KISIP 2024
Konferensi Ilmu Sosial dan Ilmu Politik

Safer Internet Lab | CSIS Indonesia | Google

Research Paper

Political Disinformation and Content Moderation "Folklore" in the Global South: Comparative Lessons for the 2024 Indonesia Elections

Panel 4
Law, Regulations, and Governance Against Disinformation and on Content Moderation in Indonesia

Matti Pohjonen

University of Helsinki, Helsinki, Finland.

✉ matti.pohjonen@helsinki.fi

Matti Pohjonen is a Senior Researcher at the Helsinki Institute of Social Sciences and Humanities (HSSH), University of Helsinki. He previously worked as a researcher for the University of Oxford and the VOX-Pol Network of Excellence, and as a Lecturer in Global Digital Media at the School of Oriental and African Studies (SOAS). He has widely published on topics related to digital communication and online extreme speech globally.

Editor: Dandy Rafitrandi

Abstract

Political disinformation usually intensifies during pivotal events like elections, when heightened social tensions underscore the need for timely and reliable information. Historically, one solution has been more effective content moderation, either through removal of content, reduction of its visibility or flagging its risks to users. Yet despite improvements in the transparency of content moderation policies globally, the process of how this happens, however, often remains misunderstood. Moreover, the mechanisms through which this happens, in turn, generates heated debate among social media users about what informs social media companies policies and how these could be circumvented. As crucially, even if such vernacular understandings of content/visibility moderation are often based on unfounded rumours, these beliefs can nonetheless pose a substantial risk to the perceived impartiality of social media platforms, particularly during critical events like elections. Drawing on the theory of extreme speech (Pohjonen and Udupa 2017), the paper argues that such user perspectives and beliefs to content moderation policies form a new kind of "algorithmic folklore" that increasingly informs how the legitimacy of political information and disinformation shared on social media is perceived. Algorithmic folklore is broadly defined as the "beliefs and narratives about moderation algorithms that are passed on informally and can exist in tension with official accounts (Savolainen 2022: 1092)." These beliefs, in turn, can have an oversized role during critical events such as elections when the need for timely and impartial information is elevated – that is, how users perceive the role of social media platforms as impartial mediators of political information significantly influence how the legitimacy of democratic processes are understood more broadly. Through examples from Kenya, Ethiopia and India, the paper proposes three preliminary meta-narratives of content moderation folklore to help better understand the role of AI-generated disinformation in preparation for the Indonesia elections.

Keywords: Disinformation, Content Moderation, Algorithmic Folklore, Extreme speech, Comparative research

Introduction

In April of 2023, a video of Indonesian President Joko Widodo performing a rendition of a popular pop song called *Asmalibrasi* went viral on social media. The video was initially shared on Twitter where it captured over 5 million viewers and garnered more than 10,000 retweets. The song also quickly gained popularity across other social media platforms, including Tik-Tok, YouTube, and Instagram, amassing hundreds of thousands of views and comments (Fikri, 2023).

The song was of course a "deep fake" – a digital artifact made using artificial intelligence (AI). Similar examples of digital manipulation involving the verisimilitude of public figures have proliferated in recent years, propelled by the arrival of new generative AI tools like *ChatGPT* (text), *MidJourney* (images) or *ElevenLabs* (speech). Although the majority of such deep fakes have been created for the purposes of satire or clickbait visibility, the ease by which artificial text, images and audio can now be generated has raised concerns about their potential misuse as vectors of political disinformation.[2] In 2024, 71% of the world population living in democracies will go to the polls (Scheier, 2023). This cascade of elections will occur against the backdrop of growing political uncertainty globally. This confluence of heightened geopolitical tension and the rapid rise of generative AI has led some observers to warn that the electoral landscape in 2024 may yet become characterized as a year of "disinformation on AI steroids (2023: para. 13)."

The growing challenge of political disinformation and AI-generated content is particularly salient for countries in the Global South where a substantial part of the voting population consists of young people who rely on audiovisual platforms such as Tik-Tok, YouTube and Instagram for their political information. Content on these platforms already frequently integrates AI-generated elements, spanning from gimmicky filters to the generation of entirely "fake" visuals, speech, or video. One proposed solution to this convergence of political disinformation and AI-generated content has been to implement better content moderation policies, either through the removal of content, reducing its visibility or flagging its risks to users. Yet, the effectiveness of existing global content moderation systems has come under repeated criticism due to their inability to prevent misleading or hateful speech in countries in the Global South. Gregorio and Stremlau (2023) write that "these regions often find themselves marginalized in current regulatory discussions, even as the global proliferation of harmful speech online is raising questions about the responsibility, and the ability, of social media companies to effectively tackle these challenges (2023: pp. 1)." Concurrently, the automated

[2] I use disinformation here in the sense defined by Wardle and Derakshan (2017) as "false information is knowingly shared to cause harm (2017: 3). That is, in the context of political information, it refers to the use of what social media platforms call "coordinated inauthentic behavior" used to advance political agendas through the illegitimate use of social media platforms such as bots, coordinated accounts and trolls.

systems required to scale up content moderation also significantly underperform in such "low-resource" contexts due to the lack of computational resources available to account for the diversity of languages and cultural idioms used.

