Article

new media & society
1–25

DOI: 10.1177/14614448231189476
journals.sagepub.com/home/nms
Sage

Shifting platform values in community guidelines: Examining the evolution of TikTok's governance frameworks

Ngai Keung Chan[*]
The Chinese University of Hong Kong, Hong Kong

Chris Chao Su[*]
Alexis Shore
Boston University, USA

Abstract

Social media can establish governance frameworks for their users through public-facing documents and policies. Such governance frameworks are value-laden and embody platform values. As a newly dominant platform in the United States, TikTok serves as an exemplary medium to study the evolution of platform values. Based on the iterations of TikTok's Community Guidelines from 2018 to 2022 ($N = 25{,}641$), we conducted longitudinal lexical analyses to determine changes in their structure and value salience. Then, through network analysis, we demonstrated how values co-exist by constructing co-occurrence networks. Our results reveal that the lexical complexity and value interconnection of these policies have increased over time. Certain values are more central in the networks than others (e.g. privacy, safety, and fairness), which may be attributed to a public outcry for change. The evolution of TikTok's governance frameworks follows three mechanisms (mediation, reversion, and founding paths) in shaping the core-periphery structures of platform values.

Keywords

Community guidelines, governance frameworks, platform values, platform, TikTok

[*] NK Chan and CC Su contributed equally to this study and share co-first authorship.

Corresponding author:
Chris Chao Su, Division of Emerging Media Studies, Boston University, 704 Commonwealth Ave., 303, Boston, MA 02215, USA.
Email: suchao@bu.edu

On January 8, 2020, TikTok published an updated version of its community guidelines, promoting transparency and user protections. Arguably, this policy revision—and other updates—respond to the increased regulatory and public attention over safety and privacy on the platform (Singer, 2020). As Lavanya Mahendran and Nasser Alsherif (2020)—members of TikTok's global trust and safety team—articulated, "These guidelines reflect our driving philosophy—providing a platform for creative self-expression while remaining safe, diverse, and authentic—and define a common code of conduct on our platform" (para 2).

Scholars have demonstrated how platforms enact and legitimize self-governance through policies such as community guidelines (Gillespie, 2018). We argue that updates on community guidelines represent a key aspect of platform evolution. Poell et al. (2022) refer to this as "governance frameworks," or standardized rules for regulating platform-based interactions. Platforms such as TikTok develop through the continuous co-evolution of governance frameworks, markets, corporate strategies, user groups (e.g. programmers, end-users, and complementors), infrastructures, and the broader social environment within which platforms operate (Barrett and Kreiss, 2019; Helmond et al., 2019; Poell et al., 2022). This process may be contingent upon the interests of divergent user groups. While critical social media research has approached platform evolution through the lens of multiple user groups—be they developers (Greene and Shilton, 2018), content creators (e.g. Arriagada and Ibanez, 2020), or lay users (e.g. Burgess and Baym, 2020)—as well as socio-technical infrastructures (e.g. Helmond et al., 2019; Kaye et al., 2022), this study offers a nuanced understanding of platform evolution at the textual and discursive levels.

Specifically, this exploratory mixed-methods study traces the evolution of TikTok's community guidelines in the United States from 2018 to 2022, focusing on the platform's corporate construction of values. Scholars have discussed values as the ideals expressed by a particular social entity, which may guide subsequent actions and judgments (Hallinan et al., 2022; Kraatz et al., 2020; Scharlach et al., 2023). Values can be conceptualized as a measurable property of platforms (i.e. nouns) or the valuation process by which ideals and expectations are articulated, negotiated, and contested (i.e. verbs) (Kraatz et al., 2020; see also Heinich, 2020; Scharlach et al., 2023). In other words, TikTok may position itself as a "safe" platform (values as nouns), whereas it can articulate "undesirable" through behaviors that threaten the very ideals about safety (values as verbs). Community guidelines—which are often written in user-friendly language and causal tone (Gillespie, 2018)—are value-laden statements (Maddox and Malson, 2020) that reveal the platform's explicit articulation of platform values (Hallinan et al., 2022), particularly concerning "the ideal experience of platforms as a community" (Scharlach et al., 2023: 6). Such idealized accounts reveal platform expectations across a diverse set of users and legitimize platform governance (Gillespie, 2018; Scharlach et al., 2023).

What is at stake here is how and which values are selectively expressed in community guidelines over time and for what practical purposes. We argue that TikTok's community guidelines serve as a compelling case study for three reasons. First, TikTok remains an emerging popular social media platform with over 700 million users worldwide (Perez, 2021). TikTok's community guidelines remained largely austere *before* January 2020. This allows for the consideration of the (re)articulation of platform values during its emergence and development. Second, though some may assume that platform values remain largely stable in platform policies, it is worth noting that well-stabilized artifacts can be renegotiated (Kline and Pinch, 1996). TikTok is under constant public scrutiny (Kaye et al., 2022) which can be considered as moments of breakdown where the platform may need to address relevant social groups' concerns (e.g. users and regulators) and re-stabilize the meanings of platform values, while ensuring the voice of community guidelines "is typically consistent with the character of the site" (Gillespie, 2018: 48). Third, community guidelines can be used to legitimize content moderation (Gerrard and Thornham, 2020; Gillespie, 2018) through platform values and the selective assignment of user responsibility (Konikoff, 2021; Scharlach et al., 2023). This case study, therefore, directs attention to how TikTok legitimizes different forms of content moderation through its community guidelines.

We aim to address the following research questions. First, how has TikTok constructed and reconstructed platform values in its community guidelines at the lexical and discursive levels? Second, how do these guidelines co-evolve with TikTok's responses to public controversies? Third, what are the implications of TikTok's community guidelines for the theorization of platform evolution and governance?

To address these questions, we combined quantitative and qualitative approaches in mapping TikTok's platform values. We first scraped and analyzed the snapshots of TikTok's community guidelines (N=25,641) between December 10, 2018, and March 7, 2022, from the Internet Archive's Wayback machine. We generated a corpus of sentence tokens for understanding the performative characters of the lexicon and the selective presentation of platform values through a series of lexical analyses and a regression-based path-dependency analysis. We enhanced the analysis of the framework by extending the quantitative results with a qualitative and thematic examination. Overall, this study contributes to platform studies in two ways. First, it offers a nuanced understanding of platform evolution through governance frameworks, an important yet relatively understudied institutional mechanism of platformization (Poell et al., 2022). Second, while previous research on platform policies and platform values (e.g. Hallinan et al., 2022; Scharlach et al., 2023) often focused on a specific timepoint, this study explores how TikTok's platform values and community guidelines evolve and its relation to public controversies. As community guidelines can outline the ideals about user activities within the platform as a community (Scharlach et al., 2023), this study affords opportunities for considering how these ideals evolve and how TikTok have used these ideals to legitimize a specific version of platform governance.

Literature review

Platform evolution

Internet historians and media scholars have directed attention to the temporality of platforms (e.g. Helmond and Van der Vlist, 2019; Helmond et al., 2019). On a discursive level, various stakeholders strategically use the term "platform" to legitimize certain user-generated content sites as well as the companies that own these sites (Gillespie, 2010). The discursive positioning of platforms evolves over time; for example, industry leaders continuously (re)articulate both the current and future-oriented vision for their platforms through public discourses (e.g. Hoffmann et al., 2018).

Recent research has drawn attention to the process of platformization (Helmond, 2015; Poell et al., 2022), which focuses on how platforms emerge as key infrastructures, markets, and institutions in various social and economic spheres. Platformization, in essence, is a story of platform evolution in which platforms format data and products to be "platform ready" (Helmond, 2015), maximize network effects (Srnicek, 2017), and gain infrastructural properties through boundary resources (e.g. APIs) and partnerships (e.g. Helmond et al., 2019). As a platform continues to evolve, multiple user groups can adapt their practices to sustain and subvert platform values and affordances (e.g. Arriagada and Ibanez, 2020; Burgess and Baym, 2020). A central throughline of these studies is that platform evolution is characterized by the entanglement between the platform firm's corporate discourses, its socio-technical infrastructures, and its users' practices (Helmond et al., 2019; Poell et al., 2022). It is noteworthy that platform evolution can be short-lived as exemplified by the notion of "platform transience" (Barrett and Kreiss, 2019). The concept suggests that ephemeral changes in platform policies and affordances often arise from external normative pressures and corporate interests.

Consider the case of TikTok, a platform that allows users to upload and engage with short videos. It was launched globally—outside of the Chinese market—in 2017. Dating back to 2014, TikTok's precursor, Musical.ly, carefully tailored the platform as a "parent-free" space for the creative expression of pre-teens while skillfully addressing parental concerns by positioning it as a utility app (Savic, 2021). As Musical.ly evolved to TikTok, it continued to strategically pitch its stimulation of creativity to users and advertisers, while attempting to distance itself from public controversies such as its connection with China (Kaye et al., 2022). Targeting youth as the key user population from the onset might explain the growing prominence of TikTok (Kaye et al., 2022). While it has been criticized for its insufficient protection of minors and user data (Badillo-Urquiola et al., 2019; Shutsko, 2020), TikTok's specific features (e.g. in-app video creation and socially creative features) encourage in-app user participation in challenges (Kaye et al., 2022) as well as positive offline lifestyle changes (Wang et al., 2022).

