not due to a general drop in quarantine-exposed user activity and is a result of a misinformation-specific process.

Prior literature on informational cascades and misinformation illuminates the mechanism of this result. Though quarantine-exposed users increase misinformation in these forums, there is no evidence of contagion to previous forum natives. This spread of misinformation on online platforms has been shown to occur through informational cascades (Vosoughi et al. 2018). A lack of contagion indicates that these cascades are not occurring. Thus, the misinformation spillover was dampened over time.

Within the operations literature, the presence of informational cascades has been studied within the context of crowdfunding (Xiao et al. 2021), product adoption (Zhou & Chen 2016), and consumer purchases (Küçükgül et al. 2022; Maglaras et al. 2022). The adoption of misinformation is distinct from other types of behaviors studied in the literature, as individuals' propensity to adopt is difficult to shift. Individuals typically prefer seeking out stances on misinformation that align with their current beliefs (van der Linden et al. 2020; Zhou & Shen 2022). Misinformation has been shown to be subject to disconfirmation biases: being presented with contradictory facts often has the effect of making individuals argue more vigorously for their views (Lodge & Taber 2013). Though previous literature states that the likelihood of an individual adopting a new behavior or information monotonically increases with the number of exposures (Hodas & Lerman 2014; Pastor-Satorras & Vespignani 2001), this does not factor in the mitigating role of individuals' propensity to adopt.

Though the neutral-vax forums were the ideologically closest active forums available to quarantine-exposed users, the user population within those forums differed significantly regarding their propensity to adopt misinformation. Neutral-vax natives posted -1.534 fewer misinformation posts ($p<0.01$) in the baseline period than quarantine-exposed users. Consequently, there were no cascades of misinformation due to the lack of a critical mass of natives receptive to misinformation in these forums.

These results complement and expand upon literature across multiple streams. Many operations management studies investigate how, in the face of informational cascades, online platforms can manipulate the visibility of information to promote desired user behavior (Küçükgül et al. 2022; Maglaras et al. 2022; Xiao et al. 2021; Zhou & Chen 2016). This set of results complements and expands upon these studies. These results show that decreasing the visibility of problematic communities can achieve the desired outcome (a drop in misinformation within those communities), but this also initiates a short-lived misinformation spillover to other related communities. Candogan & Drakopoulos (2020) study how the information cascade process for misinformation can be attenuated through a signaling mechanism, but do not include individuals' propensity to adopt within their model. This is significant, as we find that the continuation of the informational cascade process for misinformation is contingent upon exposed individuals' having a high propensity to adopt misinformation.

With regard to platform policies aimed at decreasing misinformation, prior literature has examined approaches such as labeling it (Candogan & Drakopoulos 2020; Lees et al. 2021; Papakyriakopoulos & Goodmann 2022; Saltz et al. 2021; Zanettou et al. 2021) and deleting it (Bak-Coleman et al. 2021; Papanastiasou 2020).

This study expands upon this work by empirically assessing the efficacy of prominence reduction.

These results also produce significant managerial insights. Prominence reduction policies against misinformation hold promise for user-generated content (UGC) platforms, as prior work has found that other approaches such as labeling misinformation are ineffective (Lees et al. 2021; Papakyriakopoulos & Goodmann 2022; Zannettou 2021). Furthermore, deleting misinformation raises free speech concerns and can be difficult to implement given the short time frame within which UGC becomes viral (Bak-Coleman et al. 2021; Papanastiasou 2020). This work shows that prominence reduction policies have the potential to produce misinformation spillovers. The relatively short-lived nature of the spillover in this case, as well as the decline of misinformation in directly impacted groups, leads us to conclude that prominence reduction policies can be effective in containing misinformation. However, misinformation spillovers within groups with a higher propensity to adopt misinformation have the potential to be of longer duration and more harmful. Given that the mechanisms that drive these results (homophily, informational cascades, and disconfirmation biases) have been established across multiple online platform types, as well as topical domains, we expect that these results would generalize across all types of misinformation.

2.6.3. Heterogeneity Analyses

To assess if the highest contribution users (superusers) respond differentially to quarantine, we run semiparametric DID analyses with $Superuser_i$ as the moderating variable. The results are available below in Table 6. In neutral-vax forums, being a superuser leads to an increase of 6.678 misinformation posts ($p<0.035$). When factoring in the moderating effect of being a superuser, the other users slightly increase misinformation in neutral-vax forums, but the significance is borderline (ATT =0.074; $p<0.103$).[7]

Table 6. Effect of Quarantine on Misinformation Posts in Neutral-Vax Forums for Superusers

Neutral-Vax Forums Misinformation	**ATT**	**Standard Error**	**z**	**P>z**	**95% Confidence Interval**	
Superusers	6.678	3.16	2.11	0.035	0.486	12.879
Other Users	0.074	0.045	1.63	0.103	-0.015	0.162
N: 43,670						

[7] The same analyses were run for pro-vax forums and were consistent with the main analyses shown in Table 4: neither superusers users nor other users significantly changed misinformation contribution.

We then assess if the most problematic users (i.e., those most likely to spread misinformation) respond differentially to quarantine. Problematic users are users who spread misinformation at baseline. We ran semiparametric DID analyses with $ProblematicUser_i$ as the moderating variable. The results for neutral-vax forums are available below in Table 7. In neutral-vax forums, being a problematic user leads to an increase of 9.147 misinformation posts ($p<0.011$). When factoring in this moderating effect, the other users do not increase misinformation in neutral-vax forums (ATT =0.003; $p<0.370$). [6]

Table 7. Effect of Quarantine on Misinformation Posts in Neutral-Vax Forums for Problematic Users

Neutral-Vax Forums Misinformation	**ATT**	**Standard Error**	**z**	**P>z**	**95% Confidence Interval**	
Problematic User	9.147	3.594	2.550	0.011	2.103	16.192
Other Users	0.003	0.003	0.900	0.370	-0.003	0.009
N: 43,670						

These findings show that misinformation spillovers solely come from the highest contributing users (superusers), and more strongly, users known to engage in misinformation in the past (problematic users). Given that superusers post significantly more in vaccine-related forums, it is likely that they also spend much more time monitoring said forums. The policy change of quarantine would thus be much more salient to them, and more likely to initiate a behavioral change. Regarding problematic users, it follows that they would be most likely to create misinformation spillovers due to past behavior.

Previous work has established individuals' responses to changes in online platforms are strongly moderated by user characteristics (Wang et al. 2019). These results complement this literature by assessing how different types of users respond to platform policies aimed at decreasing misinformation. Prior research established that misinformation can be sourced to a small group of problematic users (Grinberg et al. 2019; Peres Nobre et al. 2022; Shao et al. 2018). These results are consistent with this.

This work additionally produces actionable insights for managers. As superusers drive the misinformation spillover, this result suggests that these users should be removed or receive additional sanctions after prominence reduction is applied. Given the significant economic value of user-generated content for firms (Chen et al. 2016; Duan et al. 2008; Lau et al. 2018), and that platforms disproportionately rely on the contributions of superusers (Carron-Arthur et al. 2014; Chen et al. 2016; Nielsen Norman Group 2006), this creates a quandary. If these users were to be removed en masse, a platform could take a significant contribution hit. However, our results show that the spillover is even more strongly produced by problematic users. This provides a more straightforward solution for platforms: users with a history of misinformation should be monitored and potentially removed after prominence reduction is applied.

2.6.4. Effect of Quarantine on Platform Contribution

Finally, to assess the effect of quarantine on overall platform contribution, we utilize a semiparametric DID analysis. The results are below in Table 8. The number of non-misinformation posts on the Reddit platform is the dependent variable. The estimate of the average treatment effect on the treated (ATT) is insignificant (ATT= 88.716; $p<0.143$). This indicates that quarantine-exposed users are not changing their level of non-misinformation contribution on the platform due to the quarantine intervention. Reddit is comprised of thousands of topic-specific forums, and users typically participate in a wide range of different types of forums. As the quarantine intervention was only localized to a small number of vaccine-related forums, we do not see it impacting user contribution to the platform as a whole.

Table 8. Effect of Quarantine on Non-Misinformation Contribution on the Reddit Platform

Forum Type	**ATT**	**Standard Error**	**z**	**P>z**	**95% Confidence Interval**	
Reddit Platform	88.716	60.531	1.470	0.143	-29.923	207.354
N: 43,670						

Prior literature within operations management has established how online platform policies and qualities impact user contribution. Researchers have established that the size of an online social network is positively correlated with contribution frequency (Burke et al. 2009; Zhang & Zhu 2011), quality (Wang et al. 2019), and the presence of social nudges (Zeng et al. 2022). Social network structure drives contribution frequency and quality (Jin et al. 2022; Wei et al. 2021). This result expands upon these studies by showing that prominence reduction is not detrimental to non-misinformation contribution.

This result has practical implications for managers. Enacting content moderation policies runs the risk of alienating users and decreasing their contribution (Saltz et al. 2021). Indeed, these policies have been shown to be suboptimal with regard to maximizing contribution (Candogan & Drakopoulos 2020). This finding indicates that even though prominence reduction policies impact user misinformation contribution, they do not decrease otherwise welcome contributions to the platform. Given platforms' incentives to maximize contribution (Chen et al. 2016; Duan et al. 2008; Lau et al. 2018), this shows that the implementation of prominence reduction policies may not decrease firms' operational efficiencies.

2.7 Concluding Remarks

This study empirically examines the efficacy of prominence reduction policies applied to misinformation. The specific study context is the application of Reddit's quarantine policy to groups identified as high in misinformation. We find strong evidence that prominence reduction decreases misinformation in the focal groups, as expected. However, the policy prompts a misinformation spillover in other topically related groups, although this spillover is relatively short-lived as it is not contagious

to other users. This creates a dampening effect that causes the misinformation spillover to ultimately die out. This misinformation spillover is solely driven by users with a history of misinformation. Finally, enacting prominence reduction policies does not meaningfully decrease non-misinformation contribution on the platform.

This work contributes to three literature streams that address vital questions for online platforms. Many operations management studies assess how online platforms can manipulate the visibility and accessibility of information to achieve a desired outcome. This study contributes to this by assessing how decreasing the visibility of misinformation via prominence reduction policies ultimately impacts its spread while factoring in the potential presence of informational cascades. Given the disruptive nature of misinformation, prior researchers study how firms can optimally respond to, detect, and even deploy misinformation. This work contributes to this stream by empirically assessing how online platforms can best reduce the impact of misinformation. It builds upon previous work in detecting misinformation by utilizing a novel approach combining manual coding and automated artificial intelligence methods. Many studies assess how platforms can boost user contribution. This study contributes to this literature by assessing the impact of prominence reduction policies on non-problematic user contribution. Finally, prior work assesses how the effect of both platform structure and policy changes can be moderated by user characteristics. This study contributes to this literature by assessing how different groups of users differentially respond to prominence reduction.

This study produces significant managerial insights. The results indicate that prominence reduction policies can be an effective methodology for containing misinformation. However, these policies have the potential downside of producing misinformation spillovers. To address this concern, platforms should selectively monitor users in problematic groups with a history of misinformation, and potentially apply additional sanctions. Furthermore, spillovers to groups with users with a high propensity to engage in misinformation should be very closely monitored as they have the potential to be more harmful. Additionally, these results indicate that these policies are not detrimental to overall user contribution. Overall, this study indicates that hiding misinformation on an online platform is an effective method that can be deployed with limited risk.

This study has several limitations that future studies may address. First, the dataset contains information about posting behavior to focus on the spread of misinformation, not the user reaction to that information such as likes/upvotes, views, or subscriptions. Future work could examine the effect of prominence reduction strategies on user engagement with content, not just with the spread of misinformation. Second, the data lacks user demographic information. Future work could examine the heterogeneous effects of prominence reduction policies across gender, generational, and racial lines.

The main analyses show that prominence reduction produced a short-lived misinformation spillover. Subsequent analyses indicate that the short-lived nature of this spillover was due to a lack of misinformation contagion to other users, as they had a lower propensity to adopt misinformation. Future work could see if this lack of contagion holds when misinformation spills over to a group with a higher propensity to engage in misinformation, and if the spillover is of a similar duration. This study

also demonstrates that prominence reduction policies have a heterogeneous effect based on the type of user. Finally, future work could assess if these policies have a heterogeneous effect across different types of harmful content, such as illegal content, hate speech, or harassment. Overall, this study contributes to the operations research literature on misinformation in multiple ways. Due to the risks it poses to both firm operations and society at large, future research on mitigating the harm of misinformation is essential.

Chapter 3: The Impact of Platform Policy Interventions: Mitigating Verbal Aggression Online

3.1 Introduction

Harmful user-generated content (UGC) such as verbal aggression and harassment, misinformation, as well as illegal content are widely and rapidly disseminated on online platforms. This has dire implications for societal welfare, but also the operational aims of platforms (Carpenter 2021). As a result, platforms have adopted various content moderation policies. The goal of these policies is to decrease the volume and impact of objectionable content circulating on these platforms. These policies typically take one of two approaches.

The first approach, prominence reduction, reduces the visibility and spread of harmful content, rather than deleting it. This has the advantage of preserving free speech. The second approach, banning, either deletes content, removes users, or closes a group of users. This study focuses on both content moderation methodologies applied on the group level. This means that a problematic group is identified and receives an intervention, but the users remain and can continue participation elsewhere on platform. Group-level policies are advantageous for platforms as they can avoid applying an intervention to a large number of individuals, which can decrease the number of daily active users on a platform, a key metric used to set advertisement prices as well as assess firm valuation (Patel 2020). Furthermore, individual-level policies have been proven to be unpopular with users (Newell et al. 2016; Chandrasekharan et al. 2017; Gavin 2019; Shen & Rose 2019).

Prior work shows that applying group content moderation policies decreases the activity and volume of harmful content within the problematic group (Copland 2020; Chandrasekharan et al. 2022; Mudambi et al. 2022). Additionally, it has been shown that these policies have the potential to change user behavior in outside groups on a platform (Chandrasekharan et al. 2022; Trujillo & Cresci 2022). However, it is unclear if these policies are effective in decreasing harmful behaviors on a platform, as well as what externalities are associated with using them. We address this by examining how both group-level prominence reduction and banning affect verbal aggression behaviors on a UGC platform. We ask: How does applying prominence reduction to a problematic group affect verbal aggression behavior on a platform? How does applying banning to a problematic group affect verbal aggression behavior on a platform? Do the outcomes of both group prominence reduction and banning vary over the type of user and type of content?

To answer these questions, we analyze Reddit group-level content moderation policies. Reddit is a massively popular discussion platform comprised of thousands of topic-specific forums. We examine two policies applied on the forum level. The first policy is a prominence reduction policy whose goal is to hide a problematic forum from other users to lessen its impact. The second policy, forum banning, deletes a forum from the platform but leaves its participants on the platform to participate in other forums. We collect large datasets of user posts from Reddit and use machine learning prediction algorithms to analyze the text and classify verbal aggression. This

work applies econometric methods to estimate the causal impacts of these interventions.

As a first-order effect, both approaches significantly decrease the volume of verbal aggression within the problematic group. However, we find that both group prominence reduction and group banning cause users to increase verbal aggression within other ideologically similar forums on the Reddit platform, which constitutes verbal aggression spillovers. However, group banning produces a wider spillover: verbal aggression increases in both ideologically similar and dissimilar forums.

To assess what led to the differences in spillover between the two interventions, we assessed if the group banning intervention differentially affected a subgroup of users. We were able to identify multihoming users – users who concurrently operated an account on an external platform with functionality identical to Reddit. We found that the wider spillover after the group ban was driven by the behavior of multihoming users alone. Supplemental analyses indicate that having access to an external alternative platform disinhibits multihoming users. They are more aggressive in ideologically dissimilar and potentially unfriendly forums, where they risk further platform sanctions.

Moreover, we find that the verbal aggression spillovers produced by both group content moderation policies produce contagion. Our results show that verbal aggression is a contagious behavior: direct exposure to it makes users from other groups more likely to be verbally aggressive. Interestingly, this contrasts with previous work (Mudambi et al. 2022), which examined the impact of group content moderation policies on misinformation. Mudambi et al. (2022) found that group prominence reduction produced misinformation spillovers, but these were short-lived due to a lack of contagion. We surmise this difference arises from the fact that verbal aggression produce unique behavioral dynamics due to their potential to incite both social influence and reciprocity (Wheeler & Levine 1967; Boxer et al. 2005).

Our results show that applying group content moderation policies to verbal aggression produces negative spillovers and that the more drastic the policy, the wider the spillover. Furthermore, the impact of these policies is significantly different for multihoming users. Finally, our results regarding contagion show that the efficacy and externalities associated with these using policies vary significantly based on the type of harmful content they are applied to. This work shows the group content moderation policies applied to verbal aggression are significantly less effective than the same policies applied to misinformation (Mudambi et al. 2022). We conclude that the platforms must consider the type of user, type of content, as well as the potential for spillovers when designing and deploying content moderation policies.

Our findings have important implications for scholars and practitioners. This work contributes to the literature on user-generated content platforms, content moderation policies, online antisocial behavior, platform multihomers, and social contagion. Additionally, our findings have the potential to directly inform the design of future content moderation policies.

3.2 Related Work

3.2.1 Verbal Aggression

Verbal aggression is defined within the communications literature as behavior that involves attacking the self-concepts of other people through the usage of spoken or written speech (Infante & Wigley 1986). It has been proven to inflict significant psychological pain on other individuals (Infante 1987; Walsht & Clarke 2003). Individuals' propensity to engage in verbal aggression, as well as their tolerance of others' use of it, increases with their exposure to inflammatory media (Cicchirillo et al. 2015).

Verbal aggression behaviors are present in the online context as well. With the rising global penetration of the Internet, the number of Internet platforms devoted to hate groups and hate speech has correspondingly risen (Banks 2010). This growth in online hate communities has paralleled a global increase in online verbal aggression in recent years (Banks 2010; Arthur 2019). A survey of Twitter users found that users do not feel adequately protected from verbal aggression while using social media (Jhaver et al. 2018).

Some studies indicate that verbal aggression behaviors on social media platforms arise not only from the characteristics of individual users, but also from established platform-wide norms. Newcomers to toxic online communities have been shown to mimic the so-called 'toxicity norms' of these communities (Kwon & Gruzd 2017; Rajadesingan et al. 2020). Indeed, verbal aggression, especially when targeted at a specific person, has been shown to be contagious within an online community (Song & Wu 2018).

3.2.2 Content Moderation Policies

Online platforms have adopted a variety of content moderation policies to combat the proliferation of harmful UGC (Edgecomb 2019; Benjakob 2020). This study compares the effect of two policies applied to a problematic group: prominence reduction and banning. Prominence reduction aims to hide harmful UGC from other users to decrease its impact, rather than deleting or prohibiting it. Because problematic speech is not removed, the prominence reduction approach has the advantage of preserving free speech on a platform and could thus prove more palatable to users.

Banning policies involve banning problematic users, groups, or topics from a platform, or deleting problematic posts. This approach has the disadvantage of raising free speech concerns, and the potential of angering users or even prompting users to leave a platform (Newell et al. 2016; Shen & Rose et al. 2019; Chandrasekharan et al. 2017). Specifically in this study, we focus on a group-level ban of users.

In recent years there has been an explosion of interest in work studying the effects of group-level content moderation policies. Studies show that after prominence reduction is applied to a problematic group, there is a significant drop in novel users within the said group, as well as the volume of total engagement (Copland 2020; Chandrasekharan et al. 2022; Mudambi et al. 2022). However, within groups specifically hidden for verbal aggression, there is little evidence that the

proportion of problematic language decreases (Copland 2020; Chandrasekharan et al. 2022; Trujillo & Cresci 2022).

