“I’m Gonna Unalive Myself”: The Effect of AI Censorship on Internet Language

An STS Research Paper
presented to the faculty of the
School of Engineering and Applied Science
University of Virginia

by

Nathan Snyder

May 12, 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.

Nathan Snyder

STS Advisor: Peter Norton

“I’m Gonna Unalive Myself”: The Effect of AI Censorship on Internet Language

All major social media platforms use artificial intelligence to moderate the discourse on their platforms (Cobbe, 2020; Young, 2021). The cost-effective and automated nature of AI software has led to its widespread use in censoring user-generated content. Between April and June 2021, “automated flagging” identified 96% of the videos removed from YouTube (YouTube, n.d.). Despite these efforts, users have discovered ways to circumvent censorship through *censorship slang* — altering censored words to prevent their detection by AI censorship software. These linguistic innovations have revolutionized online communication and have even had implications in offline communication.

Social media platforms censor content they deem objectionable for many reasons, including removing hate speech, correcting misinformation, attracting new users, and maintaining the platform’s reputation (Cobbe, 2020; Gallo & Cho, 2021). For instance, TikTok describes its mission as “to inspire creativity and bring joy” (TikTok, n.d.-a). In alignment with this mission, TikTok’s community guidelines prohibit many videos that discuss sensitive topics, such as suicide, self-harm, and disordered eating (TikTok, n.d.-b). Violative videos are removed from TikTok, and user accounts that have posted several violative videos may be suspended or banned (Han, 2021). Between July and September 2022, TikTok removed 110 million videos for violating community guidelines (TikTok, 2022). Other major social media platforms have similar missions, community guidelines, and penalties for violators (Bateman et al., 2021; Cobbe, 2020).

Social media platforms often remove content for reasons beyond their stated guidelines. Users of these platforms have become increasingly disgruntled, claiming that guidelines are selectively and inequitably enforced (Díaz & Hecht-Felella, 2021). For example, TikTok consistently removes videos featuring people from most marginalized communities and even people that TikTok categorizes as “ugly” or “poor” (Ohlheiser, 2021; Biddle et al., 2020). Similarly, YouTube automatically demonetizes videos with titles referencing certain marginalized communities, most notably the LGBTQ+ community (Sealow, n.d.; LGBTQ+ v. Google/YouTube, 2019). Platforms often give no explanation or recourse after sanctioning user content, leaving users frustrated and confused (Delkic, 2022; Díaz & Hecht-Felella, 2021).

As a rebellion against this inequitable censorship, social media users have turned to censorship slang. Censorship slang involves specific linguistic techniques that social media users employ to disguise their messages. These disguises allow messages that would traditionally be censored to pass by AI censorship software unflagged. Users who believe that social media platforms unfairly punish them use censorship slang to communicate freely without the fear of censorship by AI censorship software.

While AI censorship software is used for most moderating tasks, human moderators are needed for situations such as when the software is unsure, when users report posts, or when users appeal removals (Young, 2021; Han, 2021). The software learns from the decisions of human moderators, allowing it to change and evolve (Sartor & Loreggia, 2020). This evolution creates a race between users and AI censorship software: what is acceptable for users to say one day may not be acceptable the next day. Social media users must change how they speak over time to account for the AI censorship software that regulates their posts. As one Twitter user explained it, “algorithms are causing human language to reroute around them in real time” (@0xabad1dea, 2021).

Censorship slang is a dynamic system of resistance against censorship that allows social media users to express themselves freely online. Internet users can employ specific linguistic techniques to modify their language so that other people instinctively understand but machines do not. Much as human languages do offline, censorship slang adapts to its environment, keeping users one step ahead of AI censorship software (Appendix I). Social media users have embraced censorship slang as a means for free expression and resistance against the AI censorship software that tries to silence them.

Review of Research

Little is publicly known about the specific algorithms used by social media platforms, but researchers have conducted several studies to understand them better. Grandinetti used press releases by Facebook and TikTok to uncover information about their censorship algorithms (Grandinetti, 2021). Biddle et al. examined leaked internal TikTok documents and found that their algorithm intentionally discriminates against certain groups of people (Biddle et al., 2020). YouTubers Sybreed and Sealow experimentally determined which words cause YouTube to automatically demonetize a video (Sealow, n.d.). The exact algorithms of these companies are still unknown, but researchers like Grandinetti, Biddle et al., Sybreed, and Sealow have provided limited insight into how these algorithms work.

