Balancing Brand Safety and User Engagement in a Two-Sided Market: An Analysis of Content Monetization on Reddit

Chris J. Vargo, Toby Hopp and Pritha Agarwal

College of Media, Communication and Information, University of Colorado Boulder

Author Note

Chris J. Vargo, Toby Hopp and Pritha Agarwal

We have no known conflict of interest to disclose.

Correspondence concerning this article should be addressed to Chris J. Vargo, College of Media, Communication and Information, University of Colorado Boulder, 478 UCB, Armory, 1511 University Ave., Boulder CO 80309-0478. Email: christopher.vargo@colorado.edu

Balancing Brand Safety and User Engagement in a Two-Sided Market: An Analysis of Content Monetization on Reddit

Abstract

Advertisers aim to avoid content containing controversy, sex, violence, and profanity, while platforms like Reddit aim to monetize as much content with ads as possible. This research explores the dynamics of content monetization in the two-sided market of social media, focusing on Reddit's brand safety system's effectiveness in shielding advertisers from unsafe user-generated content (UGC). Analyzing 2,267 active subreddits and 2.74 million submissions over three months in 2022. 55% to 66% of subreddits with relatively high toxicity deemed acceptable for advertising. Overall, brand safety on reddit seems to only affect 15% of submissions, with 85% deemed safe for advertising. Inconsistencies in blocking highly toxic subreddits suggest factors like subreddit size may impact Reddit's brand safety decisions. Together it appears that Reddit's economic motivations in content monetization outweigh advertiser pressure for brand safety. This research underscores the need for more transparent and precise brand safety solutions. It also suggests advertiser concern to appearing alongside "unsafe" content, is overblown as no advertisers have spoken out against brand safety on Reddit.

A recent incident involving X Corp., formerly known as Twitter, has brought the issue of brand safety into sharp focus for advertisers. The Center for Countering Digital Hate (CCDH), a British nonprofit, has tracked blatant speech posted to X Corp's social media platform "X," such as posts urging people to "stop race mixing" and messages stating that Black people are intrinsically violent (Field, 2023). To date, the research group has found that 86% of the content they found was still up and readily accessible (Vanian, 2023).

For advertisers, this is a stark reminder about brand safety, particularly in the context of social media platforms. Advertisers are increasingly concerned about the potential negative impact on their brand image if their ads are placed alongside hateful or harmful content on social media platforms with little advertiser protection (Lee, Kim, & Lim, 2021).

Brand safety, a concept coined by advertising agencies and ad tech corporations, refers to the mechanisms in place to protect brands from advertising alongside content that advertisers, and often their customers, find objectionable (Bishop 2021). Brand safety is available to advertisers on social media platforms as a tool in which they can configure to avoid objectionable content (Guaglione 2020). Under the hood these tools often look for words like, "pot" and "war" and block offending content from being monetized via advertisements. However, the obstacles to studying these tools from an academic perspective are immense. Ad tech companies that create these tools are not obligated nor inclined to share data with advertising agencies, the public, or academics. Together, the extent to which advertisers appear on unsafe environments on social media platforms is largely unknown.

Here, in this analysis we investigate the brand safety system designed for Reddit. We look at how the system labels a collection of over two thousand subreddits, paying close attention to whether the system labels content that advertisers consider to be unsafe. We employ

artificial intelligence developed by Alphabet to detect toxic, hateful, and crude conversations on reddit. By doing so, we reveal how subreddits that Reddit deems safe for advertisements vary from ones that are not.

Programmatic Brand Safety in the Advertising Industry

Increasingly, advertising is being bought programmatically using automated methods (Taffera-Santos 2021). Programmatic advertising forfeits direct advertiser control of exactly where ads will be placed and introduces the need for automated methods to make that decision on behalf of advertisers. However, buying massive amounts of advertising impressions using automated methods increases the possibility of fraud and unintended consequences.[1] Despite the industry's efforts, fraud remains rampant in programmatic advertising.[2] As such, advertisers spend significant portions of their advertising budget for computational systems that claim to keep them safe, either from buying fraudulent inventory, or from inventory they find unsuitable.

The Industry Self-Regulation of Brand Safety

"Practically, brand safety is a positive reproduction of a brand's ideals, an avoidance of controversy, and a circumvention of sex, violence, and profanity" (Bishop 2021, 4). The Internet Advertising Bureau (IAB) offers a taxonomy that allows different advertising computer systems to pass brand safety content labels back and forth in a uniform way (Aaron 2010). They offer no definitions for how these safety scores should be operationalized and measured, or how the brand safety tools themselves would work. The Global Alliance for Responsible Media (GARM), is a collection of advertising agencies, and ad tech vendors. Together they created a "Brand Safety

---

1 Ad fraud occurs when advertisers pay for inventory that is either not received or does not meet specifications, such as bots posing as humans to generate ad revenue (Rao, 2010). More nuanced cases include Gannett Co., parent company of USA Today, inaccurately labeling ad impressions as being served on its main website when they were appearing on less popular, and thus less costly, sites (Haggin, 2022).

2 For a review of a contemporary ad fraud scheme, see: https://adalytics.io/blog/checking-page-data-in-ad-requests

Floor and Suitability Framework" with 12 categories of unsafe content. They named the categories: adult and explicit, alcohol and drugs, child abuse, discrimination, hate speech, illegal activities, offensive language, political extremism, sensitive events and tragedy, suicide, violence, and misinformation.

The content categories still lack codebooks, and examples. Instead, just how these categories are operationalized are left to the discretion of ad tech companies who provide brand safety solutions. For example, GARM defines obscenity and profanity as ""excessive use of profane language or gestures and other repulsive actions that shock, offend, or insult." It does not provide a clear lexicon or frame of reference for what constitutes profane or repulsive language. A review of the different content types that are considered "unsafe" across different industry standards can be found in Table 1.

Content Monetization and Brand Safety in a Two-Sided Market

Madio and Quinn (2021), analyzed the trade-off faced by platforms in terms of allowing unsafe content to be advertised on. The authors identified conditions under which platforms moderate content. For instance, they put forward that platforms may very well choose not to moderate unsafe at all content if users have a strong preference for its presence and advertisers' reputation loss is limited. On the other hand, if advertisers face a higher risk from being associated with unsafe content, platforms may find it optimal to moderate unsafe content to attract more advertisers.

