How Social Media Inscribes Values into Its Users

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Khoi Pham

Spring 2023

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Advisor

Rider W. Foley, Department of Engineering and Society

The Problem With Internet Discussion

A popular internet adage is Godwin's Law, stating that as an online discussion goes longer, the probability of a comparison to Adolf Hitler being made grows (Miller, 2013). While some of the negative aspects of online discussion can be attributed to innate human nature, most of the responsibility falls on the internet as a medium (Lipinski-Harten & Tafarodi, 2013). The amount of negative language and insults thrown on online discussions have caused many users to leave social media websites, broken real-life relationships, and increased the amount of existing anger on the platform, leading to a negative feedback loop (Wang et. al, 2011). Negative online discourse is a problem and research has alluded that individual websites suffer at varying degrees, which can be seen in Figure 1, showing that certain design aspects of these platforms either exacerbate or reduce the problem (Baughan et al., 2021).

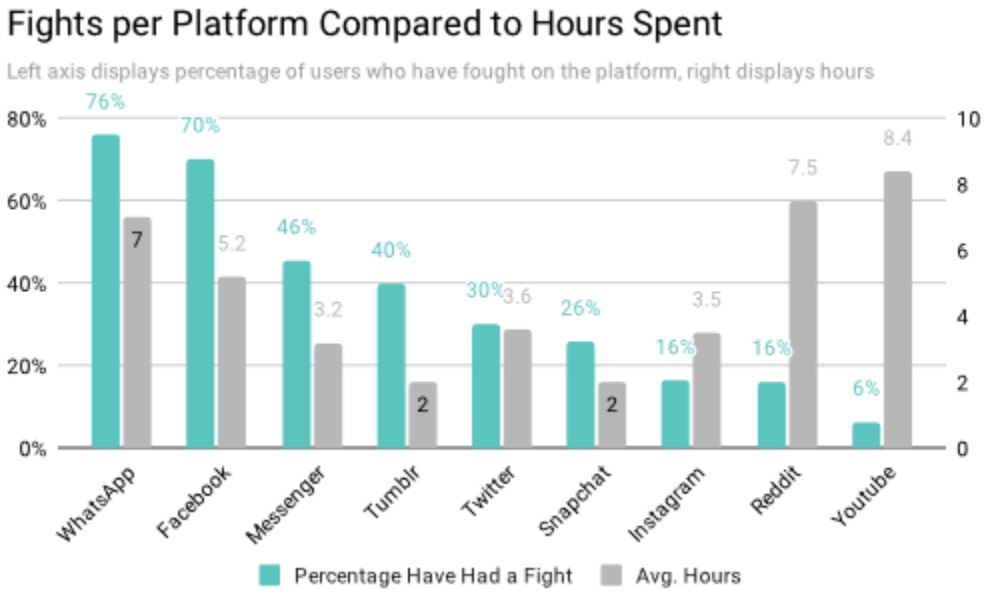


Figure 1. Graph showing the differences in the argument rate of platforms (Baughan et al., 2021)

Social media platforms are willing to put forward design decisions that encourage toxic behavior if it meant users would stay on their websites for longer. Around 2016, Facebook hired a great number of sociologists and data scientists to figure out how they could design their website to increase the amount of time a user would spend on their website (Merrill & Oremus, 2021). Consequently, Facebook's feed algorithm began showing posts with the most angry face reactions more often since users interacted more with those posts. Internal documents within the company show that employees knew the moral qualms of the algorithm and when they brought those concerns to their superiors, their concerns were dismissed. Figure 2 shows Mark Zuckerberg, the CEO of Facebook, encouraging people to angry face react. It was only after the public took notice and pressured Facebook that the algorithm changed. A similar backlash occurred after the January 6th riots, causing Facebook to deweigh live videos, which were frequently used during the event, in their algorithm.

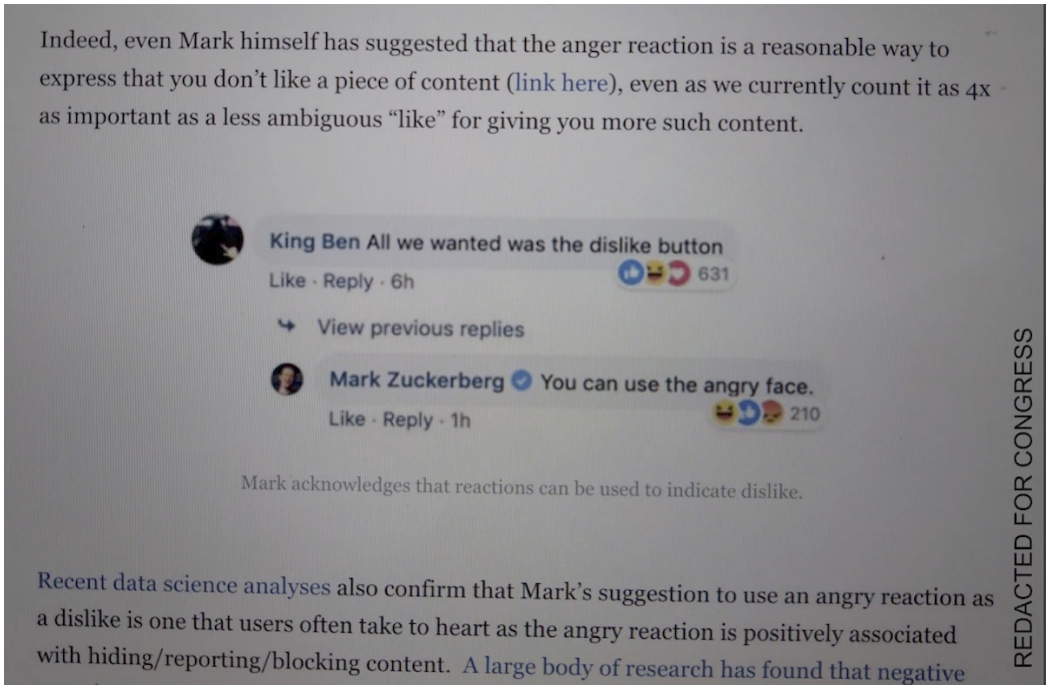


Figure 2: Zuckerberg telling a Facebook user to angry face react (Merrill & Oremus, 2021)

In addition to Facebook, YouTube's recommendation algorithm had intentionally been showing users incendiary and controversial content to gain more engagement (Munn, 2020). Again, executives ignored warnings and it was only after complaints of spreading vaccine misinformation that YouTube began flagging videos as "borderline content", videos that won't be recommended to users (Bergen, 2019). All platforms will try to thrive off inflammatory content whenever possible.