In this article, I propose a new framework to research this challenge of political disinformation and AI-generated content. Instead of approaching this object of study as a problem of *detection* (e.g. how do we more effectively identify political disinformation at scale) or *governance/moderation* (e.g. what types of content moderation policies can best help mitigate the harm of AI-generated political disinformation), I ask: *how do social media users and communities themselves understand and adapt to (what they imagine to be) the mechanisms underlying the moderation of new forms of political disinformation globally?* In other words, despite significant improvements to the transparency of content moderation policies globally, these processes still often remain misunderstood by social media users. As a result, the opaque nature of especially the algorithmic logics underpinning algorithmic ranking and recommendation systems – or what Zheng and Kaye (2022) call "visibility moderation" – has resulted in the proliferation of "folk" theories that users engage to understand the hidden forces that determine visibility on social media.

The key argument advanced in this article is thus the following: *such emerging user theories of content/visibility moderation forms a kind of "algorithmic folklore" that shapes how social media platforms are perceived as mediators of political information globally.*[3] Savolainen defines algorithmic folklore as the "beliefs and narratives about moderation algorithms that are passed on informally and can exist in tension with official accounts (Savolainen, 2022: pp. 1092)." These beliefs can play an oversized role during critical events such as elections when the need for timely and impartial is elevated – that is, *how users perceive the role of social media platforms as impartial mediators of political information significantly influences how the legitimacy of democratic processes are understood more broadly.*

With this starting point in mind, the paper explores three questions related to such entanglements of user-generated algorithmic folklore, content/visibility moderation and AI-generated political disinformation:

1. How does such moderation folklore influence the different ways social media understand the role of social media platforms during critical events such as elections?
2. How can this moderation folklore help, in turn, help better understand the different strategies social media users adopt globally to amplify their visibility or prevent the removal of their content;

[3] I use the term content/visibility moderation in this article to illustrate both algorithmic ranking of content and its removal and moderation.

3. How can policymakers better mitigate for such moderation folklore especially in situations where the legitimacy of the political information shared is already mistrusted by social media users?

To explore these questions, the article builds on the theory of "extreme speech" (Pohjonen and Udupa, 2017) to highlight examples of such algorithmic folklore from Kenya, Ethiopia and India. In particular, I propose three *meta-narratives* of algorithmic folklore that can act as a preliminary starting heuristic to structure research and thinking on these questions.

The article is divided into three parts. The first part situates the theoretical argument within the existing literature on content moderation. The second part specifies the theoretical argument with examples from Kenya, Ethiopia and India. The final part concludes with some suggestions on how this proposed framework could be applied to the 2024 elections in Indonesia, keeping especially the urgent debates on political disinformation and AI-generated content in mind.

Literature review: from platform governance to algorithmic folklore

As a part of the broader rubric of platform governance (Gorwa, 2019), content moderation is usually defined as the "process in which platforms shape information exchange and user activity through deciding and filtering what is appropriate according to policies, legal requirements, and cultural norms (Kaye and Zhing, 2022: pp. 61)." A growing body of research has highlighted different perspectives to this object of study, from legal and regulatory challenges involved in content moderation globally or the algorithmic systems required to scale up the detection of hateful or misleading speech (see Gillespie, 2018; Caplan, 2019; Gillespie, 2022).

The framework developed in this article, however, deviates from more policy- and/or governance-related perspectives. Instead, I begin with the "digital folklore" that has proliferated globally in response to the datafication of political communication infrastructures globally (de Seta, 2021). This proposed shift from "what platforms do?" to "what users *think* the platforms do?" can provide an alternative theoretical entry point by focusing on the "shifting grounds and emerging logics of algorithmic governance, not necessarily in terms of the actual practices themselves, but in terms of its *experiential dimension* (Savolainen, 2022: pp. 1092)." This emic dimension to content moderation in all its anthropological messiness, can also generate new empirical insights into the growing entanglements between political disinformation and AI-generated content globally – especially in contexts where the political information shared is *widely perceived to be corrupted by ineffective social media moderation policies and practices.*

This parallel line of inquiry builds on previous research focused on topics such as "algorithmic imaginaries" of Facebook algorithms (Butcher, 2016), "algorithmic gossip" in social media influencer communities (Bishop, 2019), "algorithmic logics" of online dating apps (Huang, Hancock, Tong, 2022), of "algorithmic lore" related to the visibility of YouTube marketing videos (MacDonald, 2023). It also picks up on research on "social media commentary" as a mechanism through which the meaning of concepts such as hate speech are contested on social media by user communities globally (Pohjonen, 2019).

These approaches share the assumption that there is a kind of information asymmetry that exists between social media platforms and their users. Cotter (2023) uses the term "black box gaslighting" to describe the different ways "platforms leverage their epistemic authority to prompt users to question what they know about algorithms, and thus destabilize the very possibility of credible criticism (2023: pp. 1227)." Savolainen (2022), in turn, argues that such user perspectives now constitute a new kind of moderation "algorithmic folklore" that affects how content/visibility moderation, and the algorithmic systems driving them, are understood by social media users. Such beliefs, she argues, do not need to reflect what platforms do; rather, they function as a kind of "*discursive gathering point for the articulation of multiple experiences and beliefs of platform governance, united by the feelings of uncertainty and not knowing* (2022: pp. 1094; my italics)."

The preponderance of research in this topic, however, has so far focused on professional social media creators or marginalized communities (Duffy & Meisner, 2023). There is less research available on how such (mythological) beliefs relate specifically to debates on political disinformation and its moderation in different global contexts. Moran, Grasso and Koltai (2022) show how the moderation of anti-vaxx information during the Covid-19 pandemic provoked audiences to "demonize, celebrate and attempt to ingratiate themselves to the mysterious algorithms to enhance their desired outcomes on social media (2022: pp. 2)." They write that

> the speed, creativity, and flexibility of folk theorization around social media match that of algorithmic change, meaning that content moderation measures spur new theories and tactics for circumvention … [that] *cultivate theories around why they are (allegedly) being targeted for moderation, how this moderation occurs and strategies to avoid it* (2022: pp. 10; my italics).