Exposure to offensive content on social media can lead to negative attitudes towards the brand and a desire to generate negative word-of-mouth, even when the brand's association with the offensive content is unintentional (Lee, Kim, & Lim, 2021). This is supported by industry reports suggesting that ad placement next to offensive content can lead to negative word-of-

mouth and brand rejection (Koch, 2019; Manatt, Daniel, & Ben, 2018; Cameron, 2017). However, a study by Bellman et al. (2018) found that the effects of program quality and content on YouTube pre-roll ads were insignificant or small, suggesting that the nature of pre-roll ads may alter the transfer of attitudes from the content itself. It is important to note that pre-roll ads are viewed before the main content, which may account for the observed minimal impact on audience perceptions.

Together, exisiting research presents a complex view of the effect of negative contextual alignment on advertising content. While some studies find weak effects of negative contextual alignment on advertising content (Lee, Kim, & Lim, 2021), theoretical explanations for these discrepancies suggest that the strength of the spillover effect may vary based on factors such as brand familiarity and the intrusiveness of the ad format. Despite these discrepancies, the prioritization of brand safety is justified by the potential for even minimal negative associations to impact consumer attitudes and the overall brand equity.

Madio and Quinn (2021) put forward that the workhorse model of two-sided markets, developed by Rochet and Tirole in 2003, is the most appropriate theoretical framework to analyze markets where two distinct user groups interact through an intermediary or platform, and the decisions of each group affect the outcomes of the other group. In the context of online advertising, the two-sided market consists of users and advertisers. The platform, in this case Reddit, serves as the intermediary. Users provide content and engage with the platform, while advertisers pay to reach these users. The value of Reddit to advertisers increases with the number of active users, and vice versa. This is an example of indirect network effects.[3]

---

3 Indirect network effects, also known as cross-side network effects, refer to the phenomenon where the value of a product or service to one group of users depends on the size or number of a different group of users (Hagiui, 2018).

Considering the social media platform Reddit, the value of the platform to advertisers (one user group) increases with the number of active users (the other user group) because a larger user base provides a wider audience for their ads. Conversely, the value of Reddit to its users can also increase with the number of advertisers if the ad revenue is used to improve the platform or provide premium features.

Rochet and Tirole (2003) argue that traditional one-sided market models cannot adequately capture the unique competition dynamics in two-sided markets, and that understanding indirect network effects is crucial for analyzing these markets. Reddit (the platform) must consider not only the direct effect of its ad monetization procedures on each user group, but also the indirect effect through the change in the size of the other user group. This can lead to complex feedback effects and make the platform's profit-maximizing prices difficult to determine. Reddit, therefore, must balance the need to monetize content with the need to maintain brand safety.

Given the mixed effects suggested by academic research, advertisers continue to prioritize brand safety. This leads us to our primary research question: To what extent does Reddit's brand safety solution protect advertisers from content that is unsafe for advertisers?

RQ1. To what extent does Reddit's brand safety solution protect advertisers from content that is unsafe for advertisers?

The Effects of Uncivil Content

While there has been little research on brand safety, there has been plenty of work that has studied toxicity and incivility online. Toxicity is conceptually defined as the presence of harmful or offensive language, including hate speech, harassment, and incivility (Chandrasekharan et al. 2017). Incivility refers to disrespectful and impolite behavior that deviates from the norms of polite social interaction. They are often used interchangeably to describe offensive behaviors such as hate speech, harassment, and cyberbullying (Papacharissi 2004; Anderson et al. 2014). Prior research on toxicity and incivility in online communication has identified various behaviors that contribute to harmful content, such as flaming and harassment (Hmielowski, Hutchens, and Cicchirillo 2014). These behaviors directly address at least two central GARM categories: hate speech, and offensive language.

As many have shown, Reddit can contain uncivil, hateful, and even threatening content that is certainly against many of the brand safety guidelines established by the advertising industry (see Table 1 for a review). Davidson, Sun, and Wojcieszak (2020) found that 9.21% of all non-political comments and 14.75% of political comments were disrespectful towards others, suggesting that incivility on subreddits were widespread. Hmielowski, Hutchens, and Cicchirillo (2014) documented that flaming, or repeatedly insulting an individual or group with the aim of starting a conflict, is quite common. Stevens, Acic, and Taylor (2021) found that news content posted to Reddit that discussed sexual assault was often met with "rape culture," or uncivil responses downplaying the severity or legitimacy of sexual assault claims. In their analysis of r/The_Donald, Gaudette et al. (2021) found Reddit's unique voting algorithm facilitated toxic "othering" discourse towards two groups, specifically Muslims and the left. Others have shown that with a clear out-group, redditors have the incentive to use inflammatory language, or low-quality, unnecessary aggressive insults (Hmielowski, Hutchens, and Cicchirillo 2014).

Given that toxic — and therefore brand unsafe — behaviors seem prevalent on Reddit, we ask the extent to which these factors correlate with decisions from Reddit to block subreddits from advertising when these behaviors are present:

RQ2. To what extent are highly uncivil subreddits subject to advertising blocks by Reddit?

Toxicity and Engagement

In recent years, researchers have examined the relationship between incivility, toxicity, and user engagement on social media platforms. On the one hand, studies have consistently reported that the presence of incivility and toxic content negatively affects user experiences and engagement on social media (Ruiz et al. 2011; Santana 2014). Gearhart and Zhang (2015) attribute their observations as a spiral of silence, where users become less willing to voice their opinions in the face of a hostile environment (Gearhart and Zhang 2015).

However, exposure to rude or uncivil comments can also trigger a disinhibition effect, leading individuals to become more prone to engage in aggressive or offensive behavior themselves (Lapidot-Lefler and Barak 2012). Similarly, a contagion effect has been observed, where uncivil conduct encourages similar behavior among other users in the online community, fostering an echo chamber of toxicity and incivility (Cheng, Danescu-Niculescu-Mizil, and Leskovec 2015). **U**ncivil or toxic behavior can result in increased user engagement in certain contexts, particularly when the controversial content triggers arousing emotional responses or amplifies polarization in online debates (Garrett et al. 2014; Bond et al. 2012).

