4-24-2023

A Look Through a Broken Window: The Relationship Between Disorder and Toxicity on Social Networking Sites

Nils Messerschmidt
University of Bamberg, nils.messerschmidt@stud.uni-bamberg.de

Jana Gundlach
Weizenbaum Institute for the Networked Society, janagundlach@uni-potsdam.de

Annika Baumann
Weizenbaum Institute for the Networked Society, annika.baumann@uni-potsdam.de

A LOOK THROUGH A BROKEN WINDOW: THE RELATIONSHIP BETWEEN DISORDER AND TOXICITY ON SOCIAL NETWORKING SITES

Research in Progress

Messerschmidt, Nils, University of Bamberg, Germany, nils.messerschmidt@stud.uni-bamberg.de.

Teitscheid, Jana, Weizenbaum Institute for the Networked Society, University of Potsdam, Germany, jana.teitscheid@uni-potsdam.de.

Baumann, Annika, Weizenbaum Institute for the Networked Society, University of Potsdam, Germany, annika.baumann@weizenbaum-institut.de.

Abstract

Toxicity has increased on social networking sites (SNSs), sparking a debate on its underlying causes. While research readily explored eligible social factors, disorder induced by the very nature of SNSs has been neglected so far. The relationship between disorder and deviant behaviors could be revealed within the offline sphere. Incorporating the theoretical lens of the Broken Windows Theory, we propose that a similar mechanism is prevalent in the online context. To test the hypothesis that perceived disorder increases toxicity on SNSs, the study compares two subcommunities on Reddit dedicated to the same topic that differ in their perceived disorder. Sampling the toxicity scores via data collection and natural language processing yields the first evidence for our hypothesis. We further outline subsequent studies that aim to investigate further the phenomenon of how disorder-related factors contribute to toxic online environments.

Keywords: Toxicity, Perceived Disorder, Reddit, Broken Windows Theory.

1 Introduction

Social media, particularly social networking sites (SNSs), promote unprecedented interactions between individuals, creating a space for social encounters. SNSs are primarily known for their potential to provide social support (Liu et al., 2018). However, where there is light, there is also shadow. Especially in recent years, disrespectful tones, hate speech, trolling, and other deviant behavior have vastly increased on SNSs (Herring et al., 2002; Lowry et al., 2017). All such deviant behaviors can be classified under the umbrella term toxicity, which refers to behavior incongruent with society's dominating values and norms (Lowry et al., 2017). While prevailing in different forms, all share that toxicity deteriorates individual interactions. In the online context, 41% of US adults have personally experienced harassment, with 25% even experiencing severe forms. Meanwhile, 79% agree that SNSs are not doing enough against cyberbullying on their platforms (Vogels, 2021).

Extant research on cyberbullying and online harassment identified social influences as drivers for toxic behavior, such as anonymity, limited control, and accountability (Barlett et al., 2016; Lowry et al., 2017). While eligible social factors have been the focus, negligibly research is yet attributed to the design features of online platforms that could drive this phenomenon. Within the offline sphere, the Broken Windows Theory (BWT; Wilson and Kelling, 1982) suggests that environmental decay (e.g., vandalism) may trigger deviant behavior. On an abstract level, it suggests that disorderliness in an environment indicates a certain norm that induces people to act similarly disorderly or even in an antisocial manner

(Cialdini et al., 1990). Transferring this to the online environment, especially SNSs, deteriorated, chaotic, and disordered SNS environments may signal a certain norm and lead users to behave antisocially by contributing toxic content or engaging in toxic interactions. Following this logic, this paper aims to investigate whether design features on SNSs promote toxic interactions by asking the following research question:

Does perceived disorder on SNSs increase toxicity?

To take a first step in investigating this research question, we pair subcommunities of an SNS aimed at the same topic but varying levels of perceived disorder against each other to analyze differences in toxicity among them using natural language processing methods. We find toxicity is observed more frequently in the subcommunity with the higher perceived disorder. We further find that the more toxic subcommunity acquires fewer new users, but its existing users become more active. Thus, the subcommunity is becoming increasingly more homogeneous, potentially leading to inherently higher toxicity levels.

This research in progress provides a first step in investigating the connection between perceived disorder within the SNS environment and their respective toxicity scores. Further, it introduces the theoretical lens of the BWT within the SNS domain. The study dives into the highly relevant topic of antisocial online behavior. It goes beyond the prevalent social explanations and investigates the design-related drivers on the systems side. By focusing on this aspect, we extend the existing knowledge to understand toxicity within the SNS environment. Suggesting a starting point for developing an objective measurement of perceived disorder, this paper further motivates future research avenues investigating the relationship between perceived disorder and antisocial behavior on SNSs. More specifically, the paper suggests how to spotlight factors that cause perceived disorder on SNSs and, therefore, could provide starting points for countermeasures. For example, if perceived disorder increases toxicity, then reversely, countering factors that cause perceived disorder may decrease the level of toxicity. This is of interest to platform operators and policymakers in emphasizing the importance of SNS design in setting certain signals in the online environment.

