Article

SM+S
social media + society

Social Media + Society
July-September 2023: 1–17

DOI: 10.1177/20563051231194586
journals.sagepub.com/home/sms
Sage

You Can (Not) Say What You Want: Using Algospeak to Contest and Evade Algorithmic Content Moderation on TikTok

Ella Steen[1*], Kathryn Yurechko[2*], and Daniel Klug[3]

Abstract

Social media users have long been aware of opaque content moderation systems and how they shape platform environments. On TikTok, creators increasingly utilize algospeak to circumvent unjust content restriction, meaning, they change or invent words to prevent TikTok's content moderation algorithm from banning their video (e.g., "le$bean" for "lesbian"). We interviewed 19 TikTok creators about their motivations and practices of using algospeak in relation to their experience with TikTok's content moderation. Participants largely anticipated how TikTok's algorithm would read their videos, and used algospeak to evade unjustified content moderation while simultaneously ensuring target audiences can still find their videos. We identify non-contextuality, randomness, inaccuracy, and bias against marginalized communities as major issues regarding freedom of expression, equality of subjects, and support for communities of interest. Using algospeak, we argue for a need to improve contextually informed content moderation to valorize marginalized and tabooed audiovisual content on social media.

Keywords

TikTok, content moderation, algorithmic literacy, communicative practice, qualitative user study

Introduction

User communication in online environments has been shown to entail spoken and written linguistic effects based on the characteristics of a platform and the user community, leading to the adaptation and creation of forms of netspeak (Bucher & Helmond, 2018; Crystal, 2001; McCulloch, 2020). As yet another form of netspeak, algospeak is commonly understood as abbreviating, misspelling, or substituting specific words, for example, "seggs" for "sex" (Curtis, 2022; Delkic, 2022) or "clock app" for "TikTok," when creating a social media post with the particular goal to circumvent a platform's content moderation systems (Levine, 2022). The analysis of such communicative user strategies aims at exploring how people adapt technology and platform environments to their personal needs in various contexts of usage. While algospeak exists as a linguistic phenomenon on many social media platforms, it is largely connected to TikTok (Lorenz, 2022), its characteristics as an audiovisual platform, and its algorithmic content moderation. Examining algospeak on TikTok allows for observing users' cognitive processes and skills to better understand social and cultural factors, such as motivation, experience, or enjoyment, in social media interaction and participation.

From a sociolinguistic perspective, algospeak can resemble orthographic, lexical, and phonetic variations of standard language; however, these are commonly motivated by geographical (Grieve et al., 2019; Tatman, 2015), socio-cultural (Ilbury, 2020), or political (Shoemark et al., 2017) influences on users' online identities to signal and retain community membership (Stewart et al., 2017). In the context of computer-mediated language, algospeak is related to similar linguistic phenomena in mobile and Internet-based communication which, however, involve different motivations and intentions:

[1]Gordon College, USA
[2]Washington and Lee University, USA
[3]Carnegie Mellon University, USA

*The two authors equally contributed to this work.

Corresponding Author:
Daniel Klug, Carnegie Mellon University, Software and Societal Systems Department, TCS Hall 430, 4665 Forbes Avenue, Pittsburgh, PA 15213, USA.
Email: dklug@cs.cmu.edu

- *Textspeak, Chatspeak*, or *SMS-language* removes vowels, capitalization, spacing, and so on that are not necessary to understand a message, motivated by the formerly limited number of characters in SMS, and the required multiple pressing of a key to generate a letter (Drouin & Davis, 2009); likewise, *Digitalk* refers to manipulations of standard written language in online communication, such as Instant Messengers (Turner et al., 2014);
- *Leetspeak* (or l337) replaces letters with numbers or adds suffixes to words as a form of playful encryption that is easy to read (Perea et al., 2008); Leetspeak originates in bulletin boards and online gaming as ironic language variation to mock new users (Blashki & Nichol, 2005) but has since become Internet mainstream; similarly, *Chanspeak* was popularized on the 4chan imageboard as sub-community related misspelling and simplifying of words (Nascimento et al., 2019);
- *LOLspeak* (or LOLcat) humorously uses incorrect grammar and spelling as language plays, primarily in cat memes on social media (Fiorentini, 2013); likewise, *DoggoLingo* is a joyful idiom in dog memes to mimic how dogs would talk in human imagination (Punske & Butler, 2019).

All these phenomena modify written communication in networked and platform communities in a way that community members would be able to easily create, read, and decipher these linguistic variations. They foremost relate to online group membership and to users' online identities and presentation of the self in social media contexts (Herring & Kapidzic, 2015; Lee, 2014; Seargeant & Tagg, 2014). In contrast, algospeak is at first not meant for establishing identity or community membership through linguistic modification, but rather is used as a communicative practice in reaction to experiencing content moderation on a platform. Although algospeak may in fact function to define social media communities and membership, and though it may as well adopt existing linguistic practices, like Leetspeak or LOLspeak, it significantly differs in its primary intention to use language to circumvent especially algorithmic content moderation.

Content moderation comprises mechanisms to govern community activities and to screen the content users generate to facilitate cooperation and to prevent harm and abuse (Grimmelmann, 2015). Content moderation heavily relies on manual content assessment toward community guidelines (Seering, 2020) and on platform users who flag content they regard as violating guidelines (Crawford & Gillespie, 2016). Concerning forms of netspeak, such content moderation practices are necessary because linguistic variations or coded language are frequently used with bad intention to avoid algorithmic detection, for example, by political extremists to spread hate (Ben-David & Fernández, 2016; Bhat & Klein, 2020), to advocate controversial health information, for example, in pro-eating disorder communities (Chancellor et al., 2016), or as forms of online harassment, hate speech, and threat (Freed et al., 2018).

TikTok evolved into a main social media outlet for teenagers and young adults (Statista, 2023a) to negotiate and present their online identity around video creation and sharing, video-based interaction, and to form content-based communities of interest (Bhandari & Bimo, 2022; Burns-Stanning, 2020; Karizat et al., 2021; Klug, 2020; Simpson & Semaan, 2021). Users largely value TikTok for its uncannily accurate, yet, compared to other platforms, highly responsive recommendation algorithm (Taylor & Choi, 2022), which is subject to common folk theories about how and why TikTok selects, pushes, and restricts videos from appearing on "for you" pages (Karizat et al., 2021; Klug et al., 2021). Such algorithmic literacy among users is an important factor to predict user behavior and to make sense of user attitudes (Oeldorf-Hirsch & Neubaum, 2021), for example, to better understand how users make practical use of algorithms (Cotter, 2022), or how their experiences shape their perceptions of digital realities (Liao & Tyson, 2021). Algorithmic literacy likewise informs TikTok users' understandings of and experiences with the platform's content moderation system, which quite often relies on user reporting of content (Zeng & Kaye, 2022) but is mainly composed of human content moderators and artificial intelligence (AI), which involves automated decision-making and machine learning but is often colloquially called algorithms or bots (Grandinetti, 2021).

On TikTok, any content posted will first pass through TikTok's algorithms and will be reviewed by human moderators if potential community guidelines violations are identified (TikTok, n.d., 2021). The number of videos removed by automation on TikTok increased from 2020 to 2022 (Statista, 2023c), with minor safety, illegal activities, and adult nudity being the main reasons for removal of content (Statista, 2023b). Despite apparently careful content moderation, TikTok is known for having previously restricted content visibility for lesbian, gay, bisexual, trans, queer (LGBTQ+), disabled, and obese users without present guideline violations (Zeng & Kaye, 2022). Such incidents largely make TikTok's content moderation system appear as inconsistent and inaccessible to users and researchers alike (Malik, 2021).

Social media users have long been aware of opaque content moderation systems and how they shape platform environments. To our knowledge, this is the first article to analyze the characteristics and the usage of algospeak as a unique social media phenomenon. We take a qualitative social and behavioral approach by conducting user interviews to explore interactions between TikTok users and the platform's content moderation system. Our goal is to understand users' motivations and practices of using algospeak in relation to their experiences with TikTok's content moderation. We examine algospeak as a user strategy on TikTok to prevent

the platform from unjust restriction in the creation of mostly tabooed, stigmatized, or unwanted yet benign video content that does not violate TikTok's community guidelines, for example, videos related to sex education, LGBTQ+ activism, and mental health. Our study addresses communicative and linguistic aspects of usability on social media platforms in relation to online expression, self-presentation, and user-generated content in general. As a written form, algospeak on TikTok can be used in any text in video, that is, text on screen, video captions, and hashtags when creating and posting a video. Our study is guided by the questions of *how and why TikTok creators use algospeak*, *to what extent creators' understandings of content moderation on TikTok influence their use of algospeak*, and *what role algospeak plays in creators' freedom of expression on TikTok.*

Background and Related Work

Computer-mediated communication depends on platforms and channels, and incorporates aspects of written and spoken language alike (Herring & Androutsopoulos, 2015). Research by now mainly focuses on user-centered and sociolinguistic perspectives to examine the interplay of technology and language practices in social and cultural contexts online (Androutsopoulos, 2006). Human-Computer Interaction research on sociolinguistic perspectives frequently applies quantitative approaches and natural language processing to analyze various aspects of online and social media communication. For example, prior studies build classification models to contextualize out-of-vocabulary terms on Twitter (Maity et al., 2016), analyze language selection regarding social capital of bilingual Twitter users (S. Kim et al., 2014), examine language switching as user strategy in search engine usage (Wang & Komlodi, 2018), or perform rhetorical analysis of commenting on political Facebook posts to understand users' perception of partisan messages (Rho et al., 2018). Some mixed-method studies analyze manipulative rhetorics in factoid online articles (Tian et al., 2022), or interpret lexical markers of minority stress experience in LGBTQ+ communities on Reddit (Saha et al., 2019). Qualitative approaches focus on, for example, examining narration and expression of grief and the role of sociotechnical features on TikTok (Eriksson Krutröck, 2021), analyzing potential misunderstandings when using animated graphics interchange formats (GIFs) in nonverbal social media communication (Jiang et al., 2018), or showing how LOLspeak is important for humorous communication in gendered interaction during Hackathons (Brooke, 2022).