There has also been some research into specific censorship slang phenomena. Knight wrote about the treatment of sex on TikTok and terms like “seggs” and “seggsd” (Knight, 2022). Barasa wrote about the usage of leetspeak on social media platforms in Kenya (Barasa, 2013). Tait investigated how suicide is discussed on TikTok, with special attention given to the term "unalive" (Tait, 2022). While these researchers provided great insight into specific examples of censorship slang, they did not address the greater picture of online censorship slang.

Most research on censorship slang has focused on slang in English, but some researchers have explored censorship slang in other languages. Cho and Kim discussed censorship slang in Korean (Cho & Kim, 2021). Xu provides an overview of censorship by the Chinese government and how Chinese internet users circumvent the censorship (Xu, 2014). Knockel measured the decentralization of Chinese censorship and the different methods that Chinese internet users employ to evade censorship (Knockel, 2018). This paper only focuses on censorship slang in English and provides a few examples from Spanish. Cho, Kim, Xu, and Knockel's research on censorship slang in other languages provides valuable insights into the broader landscape of censorship slang.

Homophones

Homophones are frequently used to disguise censored words either as acceptable preexisting words or as nonsense words. In this form of censorship slang, the disguised word can still be recognized as the original word due to the phonetic similarity between the two. However, since the spelling is different, the disguised word is not automatically flagged by AI censorship software. For example, the phrase "smoke this ounce of ouid" is not confusing to most English-speaking internet users (Rock, 2021). The English word "weed" (pronounced /wiːd/) can be reinterpreted by replacing "wee" (/wi/) with the French "oui" (/wi/). The result, "ouid" (/wiːd/), is recognizable as the original word to people familiar with the French pronunciation of "oui." This method of censorship slang is very effective in text posts, but since the disguised word is pronounced the same as the original word, this method is much less effective in audio and video posts.

It is not necessary that the original word and the disguised word have the exact same pronunciation. Approximate phonetic substitutions, like “corn” for “porn,” work almost as well as exact phonetic substitutions, while also allowing for more creative substitutions and facilitating use in audio and video posts. Users often opt for the disguised word to be an innocuous, pre-existing word. For example, the phrase “sewer slide note” is an intuitive replacement for “suicide note” (Schauer, 2021). The words “sewer” and “slide” are inconspicuous words and were intentionally chosen as to not cause suspicion upon review of the post by AI censorship software. The main benefit of using approximate phonetic substitutions is that when AI censorship software catches on to the true meaning of a disguised word, the community can easily generate a new one. However, the danger of using pre-existing words is that the disguised phrase may not be instinctively comprehensible: others may interpret phrases like “sewer slide” literally. Additional context is often necessary to communicate the intended message.

Bleeping

Bleeping is a technique used to censor audio media, achieved by emitting a *bleep* sound to obscure a word or phrase. Bleeping is a popular form of censorship because it only obscures the offending parts of a message, preserving the original phrasing. The traditional bleeping technique still works in audio and video posts, but this approach is not possible in text posts. Instead, the concept of bleeping was adapted for text-based media by altering the characters in the offending word to create a similar effect. This type of bleeping involves either removing certain characters or replacing them with a *censor character*, such as a hyphen or asterisk.

Pre-dating social media, this typographical bleeping has been a common form of character substitution since the 1600's (Liberman, 2006-a). At that time, the most common censor character was the hyphen, and it was not common to remove characters. For example, in the court transcripts of the 1698 trial of Capt. Edward Rigby, objectionable words were censored with hyphens and periods: "he had ---- in his Breeches, but notwithstanding he could F--- him. . . . [He] took hold of *Rigbys* [sic] Priv... Member" (Collins, 1698; Liberman, 2006-b). This practice of censoring words is now ubiquitous. The only major difference in modern times is that the asterisk is by far the most common censor character.