2 Theoretical Background

2.1 Toxicity

Toxicity comprises deviant behaviors incongruent with society's dominating values and norms (Lowry et al., 2017). Such deviant behavior occurs in various online environments and can manifest in multiple ways. One form of toxicity constitutes trolling, a deliberate provocation of individuals (Herring et al., 2002). Another form of toxic behavior includes hate speech, which devalues certain groups or individuals based on ethnic, religious, or gender stereotypes. Further, actions related to cyberbullying or cyberstalking (Lowry et al., 2017) occur related to various topics such as religion, celebrities, news, current events, and personal or emotional information (Statista, 2017).

While toxicity prevails in different forms, each manifestation is unfavorable and deteriorates individual interactions. Yet, despite the increasing need to investigate and understand toxicity in the online environment, there is no predominantly used definition. More recently, scholars have adapted the definition of toxicity as "a rude, disrespectful, or unreasonable comment that is likely to make one leave a discussion" (Dixon et al., 2018, p. 68; Hosseini et al., 2017, p. 2). The 'Perspective' API has also employed this definition, a research project developed by Google researchers and news outlets, including The New York Times and The Wall Street Journal (Google Jigsaw, 2022).

2.2 Broken Windows Theory and norm-setting

The Broken Windows Theory (Wilson and Kelling, 1982) originates in criminology and postulates that a neighborhood's crime rates increase due to its environmental disorder condition (Welsh et al., 2015). The theory has been developed based on an experiment by Zimbardo (1973) demonstrating that seemingly abandoned cars were quickly stripped for parts and further vandalized when left in high-crime instead of low-crime neighborhoods. In addition, numerous studies confirm that individuals increased littering behavior within an already littered environment instead of a clean environment (Cialdini et al., 1990; Keizer et al., 2008).

According to the BWT, disorder is characterized by physical and social disorder components that might cause deviant behavior. The first one, which also the name of the theory relies heavily on, is related to the immediate environmental state. As Wilson and Kelling (1982) stated, several more will likely follow if a broken window is apparent. The first pathway relates to the more indirect second pathway of social disorder. Since environmental decay signals carelessness, a deteriorated state of a physical place affects the perception of an increased likelihood of criminal activities (Costa, 1984). The swayed perception might cause law-abiding citizens to limit their presence in areas affected by disorder and simultaneously attract individuals with deviant intentions due to lessened social control (Wilson and Kelling, 1982). In turn, lessened social control might result in more criminal activities happening.

In general, the BWT states that a disordered environment indicates a prevalent norm that induces people to act similarly in a disordered manner (Cialdini et al., 1990). Norms can affect human behavior systematically (Reno et al., 1993). Related research concludes that injunctive, descriptive, and personal norms impact the littering behavior of individuals (Reno et al., 1993). Based on the results, although norms may contradict, the focal norm will dictate the individual's behavior (Reno et al., 1993).

Other fields of academia have applied the theory within their specific context, such as law (Harcourt, 2006), social psychology (Sampson et al., 2015), and public health, where the disorder of a neighborhood has been related to sexual disease transmission (Cohen et al., 2000). In Information Systems (IS) research, the BWT has been applied in the context of website design and security behavior regarding passwords (Grimes et al., 2014). The study finds that poor website design signals that insecure behavior is the norm.

In summary, within the offline context, the BWT states that disorder comprises two pathways: the direct pathway of environmental deterioration and the more indirect one relating to increased social incivilities. Therefore, adapted to the online context of SNSs, we assume disorder follows a similar composition and define perceived disorder as a combination of the two pathways of the BWT, i.e., design-related deterioration and the resultant increased incivility of users which, in turn, leads to more toxicity.

2.3 Broken Windows Theory on SNSs and toxicity

SNSs are a breeding ground for toxic behavior due to low entry barriers, such as the anonymity of their user base (Barlett et al., 2016). Consequently, the phenomenon has been researched on multiple platforms, such as Twitter (Chatzakou et al., 2017; Davidson et al., 2017) and YouTube (Chen et al., 2012; Dinakar et al., 2011). Especially Reddit has been revealed to be a flourishing platform for hate speech (Chow, 2022) and toxicity (Almerekhi et al., 2019; Chipidza, 2021; Xia et al., 2020). Despite some efforts to diminish those effects by closing known toxic communities and updating its hate speech policies (Stephan, 2020), toxicity is fostered by Reddit's fundamental characteristics (Massanari, 2017). While SNSs such as Facebook, YouTube, or Instagram rely on algorithms that direct the attention of their user base toward popular content, users on Reddit typically vote on the popularity of topics and interact in separate communities aimed at specific topics that are the so-called subreddits. As these subreddits can be created independently, highly related communities by topic can co-exist (Hessel et al., 2016). The prefix "r/" is added to the community's name to indicate a subreddit.

The design configurations of subreddits can vary tremendously in terms of color, shape, background, and arrangement. Posts need not be consistent and can vary in size, form, media type, and visual aspects. Usually, subreddits follow specific rules and are moderated. Moderators can individualize rules of usage, the degree of moderation and customize visual features (e.g., background images, colors, post flairs) within their respective subreddit. Consequently, subreddits, even the ones dedicated to similar topics, can vary heavily in appearance.

Given these unique characteristics of Reddit, we have chosen this particular SNS platform to investigate our research question. It offers the opportunity to select subreddits dedicated to similar topics while simultaneously having varying levels of perceived disorder.