However, despite the potential short-term increases in engagement, the long-term consequences of rampant incivility and toxicity in online spaces can lead to the erosion of trust, social disengagement, and eventual abandonment of the platform by users seeking more respectful interactions (Rainie et al. 2012; Glücker et al. 2018).

Overall, the literature paints a complex relationship between social media engagement and toxicity and incivility. Moreover, social media platforms themselves may be less likely to remove content that has strong engagement metrics, as it can drive user interaction, increase time spent on the platform, and ultimately result in more advertising impressions and advertising revenue. This creates a conflict of interest for social media platforms like Reddit, who must balance the need for brand safety with the desire to monetize via advertising. To this end, we seek to compare safe advertising subreddits to unsafe ones to better understand whether the most vibrant and engaging subreddits are monetized with advertising.

RQ3. To what extent does user engagement differ across ad-friendly and ad-blocked subreddits?

In recent years, there has been a growing interest in understanding the relationship between sub-community factors and toxicity levels on social media platforms, including subreddits (Chandrasekharan et al., 2017; Gaudette et al., 2021). One key factor that has been identified in the literature is the size of the sub-community, often measured by the number of subscribers or the number of posts within a subreddit. Larger subreddits have been found to be more prone to toxic behavior and incivility, potentially due to the increased anonymity and reduced social accountability that comes with a larger user base (Almerekhi et al., 2020). This suggests that sub-community factors, such as the number of subscribers and posts, may be

positively associated with toxicity levels on subreddits. To this end we ask to what extent the size and activity of a subreddit map to toxicity outcomes.

RQ4: To what extent are sub-community factors, including the number of subscribers and the number of posts, associated with toxicity levels on subreddits, and how does this relate to their ad-friendly status?

Method

Brand Safety on Reddit

A huge volume of user-generated content (UGC) is produced every minute on Reddit. Like all websites with large amounts of traffic, it is ripe for monetization via advertisements. As such, Reddit serves advertisements primarily on the main submission page for each subreddit that it hosts.[4] Therefore, ads live contextually aligned to submissions for each subreddit, and as such the submission content itself is a spillover concern for advertisers.

To address this, Reddit created its own brand safety solution that relies on human review to label subreddits as being safe for advertisements ("all ads"), not safe ("no ads"), or somewhere in between ("some ads"; Barwick 2022). This system allows for little nuance, ads are either allowed on a subreddit, or they are not. Advertisers have little control, excepting that they can extend their Reddit advertising inventory by accepting "some ads" subreddits, which are "expanded inventory," and somehow less safe than "all ads" subreddits (Reddit for Business 2022). As Joshua Lowcock, global chief media officer at UM Worldwide said of the system, "it

4 For a review of how ads are shown on Reddit: https://www.reddit.com/about/advertising/

feels like this has been done independent of any consultation with the advertising community…it probably tells you a bit more about the lack of sophistication, like the actual controls that they have in the community or their ad-approval process" (Barwick 2022, par. 20). As aforementioned, given that brand safety systems like these are out of reach from everyday advertisers on the platform, this is problematic for advertisers who want to be sure their ads are not appearing next to objectionable content.

Given that most advertising on Reddit occurs on subreddit submission pages, we set out to identify and download a large collection of submissions from a wide variety of subreddits.[5] To inspect the contextual environment of submissions, we focused on the text of titles and descriptions. Previous research identified 2,500 subreddits that had the most followers as of 2013.[6] As no other list this exhaustive was available, we downloaded all available submissions via the pushshift API for these subreddits from March 15th, 2022, to June 26th, 2022 (Baumgartner et al. 2020). This created a universe of 11.14 million. Given the size of the data, we took a 25% sample, stratifying the sample so that the percentage of data for each subreddit matched that of the full sample. In all, the sampling and annotation processes resulted in the categorization of 2,784,790 subreddit submissions. Around 85% of these submissions (n = 2,375,818) appeared in subreddits that allowed ads.[7] The examined user submissions were drawn

5 We chose to focus on submissions—and not comments—because, at the time of this paper, advertising on Reddit is limited to the main submission page of a subreddit, where comments are not visible.

6 We chose this list because it contains subreddits that have remained as established communities over an extended period. The fact that these subreddits have persisted indicates that they continue to be relevant today and are likely to be diverse in content and discussion types. Our choice also ensures that we have a more comprehensive picture of brand safety systems on Reddit, and it allows us to capture subreddits that have flourished on the platform. We also compared the list of top subreddits from 2013 with the current top subreddits as seen on redditlist.com, and we found that most of the top subreddits from our list continue to be among the most popular today. However, redditlist.com only provides the top 125 subreddits, and the data is not downloadable or scrapable. Thus, we chose to use a well-established and diverse list with 2,500 subreddits for our study. To view the list, see here: https://github.com/umbrae/reddit-top-2.5-million

7 To the best of our knowledge, Reddit does not make any information publicly available about the criteria used to distinguish between its primary (i.e., "all ads") and extended (i.e., "some ads") subreddit inventories. For that reason, we considered ad appropriateness on a binary basis (i.e., primary and extended inventories were collapsed into a single category).

from 2,267 subreddits. The mean number of submissions associated with each subreddit was 1,228 (min = 1 submission, max = 225,304 submissions).

The Detection of Unsafe Content

Given the size of the dataset, it was not possible for the researchers to manually annotate a statistically representative sample of the data for the various aspects of brand safety found in industry guidelines. As aforementioned in the literature review, the concept of brand safety varies from advertiser to advertiser, and across industry organizations. Here we did not set out to study every possible concept of brand safety that all industry organizations have put forward, but instead focus on the ones that are most associated with brand safety. Jigsaw (an Alphabet company) developed the Perspective API with the aim to help support comment moderators that remove conversations that are toxic. They define toxicity as content that abuses, harasses, or in some way silences marginalized groups.[8] Jigsaw's definition of toxicity overlaps with several GARM categories, including abuse, discrimination, hate speech, offensive language, and violence.