Users active within a hidden group decrease their overall activity on the platform (Chandrasekharan et al. 2022; Trujillo & Cresci 2022). There is conflicting evidence, however, on the effect of prominence reduction on harmful behaviors in spaces outside of the focal group. Trujillo & Cresci (2022) found that users did not significantly change their verbal aggression behaviors in outside groups. However, when looking at users from communities censured for misinformation, there is evidence that users increase misinformation in topically and ideologically similar communities. This misinformation spillover is relatively short-lived, and is not contagious to other users (Mudambi et al. 2022).

There have been studies examining the impact of group banning policies upon user behavior. Banning a problematic group on a platform has been shown to significantly decrease the activity of its users elsewhere on the platform (Chandrasekharan et al. 2017; Trujillo & Cresci 2022). However, there is no consensus regarding the effect of a group ban upon verbal aggression elsewhere on the platform. Chandrasekharan et al. (2017) found that users significantly decreased verbal aggression usage on the platform. Conversely, Trujillo & Cresci (2022) found that users did not significantly change the median toxicity of their posts.

Researchers have established a link between group-level content moderation (through either banning or prominence reduction) and user migration to external platforms with less stringent policies (Newell et al. 2016; Copland 2020; Ali et al. 2021; Ribeiro et al. 2020). There is limited evidence for pure migration behavior; instead, most users multihome, maintaining activity on the focal platform, and the new external platform (Newell et al. 2016). Multihomers have been shown to increase verbal aggression on external platforms (Ribeiro et al. 2020).

3.2.3 Research Gap

In this work we aim to contribute an understanding of how group prominence reduction and banning policies affect user verbal aggression. Though there is evidence that prominence reduction curtails engagement within directly impacted communities (Copland 2020; Chandrasekharan et al. 2022; Mudambi et al. 2022), there are conflicting findings regarding its effect upon user behavior elsewhere on a platform. Trujillo & Cresci (2022) examined the behavior of users post-group prominence reduction and found no change in verbal aggression elsewhere on the platform, but did not consider the ideology of outside groups. This is significant, because previous research shows that after prominence reduction, users factor in topic and ideology when moving to new spaces on a platform (Mudambi et al 2022). Group banning has been shown to decrease user activity (Chandrasekharan et al. 2017; Trujillo & Cresci 2022). However, there are mixed results as to the effects of a ban upon verbal aggression. Finally, previous work shows that when group content moderation policies are applied to misinformation, it does not increase misinformation contagion (Mudambi et al. 2022). It is unclear if this result will hold for verbal aggression. We address that gap in this work.

Content moderation has been shown to push users onto external platforms with laxer policies (Newell et al. 2016; Copland 2020; Ali et al. 2021; Ribeiro et al.

2020). Most of these users multihome and concurrently operate accounts on the old and new platforms (Newell et al. 2016). Previous research has examined the effect of de-platforming (Rogers 2020; Ali et. al 2021; Ribeiro et al. 2020), but multihoming behavior within this context is understudied. Our results provide a fuller understanding of how multihoming behavior impacts users' response to group content moderation.

This work makes theoretical contributions to the literature on UGC platforms, content moderation, online verbal aggression, and platform multihoming. We address key research gaps regarding the downstream effects of group prominence reduction as compared to group banning. We also address research gaps on online verbal aggression. Finally, by factoring in multihoming behavior, we contextualize online content moderation policies within the entire Internet ecosystem. This addresses a key research gap, as most previous work examines the impact of content moderation policies on the enacting platform only.

3.3 Research Context

Our study focuses on Reddit, a social news aggregator platform, with more than 430 million active users. The platform is composed of thousands of topic-specific forums, within which users can contribute conversation threads to start a discussion, or reply to existing threads. Each forum is governed by its own set of rules, which are enforced by volunteer user moderators. The platform as a whole is governed by site administrators who are employed by Reddit.

To assess the effects of group-level prominence reduction and banning in the presence of multihoming behavior, this study examines a single problematic Reddit forum: r/The_Donald, a forum that was dedicated to promoting Donald Trump. Due to violations of Reddit's content policy[8] regarding hate speech and threats, Reddit utilized a prominence reduction policy in June of 2019, applying a Reddit policy called quarantine (Robertson 2019; Vigdor & Chokshi 2019). Quarantined forums do not show up in searches, do not generate advertising revenue, and require users to see a warning and explicitly opt-in to view them.

Additionally, quarantined forums' content does not appear in other users' newsfeeds unless they were subscribed before the intervention. This policy effectively hides a problematic forum from the rest of the site. We refer to this Reddit quarantine intervention as group prominence reduction throughout the paper. We refer to Reddit's Donald Trump forum, which received both treatments of interest, as r/The_Donald.

3.4 Data

3.4.1 Data Collection

Reddit applied group prominence reduction to r/The_Donald on 6/26/19 and banned it on 2/26/20. For the prominence reduction analysis, our treatment group is users who posted in r/The_Donald in a baseline period 3 months prior to prominence reduction. Our control group posted in right-wing forums (r/Conservative and

[8] https://www.redditinc.com/policies/content-policy

r/Libertarian)[9] in this period. We scraped all posts within political forums for these users +/- 3 months around prominence reduction. We eliminated control users that had ever posted in a prominence reduced or banned forum to ensure they were not exposed to the treatments of interest. For the ban analysis, we performed the same data collection protocol +/- 3 months around the group ban date[10]. The treatment and control groups were selected in the same manner as well.

3.4.2 Identifying Verbal Aggression

The verbal aggression types of interest in this study are hate speech and threats. Though many previous studies focus on general verbal toxicity (Ali et al. 2021; Ribeiro et al. 2020; Trujillo & Cresci 2021), we focus on hate speech and threats specifically. These types of verbal aggression direct unambiguous violence towards other individuals or groups (Josey 2010; Alorainy et al. 2019). In contrast, other types of potentially benign behaviors such as profanity or insults are often grouped into general measures of verbal toxicity (Risch & Krestel 2020).

We utilized the Google Perspective API[11] to identify hate speech and threats within the dataset. This API utilizes machine learning models trained on corpora of human-annotated text and can produce probability scores for four distinct types of verbal toxicity (insults, profanity, hate speech, threats), and two general types of toxicity (toxicity, severe toxicity). There is a high correlation between the probability scores produced by Google Perspective for the six types of verbal aggression[12]. For a post to be classified as hate speech in this study, its probability score for hate speech had to be higher than the three other distinct types of verbal aggression, as well as above 75%. The same was true for threats. Any post classified as either hate speech or a threat is deemed a verbal aggression post (VA).

3.5 Empirical Analysis

We utilize difference-in-difference (DID) analysis to analyze the effect of prominence reduction, as well as a group ban, upon user behavior. In this DID analysis, we compare the behaviors of our sample of r/The_Donald users (treatment group) with a control group of other Reddit users. We have two interventions of interest: exposure to prominence reduction, and a group ban, and thus run two types of analyses. In both cases, the treatment group was exposed to the intervention, and the control group was not. For both analyses, our data comprises user-level posting data +/- 3 months before and after the intervention of interest (group prominence reduction or group ban), as well as user account features.

---

[9] These communities were selected due to a similarity with r/The_Donald in terms of political ideology, as well as the usage of verbal aggression.

[10] As in previous research (Ribeiro et al. 2020; Trujillo & Cresci 2022), we excluded the final week (May 19 – May 26, 2020) from our analysis due to the occurrence of the George Floyd murder and subsequent protests on May 25, 2020. Verbal toxicity spiked on the Reddit platform during this period (Kumar et al. 2022) and has the potential to bias the analyses.

[11] https://www.perspectiveapi.com/

[12] Correlations on verbal aggression probability scores were run on a sample of 199,014 user posts from the prominence reduction baseline period (3/26/19 – 6/26/19).

We ran DID OLS regressions with robust standard errors on a weekly panel to assess the relative effects of group prominence reduction (group ban) upon verbal aggression posting (VA). For this specification, *i* represents the user, and *j* represents the week. The dependent variable VA_{ij} represents verbal aggression posts. The dummy variable $TheDonald_i$ indicates whether user *i* is in the treated group. The variable $PostPR_j$ ($PostBan_j$) indicates whether the period is after the group prominence reduction (group ban). The interaction term $TheDonald_i\ X\ PostPR_j$ ($PostBan_j$) indicates the effect of the intervention on the behavior of the treated users. The variables $Moderator_i$, $PremiumMember_i$, $VerifiedEmail_i$, and weekly dummies, are the control variables. Visual inspection confirms the parallel trends assumption for all analyses.

$$VA_{ij} = (TheDonald_i\ X\ PostPR_j) + TheDonaldi_i + PostPR_j + Moderator_i + PremiumMember_i + VerifiedEmail_i + \sum_{j=0}^{J} Week_j + \varepsilon_{ij}$$

3.6 Results

3.6.1 The Effect of Group Prominence Reduction on Verbal Aggression

Consistent with previous work, group prominence reduction significantly decreases verbal aggression within r/The_Donald. Within our sample, the average user verbal aggression posts (VA)/week decreases from 0.0146 to 0.0100 ($p<0.00$ by t-test). To assess how this drop in activity affects user verbal aggression elsewhere within the political network, we model the number of verbal aggression posts (VA) in Reddit left-wing, centrist, and right-wing forums. The coefficient estimates are shown below in Table 9; robust standard errors are in parentheses and p-values are below. We find evidence that r/The_Donald users exposed to quarantine increase verbal aggression posts within Reddit centrist ($\beta=0.004$; $p<0.000$) and right-wing ($\beta=0.007$; $p<0.000$) forums.

To assess the impact of these verbal aggression spillovers to other users in other political forums, we calculate how many users in these forums are directly exposed to verbal aggression coming from r/The_Donald users after prominence reduction. We find that 16,418 users in right-wing forums (17.9% of non-r/The_Donald users) are directly exposed to verbal aggression spillovers. 74,681 (11.6% of non-r/The_Donald users) are directly exposed in centrist forums.

Table 9. The Effect of Group Prominence Reduction upon Verbal Aggression

	Verbal Aggression Posts		
Reddit Location	**Left-Wing Forums**	**Centrist Forums**	**Right-Wing Forums**
r/The_Donald_i X Post-Prominence $Reduction_j$	0.000 (0.000) 0.319	0.004 (-0.001) 0.000	0.007 (-0.001) 0.000
r/The_Donald_i	-0.003	-0.017	-0.022

	(0.000) 0.000	(0.000) 0.000	(-0.001) 0.000
Post-Prominence Reduction$_j$	-0.001 (0.000) 0.022	-0.009 (-0.001) 0.000	-0.009 (-0.001) 0.000
Is Moderator$_i$	0.000 (0.000) 0.000	0.000 (0.000) 0.039	0.001 (0.000) 0.000
Premium Account$_i$	0.000 (0.000) 0.284	0.004 (-0.001) 0.005	0.005 (-0.001) 0.000
Verified Email$_i$	-0.001 (0.000) 0.000	-0.003 (0.000) 0.000	-0.002 (0.000) 0.000
+ week dummies			
Constant$_{ij}$	0.004 (0.000) 0.000	0.022 (-0.001) 0.000	0.023 (-0.001) 0.000
Observations	2,875,446	2,875,446	2,875,446
R^2	0.001	0.003	0.004
Robust standard errors in parentheses; P-values below			

3.6.2 The Effect of Group Ban Reduction on Verbal Aggression

To assess how the group ban of r/The_Donald affects user verbal aggression elsewhere on the platform, we model the number of verbal aggression posts in Reddit left-wing, centrist, and right-wing forums. The coefficient estimates are shown below in Table 10; robust standard errors are in parentheses and p-values are below. We find evidence that r/The_Donald users exposed to a group ban increase verbal aggression posts within left-wing (β=0.0002; p<0.041) and right-wing (β=0.003; p<0.000) forums and decrease it in outside centrist forums (β=-0.001; p<0.017).

To assess the impact of these verbal aggression spillovers to other users in other political forums, we calculate how many users in these forums are directly exposed to verbal aggression coming from r/The_Donald users after prominence reduction. We find that 2,231 users in right-wing forums (13.4% of non-r/The_Donald users) are directly exposed to the verbal aggression spillovers. 11,401 (3.8% of non-r/The_Donald users) are directly exposed in left-wing forums.

Table 10. The Effect of Group Ban upon Verbal Aggression

	Verbal Aggression Posts		
Reddit Location	**Left-Wing Forums**	**Centrist Forums**	**Right-Wing Forums**
r/The_Donald$_i$ X Post-Ban$_j$	0.0002 (0.000)	-0.001 (0.000)	0.003 (0.000)

	0.041	0.017	0.000
r/The_Donald_i	0.000 (0.000) 0.000	-0.001 (0.000) 0.000	-0.007 (0.000) 0.000
$Post\text{-}Ban_j$	0.000 (0.000) 0.118	0.003 (0.000) 0.000	-0.001 (0.000) 0.121
$Is\ Moderator_i$	0.000 (0.000) 0.000	0.000 (0.000) 0.056	0.001 (0.000) 0.000
$Premium\ Account_i$	0.000 (0.000) 0.398	0.003 (-0.001) 0.018	0.003 (-0.001) 0.003
$Verified\ Email_i$	0.000 (0.000) 0.000	-0.001 (0.000) 0.000	-0.001 (0.000) 0.000
+ week dummies			
$Constant_{ij}$	0.001 (0.000) 0.000	0.003 (0.000) 0.000	0.007 (0.000) 0.000
Observations	2,367,025	2,367,025	2,367,025
R^2	0.0001	0.0003	0.0010
Robust standard errors in parentheses; P-values below			

3.6.2.1 The Impact of Multihoming Behavior

After group prominence reduction, we observe verbal aggression spillovers to Reddit forums that are ideologically similar (centrist and right-wing) to r/The_Donald. However, after group banning, we observe verbal aggression spillovers to Reddit forums that are ideologically similar (right-wing), as well as ideologically dissimilar (left-wing). To assess what drives these differential responses, we dig deeper into our research context.

We find that after the initial group prominence reduction event in June of 2019, several moderators of r/The_Donald began creating an external platform, entirely separate from Reddit: thedonald.win[13]. thedonald.win apes the functionality and design of the Reddit platform. However, it bypasses Reddit's administration and moderation policies, which were viewed by many users within this community as biased and overly restrictive (Ribeiro et al. 2020). This new platform went live on November 21, 2019. We henceforth refer to thedonald.win as the external platform in this paper.

We identify multihoming users – users that concurrently operate accounts on Reddit and the external platform – within our treatment group of users. We identified multihoming users by matching them by username. To ascertain if multihoming

[13] The platform has subsequently changed its name to patriots.win

behavior affects an individual's response to group ban, we run two analyses: one comparing multihoming r/The_Donald users to the control group, another comparing non-multihoming r/The_Donald users to the control group. For both analyses, we model the number of verbal aggression posts (VA)in Reddit left-wing, centrist, and right-wing forums. The coefficient estimates are shown below in Table 11 and 12; robust standard errors are in parentheses and p-values are below. We find evidence that r/The_Donald non-multihomers exposed to a group ban increase verbal aggression posts within outside right-wing (β=0.003; p<0.000) forums but decrease it in centrist forums (β=-0.001; p<0.005). They do not statistically significantly change verbal aggression within left-wing forums (β=0.0001; p<0.159).

Table 11. The Effect of Group Ban upon Non-Multihomer Verbal Aggression

	Verbal Aggression Posts		
Reddit Location	**Left-Wing Forums**	**Centrist Forums**	**Right-Wing Forums**
r/The_Donald_i X $Post\text{-}Ban_j$	0.0001 (0.000) 0.159	-0.001 (0.000) 0.005	0.003 (0.000) 0.000
r/The_Donald_i	0.000 (0.000) 0.000	-0.001 (0.000) 0.000	-0.007 (0.000) 0.000
$Post\text{-}Ban_j$	0.000 (0.000) 0.080	0.003 (0.000) 0.000	-0.001 (0.000) 0.239
$Is\ Moderator_i$	0.000 (0.000) 0.001	0.000 (0.000) 0.043	0.001 (0.000) 0.000
$Premium\ Account_i$	0.000 (0.000) 0.445	0.003 (-0.001) 0.026	0.003 (-0.001) 0.003
$Verified\ Email_i$	0.000 (0.000) 0.000	-0.001 (0.000) 0.000	-0.001 (0.000) 0.000
+ week dummies			
$Constant_{ij}$	0.001 (0.000) 0.000	0.003 (0.000) 0.000	0.007 (0.000) 0.000
Observations	2,053,925	2,053,925	2,053,925
R^2	0.0001	0.0003	0.001
Robust standard errors in parentheses; P-values below			

To contrast, we find evidence that r/The_Donald multihomers exposed to a group ban increase verbal aggression posts within Reddit left-wing (β=0.0003;

p<0.001) and right-wing (β=0.003; p<0.000) forums. They do not statistically significantly change verbal aggression within centrist forums (β=-0.0002; p<0.428).

Table 12. The Effect of Group Ban upon Multihomer Verbal Aggression

	Verbal Aggression Posts		
Reddit Location	**Left-Wing Forums**	**Centrist Forums**	**Right-Wing Forums**
r/The_Donald_i X $Post\text{-}Ban_j$	0.0003 (0.000) 0.001	-0.0002 (0.000) 0.428	0.0034 (0.000) 0.000
r/The_Donald_i	-0.001 (0.000) 0.000	-0.002 (0.000) 0.000	-0.007 (0.000) 0.000
$Post\text{-}Ban_j$	0.000 (0.000) 0.195	0.004 (-0.001) 0.000	0.000 (-0.001) 0.586
$Is\ Moderator_i$	0.000 (0.000) 0.004	0.001 (0.000) 0.019	0.002 (0.000) 0.000
$Premium\ Account_i$	0.000 (0.000) 0.566	0.005 (-0.002) 0.006	0.005 (-0.002) 0.004
$Verified\ Email_i$	0.000 (0.000) 0.000	-0.001 (0.000) 0.000	-0.002 (0.000) 0.000
+ week dummies			
$Constant_{ij}$	0.001 (0.000) 0.000	0.003 (0.000) 0.000	0.006 (-0.001) 0.000
Observations	1,048,825	1,048,825	1,048,825
R^2	0.0001	0.0004	0.001
Robust standard errors in parentheses; P-values below			

3.6.3 Assessing Contagion

Our results show that r/The_Donald users significantly increase verbal aggression outside of the focal forum after group prominence reduction. To fully assess the externalities associated with this change in behavior, we assess the presence of verbal aggression contagion within forums that received a verbal aggression spillover (Reddit right-wing and centrist forums). We first assessed which Reddit right-wing and centrist forums r/The_Donald users posted in after group prominence reduction. We refer to the other users within these right and centrist forums that were not part of the r/The_Donald treatment group as 'native users'. We then examined the native users' posts within these forums for a three-month period after group prominence

reduction (6/26/19-9/26/19). The same data collection protocol was followed to assess contagion after the group ban.