Bleeping has also allowed for the creation of new initialisms and abbreviations. In speech, the most common form of bleeping is to only say the first character of the offending word. Terms like "the F word" and "F off" were already in common usage before AI censorship was prevalent (Spears, 2020; Ice-T, 2021). This structure paved the way for initialisms to be used as disguised versions of censored multiword phrases. Common initialisms used online today are "SA" for "sexual assault" and "SW" for "sex work(er)" (@diandrasdiandra, 2023). Abbreviations are less common bleeping techniques in speech, but they have become much more common online: "There are various things involved, like 'drrrr' [drugs]. And I hope that by 'drrrr' you all know what I'm referring to, because I'm not going to say things here that later cause [YouTube] to remove this video" (Quesada, 2022).[1] Both initialisms and abbreviations have become prevalent forms of censorship slang online.

[1] Original quote in Spanish: "Hay varias cosas de por medio, como 'drrrr.' Y espero que con 'drrrr' sepáis a lo que me refiero, porque aquí no voy a decir yo cosas que luego me quitan el vídeo."

Bleeping follows certain linguistic patterns that lead to censored word forms like “fck,” “fxck,” and “f” but not “---k” (Harbeck, 2016; @bitterjaem, 2021). Even with these linguistic restrictions, users can generate many possible ways of bleeping one offending word due to the abundance of possible censor characters and the option of how many characters to censor. However, as more characters are obscured, providing context becomes more necessary to indicate the censored phrase to readers. Initialisms, which remove most characters from the censored phrase, and highly censored words like “****” may be very ambiguous and prone to confusion.

Homoglyphs

Humans perceive words by recognizing the shapes of their component characters, while computers perceive them as binary data. As a result, characters that are visually similar but differ in their binary representations are considered identical by humans but not by machines. Internet users can take advantage of this fact by replacing characters in censored words with other characters that look similar but are encoded differently. The number of character substitutions available for homoglyph substitution is always limited by the characters that are available in the platform’s chosen character encoding scheme. Technological advancements like the introductions of Unicode and emoji have greatly expanded the number of homoglyph substitutions that are available to internet users.

In the 1980’s, a homoglyph-based word obfuscation system called leetspeak rose to popularity among gamers and hackers (McFadden, 2021). At that time, the dominant character encoding scheme was ASCII, which only allowed for the representation of 128 distinct characters. This gave leetspeak users limited options for character substitutions. Common leetspeak substitutions include “3” for “e,” “1” for “l,” and “0” for “o.” The use of leetspeak fell off in the 2000’s; however, with the advent of modern social media platforms, leetspeak has seen a resurgence to evade AI censorship software. Leetspeak-encoded words, such as “fvck” or “h4te,” are often easily understood by people but can pose significant challenges to AI algorithms that have not been trained on data containing such obfuscated words (@bitterjaem, 2021; @cypheryjw, 2021). As such, leetspeak has become a prevalent censorship slang tool for people seeking to spread their messages without the threat of censorship.

After Unicode gained ubiquity on the internet in the late 2000’s, it became possible to disguise words using the expanded set of characters that Unicode provided (Oo, 2020). Due to the development of the world’s writing systems, many Unicode symbols look very similar despite being encoded differently: the traditional Latin-script “A” (U+0041) looks very similar to “Α” (U+0391), “А” (U+0410), “Ꭺ” (U+13AA), “ꓮ” (U+A4EE), “🅰” (U+1F170), etc. Unicode greatly facilitated homoglyph substitutions, as the substitute characters look much more like the original characters than their leetspeak equivalents. Many examples of Unicode homoglyph substitution, such as “fuсk,” which uses the Cyrillic “с” (U+0441) instead of the Latin “c” (U+0063), are visually indistinguishable from the original censored word (@kirawontmiss, 2022).

The Unicode Technical Standard provides a list of symbols that can be used to normalize text with homoglyphs (Unicode Consortium, 2022). However, the Unicode Technical Standard only deals with very similar homoglyphs and does not address approximate homoglyphs or leetspeak. Due to advancements in artificial intelligence, it is increasingly possible to use machine learning and computer vision techniques to detect homoglyphs and normalize texts. AI models can be trained to recognize visual characteristics of letters, such as curvature and other geometric properties, and identify more approximate homoglyphs like those used in leetspeak (Majumder et al., 2020).