The Perspective API is a set of supervised deep learning algorithms, designed for automatically reviewing text comment submissions. Jigsaw builds Perspective by leveraging a large pool of human content moderators who annotate hundreds of thousands of newspaper comments and social media posts. At least three studies reviewed in this article leveraged Jigsaw to detect toxicity in Reddit data (Almerekhi, Jansen, and Kwak 2020; Hansen 2022; Stevens, Acic, and Taylor 2021). In a recent application that married self-response survey data with social media data from participants, Hopp et al. (2020) found that not only did the toxicity API detect toxicity in social media content as humans do, but the scores also generally correlated to the

8 To access the Perspective API: https://perspectiveapi.com

perceptions that individuals had of their own incivility on social media. While we agree with researchers who have found that the tool is an imperfect measure, we conceded that for data of this scale, it is likely the most exhaustive and proven option.

To externally validate the data to our present data set, two researchers manually and independently (from both one another and from the computer-derived annotations) reviewed a random sample of 2,000 positively flagged comments for these attributes ($P > .50$) and found that the precision for each of these five algorithms exceeded 70%, and the recall was 68%, resulting in an F-1 score of .69, indicating a balanced performance between precision and recall.

The API in question returns a probability value that represents how likely a given text is to possess a specified attribute. In this case, the attribute in question is rudeness, disrespect, or general unruliness that would make people want to leave a discussion. Additionally, the API can detect identity-based attacks such as racism or xenophobia. It can also detect name-calling, profanity, threats, and sexual explicitness. In the present study, we leveraged the typology and data key put forward by Stevens, Acic, and Taylor (2021). Its "insults" measure detected negative comments toward an opposing person and its "profanity" algorithm generally detected vulgarities and clever derivatives thereof. Its "threat" measure revealed desires to harm an individual or group, and its "identity-based attack" (here referred to as IBA) algorithm revealed negative identity-based comments.

Results

Descriptive Information

For the attribute probability scores, the sample-wide mean for IBA was 0.10 (*SD* = 0.11), 0.16 (*SD* = 0.15) for threatening language, 0.11 (*SD* = 0.16) for profanity, 0.12 (*SD* = 0.14) for insulting language, and 0.11 (*SD* = 0.14) for toxicity. To create a figure that represented the top, most toxic advertisements, an average score was created from the individual (averaged) Perspective attribute scores. Only those subreddits with 100 or more comments were included in this analysis. To see the top 50, most toxic subreddits and their brand safety designation, see Table 2. For the engagement measures, the sample-wide average number of upvotes per submission was 117.91 (*SD* = 1571.95), the sample-wide number of submission crossposts was 0.04 (*SD* = 0.61), and the sample-wide number of post-related submissions was 15.76 (*SD* = 161.73).

Research Question 1

The first research question was interested in the extent to which ad-friendly subreddits feature lower levels of toxic commentary. To assess this question, we first compared the mean toxicity scores associated with submissions appearing in ad-friendly subreddits with the toxicity scores associated with submissions appearing in ad-blocked subreddits. Given the number of observations under consideration, frequentist *p*-values were judged to be generally non-informative for the purposes of assessing group-based differences. Instead, our evaluation focused on the effect size estimates (here, Cohen's *d*) describing the between-group differences of means. Generally speaking, *d* values between 0 and .10 represent negligible/no effect; values between .10 and .50 are indicative of a small effect; values between .50 and .80 represent a moderate effect, and values greater than .80 describe a strong effect. Notably, across all toxicity attributes, we observed lower scores in ad-friendly subreddits. Moreover, as can be seen in Table

3, the observed *d* values ranged from 0.26 to 0.46, indicating a set of persistent, albeit somewhat weak, effects.

TABLE 3 ABOUT HERE

Research Question 2

The second research question was interested in the extent to which highly uncivil subreddits are subject to advertising blocks by Reddit. To examine this question, we first filtered out the subreddits with less than 100 posts. This decision was made to better ensure that sub-community mean scores were plausible estimates of overarching community behavior. This filtering reduced the number of subreddits considered by 555 (n = 1,712). Next, the 100 subreddits scoring the highest on each toxicity attribute were examined. Among the 100 subreddits with the highest mean scores on the identity attack variable, 66 were subject to ad blocks. For the threatening language variable, we again found that 66% of the most offensive subreddits were subject to ad blocks. As it pertained to profanity, 56% of the most profane subreddits were ad blocked. Similarly, out of the 100 subreddits posting the highest insulting language and toxicity scores, 56 were subject to ad blocks.

Research Question 3

The third research question sought to examine the extent to which user-engagement behaviors differed across ad-friendly and ad-blocked subreddits. This inquiry—as delineated above—is important because prior research has provided a clear indication that toxicity and other forms of discursive negativity have a stimulatory effect on user behavior. Interestingly, our data failed to indicate the presence of meaningful differences in user engagement across ad-friendly and ad-blocked subreddits. As shown in Table 4, there was a general trend such that ad-friendly

subreddits had heightened levels of user engagement. That being said, the *d* values associated with these differences were negligible in nature (*d* range = 0.03–0.05). This means that user engagement across ad friendly and ad-blocked subreddits were approximately the same.

TABLE 4 ABOUT HERE

Research Question 4

Research question four assessed the extent to which sub-community factors were associated with toxicity on subreddits. To address this question, we used the reduced community-level dataset (n = 1,712) and assessed the bivariate correlations between subreddit subscribers and number of posts and the toxicity variables. Considering non-normality in the data, correlations were estimated using Kendall's τ. As seen in Table 5, we observed consistent —albeit somewhat weak—associations between the individual toxicity attributes and both number of subscribers and number of posts. An additional set of calculations indicated that highly subscribed subreddits were more like to be labeled ad-friendly, $\tau = .07$, $p < .001$. The number of posts produced in each subreddit was not, however, statistically linked to ad-friendly status, $\tau = .00$, $p = .981$. Taken as a whole, these findings suggest that larger, and presumably more complex, subreddits are subject to enhanced toxicity. At the same time, and perhaps due to market- and earnings-related pressures, Reddit is slightly more likely to attach an ad-friendly status to larger subreddit communities.

Discussion

This study suggests that Reddit, which self-polices and enforces its own brand safety solution, seems to do so with some regard to the unsafe UGC studied here. As researchers, we think it would be near impossible for the Perspective's algorithms to perfectly align with their decision-making, yet we still do observe clear correlations between Reddit's brand safety labels and our automated detection of toxicity.