We assess natives' verbal aggression behavior after direct exposure to verbal aggression coming from r/The_Donald users. Specifically, we evaluate the number of verbal aggression posts that native users make after being exposed to verbal aggression within a conversation thread. The dependent variable is a dummy variable *VA Posts*, which represents the number of verbal aggression posts a native user I contributes on their subsequent thread ($t+1$). The key independent variable is *DonaldVAExposure*. This is also a dummy variable, which is set to 1 if the user has been exposed to verbal aggression coming from an r/The_Donald user on a conversation thread *t*, and 0 otherwise. Controls are the variables $Moderator_i$, $PremiumMember_i$, $VerifiedEmail_i$ for each native user, the number of posts on thread t, weekly fixed effects to control for time trends, and forum fixed effects. We use the following specification:

$$VA\ Posts_{it+1} = DonaldVAExposure_{it} + Moderator_i + PremiumMember_i + VerifiedEmail_i + TotalPosts_t + \sum_{j=0}^{J} Week_j + \sum_{f=0}^{F} Forum_f + \varepsilon_{pi}$$

3.6.3.1 The Effect of Group Prominence Reduction Verbal Aggression Spillovers on Contagion

We observe verbal aggression spillovers from r/The_Donald users within Reddit right-wing and centrist forums after group prominence reduction. Thus, we assess contagion by examining posts coming from native users within those spaces in the post-prominence reduction period. The results are below in Table 13. We find that being directly exposed to a verbal aggression post from r/The_Donald user on a conversation thread makes a native user subsequently post 0.019 ($p<0.00$) more verbal aggression posts. Given the number of user-thread tuples with r/The_Donald verbal aggression exposures (96,151), this translates to an additional 1,797 verbal aggression posts coming from native users. This strongly suggests the presence of verbal contagion within Reddit right-wing and centrist forums after the group prominence reduction intervention.

Table 13. The Effect of Group Prominence Reduction Verbal Aggression Spillovers on Contagion

	Reddit Right-Wing & Centrist Forums
	Subsequent Verbal Aggression Posts
Donald VA Exposure$_{it}$	0.019 (0.001) 0.000
Is Moderator$_i$	-0.006

	(0.000) 0.000
Premium Account$_i$	-0.020 (0.001) 0.000
Verified Email$_i$	-0.018 (0.000) 0.000
Total Posts$_t$	0.000 (0.000) 0.000
+ week dummies	
+ forum dummies	
Constant$_{ij}$	0.034 (0.019) 0.068
Observations	3,246,205
R^2	0.005
Robust standard errors in parentheses; P-values below	

3.6.3.1 The Effect of Group Banning Verbal Aggression Spillovers on Contagion

After the group ban, we observe wider verbal aggression spillovers from r/The_Donald multihomers within Reddit right-wing forums and left-wing forums. Thus, we assess contagion by examining posts coming from control users within those spaces in the post-ban period. The results are below in Table 14. We find that being directly exposed to a verbal aggression post from r/The_Donald user on a conversation thread makes a native user subsequently post 0.006 ($p<0.00$) more verbal aggression posts. Given the number of user-thread tuples with r/The_Donald verbal aggression exposures (28,716), this translates to an additional 172 verbal aggression posts coming from native users. This strongly suggests the presence of verbal contagion within Reddit right-wing and left-wing forums after the group banning intervention.

Table 14. The Effect of Group Banning Verbal Aggression Spillovers on Contagion

	Reddit Right-Wing & Left-Wing Forums
	Subsequent Verbal Aggression Posts
Donald VA Exposure$_{it}$	0.006 (0.002) 0.000
Is Moderator$_i$	-0.002 (0.000) 0.000

Premium Account$_i$	-0.021
	(0.000)
	0.000
Verified Email$_i$	-0.01
	(0.000)
	0.000
Total Posts$_t$	0.000
	(0.003)
	0.000
+ week dummies	
+ forum dummies	
Constant$_{ij}$	0.036 (0.003) 0.000
Observations	1,970,054
R^2	0.007
Robust standard errors in parentheses; P-values below	

3.7 Discussion & Supplemental Analyses

This study uses data from users of the Reddit forum r/The_Donald to compare the effects of group-level prominence reduction with group banning policies on verbal aggression.

3.7.1 Prominence Reduction Applied to a Verbally Aggressive Group

We find that verbal aggression significantly decreases within r/The_Donald after group prominence reduction. This is consistent with changes user activity levels. This finding is consistent with previous work on prominence reduction policies (Copland 2020; Chandrasekharan et al. 2022; Mudambi et al. 2022). Consistent with the basic principles of network effects (Zhang et al. 2012), group size, as well as the number of ties within a social network have been tied to the frequency of user contribution on online platforms (Burke et al. 2009; Zhang & Zhu 2011; Shriver et al. 2013; Baek & Shore 2020). When group size diminishes, user contribution (and verbal aggression) drops with it.

We also find that r/The_Donald users increase verbal aggression within Reddit right-wing and centrist forums. Given that r/The_Donald was a right-wing forum, this suggests that group prominence reduction produces verbal aggression spillovers within ideologically similar spaces, but not within ideologically dissimilar spaces on a platform. Users' propensity to herd by personal ideology has been extensively documented in both offline (Gerber et al. 2013) and online (Colleoni et al. 2014; Boutyline & Willer 2016; Huber & Malhotra 2017) contexts. This behavior is an example of principle of homophily, which is the tendency for individuals to associate with similar others (McPherson et al. 2001).

3.7.2 Banning a Verbally Aggressive Group

After group ban, we find that users increase verbal aggression in Reddit right-wing forums, but decrease it in centrist forums. They also increase verbal aggression within left-wing forums. Although the magnitude of verbal aggression increases is smaller, the spillover is much wider. Furthermore, the increase in verbal aggression within left-wing forums runs contrary to our previous results regarding prominence reduction. Increasing verbal aggression within ideologically dissimilar forums runs contrary to the principle of homophily. We conclude that the more drastic banning policy applied to verbal aggression produces a worse outcome than prominence reduction.

To dig deeper into this result, we partitioned our analyses between multihoming and non-multihoming users. We find that non-multihoming users exhibit a stronger ideological herding behavior than that expressed after prominence reduction: these users increase verbal aggression only in Reddit right-wing forums and decrease it in centrist forums. In contrast, multihoming users increase verbal aggression in both Reddit right-wing and left-wing forums, indicating that they alone drive the observed wider verbal aggression spillover. We perform a robustness check that confirms that it is indeed multihoming behavior producing the wider spillover, not unobserved characteristics.

3.7.2 The Impact of Multihoming Behavior

We predict that multihomers' increase in ideologically dissimilar forums is driven by an increase in risk tolerance. Users continuously posting verbal aggression on UGC platforms risk further platform sanctions, namely being banned from the platform permanently (Cheng et al. 2015; Myers West 2018; Zhang et al. 2022). This is more likely for more extreme forms of verbal aggression, such as violent threats and hate speech, which we study in this work. The multihoming users within our sample operate accounts on an external platform (thedonald.win) whose functionality is highly similar to Reddit's. We reason that having access to an external Reddit alternative to use increases their risk tolerance, as these users have less fear of an individual ban. This shift in risk preferences leads them to increase verbal aggression within ideologically dissimilar (and potentially unfriendly forums), where it is more likely to be negatively received (Kelly Garrett 2009).

This explanation is supported by supplemental analyses. To support differences in risk tolerance between non-multihoming and multihoming groups, we model users' likelihood to receive a user ban. This analysis was partitioned between non-multihomer and multihomer users. We ran difference-in-differences analyses over baseline (3 months prior to group ban) and follow-up (3 months after) periods, rather than a weekly panel[14]. This specification utilized the same other user-level controls as previous analyses. The results are below in Table 15. We find that while non-multihomers are not significantly more likely to be banned than control users (β=-0.003; $p<0.127$), multihomers are 2% ($p<0.000$) more likely to receive a user ban

[14] Though we were able to ascertain with certainty which users received an individual ban using Reddit's API, we were not able to collect exact ban dates. We approximated these by identifying the date of a banned users' last post on the platform. Due to the lack of precision in this method, we opted to run the analyses on only two time periods.

after the group ban of r/The_Donald than control users. This supports our assertion that multihomers are engaging in riskier behavior.

Table 15. The Effect of Group Ban upon Likelihood of User Ban: Multihomers vs. Non-Multihomers

	Likelihood of User Ban	
User Type	**Non-Multihomers**	**Multihomers**
Ban-Exposed_i X Post-Ban_j	-0.003 (0.002) 0.127	0.020 (0.003) 0.000
Ban-Exposed_i	0.010 (0.001) 0.000	0.001 (0.002) 0.725
Post-Ban_j	-0.003 (0.001) 0.011	-0.003 (-0.001) 0.011
Is Moderator_i	-0.018 (0.000) 0.000	-0.014 (0.000) 0.000
Premium Account_i	-0.002 (0.001) 0.021	-0.002 (0.001) 0.027
Verified Email_i	-0.107 (0.000) 0.000	-0.100 (0.002) 0.000
Constant_{ij}	0.107 (0.002) 0.000	0.102 (0.002) 0.000
Observations	164,314	83,906
R^2	0.070	0.070
Robust standard errors in parentheses; P-values below		

We then model the number of posts removed by moderators within left-wing forums. When user's post explicitly breaks forum or Reddit-wide rules, forum moderators will remove it. An increase in this rule-breaking within left-wing forums from multihomers supports our risk tolerance mechanism. We ran difference-in-differences analyses on weekly panel, utilizing user-level control variables. The results are below in Table 16. We find that non-multihomers do not significantly change their behavior after the group ban. However, multihomers increase the number of removed posts within left-wing forums (β=0.0002; $p<0.040$). This shows that multihomers increase rule-breaking, along with verbal aggression within left-wing forums.

Table 16. The Effect of Group Ban upon Removed Posts: Multihomers vs. Non-Multihomers

	Moderator-Removed Posts in Left-Wing Forums	
User Type	**Non-Multihomers**	**Multihomers**
Ban-Exposed$_i$ X Post-Ban$_j$	0.000 (0.0001) 0.738	0.0002 (0.000) 0.040
Ban-Exposed$_i$	-0.0003 (0.000) 0.000	-0.001 (0.000) 0.000
Post-Ban$_j$	0.0001 (0.0001) 0.583	0.000 (0.000) 0.912
Is Moderator$_i$	0.0003 (0.000) 0.000	0.0002 (0.000) 0.002
Premium Account$_i$	-0.0004 (0.0001) 0.004	-0.0003 (0.0002) 0.148
Verified Email$_i$	-0.0002 (0.000) 0.000	-0.0004 (0.00) 0.000
+ week dummies		
Constant$_{ij}$	0.0004 (0.002) 0.000	0.0005 (0.000) 0.000
Observations	2,053,925	1,048,825
R^2	0.0004	0.0004
Robust standard errors in parentheses; P-values below		

3.7.3 Verbal Aggression Contagion

We find that both group prominence reduction and group banning produce verbal aggression spillovers into other political forums on Reddit. After group prominence reduction, r/The_Donald users increase verbal aggression within Reddit outside right-wing forums, as well as centrist forums. We assess contagion by examining the behavior of control users within these forums after being exposed to verbal aggression from a r/The_Donald user. We find very strong evidence that this exposure prompts reciprocal behavior: control users are significantly more likely to respond with verbal aggression after this exposure.

After the group ban, non-multihoming users increase verbal aggression only in Reddit right-wing forums. We find that exposure to non-multihomer verbal aggression within these spaces after the ban makes control users more likely to respond in kind. Multihomers produce a wider verbal aggression spillover: within

Reddit right-wing, but also left-wing forums. We also find that exposure to multihomer verbal aggression within these forums after the ban also prompts control user verbal aggression.

The tendency for aggressive behaviors to be contagious has been observed extensively in the offline (Wheeler & Levine 1967; Goldstein et al. 2001; Boxer et al. 2005; Warren et al. 2005) and online (Song & Wu 2018; Shen et al. 2020) contexts. This can happen because exposure to verbal aggression often prompts reciprocal behavior (Wheeler & Levine 1967), or through a social influence mechanism (Boxer et al. 2005). Thus, our results are consistent with prior literature.

However, our results deviate significantly from our previous work, which assessed the effect of group prominence reduction policies applied to misinformation (Mudambi et al. 2022). We found that group prominence reduction produced misinformation spillovers, but these were relatively short-lived due to a lack of contagion. This suggests that the unique behavioral dynamics of verbal aggression drive contagion once a spillover occurs. Taken in concert, these contrasting results show that platform-level outcomes can vary significantly based on the type of content these moderation policies are applied to.

3.7.4 Synthesis

We find that both group prominence reduction and group banning produce verbal aggression spillovers to ideologically similar forums. However, group banning produces a wider spillover: users increase verbal aggression in both ideologically similar and dissimilar forums. This wider spillover comes from the behavior of multihoming users alone. Our analysis indicates that having access to an external alternative leads these users to be more aggressive in ideologically dissimilar and potentially unfriendly forums, risking further sanctions. We can conclude that group content moderation applied to verbal aggression typically produces a spillover, and the more drastic the policy, the wider the spillover.

Furthermore, these observed verbal aggression spillovers produce verbal aggression contagion. Users who were not part of the problematic group receiving platform intervention are more likely to exhibit verbally aggressive behaviors after exposure to said spillovers. This result is consistent across both types of group content moderation policies (prominence reduction and banning), as well as across two types of user groups (multihomers and non-multihomers).

Given the significant verbal aggression spillovers, as well as ample evidence of contagion, we conclude that content moderation policies may do more harm than good when applied to groups high in verbal aggression. When factoring in prior results as to the effect of group content moderation upon misinformation (Mudambi et al. 2022), we conclude that the same content moderation policies applied to different types of harmful content produce significantly different results.

3.8 Conclusions & Future Work

Our work shows that the efficacy of content moderation policies, as well as associated externalities, vary significantly based on how they are deployed. We find that prominence reduction policies have the potential to decrease verbal aggression within directly affected spaces. However, both prominence reduction and banning produce

negative spillovers elsewhere on the platform. Our work suggests that utilizing the more drastic policy (banning) against a problematic group produces worse outcomes. We also find that multihoming users respond significantly differently to group content moderation than non-multihoming users. Finally, we observe significant verbal aggression contagion as a direct result of spillovers produced by the content moderation policies, which contrasts with previous work on misinformation. This leads us to conclude that group content moderation policies applied to groups high in verbal aggression are ineffective, and potentially very harmful. Taken together, this work shows that the same policies can produce significantly different outcomes based on the type of users and content that they are applied to.

This work has important implications for both scholars and practitioners. This work contributes to the literature on user-generated content platforms, content moderation policies, online antisocial behavior, platform multihomers, and social contagion. We contribute an understanding of how group prominence reduction and banning policies affect user verbal aggression, all while considering the effect of multihoming behavior. We empirically assess the potential for verbal aggression contagion as a result of content moderation policies and contextualize our findings with prior work on misinformation.

Our work has implications for platform administrators and designers as well. Our findings have the potential to directly inform the design of future content moderation policies. This work shows that group prominence reduction policies applied to groups high in verbal aggression produce spillovers, as well as the contagion of that behavior. Given this empirical evidence, we recommend that platforms utilize either individual or content-level policies to combat verbal aggression. Our results contrast to previous work studying misinformation (Mudambi et al. 2022), which produced significantly better outcomes. Thus, we recommend that platforms tailor their content moderation approaches based on the type of harmful content.

Platforms must consider the type of user, the type of harmful content, as well as the potential for negative spillovers and contagion when designing and deploying content moderation policies. Though the intention of content moderation policies is positive, merely enacting them is not enough. It is critical to empirically assess their efficacy, as well as point out associated negative and positive externalities to better inform future platform design.

Chapter 4: Isolating the Impact of Financial Rewards Upon Crowdsourcing Superstars

4.1 Introduction

Online platform users are not all created equal- some users have an outsized contribution to the platform, as well as other users. On user-generated content (UGC) platforms, it has been shown that contributors at the very top of the distribution create most of the content present on a platform (Lakhani & von Hippel 2003; Nielsen 2006). This translates to crowdsourcing platforms as well. High-ability and high-contribution users, known in this chapter and previous research as 'superstars', make the most contributions to a platform, and are greatly influential (Hayashi et al. 2021), positively contributing to other users' learning and future performance (Zhang et al. 2019). Furthermore, their presence attracts solution seekers to crowdsourcing platforms (Tauchert et al. 2020).

It is of great importance to crowdsourcing platforms to retain these superstar users. Crowdsourcing platforms use a mixture of features to incentivize users to participate at a high level within a challenge or contest. These can be financial (monetary rewards for performance), reputational (public rankings of performance, badges, points, and achievements), or learning (challenges give users an opportunity to learn a new skill). Previous research shows that users' performance, effort, and participation is positively influenced by all of these features (Zheng et al. 2011; Huang et al. 2012; Liu et. al 2014; Boons et al. 2015; Sun et al. 2015).

It is unclear, however, if winning a monetary reward on a crowdsourcing platform meaningfully impacts a superstar user's subsequent behavior. This research aims to address this. I ask: does winning a monetary reward significantly increase a superstar user's subsequent performance?

To answer this research question, I collect data from Kaggle, one of the foremost data science crowdsourcing competition platforms. I utilize a novel regression discontinuity design to directly compare users who win money to those who just miss out. As a primary result, I find that winning money has a significant positive impact on subsequent performance. This performance premium only exists for members who compete in groups (teams of size > 1).

This result is driven by changes in subsequent team composition. Users who win money are able to recruit more new users for the next contest. Additionally, they retain more users of their team. This boost in retention is not driven by differential rates of platform exit between money winners and non-money winners, but rather that non-money winners are significantly more likely to switch to competing as an individual after missing out on money. Due to these changes, users who win money have a significantly higher subsequent team size.

This research contributes to the literature on monetary rewards, crowdsourcing, and platform superusers. It additionally builds on previous literature by narrowing the focus and assessing the effect of winning a monetary reward on the future behavior of crowdsourcing superstars specifically. Superstars – defined in this research as high-success, high-contributing users - are disproportionately important to crowdsourcing platforms. They have been shown to contribute the vast majority of

content on platforms such as Kaggle, and have a high degree of influence with other users (Hayashi et al. 2021). Furthermore, they contribute positively to the learning of other users (Zhang et al. 2019). It is thus important to assess how winning monetary rewards affects these users specifically.

The rest of the paper proceeds as follows: I summarize related work, describe the study context and data collection, detail the empirical analysis, and present results and a discussion.

4.2 Related Work

4.2.1 Monetary Rewards

There are mixed results regarding the impact of offering monetary rewards on task effort. Performance-based monetary rewards have been shown to undermine intrinsic motivation (Murayama et al. 2010), or to have no effect on effort at all (Belle & Cantarelli 2015). To contrast, Choi et al. (2014) found that offering new hires signing bonuses positively affected worker effort, but the effect did not persist over time. There is also evidence that financial bonuses can increase worker motivation and effort when employees perceive manager allocating bonuses to have good judgement (Hewett & Leroy 2019).

Multiple studies have found that monetary rewards have a positive impact on task performance (Presslee et al. 2013; Kosfeld et al. 2017; Kralova & Kral 2019). Pay-for-performance incentives, however, have been found to restrict individuals' exploratory strategy when finding a solution, as well as decreased cooperation between solvers (Kralova & Kral 2019). There is research that suggests that the mechanism behind this increase in performance is that individuals set higher goals when there is potential for monetary rewards, and this goal-setting behavior in turn increases performance (Presslee et al. 2013).

4.2.2 Non-Monetary Incentives

Recognition as well as feedback have also been shown to have an impact on task performance (Hannan et al. 2008; Tran et al. 2012; Lourenco 2016). Some research suggests that the magnitude of performance increases from monetary incentives are similar to that from recognition alone (Lourenco 2016). Non-monetary rewards such as recognition can create significant pressure on individuals and subsequently impact their performance (Kali et al. 2017). Providing private feedback to individuals can improve their future performance (Tran et al. 2012). However, this effect may be heterogeneous; Hannan et al. (2008) found that individuals whose relative performance ranked low did worse after receiving this information. To contrast, individuals whose relative performance was high improved their performance after hearing their rank.