The field of netspeak has been extensively studied throughout the emergence of online communication and social media platforms (Baron, 2003; Crystal, 2001, 2011; McCulloch, 2020). Recent studies that focus on social media communication, users, and identity and community show, for example, that lexical changes and survival of lexical innovations in social media communication can be attributed to community network structures (Zhu & Jurgens, 2021), or how authors utilize skin-toned emojis in posts for self-expression and identity management (Robertson et al., 2021). Bhandari and Armstrong's (2019) analysis of Reddit communities finds that high affinity terms are used to signal community loyalty while also hindering new users from entering a community.

Likewise, research takes perspectives on how language and linguistic variations, such as leetspeak, are used to avoid censorship through creating noisy text (Cho & Kim, 2021), or how users substitute emojis for toxic language to evade algorithmic detection of problematic content (J. Kim et al., 2022). Gerrard (2018) identifies not using hashtags with pro-eating disorder content as an evasion strategy to make content moderation obsolete.

Only few studies specifically address sociolinguistics and language use on TikTok. Cervi et al. (2021) observe that communicative patterns, language, and emojis that are generic to TikTok are adapted into political communication on the platform; Vázquez-Herrero et al. (2022) equally find this for news presentations on TikTok. Some other studies examine the ways TikTok users utilize language in content creation to define their identities regarding communities of interest (Darvin, 2022; Simpson & Semaan, 2021), or analyze how LGBTQ+ slang is taken out of context and adapted into TikTok mainstream (Benitez, 2022). In particular, qualitative interview studies analyze how LGBTQ+ TikTokers domesticate the algorithm to manage their digital selves (Simpson et al., 2022), examine authenticity and self-presentation on TikTok within social support spaces (Barta & Andalibi, 2021), or describe motivations and creative practices in utilizing TikTok as a platform for social activism communities (Le Compte & Klug, 2021). Karizat et al. (2021) introduce the *Identity Strainer Theory*, which, based on users' observation of the content they receive on their "for you" pages, describes the user assumption that their TikTok content is being actively suppressed, filtered, or banned by the platform's algorithmic system.

Social media platforms generally apply various content moderation strategies with questionable effects. For example, deplatformization seems effective to stem toxic communication (Jhaver et al., 2021), while limiting access to toxic communities slowed down new member recruitment but did not affect established toxic communication within the community (Chandrasekharan et al., 2022). Content removal may effectively restore platform guideline compliance, yet it does not appear to affect overall platform behavior (Srinivasan et al., 2019). One solution might be a collaborative content moderation between AI and humans (Lai et al., 2022), mainly because AI-based content moderation on social media appears promising only from a technological perspective (Gillespie, 2020) but remains intransparent (Suzor et al., 2019), hard to understand, and largely unable to capture communicative context on the user side (Gorwa et al., 2020).

While the ways of communicating content moderation decisions back to users seem to depend on the type of platform (Thach et al., 2022), shadowbanning (Are, 2022), or reducing visibility of unwanted content (Gillespie, 2022) appear as popular practices on social media platforms. It has been demonstrated that content moderation decisions can influence users' social and cultural norms (Gillespie, 2018; Pilipets & Paasonen, 2022), and that social media users tend to develop "algorithmic imaginaries" (Bucher, 2017) and folk theories (DeVito et al., 2017; Lomborg & Kapsch, 2020; Savolainen, 2022) to make sense of algorithmic content moderation. In contrast, platforms usually provide dubious explanations for assumed shadowbanning (Le Merrer et al., 2021) which often leads users to blame human content moderators for platform intransparency, missing communication channels, and ultimately for content restrictions or bans (Myers West, 2018).

Research on users' understandings of and experiences with consequences of content moderation shows that users are more likely to understand content removal as fair when they are familiar with the community guidelines or when they receive explanations (Jhaver et al., 2019), yet unjust or unexplained banning of content can result in users leaving a platform (Chang & Danescu-Niculescu-Mizil, 2019). Interestingly, Vaccaro et al. (2020) find that, when given a feature to contest algorithmic decisions, users tended to question the automated content moderation system as unfair. One reason might be the perceived lack of standardization in human and AI moderation processes (Juneja et al., 2020). However, it is more crucial that while social media platforms generally enforce guidelines to remove apparent content violations, such as hate speech, they also disproportionately restrict and ban content by creators of marginalized identities, such as Black and LGBTQ+ users (Haimson et al., 2021)—a practice that was a main part of TikTok's algorithmic infrastructure (Rauchberg, 2022). For TikTok, Zeng and Kaye (2022) demonstrate that questionable content visibility based on algorithmic content moderation left creators confused and vulnerable to platform arbitrariness. Duffy and Meisner (2023) show that experiencing algorithmic invisibility motivated marginalized and stigmatized users to develop strategies to circumvent possible interventions.

Method

The goal of our interview study was to examine algospeak as a unique sociolinguistic phenomenon on social media platforms. For this, we specifically wanted to talk with TikTok users who demonstrated experience using algospeak when creating and sharing videos. We followed a qualitative research process (Flick, 2008) to analyze the creation and sharing of video content on TikTok as a way to interpret user data that emerges from everyday settings (Jensen, 2013). We first compiled a list of algospeak terms, and then used these terms to search for potential participants to ensure that we would talk to video creators who used algospeak in its intended meaning on TikTok.

Example Sampling and Participant Recruitment

In June 2022, we compiled a list of 70 commonly known algospeak examples by reviewing relevant social media news articles (e.g., Cheong, 2022; Huyghe, 2022; Lorenz, 2022), and posts on Twitter, Reddit, and TikTok in which users signified a potential use of algospeak by using a nonstandard word or emoji instead of a common word (see Table 1). We then searched each of the 70 algospeak examples in the "Videos" tab of the TikTok app, and scrolled through the results until we identified at most 10 videos that used the algospeak term in text on screen, captions, hashtags, and auto-generated captions (see Figure 1). We excluded videos that used a word or emoji only in the literal meaning, such as mentioning "cornucopia" in reference to literal cornucopias and not as algospeak for "homophobia" (see Figure 2). We also excluded videos that displayed graphic nudity, violence, crime, or extremism to protect the wellbeing of researchers and potential participants. For example, we did not select a video that used the snowflake emoji (❄) as algospeak to depict cocaine consumption. Three of the algospeak examples ("kermit sewer slide," "ouid," "sewer slide") did not return results when searched on TikTok, but displayed messages with community guidelines reminders or crisis hotline numbers from the app (see Figure 3).

We used the list of video examples to identify the creators of the videos as potential interview participants who used algospeak at least once on TikTok. We messaged creators who included their Instagram username in their TikTok profile through their Instagram profile, since TikTok does not permit direct messaging unless accounts are following each other. We messaged 198 creators and scheduled 19 interviews (9.6% participation rate) with TikTok creators in the United States (15), the United Kingdom (2), and Canada (2). The 19 participants were aged 19–32, and 73.7% identified as female which reflects the global gender distribution of TikTok users (Statista, 2023a). In addition, 16 participants identified as White, one as Black, one as Asian, and one as biracial (Black and White). As of July 2022, they had between 14k and 554.1k followers, had posted between 44 and 1.9k TikTok videos, and had in total received between 67.4k and 35.7m likes for their videos (see Table 2).

Data Collection

We conducted 19 qualitative semi-structured interviews (Longhurst, 2003) with TikTok creators in June and July 2022. The purpose of the interviews was to learn about creators' motivations of and experiences with using algospeak on TikTok. For the interviews, we provided participants with the following definition of algospeak: code words, phrases, and emojis that creators have been adopting to talk about

Table 1. This is the List of Algospeak Examples We Compiled From Websites and Social Media in Preparation for Recruiting Participants.

No.	Algospeak example	Clear word referent	No.	Algospeak example	Clear word referent
1	@b0rt!0n	abortion	36	SH	self-harm
2	accountant	sex worker	37	sh!t	shit
3	auti$m	autism	38	shmex	sex
4	Backstreet Boys reunion tour	COVID-19 pandemic	39	skripper	stripper
5	blink in lio	link in bio	40	SSA	same-sex attraction
6	blk	Black	41	str8	straight
7	bl00d	blood	42	$tripper	stripper
8	b00bs	breasts	43	SW	sex worker
9	clock app	TikTok	44	swimmers	vaccinated people
10	corn	porn	45	the vid	COVID-19
11	cornucopia	homophobia	46	tism	autism
12	cue anon	QAnon	47	unalive	dead, kill, suicide
13	depressi0n	depression	48	wh1te	White
14	ED	eating disorder	49	yt	White
15	fake body	N/A	50	🔗 in bio	link in bio
16	fork	fuck	51	💩	Black people
17	Frog	fuck	52	🐱	female genitals
18	grape	rape	53	🌽	porn
19	h0rny	horny	54	🌽⭐	pornstar
20	kermit sewer slide	commit suicide	55	🍆	male genitals
21	k1ll	kill	56	🐸	fuck
22	le dollar bean	lesbian	57	🤚🏻	White people
23	leg booty community	LGBT community	58	🥜	ejaculation
24	le$bean	lesbian	59	🟠⚫	PornHub
25	le$bian	lesbian	60	💩	shit
26	nip nops	nipples	61	🌻	Ukraine
27	not see	Nazi	62	🍒	breasts
28	opposite of love	hate	63	⚪	White people
29	Ouid	weed	64	🌶️	sex
30	panda express	COVID-19 pandemic	65	💦	ejaculation
31	Panini	COVID-19 pandemic	66	🍑	butt
32	panorama	COVID-19 pandemic	67	🎂	butt
33	SA	sexual assault	68	🤎	Black people
34	seggs	sex	69	❄️	cocaine
35	sewer slide	suicide	70	🏴	N-word

certain topics instead of using the actual terms. Interview questions were designed to better understand participants' general motivations for and routines of using algospeak, and specifically to examine if and how their understandings of TikTok's content moderation system influenced their use of algospeak. For example, we asked participants "What influences your decision to use algospeak?" In case participants mentioned content moderation as an influencing factor, we followed up with more specific questions, such as "How much does your experience with the TikTok algorithm play a role in your decision to use algospeak?" We also questioned participants about the ways in which algospeak influenced their self-expression on TikTok. Each of the interviews was conducted and recorded by one of two trained researchers after obtaining the participant's consent. The interviews lasted between 20 min and 41 min with an average duration of 30 min per interview. All interviews were anonymized and transcribed using an online transcription service.