Through a case study of TikTok, we aim to document the evolution of community guidelines and theorize the processes through which such governance frameworks change.

Why study community guidelines? Understanding platform governance and values

Community guidelines reveal the rule-setting efforts of platforms (Gerrard and Thornham, 2020; Gillespie, 2018). They are discursive performances, meaning that they construct the "reality" of platforms and their governance approaches (Bucher, 2021; Gillespie, 2018; Hoffmann et al., 2018). Echoing Ziewitz and Pentzold's (2014) discussion of performativity, platform policies not only communicate "a specific version of governance but also a version of the world in which this notion of governance has its place" (p. 307). Consider, for example, the act of publishing community guidelines (or community standards). On one hand, releasing the internal documents that were previously obscured from the public can be interpreted as self-governance (Gorwa, 2019) that increases platform transparency. On the other hand, community guidelines may legitimize a social world where platforms should have the right to govern user activities and content (DeNardis and Hackl, 2015; Gillespie, 2018). Platforms often substantiate this by emphasizing how they "successfully" remove content that violates community guidelines. Studies have discussed how platform policies may paradoxically perpetuate what they claim to regulate because of the narrowed description of "harm" (DeCook et al., 2022), biased definition of sexual content (Ruberg, 2021; Zolides, 2021), and reactive policy enforcement (Konikoff, 2021).

Community guidelines differ from other types of platform policies such as terms of service and privacy policies because the former is primarily written for multiple user groups and is less concerned with arbitration (Gillespie, 2018). Community guidelines usually define what a platform is (or ought to be) and list the types of "problematic" content and behaviors that are prohibited (Gillespie, 2018). Common categories of prohibited content include hate speech, sexual content, self-harm, harassment, and violence (Gillespie, 2018; Jiang et al., 2020). In the United States, Section 230's "Good Samaritan" provision "gives platforms the leeway to develop their own community guidelines and enforce them as they see fit" (Caplan, 2018: 27).

Community guidelines have three important performative functions. First, they perform as rulebooks that guide content moderation though they are not necessarily translated into enforcement (Gerrard and Thornham, 2020; Gillespie, 2018). Platforms might withhold their content moderation policies (Caplan, 2018). For example, Facebook only started publishing the internal guidelines that enforce community standards in April 2018 (Gorwa, 2019). These rulebooks frame opportunities for user participation (Stein, 2013) and have the potential to influence the public discourse as users may use them to decide what they should or should not post (Gillespie, 2018). Yet, users may resist platform policies through collective action (Myers West, 2017) and individual practices (Poell et al., 2022). This reveals the tensions between platforms and users in shaping platform governance.

Second, community guidelines "articulate the 'ethos' of the site, not only to lure and retain users, but also to satisfy the platform's founders, managers, and employees, who want to believe that the platform is in keeping with their own aims and values" (Gillespie, 2018: 47). Scharlach et al. (2023) conceptualize community guidelines and other types of platform policies as boundary objects which can coordinate the divergent interests of platforms and other relevant social groups without consensus. Similar to the discursive positioning of "platforms" (Gillespie, 2010), platform policies help platforms avoid liability and navigate the regulatory demands (Caplan, 2018; Gillespie, 2018; Scharlach et al., 2023). This is exemplified by how platforms strategically mobilize and shape public values for their business interests (Scharlach et al., 2023; Van Dijck et al., 2018).

Third, community guidelines are "living documents" that (re)articulate and circulate platform values among users (Gerrard and Thornham, 2020; Gillespie, 2018; Scharlach et al., 2023). Hallinan et al. (2022) have developed a theoretical model for conceptualizing and operationalizing social media platform values. They documented how users and platforms differed in terms of their articulations of engagement and authenticity on

Twitter and Instagram. The two platforms tended to favor strategic engagement (i.e. facilitating interactions with high social media metrics) over social and civic dimensions of engagement (i.e. facilitating interactions for social support and civic engagement) in their policies, whereas users were more concerned with civic engagement as compared to the other two dimensions. Pinpointing platform policies as "ideal types," Scharlach et al. (2023: 5) contend that different types of platform policies outline distinct ideals about "the relationship between platforms and users" (terms of service), "treatment of personal data" (privacy policies), and "how people should express themselves and interact with others" (community guidelines). While Hallinan et al. (2022) discussed engagement and authenticity, Scharlach et al. (2023) identified expression, community, safety, choice, and improvement as the core value (out of 10 value clusters) across platform policies on Facebook, Instagram, TikTok, Twitter, and YouTube. Notably, values can be mobilized as objects and principles; for example, while platforms emphasized the value of expression, they meanwhile used it to justify the restrictions of certain content that might threaten the realization of the value (Scharlach et al., 2023).

Inspired by this growing body of scholarship, this study examines how TikTok has expressed and reconstructed platform values in its community guidelines over time. Indeed, researchers have indicated the importance of considering the shifting of community guidelines as discursive performance (Gillespie, 2018) and certain disruptive moments when platforms modified policies (e.g. Barrett and Kreiss, 2019; Myers West, 2017; Zolides, 2021). Moreover, scholars have recently begun to build archives that collect and curate platform policies so as to understand platform governance (i.e. Platform Governance Archive; Katzenbach et al., 2021). Yet, we hope to advance the scholarship by reconstructing the evolution of TikTok's community guidelines at both lexical and discursive levels. In keeping with existing scholarly accounts, changes in community guidelines are likely to be impermanent (Barrett and Kreiss, 2019) and may be the result of unexpected events (e.g. political scandals) (Bossetta, 2020). Therefore, a systematic documentation of community guidelines allows us to empirically consider these theoretical assertions.

TikTok's platform governance

This study focuses on the construction of platform values in TikTok's community guidelines in the United States. We select the US version for two reasons. First, TikTok is widely used and continuously subject to public scrutiny in the United States. Second, we posit that the within-country comparison is a productive way for exploring the evolution of community guidelines. Yet, we are aware that platform policies may vary across countries. Although there are similarities between the infrastructures of TikTok and Douyin, platform governance varies across regulatory regimes (Kaye et al., 2021, 2022). This is reflected in Douyin's explicit reference to Chinese political ideologies in its terms of use (Kaye et al., 2021).

While all social media platforms must moderate content to a certain extent (Gillespie, 2018), scholars have argued that TikTok's young audience, short video format, virality-centric platform logic, and connection with China arguably make it more challenging to govern in the United States (Kaye et al., 2022; Zeng and Kaye, 2022). TikTok enforces its community guidelines through content removal and account suspension. TikTok's (2023) "Community Guidelines Enforcement Report" shows that more than 85 million videos[1] between October and December 2022, approximately 81.5% of which were removed because of minor safety (33.3%), illegal activities and regulated goods (27.4%), adult nudity and sexual activities (12.8%), and violent and graphic content (8%). It also blocked approximately 76 million accounts suspected to be under the age of 13, fake, or for other unspecified reasons.

While content removal is at the core of the current discussions of content moderation, platforms engage in other forms of moderation (Gillespie, 2022; Zeng and Kaye, 2022). TikTok's platform governance can be characterized by the notion of "visibility moderation" which refers to "the process through which digital platforms manipulate (i.e. amplify or suppress) the reach of user-generated content through algorithmic or regulatory means" (Zeng and Kaye, 2022: 81). TikTok's visibility moderation is enacted through its "For You Feed" algorithms. The algorithms not only recommend "personalized" content to users but also suppress the visibility of content deemed to be "problematic" from TikTok's perspective (Zeng and Kaye, 2022). Yet, it remains unclear the number of videos that TikTok decides *not* to recommend.

Methods and data

To analyze the evolution of TikTok's policies, we scraped all of TikTok's community guidelines ($N=25{,}641$, denoted by T_i) between 10 December 2018 and 7 March 2022 using the Wayback Machine. The web data were subsequently cleaned, and duplicates were eliminated to identify a collection of unique community guidelines. In addition, each document was tokenized to obtain a corpus of unique and separated sentences from the policy document collection. A batch of sentence tokens were then generated and prepared for human coding ($N=1429$).

Our analytical approach consisted of three steps. First, lexical analysis was conducted on the corpus of TikTok's community guidelines to analyze word- and sentence-level changes across policies, including counting the number of sentences, calculating word frequency, and evaluating textual complexity (Peslak and Conforti, 2019). This analysis provided a linguistic overview to measure and trace policy granularity and changes, respectively. Second, we coded platform values at the sentence level to understand how TikTok consider values as a noun (i.e. the object that is valuable) and verb (i.e. the process by which a value is constructed) (Kraatz et al., 2020). We also examined the path dependence of the evolution of platform values using the time-series network analysis technique (Kay, 2003, 2005). Based on the content analysis results and the characteristics of the constructed co-occurrence networks, we identified four major stages of policy evolution (see Table 1). Third, we conducted a qualitative thematic analysis to explore the key themes in relation to the changes across community guidelines.