3 Methodology

3.1 Pre-test and considerations

Different topics might lead to varying toxicity levels as, for example, they are more polarizing than others. Polarization can be seen as individuals or groups adopting more extremist views (Kitchens et al., 2020). Similarly, specific topics might exert more extreme viewpoints that additionally might get reflected in the way how individuals interact with each other. For example, the toxicity levels of an argument concerning politics might, under certain circumstances, not be comparable to that of an online book community. Therefore, we assume that more polarized topics bear a higher likelihood of disagreeable standpoints, with exposure to opposing views leading to aversive reactions (Minson and Dorison, 2022). Aversion can result in a higher possibility of objecting and attacking others (Mackie et al., 2000), reflecting tendencies that, ultimately, might result in more toxic behavior.

Considering this aspect and since we aim to disentangle the contributing factor of perceived disorder on toxicity on SNSs, selecting subcommunities on Reddit (i.e., subreddits) dedicated to a very similar, if not the same, topic was deemed essential. The identified subreddit pairs had to fulfill certain requirements to be considered suitable. Firstly, as we needed varying levels of perceived disorder, the subreddit pairs should differ in design configurations across our selected observation years (e.g., color, shape, background). To check that the discrepancy in terms of design-related aspects stayed as similar as possible throughout the years, we checked available snapshots of subreddit candidates on archive.org. Secondly, the subreddit pairs should have as similar descriptive metrics as possible (e.g., community size, creation date) to maximize similarity regarding the respective user base, as a newly created or small community functions differently than an established one. Additionally, all subreddits should use English as the primary language and predominantly use textual means of communication, rather than images, videos, or hyperlinks, to be accessible for natural language processing methods and minimize external influences from consuming third-party content.

In the preliminary stage of analysis, several content pairs of subreddits with a wide variety of topics (e.g., friendship, relationships, science, books) were identified and investigated. First, data were collected for all identified pairs containing 75 thousand posts to validate feasibility. Subsequently, the sentiment was analyzed using word clouds and libraries such as Vader (Hutto and Gilbert, 2014). Finally, after extensive research and monitoring of the identified pairs, we have chosen to investigate the two subreddits within the topical field of "venting" for analysis based on the mentioned requirements and preliminary results. More specifically, the subreddits r/venting and r/vent are the contexts of choice.

This decision is based on the following factors. First, both subreddits serve the same purpose and display similar structural characteristics and sentiments. The community size is similar (at the time of this study, 122 thousand members for r/vent and 75 thousand for r/venting). r/venting was created in 2011, while r/vent was created in 2008, making them both long-lasting subreddits. We further found 425 users who are active in both subreddits simultaneously, indicating that the topics are indeed overlapping. This share was the highest overlapping number of users relative to the community size across all investigated pairs.

As outlined above, the BWT considers two pathways that might lead to increased toxicity, namely environmental and social decay. Here, we will mainly focus on environmental decay through design-related aspects. Therefore, an essential requirement was a high similarity of the selected subreddit pair in content and descriptive characteristics and varying levels of perceived disorder in relation to visual clutter. An extensive discussion among the paper's authors resulted in an informal categorization of each considered subreddit on the level of perceived disorder. The vent-related pair displayed the highest discrepancy in design configurations from all pairs investigated. Regarding visual features, r/vent uses an individual header image, a customized background color, and post flairs with varying colors, whereas r/venting entirely relies on the generic visual representation. In addition, r/vent distinguishes itself by stating numerous moderating rules shown on the right side, leading to additional visual clutter, whereas r/venting does not. For both subreddits, snapshots on archive.org were available in the observation period from 2020-2022. In the case of r/venting, the only snapshots available in 2020 are from

December. Otherwise, we selected several available snapshots from July for all observation years. The subreddit r/vent had changing moderators and continued to employ several moderation rules throughout, thus leading to visual clutter, whereas r/venting had consistently the same moderation team and maintained very few, but straightforward rules. More importantly, the discrepancy in perceived disorder between the two subreddits was continuously apparent for all snapshots observed based on extensive discussions among the paper's authors.

3.2 Data collection and toxicity analysis

All published data on Reddit is publicly available. The data was collected using Communalytic (Gruzd et al., 2020) and Pushshift (Baumgartner et al., 2020) and is based on all submissions, comments, and replies to comments that were published in each of the two subreddits. As the data basis is considerably large, we selected a specific month of analysis over three years. Specifically, we accessed the data published on r/vent and r/venting from 1-28 July in the years 2020 until 2022. The analysis across multiple years was done to examine whether changes in the degree of toxicity over time have taken place and, if so, whether this trend was accompanied by structural changes of the respective subreddit (e.g., an influx of new users, average user activity, comment intensity).

The degree of toxicity is measured using the Perspective API, which contains machine learning modules to identify toxic behavior. Specifically, the model distinguishes between the following attributes: toxicity, severe toxicity, identity attack, insult, profanity, and threat, as termed by the developers (Google Jigsaw, 2022). Importantly, as we consider toxicity the umbrella term for behavior incongruent with society's dominating values and norms (Lowry et al., 2017), all attributes used by the Perspective API, even though partially having the same or similar names, constitute subcategories of our definition of toxicity. Employing the model on our data, each post gets assigned a value between 0 and 1 according to the degree of each attribute, e.g., a severe toxicity score of 0.4 indicates that 4 out of 10 people would rate this content as severely toxic. Posts within the range of 0.3 to 0.7 are classified as uncertain, while a rating between 0.7 and 1 indicates that an attribute is most likely prevalent. For our study, we selected 0.7 as a threshold that demonstrates the prevalence of an attribute, as it is also commonly employed within social sciences (Gil et al., 2017; Hua et al., 2020).