In general, the study found that ad-friendly subreddits tend to have lower levels of unsafe behaviors — including identity attacks, threats, and insults — although the effect is weak. This suggests that even with the broad strokes that Reddit takes to list an entire sub-community as safe for ads, or not safe for ads, they do seem to make these decisions with some regard to the amount of incivility that the sub-community exhibits. This means that brand safety — at least at Reddit — is conceptualized as not having characteristics of discussion that are hateful, profane, discouraging of discussion or at the expense of marginalized groups.

On average, users participate and engage with content at similar levels, regardless of whether a subreddit is ad-friendly or ad-blocked. There is no significant difference in the level of activity (e.g., upvotes, crossposts, or other submission-related activities) among users in ad-friendly and ad-blocked subreddits. This finding is important for advertisers because it shows that Reddit's brand safety decisioning process does not appear to consider the amount of engagement that exists on a given subreddit.

However, we had a slight suspicion that ad monetization would push Reddit to label the most active subreddits as brand safe, to maximize the number of ads they can serve. Affirming this, the findings here show that size of a subreddit does matter. Our results find that larger subreddits tend to be labeled as safe, even when they exhibit more toxic behavior. Turning to the competition dynamics in the two-sided market of content monetization, we offer support that

content monetization seems to be winning over advertiser concerns. While brand safety decisions at Reddit appear to be made with some regards to the standards that GARM have put forward, they are *also influenced* by the amount of advertising revenue a subreddit can generate. For advertisers, this highlights the need for a more comprehensive and transparent approach to brand safety decision-making, considering not only the content of a subreddit but also the potential social and financial implications of advertising choices.

Across the attributes explored here, we did find several highly toxic subreddits were *not blocked* (for a review, see Table 3). Overall, about 85% of the submissions posted in the subreddits studied here were indeed marked as ad safe. Reddit's current brand safety system marks most of the UGC it receives on its platform as safe and blocked "unsafe" subreddits tend to be smaller, niche subreddits. Reddit has a clear financial incentive to monetize as much submission content as possible, and here the amount of advertising impressions that a subreddit affords appears to be an important factor which ultimately drives the decision on whether Reddit allows a subreddit to be monetized.

Inspecting the top 100 most toxic subreddits, we see inconsistencies. Only two-thirds of the subreddits that we identified as having the most identity-based attacks and threatening language—two things that virtually all brand safety standards would not tolerate—were ultimately deemed unsafe. Only 56% of the most profane, insult-laden, and toxic subreddits were ad-blocked.

Turning back to what we know about advertising theory, we show that Reddit is consistently and persistently allowing advertisers to place ads in a variety of "unsafe" contexts, and therefore exposing brands to negative spillover effects. The spillover effect theory suggests that the association of a brand with negative content can affect the perception of the brand,

damaging its reputation and image (Ahluwalia, Unnava, and Burnkrant 2001). Yet, the spillover effect has been shown to be weak to null when intrusiveness is low, or when brand awareness is high. Reddit is clearly allowing a huge collection of ads to be shown on unsafe contexts. Yet, advertisers still happily place ads on Reddit, and by all accounts advertising on the platform is flourishing. This provides evidence that the spillover effect cannot be large in magnitude, but instead likely one that is minimal at best. If advertisers were being unintentionally punished by toxic Reddit content online, surely, they would be more aware of the imprecision of Reddit's brand safety tool. Yet, we could only find one news article that even discussed Reddit's brand safety approach.

Prior research suggests that ads on reddit, which look like many open web display banner ads, are likely not intrusive enough to exhibit strong spillover effects. Here we contribute to advertising theory by suggesting that spillover effects, and the general concern around brand safety, is overblown by the advertising industry, and that more conditions in which the spillover effect is null needs to be studied. This is urgent research. Open web brand safety is highly problematic in terms of the news content that it unintentionally blocks (e.g. Haggin, 2020; Parker, 2021; O'Reilly, 2021). Brand safety harms news journalism by demonetizing it. Is it a necessary evil? Toxic user generated content may not actually hurt brands and advertisers. Future research on brand safety tools, which directly affect journalistic outlets, should be studied and critically accessed for these reasons.

Social Media Technology Infrastructure Implications

Our study sheds light on the challenges and opportunities associated with implementing brand safety systems in a social media platform's technology infrastructure. On one hand, Reddit's approach to brand safety through human review and labeling of subreddits showcases

the potential for platforms to develop their own internal mechanisms to regulate advertising content. On the other hand, the inconsistencies observed in the ad-blocking of highly toxic subreddits highlight the limitations of manual review by the social media platform itself. Reddit does have a conflict of interest in that it wants to monetize as much content as possible. Both advertisers and the platforms themselves need for more sophisticated and transparent solutions.

Defining Brand Safety: The Role of Computational Detection

Brand safety, as a concept, is multifaceted and can be challenging to define precisely. In the context of advertising, brand safety generally refers to the strategies and measures implemented to prevent a brand's advertisements from appearing alongside content that could potentially harm the brand's reputation or image. This includes content that is controversial, offensive, or inappropriate, such as hate speech, explicit material, or misinformation.

However, the definition of what constitutes “unsafe” or in some cases “unsuitable” content can vary significantly between different advertisers, platforms, and industry standards. Furthermore, cultural, societal, and individual differences can also influence perceptions of what is considered offensive or inappropriate, adding another layer of complexity to the definition of brand safety.

In the era of programmatic advertising and user-generated content, the task of ensuring brand safety has become increasingly reliant on computational methods. Machine learning algorithms, as a subset of artificial intelligence (AI) systems, are commonly used to automatically review and categorize vast amounts of content, flagging potential violations of brand safety guidelines. These systems can detect certain keywords, phrases, or patterns associated with unsafe content, allowing platforms to block or restrict advertisements from appearing alongside such content.

However, while these computational methods can be highly effective in detecting explicit violations of brand safety, they are not infallible. One of the key challenges lies in the inherent limitations of AI and machine learning in understanding the nuances and context of human language. For instance, a computer may struggle to distinguish between a genuinely offensive comment and a sarcastic remark, or it may misinterpret cultural references or idioms.