Lexical analysis

To understand the linguistic changes in policy documents over time, we calculated the number of sentences, word frequency, and number of unique words per document after excluding punctuation, symbols, numbers, URLs, and other special characters.

Table 1. Descriptive statistics of the most distinctive value networks.

Statistics (T_i)	T_1 10 December 2018	T_2 8 January 2020	T_3 15 December 2020	T_4 7 March 2022
Network-level				
Nodes	8	7	8	8
Edges	14	16	22	24
Average degree	3.50	4.57	5.50	6.00
Average weighted degree	38	102.29	150.50	165.50
Density	0.5	0.76	0.79	0.86
Node-level degree centrality (normalized)				
Engagement	2.53	2.11	2.15	2.15
Authenticity	0.26	0.55	0.53	0.62
Community	2.21	1.45	1.67	1.51
Privacy	0.05	0.08	0.13	0.18
Safety	2.16	2.01	2.17	2.27
Accountability	0.37	0.68	1.10	0.95
Fairness	0.11	0.12	0.17	0.21
Self-determination	0.32	NA	0.07	0.11

Node-level degree centralities were normalized by the average weighted degrees of the value networks.

Following Peslak and Conforti's (2019), we further assessed the lexical characteristics of policy documents measuring textual readability, complexity, and richness. Readability was calculated using the Flesch–Kincaid readability score (Paasche-Orlow et al., 2003), which evaluates readability using word length and sentence length (i.e. a higher score indicates that the text is more difficult to read). The type-token ratio (TTR) was used to determine the textual richness using lexical diversity measures. A higher TTR ultimately implies that a greater proportion of unique words were used.

In addition, Hapax richness, which is defined as the number of words that appear only once divided by the total number of words, was employed in order to determine the richness of the text (Jockers and Thalken, 2020). The lexical richness index was calculated by returning a logical value for each term that occurs once in the document-feature matrix and then summing the rows to obtain the total. This value is then converted to a proportion of the overall word count. The three measures of lexical diversity were selected in order to explore the multidimensional characteristics of community guidelines in terms of their linguistic features.

Content analysis

We performed a content analysis of the values mentioned in TikTok's community guidelines to describe platform governance over time. As noted earlier, we coded a value when it was used as an attribute to describe platform governance and as a justification for legitimizing certain "desirable" attributes of the platform. Based on previous research on platform values such as engagement and authenticity (Hallinan et al., 2022), privacy (Greene and Shilton, 2018), accountability (Wieringa, 2020), fairness (Van Dijck et al., 2018), and self-determination (DeVito et al., 2021) as well as the authors' close reading of TikTok's community guidelines, we identified and examined eight platform values, including engagement, authenticity, community, privacy, safety, accountability, fairness, and self-determination (see the supplementary material for the detailed codebook). Despite the prevalence of "community" in community guidelines, we included community as a value (c.f. Scharlach et al., 2023) because TikTok might mobilize "community" for the valuation and devaluation of certain social groups. This allows us to analyze what counts as "community" on TikTok.

Notably, these platform values were not predetermined. During the initial coding process, we included three additional values, namely, "inclusion" (i.e. whether users feel welcome, supported, and safe; see DeVito et al., 2021), "transparency" (i.e. whether there is a mechanism for users to request information from the platform) and "self-expression" (i.e. whether the platform allows or prohibits users' posting and sharing behaviors). We excluded "inclusion" because it was closely associated with safety and community. Transparency was excluded from the final analysis because of its absence in the community guidelines. In addition, our intercoder reliability process (completed by two coders) revealed significant overlap between the "self-expression" and "engagement" codes. The researchers determined through further review that these codes can and should be interpreted as interchangeable, as all cases of the "engagement" (i.e. "encourage" or "promoting" activity) encouraged users to express themselves (i.e. "self-expression").

Specifically, we developed an operational definition for each value to explain how each of them serve as an attribute to describe platform governance in community guidelines. A detailed codebook and operational definitions of platform values are provided in the supplementary material. For example, we operationalized "engagement" as whether TikTok allows or prohibits interactivity and participation through social media (see Hallinan et al., 2022). Examples include what users can and cannot post on TikTok (e.g. harmful content). Similarly, we operationalized "privacy" as what TikTok allows or prohibits users from doing in terms of controlling personal information, such as "*do not disclose others' personally identifiable information*." The values are mutually exclusive to each other in definition; however, a sentence can have more than one value.

Two researchers were trained to analyze the data. They were supplied with the codebook and were instructed to code 10% of the sentence sample for the presence or absence of each of the values. In other words, each of the coders reviewed a randomly selected sentence from a community guidelines document and coded for each value as present (1) or absent (0). Intercoder reliability was run thereafter to assess agreement between the two coders, reaching 0.98 for determining sentence relevance and an average of 0.92 for determining platform values based on Krippendorff's α. Then, they coded the remaining sentences in the sample. Sentences that did not communicate a value were excluded from the analysis.

Network analysis. To understand how the community guidelines evolved, we developed a unique approach to investigating the networks of platform values. Platform values are not only considered independent and separate codes of conduct but also pairs of co-existing objects in establishing a common ground for value-system repertoires. Therefore, the ostensibly separate values co-exist and are mutually dependent when articulating imaginary boundaries of guidance on TikTok. The co-existing values compete against one another, such that certain values become redundant when new ones are created. The existence of such a contest among platform values not only makes certain values more meaningful to users but also makes the platform more valuable and visible to the public (Van Es and Poell, 2020).

Thus, data from the analyses of sentence tokens were transferred to co-occurrence matrices for conducting network analysis and path-dependency analysis. For statistical testing of the hypotheses and research questions, multiple regression quadratic assignment procedure (MRQAP) was implemented using R package asnipe. MRQAP can be used to determine how one independent matrix affects the dependent matrix by partially eliminating the effects of other predictors (Dekker et al., 2007). In this approach, residuals from the regression on each predictor (fixed effect) are randomized to calculate the *p* value.

Value co-occurrence networks. We constructed the matrices of each policy document to reflect the associations between the eight platform values. The weights were determined by analyzing how frequently the two platform values co-occur in the same tokenized sentence. In each value matrix, the entries represent the degree of association between the two corresponding platform values, with eight rows and eight columns. The more frequently the two platform values co-occurred across the tokenized sentences in the community guidelines, the stronger their connection.

Data analysis. We first calculated the node-level (i.e. normalized degree centrality for each node) and network-level statistics (i.e. nodes, edges, average degree, average weighted degree, and density) for each value co-occurrence matrix, as constructed above at each timepoint. Next, we compared the statistical results of all the matrices and focused on those networks that were most distinct over time. A total of 15 unique networks were identified between December 2018 and March 2022, and 4 of them have changed significantly compared to their predecessors and successors: (a) the value network on 10 December 2018 (T_1), (b) the value network on 8 January 2020 (T_2), (c) the value network on December 15, 2020 (T_3), and (d) the value network on 7 March 2022 (T_4). In the next step, path-dependency analysis was conducted using these four value co-occurrence matrices.

Matrices constructed at multiple timepoints were simultaneously entered into an MRQAP regression model (the matrices of precedent and succedent timepoints were also included in the model as controls to partial out the autoregression effects) to access the unique effect of each independent variable. For example, in order to test the effect of T_2 on T_3, we examined the effect of network T_2 on network T_3 while controlling for the effect of network T_1: in this case, the co-occurrence matrix was constructed on December 15, 2020.

Qualitative thematic analysis

We conducted a qualitative thematic analysis to identify the implications of platform values at the latent level (Boyatzis, 1998). We examined TikTok's discursive strategies—in other words, how values prescribed a certain relationship between TikTok and its users. For example, we considered TikTok's statement on authenticity by probing to what extent did it allow or prohibit users from using "truthful" communication, for whom, and to what end? We relied on this analytical approach to examine how TikTok's community guidelines discursively framed platform values.

Findings

We now analyze the evolution of platform values manifested in TikTok's community guidelines at the lexical and discursive levels. We first present the lexical analysis of TikTok's community guidelines. Second, we show that the co-occurrence networks of platform values have been reshaped over time and nuanced associations have been observed among the constructed matrices through network analysis. Third, based on the qualitative thematic analysis, we describe how TikTok has increasingly promoted the values of authenticity and accountability as well as framed visibility moderation (Zeng and Kaye, 2022) in its community guidelines.