Sub-reddit	Community Size	Year	Total Posts Collected	Total Active Users	Average Posts Per User	Average Post Length	Average Post Score	Average User Karma	Vader Sentiment
r/vent	47k	2020	14,777	3,726	3.97	353.06	4.01	17,076	0.06
	75k	2021	24,934	5,699	4.38	304.00	3.82	15,958	0.06
	122k	2022	49,899	11,842	4.21	330.83	3.37	10,763	0.04
r/venting	8k	2020	1,587	503	3.16	452.46	2.92	16,851	0.12
	28k	2021	8,484	2,705	3.14	345.73	4.27	11,982	0.15
	75k	2022	12,225	4,110	2.97	395.49	4.68	10,479	0.09

Table 1. Descriptives of the subreddits r/vent and r/venting.

4 Results

As seen in Table 1, each year, both subreddits have grown considerably. On average, the number of posts in r/vent approximately doubled each year, accumulating almost 50 thousand posts within one month in 2022. Despite r/vent roughly having only a 60% bigger community size in 2022, its users are significantly more active, as seen by the average post per user. Meanwhile, r/venting has attracted more new active users over time than its counterpart. Whereas in 2020, it had just 17% of the active users compared to r/venting, that percentage has increased to 35% in 2022. Whereas the average post score has slightly decreased for r/vent, it has increased in r/venting from 2.92 in 2020 to 4.68 in 2022. For the Vader sentiment compound score (ranging from -1, indicating negative sentiment, to +1, indicating positive sentiment), a downward trend can be observed for r/vent, only slightly above 0 throughout the years. For r/venting, the score is well above 0.1 and rising over time (except for 2022).

Figures 1 and 2 present the toxicity scores for the two subreddits across 2020-2022 among the toxicity attributes. While Figure 1 shows the number of total posts and therefore accounts for the growth of both subreddits, Figure 2 represents the toxicity as a percentage of all posts. Our results indicate that the toxicity scores are higher for the r/vent subreddit across all levels and that the margins between those subreddits did not decrease over time. Of all the toxicity sub-groups investigated, profanity is the most prevalent at the 0.7 threshold, affecting up to 20% of all the traffic created within the subreddits. About 10% of the traffic in r/vent is identified as toxic, with about 7% for r/venting. More than half of the toxic content is severely toxic for both subreddits. Forms of insults seem to be quite prevalent across both subreddits but more so for r/vent. Moreover, for identity attacks and threat, while the scores were relatively low in 2020, they seem to be increasing over the years for both subreddits.

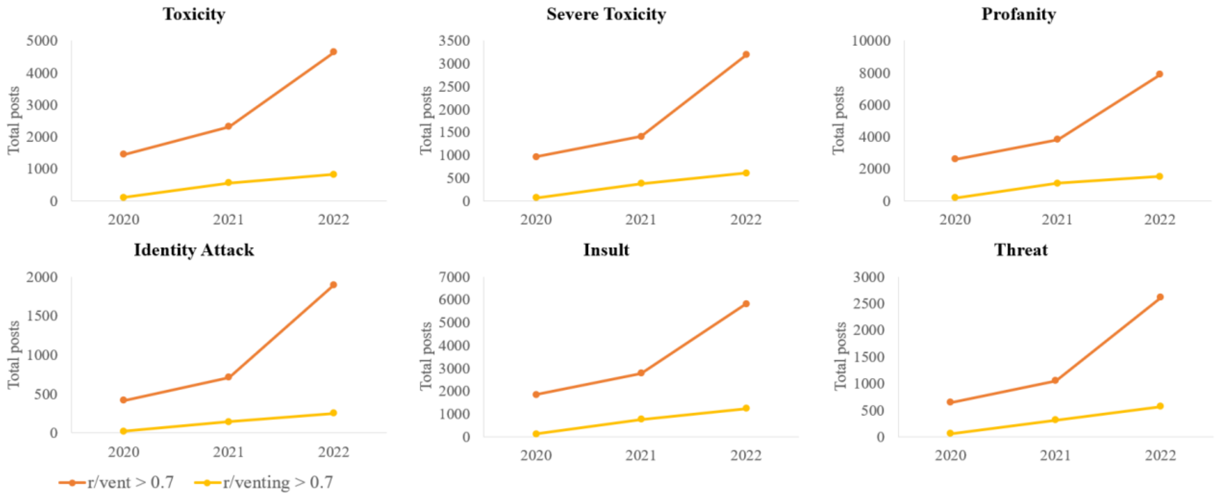

Figure 1. Comparison of toxicity scores across subreddits r/vent and r/venting (in total posts).