4.2.3 Application to Crowdsourcing

Within the context of crowdsourcing, monetary rewards have a significant impact upon worker effort and performance (Huang et al. 2012; Liu et. al 2014; Sun et al. 2015). Increasing the magnitude of a reward has been shown to significantly increase the number of submissions, as well as the overall quality of submissions (Huang et al.

2012; Liu et al. 2014). Though some research suggests that intrinsic motivation has a stronger effect on participation in crowdsourcing contests than extrinsic reward (Zheng et al. 2011), providing extrinsic incentives (such as money) does not necessarily counteract intrinsic motivation (Blashke et al. 2014). Finally, crowdsourcing participants' effort allocation increases when an individual becomes close to winning a monetary reward (Dissanayake et al. 2018).

On crowdsourcing platforms, a large number of solvers or existing solutions can lead to an average decrease in average user effort (Terwiesch et al. 2008; Huang et al. 2012) and performance (Boudreau et al. 201). This effect may be more pronounced for high-quality solvers (Liu et al. 2014). Despite this decrease in solver effort, solution seekers may benefit from a larger solver population. A larger population of users can create a more diverse set of solutions, which can outweigh this general decrease in effort (Terwiesch et al. 2008; Boudreau et al. 2011).

Various factors influence users' sustained participation on crowdsourcing platforms. Sustained participation can be positively influenced by experience with team members, the magnitude of monetary rewards, user tenure, and user feelings of pride regarding the platform (Boons et al. 2015; Wang et al. 2019; Khasragi et al. 2020). Winning a monetary reward may have an unintended effect, however: Bayus (2013) found that crowdsourcing participants with past success exhibited diminished creativity in subsequent contests.

Superstars, or users with performance and contribution at the very top of the distribution on a crowdsourcing platform, have a significant impact upon other users. As with most other online platforms (Lakhani & von Hippel 2003; Nielsen 2006), crowdsourcing platforms rely on a relatively small number of influential users with a high degree of contribution (Hayashi et al. 2021). The presence of superstars within a contest may make other participants less likely to join (Archak 2010; Boudreau et al. 2011). However, individuals who do join contests with superstars present have a much higher probability of winning subsequent contests. This may be due to the fact that in many contests, winners' solutions are made public and other users can learn from them (Zhang et al. 2019).

Previous research has shown that team formation decisions on crowdsourcing platforms are strongly driven by homophily. Individuals prefer to work with other users who speak the same language, live in the same or neighboring countries, and have comparable skills (Dissanayake et al. 2019). However, team diversity, particularly in platform tenure and subject matter interests has been shown to positively influence performance (Ren et al. 2016; Ortu et al. 2017; Dissanayake et al. 2019). As these teams do not necessarily interact in person, users' online profiles are often used to glean information about prospective team members. Team composition tends to change significantly from contest to contest, as the exit and entry barriers of teams are low on online platforms (Ren et al. 2016).

4.2.4 Research Gap

Various papers examine the impact that monetary rewards have on crowdsourcing platform users' effort and performance (Huang et al. 2012; Liu et. al 2014; Sun et al. 2015; Dissanayake et al. 2018). Evidence strongly suggests that the presence of monetary reward, as well as its magnitude, increases effort allocation and task

performance for solvers in general. However, some research shows that the magnitude of performance increases from monetary incentives are similar to that from recognition alone (Lourenco 2016). Receiving a performance review can increase task performance, but it is possible that this can be diminished when an individual also receives a monetary reward (Manthei et al. 2019). It is unclear if winning a monetary reward on a crowdsourcing platform has a significantly larger impact than the other benefits (recognition, learning, positive task feedback) that come from being a high performer.

Finally, this research focuses wholly on crowdsourcing superstars. Previous work has focused on superstars' impact on the number of submissions (Archak 2010), the degree of effort expended by other participants in a contest (Boudreau et al. 2011), as well as the performance outcomes of other participants (Zhang et al. 2019). However, this research focuses on how contest design affects superstars alone. This is an important gap, as superstars have an outsized impact on the number and quality of solutions (Hayashi et al. 2021), as well as the learning of other users (Zhang et al. 2019). It is important for all stakeholders to understand how to retain these users, and how prize structure affects their subsequent effort and performance.

4.3 Study Context

Kaggle is a platform dedicated to data analytics that was established in 2010. The platform offers its users educational services such as coding classes, discussion forums, and free code and datasets for users to practice on. However, this research focuses on Kaggle's data analytics contests, in which firms, government organizations, and researchers provide problem descriptions (and the relevant data) for Kaggle users to solve in a contest[15]. Users compete in teams that have anywhere from one to forty members.

After submitting a model to a contest, teams are ranked and evaluated based on the performance of their model in real time, using a fraction of the testing dataset. (The full testing dataset is not available to contest entrants to properly assess models' external validity.) These interim rankings are displayed on a public leaderboard. When the contest concludes, models are evaluated using the entire testing dataset, and final user rankings are displayed on what is called the private leaderboard. It is common for users to see a significant drop or ascent in their ranking between the public leaderboard (visible throughout the entire contest) and the private leaderboard (only visible at the end of the contest, displays the final entrant rankings).

Some contests provide monetary rewards for a certain number of top performers. In terms of non-monetary incentives, users can earn competition points based their contest performance. High performers (who may not necessarily win a monetary reward) can also earn medals. Competition points are used to create a platform-wide numerical ranking and are also used to assign users into five different

[15] An illustrative example of a contest on Kaggle would be the JPX Tokyo Stock Exchange Prediction contest (https://www.kaggle.com/competitions/jpx-tokyo-stock-exchange-prediction/), in which participants must build a model to predict the expected return of a sample of stocks from the Japanese market.

tiers (Novice, Contributor, Expert, Master, Grandmaster). These elements are made quite salient on a user's profile, an example of which is below in Figure 6.

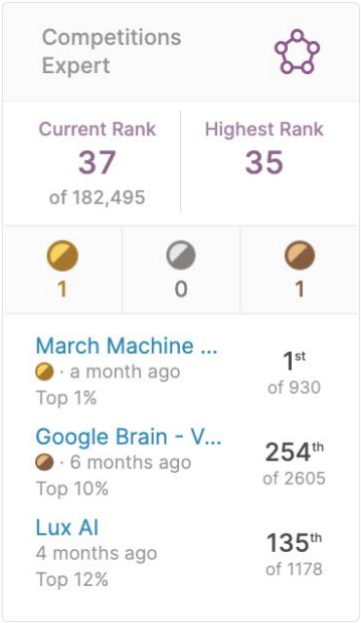


Figure 6. Kaggle Profile

4.4 Data

Archival data was collected from Kaggle.com, one of the largest data analytics crowdsourcing platforms. Kaggle provides its historical data publicly via the Kaggle Meta dataset[16]. Data was collected for all platform contests: contest features as well as their public and private leaderboards. Additionally, user profile data, as well as user contest performance data was collected.

4.4.1 Variable Construction

For each user-contest observation within our dataset, the following variables will be constructed.

4.4.1.1 Treatment Variable: Won Money

The purpose of this study is to ascertain the effect of winning a monetary prize on subsequent user behavior on a crowdsourcing platform. Each user-contest tuple within our dataset is assigned a dummy variable corresponding to 1 if they received a monetary prize for their performance in a contest, or 0 if not. This is the core treatment variable.

This variable is constructed taking into account a user's final contest rank, as well as the number of prizes offered. The final contest rank is assigned by ranking all of the models by performance on the entire dataset in descending order. If a user contributed to a model, they are assigned that rank. As multiple users can work on the

[16] https://www.kaggle.com/datasets/kaggle/meta-kaggle

same model, this means that there can be multiple users assigned the same final contest rank within our dataset.

Each contest has a fixed number of top models for which they award monetary prizes. If the user's final contest rank is less than or equal to that fixed number, they have won money. For example, if a contest awards monetary prizes for the top 2 models, if a user's model is ranked second or first, they win money.

4.4.1.2 Running Variable: Weighted Contest Rank

As the number of prizes offered varies by contest, each user's final contest rank must be weighted to generalize the analysis. For all contests, I set the cutoff for winning money to zero. Thus, for user *i* competing in contest *j*, the weighted rank is calculated as follows:

$$WeightedContestRank_{ij} = -1 * \left(FinalContestRank_{ij} - NumPrizes_j\right) \text{ if } FinalContestRank_{ij} > NumPrizes_j$$

$$WeightedContestRank_{ij} = \left(FinalContestRank_{ij} - NumPrizes_j\right) + 1 \text{ if } FinalContestRank_{ij} \leq NumPrizes_j$$

4.4.1.3 Moderating Variables

User Contest Experience

The number of contests user *i* in contest *j* has competed in (including contest *j*).

4.4.1.4 Dependent Variables

User Subsequent Performance

Contest performance for user *i* in contest *j* is calculated using the below equation. This corresponds to each user's performance percentile in each contest. Each user's subsequent performance is their performance in the next contest they enter chronologically.

$$ContestPerformance_{ij} = 1 - \frac{FinalContestRank_{ij}}{Maximum\ Rank_j}$$

Subsequent Group Size

This is the total number of users present in the user's group for the subsequent contest.

Likelihood of Not Competing Again

Whether the user exits the platform and does not enter a subsequent contest.

Likelihood of Switching to Individual

Whether the user (who competed in a team in the focal contest) competes as an individual in the subsequent contest.

Number of New Users Added

The number of new users (users who were not in the user's team for the focal contest) added to the user's team for the subsequent contest.

Number of Users Retained
The number of users who were in the user's team for the focal contest that remain in the user's team for the subsequent contest.

Likelihood of Subsequent Contest Monetary
Whether the user's subsequent competition offers monetary rewards.

Subsequent Contest Reward Quantity
The total prize money amount (ranging from 0 to 1,500,000 USD) awarded in the user's subsequent contest.

Subsequent Contest Similarity
The cosine text similarity between the user's focal contest and subsequent contest descriptions.

Relative User Effort
Within each contest, each user can submit as many models as they like, to improve upon their score. This has been used as a proxy for effort in many previous studies (Huang et al. 2012; Liu et. al 2014; Sun et al. 2015; Dissanayke et al. 2018). For users competing in groups (teams of size >1), any member can submit a model. To assess each team member's effort relative to the group, their effort is the number of contest submissions they personally submitted divided by the total number of submissions submitted by the team.

4.5 Empirical Strategy

The empirical strategy used in this paper takes advantage of the quasi-random rank cutoffs used to grant monetary rewards in Kaggle contests. This strategy hinges on the fact that users cannot precisely manipulate their final rank in the contests. Users should be similar in their pre-intervention characteristics around these rank cutoff points. Thus, I utilize a regression discontinuity design to estimate the effect of receiving a monetary prize on subsequent user behavior. The unit of analysis is each user-contest tuple in our dataset.

It is critical to be certain that users cannot precisely manipulate their ultimate rank in the contests. If this were the case, users could make small improvements to their models to ensure that they fall on the appropriate side of a cutoff to receive money. This process of self-selection could mean that the users on either side of a cutoff point are not similar. However, there are multiple reasons that suggest that this sort of manipulation is impossible within the context of our study.

First, the contestants train and optimize their models only using a small percentage of testing data; final scores are calculated using the entire dataset and are kept private until the end of the contest. It is impossible for a user to know with certainty what their ultimate score will be. Second, to manipulate their rank, users would need to be able to accurately forecast the scores of all their rivals within a

contest, as the rankings depend on relative performance within all entrants. Since a user's ultimate score is unknown, as well as the ultimate scores of their rivals, this means that they cannot know with certainty where they will fall within the rankings before a contest concludes. Therefore, the identifying assumption of no precise manipulation near a cut-off point is plausible in our study.

Within our dataset, the cutoff points vary for monetary reward. Some contests only award monies to the top-performing model; others award money to as many as the top ten. Within each contest that awards monetary rewards to the top p models, users who contribute to the top (p*2) models are included in subsequent empirical analyses. This bandwidth ensures an equal number of models above and below the monetary cutoff are analyzed for each contest. This is illustrated below in Figure 7; the contest awarded monetary prizes to the top three models, so the members of the top six teams are included in the analyses. After constructing this dataset, I run a density discontinuity test on the dataset which shows no evidence of manipulation around cutoff points ($p < 0.3592$). This supports the assertion that individuals above and below the cutoffs for earning money are similar and can be directly compared.

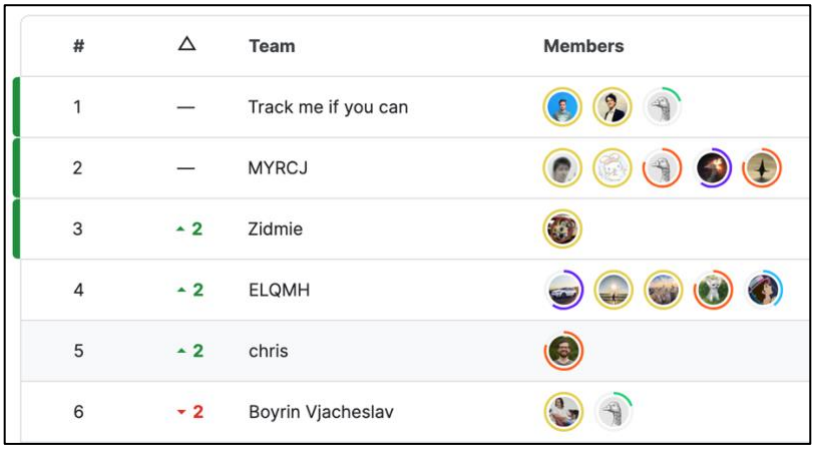


Figure 7. Contest Bandwidth Selection Example

The generalized specification for the regression discontinuity analysis is below for user i in contest j. The focus of this work is to assess the impact of winning a monetary reward in contest j on user i's subsequent behavior. Thus, the dependent variables of interest correspond to user i's behavior in the subsequent (j +1) contest. I run ordinary least squares regressions with robust standard errors clustered by each contest utilizing this specification to obtain results.

$$DependentVariable_{ij+1} = WonMoney_{ij} + WeightedContestRank_{ij} + \varepsilon_{ij}$$

4.6 Results

4.6.1 Subsequent User Performance

I first model each user's subsequent contest performance. The coefficient estimates are shown below in Table 17a. I find evidence that users who perform well enough to win money perform significantly better than other high-performing users. The

coefficient on the *Won Money*$_{ij}$ coefficient is positive (0.038; corresponding to a 3.8% premium in subsequent contest rank) and significant ($p<0.016$). This shows that financial rewards significantly improve subsequent user performance in crowdsourcing contests.[17]

This is borne out by descriptive data. On average, both users who win money and users who miss out experience a drop in performance for the subsequent competition. This can be attributed to regression to the mean. However, users who win money experience a smaller drop: they drop their performance by 19.3% on average, as compared to users who miss out, who drop it by 21.8%. This is evidenced below in Figure 8. The jump in performance at the monetary cutoff also justifies the regression discontinuity approach.

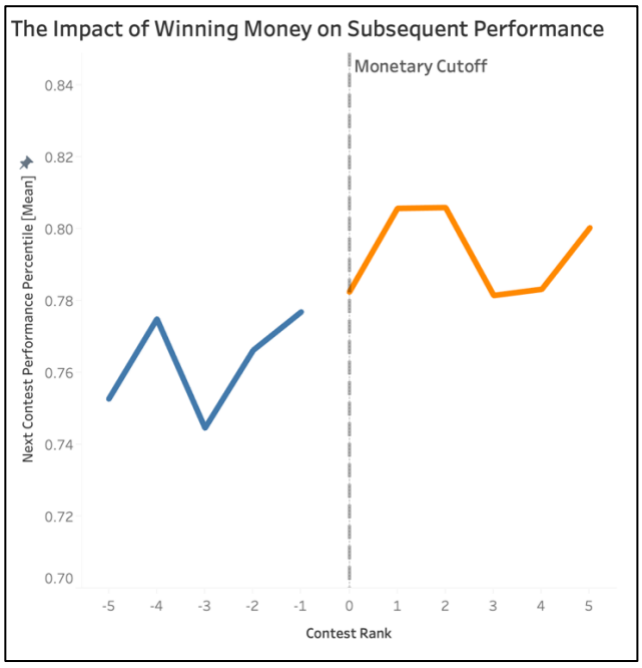


Figure 8. Descriptive Evidence of Performance Premium

I also look at moderating effects. To assess how each user's competition experience affects their response to winning money, I interact each user's competition number into the specification. The results are found in Table 17b. I find that the performance premium decreases with the user's competition experience. However, when assessing if the number of places above or below a monetary cutoff impact the magnitude of the performance premium, no significant result is found. This is available in Table 17c.

To assess the source of this performance premium, I partition the dataset between users who competed as individuals (teams of size 1) and users who competed in groups (teams of size > 1) and re-run the analysis. The results for users who competed in groups (individuals) are available in Table 18 (19). These show that users who competed in groups see a significant increase in subsequent performance

[17] Non-parametric regression discontinuity was carried out as a robustness check; it found that users that win money increase their subsequent performance by 2.97 percent ($p < 0.008$).

after winning money (0.058; $p < 0.003$). However, users who competed individually do not see a significant increase in subsequent performance (-0.009; $p < 0.764$). This indicates that the subsequent boost in performance after winning money comes from a group-specific process.

Table 17a. Effect of Winning Money on Subsequent User Performance

Subsequent User Performance	**Coefficient**	**Robust Standard Error**	**t**	**P>|t|**	**95% Confidence Interval**	
Won Money$_{ij}$	0.038	0.016	2.42	0.016	0.007	0.069
Weighted Contest Rank$_{ij}$	-0.002	0.002	-0.91	0.366	-0.007	0.003
Constant$_{ij}$	0.760	0.010	73.51	0.000	0.740	0.781
N: 3,417						

Table 18b. Effect of Winning Money on Subsequent User Performance, Interacting Contest Experience

Subsequent User Performance	**Coefficient**	**Robust Standard Error**	**t**	**P>|t|**	**95% Confidence Interval**	
Won Money$_{ij}$ X Contest Experience$_{ij}$	-0.001	0.0.001	-1.82	0.069	-0.002	0.000
Won Money$_{ij}$	0.049	0.017	2.84	0.005	0.015	0.084
Contest Experience$_{ij}$	0.001	0.000	2.83	0.005	0.000	0.002
Weighted Contest Rank$_{ij}$	-0.002	0.002	-0.84	0.401	-0.007	0.003
Constant$_{ij}$	0.750	0.012	62.86	0.000	0.724	0.771
N: 3,417						

Table 19c. Effect of Winning Money on Subsequent User Performance, Interacting Weighted Contest Rank

Subsequent User Performance	**Coefficient**	**Robust Standard Error**	**t**	**P>|t|**	**95% Confidence Interval**

Won Money$_{ij}$ X Weighted Contest Rank$_{ij}$	-0.001	0.007	-0.21	0.069	-0.002	0.000
Won Money$_{ij}$	0.038	0.016	-0.31	0.757	-0.010	0.007
Weighted Contest Rank$_{ij}$	-0.001	0.005	-0.31	0.757	-0.010	0.007
Constant$_{ij}$	0.762	0.014	55.64	0.000	0.736	0.789
N: 3,417						

Table 20. Effect of Winning Money on Subsequent User Performance, Groups Only

Subsequent User Performance, Groups Only	**Coefficient**	**Robust Standard Error**	**t**	**P>|t|**	**95% Confidence Interval**	
Won Money$_{ij}$	0.058	0.19	3.00	0.003	0.020	0.0957
Weighted Contest Rank$_{ij}$	-0.004	0.003	-1.64	0.101	-0.009	0.009
Constant$_{ij}$	0.756	0.013	58.53	0.000	0.731	0.782
N: 2,377						

Table 21. Effect of Winning Money on Subsequent User Performance, Individuals Only

Subsequent User Performance, Groups Only	**Coefficient**	**Robust Standard Error**	**t**	**P>|t|**	**95% Confidence Interval**	
Won Money$_{ij}$	-0.009	0.028	-0.30	0.764	-0.064	0.047
Weighted Contest Rank$_{ij}$	0.003	0.005	0.60	0.550	-0.007	0.12
Constant$_{ij}$	0.769	0.017	45.99	0.000	0.736	0.802
N: 1,040						

4.6.2 Subsequent Team Composition

To ascertain the mechanism of the subsequent performance premium associated with winning money, I run analyses assessing the impact of winning money on team composition. The analyses detailed in this section are run only on users who compete in groups (teams of size > 1), as no performance premium was observed for users who compete as individuals. First, I model the team size for the subsequent competition for each user. The coefficient estimates are below in Table 20. Users who perform well enough to win money have a subsequent team size of 0.666 greater than other high-performing users ($p < 0.000$). This shows that users who win financial rewards are increasing their team size as a result.