Moreover, computational detection methods are typically based on predefined rules or patterns, which may not fully capture the evolving nature of online discourse. As language use changes and new forms of offensive or harmful content emerge, these systems may fail to recognize and flag such content as unsafe.

In addition, computational methods are often unable to account for the subjective nature of brand safety. What one brand considers unsafe may be perfectly acceptable to another, and vice versa. As such, a one-size-fits-all approach to computational detection may not adequately cater to the diverse needs and preferences of different brands.

Taken together, we urge the advertising industry to adopt approaches where humans are in the loop, alongside computers. If the industry approached the concept more like academics to content analysis, where concepts like intercoder reliability and external validity reign supreme, they would increase the likelihood that their systems would label content appropriately.

User Engagement Across Ad-Friendly and Ad-Blocked Subreddits

Our findings reveal that user engagement, as measured by upvotes, crossposts, and other submission-related activities, does not significantly differ between ad-friendly and ad-blocked subreddits. This has important implications for advertisers and social media platforms. For advertisers, it suggests that the ad-friendly status of a subreddit does not necessarily indicate a

higher level of user engagement. This challenges the common assumption that ad-friendly environments automatically equate to more active and engaged audiences.

For social media platforms, this finding underscores the complexity of managing user engagement and advertising revenue. It suggests that the decision to block or allow ads on a subreddit cannot be based solely on user engagement metrics. Other factors, such as the quality of content and the level of toxicity, must also be considered.

Avenues for Future Research

One potential area of exploration could be to investigate the impact of brand safety algorithms on other platforms and compare their effectiveness with Reddit's system. This could help identify best practices and potential areas of improvement in the field of brand safety.

Another interesting area of research could be to examine the influence of capitalistic motivations on brand safety decisions. This would involve exploring the trade-offs platforms make between advertising revenue and maintaining a safe environment for advertisers. It would be interesting to see how these decisions impact the overall brand safety landscape and whether there are better ways to balance these competing interests.

A third potential research direction could be to study the potential biases in the implementation of brand safety measures. This could involve looking at whether certain topics, communities, or demographics are under- or over-represented in ad-friendly subreddits. Such a study could provide valuable insights into the fairness and inclusivity of brand safety measures.

Future research could also aim to develop and test alternative brand safety models that incorporate more nuanced and context-aware approaches to the spillover effect. This could leverage advances in natural language processing and machine learning techniques to create more sophisticated and effective brand safety systems.

Finally, it would be valuable to analyze the long-term effects of brand safety measures on user behavior, advertiser trust, and platform sustainability. This could provide insights into the consequences of different brand safety strategies and help platforms make more informed decisions about their brand safety policies.

Limitations

This study utilized Jigsaw's Perspective API to detect toxic content. While effective and comprehensive, the Perspective API is not perfect and may miss out on nuanced or context-dependent instances of toxicity. It's also worth noting that our manual validation of the Perspective API's outputs had a precision range of 70 percent, leaving a relatively significant margin of error. We have taken the size and activity of a subreddit as indicators of a sub-community's function, disregarding other potential factors such as subreddit-specific moderation policies, community norms, or the age of the subreddit. The complex nature of Reddit's community dynamics may not be fully encapsulated in the factors we have considered.

Lastly, while our study observed patterns among toxicity, brand safety, and user engagement, it is largely correlational. It does not establish causal relationships between these variables. Further experimental or longitudinal studies could lend more light on how these factors influence each other over time. Despite these limitations, this study provides a foundational understanding of Reddit's brand safety system and its relationship with toxicity, user engagement, and subreddit characteristics. Future research should build on these findings to devise strategies that can improve the efficacy of brand safety systems in user-generated content platforms.

Tables and Figures

Table 1

A review of various "unsafe" advertising environments as defined by advertising industry groups.

Category	GARM (Global Alliance for Responsible Media)	4A's Advertiser Protection Bureau (APB)
Crime & Harmful Acts to Individuals and Society and Human Rights Violations	Graphic promotion, advocacy, and depiction of willful harm and actual unlawful criminal activity – Explicit violations/demeaning offenses of Human Rights (e.g., human trafficking, slavery, self-harm, animal cruelty etc.), Targeted harassment of individuals and groups	Unlawful criminal activity- murder, manslaughter and harm to others. Explicit violations- trafficking and slavery.
Debated Sensitive Social Issues	Insensitive, irresponsible and harmful treatment of debated social issues and related acts intended to demean a particular group or incite greater conflict	✕
Sensitive Social Issues/Violations of Human Rights	✕	Disrespectful and harmful treatment of sensitive social topics (e.g., abortion, extreme political positions, etc.) Acts, language, and gestures deemed illegal, not otherwise outlined in the framework (e.g., harm to self or other and animal cruelty) Targeted harassment of individuals and groups.