Data Analysis

The interview data were analyzed by three trained researchers following a qualitative open coding process (Roulston, 2014). In order to retain their roles of unbiased researchers throughout the research process, all three researchers continually reflected on their own biographical, educational, and socio-cultural backgrounds, and stayed conscious of their gender, racial, and sexual identities as White cisgender males and females in relation to the identity-based subjects that participants discussed. All three researchers recognize that

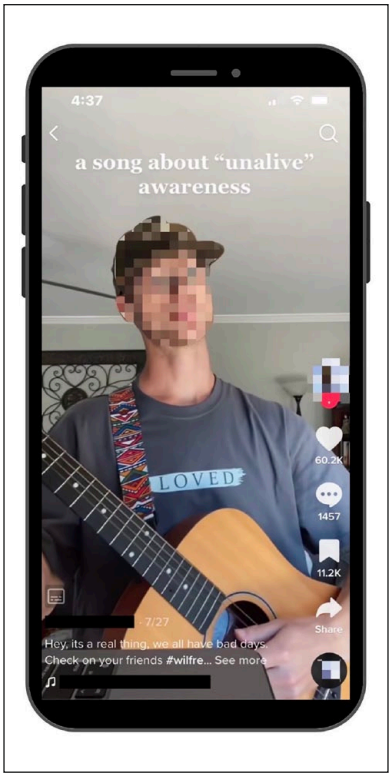

Figure 1. This is an example for a video that was included in the list; it uses the algospeak "unalive" to reference suicide awareness.

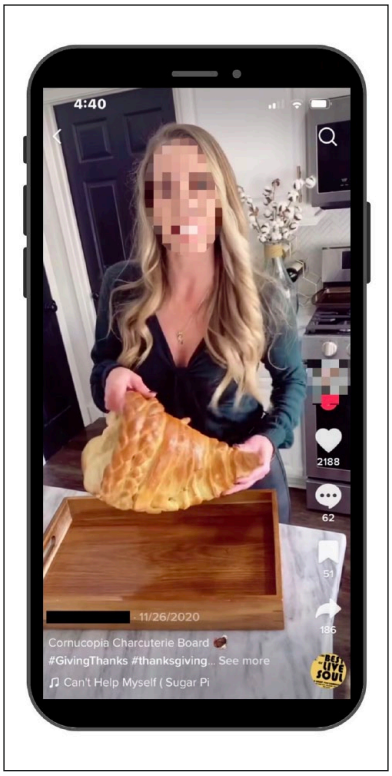

Figure 2. This is an example for a video that was not included; it uses the algospeak term "cornucopia" in a literal sense.

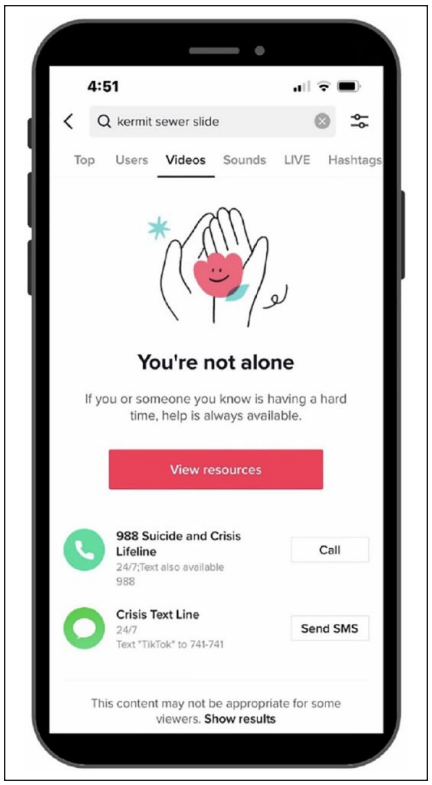

Figure 3. This is an example of an algospeak term ("kermit sewer slide") that did not result in a list of videos, but provided a Helpline as the algospeak term refers to suicide.

their personal backgrounds and identities are limitations that may have affected the research process. Each researcher coded the first interview, then all three researchers met to compare and discuss the naming and application of each others' codes (Campbell et al., 2013). In this first step of the coding process, we found a high level of agreement between all researchers regarding the identification of relevant interview segments and the creation and application of codes. This process of finding agreement ensured validity and reliability of the results that emerged in the qualitative coding across all coders (McDonald et al., 2019). We subsequently created a code book as a reference guide for the coding of the remaining interviews which were divided equally between the researchers. Each researcher used the code book accordingly to apply existing codes or added new codes if necessary (Benaquisto, 2008). Throughout the process, codes were grouped into categories based on similarities between codes to further organize the data and to designate central ideas that emerged through the coding process. For example, the codes "using algospeak in hashtags" and "using algospeak in autogenerated captions" were organized into the category "practical use of algospeak." After the coding process, all three researchers met again to review and revise the final code list. In this, 38 codes that were only applied three times or fewer were revisited for coding rigor. Of these, 23 codes were merged into 10 new and more comprehensive codes, and 15

Table 2. This Table Shows the Demographic Information for Participants (as of July 2022), Which Algospeak Term was Used to Recruit Participants, and Which Theme Participants Primarily Address in their Videos on TikTok. Participants Used a Great Variety of Algospeak Terms to Create Benign Videos on Subjects that Apparently were in Some Way Restricted by TikTok.

Participant	Followers	Likes	Videos	Recruitment through algospeak example	Algospeak example meaning	Video content theme
P01	68.2k	4.1m	2.0k	auti$m	autism	daily life
P02	255.7k	7.4m	1.9k	shmex	sex	pole dancing
P03	20.6k	1.7m	44	le$bean	lesbian	LGBTQ+
P04	14k	67.4k	161	blink in lio	link in bio	business promotion
P05	175.1k	8.9m	498	blk	Black	social commentary
P06	73.9k	5.4m	620	le$bean	lesbian	LGBTQ+
P07	86.3k	2.0m	204	clock app	TikTok	books
P08	268.1k	17.7m	1.1k	🔗 in bio	link in bio	sex education
P09	55.6k	1.9m	425	SA	sexual assault	sex education
P10	289.4k	2.9m	274	seggs	sex	sex education
P11	39.5k	2.9m	751	✋	White people	social commentary
P12	20.9k	389k	184	str8	straight	comedy
P13	59.9k	4.9m	1.4k	blink in lio	link in bio	product promotion
P14	103.2k	6.9m	310	SA	sexual assault	daily life
P15	554.1k	35.7m	1.9k	h0rny	horny	comedy
P16	264.1k	6.9m	163	corn	porn	sex education
P17	316.2k	5.0m	1.1k	blink in lio	link in bio	TikTok coaching
P18	80.3k	2.2m	586	h0rny	horny	daily life
P19	22.0k	1.5m	303	clock app	TikTok	LGBTQ+

codes were merged with existing and more frequently applied similar codes. Overall, the coding process resulted in 197 codes across the following 10 categories: *definition of algospeak*, *motivations for using algospeak*, *practical use of algospeak*, *effects of algospeak*, *opinions on algospeak*, *opinions on content moderation*, *experiences with content moderation consequences*, *understandings of content moderation*, *relationship with TikTok as a platform*, and *video creation topics*.

Results

In general, we find participants either adopted algospeak after seeing it on TikTok, or invented their own algospeak in order to avoid consequences for posting about topics they had previously felt were secretly unwanted but were not explicitly violating TikTok's community guidelines. Participants largely anticipated how TikTok's algorithm would read their video content, and used algospeak to primarily evade unjustified algorithmic content moderation while also making sure that target audiences could still find their videos and understand the video's context. Participants' experiences suggest non-contextuality, randomness, inaccuracy, and biases against marginalized communities are major issues regarding freedom of expression, equality of subjects, and support for communities of interest on TikTok.

Effects of Content Moderation Understandings on Algospeak Use

User Understandings of and Reactions to Content Moderation Procedures. Participants shared many experiences and practices that led them to assume how TikTok assigned consequences to unwanted videos. Many participants (P01–P04, P07, P10, P11, P14, P18) learned about TikTok's opaque content moderation system when applying trial-and-error strategies to see which subjects they could post about without consequences. Some analyzed the community guidelines (P10, P12, P15, P17) or looked to other TikTok creators (P01, P08, P13, P18) to find out which benign topics were non-publicly restricted by the app. Based on such experiences, participants developed tentative understandings of what video content is unwanted on TikTok, and used algospeak accordingly to alter videos that they felt might trigger consequences. We generally find that participants' perceptions of which parts of their video TikTok observed informed which parts of the video they applied algospeak to.