The key point to pick up from this burgeoning literature is thus that, whether such beliefs and theories are true or not, is often of secondary importance. Such beliefs operate on a different epistemological register. In an anthropological sense, they constitute a kind of modern folk-mythological structure of knowledge that people

rely on to explain the mysterious and hidden forces that (they believe) control their lives (see West and Sanders, 2003; Cotter et al, 2022). At the same time, regardless of their factuality, technical or otherwise, the experiential dimension of content/visibility moderation can nonetheless significantly *influence how users and user communities perceive the impartiality of social media platforms as mediators of political information during critical events such as elections*. As we have seen, for instance, during the 2016 US presidential elections (e.g. was it Cambridge Analytica or Russia that manipulated the election results?) and other subsequent elections (e.g. was it actually AI-generated fake audio that determined the Slovenian elections?) allegations of widespread social media manipulation – whether AI-generated or "shallow fakes" – can have a detrimental effect on the trust people have in the democratic process. With the arrival of generative AI as a new vector for producing content on social media at a speed and scale never possible before, the pondering question of how users perceive the authenticity and veracity of political information will thus become crucial to understand (see Seta, Pohjonen and Knuutila, 2023).

Toward "folk" theories of content/visibility moderation in the Global South

How can we then best understand such mythopoetic perspectives to political disinformation and its moderation? One useful entry point into these debates is to see algorithmic folklore as a form of *extreme speech* that has proliferated in response to the pervasive datafication of political communication infrastructures globally. The extreme speech framework was developed to provide a more anthropological perspective to global debates on hate speech and disinformation by foregrounding "the situatedness of online speech forms in different cultural and political milieus (Pohjonen & Udupa 2017: pp. 1174)." This framework has been subsequently used to produce theoretical and ethnographic insight into various types of extreme speech cultures globally (see Udupa & Pohjonen, 2019; Udupa, Gagliardone & Hervik, 2021).

To illustrate what this conjuncture could mean in practice, I will highlight three types of examples from Kenya, Ethiopia and India that I have found useful in structuring my thoughts on the topic – or what I tentatively call in this article *three meta-narratives of algorithmic folklore* The purpose of these examples is not to provide a comprehensive account of different types of algorithmic folklore related to content/visibility moderation globally – it would be impossible to do this in a short article. Rather, the aim is to show how this analytical shift from "what platforms do" to "what users *think* the platforms do" can potentially open up new insights to aid policy makers and other stakeholders mitigate the risk of political disinformation.

Algorithmic folklore and disinformation influencers for hire

The first example from Kenya illustrates how such moderation folklore related to the algorithmic logics of content/visibility moderation system has now engendered a burgeoning industry of "disinformation influencers" globally. While the use of bots, sock puppet accounts, and other forms of coordinated inauthentic behavior has been widely documented (Assemacher et, 2020; Olaniran, 2022), the Kenyan example suggests how pervasive this form of political communication has become. Focusing on this class of shadow influencers can thus provide researchers with new insights into what factors drive the adoption, and/or rejection, of different types of political disinformation globally and how users perceive the role of algorithms that determine the visibility of such content (see Rudyansjah and Rasidi, 2022).

Kenya has historically had one of the most dynamic social media communities in Africa. A celebrated example of this is the Twitterati who loosely assemble behind the gathering call of *#KoT* – Kenyans on Twitter. This active and often rambunctious assemblage of social media users has historically campaigned against corruption and other social causes (Ogola, 2022). At the same time, Kenya is also a volatile democracy with political factions divided by ethnic faultlines. Its social media communities have thus played a positive role but social media has also been widely used to ferment inter-ethnic conflict and communal violence especially in the aftermath of the presidential elections in 2012 and 2017. It is within this congruence of conflictual politics and tech-savvy social media users that researchers have also noted the emergence of what one critic calls a "*booming and shadowy industry of Twitter influencers for political hire* (Madung and Obilo, 2021: pp. 3; my italics)." This industry largely operates through social media influencers (claiming to) sell their expertise in being able to "game" the algorithmic ranking and recommender systems of social media platforms for political gain – or in the words of one interviewed influencer: *"the main goal is to go trending on Twitter. I'm not sure what our jobs would look like without that target* (Madung and Obilo, 2021: pp. 12; my italics).

The growth of this disinformation-for-hire industry has been attributed to many factors, such as a large tech-savvy unemployed youth population, a polarized and money-driven political culture known for its winner-takes-all political campaigning, and the scarcity of resources social media companies such as Twitter have allocated to countries such as Kenya for monitoring coordinated inauthentic behavior (Odenga, 2021; Odenga, 2022). At the same time, these allegations of being able to "game" the algorithmic logics of content/visibility moderation, critics argue, have also led to a growing skepticism among social media users about what are the "actual" mechanisms that drive political visibility

and what social media companies are doing about such coordinated behavior. Madung (2022) remarks that

> Many accounts and individuals involved promote brands, causes and political ideologies without disclosing that they are part of paid campaigns. This is a lucrative, well-oiled machine with very clear targets and *as a result it is chilling good faith activism.* Twitter's features are being exploited to achieve the goals of these campaigns. Its trending algorithm is amplifying these campaigns and accounts verified by the platform are complicit in leading these attacks. *The goal of these campaigns is to exhaust critical thinking and poison the information environment by annihilating truth* (2022: pp. 2).