Semantic manipulation

Social media users can disguise phrases by manipulating the phrase's component parts, either through replacement or rearrangement. Replacement involves substituting one component with a different one that has a similar meaning, while rearrangement involves switching the order of components. The new components are not required to be transmitted in the same mode of communication as the original: component parts can switch between written words, emoji, gestures, etc. Text-based transformations frequently result in the creation of neologisms that widely replace the original word.

Neologisms are typically created to represent new concepts, but in the age of social media censorship, neologisms are sometimes created to replace existing censored concepts. A common semantic substitution tactic is double negation, where users replace a word by combining an antonym of the word with a negation affix (e.g., "non-," "anti-," "-n't"). For instance, instead of the word "death," users may blend opposites such as "life" and "alive" with negation affixes to create neologisms like "lifen't" and "unalive." In particular, the term "unalive" has gained widespread usage on the internet, resulting in phrases such as "unalive watch" to refer to "suicide watch" and "unalive gas pod" to refer to a certain suicide machine (Khalifa, 2021; @mo_sky_toe, 2022). These neologisms are not found in dictionaries, but since their components already exist in the language, they are typically understandable to anyone familiar with the language. Due to the lack of training data on these words, these neologisms bypass AI censorship software, making this method of censorship slang highly effective.

In addition to lexical and orthographic strategies, alternative modes of communication can also be used to disguise messages. In video posts, users often employ visual cues or gestures to convey prohibited ideas without using censored language: “Look up his face on Google and tell me you don’t want to just … uh … YouTube doesn’t let me say this … uh … take a … [speaker raises their hands up] … and … in his face … [speaker mimics swinging an ax]” (Quesada, 2023).[2] Since these cues are not based on traditional, textual language structures, they are more difficult for AI algorithms to recognize and classify.

X-phemisms

In all cultures, past and present, euphemisms and dysphemisms (together called “X-phemisms”) are common techniques for replacing taboo words (Burridge, 2017). As such, X-phemisms are popular choices for internet users looking to evade censorship. AI censorship algorithms have been trained on traditional X-phemisms, so new ones have been created to evade detection. Well-known examples include “do the nasty” for “have sex” and “accountant” for “sex worker” (@RobynDMarley_, 2022; Green, 2021). Emoji have also developed additional X-phemistic meanings to express censored ideas. “🍃” can be used as a substitution for “marijuana” or “weed,” and several emoji, such as “🍆” and “🍑,” have taken on sexual meaning due to their shapes. The creation of new X-phemisms as a response to AI censorship demonstrates the same fluidity and creativity of human language that allowed for the creation of traditional X-phemisms.

[2] Original quote in Spanish: “Busca su cara en Google y decidme que no tenéis ganas de … eh … YouTube no me deja decir esto … de … coger un … y … en la cara ….”

Since these X-phemisms are not always intuitive, they are much less likely to be caught by AI content moderation software. However, they also may not be immediately transparent to other users who are not already familiar with them. For instance, the term “mascara” caused significant confusion in early 2023 when a TikTok user posted a video mentioning how two people had tried his “mascara . . . without [his] consent” (Whipple, 2023). Actress and model Julia Fox commented on the video, stating that she did not feel bad for the user, without realizing that “mascara” was a code word for “sex” (@makenac12, 2023). This incident sparked a discussion about the appropriateness of X-phemisms and the potential confusion they may cause (@cinemamilf, 2023; Harrington, 2023). Even though X-phemisms work very well against AI censorship software, other users may misinterpret them or find them confusing.