Table 22. Effect of Winning Money on Subsequent Team Size

Subsequent Team Size	**Coefficient**	**Robust Standard Error**	**t**	**P>|t|**	**95% Confidence Interval**	
Won Money$_{ij}$	0.666	0.189	3.53	0.000	0.295	1.038
Weighted Contest Rank$_{ij}$	0.022	0.066	0.33	0.740	-0.107	0.151
Constant$_{ij}$	2.128	0.182	11.72	0.000	1.771	2.486
N: 2,377						

Users competing in groups have many potential options after completing a competition: they can leave the platform and cease competing, leave their team and compete as an individual, join another group (team of size > 1), or remain in their team. To assess what is driving the boost in team size that money-winners experience, I examine the impact a financial reward has on all these outcomes for users competing in groups.

First, I model the likelihood of a user leaving the platform and ceasing competition. The coefficient estimates are below in Table 21. Users who perform well enough to win money are not significantly more likely to not compete again than other high-performing users ($p < 0.115$). Though this significance is close to $p < 0.10$, the coefficient is positive ($\beta = 0.055$), indicating that users who win monetary rewards would be more likely to cease competing. Platform exit thus does not explain monetary winners' larger subsequent team sizes.

Next, I model the likelihood of a user leaving their team and competing next as an individual. The coefficient estimates are below in Table 22. Users who perform well enough to win money are 9.73% ($p < 0.014$) less likely to subsequently compete as an individual than other high-performing users. As users who won money are

significantly less likely to compete as individuals subsequently, this contributes to their subsequent higher team sizes.

I then assess the impact of winning money on recruiting new members to a team. I model the number of individuals added to the user's team for the subsequent competition. The coefficient estimates are below in Table 23. Users who perform well enough to win money add 0.245 ($p < 0.031$) greater members than other high-performing users to their team. As users who won money are recruiting more individuals to join their subsequent team, this contributes to their subsequent higher team sizes.

Finally, I assess the impact of winning money on how many team members a user is able to retain. I model the number of individuals retained within the user's team for the subsequent competition. The coefficient estimates are below in Table 24. Users who perform well enough to win money retain 0.421 ($p < 0.003$) greater members than other high-performing users in their team. As users who won money are retaining more individuals in their subsequent team, this also contributes to their subsequent higher team sizes.

Table 23. Effect of Winning Money on the Likelihood of Not Competing Again

Likelihood of Not Competing Again	**Coefficient**	**Robust Standard Error**	**t**	**P>|t|**	**95% Confidence Interval**	
Won Money$_{ij}$	0.055	0.035	1.58	0.115	-0.014	0.124
Weighted Contest Rank$_{ij}$	-0.007	0.005	-1.43	0.154	-0.017	0.003
Constant$_{ij}$	0.214	0.023	9.48	0.000	0.169	0.258
N: 2,377						

Table 24. Effect of Winning Money on the Likelihood of Switching to Individual

Likelihood of Switching to Individual	**Coefficient**	**Robust Standard Error**	**t**	**P>|t|**	**95% Confidence Interval**	
Won Money$_{ij}$	-0.097	0.039	-2.48	0.014	-0.175	-0.20
Weighted Contest Rank$_{ij}$	-0.005	0.005	1.01	0.315	-0.005	0.016
Constant$_{ij}$	0.555	0.023	23.87	0.000	0.509	0.600
N: 2,377						

Table 25. Effect of Winning Money on the Number of Users Added

Number of Users Added	**Coefficient**	**Robust Standard Error**	**t**	**P>|t|**	**95% Confidence Interval**	
Won Money$_{ij}$	0.245	0.113	2.17	0.031	0.023	0.468
Weighted Contest Rank$_{ij}$	-0.012	0.160	-0.72	0.470	-0.043	0.020
Constant$_{ij}$	0.675	0.069	9.73	0.000	0.538	0.812
N: 2,377						

Table 26. Effect of Winning Money on the Number of Users Retained

Number of Users Retained	**Coefficient**	**Robust Standard Error**	**t**	**P>|t|**	**95% Confidence Interval**	
Won Money$_{ij}$	0.421	0.142	2.95	0.003	0.140	0.701
Weighted Contest Rank$_{ij}$	0.033	0.064	0.53	0.600	-0.092	0.158
Constant$_{ij}$	1.453	0.172	8.45	0.000	1.115	1.792
N: 2,377						

4.7 Robustness Checks

Previous analyses indicate that the observed increase in subsequent performance for money-winners can be tied to subsequent changes in team composition. However, there are many alternative explanations for the increase in performance. In this section, I assess and rule out the impact of subsequent user effort and subsequent competition selection on this result.

4.7.1 Subsequent User Effort

First, I ascertain if money winners' increased performance is due to a subsequent increase in effort. This analysis is run on all users, including those who competed individually and those who competed in groups. The result of this analysis is available below in Table 25. I find evidence that users who perform well enough to win money significantly decrease their effort in the subsequent contest. The coefficient on the *Won Money$_{ij}$* coefficient is negative (-0.081; corresponding to an 8.1% drop in the percentage of submissions made for their team) and significant

(p<0.000). This shows evidence that financial rewards have a negative impact on user effort. The observed increase in subsequent contest performance cannot be attributed to an increase in effort.

Table 27. Effect of Winning Money on Subsequent User Effort

Subsequent User Effort	**Coefficient**	**Robust Standard Error**	**t**	**P>|t|**	**95% Confidence Interval**	
Won Money$_{ij}$	-0.081	0.023	-3.55	0.000	-0.126	-0.036
Weighted Contest Rank$_{ij}$	0.002	0.004	0.57	0.570	-0.005	0.009
Constant$_{ij}$	0.777	0.013	57.58	0.000	0.751	0.804
N: 3,417						

4.7.2 Subsequent Competition Type

Next, I ascertain if money winners' increased performance is due to them selecting different competitions than users who miss out on a financial reward. To do this, I model various attributes of a user's subsequent competition: whether it offers monetary rewards, the quantity of monetary reward offered (ranging from 0 to 1,500,000 USD), and its similarity to the previous competition. The coefficient estimates are shown below in Tables 26-28.

The coefficients for subsequent competitions offering monetary rewards ($p < 0.830$), monetary reward quantity ($p < 0.219$), and similarity[18] ($p < 0.218$) were all insignificant. Thus, there is no evidence that winning a monetary reward affects a user's choice of subsequent competition.

Table 28. Effect of Winning Money on the Likelihood of Choosing a Monetary Subsequent Competition

Subsequent Competition Monetary	**Coefficient**	**Robust Standard Error**	**t**	**P>|t|**	**95% Confidence Interval**	
Won Money$_{ij}$	-0.003	0.013	-0.22	0.830	-0.0286	0.023
Weighted Contest Rank$_{ij}$	0.001	0.002	0.63	0.531	-0.003	0.005

[18] This analysis is only run on 3,400 user-contest tuples instead of 3,417. The cosine similarity was not able to be calculated for 14 contests as their descriptions contained only stop words.

$Constant_{ij}$	0.939	0.008	112.98	0.000	0.923	0.956
N: 3,417						

Table 29. Effect of Winning Money on Subsequent Competition Reward Quantity

Subsequent Competition Reward Quantity	**Coefficient**	**Robust Standard Error**	**t**	**P>|t|**	**95% Confidence Interval**	
$Won\ Money_{ij}$	12,444.79	10,108.10	1.23	0.219	-7,346.28	32,325.85
Weighted Contest $Rank_{ij}$	1,193.73	1,759.08	0.68	0.498	-2,266.11	4,653.57
$Constant_{ij}$	52,777.49	5,958.80	8.86	0.000	41,057.47	64,497.51
N: 3,417						

Table 30. Effect of Winning Money on Subsequent Competition Similarity

Subsequent Competition Similarity	**Coefficient**	**Robust Standard Error**	**t**	**P>|t|**	**95% Confidence Interval**	
$Won\ Money_{ij}$	0.010	0.008	1.23	0.218	-0.006	0.027
Weighted Contest $Rank_{ij}$	0.001	0.001	0.97	0.331	-0.001	0.004
$Constant_{ij}$	0.427	0.007	57.44	0.000	0.413	0.442
N: 3,400						

4.8 Summary of Results & Discussion

As a primary result, I find that winning money has a significant positive impact on subsequent performance. This performance premium only exists for members who compete in groups (teams of size > 1). This result is driven by changes in subsequent team composition. Users who win money are able to recruit more new users for the next contest. Additionally, they retain more users of their team. This boost in retention is not driven by differential rates of platform exit between money winners and non-money winners, but rather that non-money winners are significantly more likely to switch to competing as an individual after missing out on money. Due to these changes, users who win money have a significantly higher subsequent team size.

I propose that users are able to recruit more new members due to a signaling mechanism (Spence 1978). The money winners of each contest are highly salient

through the leaderboards, which the Kaggle platform makes public. Furthermore, users' past contest participation is made public on their profiles (see Figure 6). Thus, potential new team members seek out past money-winning teams to join.

Money winners are able to retain more members on their team as well, as users who just missed out on money are significantly more likely to leave their team to compete as individuals. Though competing in a group (team of size > 1) has its advantages, it also has the disadvantage that users have to split any potential monetary reward (Ren et al. 2016). Just missing out on money may be de-motivating and prompt users to lose confidence in their team.

Previous research shows that a higher team size is positively correlated to performance on team performance (Girotra et al. 2010). The above changes in team composition lead to monetary winners having higher team sizes in subsequent competitions. It is highly likely that this contributes to their subsequent higher performance. This is further supported by robustness checks that rule out an increase in effort or differences in subsequent competition selection as driving the performance differential.

4.9 Concluding Remarks

This research contributes to the literature on monetary rewards, crowdsourcing, and platform superusers. Much of the prior work examining monetary rewards and crowdsourcing focuses on the impacts of the presence of monetary rewards on general user performance and effort within a contest (Huang et al. 2012; Liu et. al 2014; Sun et al. 2015); Dissanayake et al. 2018). There is a lack of consensus, however, regarding the impact of winning a monetary reward on future behavior, and whether financial gain meaningfully shifts user behavior in the presence of significant reputational and learning benefits (Archak & Ghose 2010; Wang et al. 2019).

My research builds on previous literature by narrowing the focus and assessing the effect of winning a monetary reward on the future behavior of crowdsourcing superstars specifically. Superstars – defined in this research as high-success, high-contributing users - are disproportionately important to crowdsourcing platforms. They have been shown to contribute the vast majority of content on platforms such as Kaggle, and have a high degree of influence with other users (Hayashi et al. 2021). Furthermore, they contribute positively to the learning of other users (Zhang et al. 2019). It is thus important to assess how winning monetary rewards affects these users specifically.

This work has potential to generate findings of great significance to both researchers and practitioners. Understanding the impact that winning a monetary reward has on the future behavior of superstars could illuminate how to optimize contest design to ensure a continued level of high performance, effort, and participation.

Appendices

Appendix 1: Anti-Vaccine Misinformation Typology

No.	Myth	Total Posts	Scientific Consensus	Sources in Literature	Selected ?
1	Vaccines cause autism	62	Vaccines do not cause autism.	Kata 2010; Kata 2012	Y
2	Natural remedies (homeopathy, plant-based nutritional supplements, vitamins, injecting or ingesting disinfectant, essential oils) are more effective for treating illnesses addressed by vaccines.	43	Natural remedies are not effective against vaccine-preventable diseases.	Kata 2010; Kata 2012	Y
3	Vaccines contain toxic additives that are injurious to health	29	Vaccines contain additives, but at a trace level which is not toxic	Kata 2010; Kata 2012	Y
4	Measles is a harmless childhood disease and/or measles can be beneficial to children.	23	Measles can cause serious complications and lead to death and is not beneficial to children.	Kata 2010	Y
5	Those supporting vaccines do so because they are hired by pharmaceutica	22	There is significant scientific evidence for vaccines beyond	Kata 2010; Kata 2012	Y

	l companies/ part of a medical/pharmaceutical/government conspiracy		pharmaceutical companies' financial motives		
6	Vaccines are carcinogenic	14	Vaccines do not cause cancer	Kata 2010	Y
7	Vaccines are medically unnecessary	10	Vaccines are necessary to prevent many preventable diseases; their efficacy has been empirically demonstrated many times	Kata 2010	Y
8	Vaccines cause sudden infant death syndrome (SIDS)	4	Vaccines do not cause SIDS; SIDs has been significantly reduced in recent years due to an uptake in safe sleeping practices	Kata 2010	Y
9	Vaccines cause celiac disease	2	Vaccines do not cause celiac disease	Offitt & Hackett 2003	Y
10	Vaccines cause asthma	2	Vaccines do not cause asthma	Offitt & Hackett 2003; Kata 2010	Y
11	Face masks are ineffective in preventing the spread of respiratory illnesses	1	Masks are effective in preventing the transfer of respiratory illnesses	Ayers et al. 2021	Y
12	Vaccines cause multiple sclerosis	1	Vaccines do not cause	Offitt & Hackett	Y

			multiple sclerosis	2003; Kata 2010	
13	The flu vaccine is unnecessary and ineffective	21	The flu vaccine's efficacy varies significantly year-to-year due to the high degree of mutation in the virus	Guidry 2020	N
14	Acute vaccine allergic reactions are underreported	16	Acute vaccine allergic reactions occur and are tracked using the CDC's VAERS database, but its accuracy is difficult to assess.	Kata 2010	N
15	Vaccines cause general health problems later in life	14	No scientific consensus as myth is quite vague and wide-reaching	Kata 2010	N
16	The polio vaccine causes polio, not the virus.	8	The oral polio vaccine has caused polio in the past, but the inactive polio vaccine (which is more commonly used today) does not.	Kata 2010	N
17	Vaccines cause allergies	4	Vaccines do not cause chronic allergies, but have caused	Offitt & Hackett 2003; Kata 2010	N

			allergic reactions		
18	Vaccines are grown in aborted fetal tissue	2	Some vaccines have been grown in fetal tissue in the past.	Kata 2010	N
19	Diseases have declined from improved hygiene, not vaccines	2	Vaccines have significantly declined the prevalence of many illnesses, and improved hygiene has also played a role	Kata 2010	N

Appendix 2: Anti-Vaccine Myth Keywords

No.	Myth	Associated Keywords
1	Vaccines cause autism	autism, autistic, autist, aspergers, asperger, asperger's, aspie, spectrum, brain injury, brain damage, neurological damage, encephalopathy, childhood disintegrative disorder, disintegrative disorder, rett's disorder, retts disorder, retts, rett's, pervasive developmental disorder, developmental disorder, developmental disability, developmental disabilities, developmental delay, high-functioning, high functioning, regression Acronyms: ASD, AD, CDD, RD, PDD, DD, HFA
2	Natural remedies (homeopathy, plant-based nutritional supplements, vitamins, injecting or ingesting disinfectant, essential oils) are more effective for treating illnesses addressed by vaccines.	naturopath, essential oil, oil, homeopath, homeopathy, lavender, vitamin, supplement, disinfectant, holistic, natural remedy, alternative medicine, vegan, materia medica, tincture, flower essences, wellness

3	Vaccines contain toxic additives that are injurious to health	antifreeze, anti-freeze, ethylene glycol, methanol, propylene glycol, deicer, de-icer, ether, diethyl oxide, formaldehyde, mercury, quicksilver, hydrargyrum, nano-bacteria, nanobacteria, toxic, toxin, poison, contaminant, contaminate, contamination, adjuvant, additive, preservative, thimerosal, aluminum, alum, methanal, methylene oxide, oxymethyline, methylaldehyde, oxomethane, metal, metals, merthiolate, sodium ethylmercurithiosalicylate, AS01, AS04, CpG 1018, MatrixM, MF59, squalene
4	Measles is a harmless childhood disease and/or measles can be beneficial to children.	Measles, MMR, morbilli, rubeola, maculopapular, coryza, conjunctivitis
5	Those supporting vaccines do so because they are hired by pharmaceutical companies/ part of a medical/pharmaceutical/government conspiracy	conspiracy, shill, CDC, payments, financial, money, fraud, big pharma, collusion, conspiracist, hoax, collude, colluding
6	Vaccines are carcinogenic	cancer, cancerous, carcinoma, leukemia, lymphoma, mass, metastasis, metastatic, tumor, carcinogen, malignant, malignancy, polyp, benign, precancerous, sarcoma, nodule
7	Vaccines are medically unnecessary	Unnecessary, unneeded, useless, worthless, needless, futile, inessential, redundant, superfluous
8	Vaccines cause sudden infant death syndrome (SIDS)	sudden infant death syndrome, crib death, cot death, sudden unexpected infant death, accidental suffocation and strangulation in bed Acronyms: SIDS, SUID, ASSB
9	Vaccines cause celiac disease	Celiac, gluten, sprue, Dermatitis herpetiformis
10	Vaccines cause asthma	Asthma, inhaler, HFA, leukotriene modifier, MDI, theophylline, spacer
11	Face masks are ineffective in preventing the spread of respiratory illnesses	Mask, KN95, N95, respirator
12	Vaccines cause multiple sclerosis	MS, multiple sclerosis, myelin sheath, lhermitte, sclerosis

Appendix 3: Misinformation Classification & Verification Protocols

A3.1 Parent Post Labels

Parent posts are thread conversation starter posts and level-one replies. These were manually labeled as to their stance towards misinformation. We randomly selected 1,006 posts from our dataset of 43,484 manually coded parent posts. To validate these manual labels, we provided ChatGPT-3.5 with a prompt and asked it to classify each post as agrees with misinformation, disagrees with misinformation, neutral to misinformation, or unrelated to misinformation. On our random sub-sample, we achieved an accuracy of 90.7% when comparing the ChatGPT labels to our manual labels.

A3.2 Reply Post Labels

We provided ChatGPT-3.5 with a Reddit parent post with its stance towards misinformation annotated, and all of its responses in a conversation branch. ChatGPT then parsed the conversation branch and coded each reply as 'agrees with misinformation' or not. In total, ChatGPT coded 29,518 replies and yielded 2,094 'agrees with misinformation' replies. To validate these ChatGPT labels, two human coders coded 1,029 randomly selected replies in conversation branches. Where the two human coders disagreed, a third coder assigned a label as a tiebreaker. Given these human-produced labels, the ChatGPT model had an accuracy of 90.0% on the random sample.