Experiences With Unwanted Content on TikTok. When we asked participants what video content they thought TikTok does not want on its platform, they mentioned videos related to violence (P07, P09, P14, P15, P17–P19), controversial events or beliefs (P03, P05, P09, P13, P16, P17, P19), and depictions of or references to sexual activities (P03, P08,

P10, P17, P18). We can see a causal relationship between these understandings and the algospeak participants used on TikTok. For example, P15, who creates comedy videos, noticed that videos dealing with anything related to death or violence were moderated. They described that pointing a fork toward an outlet or talking about "unaliving" oneself is algospeak they used to discuss "suicide." Likewise P17, who observed that videos involving controversial social or political topics were unwanted on TikTok, explained that creators in the community would, for example, use "shmortions" for "abortions" to discuss the 2022 overturning of Roe v. Wade. In addition, from their experience as sex education creator, P16 realized that all references to sexuality or sexual activities were heavily moderated which is why they would use the corn emoji (🌽) to share their thoughts on "porn." All participants were experienced TikTok users but the opacity in the platform's content moderation procedures often meant that they were merely "speculating" (P02) about what the app censored. For example, P19 openly talked abortion rights from a LGBTQ+ perspective in their videos and was surprised to see that others felt the need to use algospeak to discuss Roe v. Wade, saying, "I was like, you really think that they're checking that?" In any case, to their knowledge most participants (P01, P04–P14, P16–P19) concluded that TikTok looked for keywords in text on screen, video captions, and hashtags to identify unwanted video content. For example, P02, who frequently use the algospeak "shmex" for "sex" in their pole dancing videos, felt that TikTok kept a non-public "list of unapproved words," and explained that this belief came from personal experience with content moderation: "I referenced a sex toy in a video by its name [. . .] They were like, 'Nope, we've deleted the audio [. . .] we got rid of it cuz you said a word'." P18 described similar experiences when saying benign words such as "horny" in simple videos about various daily life subjects:

> I constantly had my videos taken down, and I wasn't filtering anything. [. . .] and then I realized I need to sort of put a cap on it and try to work my way around it [. . .] I've just had to obviously change the words slightly.

Such encounters gradually shaped participants' beliefs about content limitations and bans that would otherwise go unnoticed in the background and led them to replace particular words with algospeak: "when discussing anything of an adult topic, like sexual relationships, things of that nature, I have leaned on keywords to replace what I'm actually discussing" (P02). This strategy of swapping out trigger words demonstrates that participants primarily started to utilize algospeak as single-word replacements such as "awetistic" for "autistic" (P01), before moving on to more complex variations. Overall, these examples demonstrate that uncertainty regarding which subjects TikTok censors leads to differing assumptions of when algospeak is truly necessary to evade content moderation and subsequently to different practices of applying algospeak.

Variations of Using Algospeak. All participants assessed the details of videos they had previously restricted to infer that TikTok's content moderation mainly scrutinized written text. This encouraged them to use algospeak mostly in text rather than in spoken language ("I usually can get away with saying the actual word, but when it comes to any kind of text in the video, I do have to use algospeak" [P05]). Several participants (P01, P03, P05, P06, P08, P09, P12, P14, P16, P17) mentioned using written algospeak in more than one part of their video content as precaution. For example, when creating LGBTQ+ videos, P19 described how they would apply algospeak terms in text on video and also in the video captions: "[. . .] onscreen text [and] I would put it in the caption too [. . .] It's gonna be the algorithm watching it. It's gonna pay attention to both." Such multiple use of written algospeak illustrates a common strategy among all participants in reaction to their perception of TikTok's content moderation system as a watchful entity. It further informed participants' adaptation of algospeak beyond just written language.

Many participants (P02, P04, P08, P10, P11, P16, P18) also suspected that TikTok moderates video content based on audio scanning, meaning the identification of spoken language in videos. For example, P02 described several instances when creating pole dancing videos in which their audio was restricted because of sex- or sexuality-related language: "I said 'dildo' in a video and they're like, 'Nope'. And they cut, they just deleted the entire audio." Interestingly, some participants reacted to their suspicion of audio scanning by saying algospeak terms out loud in their videos. P08, a sex educator, had their videos taken down numerous times because of benign language which prompted them to instead encourage viewers to use condoms by saying, "Hey, use these rubber bands on your eggplant." Such strategies of verbalizing algospeak or emojis aimed at "dodging this little community guidelines bot" (P08) that they believed was wrongly inspecting their audio for violations. Another sex education creator (P16) followed another common strategy and whispered words like "vulva" to prevent incurring verbally triggered consequences. They explained their strategy through their observation that TikTok's algorithmic content moderation would sometimes even pick up on spoken algospeak: "Sometimes I'll whisper words because I've noticed even verbally saying things, TikTok will pick up on them." These advanced practices demonstrate creative ways of modifying spoken language as variants of algospeak to accommodate unjust content moderation of benign yet stigmatized video content.

Some participants (P02, P07, P12, P17, P18) even described instances in which they assumed that TikTok also wrongfully moderates unwanted visual components of videos on daily life subjects like books or business coaching. In order to avoid such unexplainable restrictions, participants explained they would add algospeak into the captions of videos that they thought would be mistakenly banned for their visual content, or directly as text-on-screen to certain

visual elements of a video. Participants indicated that adding written algospeak to such visual elements would prevent the video from being taken down. P12, a creator of comedy videos, said, "If you show yourself wearing a revealing outfit, it'll get suppressed or even banned for nudity. So if you say 'fake body', they don't count it" (P12). P18 who created videos on random everyday subjects followed the same strategy explaining, "obviously if you say it's a fake body, then they can't really take that down." These examples show how adding algospeak ("fake body") to harmless videos of people in sparse clothing helped to counteract the content moderation's mistake of interpreting them as being nude. It demonstrates how participants used algospeak as a necessary way to provide clarification and contextualization to TikTok's algorithmic content moderation. Moreover, it shows the inevitability of using algospeak to clarify visual content that might be misinterpreted because using clear word referents, for example "no nudity," would again wrongfully trigger content moderation. Such use of written algospeak further demonstrates that participants trusted algorithmic content moderation on TikTok to consider semantic relations between a video's text and image, and thus to understand that no violation exists when they add clarifying algospeak to their videos.

User Experiences with and Identification of Algorithmic Content Moderation Flaws. The majority of participants described how their perceptions of how accurately TikTok's human and AI content moderation identified unwanted video content informed their decision to use algospeak. Based on previous instances in which their videos were restricted, many participants (P02, P05–P07, P09–P14, P17, P19) considered that other users would report videos "out of spite" (P07) without present violations. In addition, some participants (P07, P09, P11–P13, P17) suggested that human content moderators would simply restrict videos that users or AI had previously reported or marked. However, almost all participants (P01, P02, P05–P19) seemed capable of differing between human and AI content moderation and suspected that AI is the main agent that their videos must appease. All participants were certain that TikTok had a pre-posting period during which AI determined whether a video would be allowed for upload or not. This encouraged participants to use algospeak as a means to circumvent possible faulty AI pre-checking and ensure their videos would pass this first stage of content moderation. As a result, participants named four main algorithmic content moderation flaws that they experienced and that influenced their uses of algospeak on TikTok: *non-contextuality*, *randomness*, *inaccuracy*, and *bias against marginalized communities*.

Non-contextuality. Overall, participants experienced content moderation on a great variety of videos and subjects. Most participants (P02, P05, P07, P09, P10, P12, P16–P18) explained this with their observation that TikTok's AI content moderation system does not consider the contextual meanings of single words or phrases when identifying unwanted video content. Some participants (P05, P07, P09, P16) described particular instances, for example, P09, who created educational videos about sexual assault, recounted several uses of algospeak in response to algorithmic non-contextuality: "I'd written out the word 'harassment', and it was flagged as a video for harassment or bullying and the video got taken down, whereas it survives on TikTok with me kind of censoring the word 'harassment'." Another of their videos got banned "obviously completely out of context, they took the video down cuz I said 'Nazi'." These examples demonstrate participants' struggle with AI not understanding their contextual use of a word in a video and hence incorrectly interpreting it as community guidelines violation, or as P09 described it: "You said Nazi, so I think that you are rooting for the Nazis." We find that all participants critically evaluated the fact that TikTok's algorithmic content moderation often identified words detached from their contextual meanings and in turn assigned unjust consequences. This realization demonstrates participants' algorithmic literacy, and motivated them to replace innocuous terms with algospeak in order to compensate for TikTok's failure to accurately capture the context and, hence, prevent the unreasonable banning of a video. P05 described this strategy for creating race-related videos in which they would use words like "white," but rather thought "maybe I should try the algospeak version of that word." Even in non-racial contexts, TikTok banned some videos in which P05 mentioned (skin) colors, leading them to realize that "If I were to say, 'my favorite color is white', I'd still have to use some sort of algospeak." We find that participants rightly anticipated high chances of non-contextuality in TikTok's algorithmic content moderation and therefore felt almost obligated to replace potentially triggering words like "white" with algospeak, such as "yt" (P05) in order to evade wrongful consequences.

Randomness. Most participants (P06–P08, P10–P12, P14–P19) also observed that TikTok's algorithmic content moderation was often random and did not provide any apparent reason for incurred consequences. For example, P11 perceived content moderation as "luck of the draw," since they had several of their social commentary videos taken down but were not able to match the restrictions to any of the details of their videos. In a few instances, such random banning seemed completely arbitrary and almost laughable, as P06 described,

> I made a joke about going to the Chicago bean and flicking it. [. . .] it wasn't like anything explicit was written or anything like that, but my video got taken down. And I thought that was hilarious, but also ridiculous.

One explanation could be that P06 primarily created LGBTQ+ videos which in our findings seem to generally

have a higher chance of receiving unjust restrictions. However, more participants (P10, P14, P16, P17) shared similar experiences with completely different videos and described how it led them to "play it safe" (P16), and use algospeak even when their video content would not violate, for example, adult nudity guidelines. For example, P14 had received several random unjust violations for benign daily life videos and concluded, "random things can just set off the censor, which is why you see people putting things in their captions like 'fake body'." In regards to receiving an arbitrary minor safety violation, they said, "some of the videos that I've had restricted for minor safety, it makes no sense. There's no skin showing, there's no minors in the video [. . .] it wasn't even a sensitive topic." This shows how creators of various video subjects adopt algospeak in order to evade a moderation system that they believe does not follow a logical procedure. By using algospeak strategically, they feel that they can outmaneuver the system's disorderly decision-making and safely post videos that do not violate the community guidelines.