Table 2
The top 50 most toxic subreddits as derived by average toxicity score

subreddit	Avg_Inciv	Identity_Attack	Profanity	Threat	Insult	Toxicity	Ads present?	Submissions	Subscribers
bitchimabus	0.62	0.43	0.85	0.27	0.79	0.78	1	140	201081
nocontext	0.34	0.23	0.41	0.34	0.32	0.39	1	114	371208
Sexy	0.29	0.16	0.41	0.27	0.25	0.38	0	12425	394793
FiftyFifty	0.29	0.19	0.32	0.40	0.24	0.28	0	451	2077050
racism	0.28	0.40	0.21	0.20	0.29	0.28	0	248	41447
sex	0.27	0.19	0.39	0.22	0.22	0.35	0	12166	2172105
rant	0.27	0.22	0.31	0.22	0.30	0.31	0	3272	256761
confession	0.27	0.20	0.29	0.33	0.23	0.29	1	3641	3271849
emogirls	0.26	0.16	0.34	0.25	0.24	0.32	0	2750	396098
confessions	0.26	0.21	0.29	0.29	0.22	0.28	1	5191	951712
circlejerk	0.25	0.20	0.29	0.24	0.26	0.28	0	960	459519
TwoXSex	0.25	0.18	0.36	0.20	0.20	0.33	0	222	120703
FloridaMan	0.25	0.29	0.19	0.34	0.25	0.19	0	219	751364
exmuslim	0.25	0.36	0.19	0.22	0.23	0.23	1	1739	115915
bigdickproblems	0.25	0.15	0.35	0.18	0.23	0.32	0	842	234694
askgaybros	0.24	0.25	0.27	0.19	0.24	0.27	0	5930	325292
ProRevenge	0.24	0.15	0.21	0.33	0.24	0.27	1	125	1279959
IdiotsFightingThings	0.24	0.16	0.24	0.31	0.25	0.26	1	140	1016928
JessicaNigri	0.24	0.19	0.32	0.19	0.22	0.26	0	308	224614
Hot_Women_Gifs	0.23	0.15	0.31	0.28	0.19	0.24	0	132	97189
GunsAreCool	0.23	0.17	0.13	0.51	0.17	0.19	0	288	43802
straya	0.23	0.17	0.29	0.20	0.25	0.25	1	209	81539
shittyaskreddit	0.23	0.16	0.26	0.21	0.25	0.27	0	424	68185
tifu	0.23	0.15	0.23	0.27	0.19	0.29	1	3183	18016834
Bad_Cop_No_Donut	0.22	0.20	0.16	0.37	0.21	0.19	0	640	531305
pranks	0.22	0.16	0.22	0.27	0.24	0.22	1	151	34365
gay	0.22	0.25	0.23	0.19	0.21	0.22	0	1421	309064
shittyama	0.22	0.16	0.27	0.20	0.21	0.25	1	164	45268
MensRights	0.21	0.26	0.16	0.22	0.22	0.20	0	1190	332028
atheism	0.21	0.31	0.16	0.20	0.20	0.19	1	2534	2731332
fightporn	0.21	0.15	0.20	0.31	0.19	0.21	0	878	1546778
howtonotgiveafuck	0.21	0.13	0.28	0.17	0.23	0.25	1	244	663684
SuicideWatch	0.21	0.13	0.20	0.36	0.16	0.21	0	10466	361074
pettyrevenge	0.21	0.13	0.19	0.28	0.21	0.23	1	510	1722438
happygirls	0.21	0.12	0.28	0.21	0.18	0.26	0	222	196862
Feminism	0.21	0.27	0.16	0.22	0.20	0.19	1	512	241308
fitnesscirclejerk	0.21	0.14	0.23	0.20	0.24	0.23	0	214	33211
ShittyLifeProTips	0.20	0.14	0.18	0.26	0.24	0.20	1	624	1633791
offmychest	0.20	0.16	0.21	0.25	0.18	0.21	0	18808	2761944
nottheonion	0.20	0.20	0.16	0.28	0.19	0.18	1	3243	21549277
Palestine	0.20	0.28	0.12	0.27	0.16	0.17	0	432	123686
bigboobproblems	0.20	0.12	0.27	0.18	0.17	0.26	0	218	125562
popping	0.20	0.15	0.19	0.24	0.19	0.21	1	655	530202
transgender	0.20	0.29	0.15	0.18	0.18	0.19	1	239	143201
serialkillers	0.19	0.15	0.12	0.33	0.19	0.17	0	218	559157
actuallesbians	0.19	0.24	0.19	0.17	0.18	0.19	1	2621	412956
TwoXChromosomes	0.19	0.19	0.18	0.24	0.18	0.18	1	4455	13388528
SubredditDrama	0.19	0.18	0.16	0.22	0.21	0.18	1	393	868011
trailerparkboys	0.19	0.13	0.24	0.17	0.21	0.20	1	270	218214
gaybros	0.19	0.21	0.19	0.17	0.18	0.19	0	1308	348761

Table 3

Mean toxicity scores across ad-friendly and ad-blocked subreddits

	Ad-Friendly	Ad-Blocked	
Toxicity Attribute	M_P	M_P	*d*
Identity Attack	0.10 (0.11)	0.13 (0.14)	0.28
Threatening Language	0.16 (0.15)	0.20 (0.18)	0.26
Profanity	0.10 (0.14)	0.17 (0.23)	0.43
Insulting Language	0.11 (0.14)	0.15 (0.18)	0.26
Toxicity	0.10 (0.13)	0.16 (0.20)	0.46

Notes: M_P = mean probability score for group; *d* = Cohen's *d*; standard deviations are in parentheses next to group-level means; all *p* values < .001.

Table 4

Mean engagement scores across ad-friendly and ad-blocked subreddits

	Ad-Friendly	Ad-Blocked	
Engagement Measure	M	M	*d*
Number of Upvotes	130.00 (1694.00)	46.60 (396.00)	0.05
Number of Responses	16.50 (173.00)	11.30 (67.60)	0.03
Number of Crossposts	0.04 (0.65)	0.02 (0.26)	0.04

Notes: M = mean score for group; *d* = Cohen's *d*; standard deviations are in parentheses next to group-level means; all *p* values < .001.

Table 5

Correlations between subreddit attributes and toxicity factors

	Number of Subscribers	Number of Posts
Identity Attack	.06***	.04*
Threat	.08***	.06***
Profanity	.06***	.07***
Insult	.07***	.08***
Toxicity	.06***	.09***

Notes: estimates are τ correlations; * = *p* < .05, *** *p* < .001

References

Aaron, M. (2010). IAB Networks & Exchanges Committee Develops Guidelines and Proposes Compliance Program. IAB. https://www.iab.com/news/iab-networks-exchanges-committee-develops-guidelines-proposes-compliance-program/

Ahluwalia, R., Unnava, H. R., & Burnkrant, R. (2001). The Moderating Role of Commitment on the Spillover Effect of Marketing Communications. Journal of Marketing Research, 38(4), 458–470. https://doi.org/10.1509/jmkr.38.4.458.18903

Almerekhi, H., Jansen, B., & Kwak, H. (2020). Investigating Toxicity across Multiple Reddit Communities, Users, and Moderators. WWW '20: Companion Proceedings of the Web Conference 2020, 294–298. https://doi.org/10.1145/3366424.3382091

Barwick, R. (2022). Reddit's Advertising Policy Seems to Differ from Subreddit to Subreddit. Marketing Brew. https://www.marketingbrew.com/stories/2022/07/19/reddit-s-advertising-policy-seems-to-differ-from-subreddit-to-subreddit

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. (2020). The Pushshift Reddit Dataset. Proceedings of the Fourteenth International AAAI Conference on Web and Social Media, 830–839.