Appendix 4: Non-Quarantined Anti-Vax Forum: Reasoning for Exclusion from Analyses

We included one non-quarantined anti-vax forum, r/ThingsProVaxxersSay, within our dataset as the forum was centered around vaccine-related discussion. However, we chose to exclude this forum from our user-level analyses as the results were likely to be misleading and not representative of general trends in user behavior. The reasons for this are twofold.

First, we assess the activity level of the non-quarantined forum. This forum was created in March of 2019, but never reached a critical mass of users that could sustain prolonged activity. This is evidenced below in Figures 8 and 9, which compare activity levels between the quarantined anti-vax forums and the non-quarantined anti-vax forum. Figure 8 shows that by the quarantine date (May 23, 2019), the flow of new users to the non-quarantined forum had dropped to zero. Additionally, Figure 9 shows that misinformation posts had also gone to zero by the time of quarantine. Both trends are in opposition to trends in the quarantined anti-vax forums. This descriptive evidence shows that the anti-vax non-quarantined forum had ceased to be active by the time of quarantine. For this reason, it was an unlikely site for any spillover effects or changes in user behavior.

Second, we assess the type of discourse present in the anti-vax non-quarantined forum during the baseline period. A qualitative and descriptive analysis of the topics present in the non-quarantined anti-vax forum indicates that posts are

mainly focused on interpersonal conflicts between anti- and pro-vaxxers rather than debate around misinformation topics. The anti-vax non-quarantined forum had a significantly lower (Difference: - 11.91%; $p<0.000$) percentage of conversation threads with misinformation-relevant keywords than quarantined anti-vax forums. To contrast, neutral-vax forums had a percentage that was not significantly different than in quarantined anti-vax forums (Difference: -2.00%; $p<0.373$). When looking at the correlation between a post being misinformation and its karma (a Reddit measure for how positively a post is received), we find that in quarantined anti-vax and neutral forums, the correlation is positive (0.089; 0.010). To contrast, in the non-quarantined anti-vax forum and pro-vax forums, there is a negative correlation (-0.057; -0.003). This is notable as karma also determines which posts are the most prominent within a forum due to algorithmic sorting. Though the stated ideology of the forum was anti-vax, it was not a fertile space for misinformation discourse; as such anti-vaccine misinformation was unlikely to spill over there.

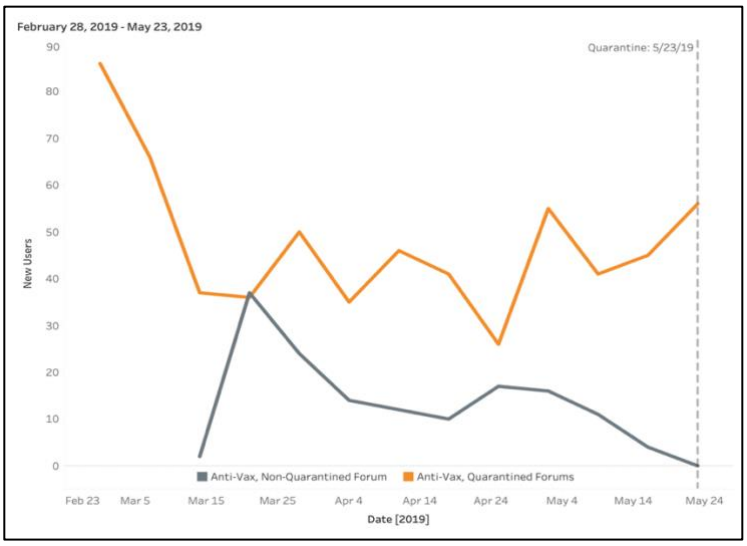


Figure 9. New Users/ Week in Anti-Vax Forums

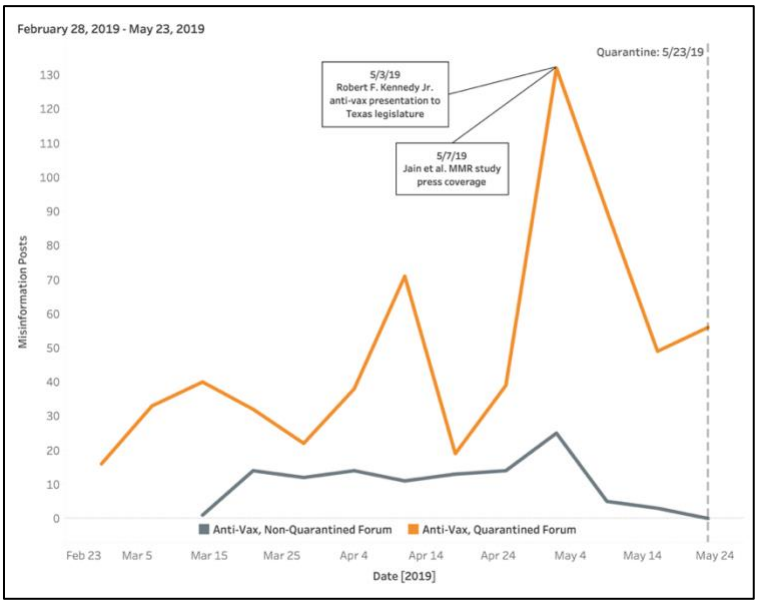


Figure 10. Misinformation Posts/ Week in Anti-Vax Forums

Appendix 5: Robustness Check, Effect of Quarantine on Other Vaccine-Related Forums

To assess the effect of quarantine on users' misinformation spreading within vaccine-related forums, we ran a difference-in-differences analysis on a matched sample of control and quarantine-exposed users. These users were matched on the same control variables utilized in the main semiparametric DID analyses (misinformation posts at baseline, posts in vaccine-related forums at baseline, number of unique vaccine-related forums posted in at baseline, posts in non-vaccine related forums at baseline, number of unique non-vaccine-related forums posted in at baseline, user tenure,

whether the user is a moderator, whether the user has a premium account, and whether the user has verified their email). A regression was then run with robust standard errors.

We find that these results are consistent with our main analyses. Table 29 shows that quarantine induces users to increase misinformation posts in neutral-vax forums (1.075 posts; p<0.029). Table 30 shows that quarantine has no significant effect on users' misinformation behavior in pro-vax forums (p<0.403).

Table 31. Effect of Quarantine on Misinformation Posts in Neutral-Vax Forums, Matched Difference-in-Differences Specification

Neutral-Vax Forums: Misinformation Posts	**Coefficient**	**Robust Standard Error**	**t**	**P>|t|**	**95% Confidence Interval**	
Quarantine-Exposed X Post-Quarantine	1.075	0.491	2.19	0.029	0.113	2.037
Quarantine-Exposed	0.212	0.088	2.40	0.016	0.039	0.385
Post-Quarantine	-0.005	0.043	-0.11	-.914	-0.088	0.079
Constant	0.037	0.030	1.21	0.225	-0.023	0.096
N: 1,954; **F(3, 1950)**: 6; **Prob>F**: 0.0004; **R^2**: 0.008; **Root MSE**: 6.018						

Table 32. Effect of Quarantine on Misinformation Posts in Pro-Vax Forums, Matched Difference-in-Differences Specification

Pro-Vax Forums: Misinformation Posts	**Coefficient**	**Robust Standard Error**	**t**	**P>|t|**	**95% Confidence Interval**	
Quarantine-Exposed X Post-Quarantine	-0.052	0.063	-0.84	0.403	-0.175	0.071
Quarantine-Exposed	0.114	0.057	2.00	0.045	0.002	0.226
Post-Quarantine	-0.078	0.024	-3.33	0.001	-0.125	-0.032
Constant	0.092	0.023	4.03	0.000	0.047	0.137
N: 1,954; **F(3, 1950)**: 9.46; **Prob>F**: 0.000; **R^2**: 0.009; **Root MSE**: 0.751						

Appendix 6: Misinformation Contagion within Neutral-Vax Forums

We run this analysis on all posts from neutral-vax native users made during the long-term follow-up period (May 23, 2019 – March 1, 2020). The goal is to assess if direct interaction with misinformation on a conversation thread makes these control users more likely to post misinformation themselves subsequently. If this is true, this is strong evidence for misinformation contagion. We aggregate each users' activity by conversation thread and arrange the threads chronologically. We then assess if the

user directly interacted with misinformation on the thread. We define direct interaction as being in the same comment branch as another post. An example is below in Figure 10.

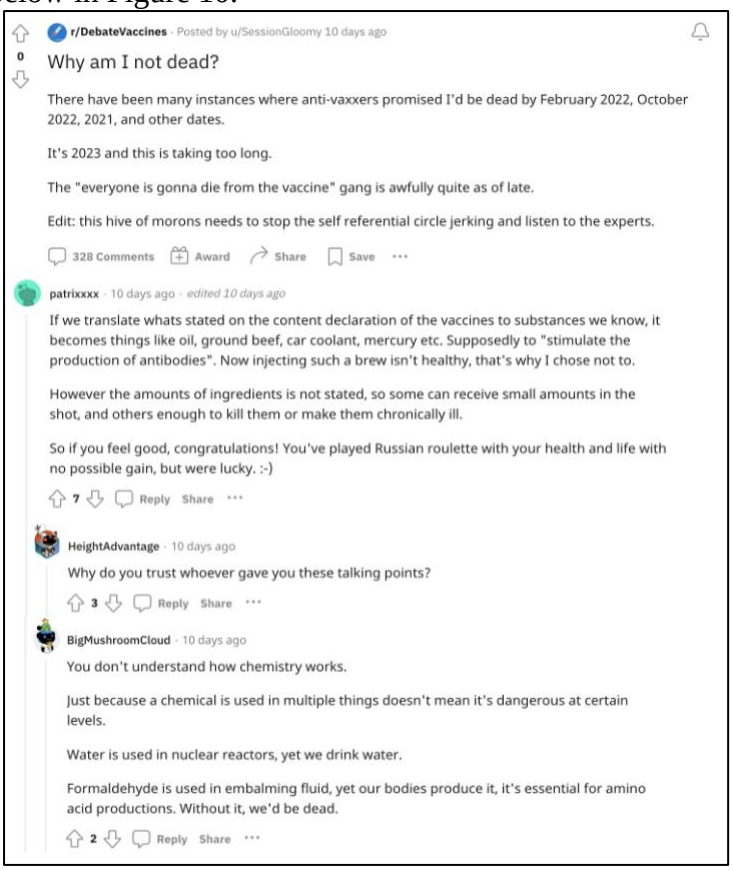


Figure 11. Direct Interaction with Misinformation on Conversation Thread

We then run an ordinary least squares regression with robust standard errors on all neutral-vax natives' conversation thread activity. The results are available below in Table 31. Control variables are the number of posts the native has interacted with on the thread, weekly time dummies, as well as variables constructed based on the native's baseline activity[19] and account features[20]. The dependent variable is a dummy variable that takes the value of 1 if the neutral-vax native subsequently posted misinformation, 0 otherwise. The independent variable is whether that user directly interacted with misinformation on that thread. The specification of the regression user *i* on thread *t* is below:

[19] Misinformation posts, posts in vaccine-related forums, number of unique vaccine-related forums posted in, posts in non-vaccine related forums, number of unique non-vaccine-related forums posted in

[20] User tenure, whether the user is a moderator, whether the user has a premium account, and whether the user has verified their email

$$Subsequent\ Misinformation\ Post_{it+1} = \beta_0 + \beta_1 Misinformation\ Interaction_{it} + \beta_2 Thread\ Interactions_{it} + \sum_{k=3}^{K} \beta_k Baseline\ Activity\ Variable_i + \sum_{j=K+1}^{J} \beta_j Account\ Feature\ Variable_i + \beta_{J+1} Week_t + \varepsilon_{it}$$

We find that neutral-vax natives are not significantly more or less likely to spread misinformation ($p<0.291$) after directly interacting with it on a conversation thread. Thus, there is not strong evidence for misinformation contagion to neutral-vax natives.

Table 33. Assessment of Misinformation Contagion to Neutral-Vax Natives

Likelihood of Misinformation Next Post	**Coefficient**	**Robust Standard Error**	**t**	**P>|t|**	**95% Confidence Interval**	
Interacted with Misinformation	0.015	0.014	1.060	0.291	-0.013	0.042
Conversation Interactions	0.000	0.000	1.200	0.228	0.000	0.001
Baseline Non-Vaccine Related Posts	0.000	0.000	2.490	0.013	0.000	0.000
Baseline Non-Vaccine Related Forums	0.000	0.000	-4.500	0.000	0.000	0.000
Baseline Vaccine Related Posts	-0.001	0.000	-4.320	0.000	-0.002	-0.001
Baseline Vaccine Related Forums	0.026	0.007	3.890	0.000	0.013	0.040
Baseline Misinformation Posts	0.007	0.001	5.920	0.000	0.005	0.010
Is Moderator	0.030	0.010	2.940	0.003	0.010	0.050
Has Premium Account	-0.075	0.017	-4.380	0.000	-0.109	-0.042
Tenure	0.000	0.000	4.190	0.000	0.000	0.000
Has Verified Email	-0.047	0.013	-3.740	0.000	-0.072	-0.022
+ WEEKLY DUMMIES						

Constant	0.014	0.016	0.910	0.365	-0.016	0.045
N: 2,864; **F(55, 2808)**: 2.83; **Prob>F**: 0.000; **R^2**: 0.070; **Root MSE**: 0.223						

Appendix 7: Non-Misinformation Posts in Neutral-Vax Forums: Quarantine-Exposed Users vs. Neutral-Vax Natives

Both descriptive and empirical analyses show that quarantine-exposed users produce a misinformation spillover in neutral-vax forums, but that it does not persist. We assess if non-misinformation posting follows the same trend by directly comparing quarantine-exposed users with neutral-vax natives. We first descriptively compare the two groups. Figure A7 shows that quarantine-exposed users significantly increase non-misinformation contribution after quarantine. This non-misinformation spillover peaks in September 2019, but persists at a high level until 3/1/20. Even though the misinformation spillover to these forums is relatively short-lived, the non-misinformation spillover persists.

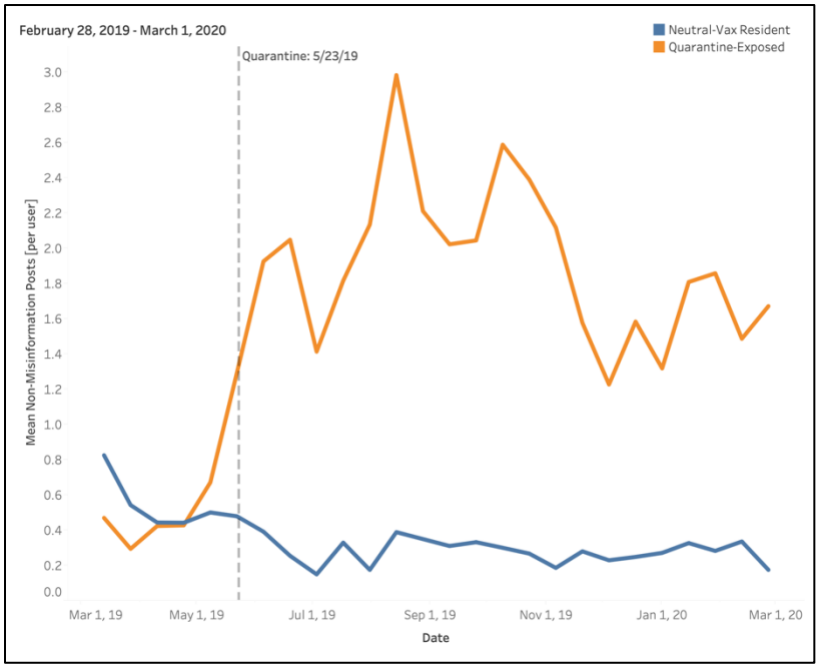


Figure 12. Non-Misinformation Posts in Neutral-Vax Forums: Quarantine-Exposed Users vs. Neutral-Vax Natives

These descriptive results are confirmed by semiparametric DID analyses. Table 32 shows that quarantine-exposed users significantly increase non-misinformation contribution compared to neutral-vax natives in the short-term (ATT= 7.864 non-misinformation posts; $p<0.005$). In the long term, the difference between quarantine-exposed users and neutral-vax natives persists and is highly significant (ATT = 15.663; $p<0.018$).

Table 34. Comparison between Quarantine-Exposed Users and Neutral-Vax Natives, Non-Misinformation

Forum Type	**ATT**	**Standard Error**	**z**	**P>z**	**Confidence Interval**	
Neutral-Vax Forums, Short Term	7.864	2.790	2.82	0.005	2.396	13.332
Neutral-Vax Forums, Long Term	15.663	6.613	2.37	0.018	2.701	28.624
N: 1,305						

Chapter 5: Conclusions

The work outlined in this dissertation proposal answers one question- how platform design choices and governance strategies impact user behavior- within three different contexts. In the Chapter Two, I discuss how prominence reduction strategies impact misinformation spreading behavior on UGC platforms. It is critical to understand both the efficacy and downstream effects of these widely utilized strategies. The unchecked spread of misinformation has been tied to skewed electoral results, negative public health outcomes, and violence worldwide (Hussain et al. 2018; McLaughlin 2018; Grinberg et al. 2019). Additionally, it brings negative media attention to platforms themselves (Pariser 2011; Schulze 2019).

I find that when prominence reduction strategies are applied to a problematic group, they misinformation decreases in certain spaces on a platform. However, it creates a short-lived spillover. This finding has important theoretical implications regarding the spread of misinformation and how it relates to platform design. Furthermore, it provides concrete recommendations for practitioners on how to apply prominence reduction strategies most effectively to groups endemic with misinformation.

In Chapter Three, I discus how both prominence reduction and banning strategies impact verbal aggression behaviors online. With the rising global penetration of the Internet, the number of Internet platforms devoted to hate groups and hate speech has correspondingly risen (Banks 2010). This growth in online hate communities has paralleled a global increase in hate speech, harassment, and cyberbullying in recent years (Banks 2010; Arthur 2019). A survey of Twitter users found that users do not feel adequately protected from harassment while using social media (Jhaver et al. 2018). It is thus important for all stakeholders to understand how popular content moderation strategies impact verbal aggression behaviors.

I find that prominence reduction strategies have limited efficacy in decreasing verbal aggression outside of the focal problematic group. Indeed, both strategies produce spillovers. Banning strategies produce a wider spillover, which is driven by multihoming users. These findings have theoretical implications online platform design and informational cascades. In addition, they provide suggestions to practitioners as how to best control verbal aggression.

Finally, in Chapter Four, I examine a different facet of platform design strategies: incentive structures on crowdsourcing platforms. Previous work has examined how the presence of financial incentives impacts user behavior ex ante (Presslee et al. 2013; Kosfeld et al. 2017; Kralova & Kral 2019), but it is unclear how winning a monetary reward impacts the behavior of high-contribution, high-ability superstar users. I find that monetary rewards significantly improve subsequent user performance. The mechanism is improved team composition.

The work outlined in this dissertation proposal digs into how specific choices made my platform administrators and designers trickle down and impact user behavior, and general outcomes in turn. I study the impact of platform strategies on the so-called ‘dark side’ of Internet behavior: misinformation, as well as verbal aggression. Additionally, I analyze how platform design can impact the ability of crowdsourcing platforms to create publicly source solutions. With this dissertation, I

hope to contribute to a rich stream of literature that illuminates how online platforms impact myriad aspects of modern life, through their direct impact on human behavior.