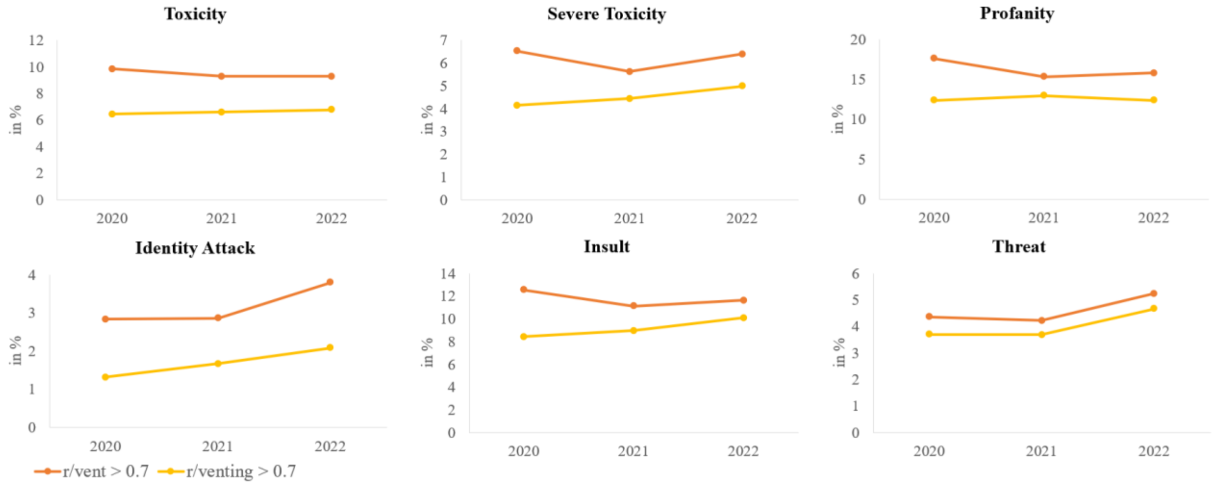

Figure 2. Comparison of toxicity scores across subreddits r/vent and r/venting (as a percentage).

5 Discussion

Our results present initial evidence for the prevalence of the BWT phenomenon occurring within the digital sphere based on the SNS Reddit. Across all toxicity measures, r/vent – the subreddit identified as more disordered – exceeds the subreddit r/venting in toxicity scores across 2020-2022. Moreover, additional robustness for that statement is provided, as the severe toxicity attribute displays the same margins as the basic toxicity attribute when comparing both subcommunities over time. This observation yields initial evidence for the BWT. Perceived decay and chaos on the SNS platform may, according to the theory, foster antisocial behavior that manifests in toxic expressions.

In addition to evidencing this hypothesized link, the results also confirm extant research findings. More specifically, identity attacks are increasing over the years, confirming studies establishing this link in political communities (Mittos et al., 2020; Tahmasbi et al., 2021). Furthermore, the average number of posts per user steadily increases in the community with higher toxic behavior. Combining this with the results from Figures 1 and 2, our research shows that the average toxicity exerted by each user is increasing. Simultaneously, our structural analysis reveals that the subreddit with lower toxic behavior receives significantly more new users over time. This could indicate that toxic communities get smaller and more homogenous and thus increase their toxicity inherently, as suggested by prior research (Horta Ribeiro et al., 2021; Atari et al., 2022). Finally, our data indicate that among the 425 users present in both communities, those users generally receive fewer upvotes in the less toxic subcommunity. This could indicate norm-setting and that these users are not as accepted within the community due to their deviant behavior toward the norm.

Theoretically, insights from our research will add to the understanding of the body of literature on dark sites of SNSs (O'Donnell and O'Donnell, 2022). More specifically, by offering first exploratory empirical insights, the study has the potential to enhance our understanding of the factors that cause toxicity on SNSs beyond structural factors identified through prominent social network analysis approaches (Almerekhi et al., 2020; Mittos et al., 2020) or social factors (Chan et al., 2021) by particularly considering disorder perceptions. Additionally, it generates a more nuanced understanding of how toxicity can lead to even more toxicity, creating a vicious circle of antisocial behavior. While this is in fashion with Horta Ribeiro et al. (2021), our work argues in terms of the BWT and hence, sheds light on a novel perspective by considering the design features of SNSs. So far, the BWT has not been prominently used in IS research (except for Grimes et al., 2019), albeit the circumstance that the digital environment is increasingly replacing our external environment in terms of social interactions and other advances. This study thus provides an exploratory first step and paves the way for linking disorder and toxicity in the online context in subsequent studies.

Practically, our insights add to the understanding of the effects of dynamic interactions on the platform Reddit beyond well-researched social factors. The call for the importance of considering the design is predominantly critical for three stakeholders. First, platform providers should rethink their responsibility to provide adequate structures that allow a clean comprehension of information. Similarly, in the offline context, the owner of an offline environment, like a parking lot, may be liable for keeping the environment clean and orderly and getting rid of any litter. Whether this responsibility needs to be shared with policymakers is yet to be debated. Similar to measurements taken in New York in the sixties to counter criminal behaviors by cleaning neighborhoods (Harcourt and Ludwig, 2006), policymakers and platform providers may consider the role of a clean and orderly online environment in preventing toxic interactions. Finally, users are responsible for contributing to the community by not littering, loitering, or vandalizing – offline and online.