Inaccuracy. From consuming videos on TikTok, the majority of participants (P01, P02, P05, P06, P08–P19) also suspected that TikTok's content moderation system does not consistently assign consequences to videos that actually violate community guidelines. P17, for example, observed discrepancies regarding explicit content: "I see a lot of sex workers on TikTok live streaming, which is insane that they're getting away with that." They added that such inaccuracy engendered a double standard in TikTok's content moderation, and other participants agreed that "[. . .] more good content like educational, safe, health content is being repressed than the actual harmful content being taken down" (P08). This shows how participants perceived TikTok's distribution of consequences as unjust, foremost concerning videos that deal with sex education, gender queer themes, and race-related subjects. We see that all participants who created sex education content experienced inaccurate banning which made them question moderation practices: "I've seen hate speech and actual explicit media that I'm like, 'How is this not taken down?' But my diagram of a vulva was" (P16). P02 described similar inaccuracies between their pole dancing content and sexually explicit videos:

> I was standing in eight-inch heels in a T-shirt with shorts, but I was holding onto a pole, and it got taken down immediately for being sexually explicit [. . .] versus I see videos of people, typically women in bikinis, like very, very small bikinis, but that somehow exists under the radar.

P05 observed the same issues when creating social commentary videos about race-related subjects: "Somebody using the N-word in a very derogatory way, 9 times out of 10 that video will stay up. But whenever people speak on that issue without using said word, our videos get taken down." We find that seeing such inaccurate content restrictions when consuming TikTok videos increased participants' awareness of covert algorithmic content moderation which in turn motivated them to adopt algospeak as an evasion tactic. P16 explained this as "when I'm doing my closed captioning for a vulva and words like that, I always have to algospeak them." For talking about race, P02 said, "you have to use some sort of algospeak. So, for example, when it comes to the word 'Black,' I use 'Bl@ck'." Overall, these experiences with unjust consequences motivated participants to share benign video content more covertly by utilizing algospeak to avoid being affected by the observed inaccurate moderation of content on TikTok.

Bias Against Marginalized Communities. Some participants (P01–P03, P05–P07, P11, P12, P19) experienced TikTok's content moderation procedures as biased against marginalized communities that they supported or personally identified with, including the LGBTQ+, Black, and disability communities. Participants who created LGBTQ+ videos experienced, for example, rather unreasonable restrictions, such as blocking videos that have the word "lesbian" in it as "[. . .] not safe for children," (P12), or censoring and shadowbanning of videos in which they used the term "lesbian":

> I made a video [. . .] I just said the word "lesbian" [. . .] that was taken down once or it just was [. . .] you know, if you've been shadowbanned, if there's just a "0" on your video, and you're not getting any views. (P06)

P07 witnessed numerous occurrences of content moderation consequences when watching other users' LGBTQ+ content, saying, "Videos I find sometimes just during Pride Month, sometimes weird things will happen with videos like that." In addition, P12 observed that funny videos by a creator they followed "got taken down [. . .]. And it felt almost like either somebody was directly targeting her, or she was being suppressed for being a Black lesbian." This led, especially participants who identified with marginalized communities, to believe that "TikTok is secretly quite homophobic, or not even secretly" (P07) which then informed the way they would use algospeak as a response to their experiences. For example, as a result of TikTok banning their comedy videos that dealt with LGBTQ+ subjects, P12 used the algospeak "le$bean," and P07 spelled "queer" with "three's instead of e's" (as "*qu33r*," the authors) in BookTok videos to prevent consequences for "queer language." In addition, P05 who posted social commentary videos on TikTok said,

> I made a video about LGBTQ+ issues [. . .] All I can think of is that I said "gay" and "trans," and TikTok shadowbanned that video 10 times. And it wasn't until the 11th time that they finally put it on the For You Page and didn't mess with it. And I had to use a lot of algospeak in that video.

As a result of experiencing these restrictions for their innocuous videos, participants who identified with the LGBTQ+ community concluded that TikTok might fundamentally restrict the visibility of any of their videos. Such impressions further motivated participants to adapt algospeak as a general way "[for] marginalized people to talk about issues that regular people already were getting to talk about" (P11). This shows how algospeak served as a means to regain visibility and to counteract algorithmic content moderation that participants felt was biased against them because of their identity and identification with marginalized communities.

Impact of Algospeak on TikTok's Content Moderation

We also asked participants if and in which way they thought algospeak could influence or even change algorithmic content moderation on TikTok. Participants largely viewed algospeak as an effective evasion tactic, but observed that TikTok would evolve its content moderation system in response to it and therefore compel creators to become more creative with their use of algospeak to continue to prevent consequences.

TikTok Adapts to Algospeak. All participants observed far less issues when they used algospeak instead of wording they identified as potential reason for previous restrictions. Most participants (P01, P03–P06, P08, P10, P11, P13–P16, P18, P19) considered algospeak to be effective in preventing consequences on TikTok, saying, for example, that it is "the number one reason why a lot of my videos haven't been taken down" (P07); however, they also expressed that algospeak often is a hit or miss. According to P05, for subjects like social commentaries, algospeak use is "90 percent effective, but there is that 10 percent of times where TikTok can catch onto what you're saying." In addition, many participants (P01, P02, P04, P05, P08, P09, P10, P14–P18) noticed that over time, the TikTok algorithm is learning and understanding the intended meaning behind algospeak, and therefore of moderating videos accordingly. P08 experienced this when posting sex education videos saying, "the little community guidelines bot is gonna keep following you, and it'll eventually be like, 'Oh, you've been using seggs a lot [. . .] I recognize this. This means sex'." Such experiences suggest that in order to keep evading unjust consequences creators must draw upon their algorithmic literacy and adapt their algospeak in response to TikTok's continuously improving algorithmic content moderation system. Moreover, affected communities of interest need to constantly negotiate new or revised algospeak in order to secure communication structures and mutual understanding.

Varying Effectiveness of Algospeak. Participants used algospeak for all kinds of subjects and types of videos, and some felt that certain forms of algospeak were more effective at circumventing unjust content moderation than others. We find that when participants used entirely new words such as "unalive," it proved to be more effective at evading content moderation than just editing, for example, the spelling or orthography of unwanted words like "suicide" (see Figure 4). For example, on creating sexual assault awareness videos, P14 noted, "I would say for sure algospeak in which you're changing the words like 'unalive' is completely different, works better than putting the word in punctuation." Moreover, many creators of sex education videos realized that TikTok's algorithmic content moderation easily figured out Leetspeak whereas new words were more difficult to comprehend: "if you put s3x, TikTok can figure out that you're trying to say sex [. . .] so now people have to say 'seggs'" (P05). We can see that participants carefully evaluated the algospeak they used and eventually used different algospeak or invented new terms to improve its effectiveness. Some participants pointed out that some algospeak terms had gotten popular on TikTok in general and had spread outside the original community which made them obvious to TikTok's content moderation system: "explicit algospeak, unalive, le dollar bean [. . .] actual words, I think that that is not far off from just getting added and lumped into other words they don't want to hear" (P02). As a result, especially many variations of written algospeak had become ubiquitous and easier for TikTok's content

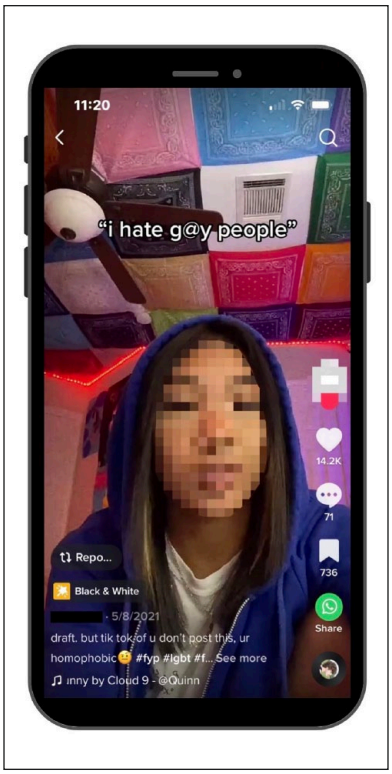

Figure 4. This is an example for a lexical variation ("g@y") as algospeak for "gay" to comment on LGBTQ+ hate.

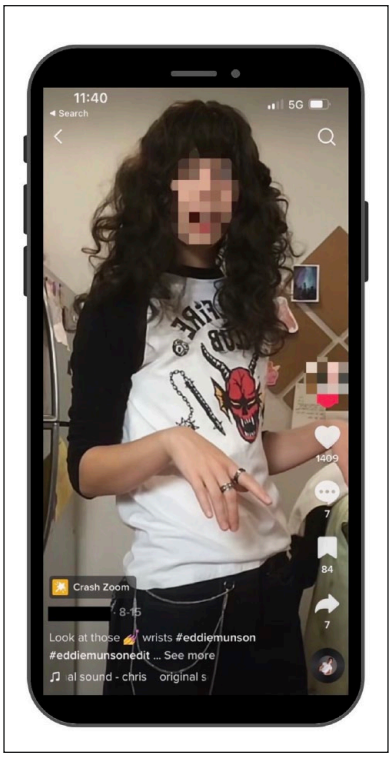

Figure 5. This is an example for the extended algospeak forms of using emojis (💅) and gestures (bend wrist) to replace the written word "gay."

moderation system to find and censor videos that contain it, hence making these algospeak terms less effective in evading unjust consequences.