Multiple methods

Users of censorship slang are concerned about the constantly evolving nature of AI censorship software. These users worry that their posts may be removed if the software uncovers the messages that they disguised with censorship slang. To address this issue, many users employ multiple forms of censorship slang simultaneously. For example, the phrase “porn star” can be disguised by applying a homophonic substitution to the word “porn” (resulting in “corn star”) and then a semantic substitution with the emoji equivalents of these words (resulting in “🌽⭐”) (@naui553sputnik, 2023). This transformation is very effective, because the characters “🌽” and “⭐” are generally innocuous symbols. Another salient example is the spelling of “lesbian” as “ledollarbean.” This transformation occurred in three steps: first a homoglyph substitution to “s” with leetspeak (“le$bian”), then an approximate homophone substitution to “bian” (“le$bean”), and finally a semantic substitution that spelled out the symbol “$” (“ledollarbean”)

(@vngelsarabia, 2021; @kyndikading, 2022). By layering different types of censorship slang, users can disguise their messages more effectively than if they only used one method.

Internet users can generate an endless array of substitutes for censored word by layering different methods of censorship slang. AI censorship software experiences difficulties with single forms of censorship slang, and using multiple forms of censorship slang compounds those difficulties. As such, using multiple layered methods of censorship slang is very effective at curtailing censorship by AI censorship software. However, like using single methods of censorship slang, using multiple methods of censorship slang is susceptible to incomprehensibility. When the disguised phrase is very different than the original, it might be unrecognizable to the people that the user is trying to communicate with.

Conclusion

Techniques like homophones, bleeping, homoglyphs, semantic manipulation, and X-phemisms let users deceive the AI censorship software that social media platforms use to silence them. However, different forms of censorship slang are used depending on the medium of communication. Homophones are highly effective in text but are usually less effective in audio and video due to the similarity in pronunciation between the original and disguised word. Homoglyphs are only effective in text because they only change the spelling of a word not the pronunciation. Semantic manipulation may be effective in any medium depending on the specific type of manipulation Bleeping and X-phemisms are effective across all media types. Due to these differences, content on text-based platforms, like Twitter and Reddit, experiences different types of censorship slang than content on video-based platforms like TikTok and YouTube.

The most prominent drawback to using censorship slang is that more context is often necessary to communicate the intended message. This need for additional context is seen throughout most forms of censorship slang. People who are not active on social media may not be able to understand some examples of censorship slang because censorship slang involves non-normative linguistic processes. In general, as disguised phrases differ more from their original phrases, more context is necessary. X-phemisms may even require an explicit explanation, such as in the case of "mascara."

Advancements in technology are a double-edged sword for internet users. Certain advancements, such as the introduction of emoji, have given internet users many new ways to create censorship slang. On the other hand, other advancements in technology, most notably the rise and eventual ubiquity of artificial intelligence, have facilitated the detection of censored topics even when disguised by censorship slang. This phenomenon has resulted in a race between internet users and AI censorship software, with the former striving to stay one step ahead of the latter (Appendix I).

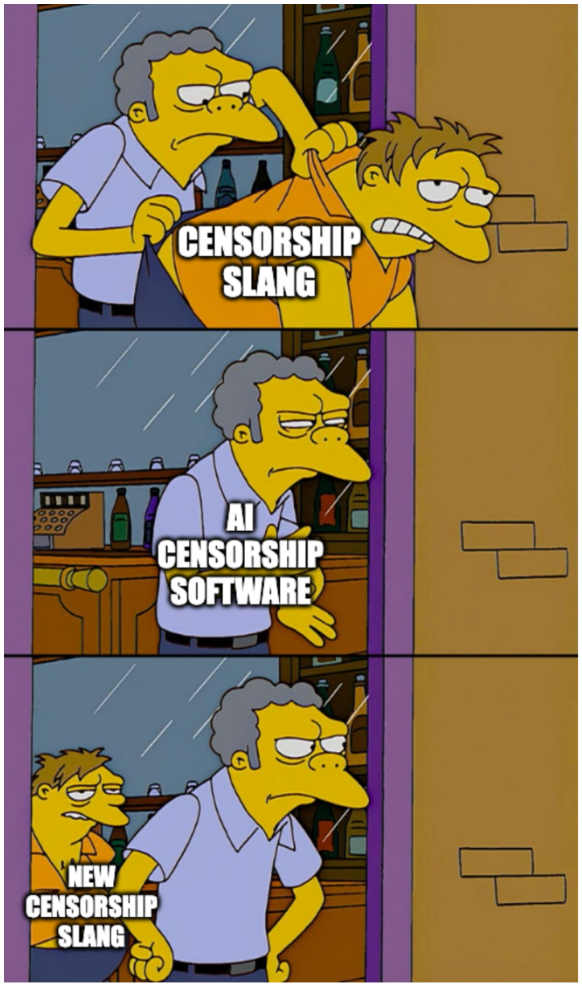
It is likely that these same processes happen in every language, but much more research is needed. Most of the evidence in this research paper came from English, with some examples from Spanish. Similar processes are known to happen in Korean and Chinese, but few other languages have been studied in this context because of their limited use online (Cho & Kim, 2021; Xu, 2014). An important next step in censorship slang research is the analysis of censorship slang in other languages, which may reveal additional methods of censorship slang not used in English.