6 Future Study Plan

The study plan is three-fold. In the first step, we aim to administer a survey to 300 panel members at Prolific. The survey's main goals will be to (a) develop an objective measure of perceived disorder and (b) gain more empirical evidence for the perceived disorder. The study consists of two parts. Participants are asked to rate subreddits regarding their perceived disorder in the first part. After briefly browsing subreddits for at least 2 minutes in randomized order, participants are asked to answer the following item, *"I feel that the subreddit is disordered."* on a 7-point Likert scale (Strongly agree – Strongly disagree). Following this rating task, in the second part, participants are requested to answer the following open-ended questions: "How come the subreddit was perceived as disordered by you?" and

"What has caused this perception?". Participants' answers are then coded inductively using open coding. These codes are then clustered in higher-level codes and overarching categories. The answers to these questions will contribute towards answering our research question and provide dimensions of perceived disorder that will help develop a measurement instrument.

In the second step, we aim to utilize the newly developed measurement instrument to use it in an experimental design with two groups administered to panel members at Prolific. Both groups will receive the task of browsing an artificial but real-life-inspired subreddit page. We aim to keep as many aspects constant between the two groups as possible, e.g., the subreddit will contain duplicate threads. However, both groups will differ regarding the perceived disorder related to design-related aspects. The differing levels of perceived disorder will be pre-tested using the measurement instrument developed in the first study. In this regard, whereas the control group receives an ordered subreddit, the experimental group will receive the disordered version. After browsing the subreddit for a few minutes, both experimental groups will be requested to leave a comment below one of the threads of the subreddit. The comment will then be coded according to their level of toxicity.

We aim to return to the data-driven approach in the third and final step. This time, we aim to develop an automated method to gather disorder levels of an SNS such as Reddit over time. We then seek to monitor specific subcommunities with varying levels of perceived disorder over a prolonged period to track the perceived disorder and toxicity levels in an uninterrupted and controlled manner.

7 Challenges and Outlook

As our explorative study is not without limitations, it offers several additional important angles for future research. Firstly, next to significantly extending the underlying dataset, we believe that variations of different subreddit pairs in combination with varying facets of toxicity would yield promising results. In line with this, the various nuances of toxicity may be affected differently by their perceived disorder. A distinction will enhance our understanding of the phenomenon but is beyond this study's scope. Secondly, by comparing subreddits, additional information on individual participants will generate more insights. Within our analysis, we have neglected participant overlaps, but those 425 participants make up ten percent of all the active participants in the subreddit with lower toxicity observed. This allows further investigation as interdependency effects could occur (Hessel et al., 2016). Potentially, these users are more toxic in one subreddit as they adapt to the different climates to fit in. A further survey for this specific subgroup could shed light on their behaviors. In addition, moderation could serve as an inhibitor of toxicity. However, in the study, the subcommunity with higher toxicity had specified more usage rules, especially related to the structure and content of participants' posts. This observation is potentially explained by rules not actively enforced, but further research should be conducted to explore moderation's effect on the link between perceived disorder and toxicity. This research should seek to distinguish between the existence of moderation rules and the intensity of actual moderation. Similarly, we have not considered the effect of low interaction with posts and how this may affect toxic behaviors. Finally, while we endeavored to identify two subcommunities as similar as possible, finding a match where the underlying characteristics (e.g., creation date, community size) were identical was impossible. Instead, it was only possible to select a subcommunity pair that resembled as similar characteristics as possible to match the main factor we wanted to rule out to be responsible for varying toxicity levels, i.e., the topic. Since Reddit has subcommunities ranging from 1 to more than 50 million members, we assume the two chosen subreddits, r/vent, and r/venting, are similar enough to be comparable. However, future research should consider this aspect and match subcommunities in a more fine-grained way.

Overall, the study offers potential for extensions in various directions and initially explores the phenomenon. It signifies the first step of a long-term project to measure the link between perceived disorder and toxicity over time. More specifically, when toxicity and perceived disorder can be gathered over time, the connections between the level of disorder and toxicity may be investigated to derive countermeasures for ameliorating interaction between users on SNSs.

References

Almerekhi, H., Kwak, H., Jansen, B. J., and Salminen, J. (2019). "Detecting Toxicity Triggers in Online Discussions," in: Proceedings of the 30th ACM Conference on Hypertext and Social Media (HT'19), 291-292, DOI: 10.1145/3342220.3344933.

Almerekhi, H., Jansen, B.J., and Kwak, H. (2020). "Investigating Toxicity Across Multiple Reddit Communities, Users, and Moderators," in: Companion Proceedings of the Web Conference 2020 (WWW'20), 294–298, DOI: 10.1145/3366424.3382091.

Atari, M., Davani, A.M., Kogon, D., Kennedy, B., Ani Saxena, N., Anderson, I., and Dehghani, M. (2022). "Morally Homogeneous Networks and Radicalism," *Social Psychological and Personality Science* 13 (6), 999-1009, DOI: 10.1177/19485506211059329.

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., and Blackburn, J. (2020). "The Pushshift Reddit Dataset," in: Proceedings of the International AAAI Conference on Web and Social Media (ICWSM 2020), 830-839, DOI: 10.1609/icwsm.v14i1.7347.

Barlett, C.P., Gentile, D.A., and Chew, C. (2016). "Predicting Cyberbullying from Anonymity," *Psychology of Popular Media Culture* 5 (2), 171-180. DOI: 10.1037/ppm0000055.

Chipidza, W. (2021). "The Effect of Toxicity on COVID-19 News Network Formation in Political Subcommunities on Reddit: An Affiliation Network Approach," *International Journal of Information Management* 61, 1-14, DOI: 10.1016/j.ijinfomgt.2021.102397.