The first type of meta-narrative thus suggests that such user folklore related to the *vulnerability of social media platforms and their ranking and recommender algorithms to manipulation* by tech-savvy disinformation influencers manifests in two ways. Firstly, there are the different theories that users create to explain how such influencers are able to exploit the inner workings of social media moderation systems for financial and political gain. As importantly, it also manifests in the growing skepticism about what, in fact, can be trusted as legitimate political information if *the companies are unable to effectively respond to such coordinated inauthentic behavior.* At worst, such algorithmic folklore of mistrust can risk poisoning the information environment and prevent "good faith activism" in countries where social media has been historically celebrated also for its positive role in campaigns against corruption and for mitigating outbreaks of violence during elections.

Algorithmic folklore and weaponization of content/visibility moderation

The second example from Ethiopia, in turn, suggests that this widespread mistrust of social media content/visibility moderation policies globally can contribute to the "weaponization" of moderation mechanisms such as users flagging of content (Meisner, 2023; see also Conteras, 2021; Crawford & Gillespie, 2016; Kayser-Bril, 2021). Particularly in contexts where the effectiveness of existing moderation systems is already widely questioned, the use of coordinated behaviors such as "mass reporting" can provide a viable option when no other alternatives are seen to be available.

The use of social media in Ethiopia has been relatively slow to develop compared to countries such as Kenya. This has been partially the outcome of low

internet penetration rates but also because of the strong-armed control the Ethiopian government has historically maintained on its social media environment (Gagliardone et al, 2016: Gagliardone and Pohjonen, 2016). The growing use of social media in Ethiopia, however, has recently contributed to political protests and it has been widely criticized as a forum for spreading violent ethnic hate speech (see Worknet, 2020). The case of Ethiopia is especially important to highlight because of the visible role the country has had in global debates on content/visibility moderation. Alongside Myanmar, Ethiopia is perhaps the country most cited as an example of the *failure of social media platforms* in preventing hate speech and violence in countries in the Global South (see Schemm, 2016; Taye & Paller, 2020).

Similar moderation folklore was also popular during the Tigray War in Ethiopia in 2020. The conflict was quickly accompanied by a kind of "digitally mediated *epistemic proxy war* that accompanied the rapidly changing events on the ground (Pohjonen 2022: pp. 242; italics in original)." Moreover, this proxy war also incorporated elements from global debates on content/visibility moderation into the rhetoric used by both sides of the conflict. Figure 1, for instance, illustrates how the widely shared testimony of whistleblower Francit Haugen in the US congress was co-opted to support narratives related to conflict (see Milmo, 2021).

Figure 1: Twitter post about the Haugen revelations is linked to accusations of the use of mass reporting as a tactic of information warfare

Similarly, Figure 2 shows another example of how accusations about the use of "mass reporting" – together with the allegations of the indifference or inefficiency of social media platforms in preventing such deceptive content – were used as a rhetorical tool to try to control the narrative and/or to discredit or remove content coming from the opposing side of the conflict (see Jeffrey, 2019; Drew & Wilmot, 2021).