Times Newer Roman is a parody social media website created to determine the causes of an inflammatory social media environment. By mimicking design aspects from websites such as Facebook, Twitter, and Instagram, Times Newer Roman can show how design features can facilitate the negative feedback loop common in social media platforms. Using the framework of Actor Network Theory, this paper will examine how algorithms, voting systems, and other features inscribe negative values into users, how users prescribe their own values into platforms, and how social media thrives off and facilitates this cycle.

The Components of a Crowdsourced Font

I made an online social media parody called Times Newer Roman that can be viewed at crowseeds.com/font. The goal of the platform is to create the best font imaginable through crowdsourcing. Though the project was originally built for my entertainment, it has been repurposed for educational and research use. The project will hopefully help contribute to research in social media design and interactive media by acting as a case study.

Users get assigned a letter to draw and submit to a database, see Figure 3. Users can also like and dislike other artists' drawings, as shown in Figure 4, with the highest rated drawings becoming a part of the font.

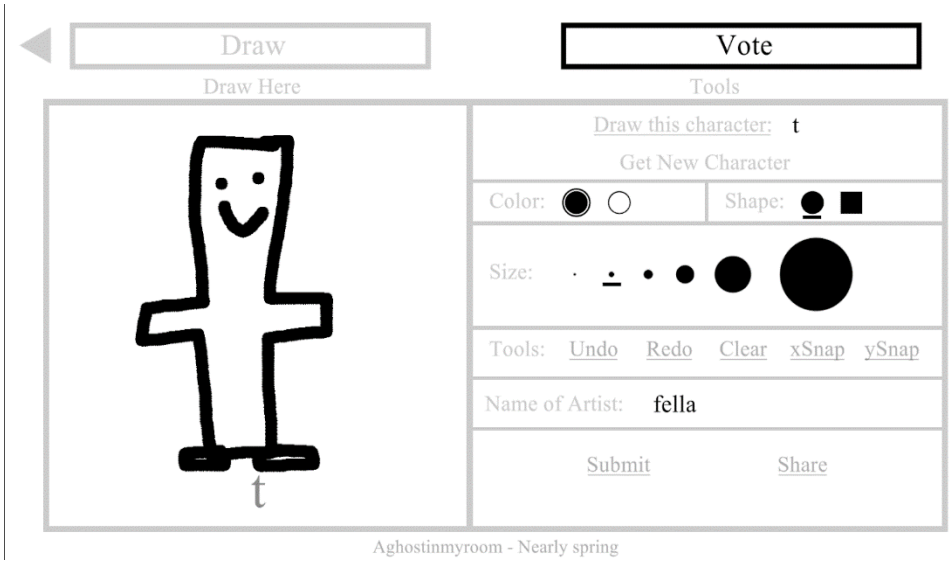


Figure 3: Drawing Screen of Times Newer Roman (Pham, 2022)

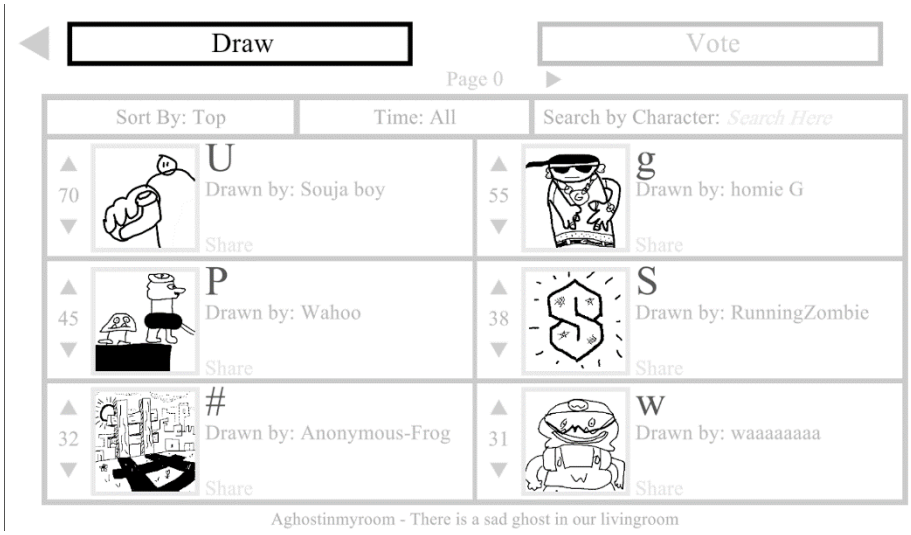


Figure 4: Voting Screen of Times Newer Roman (Pham, 2022)

Design aspects of the platform are made to mirror features found in other social media platforms. Artists can name themselves anything and keep themselves anonymous. The voting feed displays the highest rated posts. Users can like and dislike each other's posts. The results of

Times Newer Roman will be analyzed in this paper to determine what features cause inflammatory drawings.

Using Actor Network Theory To Analyze Intentional Social Media Design

Actor Network Theory, or ANT, will be used to analyze the relationships between social media companies, users, and design choices to explain how they all relate to the problem of companies engineering their platforms to encourage toxic discourse. The framework breaks down the problem into a series of actors, which can be human or non-human, and describes how they interact. Actors are independent and are not "the subject of the other, more important actors' whims" (Latour, 2005). This is important as it separates the non-human elements of the system, such as the feed sorting algorithms on social media or a game an activist creates, from their creators, making them contributors to the system with as much as a role their human counterparts.