Expanding the Definition of Algospeak. The fact that some written algospeak became ineffective over time inspired participants to create "unique and novel" (P19) forms of algospeak which they believed would be more difficult for TikTok to moderate. For example, using emojis (see Figure 5) which P05 described as a more impactful form than written text that allowed them to comment on social events. Likewise, P02 believed TikTok's content moderation system could not recognize gestures as readily as words, saying, "rather than coming up with different keywords, it became a game of omission for me [. . .] Like sometimes people, you know, bend wrists to indicate you're part of the queer community." These examples further demonstrate users' advanced algorithmic literacy and the creative ways in which they would extend and innovate algospeak beyond textual elements into skillful visual and audiovisual variations. These constantly improving user practices indicate that algospeak is more than the simple replacement of words. Therefore, the existing definition of algospeak needs to be enhanced, and algospeak needs to be understood as code words and linguistic variations, visual and multimodal communication, and audiovisual coherences in social media interaction.

Conclusion

The results of our interview study demonstrate that using algospeak on TikTok is largely a result of users observing and experiencing its algorithmic content moderation as being non-contextual, random, inaccurate, or biased and therefore unjust. Participants used algospeak primarily to circumvent what they perceived to be faulty content moderation of benign subjects, to prevent what they perceived as undeserved and irrational consequences, and ultimately to achieve and ensure adequate freedom of speech and equality on TikTok.

Our analysis illustrates how recurring experiences of receiving unjust content moderation leads to increased and improved use of algospeak. Participants first utilized algospeak to substitute or change text-on-screen, video captions, or hashtags that they anticipated as possibly inciting algorithmic content moderation. When they later observed that TikTok's content moderation algorithm appeared to learn the semantic referent behind algospeak terms and subsequently restricted it, they extended algospeak into using emojis, making gestures, or whispering or miming certain words. This progression of improving textual, visual, and audiovisual variants of algospeak is a creative user practice to work against unjust content restrictions that requires advanced algorithmic literacy and profound skills.

TikTok relies especially on machine learning and algorithmic detection to review and moderate uploaded videos (TikTok, n.d.), and foremost to automatically remove videos that are identified as community guideline violations (TikTok, 2021). However, our analysis demonstrates that TikTok's algorithmic content moderation is not capable of understanding the particular contexts in which participants discussed benign subjects. By default, this poses unjust restrictions to a large number of TikTok users who are talking about benign subjects like sex education, sexualities and gender identities, ethnicity, or social and political activism without violating community guidelines. Rather, TikTok's algorithmic content moderation system further traces, restricts, and suppresses user communication about unwanted themes, even when they are mentioned in harmless contexts. Algospeak appeared to be the main way for TikTok users to deal with this discrepancy of community guidelines, content moderation, and video creation. However, we find that TikTok's content moderation algorithm seems to learn algospeak and instead of taking into account the still benign context, it subsequently censors learned algospeak terms and limits their effectiveness for affected users. As an act of platform governance, this further hampers users' communicative needs and freedom of speech to address relevant social, cultural, and political subjects on the platform. In addition to previous research that found TikTok's content moderation system to be inconsistent and inaccessible to users (Malik, 2022), we show that such non-contextuality, randomness, inaccuracy, and biases also negatively affect users' attitudes

toward platform mechanisms, such as content moderation or algorithms, in general.

It would greatly improve user experience and usability on TikTok if automated content moderation were able to consider platform-internal contexts such as the ones we analyze in this study. Algospeak was largely applied by creators who belong to marginalized communities, such as LGBTQ+, or within communities of interest that address harmless yet unwanted subjects, such as sex education to evade further restraints from the platform. This shows that TikTok's general claim that their algorithms monitor and evaluate user behavior to optimize the user experience is not true for users who create videos about covertly unwanted subjects. In this regard, our analysis also provides further insights regarding restricted content visibility on TikTok (Zeng & Kaye, 2022), and confirms that TikTok creators of niche communities show sufficient digital skills and algorithmic literacy to manage their digital selves (Barta & Andibili, 2021; Simpson et al., 2022). In addition, considering communicative contexts might additionally help to counteract any malicious use of algospeak or similar practices.

Based on analyzing algospeak as a holistic communication phenomenon, we enhance the definition of algospeak from code words and linguistic variations to visual communication, multimodal communication, and audiovisual coherences. Compared to other forms of netspeak, like Chatspeak, Leetspeak, or LOLspeak, algospeak is a community-based means of communication that exceeds playfulness, encryption, or adaptation to platform features and environments to specifically combat experienced non-contextuality, randomness, inaccuracy, and bias against marginalized communities in algorithmic content moderation.

Future research could examine additional socio-technological perspectives, such as implications of algospeak for governance on social media platforms beyond TikTok, for example, similar linguistic adaptations to algorithmic content moderation on YouTube or Instagram. Socio-cultural research could further investigate the relationship between marginalized communities and algospeak, and take a non-Western perspective on TikTok and algospeak, for example, by analyzing user communication practices and content moderation in the context of political censorship and civic engagement. Based on our refined definition of algospeak, sociolinguistic approaches can study the effectiveness of different forms of algospeak (e.g., words versus gestures), or perform a language analysis of word-form algospeak and analyze what motivates users to adopt different variations of it.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Daniel Klug https://orcid.org/0000-0003-2320-3321

References

Androutsopoulos, J. (2006). Introduction: Sociolinguistics and computer-mediated communication. *Journal of Sociolinguistics*, *10* (4), 419–438. https://doi.org/10.1111/j.1467-9841.2006.00286.x

Are, C. (2022). The shadowban cycle: An autoethnography of pole dancing, nudity and censorship on Instagram. *Feminist Media Studies*, *22*, 2002–2019. https://doi.org/10.1080/14680777.2021.1928259

Baron, N. S. (2003). Language of the internet. In A. Farghali (Ed.), *The Stanford handbook for language engineers* (pp. 59–127). CSLI.

Barta, K., & Andalibi, N. (2021). Constructing authenticity on TikTok: Social norms and social support on the "fun" platform. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1–29. https://doi.org/10.1145/3479574

Benaquisto, L. (2008). Codes and coding. In L. M. Given (Ed.), *The SAGE encyclopedia of qualitative research methods* (pp. 85–88). Sage. https://doi.org/10.4135/9781412963909.n48

Ben-David, A., & Fernández, A. M. (2016). Hate speech and covert discrimination on social media: Monitoring the Facebook pages of extreme-right political parties in Spain. *International Journal of Communication*, *10*, Article 27.

Benitez, K. (2022). A content analysis of Queer Slang on Tik Tok. *Student Research Submissions*, *476*, 13. https://scholar.umw.edu/student_research/476

Bhandari, A., & Armstrong, C. (2019, November). Tkol, Httt, and r/radiohead: High affinity terms in Reddit communities. In *Proceedings of the 5th workshop on noisy user-generated text (W-NUT 2019)* (pp. 57–67). Association for Computational Linguistics.

Bhandari, A., & Bimo, S. (2022). Why's everyone on TikTok now? The algorithmized self and the future of self-making on social media. *Social Media+ Society*, *8*(1), 20563051221086241. https://doi.org/10.1177/20563051221086241

Bhat, P., & Klein, O. (2020). Covert hate speech: White nationalists and dog whistle communication on twitter. In *Twitter, the public sphere, and the chaos of online deliberation* (pp. 151–172). Palgrave Macmillan. https://doi.org/10.1007/978-3-030-41421-4_7

Blashki, K., & Nichol, S. (2005). Game geek's goss: Linguistic creativity in young males within an online university forum. *Australian Journal of Emerging Technologies and Society*, *3*(2), 71–80. https://hdl.handle.net/10536/DRO/DU:30003258

Brooke, S. J. (2022, April). Nice guys, virgins, and incels: Gender in remixing and sharing memes at hackathons. In *CHI conference on human factors in computing systems* (pp. 1–14). Association for Computing Machinery. https://doi.org/10.1145/3491102.3517627

Bucher, T. (2017). The algorithmic imaginary: Exploring the ordinary affects of Facebook algorithms. *Information, Communication & Society*, *20*(1), 30–44. https://doi.org/10.1080/1369118X.2016.1154086

Bucher, T., & Helmond, A. (2018). The affordances of social media platforms. In J. Burgess, A. Marwick, & T. Poell (Eds.), *The SAGE handbook of social media* (pp. 233–254). Sage.

Burns-Stanning, K. (2020). Identity in communities and networks TikTok social networking site empowering youth civic engagement. In *The 11th debating communities and networks conference* (Vol. 27, pp. 1–11). Debating Communities and Networks 11.

Campbell, J. L., Quincy, C., Osserman, J., & Pedersen, O. K. (2013). Coding in-depth semistructured interviews: Problems of unitization and intercoder reliability and agreement. *Sociological Methods & Research*, *42*(3), 294–320. https://doi.org/10.1177/0049124113500475

Cervi, L., Tejedor, S., & Lladó, C. M. (2021). TikTok and the new language of political communication. *Cultura, Lenguaje y Representación*, *26*, 267–287. https://doi.org/10.6035/clr.5817

Chancellor, S., Pater, J. A., Clear, T., Gilbert, E., & De Choudhury, M. (2016, February). #thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing* (pp. 1201–1213). Association for Computing Machinery. https://doi.org/10.1145/2818048.2819963

Chandrasekharan, E., Jhaver, S., Bruckman, A., & Gilbert, E. (2022). Quarantined! Examining the effects of a community-wide moderation intervention on Reddit. *ACM Transactions on Computer-human Interaction (TOCHI)*, *29*(4), 1–26.