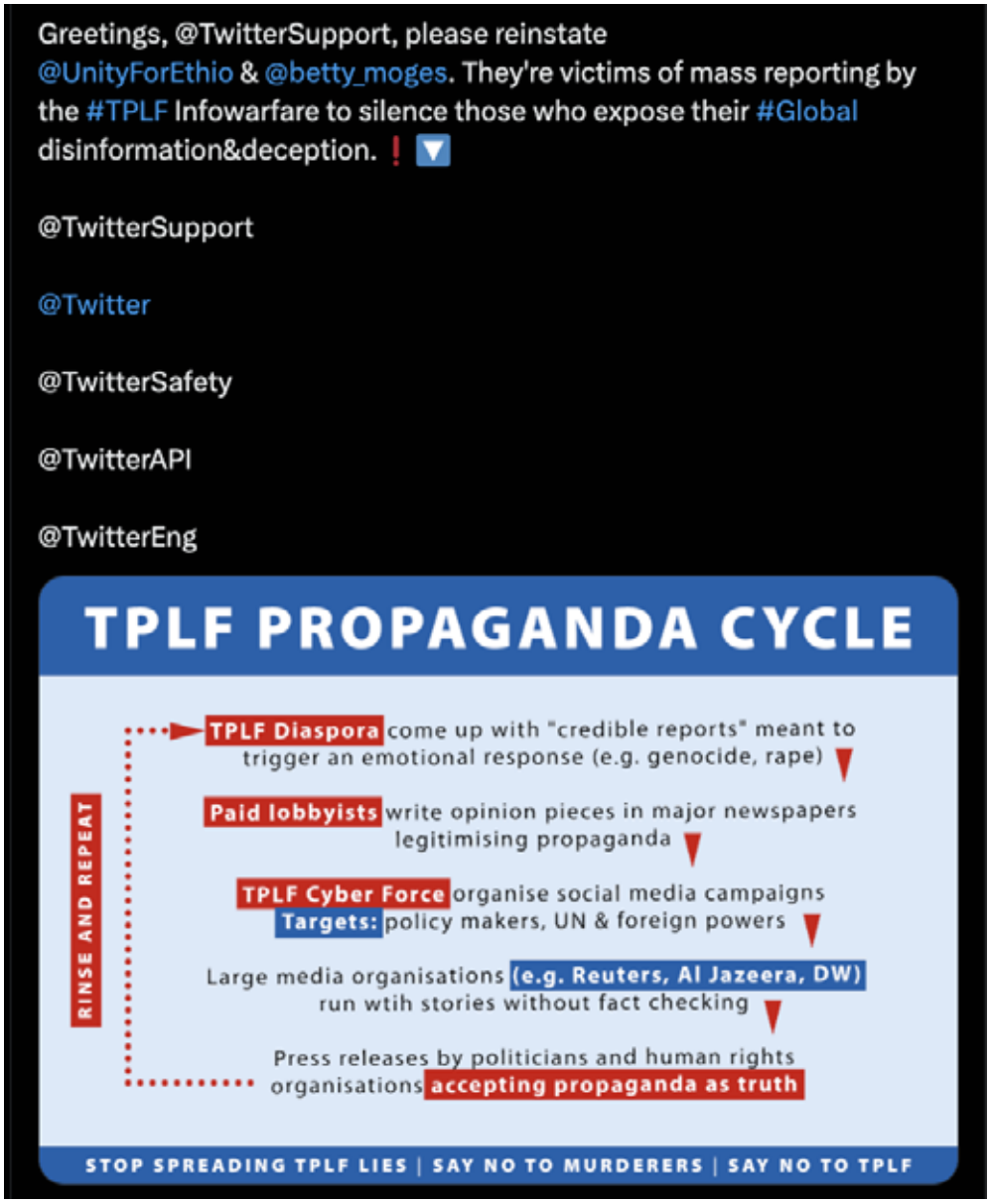

Figure 2: Algorithmic folklore linking war propaganda with content/visibility moderation mechanisms such as mass reporting

Allegations related to the manipulation of social media for war propaganda are, of course, not unique to Ethiopia; these have been also widely documented in other conflicts such as the Ukraine War (see Mehndi, 2023). These examples of moderation folklore from Ethiopia, however, suggest that one consequence of this *growing discourse of failure* related to social media moderation policies and practices globally is the growing popularity of alternative tactics, such as coordinated mass reporting as a way for users to "game" the moderation systems of social media platforms. The user folklore related to such practices of coordinated reporting of oppositional content – including allegations of manipulation of social media moderation systems through coordinated mass reporting – can thus provide further insights into how users debate and share best

strategies to prevent political disinformation in response to the perceived *failure of content/visibility moderation globally*.

Algorithmic folklore and the rise of digital authoritarianism

The final example from India signals how the growing power of governments in controlling political information on social media, or what some critics have called the rise of "digital authoritarianism" globally (Wilson, 2022). Freedom House notes that internet freedoms have been in decline now for 13 years in a row (Funk et al, 2023). One of the contributing factors to this decline has been the increased popularity of legislative and punitive actions aimed at forcing social media companies to remove objectionable content in different national contexts. Marchant and Stremlau (2019) contend that what social media companies do is just one among the many solutions governments have available to respond to a "growing frustration with how difficult it is to control extreme speech and misinformation on social media and the *perceived inaction and inability of large companies to effectively address those challenges* (Marchant & Stremlau: 2019: pp. 4218)."

A panoply of other tactics such as internet shutdowns and social media bans have thus increasingly also contributed to debates on content/visibility moderation globally (also see Ruijgrok, 2022; Yilmaz et al., 2022; Eichard and Linnart, 2023; Alkiviadou, 2023).[4] An often cited example of this is the *Intermediary Guidelines and Digital Media Code* law passed in India in 2021 that compelled social media companies to pre-screen all the content published on social media for objectionable content and maintain a record of the "first originator" of the content (Ashwini, 2021). The legislation emerged out of the backdrop of massive protests that erupted early in 2021 in India in response to a set of *Farm Bills* that were seen by critics to significantly restrict the rights of farmers (Behl, 2022). These escalating protests includes a mass farmer blockade of the capital of India, New Delhi, and resulted in the Indian government issuing orders to Twitter to remove all content that used hashtags such as *#farmergenocide* in support of the protests. In response to the initial recalcitrance of Twitter to comply with the takedown request, the Indian government threatened to jail employees of Twitter. Twitter responded, arguing that they were simply defending the right to free speech but were forced to initially comply with the takedown requests (Rajvanhsi, 2023).

While the farmers' protests were just one among many examples of the increasingly antagonistic relationship the Indian government has with social media companies – including banning Tik-Tok entirely from the country – digital activists

[4] Facebook has been temporarily or partially banned by 30 countries globally; YouTube has been temporarily banned in 23 countries and permanently in 5. Twitter is blocked in seven countries, and temporarily banned in Egypt, Nigeria and Turkey after government's requests to remove content.

have argued that the growing power over social media platforms has had severe consequences to freedom of speech, the right to protest, and the ability to criticize the government in India (Bhatia, 2023). Subsequently, social media companies have largely complied when the Indian government has requested to remove content, including Twitter blocking hundreds of accounts belonging to critical journalists, authors and politicians in 2023 (Sakunia, 2023). The example from India is thus interesting because it illustrates how the increasingly antagonistic relationship between governments and social media companies influences the type of algorithmic folklore created to explain the visibility of political content available on social media and possible government interference into this. Bhatia (2023) calls such user perspectives "bottom-up imaginaries" of platform governance, which link together India's idiosyncratic historical and political conditions with global debates on content/visibility moderation and extreme speech. He writes that such

> bottom-up imaginaries of social media are not independent of the government's imagination for the country's future .. for Indian citizens, memories of colonization, histories of violence and oppression, the partition based on religious identities, diverse ideologies, policies, and actions of the past governments, and fear of being overtaken by Western powers again through data and technologies ... *all of these get intertwined in the quotidian online discourse surrounding episodes that sharply divide the nation* (Bhatia, 2023: pp. 252).