Lexical characteristics of TikTok's community guidelines

The results indicate that the number of sentences per text has increased since the end of 2018, with the two most significant increases observed on January 8, 2020 (N=99, a 33% increase compared to the nearest precedent version) and on December 15, 2020 (N=136, a 29% increase compared to the nearest precedent version). There was a substantial increase in the number of words per document and the number of unique words per document on both the aforementioned dates. Taking the example of January 8, 2020, the number of words in the community guideline was increased from 618 to 2001 words, and the number of unique words was almost doubled (from 344 to 847 words) compared to the previous version. A similar trend was found in the community guidelines updated on 15 December 2020 and 7 March 2022. Figure 1 illustrates the trend of lexical compositions in TikTok's community guidelines.

Figure 2 illustrates how lexical diversity has changed over time in TikTok's community guidelines. Accordingly, lexical readability—as measured by the Flesch-Kincaid score—has increased over time, with one significant increase on 8 January 2020 and becoming stable thereafter. Throughout the entire period, there was a downward trend in lexical complexity (measured by the TTR) and richness (measured by the Hapax proportion), both of which decreased over time at a similar rate of continual decline in each case. For each of the three lexical measures, the variances of change are most pronounced for readability (M=16.89, SD=4.07), followed by lexical complexity (M=0.45, SD=0.09) and lexical richness (M=0.15, SD=0.05).

To investigate how TikTok's platform values changed with different iterations of their community guidelines, we plotted the number of occurrences of each value per day

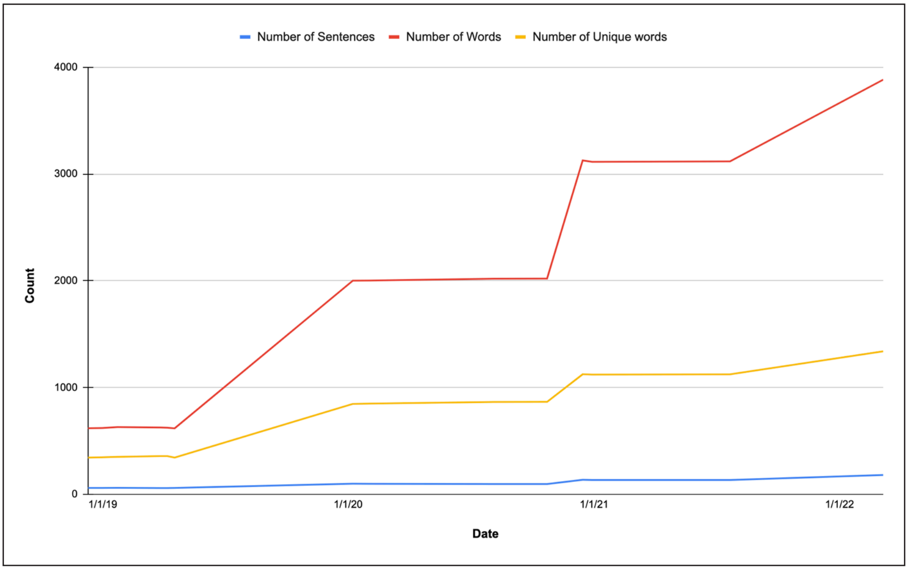

Figure 1. Lexical overview of TikTok's community guidelines.

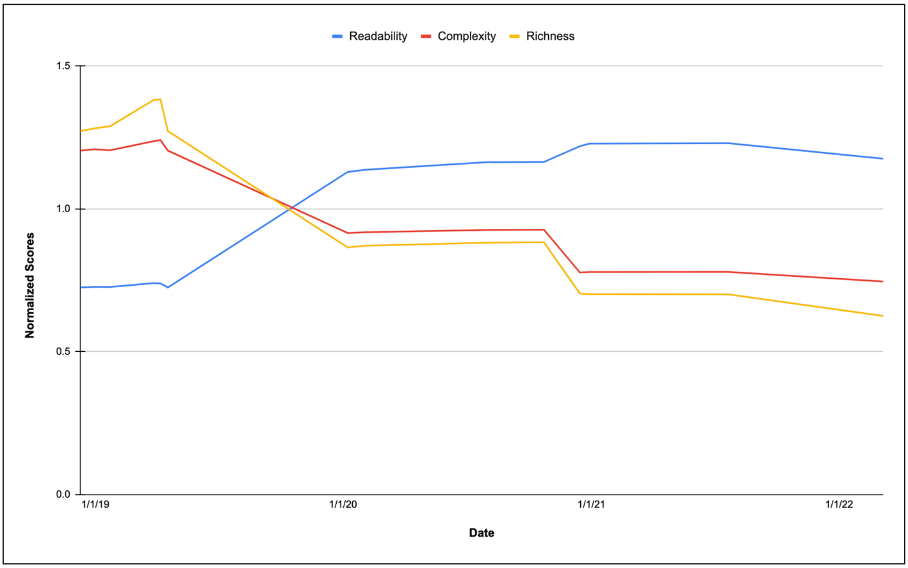

Figure 2. Lexical diversity of TikTok's community guidelines.

between December 2018 and March 2022 in Figure 3. Compared to other values in community guidelines, engagement and safety were the most coded values, with similar trends and increases from the end of 2019 to the beginning of 2020, as well as another significant increase at the end of 2020 (N=89 and 87, respectively). Similarly, the

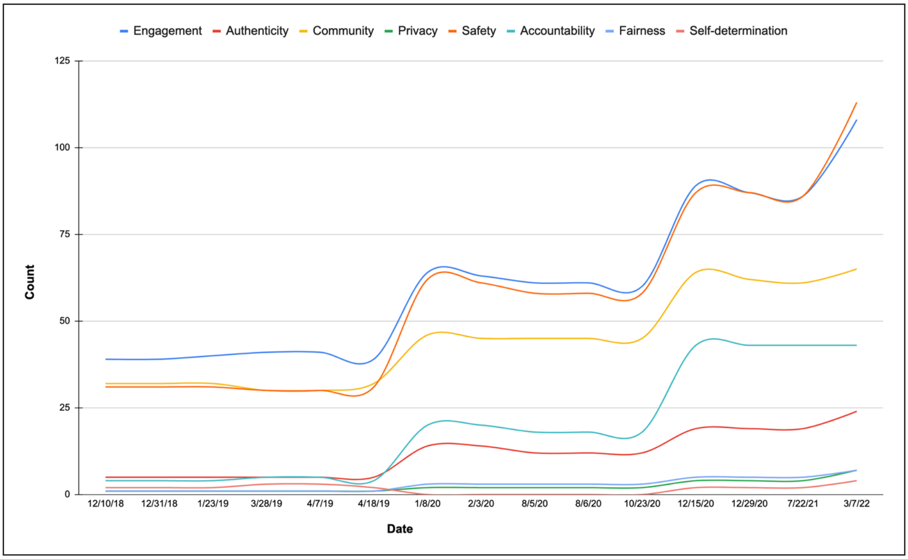

Figure 3. The evolution of platform values of TikTok's community guidelines over time.

frequency of platform values in community guidelines gradually increased over the same period of time. However, authenticity and accountability revealed a different pattern. While they were seldom found in 2018 and 2019 ($N \leqslant 5$), both values significantly increased in 2020. A relatively small number of occurrences were also found for privacy, fairness, and self-determination. Furthermore, there were no observations of self-determination during the year 2020.

Evolution of TikTok's value networks in its community guidelines

Based on lexical diversity and the characteristics of value networks, we identified four distinct versions of community guidelines for network analysis. Table 1 presents the descriptive statistics of the value networks at the following four timepoints: (1) T_1 on 10 December 2018; (2) T_2 on 8 January 2020; (3) T_3 on 15 December 2020; and (4) T_4 on 7 March 2022. At the network level, three of the four networks reported eight nodes, with the exception of T_2, which was missing the node of self-determination. Throughout the time period, the number of edges, the average weighted degrees, and the density scores have all increased, thereby indicating that the platform value networks have become more condensed and interconnected. As depicted in Figure 4, we visualized the interconnections among the platform values to better observe the evolution of value co-occurrence networks over discrete periods. Generally, value networks have become more interconnected and more closely correlated over time.