Platforms have also delegated the responsibility of content moderation and prioritization to automated algorithms. Companies have inscribed the value of keeping users on their website as long as possible, which has been responsible for inadvertently popularizing controversial content due to their high user retention rates (Merrill & Oremus, 2021). Ultimately, the prescription, the behaviors reflected back onto human actors, of social media algorithms causes users of the site to engage in bad behaviors, with the feed algorithms of Facebook playing a part of the January 6th Capitol Riots and vaccine hesitancy during the COVID-19 pandemic. As misinformation received lots of reactions and comments on Facebook, the algorithm favored these posts, causing many to believe claims of a stolen election or fake information about the vaccine (Merrill & Oremus, 2021). The behaviors prescribed as a result of the inscriptions on

social media platforms are the crux of this problem and ANT allows us to break down the back-and-forth relationship between non-human and human actors.

Platforms thrive off the negative feedback loop of inscription and prescription, with algorithms showing users inflammatory content and users staying on longer to post inflammatory content. What are the actors, non-human or human, that help facilitate this cycle? This paper will use ANT to figure out what features cause this cycle and if there is any way to break it.

What Facilitates a Comparison to Hitler?

While it is known that different social media platforms have varying levels of toxicity, the causes of these differences are still being researched (Baughan et al., 2021). Features such as the recommendation algorithm and anonymity have often been cited as candidates for facilitating bad online discourse (Munn, 2020) (Christopherson, 2007). Though correlations have been noticed (Baughan et al., 2021), the field of analyzing social media is still developing and no consensus has been made about why certain platforms are worse than others. This research paper will attempt to answer the question: Why are certain social media platforms worse than others and what can be done to fix them?

Dissecting a Social Media Website

In an attempt to answer this question, I have conducted two different experiments that tie into each other.

"Times Newer Roman" was created as a parody social media platform to help answer this question. The website can be viewed at https://crowseeds.com/font and serves as the first part of

the study. Users would draw letters of the alphabet to help create a crowdsourced font. Drawings are meant to act as user-made content. Users can then vote on drawings made by other people. The website employs a set of features commonly seen in modern platforms. This includes the ability to be anonymous, liking and disliking submissions, and seeing the highest rated posts first. A screenshot of the website can be seen in Figure 5.

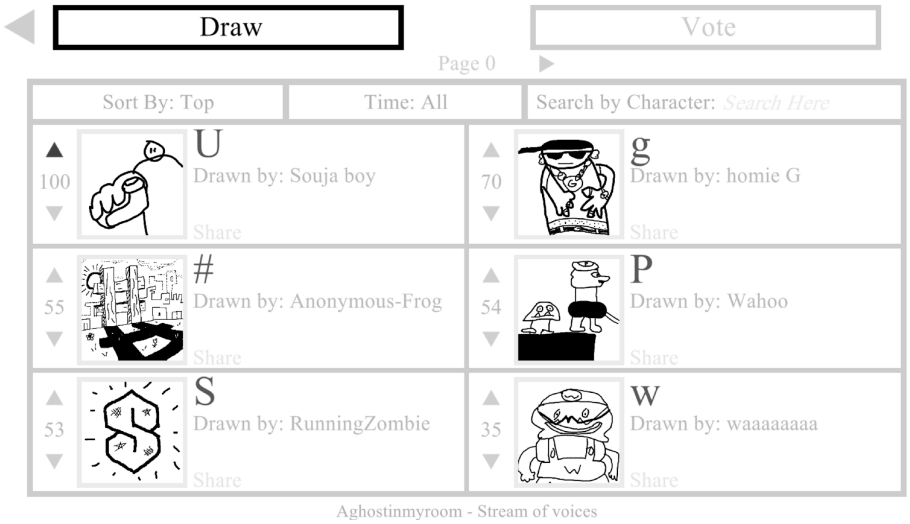
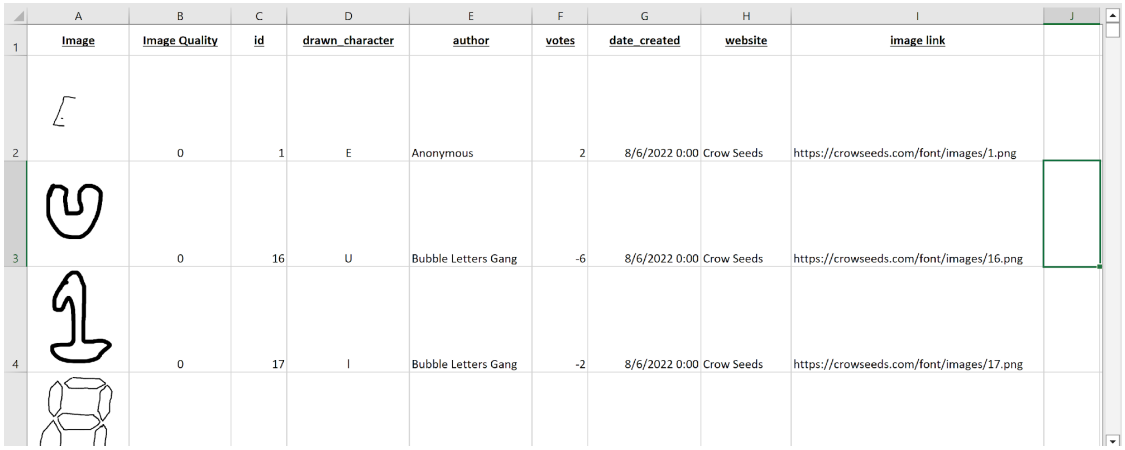


Figure 5. Screenshot of the Times Newer Roman feed (Pham, 2023)

After seven months of letting users interact with the platform, the drawings and votes were then analyzed. If the features put in place were effective at limiting toxic discourse, then drawings with inappropriate content such as hate speech and phallic imagery should be voted lower than images of a higher quality. A randomized sample of user-made images was added to a spreadsheet and are rated as either high or low quality. By correlating the number of votes an image receives and the image quality, it can be determined if the features on the site increase or decrease the visibility of toxic content. A section of the spreadsheet can be seen in Figure 6.