Bibliography

Abadie, A. 2005. "Semiparametric Difference-in-Differences Estimators," *Review of Economic Studies* (72), pp. 1-19.

Acemoglu, D., A. Ozdaglar, J. Siderius. 2021. Misinformation: Strategic sharing, homophily, and endogenous echo chambers. *Technical Report, National Bureau of Economic Research*.

Adar, E., B. A. Huberman. 2000. Free Riding on Gnutella. *First Monday,* 5 (10), Digital Edition.

Ali, S., Saeed, M.H., Aldreabi, E., Blackburn, J., De Cristofaro, El, Zannettou, S., and Gianluca, S. 2021. "Understanding the Effect of Deplatforming on Social Networks", *Proceedings of the 13th ACM Web Science Conference*.

Allon, G., A. Bassamboo, I. Gurvich. 2011. "We Will Be Right with You" Managing Customer Expectations with Vague Promises and Cheap Talk. *Operations Research*, 59 (6), 1382-1394.

Alorainy, W., Burnap, P., Liu, H., Williams, M.L. 2019. "The Enemy Among Us": Detecting Cyber Hate Speech with Threats-based Othering Language Embeddings," *ACM Trans. Web* (13:3:14), pp. 1-26.

Aratani, L. 2020. "How did face masks become a political issue in America?" *The Guardian,* https://www.theguardian.com/world/2020/jun/29/face-masks-us-politics-coronavirus

Archak, N. 2010. "Money, glory and cheap talk: Analyzing strategic behavior of contestants in simultaneous crowdsourcing contests on TopCoder.com," *Proceedings of the International World Wide Web Conference,* Raleigh, NC, pp. 21-30.

Arthur, R. 2019. We Analyzed More Than 1 Million Comments on 4chan. Hate Speech There Has Spiked by 40% Since 2015. Vice News. Retrieved from https://www.vice.com/en_us/article/d3nbzy/we-analyzed-more-than-1-million-comments-on-4chan-hate-speech-there-has-spiked-by-40-since-2015 on October 2, 2019.

Asarch, S. 2018. "Reddit quarantines a new wave of subreddits, including CringeAnarchy and WatchPeopleDie," *Newsweek,* https://www.newsweek.com/reddit-quarantine-subs-toxic-controversial-moderators-1144663.

Avram, M., N. Micallef, S. Patil, F. Menczer. 2020. Exposure to social engagement metrics increases vulnerability to misinformation. *Harvard Kennedy School Misinformation Review*, 1, 1-11.

Ayers, J. W., B. Chu, Z. Zhu. 2021. Spread of Misinformation About Face Masks and COVID-19 by Automated Software on Facebook. *JAMA Internal Medicine,* 181 (9), 1251-1253.

Baek, J. and Shore, J. 2020. Forum Size and Content Contribution per Person: A Field Experiment. *Management Science* (66:12), 5906-5924.

Bak-Coleman, J. B., I. Kennedy, M. Wack, A. Beers, J. S. Schafer, E. S. Spiro, K. Starbird, J. D. West. 2022. Combining interventions to reduce the spread of viral misinformation. *Nature Human Behavior*, 6, 1372–1380.

Banerjee, A. 1992. A simple model of herd behavior. *Quarterly Journal of Economics*, 107 (3), 797–817.

Banks, J. 2010. Regulating hate speech online. *International Review of Law, Computers and Technology* 24:3, 233-239.

Bayus, B. 2013. "Crowdsourcing New Product Ideas over Time: An Analysis of the Dell IdeaStorm Community," *Management Science* (59:1), pp. 226-244.

Belle, N., and Cantarelli, P. 2015. "Monetary Incentives, Motivation, and Job Effort in the Public Sector: An Experimental Study With Italian Government Executives," *Review of Public Personnel Administration* (25:2), pp. 99-123.

Benjakob, O. 2020. "Why Wikipedia Is Much More Effective Than Facebook at Fighting Fake News," *Haaretz*, https://www.haaretz.com/us-news/.premium-why-wikipedia-is-much-more-effective-than-facebook-at-fighting-fake-news-1.8378622.

Bernstein, M.S., Bakshy, E., Burke, M., and Karrer, B. 2013. "Quantifying the Invisible Audience in Social Networks," *CHI: Conference on Human Factors in Computing Systems Proceedings*, Paris, France.

Bikhchandani, S., D. Hirshleifer, I. Welch. 1992. A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades. *Journal of Political Economy*, 100 (5), 992-1026.

Blaschke, S.M., Carroll, P.P., Chaves, D.R., Findley, M.G., Gleave, M.C., Morello, R.N., and Nielson, D.L. 2013. "Extrinsic, Intrinsic, and Social Incentives for Crowdsourcing Development Information in Uganda: a Field Experiment," *Mimeo*.

Boons, M., Stam, D., and Barkema, H.G., 2015. "Feelings of Pride and Respect as Drivers of Ongoing Member Activity on Crowdsourcing Platforms," Journal of Management Studies (52:6), pp. 717-741.

Boudreau, K.J., Lacetera, N., Lakhani, K.R. 2011. "Incentives and Problem Uncertainty in Innovation Contests: An Empirical Analysis," *Management Science* (57:5), pp. 843-863.

Boutyline, A., Willer, R. 2016. "The Social Structure of Political Echo Chambers: Variation in Ideological Homophily in Online Networks," *Political Psychology* (38:3), pp. 551-569.

Boxer, P. et al. 2005. "Proximal Peer-Level Effects of a Small-Group Selected Prevention on Aggression in Elementary School Children: An Investigation of the Peer Contagion Hypothesis," *Journal of Abnormal Child Psychology* (33:3), pp. 325-338.

Brandtzæg, P.B., and Heim, J. 2011. "A typology of social networking sites users," *International Journal of Web Based Communities* (7:1), pp. 28-51

Burke, M., Marlow, C., and Lento, T. 2009. "Feed Me: Motivating Newcomer Contribution in Social Network Sites," *CHI: Conference on Human Factors in Computing Systems Proceedings*, Boston, MA.

Calonico, S., Cattaneo, M.D., Farrell, M.H., and Titiunik, R. 2017. "rdrobust: Software for regression-discontinuity designs," The Stata Journal (17:2), pp. 372-404.

Candogan, O., K. Drakopoulos. 2020. Optimal Signaling of Content Accuracy: Engagement vs. Misinformation. *Operations Research,* 68 (2), 497-515.

Carron-Arthur, B., J. A. Cunningham, K. M. Griffiths. 2014. Describing the distribution of engagement in an Internet support group by post frequency: A comparison of the 90-9-1 Principle and Zipf's Law. *Internet Interventions,* 1 (4), 65-68.

Carpenter, S. 2021. Zuckerberg Loses $6 Billion in Hours as Facebook Plunges. Bloomberg. https://www.bloomberg.com/news/articles/2021-10-04/zuckerberg-loses-7-billion-in-hours-as-facebook-plunges

Cattaneo, M.D., Titiunik, R., and Vazquez-Bare, G. 2020. "Analysis of regression-discontinuity designs with multiple cutoffs or multiple scores," The Stata Journal (20:4), pp. 866-691.

Cezar, A., S. Raghunathan, S. Sarkar. 2020. Adversarial Classification: Impact of Agents' Faking Cost on Firms and Agents. *Production and Operations Management,* 29 (12), 2789-2807.

Chandrasekharan, E., Pavalanathan, U., Srinivasan, A., Glynn, A., Eisenstein, J., and Gilbert, E. 2017. "You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech," *Proceedings of the ACM on Human-Computer Interaction* (2:31).

Chandrasekharan, E., Jhaver, S., Bruckman, A., and Gilbert, E. 2022. "Quarantined! Examining the Effects of a Community-Wide Moderation Intervention on Reddit," *ACM Transactions on Computer-Human Interaction* (29:4), 29.

Chang, J., C. Danescu-Niculescu-Mizil. 2019. Trajectories of Blocked Community Members: Redemption, Recidivism and Departure. *Proceedings of The World Wide Web Conference*, San Francisco, CA (184-195).

Chen, Z., and Berger, J. 2013. "When, Why, and How Controversy Causes Conversation," *Journal of Consumer Research* (40:3), pp. 580-593.

Chen, H., Z. E. Zheng, Y. Ceran. 2016. De-Biasing the Reporting Bias in Social Media Analytics. *Production and Operations Management,* 25 (5), 849-865.

Cheng, J., Danescu-Niculescu-Mizil, C., and Leskovec, J. 2015. "Antisocial Behavior in Online Discussion Communities," *Proceedings of the Ninth International AAAI Conference on Web and Social Media,* pp. 61-70.

Choi, J. 2014. "Can Offering a Signing Bonus Motivate Effort? Experimental Evidence of the Moderating Effects of Labor Market Competition," The Accounting Review (89:2), pp. 545-570.

Chung Ng, K., P. F. Ke, M. K. P. So, K. Y. Tam. 2023. Augmenting fake content detection in online platforms: A domain adaptive transfer learning via adversarial training approach. *Production and Operations Management*, forthcoming.

Cicchirillo, V., Hmielowski, J., and Hutchens, M. 2015. "The Mainstreaming of Verbally Aggressive Online Political Behaviors," *Cyberpsychology, Behavior, and Social Networking* (18:5), pp. 253-259.

Colleoni, E., Rozza, A., and Arvidsson, A. 2014. "Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data," *Journal of Communication* (64:2), pp. 317-332.

Copland, S. 2020. "Reddit quarantined: Can changing platform affordances reduce hateful material online?" *Internet Policy Review* (9:4), pp. 1-26.

Corner, J. 2017. Fake news, post-truth, and media-political change. *Media, Culture & Society*, 39 (7), 1100-1107.

Crapis, D., B. Ifrach, C. Maglaras, M. Scarsini. 2017. Monopoly Pricing in the Presence of Social Learning. *Management Science*, 63 (11), 3586-3608.

Cusumano, M. A., A. Gawer, D. B. Yoffe. 2021. Social Media Companies Should Self-Regulate. Now. *Harvard Business Review*, Digital Edition. Available at https://hbr.org/2021/01/social-media-companies-should-self-regulate-now.

Davis, A.M., V. Gaur, D. Kim. 2021. Consumer Learning from Own Experience and Social Information: An Experimental Study. *Management Science*, 67 (5), 2924-2943.

Del Vicario, M., Bessi, A., Zollo, F., Scala, A., Caldarelli, G., Stanley, H.E., and Quattrociocchi, W. 2016. "The spreading of misinformation online," *PNAS* (113: 3), pp. 554-559.

Dissanayake, I., Nerur, Sridhar, and Zhang, Jie. 2019. "Team Formation and Performance in Online Crowdsourcing Competitions: The Role of Homophily and Diversity in Solver Characteristics," *Proceedings of the International Conference on Information Systems*, 5.

Duan, W., B. Gu, A. B. Whinston. 2008. Do online reviews matter? An empirical investigation of panel data. *Decision Support* Systems, 45 (4), 1007–1016.

Edgecomb, C. 2019. "Facebook and Instagram starting to identify and label 'fake news'," *Impact BND*, https://www.impactbnd.com/blog/facebook-and-instagram-starting-to-identify-and-label-fake-news-before-it-goes-viral.

Feldman, P., Y. Papanastasiou, E. Segev. 2019. Social Learning and the Design of New Experience Goods. *Management Science*, 65 (4), 1502-1519.

Gavin, J. 2019. "Twitter bans political advertising: the beginning of the fight-back against fake news?" *FIPP*, https://www.fipp.com/news/industrynews/twitter-bans-political-advertising.

Gerber, E.R., Henry, A.D., and Lubell, M. 2013. "Political Homophily and Collaboration in Regional Planning Networks," *American Journal of Political Science* (57:3), pp. 598-610.

Gergen, K.J. 1965. "The effects of interaction goals and personalistic feedback on the presentation of self," *Journal of Personality and Social* Psychology (1), pp. 413-424.

Geurin-Eagleman, and A.N., Burch, L.M., 2015. "Communicating via photographs: A gendered analysis of Olympic athletes' visual self-presentation on Instagram," *Sport Management Review*.

Ghosh, B. P., M. R. Galbreth. 2022. The weight of the crowd, social information credibility, and firm strategy. *Production and Operations Management*, 32 (4), 1079-1095.

Goffman, E. 1959. *The presentation of self in everyday life*. New York: Doubleday Anchor.

Goldstein et al. 2001. "Contagion of Aggression in Day Care Classrooms as a Function of Peer and Teacher Responses," *Journal of Educational Psychology* (93:4), pp. 708-719.

Granovetter, M. 1978. Threshold Models of Collective Behavior. *American Journal of Sociology*, 83 (6), 1420-1443.

Granovetter, M., Soong, R. 1983. Threshold Models of Diffusion and Collective Behavior. *Journal of Mathematical Sociology*, 9, 165-179.

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., and Lazer, D. 2019. "Fake news on Twitter during the 2016 U.S. presidential election," *Science* (363:6425), pp. 374-378.

Guidry, J. P. D., L. L. Austin, N. H. O'Donnell, I. A. Coman, A. Lovari, M. Messner. 2020. Tweeting the #flushot: Beliefs, Barriers, and Threats During Different Periods of the 2018 to 2019 Flu Season. *Journal of Primary Care & Community Health*, 11, 1-10.

Haaretz. 2020. Why Wikipedia Is Much More Effective Than Facebook at Fighting Fake News. January 9, 2020. Available at https://www.haaretz.com/us-news/.premium-why-wikipedia-is-much-more-effective-than-facebook-at-fighting-fake-news-1.8378622 (last accessed June 16, 2023).

Habib, H., R. Nithyanand. 2022. Exploring the Magnitude and Effects of Media Influence on Reddit Moderation. *Proceedings of the International AAAI Conference on Web and Social Media*, Atlanta, GA (275-286).

Habib, H. M. Bin Musa, F. Zaffar, R. Nithyanand. 2022. Are Proactive Interventions for Reddit Communities Feasible? *Proceedings of the International AAAI Conference on Web and Social Media*, Atlanta, GA (264-274).

Hannan, R.L., Krishnan, R., and Newman, A.H. 2008. "The Effects of Disseminating Relative Performance Feedback in Tournament and Individual Performance Compensation Plans," *The Accounting Review* (83:4), pp. 893-913.

Hausken, K. 2020. Game Theoretic Analysis of Ideologically Biased Clickbait or Fakes News, and Real News. *Operations Research and Decisions,* 30 (2), 39-57.

Hayashi, T., Shimizu, T., and Fumaki, Y. 2021. "Collaborative Problem Solving on a Data Platform Kaggle," IEICE Technology Report (120:362), pp.37-40.

Hewett, R., and Leroy, Hannes. 2019. "Well It's Only Fair: How Perceptions of Manager Discretion in Bonus Allocation Affect Intrinsic Motivation," *Journal of Management Studies* (56:6), pp. 1105-1137.

Hirano, K., Imbens, G.W., and Ridder, G. 2003. "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score", *Econometrica* (71:4), pp. 1161-1189.

Hodas, N. O., K. Lerman. 2014. The Simple Rules of Social Contagion. *Nature Scientific Reports,* 4 (4343), 1-7.

Hogan, B. 2010. "The Presentation of Self in the Age of Social Media: Distinguishing Performances and Exhibitions Online," *Bulletin of Science, Technology, and Society* (30:6), pp. 377-386.

Hossain, M. A., M. M. H. Chowdhury, I. O. Pappas, B. Metri, L. Hughes, Y. K. Dwivedi. 2022. Fake news on Facebook and their impact on supply chain disruption during COVID-19. *Annals of Operations Research*, https://doi.org/10.1007/s10479-022-05124-1.

Houngbedjie, K. 2016. "Abadie's semiparametric difference-in-differences estimator," *The Stata Journal* (16:2), pp. 482-490.

Huang, Y., Vir Singh, P., and Mukhopadhyay, T. 2012. "Crowdsourcing Contests: A Dynamic Structural Model of the Impact of Incentive Structure on Solution Quality," *Proceedings of the International Conference on Information Systems*, Orlando, FL.

Huber, G.A. and Malhotra, N. 2017. "Political Homophily in Social Relationships: Evidence from Online Dating Behavior," *The Journal of Politics* (79:1).

Hussain, A., Ali, S., Ahmed, M., and Hussain, S. 2018. "The Anti-vaccination movement: A Regression in Modern Medicine," *Cureus* (10:7), e2919.

Hwang, E. H., D. Krackhardt. 2020. Online Knowledge Communities: Breaking or Sustaining Knowledge Silos?. *Production and Operations Management,* 29 (1), 138-155.

Infante, D.A., Wigley C.J. 1986. "Verbal aggressiveness: An interpersonal model and measure," *Communication Monographs* (53:1), pp. 61-69.

Jellison, J.M., and Gentry, K.W. 1978) A self-presentation interpretation of the seeking of approval. *Personality and Social Psychology Bulletin* (4), pp. 227-230.

Jhaver, S., Ghoshal, S., Bruckman, A., and Gilbert, E. 2018. Online Harassment and Content Moderation: The Case of Blocklists. *ACM Transactions on Computer-Human Interaction* 25:2, Article 12.

Jin, Y., Y. Tan., J. Huang. 2022. Managing contributor performance in knowledge-sharing communities: a dynamic perspective. *Production and Operations Management,* 31 (11), 3945-3962.

Jolley, D., and Douglas, K.M. 2014. "The Effects of Anti-Vaccine Conspiracy Theories on Vaccination Intentions," *PloS ONE* (9:2), e89177.
Jones, R.G., and Jones, E.E. 1964. "Optimum conformity as an ingratiation tactic," *Journal of Personality* (32), pp. 436-458.
Josey, C.S. 2010. "Hate speech and identity: An analysis of neo racism and the indexing of identity," *Discourse & Society* (21:1), pp. 27–39.
Jouini, O., Z. Aksin, Y. Dallery. 2011. Call Centers with Delay Information: Models and Insights. *Manufacturing & Service Operations Management*, 13 (4), 534-548.
Kali, R., Pastoriza, D., and Plante, J-F. "The burden of glory: Competing for nonmonetary incentives in rank-order tournaments," *Journal of Economics & Management Strategy* (27), pp 108-118.
Kalinowski, C., and Matei, S. 2011. "Goffman meets online dating: Exploring the 'virtually' socially produced self,' *Journal of Social Informatics* (16), pp. 6-20.
Karlsen, R., Steen-Johnsen, K., Wollebaek, D., and Enjolras, B. 2017. "Echo chamber and trench warfare dynamics in online debates," *European Journal of Communication* (32:3), pp. 257-273.
Kata, A. 2010. A postmodern Pandora's box: Anti-vaccination misinformation on the Internet. *Vaccine*, 28 (7), 1709-1716.
Kata, A. 2012. "Anti-vaccine activists, Web 2.0, and the postmodern paradigm – An overview of tactics and tropes used online by the anti-vaccination movement," *Vaccine* (30:25), pp. 3778-3789.
Kelly Garrett, R. 2009. "Echo chambers online?: Politically motivated selective exposure among Internet news users," *Journal of Computer-Mediated Communication* (14:2), pp. 265-285.
Khasragi, H.J., Wang, X., and Li, Y. 2020. "Crowdsourcing Contests: Understanding the Effect of Environment and Organization Specific Factors on Sustained Participation," *Proceedings of the 53rd Hawaii International Conference on System Sciences*, pp. 178-87.
Klofstad, C.A., Uscinski, J.E., Connolly, J.M., and West, J.P. 2019. "What drives people to believe in Zika conspiracy theories?" *Palgrave Communications* (5:36), pp. 1-8.
Konstantakis, K., P. T. Cheilas, I. G. Melissaropoulos, P. Xidonas, P. G. Michaelides. 2022. Supply chains and fake news: a novel input–output neural network approach for the US food sector. *Annals of Operations Research*, https://doi.org/10.1007/s10479-022-04817-x.
Kosfeld, M., Neckermann, S., and Yang, X. 2017. "The Effects of Financial and Recognition Incentives Across Work Contexts: The Role of Meaning," *Economic Inquiry* (55:1), pp. 237-247.
Kralova, V., and Kral, P. 2019. "Performance Myopia: The Effect of Pay-For-Performance Incentives on Exploration and Coordination," Acta Oeconomica Pragensia (27:1), pp. 50–69.