Table 1 presents the normalized degree centrality of each node for each value network. Compared to values such as privacy and fairness, the values of engagement, community, and safety tended to be more central in the networks, thereby suggesting a larger

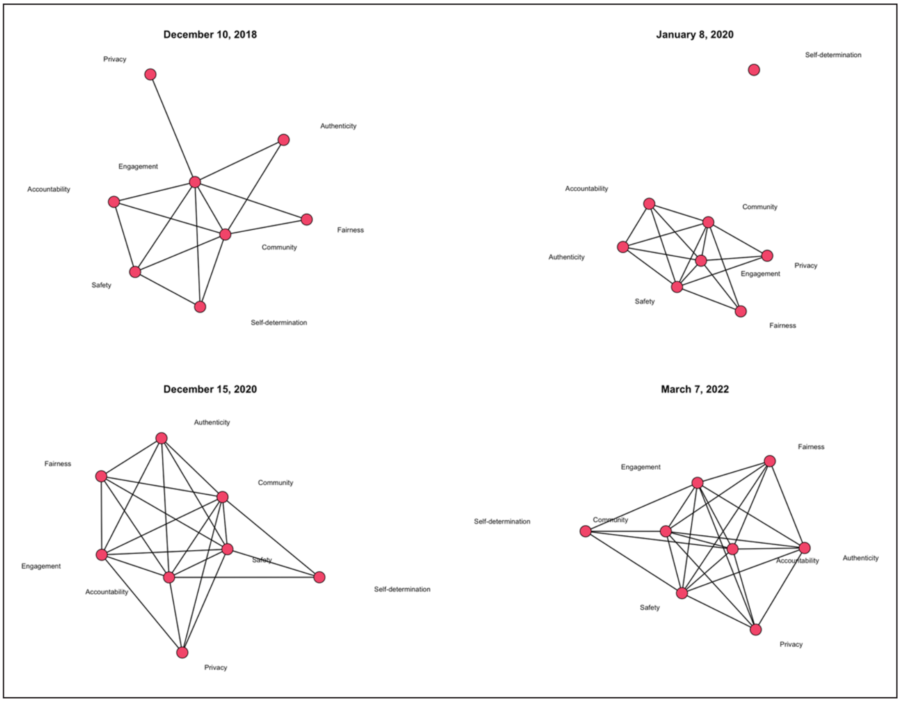

Figure 4. The evolution of value co-occurrence networks across four discrete timepoints. Graph features correspond to descriptive statistics in Table 1.

number of connections with other nodes. Nevertheless, there was a slight decline in centrality at the nodes of engagement, community, and self-determination, whereas increases in centrality were observed at the nodes of authenticity, privacy, accountability, and fairness.

In terms of the path-dependency among the four distinct value networks, the value network of T_2 was significantly correlated with the value network of T_1 (β=0.95, $p < 0.001$), as presented in Table 2 (Model 1). A similar pattern of path was observed in the value networks of T_3 and T_4 when precedent networks were used to predict the succedent networks (see Table 2 (Models 2a–2b) and Table 3 (Models 3a–3c). However, as indicated in Table 2, Model 2c, the impact of network T_1 (β=−0.14, $p > 0.05$) on the value network of T_3 was eliminated when the value network of T_2 was controlled, thereby indicating that the impact of T_1 on T_3 was mediated entirely by the value network of T_2 (β=1.12, $p < 0.001$).

To further examine the sophisticated influence of precedent networks on succedent networks, we used network T_4 as the dependent variable and the other three precedent networks (T_1—T_3) as independent variables in a series of MRQAP regressions. Different sets of IVs were entered into the regression models. Table 3 (Model 3) summarizes the results; in the table, each precedent network has been used to predict network T_4 and has then been paired with one of the other two precedent networks to predict network T_4.

Table 2. Results of multiple quadratic assignment procedure analysis on value networks T_2 and T_3.

Value co-occurrence networks T_i	Value co-occurrence networks T_j	
	β	Adjusted R^2
Model 1	Value co-occurrence network T_2	
Value co-occurrence network T_1	0.95***	0.78**
Model 2	Value co-occurrence network T_3	
Model 2a		
Value co-occurrence network T_1	0.92***	0.64***
Model 2b		
Value co-occurrence network T_2	0.98***	0.76***
Model 2c		
Value co-occurrence network T_1	–0.14	0.76***
Value co-occurrence network T_2	1.12***	

Standardized coefficients were presented.
$^{**}p < .01$. $^{***}p < .001$.

First, each of the precedent networks showed a positive and significant correlation with network T_4 ($\beta = 0.91$, 0.98, 0.99, respectively; $p < 0.001$). Compared to the two other networks, network T_1 explained the least variance (63%), but still played a significant role in explaining network T_4. However, as depicted in Table 3 (Model 3d), the impact of network T_1 on T_4 was reversed, thereby negatively and significantly correlating with network T_4 ($\beta = -0.27$, $p < 0.01$), when only network T_2 was controlled.

Similar to Model 2c, a full mediation effect was observed when network T_1 ($\beta = 0.03$, $p > 0.05$) and network T_3 ($\beta = 0.96$, $p < 0.001$) were used simultaneously to predict network T_4 (see Table 3, Model 3e). Furthermore, both effects of network T_2 and network T_3 on the formation of network T_4 were evidently decreased in Model 3f, when T_2 ($\beta = 0.45$, $p < 0.001$) and T_3 ($\beta = 0.54$, $p < 0.001$) were used simultaneously to predict network T_4. Finally, we predicted network T_4 using all three precedent networks. The results are mixed. As depicted in Table 3 (Model 4), network T_1 ($\beta = -0.020$, $p < 0.001$) was negatively related to the dependent network T_4, whereas network T_2 ($\beta = 0.72$, $p < 0.01$) and network T_3 ($\beta = 0.46$, $p < 0.01$) were positively related to network T_4.

The making and remaking of TikTok's rulebooks

While the lexical and network analyses illustrated how the linguistic characteristics of TikTok's community guidelines changed and how the value networks correlated with one another, the qualitative analysis identified two emergent themes. First, while the quantitative analysis indicates that the values of authenticity and accountability significantly increased in 2020 compared to other values, the qualitative analysis illustrates the meanings of authenticity and accountability. Second, the analysis reveals a discursive shift from content removal to visibility moderation.

Table 3. Results of multiple quadratic assignment procedure analysis on value network T_4.

Value co-occurrence networks T_i	Value co-occurrence networks T_j	
	β	Adjusted R^2
Model 3	Value co-occurrence network T_4	
Model 3a		
Value co-occurrence network T_1	0.91***	0.63***
Model 3b		
Value co-occurrence network T_2	0.98***	0.78***
Model 3c		
Value co-occurrence network T_3	0.99***	0.78***
Model 3d		
Value co-occurrence network T_1	-0.27**	0.79***
Value co-occurrence network T_2	1.24***	
Model 3e		
Value co-occurrence network T_1	0.03	0.78***
Value co-occurrence network T_3	0.96***	
Model 3f		
Value co-occurrence network T_2	0.45***	0.78***
Value co-occurrence network T_3	0.54***	
Model 4		
Value co-occurrence network T_1	–0.20**	0.79***
Value co-occurrence network T_2	0.72**	
Value co-occurrence network T_3	0.46**	

Standardized coefficients were presented.
**$p < .01$. ***$p < .001$.

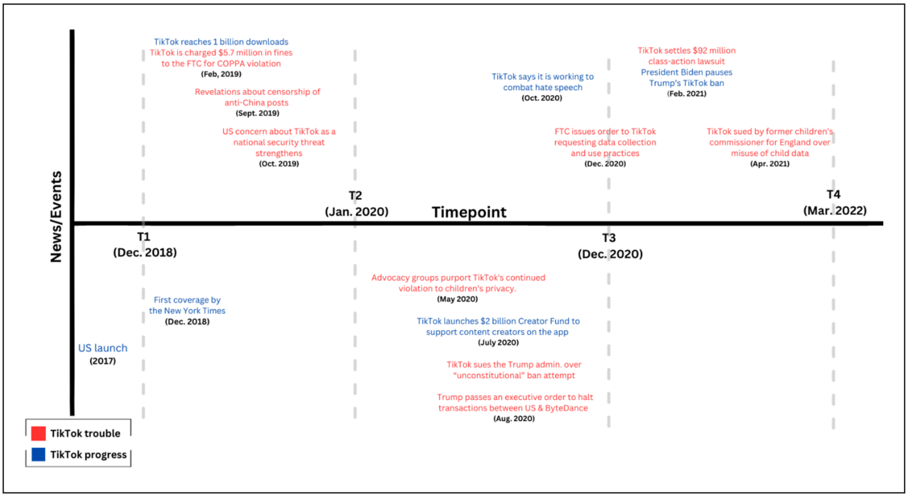
News/Events
Timepoint
T1
(Dec. 2018)
T2
(Jan. 2020)
T3
(Dec. 2020)
T4
(Mar. 2022)
TikTok reaches 1 billion downloads
TikTok is charged $5.7 million in fines to the FTC for COPPA violation
(Feb, 2019)
Revelations about censorship of anti-China posts
(Sept. 2019)
US concern about TikTok as a national security threat strengthens
(Oct. 2019)
TikTok says it is working to combat hate speech
(Oct. 2020)
FTC issues order to TikTok requesting data collection and use practices
(Dec. 2020)
TikTok settles $92 million class-action lawsuit
President Biden pauses Trump's TikTok ban
(Feb. 2021)
TikTok sued by former children's commissioner for England over misuse of child data
(Apr. 2021)
US launch
(2017)
First coverage by the New York Times
(Dec. 2018)
Advocacy groups purport TikTok's continued violation to children's privacy.
(May 2020)
TikTok launches $2 billion Creator Fund to support content creators on the app
(July 2020)
TikTok sues the Trump admin. over "unconstitutional" ban attempt
Trump passes an executive order to halt transactions between US & ByteDance
(Aug. 2020)
🟥 TikTok trouble
🟦 TikTok progress

Figure 5. Noteworthy TikTok news corresponding to community guidelines timepoints.
Authors' summary of TikTok-related news event.