	A	B	C	D	E	F	G	H	I	J
1	Image	Image Quality	id	drawn_character	author	votes	date_created	website	image link	
2		0	1	E	Anonymous	2	8/6/2022 0:00	Crow Seeds	https://crowseeds.com/font/images/1.png	
3		0	16	U	Bubble Letters Gang	-6	8/6/2022 0:00	Crow Seeds	https://crowseeds.com/font/images/16.png	
4		0	17	I	Bubble Letters Gang	-2	8/6/2022 0:00	Crow Seeds	https://crowseeds.com/font/images/17.png	

Figure 6. Section of Times Newer Roman spreadsheet (Pham, 2023)

In addition to “Times Newer Roman”, I created a survey via Google Forms to ask social media users about their least and most favorite platforms. In particular, the survey is asking the participant about what features of the sites they dislike and like the most and how they would improve the platform. The first part of the survey focuses on their least favorite platform and the second part focuses on their favorite platform. The survey asks the following:

- What is your least/most favorite social media platform?
- How much time do you spend daily on the platform?
- What is your least/most favorite feature of the platform?
- (Optional) Rant/gush about the platform.
- (Optional) What would you do to improve the platform?

The survey can be viewed at https://forms.gle/uGMMwS9ywdrJdiNjB. The survey will be posted on the Times Newer Roman platform and my personal social media accounts. The

hope is that by looking at the features commonly mentioned in the survey, it can be shown what aspects of social media cause or limit toxicity. A portion of the survey can be seen in Figure 6.

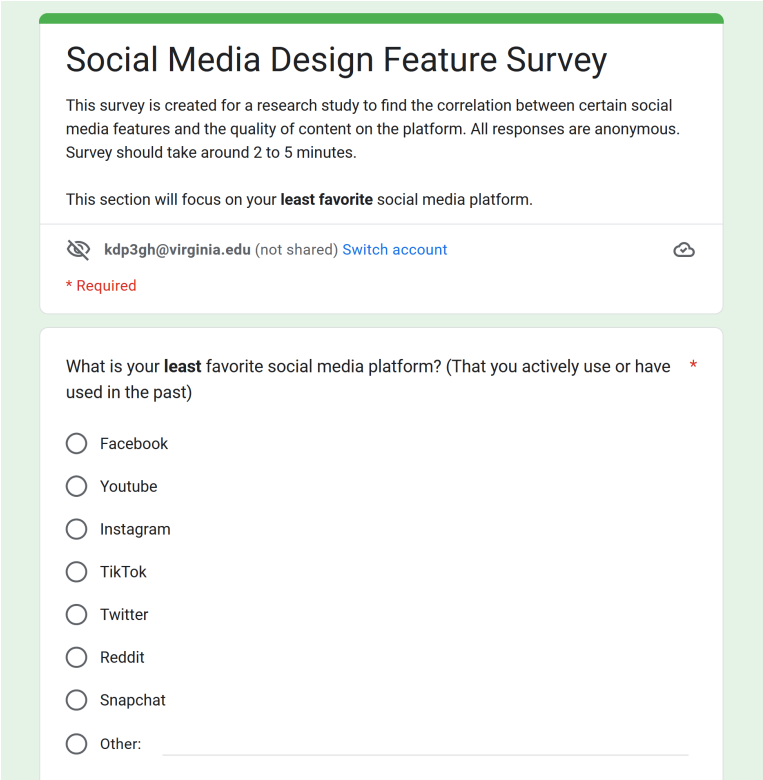


Figure 6. Portion of the Google Forms survey (Pham, 2023)

Results of Times Newer Roman

There were two key social media features that shaped the content of Times Newer Roman. Anonymity correlated strongly with lower quality works, while a like-dislike voting system encouraged higher quality works. Actor Network Theory shows how these features can give incentives to draw images of certain qualities. The voting system prescribed a value of quality and creativity in users, assigning each drawing a score, giving users the motivation to put more effort. Just as how Latour described the responsibility of closing a door was given to an

automatic groom, the responsibility of critiquing art is given to a voting system. The voting system serves as an actor that influences the way users draw letters in Times Newer Roman. Anonymity prescribes the value of a lack of responsibility, acting as a shield that removes the incentive to put effort. These findings were found in a long study analyzing the drawings people have made on the website.

Over seven months from August 2022 to March 2023, there have been 687 drawings made from 288 unique authors. A random sample of 140 drawings was added to a spreadsheet and graded manually by their subjective quality individually by the author of this paper. Submissions of a lower quality were given a score of -1. These are drawings that are off-topic or contain inappropriate content such as phallic imagery and hate speech. Drawings with substantial effort were given a score of 1. The rest were given a score of 0. Drawings of each quality score can be seen in Figure 7.

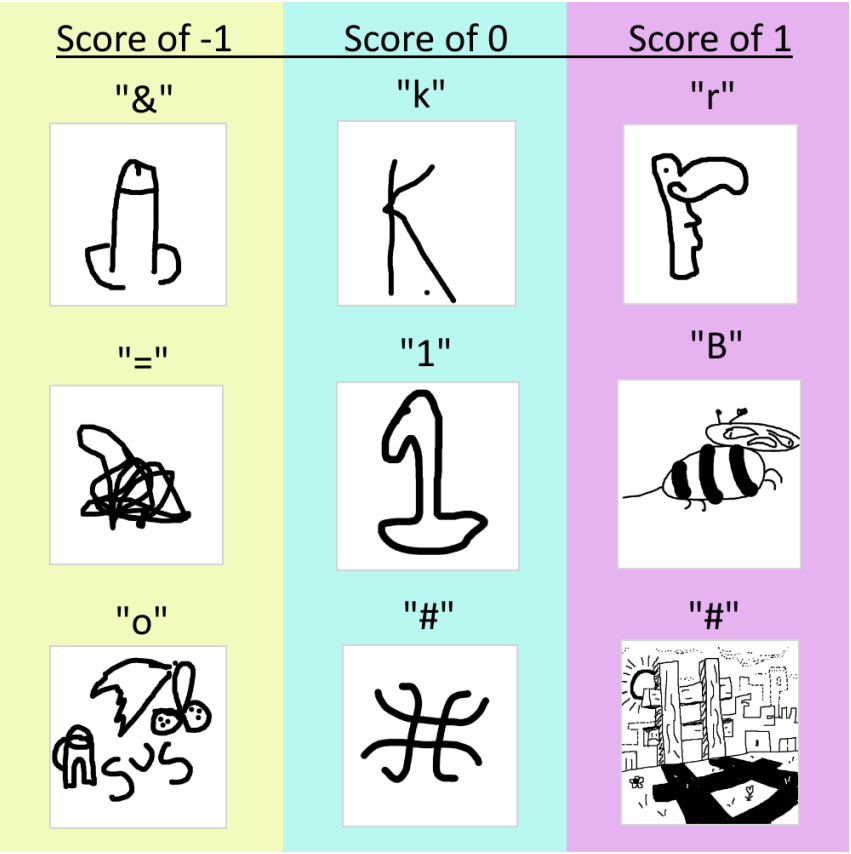


Figure 7. Quality grade examples (Pham, 2023)