Chang, J., & Danescu-Niculescu-Mizil, C. (2019, May). Trajectories of blocked community members: Redemption, recidivism and departure. In *The world wide web conference* (pp. 184–195). Association for Computing Machinery. https://doi.org/10.1145/3308558.3313638

Cheong, C. (2022, February 8). The phrase 'fake body' is spreading on TikTok as users think it tricks the app into allowing semi-nude videos. *Insider*. https://www.insider.com/fake-body-tiktok-hashtag-meaning-nudity-violations-trending-2-2022, accessed 16 August 2023.

Cho, W. I., & Kim, S. (2021, November). 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)* (pp. 56–61). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.wnut-1.7

Cotter, K. (2022). Practical knowledge of algorithms: The case of BreadTube. *New Media & Society*. 14614448221081802. https://doi.org/10.1177/14614448221081802

Crawford, K., & Gillespie, T. (2016). What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society*, *18*(3), 410–428. http://doi.org/10.1177/1461444814543163

Crystal, D. (2001). *Language and the internet*. Cambridge University Press. https://doi.org/10.1017/CBO9781139164771

Crystal, D. (2011). *Internet linguistics: A student guide*. Routledge.

Curtis, S. (2022, September 29). How TikTok changes the way we SPEAK. *Daily Mail*. https://www.dailymail.co.uk/sciencetech/article-11262889/TikTok-changing-way-SPEAK-phrases-like-quiet-quitting-le-dollar-bean.html, accessed 16 August 2023.

Darvin, R. (2022). Design, resistance and the performance of identity on TikTok. *Discourse, Context & Media*, *46*, Article 100591. https://doi.org/10.1016/j.dcm.2022.100591

Delkic, M. (2022, November 21). Leg booty? Panoramic? Seggs? How TikTok is changing language. *The New York Times*. https://www.nytimes.com/2022/11/19/style/tiktok-avoid-moderators-words.html, accessed 16 August 2023.

DeVito, M. A., Gergle, D., & Birnholtz, J. (2017, May). "Algorithms ruin everything." # RIPTwitter, folk theories, and resistance to algorithmic change in social media. In *Proceedings of the 2017 CHI conference on human factors in computing systems* (pp. 3163–3174). Association for Computing Machinery. https://doi.org/10.1145/3025453.3025659

Drouin, M., & Davis, C. (2009). R u txting? Is the use of text speak hurting your literacy? *Journal of Literacy Research*, *41*(1), 46–67. https://doi.org/10.1080/10862960802695131

Duffy, B. E., & Meisner, C. (2023). Platform governance at the margins: Social media creators' experiences with algorithmic (in) visibility. *Media, Culture & Society*, *45*(2), 285–304. https://doi.org/10.1177/01634437221111923

Eriksson Krutrök, M. (2021). Algorithmic closeness in mourning: Vernaculars of the hashtag #grief on TikTok. *Social Media+ Society*, *7*(3), 20563051211042396. https://doi.org/10.1177/20563051211042396

Fiorentini, I. (2013). "Zomg! Dis iz a new language": The case of lolspeak. Selected Papers from Sociolinguistics Summer School, *4*, 90–108.

Flick, U. (2008). *An introduction to qualitative research*. Sage.

Freed, D., Palmer, J., Minchala, D., Levy, K., Ristenpart, T., & Dell, N. (2018, April). "A stalker's paradise." How intimate partner abusers exploit technology. In *Proceedings of the 2018 CHI conference on human factors in computing systems* (pp. 1–13). Association for Computing Machinery. https://doi.org/10.1145/3173574.3174241

Gerrard, Y. (2018). Beyond the hashtag: Circumventing content moderation on social media. *New Media & Society*, *20*(12), 4492–4511. https://doi.org/10.1177/1461444818776611

Gillespie, T. (2018). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Gillespie, T. (2020). Content moderation, AI, and the question of scale. *Big Data & Society*, *7*(2), 2053951720943234. https://doi.org/10.1177/2053951720943234

Gillespie, T. (2022). Do not recommend? Reduction as a form of content moderation. *Social Media+ Society*, *8*(3), 20563051221117552. https://doi.org/10.1177/20563051221117552

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, *7*(1), 2053951719897945. https://doi.org/10.1177/2053951719897945

Grandinetti, J. (2021). Examining embedded apparatuses of AI in Facebook and TikTok. *Ai & Society*, 1–14. https://doi.org/10.1007/s00146-021-01270-5

Grieve, J., Montgomery, C., Nini, A., Murakami, A., & Guo, D. (2019). Mapping lexical dialect variation in British English using Twitter. *Frontiers in Artificial Intelligence*, *2*, Article 11. https://doi.org/10.3389/frai.2019.00011

Grimmelmann, J. (2015). The virtues of moderation. *Yale JL & Tech*, *17*, 42.

Haimson, O. L., Delmonaco, D., Nie, P., & Wegner, A. (2021). Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas.

Proceedings of the ACM on Human-Computer Interaction, *5*(CSCW2), 1–35. https://doi.org/10.1145/3479610

Herring, S. C., & Androutsopoulos, J. (2015). Computer-mediated discourse 2.0. In D. Tannen, H. E. Hamilton, & D. Schiffrin (Eds.), *The handbook of discourse analysis* (*Vol. 2*, pp. 127–151). Wiley. https://doi.org/10.1002/9781118584194.ch6

Herring, S. C., & Kapidzic, S. (2015). Teens, gender, and self-presentation in social media. In J. D. Wright (Ed.), *International encyclopedia of social and behavioral sciences* (2nd ed., pp. 1–16). Elsevier.

Huyghe, S. (2022). Algospeak: A new language to circumvent AI-powered content moderation: *Multilingual*. https://multilingual.com/algospeak-a-new-language-to-circumvent-ai-powered-content-moderation/, accessed 16 August 2023.

Ilbury, C. (2020). "Sassy queens": Stylistic orthographic variation in Twitter and the enregisterment of AAVE. *Journal of Sociolinguistics*, *24*(2), 245–264. https://doi.org/10.1111/josl.12366

Jensen, K. B. (2013). The qualitative research process. In K. B. Jensen (Ed.), *A handbook of media and communication research* (pp. 247–265). Routledge.

Jhaver, S., Appling, D. S., Gilbert, E., & Bruckman, A. (2019). "Did you suspect the post would be removed?" Understanding user reactions to content removals on Reddit. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–33. https://doi.org/10.1145/3359294

Jhaver, S., Boylston, C., Yang, D., & Bruckman, A. (2021). Evaluating the effectiveness of deplatforming as a moderation strategy on Twitter. *Proceedings of the ACM on Human-Computer Interaction, 5(CSCW2), 1–30*. https://doi.org/10.1145/3479525

Jiang, J. A., Fiesler, C., & Brubaker, J. R. (2018). The perfect one: Understanding communication practices and challenges with animated GIFs. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 1–20. https://doi.org/10.1145/3274349

Juneja, P., Rama Subramanian, D., & Mitra, T. (2020). Through the looking glass: Study of transparency in Reddit's moderation practices. *Proceedings of the ACM on Human-Computer Interaction*, *4*(GROUP), 1–35. https://doi.org/10.1145/3375197

Karizat, N., Delmonaco, D., Eslami, M., & Andalibi, N. (2021). Algorithmic folk theories and identity: How TikTok users co-produce knowledge of identity and engage in algorithmic resistance. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1–44. https://doi.org/10.1145/3476046

Kim, J., Wohn, D. Y., & Cha, M. (2022). Understanding and identifying the use of emotes in toxic chat on Twitch. *Online Social Networks and Media*, *27*, Article 100180. https://doi.org/10.1016/j.osnem.2021.100180

Kim, S., Weber, I., Wei, L., & Oh, A. (2014, September). Sociolinguistic analysis of Twitter in multilingual societies. In *Proceedings of the 25th ACM conference on hypertext and social media* (pp. 243–248). Association for Computing Machinery. https://doi.org/10.1145/2631775.2631824

Klug, D. (2020). "It took me almost 30 minutes to practice this": Performance and production practices in dance challenge videos on TikTok. *ArXiv Preprint: arXiv:2008.13040*. https://doi.org/10.33767/osf.io/j8u9v

Klug, D., Qin, Y., Evans, M., & Kaufman, G. (2021, June). Trick and please. A mixed-method study on user assumptions about the TikTok algorithm. In *13th ACM web science conference 2021* (pp. 84–92). Association for Computing Machinery. https://doi.org/10.1145/3447535.3462512

Lai, V., Carton, S., Bhatnagar, R., Liao, Q. V., Zhang, Y., & Tan, C. (2022, April). Human-AI collaboration via conditional delegation: A case study of content moderation. In *CHI conference on human factors in computing systems* (pp. 1–18). Association for Computing Machinery. https://doi.org/10.1145/3491102.3501999

Le Compte, D., & Klug, D. (2021, October). "It's viral!": A study of the behaviors, practices, and motivations of TikTok users and social activism. In *Companion publication of the 2021 conference on computer supported cooperative work and social computing* (pp. 108–111). Association for Computing Machinery. https://doi.org/10.1145/3462204.3481741

Lee, C. (2014). Language choice and self-presentation in social media: The case of university students in Hong Kong. In P. Seargeant & C. Tagg (Eds.), *The language of social media* (pp. 91–111). Palgrave Macmillan. https://doi.org/10.1057/9781137029317_5

Le Merrer, E., Morgan, B., & Trédan, G. (2021, May). Setting the record straighter on shadow banning. In *IEEE INFOCOM 2021-IEEE conference on computer communications* (pp. 1–10). IEEE. https://doi.org/10.1109/INFOCOM42981.2021.9488792

Levine, A. (2022, September 19). From camping to cheese pizza. 'Algospeak' is taking over social media. *Forbes*. https://www.forbes.com/sites/alexandralevine/2022/09/16/algospeak-social-media-survey/?sh=863232355e10, accessed 16 August 2023.