Authenticity is a prevalent value in the corporate discourses pertaining to social media (Hallinan et al., 2022). TikTok (2021) promoted that "authenticity and joy are unique and ownable aspects of the TikTok community, not only for users but also for brands and businesses looking to make an impact." While the narrative appears to proclaim that the value of authenticity is rooted in TikTok's community, it did not exist in the platform's community guidelines *before* T_2. Indeed, TikTok's mission—which aims to "inspire creativity and bring joy" in a global community—has remained largely the same in the distinct versions of its community guidelines. Nonetheless, in T_2, TikTok's global community was no longer just about "fun" and social connection but added that it allows users to "create and share *authentically*" (emphasis added). TikTok's value of authenticity was primarily an informational one (Hallinan et al., 2022), which targeted "deceptive content and accounts" on the platform. While the T_1 version had discouraged spamming, fake engagement, and misleading content, the T_2 version grouped all of these behaviors into a new section called "Integrity and authenticity." This value was stabilized in T_3 and T_4, as exemplified by the following statement "At TikTok, we prioritize safety, diversity, inclusion, and authenticity."

With the advent of time, TikTok's community guidelines emphasized accountability. Accountability can be characterized by self-governance (Gorwa, 2019). It was used to describe how TikTok actively removed problematic content, banned accounts that repeatedly violated the rules, and reported to relevant legal authorities. The realization of accountability requires not only the platform's proactive algorithmic and human moderation (being included since T_3) but also users' efforts to use the reporting tools. Such reporting tools can enable platforms to strategically legitimize the regulation of content (Gillespie, 2018).

The second theme concerns TikTok's framing of content regulation. Consistent with the findings of Zeng and Kaye (2022), we observed that TikTok conveyed the manipulation of content visibility, or what it termed "discoverability," of inappropriate content in T_3 and T_4. Leaked reports already discussed how TikTok marked certain content as "visible to self" in September 2019 (Hern, 2019). However, TikTok explicitly mentioned visibility moderation only until T_3. A common means of framing content regulation is reflected in this statement: "We remove content, including video, audio, image, and text that violates our Community Guidelines, and suspend or ban accounts involved in severe or repeated violations." Since T_3, TikTok has begun re-purposing this as "For You Feed"—an algorithmic system that supposedly recommends "personalized" content to users based on user interactions, video information (e.g. hashtags), and device and account settings for visibility moderation. While the T_3 version only included one statement to note how the platform could redirect search results or suppress the discoverability of content like spam, the T_4 version added a new section called "Ineligible for the For You Feed." This expanded section stated that a variety of content might remain searchable and viewable on the site but be ineligible for algorithmic recommendation. Indeed, TikTok has long regulated much of the ineligible content related to safety of minors, sexualized depictions, violence, tobacco and alcohol products, and misleading content. Nevertheless, TikTok now also moderates the visibility of "unoriginal, low-quality, and QR code content," including content from other platforms, legacy media, and "extremely short clips, static images, and exclusively-GIF based videos." The broad yet ambiguous scope of such content could potentially exert control over content creators and users through "the threat of invisibility" (Zeng and Kaye, 2022: 81).

Discussion

This study examined the evolution of TikTok and its governance frameworks at the lexical and discursive levels. Through our lexical analysis, we identified four distinct versions of community guidelines for further examining the construction, maintenance, and reconstruction of platform values between 2018 and 2022. Overall, TikTok's community guidelines have become more readable and less complex despite containing longer sentences and more complex vocabulary. Among the eight identified values, engagement, community, and safety were the most important values in terms of their frequency, followed by accountability and authenticity. These three values occupied central positions in the four identified value networks because community guidelines are concerned with user engagement, and TikTok has increasingly branded itself as a safe community for users, particularly minors. It is noteworthy, however, that values are often used to justify both what users *can* and *cannot* do (Scharlach et al. et al., 2023). For instance, in order for TikTok to "promote safety," it often claims to prohibit certain forms of "harmful," "dangerous," and "illegal" engagement on the site. By contrast, privacy, fairness, and self-determination were the least important values in TikTok's community guidelines. One reason is that privacy is more likely to be invoked in privacy policies than community guidelines. Furthermore, our operationalization of self-determination focuses on users' ability to make decisions about TikTok's technical structure (DeVito et al., 2021). Echoing Scharlach et al. (2023), our findings indicate that some public and user-centric values were largely absent in platform policies.

Importantly, we found that the occurrences of authenticity and accountability significantly increased in 2020, compared to the other values. In TikTok's community guidelines, being authentic was largely confined to avoid posting and sharing "harmful misinformation" that might misinform public opinion. In this vein, TikTok's appeals for authenticity might be better understood as a strategic discursive performance in response to the growing public concern over misinformation (Ng, 2020). Notably, the growth of accountability must be interpreted with caution. We used a binary quantitative measure of whether there exists a mechanism for the platform or users to hold the platform accountable. While the ideal of accountability suggests that it is obligatory for the platform to explain and justify its action to users and other stakeholders at various stages (Wieringa, 2020), our qualitative analysis reveals that TikTok rather pointed to the appeal to self-governance. Simply put, every updated version of the guidelines was accompanied by an expanded list of restrictions that TikTok would regulate to protect users. For example, while the T_1 version had briefly stated types of misbehavior, the T_2 version grouped and defined 10 categories of misbehaviors (e.g. minor safety, illegal activities and regulated goods, threats to platform security). Taken together, TikTok's community guidelines might selectively reveal and obfuscate platform values in its governance frameworks.

Empirically, we identified three mechanisms by which TikTok's community guidelines evolve over time using path-dependency analysis: (1) the mediation path, (2) the reversion path, and (3) the confounding path. The first mechanism of the mediation path-dependency is specifically capable of indirectly implementing previous policies into a new version of the policy through value transference, thereby significantly decreasing public invisibility when dealing with sensitive matters (e.g. privacy and safety). In other words, rather than making substantial changes that are distinctive from the last update, the platform can implicitly incorporate the intended changes within an accountable period of time instead of a subsequent update. As an example based on the mediation path, TikTok did not immediately update its community guidelines to respond to the fines by the Federal Trade Commission for violating the Children's Online Privacy Protection Act (COPPA) in February 2019 (Matsakis, 2019). Instead, values that TikTok would want to prioritize following this news (e.g. privacy and safety) gradually and quietly prevailed on its community guidelines nearly 2 years later (i.e. December 2020). Second, after several "hops" in the platform's policy changes, the values can be completely reversed. For example, our results found that after three updates, the value co-occurrence network within community guidelines had been subverted. While such a reversion path may take a considerable amount of time to achieve, it is more effective and imperceptible than the mediation path. For instance, the evolution of *privacy* may be dictated by its interconnections with other values at least partially, due to the reversion path. In the T_1 version, privacy was fairly isolated from other values, as it was only associated with engagement. As part of the T_2 version, it began to develop a strong connection with the values community and safety; however, in the T_3 version, it exhibited a strong connection with accountability as well. Consequently, these co-associations positioned *privacy* as one of the most prominent clusters in the T_4 version, along with safety and accountability, but became distanced from engagement, which was prevalent at the outset. Finally, the construction–maintenance–reconstruction mechanism, as depicted through mediation and reversion in path-dependency analysis, also illustrates a confounding route in the evolution of platform values. Within TikTok, policymakers and stakeholders are arguably complicating, institutionalizing, and confounding the production process of value selection by determining the visibility and invisibility, the core and peripheral, as well as the prominence and obscurity of platform values. Such selective production in the community guidelines and platform policies could lead to selective exposure and perception of users through the institutionally crafted dominant platform imaginary (Van Es and Poell, 2020).

A key question remains unresolved: What explains the dynamic evolution of TikTok's community guidelines? We argue that TikTok might adopt a reactive approach that strategically re-articulated their values in response to public outcry over the platform in the United States. This argument builds upon Barrett and Kreiss's (2019) concept of platform transience which suggests that platforms' rapid changes might be attributed to external normative pressures. Nonetheless, we acknowledge that it would be difficult to fully address this question without knowing the *intentionality* of the platform. Figure 5 identifies the noteworthy TikTok-related news in the United States corresponding to its updates on its community guidelines.