The use of censorship slang terms does not stop at social media. Many disguised words that started as online censorship slang have begun to be used offline as euphemisms for their corresponding original words. Since the original word is considered too harsh or offensive on social media, the disguised word becomes a euphemism for it. Some people have noticed this censorship slang being used in situations that are not subject to the same level of censorship that creating the slang in the first place, calling this phenomenon the "unnecessary, voluntary sterilization of language" (McGonagall, 2023).

Despite opposition from social media platforms, internet users have created an ample collection of censorship slang techniques that allow them to freely express themselves online. The use of censorship slang by internet users demonstrates the inherent flexibility and adaptability of human language. This slang changes over time and adapts to its environment, staying one step ahead of AI censorship software. Censorship slang will continue to evolve, bringing with it the promise of free expression.

Appendix I

Illustration of the cyclical relationship between censorship slang and AI censorship software



(text by author, images by Persi et al., 2006)

References

@0xabad1dea. (2021, Dec. 15). Algorithms are causing human language to reroute around them (tweet). twitter.com/0xabad1dea/status/1471054750531702785

@bitterjaem. (2021, Feb. 11). “fck” “fvck” “fxck” “f” do it (tweet). twitter.com/bitterjaem/status/1359889148371468293

@cinemamilf. (2023, Jan. 27). you can’t refer to sexual assault as “stealing mascara” (tweet). twitter.com/cinemamilf/status/1618993990120132610

@diandrasdiandra. (2023, Mar. 1). we don’t talk enough about how he used (tweet). twitter.com/diandrasdiandra/status/1630966606393229315

@kirawontmiss. (2022, Jul. 11). who gives a fuck?? (tweet). twitter.com/kirawontmiss/status/1546531300156350465

@kyndikading. (2022, Aug. 1). My mans knew before i did (TikTok). www.tiktok.com/@kyndikading/video/7126803657034583338

@makenac12. (2023, Jan. 25). @juliafox hates men. ....this is gross (TikTok). www.tiktok.com/@makenac12/video/7192677389636472110

@mo_sky_toe. (2022, Jun. 25). No you can’t come to Switzerland only for the unalive gas pod (tweet). twitter.com/mo_sky_toe/status/1540810694576283652

@naui553sputnik. (2023, Jan. 11). I’m so sorry to break the mood (tweet). twitter.com/naui553sputnik/status/1613190401401450503

@RobynDMarley_. (2022, Dec. 1). It’s giving best friends, it’s giving soul mates (tweet). twitter.com/RobynDMarley_/status/1598403353641644032

@vngelsarabia. (2021, Feb. 21). Just because people are hating on me I’m reposting this (TikTok). www.tiktok.com/@vngelsarabia/video/6931588582570233094

Barasa, S. N. (2013, Jan.). Leetspeak in linguistic taboos: A study of Social Network Sites in Kenya. www.researchgate.net/publication/261133257_Leetspeak_in_linguistic_taboos_A_study_of_Social_Network_Sites_in_Kenya

Bateman, J., Thompson, N., & Smith, V. (2021, Apr. 1). How Social Media Platforms’ Community Standards Address Influence Operations. *Carnegie Endowment for International Peace*. carnegieendowment.org/2021/04/01/how-social-media-platforms-community-standards-address-influence-operations-pub-84201

Biddle, S., Ribeiro, P. V., & Dias, T. (2020, Mar. 16). Invisible Censorship: TikTok Told Moderators to Suppress Posts by “Ugly” People and the Poor to Attract New Users. *The Intercept*. theintercept.com/2020/03/16/tiktok-app-moderators-users-discrimination/