A dot plot showing the distribution of likes can be seen in Figure 8. As seen in Figure 8, there is a positive correlation between an image's quality with the number of likes received. The graph shows that images with a quality score of -1 tended to have a like amount in the negatives due to the dislike feature on the website. Images with a quality score of 1 on average receive a much higher amount of likes than images with a quality score of 0 or -1. Vote averages can be seen in Figure 9.

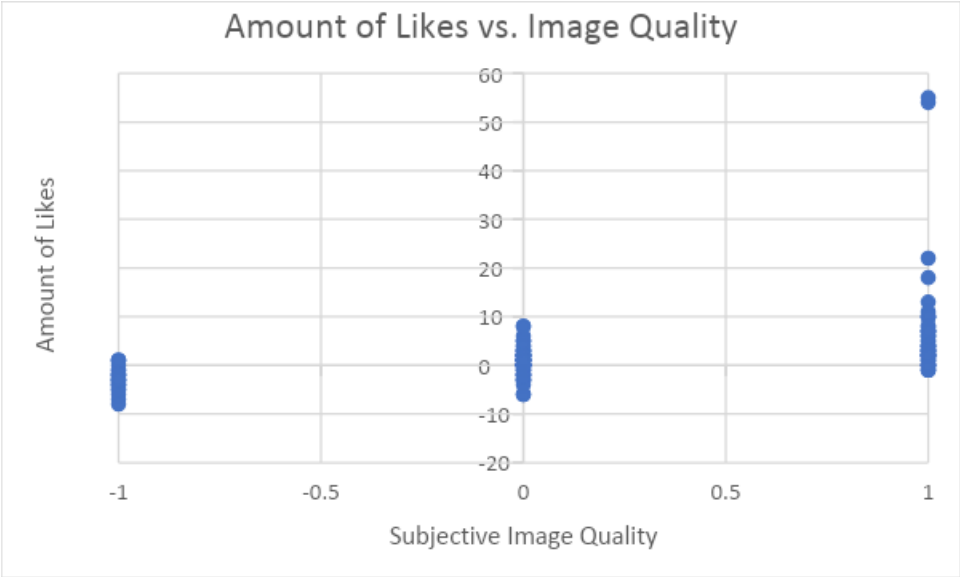


Figure 8. Amount of likes vs. image quality graph (Pham, 2023)

Image Quality	Average Votes	Std Deviation	Count
-1	-2.818181818	2.481097369	22
0	0.796875	2.743995971	64
1	6.444444444	10.55384517	54

Figure 9. Amount of likes vs. image quality table (Pham, 2023)

As Times Newer Roman shows posts with the highest likes first, it successfully hides images with a lower quality. By having users like and dislike posts, content with more effort retains a high visibility. Since drawings of higher quality are the first thing a new user sees, this can encourage them to create posts with more effort. This can explain how there is a greater

number of posts with an image quality of 0 or 1 than posts with an image quality of -1, as seen in Figure 9.

Does anonymity have an effect on the image quality? As seen in Figure 10, anonymous posts had a higher likelihood of being low quality. 27.87% of anonymous posts had a quality of -1 compared to 6.33% of non-anonymous posts. Posts with a named author also had a greater chance of being higher quality. 50.63% of named posts had an image quality of 1, more than double that of the 22.95% of anonymous posts. There are some possible explanations. Users that post low-quality content could not care enough to put a name or be using anonymity as a shield to protect themselves from backlash. Users that create higher quality content may be proud of their work and are more inclined to attach a name to their work. While a cause isn't confirmed, there is a strong correlation between anonymity and the quality of a post.

Image Quality	Number of Anonymous Posts	Number of Non-Anonymous Posts	Proportion of Anonymous Posts	Proportion of Non-Anonymous Posts
-1	17	5	27.87%	6.33%
0	30	34	49.18%	43.04%
1	14	40	22.95%	50.63%
All	61	79	100.00%	100.00%

Figure 10. Anonymity vs. image quality table (Pham, 2023)

Times Newer Roman had successfully fostered a creative community. A majority of the submissions were of high quality which can be ascribed to a positive feedback loop. As new artists enter the website, they are first prescribed values of creativity and humor by the platform. The first thing they see are the highest rated posts of Times Newer Roman. With those values in mind, they use the platform as they see fit, inscribing the website with their drawings and votes.

Actor Network Theory shows this feedback loop using the relationships of inscription and prescription. The voting system and feed sorting algorithm help keep this loop running by incentivizing higher quality posts. This is the inverse of many social media platforms such as YouTube that create a negative feedback loop, recommending videos a user would get mad at to facilitate a response (Chaslot, 2019).

Results of the Survey

With Actor Network Theory as the framework used, specific features must be singled out as actors that prescribe values of hate and uncivil discussion. The responses of the survey are combed for these features.

In two weeks the survey garnered 43 responses. As seen in Figure 11, opinions on the worst social media were split. All seven possible responses were chosen by at least 3 people. Facebook and Twitter had the highest proportion, with 23.3% and 25.6% respectively. This shows that the problems people have with social media aren't constrained to a specific feature of a platform, but rather features present in all social media. In contrast, the most loved social media was YouTube by a significant margin, with 44.2% of the vote, as seen in Figure 12.