Chan, T. K., Cheung, C. M., and Lee, Z. W. (2021). "Cyberbullying on Social Networking Sites: A Literature Review and Future Research Directions," *Information & Management* 58 (2), 103411, DOI: 10.1016/j.im.2020.103411.

Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., and Vakali, A. (2017). "Mean Birds: Detecting Aggression and Bullying on Twitter," in: Proceedings of the 2017 ACM on Web Science Conference (WebSci'17), 13-22, DOI: 10.1145/3091478.3091487.

Chen, Y., Zhou, Y., Zhu, S., and Xu, H. (2012). "Detecting Offensive Language in Social Media to Protect Adolescent Online Safety," in: 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing, DOI: 10.1109/SocialCom-PASSAT.2012.55.

Chow, A. R. (2022). *Reddit Allows Hate Speech to Flourish in Its Global Forums, Moderators Say*. URL: https://time.com/6121915/reddit-international-hate-speech/ (visited on November 15, 2022).

Cialdini, R. B., Reno, R. R., and Kallgren, C. A. (1990). "A Focus Theory of Normative Conduct: Recycling the Concept of Norms to Reduce Littering in Public Places," *Journal of Personality and Social Psychology* 58 (6), 1015–1026, DOI: 10.1037/0022-3514.58.6.1015.

Cohen, D., Spear, S., Scribner, R., Kissinger, P., Mason, K., and Wildgen, J. (2000). " 'Broken Windows' and the Risk of Gonorrhea," *American Journal of Public Health* 90 (2), 230-236, DOI: 10.2105/ajph.90.2.230.

Costa, J. J. (1984). *Abuse of the Elderly*, Boston, MA: Lexington Books.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). "Automated Hate Speech Detection and the Problem of Offensive Language," in: Proceedings of the International AAAI Conference on Web and Social Media (ICWSM 2017), 512-515, DOI: 10.1609/icwsm.v11i1.14955.

Dinakar, K., Reichart, R., and Lieberman, H. (2011). "Modeling the Detection of Textual Cyberbullying," in: Proceedings of the International AAAI Conference on Web and Social Media (ICWSM 2011), 5 (3), 11-17, DOI: 10.1609/icwsm.v5i3.14209.

Dixon, L., Li, J., Sorensen, J., Thain, N., and Vasserman, L. (2018). "Measuring and Mitigating Unintended Bias in Text Classification," in: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES'18), 67-73, DOI: 10.1145/3278721.3278729.

Gil, Y., Chai, Y., Gorodissky, O., and Berant, J. (2019). *White-to-black: Efficient Distillation of Black-Box Adversarial Attacks*, arXiv, DOI: 10.48550/arXiv.1904.02405.

Google Jigsaw (2022). *Perspective API.* URL: https://perspectiveapi.com/ (visited on November 15, 2022).

Grimes, M., and Marquardson, J. (2019). "Quality Matters: Evoking Subjective Norms and Coping

Appraisals by System Design to Increase Security Intentions," *Decision Support Systems* 119, 23-34, DOI: 10.1016/j.dss.2019.02.010.

Grimes, G. M., Marquardson, J., and Nunamaker, J. F. (2014). "Broken Windows, Bad Passwords: Influencing Secure User Behavior via Website Design," in: Proceedings of the Americas Conference on Information Systems (AMCIS2014), Savannah, Georgia, USA.

Gruzd, A., Mai, P., and Vahedi, Z. (2020). *Studying Antisocial Behaviour on Reddit with Communalytic*, Preprint. DOI: 10.31124/advance.12453749.v1.

Harcourt, B. E. (2006). "Reflecting on the Subject: A Critique of the Social Influence Conception of Deterrence, the Broken Windows Theory, and Order-Maintenance Policing New York Style," *Michigan Law Review* 97 (2), 291-389.

Harcourt, B. E., and Ludwig, J. (2006). "Broken Windows: New Evidence from New York City and a Five-city Social Experiment," *University of Chicago Law Review* 73 (1), 271-320.

Hessel, J., Tan, C., and Lee, L. (2016). "Science, Askscience, and Badscience: On the Coexistence of Highly Related Communities," in: Proceedings of the International AAAI Conference on Web and Social Media (ICWSM 2016), 171-180, DOI. 10.1609/icwsm.v10i1.14739.

Horta Ribeiro, M., Jhaver, S., Zannettou, S., Blackburn, J., Stringhini, G., De Cristofaro, E., and West, R. (2021). "Do Platform Migrations Compromise Content Moderation? Evidence from r/the_donald and r/incels," in: Proceedings of the ACM on Human-Computer Interaction, 5 (CSCW2), 1-24, DOI: 10.1145/3476057.

Hosseini, H., Kannan, S., Zhang, B., and Poovendran, R. (2017). *Deceiving Google's Perspective API Built for Detecting Toxic Comments*, arXiv, DOI: 10.48550/arXiv.1702.08138.

Herring, S., Job-Sluder, K., Scheckler, R., and Barab, S. (2002). "Searching for Safety Online: Managing 'Trolling' in a Feminist Forum," *The Information Society* 18 (5), 371–384, DOI: 10.1080/01972240290108186.