Udupa and Pohjonen note that understanding such socio-political and cultural factors behind extreme speech can help "examine the specific contexts that instigate and shape online extreme speech as violence, and its divergent and often unforeseen implications (Udupa and Pohjonen, 2019: pp. 3053)." Similarly, the growing stronghold the Indian government has over social media platforms has led to widespread mistrust about whether social media platforms can be seen as neutral or impartial mediators of political information. The third example of algorithmic folklore thus suggests that such debates on content/visibility moderation are not only linked to what platforms do, or the technical details of the algorithms structuring visibility – but also to the *broader discourse* behind how users respond to what critics call a growing "digital authoritarianism" globally and its impact of political information shared on social media (Wilson, 2022).

Conclusion: Interpreting Algorithmic Folklore in the Context of Indonesia's 2024 Elections

I have argued in this article that such an analytical reorientation from the actions of platforms to the perceptions users hold about these actions can offer new

insights into the interplay of political disinformation and the mechanisms of content moderation and visibility on social media. While case studies from Kenya, Ethiopia, and India can start this conversation, they only scratch the surface of a wider phenomenon that still lacks empirical attention (see Rasidi 2023a). So, what could a closer examination of moderation folklore tell us about the impending Indonesian (and other) elections? In what ways could a more systematic understanding of such user perspectives enhance our comprehension of the implications and fear of AI-driven disinformation?

The 2023 *Freedom on the Net* report, "The Repressive Power of Artificial Intelligence," warns that at least 47 countries have documented uses of social media activists who use deceitful tactics to shape online discourse. These actors, the report also warns, are now also increasingly "employing AI-generated images, audio, and text, making the truth easier and harder to discern (Funk et al, 2023: para 3)." Alvarez (2023) suggests that examples of such AI-mediated disinformation during election times could include, among other things, the production of false but believable disinformation aimed at swaying voters, manipulated media to present confusing representations of public figures, and/or the generation negative, misleading and inflammatory content to target candidates and their voters (2023: pp. 2). Some of the suggested counter-measures, in turn, include fact-checking, flagging or removing false content, warning users of content that is AI-generated and/or banning pages spreading AI-generated disinformation and false material (2023: pp. 3).

Yet many current recommendations for mitigating the risk of AI-generated political disinformation focus on the actions of legislators, policymakers and social media companies. There is still less knowledge on how social media users themselves perceive or understand the significance of such AI-generated content or the parallel attempts to moderate its use? *How do users, for instance, define the authenticity of different types of political information against the backdrop of the growing use of AI for generating content online? What are the global patterns of resistance to moderation efforts within diverse online communities?* The three preliminary meta-narratives I have suggested in this article – a sustained focus on disinformation influencers, on the weaponization of content moderation systems and a focus on the consequences of the rise of digital authoritarianism globally – can provide a preliminary starting point for such research.

The examples highlighted from Kenya, Ethiopia and India are tied to the unique socio-political histories and media environments of these countries. Yet similar experiences are common to other countries as well. In Indonesia, for instance, researchers have widely documented how the electoral landscape is similarly characterized by the growth of disinformation influencers or "buzzers" spurred on by a growing mistrust in the mainstream media or the government (Panditharatne, 2023; Rasidi, 2023a; Umami & Al Windy, 2023; Sastramidjaja et al

(2022). *How, where and when are such "buzzers," then, engaging with especially different types of AI-generated content in preparation for the 2024 election (see Rasidi, 2023b)? What types of algorithmic folklore are other social media users engaging in to counter the growing power of buzzers in a media environment maculated by inauthentic coordinated behavior and mistrust in the government media?*

In conclusion, I propose that a deeper understanding of the algorithmic folklore users engage with to make sense of the increasing complexities surrounding political misinformation, content regulation, and the proliferation of AI-produced content can provide one starting point. Such new empirically grounded research can, in part, also guide policymakers and social media platforms in devising new approaches to counteract the challenges posed by political disinformation and AI-generated content globally.

References

Alkiviadou, N. (2023). "The internet, internet intermediaries and hate speech: freedom of expression in decline?" *SCRIPTed: Journal of Law, Technology and Society*, 20(1), 243-268.

Ashwini, S. (2021). "Social Media Platform Regulation in India–A Special Reference to The Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021." *Perspectives on Platform Regulation*, 215-232.

Assenmacher, D., Clever, L., Frischlich, L., Quandt, T., Trautmann, H., & Grimme, C. (2020). "Demystifying Social Bots: On the Intelligence of Automated Social Media Actors." *Social Media + Society*, *6*(3). https://doi.org/10.1177/2056305120939264

Behl, N. (2022). "India's Farmers' Protest: An Inclusive Vision of Indian Democracy." *American Political Science Review*, 116(3), 1141-1146. Doi:10.1017/S0003055422000156

Bishop, S. (2019). "Managing visibility on YouTube through algorithmic gossip." *New Media & Society*, 21(11-12), 2589-2606. https://doi.org/10.1177/1461444819854731

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

Caplan R. (2019). "Content or context moderation? Artisanal, community-reliant, and industrial approaches." *Data & Society Research Institute.* https://datasociety.net/library/content-or-context-moderation/

Contreras, B. (2021). "TikTok creators say they lose videos through mass reporting." *The Los Angeles Times.* , December 3, 2021. https://www.latimes.com/business/technology/story/2021-12-03/inside-tiktoks-mass-reporting-problem

Cotter, K., DeCook, J., Kanthawala, S., & Foyle, K. (2022). “In FYP We Trust: The Divine Force of Algorithmic Conspirituality.” I*nternational Journal Of Communicatio*n, 16, 24. https://ijoc.org/index.php/ijoc/article/view/19289