Liao, T., & Tyson, O. (2021). "Crystal is creepy, but cool": Mapping folk theories and responses to automated personality recognition algorithms. *Social Media+ Society*, *7*(2), 20563051211010170. https://doi.org/10.1177/20563051211010170

Lomborg, S., & Kapsch, P. H. (2020). Decoding algorithms. *Media, Culture & Society*, *42*(5), 745–761. https://doi.org/10.1177/0163443719855301

Longhurst, R. (2003). Semi-structured interviews and focus groups. *Key Methods in Geography*, *3*(2), 143–156.

Lorenz, T. (2022, April 8). Internet 'algospeak' is changing our language in real time, from 'nip nops' to 'le dollar bean'. *The Washington Post*. https://www.washingtonpost.com/technology/2022/04/08/algospeak-tiktok-le-dollar-bean/, accessed 16 August 2023.

Maity, S., Chaudhary, A., Kumar, S., Mukherjee, A., Sarda, C., Patil, A., & Mondal, A. (2016, February). Wassup? lol: Characterizing out-of-vocabulary words in Twitter. In *Proceedings of the 19th ACM conference on computer supported cooperative work and social computing companion* (pp. 341–344). Association for Computing Machinery. https://doi.org/10.1145/2818052.2869110

Malik, A. (2022, July 27). TikTok will provide select researchers with more transparency about its platform and moderation system. *TechCrunch*. https://techcrunch.com/2022/07/27/tiktok-select-researchers-more-transparency-about-platform-moderation-system/, accessed 16 August 2023.

McCulloch, G. (2020). *Because internet: Understanding the new rules of language*. Penguin.

McDonald, N., Schoenebeck, S., & Forte, A. (2019). Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proceedings of the*

ACM on Human-Computer Interaction, 3(CSCW), 1–23. https://doi.org/10.1145/3359174

Myers West, S. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, *20*(11), 4366–4383. https://doi.org/10.1177/1461444818773059

Nascimento, G., Carvalho, F., Cunha, A. M. D., Viana, C. R., & Guedes, G. P. (2019, October). Hate speech detection using Brazilian imageboards. In *Proceedings of the 25th Brazillian symposium on multimedia and the web* (pp. 325–328). Association for Computing Machinery. https://doi.org/10.1145/3323503.3360619

Oeldorf-Hirsch, A., & Neubaum, G. (2021). What do we know about algorithmic literacy? The status quo and a research agenda for a growing field. https://doi.org/10.1177/14614448231182662

Perea, M., Duñabeitia, J. A., & Carreiras, M. (2008). R34d1ng w0rd5 w1th numb3r5. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(1), 237–241. https://doi.org/10.1037/0096-1523.34.1.237

Pilipets, E., & Paasonen, S. (2022). Nipples, memes, and algorithmic failure: NSFW critique of Tumblr censorship. *New Media & Society*, *24*(6), 1459–1480. https://doi.org/10.1177/1461444820979280

Punske, J., & Butler, E. (2019). Do me a syntax: Doggo memes, language games and the internal structure of English. *Ampersand*, *6*, Article 100052. https://doi.org/10.1016/j.amper.2019.100052

Rauchberg, J. S. (2022). #Shadowbanned: Queer, trans, and disabled creator responses to algorithmic oppression on TikTok. In P. Pain (Ed.), *LGBTQ digital cultures* (pp. 196–209). Routledge.

Rho, E. H. R., Mark, G., & Mazmanian, M. (2018). Fostering civil discourse online: Linguistic behavior in comments of #MeToo articles across political perspectives. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 1–28. https://doi.org/10.1145/3274416

Robertson, A., Magdy, W., & Goldwater, S. (2021). Black or White but never neutral: How readers perceive identity from yellow or skin-toned emoji. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1–23. https://doi.org/10.1145/3476091

Roulston, K. (2014). Analysing interviews. *The SAGE Handbook of Qualitative Data Analysis*, 297–312.

Saha, K., Kim, S. C., Reddy, M. D., Carter, A. J., Sharma, E., Haimson, O. L., & De Choudhury, M. (2019). The language of LGBTQ+ minority stress experiences on social media. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–22. https://doi.org/10.1145/3361108

Savolainen, L. (2022). The shadow banning controversy: Perceived governance and algorithmic folklore. *Media, Culture & Society*, *44*(6), 1091–1109. https://doi.org/10.1177/01634437221077174

Seargeant, P., & Tagg, C. (Eds.). (2014). *The language of social media: Identity and community on the internet*. Springer.

Seering, J. (2020). Reconsidering community self-moderation: The role of research in supporting community-based models for online content moderation. *Proceedings of the ACM on Human-Computer Interaction*, *4*, Article 107. https://doi.org/10.1145/3415178

Shoemark, P., Sur, D., Shrimpton, L., Murray, I., & Goldwater, S. (2017, April). Aye or naw, whit dae ye hink? Scottish independence and linguistic identity on social media. In *Proceedings of the 5th conference of the European chapter of the association for computational linguistics: Volume 1, long papers* (pp. 1239–1248). Association for Computational Linguistics.

Simpson, E., Hamann, A., & Semaan, B. (2022). How to tame "your" algorithm: LGBTQ+ users' domestication of TikTok. *Proceedings of the ACM on Human-Computer Interaction*, *6*(GROUP), 1–27. https://doi.org/10.1145/3492841

Simpson, E., & Semaan, B. (2021). For you, or for "you"? Everyday LGBTQ+ encounters with TikTok. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW3), 1–34. https://doi.org/10.1145/3432951

Srinivasan, K. B., Danescu-Niculescu-Mizil, C., Lee, L., & Tan, C. (2019). Content removal as a moderation strategy: Compliance and other outcomes in the changemyview community. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–21. https://doi.org/10.1145/3359265

Statista. (2023a). *Distribution of TikTok users worldwide as of January 2023, by age and gender*. https://www.statista.com/statistics/1299771/tiktok-global-user-age-distribution/

Statista. (2023b). *Distribution of videos removed from TikTok worldwide Q3 2022, by reason*. https://www.statista.com/statistics/1249178/distribution-of-videos-removed-from-tiktok-worldwide-by-reason/

Statista. (2023c). *TikTok: Videos removed by automation 2020-2022*. https://www.statista.com/statistics/1300020/tiktok-videos-removed-by-automation/

Stewart, I., Chancellor, S., De Choudhury, M., & Eisenstein, J. (2017, December). #anorexia, #anarexia, #anarexyia: Characterizing online community practices with orthographic variation. In *2017 IEEE international conference on Big Data (Big Data)* (pp. 4353–4361). IEEE. https://doi.org/10.1109/BigData.2017.8258465

Suzor, N. P., West, S. M., Quodling, A., & York, J. (2019). What do we mean when we talk about transparency? Toward meaningful transparency in commercial content moderation. *International Journal of Communication*, *13*, 18.

Tatman, R. (2015). #go awn: Sociophonetic variation in variant spellings on Twitter. *Working Papers of the Linguistics Circle*, *25*(2), 97–108.

Taylor, S. H., & Choi, M. (2022). An initial conceptualization of algorithm responsiveness: Comparing perceptions of algorithms across social media platforms. *Social Media+ Society*, *8*(4), 20563051221144322. https://doi.org/10.1177/20563051221144322

Thach, H., Mayworm, S., Delmonaco, D., & Haimson, O. (2022). (In)visible moderation: A digital ethnography of marginalized users and content moderation on Twitch and Reddit. New Media & Society. 14614448221109804. https://doi.org/10.1177/14614448221109804

Tian, H., Ma, X., Bardzell, J., & Patil, S. (2022). Non-literal communication in Chinese internet spaces: A case study of fishing. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW1), 1–32. https://doi.org/10.1145/3512951

TikTok. (n.d). *Our approach to content moderation*. https://www.tiktok.com/transparency/en-us/content-moderation/

TikTok. (2021). Advancing our approach to user safety. https://newsroom.tiktok.com/en-us/advancing-our-approach-to-user-safety/

Turner, K. H., Abrams, S. S., Katíc, E., & Donovan, M. J. (2014). Demystifying digitalk: The what and why of the language teens use in digital writing. *Journal of Literacy Research*, *46*(2), 157–193. https://doi.org/10.1177/1086296X14534061

Vaccaro, K., Sandvig, C., & Karahalios, K. (2020). "At the end of the day Facebook does what itwants." How users experience contesting algorithmic content moderation. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW2), 1–22. https://doi.org/10.1145/3415238

Vázquez-Herrero, J., Negreira-Rey, M. C., & López-García, X. (2022). Let's dance the news! How the news media are adapting to the logic of TikTok. *Journalism*, *23*(8), 1717–1735. https://doi.org/10.1177/1464884920969092

Wang, J., & Komlodi, A. (2018, March). Switching languages in online searching: A qualitative study of web users' code-switching search behaviors. In *Proceedings of the 2018 conference on human information interaction & retrieval* (pp. 201–210). Association for Computing Machinery. https://doi.org/10.1145/3176349.3176396

Zeng, J., & Kaye, D. B. V. (2022). From content moderation to visibility moderation: A case study of platform governance on TikTok. *Policy & Internet*, *14*(1), 79–95. https://doi.org/10.1002/poi3.287

Zhu, J., & Jurgens, D. (2021). The structure of online social networks modulates the rate of lexical change. *ArXiv Preprint: arxiv: 210405010*. https://doi.org/10.48550/arXiv.2104.05010

Author Biographies

Ella Steen is a Student of Linguistics and Computer Science at Gordon College. Her research interests include computational linguistics and the sociolinguistic investigation of social media.

Kathryn Yurechko is a Student of Computer Science and Philosophy at Washington and Lee University. Her research interests include social computing, social media content moderation, and the support of marginalized communities online.

Daniel Klug (PhD, University of Basel) is a Faculty Member at the Software and Societal Systems Department at Carnegie Mellon University. His research interests include sociotechnical systems, digital media studies, algorithmic literacy, and qualitative user studies.