Specifically, TikTok included an expanded discussion of political misinformation in T_2 after news reports emerged about TikTok's ban on content that was deemed sensitive to the Chinese government (i.e. revelations about censorship of anti-China posts). T_2 was also the time that TikTok significantly emphasized the value of informational authenticity, tackling the issue of misinformation. In addition, TikTok mentioned the prohibition of "hate speech" including "attacks on protected groups," "slurs," and "hateful ideology" in T_2. Yet, as TikTok proclaimed it is working to combat hate speech in October 2020 following reports on its "White supremacy problem" (Fung, 2020), the T_3 version added "claims of supremacy over a group of people with reference to other protected attributes" and "conspiracy theories used to justify hateful ideologies" as kinds of "hateful ideology" as well as a new subsection about "organized hate," TikTok, accordingly, would remove such content to "protect the community." This does not mean that the revised community guidelines systematically detailed how they would deal with these perceived social problems; in most cases, TikTok simply added a few sentences to prohibit users from posting content related to specific misbehaviors to preserve platform values such as safety and community. In this vein, TikTok strategically adapts its community guidelines to perform ideals that suit different political and social needs.

Indeed, one may reasonably question whether platform values are mere "buzzwords" in parallel to the practice of ethics washing. However, buzzwords are performative in the sense that they can attract different actors, determine agendas, and enable the construction of numerous unstable collectives in a moment (Vincent, 2014). Echoing Gillett et al.'s (2022) analysis of the discursive construction of safety and harm in social media platforms' newsroom posts, TikTok's community guidelines have created a social world where certain users threaten platform values, especially of community, safety, and engagement.

TikTok has framed algorithmic and human moderation as solutions because they could help to filter much of the "potentially violative content" before users report the content to them. Since T_3, visibility moderation has been framed as one of the solutions that police content that is not necessarily removed yet becomes ineligible for algorithmic recommendation. Although the guidelines have indicated that individuals can dispute TikTok's moderation decisions, they do not detail the procedures for doing this. There are two key insights. First, users are framed to be proactive content gatekeepers (Konikoff, 2021) who are responsible for sustaining platform values. Users are instructed to avoid engaging in the ever-expanding list of restrictions and report potentially violative content to TikTok. Relatedly, those who violate community guidelines become the ones who threaten platform values. Second, TikTok's self-governance is legitimized by appealing to a variety of techniques that help to police a few bad actors and violative content.

Conclusion

Using a mixed-methods approach, this study explores the manifestation of platform values in TikTok's community guidelines in the United States between 2018 and 2022. There are, however, limitations of the study. First, the study is limited in its generalizability because we only analyzed TikTok's US version of community guidelines. A comparative study of how a platform's community guidelines vary across countries would enrich the current findings by examining how the platform adapts to distinct regulatory frameworks in different countries on a discursive level. Second, as we only focused on TikTok's community guidelines, the identified mechanisms may not be generalized to other types of platform policies across platforms. Third, we primarily focused on TikTok's construction of platform values rather than the enforcement of community guidelines. Fourth, while community guidelines represent the platform-initiated values, multiple user groups can both sustain and disrupt platform values. Future research should examine how different groups of users negotiate platform values in the evolving context of platform governance.

While acknowledging the limitations of the study, this research contributes to understanding the temporality and discursive performativity of platform governance. The TikTok of 2022 differs from the TikTok of 2018 and will continue to evolve rapidly. In fact, in March 2023, TikTok announced a revamp to its community guidelines 2 days before TikTok CEO Shou Zi Chew appeared before US Congress. The purpose of this revamp was "to help people understand our decisions about how we work to keep TikTok safe and build trust in our approach" (Bailliencourt, 2023, para 2). The overhauled community guidelines, effective from April 2023, introduce new policies regarding AI-generated or modified content, "tribe" as a protected attribute in their hate speech and hateful behavior policies, and more information on how the platform safeguards civil and election integrity (Bailliencourt, 2023). While the community guidelines and platform values highlighted in this article are likely to change, analyzing the evolution of community guidelines through platform values (Hallinan et al., 2022; Scharlach et al., 2023) helps us to question how certain values (e.g. community, safety, and engagement) have been selectively mobilized in a platform's governance framework to create a social reality in which certain content must be prohibited, thereby rendering self-governance (DeNardis and Hackl, 2015; Gillespie, 2018).

The meanings behind platform values are both malleable and contingent upon external normative pressures. However, the constant (minor) revision of community guidelines often creates the illusion of neutrality in platform governance. In our case study, the rhetoric of community guidelines hints at the ways that platform governance (e.g. content moderation) is a preferred and impartial solution to uphold a specific set of platform values. Nevertheless, these values are not inherently stable; they are socially constructed and can be subject to politicization. Consequently, the overhaul of community guidelines reveals the adaptive and performative nature of platform governance. Indeed, as Caplan (2023) writes about network governance, there is a need to examine how platforms communicate their relational power and delegate the responsibility to other actors through public statements. By continuously shifting platform values and selectively assigning user responsibility in community guidelines, platforms attempt to enroll various relevant social groups in a network of social relations within which platforms are in control. Investigating the evolution of community guidelines, therefore, invites us to understand and problematize the situated discursive constructions of platforms, their ideals about governance, and the performative and relational nature of platform governance.

Acknowledgements

The authors wish to thank research assistants Wanjiang Jacob Zhang and Lyon You for their work and the anonymous reviewers for their productive recommendations.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by The Chinese University of Hong Kong's Improvement on Competitiveness in Hiring New Faculties Funding Scheme (Project code: 4937060) and Boston University's East Asia Studies Career Development Professorship.

ORCID iDs

Ngai Keung Chan https://orcid.org/0000-0002-5848-3098
Alexis Shore https://orcid.org/0000-0002-8085-2355

Supplemental material

Supplemental material for this article is available online.

Note

1. Note that TikTok (2023) does not provide the country breakdown of these statistics. It, however, provides the removal volume and rates in 50 markets that represent roughly 90% of overall removal volume. By December 2022, TikTok removed 13,052,932 videos in the United States.

References

Arriagada A and Ibanez F (2020) "You need at least one picture daily, if not, you're dead": content creators and platform evolution in the social media ecology. *Social Media + Society*. Epub ahead of print 18 August. DOI: 10.1177/2056305120944624.

Badillo-Urquiola K, Smriti D, McNally B, et al. (2019) Stranger danger! social media app features co-designed with children to keep them safe online. In: *Proceedings of the th18 ACM international conference on interaction design and children*, Boise, ID, 12 June, pp. 394–406. New York: ACM.

Bailliencourt JD (2023) Helping creators understand our rules with refreshed community guidelines. *TikTok Newsroom*, 21 March. Available at: https://newsroom.tiktok.com/en-us/community-guidelines-update (accessed 4 July 2023).

Barrett B and Kreiss D (2019) Platform transience: changes in Facebook's policies, procedures, and affordances in global electoral politics. *Internet Policy Review* 8(4): 1–22.

Bossetta M (2020) Scandalous design: how social media platforms' responses to scandal impacts campaigns and elections. *Social Media + Society*. Epub ahead of print 17 June. DOI: 10.1177/2056305120924777.

Boyatzis RE (1998) *Transforming Qualitative Information; Thematic Analysis and Code Development*. Thousand Oaks; CA: Sage.

Bucher T (2021) *Facebook*. Cambridge: Polity Press.

Burgess J and Baym NK (2020) *Twitter: A Biography*. New York: New York University Press.

Caplan R (2018) *Content or Context Moderation? Artisanal, Community-Reliant, and Industrial Approaches*. New York: Data and Society Institute.

Caplan R (2023) Networked platform governance: the construction of the democratic platform. *International Journal of Communication* 17: 3451–3472.

DeCook JR, Cotter K, Kanthawaia S, et al. (2022) Safe from "harm": the governance of violence by platforms. *Policy & Internet* 14(1): 63–78.

Dekker D, Krackhardt D and Snijders TAB (2007) Sensitivity of MRQAP tests to collinearity and autocorrelation conditions. *Psychometrika* 72(4): 563–581.

DeNardis L and Hackl AM (2015) Internet governance by social media platforms. *Telecommunications Policy* 39(9): 761–770.

DeVito MA, Walker AM and Fernandez JR (2021) Values (mis)alignment: exploring tensions between platform and LGBTQ+ community design values. *Proceedings of the ACM on Human-Computer Interaction* 5: 1–27.

Fung B (2020) Even TikTok has a white supremacy problem. *CNN*, 14 August. Available at: https://edition.cnn.com/2020/08/14/tech/tiktok-white-supremacists/index.html (accessed 4 July 2023).

Gerrard Y and Thornham H (2020) Content moderation: social media's sexist assemblages. *New Media & Society* 22(7): 1266–1286.

Gillespie T (2010) The politics of "platforms." *New Media & Society* 12(3): 347–364.

Gillespie T (2018) *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. New Haven, CT: Yale University Press.

Gillespie T (2022) Do not recommend? Reduction as a form of content moderation. *Social Media + Society*. Epub ahead of print 19 August. DOI: 10.1177/20563051221117552.