Burridge, K. (2017, Sep. 27). Euphemisms and Dysphemisms. doi:10.1093/OBO/9780199772810-0210

Cho, W. I. & Kim, S. (2021, Nov. 11). Google-trickers, Yaminjeongeum, and Leetspeak: An Empirical Taxonomy for Intentionally Noisy User-Generated Text. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 56–61. Association for Computational Linguistics. doi:10.18653/v1/2021.wnut-1.7

Cobbe, J. (2020, Oct. 7). Algorithmic Censorship by Social Platforms: Power and Resistance. *Philosophy & Technology*. doi:10.1007/s13347-020-00429-0

Collins, F. (1698, Dec. 7). An Account of the Proceedings against Capt. Edward Rigby. *Old Bailey*. Accessed from quod.lib.umich.edu/e/eebo2/A25634.0001.001

Delkic, M. (2022, Nov. 19). Leg Booty? Panoramic? Seggs? How TikTok Is Changing Language. *The New York Times*. www.nytimes.com/2022/11/19/style/tiktok-avoid-moderators-words.html. Accessed from archive.is/tBYVK

Díaz, Á. & Hecht-Felella, L. (2021, Aug. 4). Double Standards in Social Media Content Moderation. *Brennan Center for Justice at New York University School of Law*. https://www.brennancenter.org/our-work/research-reports/double-standards-social-media-content-moderation

Gallo, J. A. & Cho, C. Y. (2021, Jan. 27). Social Media: Misinformation and Content Moderation Issues for Congress. *Congressional Research Service*. https://crsreports.congress.gov/product/details?prodcode=R46662

Grandinetti, J. (2021, Sep. 12). Examining embedded apparatuses of AI in Facebook and TikTok. *AI & Society*. doi:10.1007/s00146-021-01270-5

Green, H. (2021, Apr. 3). Intel from TikTok (tweet). twitter.com/hankgreen/status/1378473387341668354

Han, E. (2021, Jul. 9). Advancing our approach to user safety. *TikTok Newsroom*. newsroom.tiktok.com/en-us/advancing-our-approach-to-user-safety

Harbeck, J. (2016, May 25). Why the f— do we do this and why the —k don’t we do that? *Strong Language: a sweary blog about swearing*. stronglang.wordpress.com/2016/05/25/why-the-f-do-we-do-this-and-why-the-k-dont-we-do-that/

Harrington, D. (2023, Jan. 27). Yes—other examples like seggs, corn etc are words that are close either in written or verbal form to the original form (tweet). twitter.com/DeliaMary/status/1619157826122940416

Ice-T. (2021, Aug. 12). People should know by now that Coco and I are far from normal parents (tweet). twitter.com/FINALLEVEL/status/1425898431143424004

Khalifa, M. (2021, Apr. 7). Jake Gyllenhaal should be on unalive watch (tweet). twitter.com/miakhalifa/status/1379801468564295686?s=20

Knight, M. (2022, May 6). #seggsd: Sex, Safety, and Censorship on TikTok. *San Diego State University*. digitallibrary.sdsu.edu/islandora/object/sdsu%3A200765

Knockel, J. (2018, Jan. 31). Measuring Decentralization of Chinese Censorship in Three Industry Segments. *University of New Mexico.* digitalrepository.unm.edu/cs_etds/90/

LGBTQ+ v. Google/YouTube (2019, Aug. 13). *United States District Court, Northern District of California, San Jose Division* (case 5:19-cv-04749, document 1). www.courthousenews.com/wp-content/uploads/2019/08/Censorship.pdf

Liberman, M. (2006-a, Jun. 10). The history of typographical bleeping. *University of Pennsylvania Institute for Research in Cognitive Science Language Log.* itre.cis.upenn.edu/~myl/languagelog/archives/003244.html

Liberman, M. (2006-b, Jun. 15). The earliest typographically bleeped F-word? *University of Pennsylvania Institute for Research in Cognitive Science Language Log.* itre.cis.upenn.edu/~myl/languagelog/archives/003253.html