Bellman, S., Abdelmoety, Z., Murphy, J., Arismendez, S., & Varan, D. (2018). Brand Safety: The Effects of Controversial Video Content on Pre-Roll Advertising. Heliyon, 4(12). https://doi.org/10.1016/j.heliyon.2018.e01041

Bishop, S. (2021). Influencer Management Tools: Algorithmic Cultures, Brand Safety, and Bias. Social media + Society, 7(1). https://doi.org/10.1177/20563051211003066

Cameron, N. (2017). Consumers Boycotting Brands for Placing Ads Near “Offensive” Content. CMO. https://www.cmo.com.au/article/620574/report-consumers-boycotting-brands-placing-ads-near-offensive-content/

Chandrasekharan, E., Pavalanathan, U., Srinivasan, A., Glynn, A., Eisenstein, J., & 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, 1(CSCW), 1-22.

Davidson, S., Sun, Q., & Wojcieszak, M. (2020). Developing a New Classifier for Automated Identification of Incivility in Social Media. Proceedings of the Fourth Workshop on Online Abuse and Harms, 95–101. https://doi.org/10.18653/v1/2020.alw-1.12

Field, H. (2023, August 1). Twitter, now called X, sues researchers who showed rise in hate speech on platform after Musk takeover. CNBC. https://www.cnbc.com/2023/08/01/x-sues-ccdh-for-showing-hate-speech-rise-on-twitter-after-musk-deal.html

Gaudette, T., Scrivens, R., Davies, G., & Frank, R. (2021). Upvoting Extremism: Collective Identity Formation and the Extreme Right on Reddit. New Media & Society, 23(12), 3491–3508. https://doi.org/10.1177/1461444820958123

Guaglione, S. (2020). Incorrect Keyword Blocking Costs U.S. Publishers $2.8 Billion. MediaPost. https://www.mediapost.com/publications/article/345966/incorrect-keyword-blocking-costs-us-publishers.html

Haggin, P. (2020). Target, MTV Blocked Ads from News Mentioning “George Floyd” and “Protests”. The Wall Street Journal. https://www.wsj.com/articles/target-mtv-blocked-ads-from-news-mentioning-george-floyd-and-protests-11594576272

Haggin, P. (2022). USA Today Owner Gannett Co. Gave Advertisers Inaccurate Information for Nine Months. The Wall Street Journal. https://www.wsj.com/articles/usa-today-owner-gannett-co-gave-advertisers-inaccurate-information-for-nine-months-11646784745

Hagiui, Andrei (2018). *The Palgrave Encyclopedia of Strategic Management*. Cambridge, Mass.: Macmillan Publishers Ltd. pp. 1104–1107.

Hmielowski, J., Hutchens, M., & Cicchirillo, V. (2014). Living in an Age of Online Incivility: Examining the Conditional Indirect Effects of Online Discussion on Political Flaming. Information, Communication, & Society, 17(10), 1196–1211. https://doi.org/10.1080/1369118X.2014.899609

Hopp, T., Vargo, C. J., Dixon, L., & Thain, N. (2020). Correlating Self-Report and Trace Data Measures of Incivility: A Proof of Concept. Social Science Computer Review, 38(5), 584–599. https://doi.org/10.1177/0894439318814241

Janssens, W., & De Pelsmacker, P. (2007). Fear Appeal in Traffic Safety Advertising: The Moderating Role of Medium Context, Trait Anxiety, and Differences Between Drivers and Non-Drivers. Psychologica Belgica, 47(3), 173–193. https://doi.org/10.5334/pb-47-3-173

Koch, L. (2019). Offensive Content on Platforms Turns Off Consumers, Tarnishes Ads. eMarketer. https://www.emarketer.com/content/does-bad-content-affect-consumer-perceptions-of-brand-safety

Lee, C., Kim, J., & Lim, J. S. (2021). Spillover Effects of Brand Safety Violations in Social Media. Journal of Current Issues & Research in Advertising, 42(4), 354–371. https://doi.org/10.1080/10641734.2021.1905572

Madio, L., & Quinn, M. (2021). Content moderation and advertising in social media platforms. Managerial Marketing eJournal.

Manatt, K., Avital, D., & Ofer, B. (2018). The Brand Safety Effect: How Unsafe Ad Placement Impacts Consumer Brand Perception. MAGNA. https://magnaglobal.com/wp-content/uploads/2021/04/The-Brand-Safety-Effect-CHEQ-Magna-IPG-Media-Lab-BMW-Logo-101018.pdf

O'Reilly, L. (2021). Advertisers are Putting Climate Crisis Content on their Blocklists. Publishers Fear it's "Defunding" their News Coverage. Insider. https://www.businessinsider.com/publishers-advertisers-keyword-blocking-climate-crisis-news-2021-9

Parker, B. (2021). How Advertisers Defund Crisis Journalism. The New Humanitarian, January 27. https://www.thenewhumanitarian.org/analysis/2021/01/27/brand-safety-ad-tech-crisis-news

Reddit for Business. (2020). Introducing Inventory Types. Reddit, September 24. https://www.redditinc.com/blog/introducing-inventory-types/

Rochet, J.-C., & Tirole, J. (2003). Platform Competition in Two-sided Markets. *Journal of the European Economic Association*, *1*(4), 990–1029. http://www.jstor.org/stable/40005175

Stevens, H., Acic, I., & Taylor, L. D. (2021). Uncivil Reactions to Sexual Assault Online: Linguistic Features of News Reports Predict Discourse Incivility. Cyberpsychology, Behavior, and Social Networking, 24(12), 815–821. https://doi.org/10.1089/cyber.2021.0075

Taffera-Santos, N. (2021). US Programmatic Digital Display Advertising Outlook 2021. https://on.emarketer.com/rs/867-SLG-901/images/eMarketer%20US%20Programmatic%20Digital%20Display%20Advertising%20Outlook%202021%20Report.pdf

Vanian, J. (2023, September 13). Elon Musk's X sued this nonprofit after it exposed hate speech, and its new research shows little has changed. CNBC. https://www.cnbc.com/2023/09/13/x-formerly-twitter-hate-speech-running-rampant-ccdh.html