Gillett R, Stardust Z and Burgess J (2022) Safety for whom? Investigating how platforms frame and perform safety and harm interventions. *Social Media + Society*. Epub ahead of print 15 December. DOI: 10.1177/20563051221144315.

Gorwa R (2019) What is platform governance? *Information, Communication & Society* 22(6): 854–871.

Greene D and Shilton K (2018) Platform privacies: governance, collaboration, and the different meanings of "privacy" in iOS and Android development. *New Media & Society* 20(4): 1640–1657.

Hallinan B, Scharlach R and Shifman L (2022) Beyond neutrality: conceptualizing platform values. *Communication Theory* 32(2): 201–222.

Heinich N (2020) Ten proposals on values. *Cultural Sociology* 14(3): 213–232.

Helmond A (2015) The platformization of the web: making web data platform ready. *Social Media + Society*. Epub ahead of print 30 September. DOI: 10.1177/2056305115603080.

Helmond A and Van der Vlist FN (2019) Social media and platform historiography: challenges and opportunities. *TMG Journal for Media History* 22(1): 6–34.

Helmond A, Nieborg DB and Van der Vlist FN (2019) Facebook's evolution: development of a platform-as-infrastructure. *Internet Histories* 3(2): 123–146.

Hern A (2019) Revealed: how TikTok censors videos that do not please Beijing. *The Guardian*, 25 September. Available at: https://www.theguardian.com/technology/2019/sep/25/revealed-how-tiktok-censors-videos-that-do-not-please-beijing (accessed 4 July 2023).

Hoffmann AL, Proferes N and Zimmer M (2018) "Making the world more open and connected": Mark Zuckerberg and the discursive construction of Facebook and its users. *New Media & Society* 20(1): 199–218.

Jiang J, Middler S, Brubaker JR, et al. (2020) Characterizing community guidelines on social media platforms. In: *Conference companion publication of the 2020 on computer supported cooperative work and social computing, Virtual Event*, 17 October, pp. 287–291. New York: ACM.

Jockers ML and Thalken R (2020) Hapax richness. In: Jockers ML and Thalken R (eds) *Text Analysis with R*. Cham: Springer, pp. 93–97.

Katzenbach C, Magalhães JC, Kopps A, et al. (2021) *The Platform Governance Archive*. Alexander von Humboldt Institute for Internet and Society. Available at: https://doi.org/10.17605/OSF.IO/XSBPT (accessed 4 July 2023).

Kay A (2003) Path dependency and the CAP. *Journal of European Public Policy* 10(3): 405–420.

Kay A (2005) A critique of the use of path dependency in policy studies. *Public Administration* 83(3): 553–571.

Kaye DBV, Chen X and Zeng J (2021) The co-evolution of two Chinese mobile short video apps: parallel platformization of Douyin and TikTok. *Mobile Media & Communication* 9(2): 229–253.

Kaye DBV, Zeng J and Wikstrom P (2022) *TikTok: Creativity and Culture in Short Video*. Cambridge: Polity Press.

Kline R and Pinch T (1996) Users as agents of technological change: the social construction of the automobile in the rural United States. *Technology and Culture* 37(4): 763–795.

Konikoff D (2021) Gatekeepers of toxicity: reconceptualizing Twitter's abuse and hate speech policies. *Policy & Internet* 13(4): 502–521.

Kraatz MS, Flores R and Chandler D (2020) The value of values for institutional analysis. *Academy of Management Annals* 14(2): 474–512.

Maddox J and Malson J (2020) Guidelines without lines, communities without borders: the marketplace of ideas and digital manifest destiny in social media platform policies. *Social Media + Society*. Epub ahead of print 19 June. DOI: 10.1177/2056305120926622.

Mahendran L and Alsherif N (2020) Adding clarity to our community guidelines. *TikTok Newsroom*, 8 June. Available at: https://newsroom.tiktok.com/en-us/adding-clarity-to-our-community-guidelines (accessed 4 July 2023).

Matsakis L (2019) FTC hits TikTok with record $5.7 million fine over children's privacy. *WIRED*, 27 February. Available at: https://www.wired.com/story/tiktok-ftc-record-fine-childrens-privacy/ (accessed 4 July 2023).

Myers West S (2017) Raging against the machine: network gatekeeping and collective action on social media platforms. *Media and Communication* 5(3): 28–36.

Ng A (2020) US officials in contact with TikTok over political disinformation. *CNET*, 3 March. Available at: https://www.cnet.com/news/politics/us-officials-in-contact-with-tiktok-over-political-disinformation/ (accessed 4 July 2023).

Paasche-Orlow M, Taylor HA and Brancati FL (2003) Readability standards for informed-consent forms as compared with actual readability. *The New England Journal of Medicine* 348: 721–726.

Perez S (2021) TikTok to rank as the third largest social network, 2022 forecast notes. *TechCrunch*, 21 December. Available at: https://techcrunch.com/2021/12/20/tiktok-to-rank-as-the-third-largest-social-network-2022-forecast-notes/ (accessed 4 July 2023).

Peslak A and Conforti M (2019) A longitudinal study of Facebook privacy policies. *Issues in Information Systems* 20(1): 213–223.

Poell T, Nieborg DB and Duffy BE (2022) *Platforms and Cultural Production*. Cambridge: Polity Press.

Ruberg B (2021) "Obscene, pornographic, or otherwise objectionable": biased definitions of sexual content in video game live streaming. *New Media & Society* 23(6): 1681–1699.

Savic M (2021) From Musical.ly to TikTok: social construction of 2020's most downloaded short video app. *International Journal of Communication* 15: 3173–3194.

Scharlach R, Hallinan B and Shifman L (2023) Governing principles: articulating values in social media platform policies. *New Media & Society*. Epub ahead of print 7 March. DOI: 10.1177/14614448231156580.

Shutsko A (2020) User-generated short video content in social media: a case study of TikTok. In: *International conference on human-computer interaction*, Copenhagen, 19–24 July, pp. 108–125. Cham: Springer.

Singer N (2020) TikTok broke privacy promises, children's groups say. *The New York Times*, 14 May. Available at: https://www.nytimes.com/2020/05/14/technology/tiktok-kids-privacy.html (accessed 4 July 2023).

Srnicek N (2017) *Platform Capitalism*. Cambridge: Polity Press.

Stein L (2013) Policy and participation on social media: the case of YouTube, Facebook, and Wikipedia. *Communication, Culture & Critique* 6(3): 353–371.

TikTok (2021) *Nielsen Study Shows TikTok Ideal Place for "Discovery," Content More "Authentic."* TikTok for Business. Available at: https://www.tiktok.com/business/en/blog/nielsen-study-tiktok-discovery-content-authentic (accessed 4 July 2023).

TikTok (2023). Community guidelines enforcement report, 31 March. Available at: https://www.tiktok.com/transparency/en-us/community-guidelines-enforcement-2022-4/ (accessed 4 July 2023).

Van Dijck J, Poell T and De Waal M (2018) *The Platform Society: Public Values in a Connective World*. New York: Oxford University Press.

Van Es K and Poell T (2020) Platform imaginaries and Dutch public service media. *Social Media + Society*. Epub ahead of print 23 June. DOI: 10.1177/2056305120933289.

Vincent BB (2014) The politics of buzzwords at the interface of technoscience, market and society: the case of "public engagement in science." *Public Understanding of Science* 23(3): 238–253.

Wang C-H, Sher ST-H, Salman I, et al. (2022) "Tiktok made me do it": teenagers' perception and use of food content on TikTok. In: *Proceedings of the 21st annual ACM interaction design and children conference*, Braga, 27 June, pp. 458–463. New York: ACM.

Wieringa M (2020) What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability. In: *Proceedings of the 2020 conference on fairness, accountability, and transparency*, New York, 27 January, pp. 1–18. New York: ACM.

Zeng J and Kaye DBV (2022) From content moderation to visibility moderation: a case study of platform governance on TikTok. *Policy & Internet* 14(1): 79–95.

Ziewitz M and Pentzold C (2014) In search of internet governance: performing order in digitally networked environments. *New Media & Society* 16(2): 306–322.

Zolides A (2021) Gender moderation and moderating gender: sexual content policies in Twitch's community guidelines. *New Media & Society* 23(10): 2999–3015.

Author biographies

Ngai Keung Chan is an assistant professor at the School of Journalism and Communication at The Chinese University of Hong Kong. His research examines the intersection of platform governance, algorithms, and service work. His work has appeared in such journals as *New Media & Society*, *Information, Communication & Society*, *Big Data & Society*, *Social Movement Studies*, and *Media and Communication*, among others.

Chris Chao Su is an assistant professor of Emerging Media Studies at the College of Communication, Boston University, USA. His research examines the audience consumption of digital media through computational methods and passively measured data.

Alexis Shore is a PhD candidate in the Division of Emerging Media Studies at Boston University. Her research leverages quantitative and policy-based methods to understand the influence of interface design and interpersonal surveillance on information boundary management.