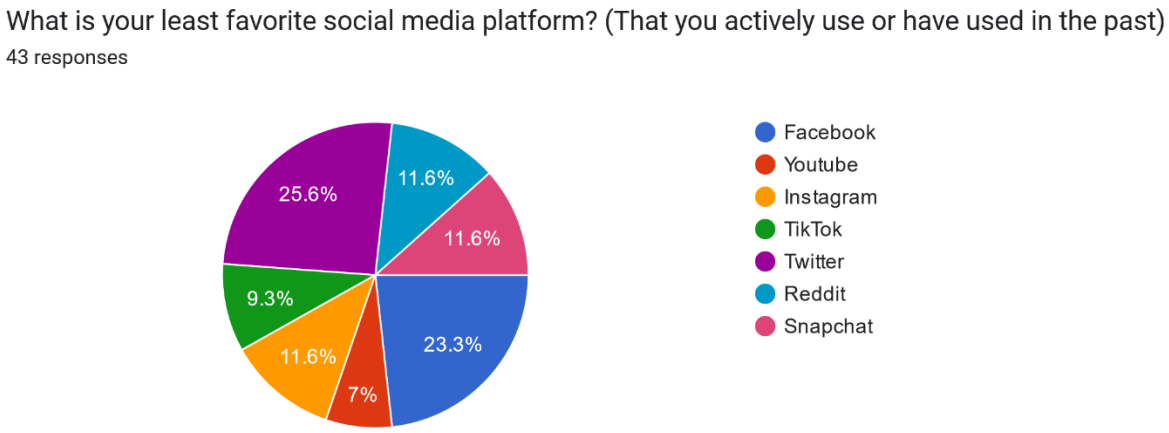


Figure 11. Least favorite social media pie chart (Pham, 2023)

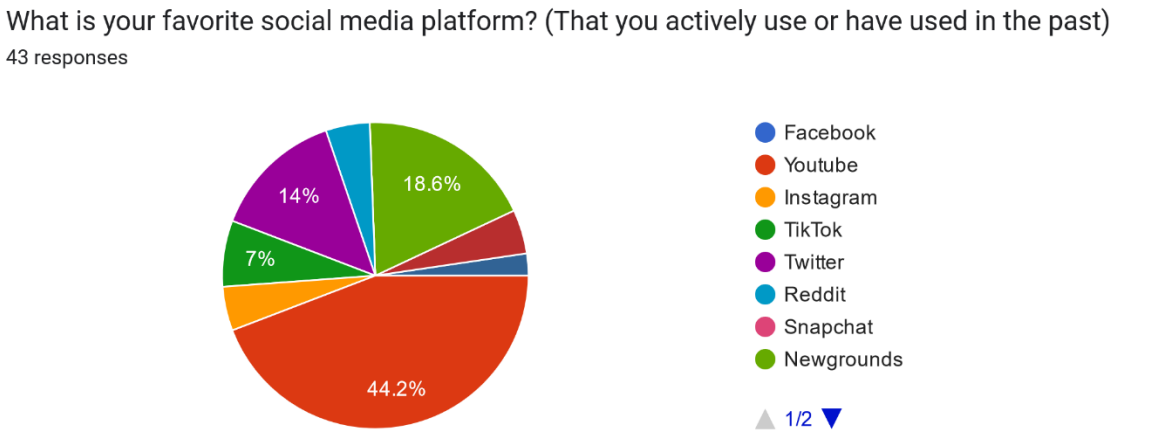


Figure 12. Most favorite social media pie chart (Pham, 2023)

There is an incentive to create good social media. As seen in Figure 13, an overwhelming number of users went on their least favorite social media for less than 30 minutes a day. Though it might seem that this conclusion is obvious, many studies in the field of social media point

towards the opposite effect. It has been claimed that anger causes a user to use a platform for longer (Merrill & Oremus, 2021). Figure 14 confirms that people are significantly more likely to use their favorite social media.

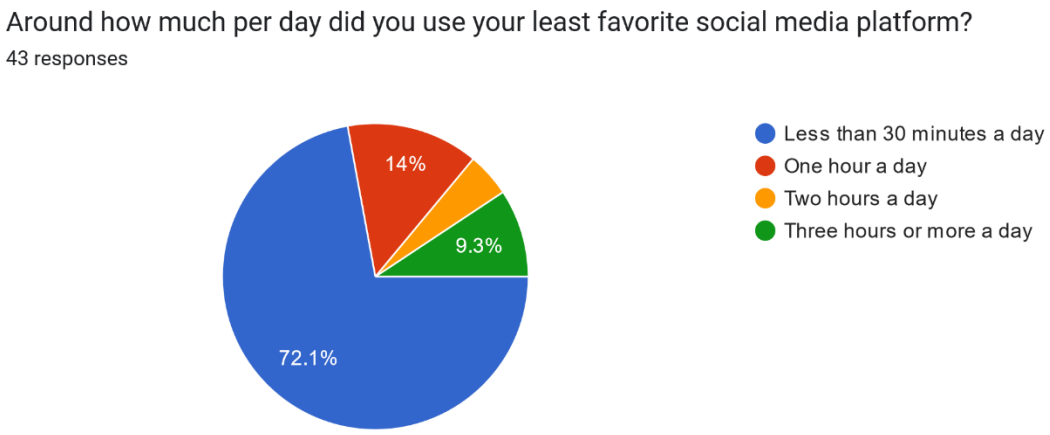


Figure 13. Least favorite social media usage pie chart (Pham, 2023)

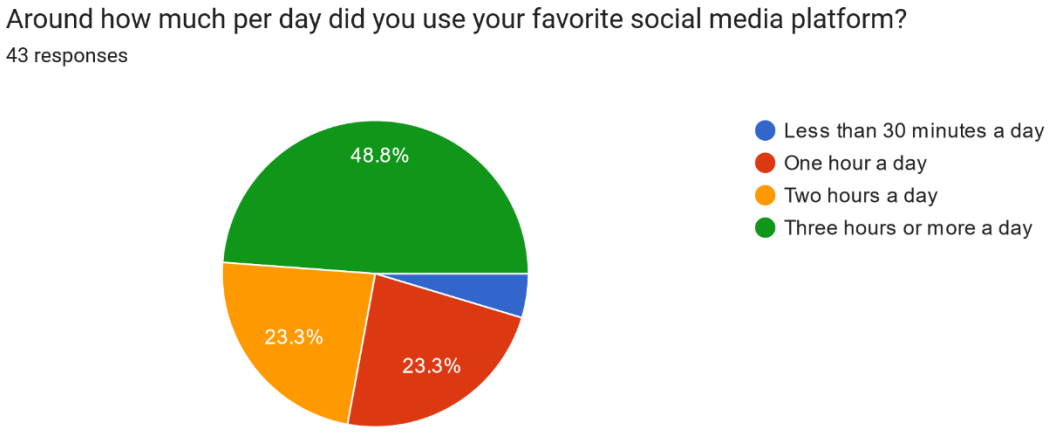


Figure 14. Most favorite social media usage pie chart (Pham, 2023)

Recommendation algorithms and conversations were the two features most disliked about social media, as seen in Figure 15. This helps confirm existing theories in the field that social media intentionally displays inflammatory content (Bergen, 2019). Twitter and Facebook, the most disliked websites according to Figure 11, have been shown to intentionally spread content that would rile up the user (Merrill & Oremus, 2021). Toxic discourse also plays a huge role in a user's distaste of a website, being the most common feature listed with 62.8%.