Cotter, K. (2023) “‘Shadowbanning is not a thing’: black box gaslighting and the power to independently know and credibly critique algorithms.” *Information, Communication & Society*, 26:6, 1226-1243, DOI: 10.1080/1369118X.2021.1994624

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. https://doi.org/10.1177/1461444814543163

Drew, A., & Wilmot, C. (2021). “In Ethiopia’s digital battle over the Tigray region, facts are casualties.” February 5, 2021. *Washington Post*. https://www.washingtonpost.com/politics/2021/02/05/ethiopias-digital-battle-over-tigray-region-facts-are-casualties/

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

Eichhorn, K., & Linhart, E. (2023). “Election-related internet-shutdowns in autocracies and hybrid regimes,” *Journal of Elections, Public Opinion and Parties*, 33:4, 705-725, DOI: 10.1080/17457289.2022.2090950

Fikri, H. (2023). Fikri, H. “Jokowi deepfakes? Fears grow over AI-generated election hoaxes.” *Jakarta Post.* June 7, 2023. https://www.thejakartapost.com/indonesia/2023

Funk, A., Shahbaz, A., and Vesteinnson, K. (2023). “Freedom on the Net 2023: The Repressive Power of Artificial Intelligence.” *Freedom House*. Accessed from: https://freedomhouse.org/report/freedom-net/2023/repressive-power-artificial-intelligence#the-repressive-power-of-artificial-intelligence

De Gregorio, G. & Stremlau, N. (2023). “Inequalities and content moderation.” *Global Policy,* 00, 1–10.

De Seta, G. (2019). “Digital Folklore.” In: Hunsinger, J., Klastrup, L., Allen, M. (eds) *Second International Handbook of Internet Research*. Springer, Dordrecht. https://doi.org/10.1007/978-94-024-1202-4_36-1

de Seta, G., Pohjonen, M., & Knuutila, A. (2023). “Synthetic ethnography: Field devices for the qualitative study of generative models.” *SoCArxiv Papers preprint*. Accessed from: https://doi.org/10.31235/osf.io/zvew4

Gagliardone, I. & Pohjonen, M. (2016).”Engaging in polarized society: social media and political discourse in Ethiopia'' In Mutsavairo, B (eds). *Digital Activism in the Social Media Era*. Pp. 25-44. London: Palmgrave McMillan

Gagliardone I., Pohjonen M., et al (2016). “Mechachal – Online debates and elections in Ethiopia. Final Report: From hate speech to engagement in social media.” *The Programme in Comparative Law in Media and Policy (PCLMP), University of Oxford & Addis Ababa University*. https://ora.ox.ac.uk/objects/uuid:da10c4ee-2726-41cf-ba21-

72c9f7cbc440/download_file?file_format=application%2Fpdf&safe_filename=Mechachal_-_Online_Debates_and_Elections.pdf&type_of_work=Report
Gilbert, D. (2020). "Hate Speech on Facebook Is Pushing Ethiopia Dangerously Close to a Genocide." *Vice News*, 14 September, 2020. https://www.vice.com/en/article/xg897a/hate-speech-on-facebook-is-pushing-ethiopia-dangerously-close-to-a-genocide
Gillespie T. (2018). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.
Gillespie, T. (2022). "Do Not Recommend? Reduction as a Form of Content Moderation." *Social Media + Society*, 8(3). https://doi.org/10.1177/20563051221117552
Gorwa, R. (2019). "What is platform governance?" *Information, Communication & Society*, 22:6, 854-871, DOI: 10.1080/1369118X.2019.1573914
Huang, S. A., Hancock, J., & Tong, S. T. (2022). "Folk Theories of Online Dating: Exploring People's Beliefs About the Online Dating Process and Online Dating Algorithms." *Social Media + Society,* 8(2). https://doi.org/10.1177/20563051221089561
Hui, J. Y. (2020). "Social Media and the 2019 Indonesia Elections: Hoax Takes the Centre Stage." *Southeast Asian Affairs*, 155–172. https://www.jstor.org/stable/2693889
Jeffrey, J. (2019). "The challenges of navigating Ethiopia's new media landscape". *Al-Jazeera.* October 29, 2019. https://www.aljazeera.com/news/2019/10/29/the-challenges-of-navigating-ethiopias-new-media-landscape
Kayser-Bril, N. (2021, February 1). "The insta-mafia: How crooks mass-report users for profit." *AlgorithmWatch*. https://algorithmwatch.org/en/facebook-instagram-mass-report/
Kessler, G. (2023). "The truth about Russia, Trump and the 2016 election." *Washington Post.* May 17, 2023. https://www.washingtonpost.com/politics/2023/05/17/truth-about-russia-trump-2016-election
Levine, D. N. (2011). "Ethiopia's nationhood reconsidered." *Análise Social*, 46(199), 311–327. Retrieved from http://www.jstor.org/stable/41494856
MacDonald, T. W. (2023). "How it actually works": Algorithmic lore videos as market devices. *New Media & Society*, 25(6), 1412-1431. https://doi.org/10.1177/14614448211021404
Madung, O. (2022). "How Twitter's Negligence is Harming Kenya's Democracy." *The Elephant,* July 1, 2022. https://www.theelephant.info/op-eds/2022/07/01/how-twitters-negligence-is-harming-kenyas-democracy/?print=pdf
Mahedi, H. (2023). "Russia Ukraine Propaganda on Social Media: A Bibliometric Analysis." *SSRN*: https://ssrn.com/abstract=4522077
Meaker, M. (2023). "Slovakia's Election Deepfakes Show AI Is a Danger to Democracy." *The Wired*. Oct 3, 2023.

https://www.wired.com/story/slovakias-election-deepfakes-show-ai-is-a-danger-to-democracy/

Moran, R. E., Grasso, I., & Koltai, K. (2022). "Folk Theories of Avoiding Content Moderation: How Vaccine-Opposed Influencers Amplify Vaccine Opposition on Instagram." *Social Media + Society*, 8(4). https://doi.org/10.1177/20563051221144252

Nyambola, N. (2020). *Digital Democracy, Analogue Politics: How the Internet Era Is Transforming Politics in Kenya*. London: Zed Books.