Küçükgül, C., Ö. Özer, S. Wang. 2022. Engineering Social Learning: Information Design of Time-Locked Sales Campaigns for Online Platforms. *Management Science,* 68 (7), 4899-4918.

Kulp, P. 2016. "Lawless forum site 4chan is going broke as ad sales dry up," *Mashable,* https://mashable.com/2016/10/06/4chan-advertiser-struggle-nearly-broke/

Kumar, D., Hancock, J., Thomas, K., Durumeric, Z. 2022. "Understanding Longitudinal Behaviors of Toxic Accounts on Reddit," *Working Paper*.

Kwan, A., S. A. Yang, A. H. Zhang. 2023. Crowd-judging on Two-sided Platforms: An Analysis of In-group Bias. *Management Science,* forthcoming.

Kwon, K.H., and Gruzd, A. 2017. "Is Aggression Contagious Online? A Case of Swearing on Donald Trump's Campaign Videos on Youtube," *Proceedings of the 50th Hawaii International Conference on System Sciences*.

Lakhani, K.R., and von Hippel, E. 2003. "How open source software works: "free" user-to-user assistance," *Research Policy* (32:6), pp. 923-943.

Lau, R. Y. K., W. Zhang, W. Xu. 2018. Parallel aspect-oriented sentiment analysis for sales forecasting with big data. *Production and Operations Management,* 27 (10), 1775–1794.

Leary, M.R., and Kowalski, R. M. 1990. "Impression management: A literature review and two component model," *Psychological Bulletin* (107), pp. 34–47.

Lee, S., L. Qiu, A. Whinston. 2018. Sentiment Manipulation in Online Platforms: An Analysis of Movie Tweets. *Production and Operations Management,* 27 (3), 393-416.

Lees, J., A. McCarter, D. M. Sarno. 2021. Twitter's disputed tags may be ineffective at reducing belief in fake news and only reduce intentions to share fake news among Democrats and Independents. *Working Paper.*

Liu, T.X., Yang, J., Adamic, L.A., and Chen, Y. 2014. "Crowdsourcing with All-Pay Auctions: A Field Experiment on Taskcn," *Management Science* (60:8), pp. 2020-2037.

Litt, E. 2012. 'Knock, Knock. Who's There? The Imagined Audience," *Journal of Broadcasting & Electronic Media* (56:3), pp. 330-345,

Litt, E., and Hargittai, E. 2016. "The Imagined Audience on Social Network Sites," *Social Media and Society (January-March),* pp. 1-12.

Lodge, M., C. S. Taber. 2013. *The Rationalizing Voter*. Cambridge, UK: Cambridge Univ. Press

Lourenço, S.M. 2016. "Monetary Incentives, Feedback, and Recognition—Complements or Substitutes? Evidence from a Field Experiment in a Retail Services Company," *The Accounting Review* (91:1), pp. 279-297.

Luo, X., Gu, B., Zhang, J., and Phang, C.W. 2017. "Expert Blogs and Consumer Perceptions of Competing Brands," *MIS Quarterly* (41:2), pp. 371-395.

Maglaras, C., M. Scarsini, D. Shin, S. Vaccari. 2022. Product Ranking in the Presence of Social Learning. *Operations Research*, forthcoming.
Manthei, K., Sliwka, D., and Vogelsang, T. 2019. "Talking about Performance or Paying for it? Evidence from a Field Experiment," *IZA Discussion Papers*, No. 12446.
Marwick, A.E., and Boyd, D. 2010. "I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience," *New Media & Society* (13:1), pp. 114-133.
McLaughlin, T. 2018. "How WhatsApp Fuels Fake News and Violence in India," *Wired*, https://www.wired.com/story/how-whatsapp-fuels-fake-news-and-violence-in-india/.
McPherson, M., Smith-Lovin, L., Cook, J.M. 2001. "Birds of a Feather: Homophily in Social Networks," *Annual Review of Sociology* (27), pp. 415-444.
Merunkova, and L., Slerka, J. 2019. "Goffman's Theory as a Framework for Analysis of Self Presentation on Online Social Networks," *Masaryk University Journal of Law and Technology* (13:2), pp. 243-276.
Mudambi, M., Clark, J., Rhue, L, Viswanathan, S. 2022. "Fighting Misinformation on Social Media with Prominence Reduction Strategies," *Working Paper*.
Murayama, K., Matsumoto, M., Izuma, K., and Matsumoto, K. 2010. "Neural basis of the undermining effect of monetary reward on intrinsic motivation," *PNAS* (107:49), pp. 20911–20916.
Myers West, S. 2018. "Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms," *New Media & Society* (20:11), pp. 4366-4383.
Newell, E., Jurgens, D., Saleem, H.M., Vala, H., Sassine, J., Armstrong, C., Ruths, D. 2016. "User Migration in Online Social Networks: A Case Study on Reddit During a Period of Community Unrest," *Proceedings of the Tenth International AAAI Conference on Web and Social Media*, Cologne, Germany, pp. 279-289.
Newsweek. 2018. Reddit quarantines a new wave of subreddits, including CringeAnarchy and WatchPeopleDie. September 28, 2018. Available at https://www.newsweek.com/reddit-quarantine-subs-toxic-controversial-moderators-1144663 (last accessed June 16, 2023).
New York Times. 2021. YouTube bans all anti-vaccine misinformation. September 29, 2021. Available at https://www.nytimes.com/2021/09/29/technology/youtube-anti-vaxx-ban.html (last accessed June 16, 2023).
New York Times. 2023. Reddit Wants to Get Paid for Helping to Teach Big A.I. Systems. April 18, 2023. Available at https://www.nytimes.com/2023/04/18/technology/reddit-ai-openai-google.html (last accessed May 27, 2023).
Nielsen Norman Group. 2006. The 90-9-1 Rule for Participation Inequality in Social Media and Online Communities. Available at

https://www.nngroup.com/articles/participation-inequality/ (last accessed June 16, 2023).
Offitt, P., C. J. Hackett. 2003. Addressing parents' concerns: do vaccines cause allergic or autoimmune diseases? *Pediatrics*, 111 (3), 653-659.
Ortu, M., Destefanis, G., Counsell, S., Swift, S., Tonelli, R., Marchesi, M. 2017. "How diverse is your team? Investigating gender and nationality diversity in GitHub teams," *Journal of Software Engineering Research and Development* (5:9).
Pariser, E. 2011. *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*. London, UK: Penguin.
Papakyriakopoulos, O., E. Goodmann. 2022. The Impact of Twitter Labels on Misinformation Spread and User Engagement: Lessons from Trump's Election Tweets. *Proceedings of the ACM Web Conference,* Lyon, France (2541–2551).
Papanastiasou, Y. 2020. Fake News Propagation and Detection: A Sequential Model. *Management Science*, 66 (5), 1826-1846.
Papanastiasou, Y., K. Bimpikis, N. Savva. 2018. Crowdsourcing Exploration. *Management Science*, 64 (4), 1727-2746.
Pastor-Satorras, R., Vespignani, A. 2001. Epidemic Spreading in Scale-Free Networks. *Physical Review Letters*, 86 (14), 3200-3203.
Patel, S. 2020. "Reddit Claims 52 Million Daily Users, Revealing a Key Figure for Social-Media Platforms," *The Wall Street Journal*, https://www.wsj.com/articles/reddit-claims-52-million-daily-users-revealing-a-key-figure-for-social-media-platforms-11606822200.
Peres Nobre, G., C. H. G. Ferreira, J. M. Almeida. 2022. A hierarchical network-oriented analysis of user participation in misinformation spread on WhatsApp. *Information Processing & Management*, 59 (1), 102757.
Perez, S. 2019. "Reddit's monthly active user base grew 30% to reach 430M in 2019," *Tech Crunch*, https://techcrunch.com/2019/12/04/reddits-monthly-active-user-base-grew-30-to-reach-430m-in-2019/.
Phadke, S., M. Samory, T. Mitra. 2022. Pathways through Conspiracy: The Evolution of Conspiracy Radicalization through Engagement in Online Conspiracy Discussions. *Proceedings of the International AAAI Conference on Web and Social Media*, Atlanta, GA (770-781).
Presslee, A., Vance, T.W., and Webb, R. Alan. 2013. "The Effects of Reward Type on Employee Goal Setting, Goal Commitment, and Performance," *The Accounting Review* (88:5), pp. 1805-1831.
Qiu, L., S. Kumar. 2017. Understanding Voluntary Knowledge Provision and Content Contribution Through a Social Media-Based Prediction Market: A Field Experiment. *Information Systems Research*, 28 (3), 529-546.
Rajadesingan, A., Resnick, P., and Budak, C. 2020. "Quick, Community-Specific Learning: How Distinctive Toxicity Norms Are Maintained in Political Subreddits," *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media*.

Reddit Administration. 2018. What is a Quarantined Subreddit? Available at https://www.reddit.com/r/help/comments/aayoxb/what_is_a_quarantined_subreddit/ (last accessed June 16, 2023).

Ren, Y., Chen, J., Riedl, J. 2016. "The Impact and Evolution of Group Diversity in Online Open Collaboration," *Management Science* (62:6), pp. 1668–1686.

Reuters. 2022. PayPal says policy to fine customers for 'misinformation' was an 'error'. October 10, 2022. Available at https://www.reuters.com/business/finance/paypal-says-it-never-intended-fine-users-misinformation-bloomberg-news-2022-10-10/ (last accessed June 3, 2023).

Ribeiro, M.H., Jhaver, S., Zannettou, S., Blackburn, J., De Cristofaro, E., Stringhini, G., and West, R. 2020. "Does Platform Migration Compromise Content Moderation? Evidence from r/The_Donald and r/Incels", *arXiv:2010.10397v2*.

Risch, J., and Krestel, R. 2020. "Toxic comment detection in online discussions," *Deep Learning-Based Approaches for Sentiment Analysis*, pp. 85-109.

Robertson, A. 2019. "Reddit quarantines Trump subreddit r/The_Donald for violent comments," *The Verge*, https://www.theverge.com/2019/6/26/18759967/reddit-quarantines-the-donald-trump-subreddit-misbehavior-violence-police-oregon.

Rogers, R. 2020. "Deplatforming: Following extreme Internet celebrities to Telegram and alternative social media," *European Journal of Communication* (35:3), pp. 213-229.

Saltz, E., C. Leibowicz, C. Wardle. 2021. Encounters with Visual Misinformation and Labels Across Platforms. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, Yokohama, Japan (1-6).

Sathyanarayana Rao, T.S., and Andrade, C. 2011. "The MMR vaccine and autism: Sensation, refutation, retraction, and fraud," *Indian Journal of Psychiatry* (53:2), pp. 95-96.

Schaubroeck, J., Shaw, J.D., Duffy, M.K., and Mitra, A. 2008. "An Under-Met and Over-Met Expectations Model of Employee Reactions to Merit Raises," *Journal of Applied Psychology* (93:2), pp. 424-434.

Schmidt, A. L., F. Zollo, A. Scala, C. Betsch, W. Quattrociocchi. 2018. Polarization of the Vaccination Debate on Facebook. *Vaccine*, 36 (25), 3606-3612.

Schulze, E. 2019. "EU tells Facebook, Google and Twitter to take more action on fake news," *CNBC News*, https://www.cnbc.com/2019/10/29/eu-tells-facebook-google-and-twitter-to-take-more-action-on-fake-news.html.

Shao, C., G. L. Ciampaglia, O. Varol, K. Yang, A. Flammini, F. Menczer. 2018. The spread of low-credibility content by social bots. *Nature Communications*, 9 (1), 4787.

Sharevski, F., R. Alsaadi, P. Jachim, E. Pieroni. 2022. Misinformation warnings: Twitter's soft moderation effects on COVID-19 vaccine belief echoes. *Computers & Security*, 114, 102577.

Shen, Q., and Rose, C.P. 2019. "The Discourse of Online Content Moderation: Investigating Polarized User Responses to Changes in Reddit's Quarantine Policy," *Proceedings of the Third Workshop on Abusive Language Online*, Florence, Italy, pp. 58-69.

Shen et al. 2020. "Viral vitriol: Predictors and contagion of online toxicity in World of Tanks," *Computers in Human Behavior* (108), 106343.

Shriver, S.K., Nair, H.S., and Hofstetter, R. 2013. "Social Ties and User-Generated Content: Evidence from an Online Social Network," *Management Science* (59:6), pp. 1425-1443.

Smith, N., T. Graham. 2019. Mapping the anti-vaccination movement on Facebook. *Information, Communication & Society*, 22 (9), 1310-1327.

Song, Y., and Wu, Y. 2018. "Tracking the viral spread of incivility on social networking sites: The case of cursing in online discussions of Hong Kong-Mainland China conflict," *Communication and the Public* (31), pp. 46-61.

Spence, M. 1978. "Job market signaling," *Quarterly Journal of Economics* (87:3), pp. 355-374.

Stolton, S. 2019. "In the fight against fake news, YouTube has a 'bias toward keeping content up'," *Euractiv*. https://www.euractiv.com/section/digital/news/in-the-fight-against-fake-news-youtube-has-a-bias-toward-keeping-content-up/.

Sun, Y., Wang, N., Yin, C., and Zhang, J.X. 2015. "Understanding the relationships between motivators and effort in crowdsourcing marketplaces: A nonlinear analysis," *International Journal of Information Management* (35), pp. 267-276.

Tandoc Jr., E.C., Lou, C., and Min, V.L.H. 2019. "Platform-swinging in a poly-social-media context: How and why users navigate multiple social media platforms," *Journal of Computer-Mediated Communication* (24), 21-35.

Tauchert, C., Buxmann, P., and Lambinus, J. 2020. "Crowdsourcing Data Science: A Qualitative Analysis of Organizations' Usage of Kaggle Competitions," *Proceedings of the 53rd Hawaii International Conference on System Sciences*. Hawaii, US.

Terwiesch, C., Xu, Y. 2008. "Innovation Contests, Open Innovation, and Multiagent Problem Solving," *Management Science* (54:9), pp. 1529-1543.

Tollefson, J. 2021. "The race to curb the spread of COVID vaccine disinformation," *Nature News*. https://www.nature.com/articles/d41586-021-00997-x.

Tran, A., and Zeckhauser, R. 2012. "Rank as an inherent incentive: Evidence from field experiment," Journal of Public Economics (96:9-10), pp. 645-650.

Trujillo, A., and Cresci, S. 2022. "Make Reddit Great Again: Assessing Community Effects of Moderation Interventions on r/The_Donald," *arXiv: 2201.06455*.

van der Linden, S., C. Panagopoulos, J. Roozenbeek. 2020. You are fake news: political bias in perceptions of fake news. *Media, Culture & Society*, 42 (3), 460-470.

von Baeyer, C.L., Sherk, D.L., and Zanna, M.P. 1981. "Impression management in the job interview: When the female applicant meets the male (chauvinist) interviewer," *Personality and Social Psychology Bulletin* (7), pp. 45-51.

Vigdor, N., and Chokshi, N. 2019. "Reddit Restricts Pro-Trump Forum Because of Threats", *The New York Times*,

https://www.nytimes.com/2019/06/26/us/politics/reddit-donald-trump-quarantined.html.
Vosoughi, S., Roy, D., and Aral, S. 2018. "The spread of true and false news online," *Science* (359:6380), pp. 1146-1151.
Walsht, B.R., Clarke, E. 2003. "Post-trauma symptoms in health workers following physical and verbal aggression," *Work & Stress* (17:2), pp. 170-181.
Wang, M., and Wang, J. 2019. "Understanding Solvers' Continuance Intention in Crowdsourcing Contest Platform: An Extension of Expectation-Confirmation Model," *Journal of Theoretical and Applied Electronic Commerce Research* (14:3), pp. 17-33.
Wang, X., Khasragi, H.J., and Schneider, H. 2019. "What Sustains Individuals' Participation in Crowdsourcing Contests?" *Proceedings of the 52nd Hawaii International Conference on System Sciences*, pp. 136-145.
Warren et al. 2005. "A Model of Contagion Through Competition in the Aggressive Behaviors of Elementary School Students," *Journal of Abnormal Child Psychology* (33:3), pp. 283-292.
Wei, Z., M. Xiao, R. Rong. 2021. Network Size and Content Generation on Social Media Platforms. *Productions and Operations Management*, 30 (5), 1406-1426.
Wheeler, L, and Levine, L. 1967. "Observer-Model Similarity in the Contagion of Aggression," *Sociometry* (30:1), pp. 41-49.
Xiao, S., Y. Ho, H. Che. 2021. Building the Momentum: Information Disclosure and Herding in Online Crowdfunding. *Production and Operations Management*, 30 (9), 3213-3230.
Yiannakoulias, N., Slavik, C.E., and Chase, M. 2019. "Expressions of pro- and anti-vaccine sentiment on YouTube," *Vaccine* (37:15), pp. 2057-2064.
Young, H.P. 2008. "Social norms," in *The New Palgrave Dictionary of Economics (2nd ed.)*, London: Macmillan.
Zanna, M.P., and Pack, S.J. 1975. "On the self-fulfilling nature of apparent sex differences in behavior," *Journal of Experimental Social Psychology* (77), pp. 583-591.
Zeng, Z., H. Dai, D. J. Zhang, H. Zhang, R. Zhang, Z. Xu, Z. M. Shen. 2022. The Impact of Social Nudges on User-Generated Content for Social Network Platforms. *Management Science*, forthcoming.
Zhang, X., Q. Du, Z. Zhang. 2022a. A theory-driven machine learning system for financial disinformation detection. *Production and Operations Management*, 31 (8), 3160-3179.
Zhang, X., Z. Wei, Q. Du, Z. Zhang. 2022b. Social Media Moderations, User Ban, and Content Generation: Evidence from Zhihu. *Proceedings of the Hawaii International Conference on System Sciences*, Honolulu, HI (4432-4441).
Zhang, X., F. Zhu. 2011. Group Size and Incentives to Contribute: A Natural Experiment at Chinese Wikipedia. *American Economic Review*, 101, 1601-1615.

Zhang, S., Vir Singh, P, and Ghose, A. 2019. “A Structural Analysis of the Role of Superstars in Crowdsourcing Contests,” *Information Systems Research* (30:1), pp. 15-33.
Zheng, H., Li, D., and Hou, W. 2011. “Task Design, Motivation, and Participation in Crowdsourcing Contests,” International Journal of Electronic Commerce (15:4), pp. 57-88.
Zhou, J., Y. Chen. 2016. Targeted Information Release in Social Networks. *Operations Research*, 64 (3), 721-735.
Zhou, Y., L. Shen. 2022. Confirmation Bias and the Persistence of Misinformation on Climate Change. *Communication Research*, 49 (4), 500-521.
Zuhair-Al-Taie, M., S. Kadry. 2017. Information Diffusion in Social Networks. In *Python for Graph and Network Analysis* (165-184)