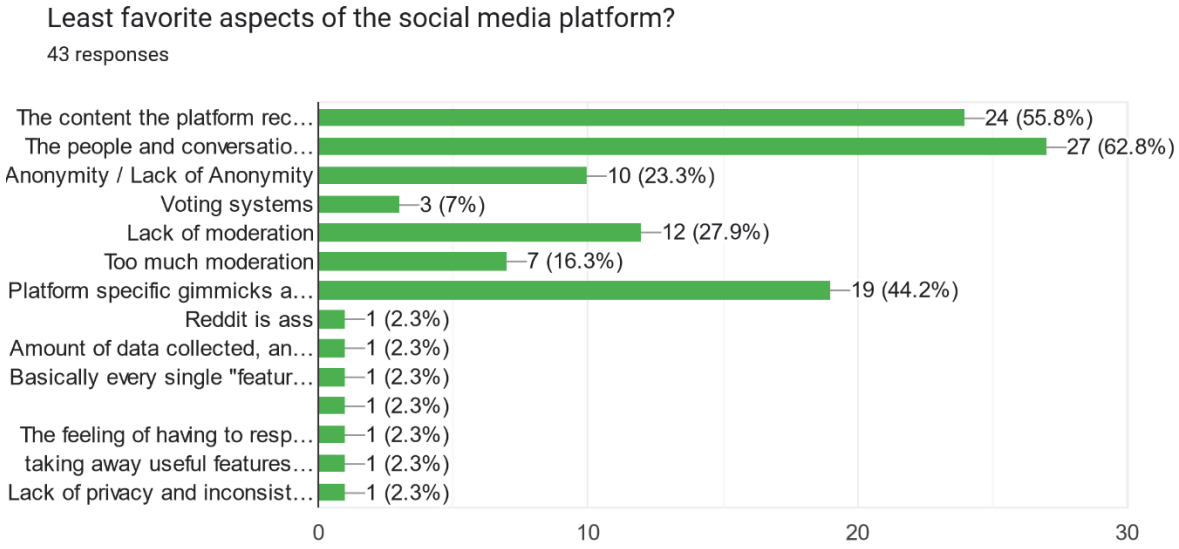


Figure 15. Least favorite social media features of survey participants (Pham, 2023)

When asked to rant about their least favorite social media, respondents brought up the intentional spreading of misinformation and toxic speech. Survey participants claimed that "Twitter breeds negativity in a way like no other. Twitter is like if Youtube comments got an

app." and "the content on [Facebook] seems designed to spread hate or misinformation, or to just be stolen content." However, many complained about the opposite effect. One participant wrote this about Reddit: "If you want to know what staring into the abyss feels like, just browse Reddit for about 5-10 minutes. The upvote system ensures that people are only exposed to ideas that are already popular, thus suppressing dissenting ideas and meaningful discussion." Another respondent added: "I'm aware almost every community will fall into echo chambers, but subreddits take it to a whole new level."

Recommendation algorithms need to keep a delicate balance. If recommended content is tailored to never make anyone angry and always be palatable to the user, it could lead to echo chambers and suppress discussion. On the other hand, encouraging arguments and giving the user a diverse array of content could lead to spreading hate and fuel toxic discourse. A participant said this about Newgrounds when asked to gush about their favorite social media: "You'll notice I selected both lack of moderation and too much moderation as things I like; Newgrounds has moderation down to a tee I'd say. The way they allow self-moderation is so flexible and liberating for those who do not wish to see certain content."

Anonymity and voting systems had little impact on a user's disdain for a platform, making up a small portion of respondents' rants. One participant stated that "people behave like twitter is an anonymous platform despite the fact it isn't." As seen in Figure 15, issues about anonymity comprised only 23.3% of responses.

Many respondents cited YouTube as their favorite social media, being the most well liked in the survey. When talking about their positive experiences with the site, many avoid talking about social interactions and discussions and instead focused on YouTube as an entertainment medium. One respondent stated: "I have a love-hate relationship with Youtube as most people do

but there is just so much great content and videos on there that I can just sit and watch for hours. I've learned a lot from video documentaries on shit like New Coke and where the Roblox oof sound comes from." Another added: "The only good things about Youtube are funny videos and videos created with a lot of effort, like animations, live-action, documentaries, video essays, and so on."

It can be concerning that the best social media lacks the "social" component and is more focused on being just "media". Almost every single gush response avoids the topic of interacting and talking to others. One respondent put it nicely when talking about Newgrounds: "It's not *necessarily* a social media platform, I understand. Moreso it's just a 'media platform' if anything but that's what I like about it. The 'social' part is very minimal with more of a focus on the content and that's pretty neat I think."

Users did not attribute the failings of social media to anonymity and focused more on recommendation algorithms. Feed algorithms are often cited as the main cause of social media toxicity in research papers (Munn, 2020). The survey results help confirm the notion.

Relating back to Actor Network Theory, recommendation algorithms prescribed values of toxicity by showing content that makes them angry. By acting as a tour guide for the internet, these algorithms reflect the most inflammatory content and hateful values users have inscribed into the internet. By prescribing these values into the users, it encourages them to exhibit toxic behaviors, which is why the survey responses also cited the users of social media being one of its worst aspects. This creates a negative feedback loop that echoes anger throughout the system.