Majumder, T. H., Rahman, M., Iqbal, A., & Rahman, M. S. (2020, Sep. 23). Convolutional Neural Network Based Ensemble Approach for Homoglyph Recognition. *Mathematical and Computational Applications*. doi:10.3390/mca25040071

McFadden, C. (2021, Jul. 19). 'Leetspeak' 101: What Exactly Is It? *Interesting Engineering*. interestingengineering.com/culture/leetspeak-101-what-exactly-is-it

McGonagall, S. (2023, Jan. 5). "unalive" "SA" "secs" "bd$m" "🌽ography" the unnecessary, voluntary sterilization of language (tweet). twitter.com/gothspiderbitch/status/1611128766054400002

Ohlheiser, A. (2021, Jul. 13). Welcome to TikTok's endless cycle of censorship and mistakes. *MIT Technology Review*. www.technologyreview.com/2021/07/13/1028401/tiktok-censorship-mistakes-glitches-apologies-endless-cycle/

Oo, M. T. (2020, Feb. 18). Unicode: the journey from standardizing texts to emojis. *Translation Royale*. www.translationroyale.com/the-history-of-unicode/

Persi, R. S. (Director), Maxtone-Graham, I. (Director), Groening, M. (Writer), Brooks, J. L. (Writer), & Simon, S. (Writer). (2006, Mar. 12). The Seemingly Never-Ending Story (Season 17, Episode 13) [TV series episode]. *The Simpsons*. Fox Broadcasting Company.

Quesada, E. (2022, Mar. 4). Euphoria - Soy Una Pringada (YouTube video). www.youtube.com/watch?v=ScwHpIpZqj0

Quesada, E. (2023, Jan. 19). Trisha Paytas, el ICONO (YouTube video). www.youtube.com/watch?v=dt_rFQbXgeY

Rock, B. (2021, Nov. 17). Okay I promise after I smoke this ounce of ouid (tweet). twitter.com/bretmanrock/status/1461022965387067392

Sartor, G. & Loreggia, A. (2020, Sep. 15). The impact of algorithms for online content filtering or moderation. *Committee on Citizens' Rights and Constitutional Affairs*. https://www.europarl.europa.eu/thinktank/en/document/IPOL_STU(2020)657101

Schauer, S. (2021, Jun. 1). I tried to start journaling but my hand got too tired and cramped (tweet). twitter.com/sarahschauer/status/1399917345171152899

Sealow. (n.d.). Demonetization report. docs.google.com/document/d/18B-X77K72PUCNIV3tGonzeNKNkegFLWuLxQ_evhF3AY. Accessed on Mar. 19, 2023.

Spears, B. (2020, May 4). Happy B day mamma (tweet). twitter.com/britneyspears/status/1257482571379699712

Tait, A. (2022, May 27). Are TikTok algorithms changing how people talk about suicide? *Wired*. www.wired.com/story/algorithms-suicide-unalive/

TikTok. (n.d.-a). About. www.tiktok.com/about. Accessed on Mar. 19, 2023.

TikTok. (n.d.-b). Community Guidelines. www.tiktok.com/community-guidelines. Accessed on Mar. 19, 2023.

TikTok. (2022, Dec. 19). Community Guidelines Enforcement Report. www.tiktok.com/transparency/en-gb/community-guidelines-enforcement-2022-3/

Unicode Consortium. (2022, Aug. 26). Unicode Technical Standard #39: Unicode Security Mechanisms. www.unicode.org/reports/tr39/

Whipple, C. (2023, Jan. 22). Idk if y'all will get it but (TikTok video). www.tiktok.com/@big_whip13/video/7191561464363158826

Xu, B. (2014, Sep. 25). Media censorship in China. *Counsel on Foreign Relations*. www.files.ethz.ch/isn/177388/media%20censorship%20in%20china.pdf

Young, G. K. (2021, Mar. 26). How much is too much: the difficulties of social media content moderation. *Information & Communications Technology Law*. doi:10.1080/13600834.2021.1905593

YouTube. (n.d.). Progress on managing harmful content. www.youtube.com/howyoutubeworks/progress-impact/responsibility/. Accessed on Mar. 19, 2023.