Hua, Y., Ristenpart, T., and Naaman, M. (2020). "Towards Measuring Adversarial Twitter Interactions Against Candidates in the US Midterm Elections," in: Proceedings of the International AAAI Conference on Web and Social Media (ICWSM 2020), 272-282, DOI: 10.1609/icwsm.v14i1.7298.

Hutto, C., and Gilbert, E. (2014). "Vader: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text," in: Proceedings of the International AAAI Conference on Web and Social Media (ICWSM 2014), 8 (1), 216-225, DOI: 10.1609/icwsm.v8i1.14550.

Keizer, K., Lindenberg, S., and Steg, L. (2008). "The Spreading of Disorder," *Science* 322 (5908), 1681–1685, DOI: 10.1126/science.1161405.

Kitchens, B., Johnson, S. L., and Gray, P. (2020). "*Understanding Echo Chambers and Filter Bubbles: The Impact of Social Media on Diversification and Partisan Shifts in News Consumption*," MIS Quarterly 44 (4), 1619-1649, DOI: 10.25300/MISQ/2020/16371.

Liu, D., Wright, K. B., and Hu, B. (2018). "A Meta-Analysis of Social Network Site Use and Social Support," *Computers & Education* 127, 201–213, DOI: 10.1016/j.compedu.2018.08.024.

Lowry, P. B., Moody, G. D., and Chatterjee, S. (2017). "Using the Control Balance Theory to Explain Social Media Deviance," in: Proceedings of the Hawaii International Conference on System Sciences (HICSS-50), Big Island, Hawaii.

Mackie, D. M., Devos, T., and Smith, E. R. (2000). "*Intergroup Emotions: Explaining Offensive Action Tendencies in an Intergroup Context*," Journal of Personality and Social Psychology, 79 (4), 602–616, DOI: 10.1037/0022-3514.79.4.602

Massanari, A. (2017). "#Gamergate and The Fappening: How Reddit's Algorithm, Governance, and Culture Support Toxic Technocultures," *New Media & Society* 19 (3), 329-346, DOI: 10.1177/1461444815608807.

Minson, J. A., and Dorison, C. A. (2022). "Why is Exposure to Opposing Views Aversive? Reconciling Three Theoretical Perspectives," *Current Opinion in Psychology* 101435, DOI: 10.1016/j.copsyc.2022.101435.

Mittos, A., Zannettou, S., Blackburn, J., and De Cristofaro, E. (2020). "'And We Will Fight for Our Race!' A Measurement Study of Genetic Testing Conversations on Reddit and 4chan," in: Proceedings of the International AAAI Conference on Web and Social Media (ICWSM 2020), 14 (1), 452-463, DOI: 10.1609/icwsm.v14i1.7314.

O'Donnell, E., and O'Donnell, L. (2022). *The Dark Side of Engaging With Social Networking Sites (SNS)*, in: Khosrow-Pour, M. (ed.), Encyclopedia of Criminal Activities and the Deep Web 2, 615-627, Hershey, Pennsylvania: Information Resources Management Association, IGI Global, USA.

Reno, R. R., Cialdini, R. B., and Kallgren, C. A. (1993). “The Transsituational Influence of Social Norms,” *Journal of Personality and Social Psychology* 64 (1), 104-112, DOI: 10.1037/0022-3514.64.1.104.

Sampson, R. J., Raudenbush, S. W., Sampson, R. J., and Raudenbush, S. W. (2015). “Neighborhood Seeing Disorder : Stigma and the Social Construction of 'Broken Windows'," *Social Psychology Quarterly* 67 (4), 319–342, DOI: 10.1177/019027250406700401.

Statista (2017). *On Which of the Following Topics Have You Seen Trolling Behavior on the Internet?* URL: https://www.statista.com/statistics/380051/topics-witness-trolling-behavior-internet/ (visited on November 15, 2022).

Stephan, A. (2020). *Comparing Platform Hate Speech Policies: Reddit's Inevitable Evolution*. URL: https://fsi.stanford.edu/news/reddit-hate-speech (visited on November 11, 2022).

Tahmasbi, F., Schild, L., Ling, C., Blackburn, J., Stringhini, G., Zhang, Y., and Zannettou, S. (2021). “'Go eat a bat, Chang!': On the Emergence of Sinophobic Behavior on Web Communities in the Face of COVID-19,” in: Proceedings of the Web Conference 2021 (WWW'21), 1122-1133, DOI: 10.1145/3442381.3450024.

Vogels, E. A. (2021). *The State of Online Harassment. Pew Research Center*. URL: https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/ (visited on November 11, 2022).

Welsh, B. C., Braga, A. A., and Bruinsma, G. J. N. (2015). “Reimagining Broken Windows: From Theory to Policy,” *Journal of Research in Crime and Delinquency* 52 (4), 447–463, DOI: 10.1177/0022427815581399.

Wilson, J. Q., and Kelling, G. L. (1982). “Broken Windows,” *Atlantic Monthly* 249 (3), 29–38.

Xia, Y., Zhu, H., Lu, T., Zhang, P., and Gu, N. (2020). “Exploring Antecedents and Consequences of Toxicity in Online Discussions: A Case Study on Reddit,” in: Proceedings of the ACM on Human-Computer Interaction 4 (CSCW2), 1–23, DOI: 10.1145/3415179.

Zimbardo, P. G. (1973). “A Field Experiment in Auto Shaping,” *Vandalism*, 85–90.