The Implications of the Results

Social media is dependent on people. The drawings on Times Newer Roman only exist because people inscribe their creative values into the website. The complaints and praises of social media platforms in the survey were exclusively about people and their content, rather than non-human elements such as the user interface or monetization practices. Using Actor Network Theory, it can be determined that features of social media play as much as a role as users in terms of being actors in the system. Users will act on these features, inscribing values of hate, creativity, and all other human emotions onto the software. Features such as voting systems and recommendation algorithms amplify these values, which prescribes them back to more users. Actor Network Theory shows us this feedback loop. The results of the survey show the existence of negative feedback loops, those responsible for the uncivility of internet discussion. Times Newer Roman however, shows that positive feedback loops are possible on the internet and that creativity and effort can be incentivized with social media features.

While Times Newer Roman can show that the features put in place had successfully fostered a creative and civil environment, there are some limitations. Most notably, there isn't a control or any other versions of the website to compare the results to. There would be no way to know if the dislike feature improved the platform if there isn't a version of Times Newer Roman without the dislike feature to compare to. Also, the website only received 687 drawings, a small amount compared to the millions of comments and posts mainstream platforms receive. Drawings are also much different than text discussions. Still, the website helps confirm the existing notion that anonymity correlates with more toxic online environments (Davenport, 2002). It also supports the idea of a dislike button, a feature that still has no consensus in the field (Lutz & Schneider, 2021).

In the future, more versions and variations of social media need to be tested. Implementing different versions of Times Newer Roman would be helpful. Spreading the platform and the survey more to increase the number of participants would help with the accuracy of the results. Utilizing a different form of communication more similar to modern websites rather than a drawing website would yield more applicable results.

The survey results also lacked in sample size, with only 43 respondents. They were also answered by a specific demographic of people who play my games, which skews towards an age range of 18-35, who are not representative of all internet users. Most of the participants came from the website Newgrounds, which could be responsible for the praise the website has received in the survey.

The results found in the study are helpful if I decide to work for an internet communications company. As shown in the survey results, there is a high incentive to mitigate toxicity online. Knowing how certain social media features such as anonymity, feed algorithms, and voting systems affect users is highly important when implementing them. Hopefully, as a software engineer, I can improve the quality of online discussion and prevent companies from encouraging toxic discourse, which has been a problem in the past decade (Crockett, 2017).

Conclusion

Social media allows people to communicate with the world. It gives muted voices a microphone and facilitates a meshing of diverse thoughts. However, the design of these platforms either create echo chambers that restrict discussion or burn users out by feeding them inflammatory content. The results of the survey point out that no social media platform is

immune to these issues. Developers should be cognizant of the social media features addressed by Times Newer Roman and the survey. Hopefully, the silly nature of Times Newer Roman brings attention to bad social media design. By raising awareness about these issues, companies and developers can be pushed to alter their platforms.

It is incorrect to solely attribute toxic online environments to the values users inscribe and shift responsibility from the developers to the users (Lipinski-Harten & Tafarodi, 2013). While you can't change the behavior of people directly, the features and algorithms developers implement can help curb inflammatory content and help prescribe positive values into a platform's users. Disregarding anonymity for instance can get commenters to behave better. Features and algorithms play a major role in creating an online environment.

References

Allcott, H., & Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives, 31*(2), 211–236. https://doi.org/10.1257/jep.31.2.211

Baughan, A., Petelka, J., Yoo, C., Lo, J., Wang, S., Paramasivam, A., Zhou, A., & Hiniker, A. (2021). Someone Is Wrong on the Internet: Having Hard Conversations in Online Spaces. *Proc. ACM Hum.-Comput. Interact., 5(CSCW1).*

Bergen, M. (2019, April 02). YouTube executives ignored warnings, let toxic videos run rampant. Retrieved from https://www.bloomberg.com/news/features/2019-04-02/youtube-executives-ignored-warnings-letting-toxic-videos-run-rampant

Chaslot, G. (2019) The toxic potential of YouTube's feedback loop. Wired, July 13, 2019. https://www.wired.com/story/the-toxic-potential-of-youtubes-feedback-loop/

Christopherson, K. M. (2007). The positive and negative implications of anonymity in Internet social interactions:"On the Internet, Nobody Knows You're a Dog". Computers in Human Behavior, 23(6), 3038-3056.

Crockett, M. (2017) How social media makes us angry all the time. Big Think. https://www.youtube.com/watch?v=fE_QoebLUFQ

Davenport, D. (2002). Anonymity on the Internet: why the price may be too high. *Communications of the ACM*, *45*(4), 33-35.

Latour, B. (2005). Reassembling the social: An introduction to actor-network-theory. Oxford University Press.

Lipinski-Harten, M., & Tafarodi, R. (2013, June 28). Attitude moderation: A comparison of online chat and face-to-face conversation. Retrieved October 25, 2022, from https://doi.org/10.1016/j.chb.2013.06.004.

Lutz, S., & Schneider, F. M. (2021). Is receiving dislikes in social media still better than being ignored? The effects of ostracism and rejection on need threat and coping responses online. *Media Psychology*, *24*(6), 741-765.

Merrill, J. B., & Oremus, W. (2021, October 26) "Five points for anger, one for a 'like': How Facebook's formula fostered rage and misinformation". The Washington Post, https://www.washingtonpost.com/technology/2021/10/26/facebook-angry-emoji-algorithm/

Miller, A. (2013). On Godwin's Law: A Statistical Analysis on the Distribution of Nazi Analogies in Online Discussion https://alex.miller.im/files/Godwins-Law-Statistical-Analysis-Alex-Miller.pdf.

Munn, L. (2020). Angry by design: toxic communication and technical architectures. *Humanities and Social Sciences Communications, 7(1), 53.*

Waters, J. K., & Lester, J. (2010). *The Everything Guide to Social Media: All you need to know about participating in today's most popular online communities*. Simon and Schuster.

Wang, Y., Norcie, G., Komanduri, S., Acquisti, A., Leon, P. G., & Cranor, L. F. (2011). 'I Regretted the Minute I Pressed Share': A Qualitative Study of Regrets on Facebook. *Proceedings of the Seventh Symposium on Usable Privacy and Security*. Pittsburgh, Pennsylvania. doi:10.1145/2078827.2078841