Ogenga, F. (2022) "Mitigating Election Violence through Social Media Micro-Influencers: Baseline Report," *Center for Media Democracy, Peace, and Security*, Rongo University. http://repository.rongovarsity.ac.ke/handle/123456789/2417.

Ogenga, F. (2021). "Social Media, Ethnicity and Peacebuilding in Kenya" in *The Techtonic Shift – Social Media Impacts on Conflicts and Democracy*. London, New York. Routledge, 2021, Routledge, pp 131-140.

Ogola, G.O. (2023). "Digital (Dis)order, Twitter Hashtags, and the Performance of Politics in Kenya." In *Cryptopolitics: Exposure, Concealment, and Digital Media.* Berghahn Books.

Ong, J.C: & Tapsell, R. (2022). "Demystifying disinformation shadow economies: fake news work models in Indonesia and the Philippines," *Asian Journal of Communication*, 32:3, 251-267, DOI: 10.1080/01292986.2021.1971270akes-show-ai-is-a-danger-to-democracy/

Olaniran, S. (2022). "Disinformation: Exploring the nexus between politics and technology in Nigeria". *Phd Thesis*. University of Witwatersrand. Accessed from: https://wiredspace.wits.ac.za/bitstreams/80083de1-e698-4a84-947a-7ce835ec6fbe/download

Panditharatne, M. (2023). "How AI Puts Elections at Risk — And the Needed Safeguards." *Brennan Centre for Justice*. July 21, 2023. https://www.brennancenter.org/our-work/analysis-opinion/how-ai-puts-elections-risk-and-needed-safeguards

Pohjonen, M. (2019). "A comparative approach to social media extreme speech: Online hate speech as media commentary". *International Journal of Communication*, Vol 13: 3088–3103.

Pohjonen, M. & Udupa, S. (2017). "Extreme speech online: an anthropological critique of hate speech debates." *International Journal of Communication*. Vol 11: pp. 1173–1191

Rajvanshi, A. (2023). "Why Twitter Co-Founder Jack Dorsey Is Clashing with India's Government." *Time*. June 13, 2023. https://time.com/6286814/india-twitter-jack-dorsey-clash/

Rasidi, P.P. (2023a). "Ludic cybermilitias: shadow play and computational propaganda in the Indonesian predatory state," *Communication, Culture and Critique* https://doi.org/10.1093/ccc/tcad020

Rasidi, P.P. (2023b). "Transformative Working-Class Labor in Indonesia's Political Influence Operations." *Tactical Technology Collective*. https://influenceindustry.org/en/explorer/case-studies/indonesia-political-influence-operations/

Rudyansjah, T. & Rasidi, P.P.(2022). "Virtual embodiment in physical realities: Brand buzzers and disciplined bodies in an Indonesian cyberscape." *HAU journal of ethnographic theory.* https://www.haujournal.org/index.php/hau/article/view/1699

Ruijgrok, K. (2022). "The authoritarian practice of issuing internet shutdowns in India: the Bharatiya Janata Party's direct and indirect responsibility", *Democratization,* 29:4, 611-633, DOI: 10.1080/13510347.2021.1993826

Savolainen, L. (2022). "The shadow banning controversy: perceived governance and algorithmic folklore." *Media, Culture & Society,* 44(6), 1091-1109

Sakunia, S. (2023). "Twitter blocked 122 accounts in India at the government's request." *Rest of the world*. 24 March, 2023. https://restofworld.org/2023/twitter-blocked-access-punjab-amritpal-singh-sandhu/

Sastramidjaja, Y., Rasidi, P.P and Elsitra, G.N. (2022). "Peddling Secrecy in a Climate of Distrust: Buzzers, Rumours and Implications for Indonesia's 2024 Elections." *ISEAS perspective*. Issue: 2022 No. 8524. https://www.iseas.edu.sg/articles-commentaries/iseas-perspective/2022-85-peddling-secrecy-in-a-climate-of-distrust-buzzers-rumours-and-implications-for-indonesias-2024-elections-by-yatun-sastramidjaja-pradipa-p-rasidi-and-gita-n-elsitra/

Scheier, B. (2023). "AI disinformation is a threat to elections – learning to spot Russian, Chinese and Iranian meddling in other countries can help the US prepare for 2024." *The Conversation*. September 29, 2023. https://theconversation.com/ai-disinformation-is-a-threat-to-elections-learning-to-spot-russian-chinese-and-iranian-meddling-in-other-countries-can-help-the-us-prepare-for-2024-214358

Schemm, P. (2016). "In Ethiopia's war against social media, the truth is the main casualty." *The Washington Post*. , October 14, 2016. https://www.washingtonpost.com/news/worldviews/wp/2016/10/14/in-ethiopias-war-against-social-media-the-truth-is-the-main-casualty/

Taye, B., & Paller, J. (2020). "Open letter to Facebook on violence-inciting speech: Act now to protect Ethiopians". *Access Now*. July 27, 2020 https://www.accessnow.org/open-letter-to-facebook-protect-ethiopians/

Umami, A. M., & Al Qindy, F. H. (2023). "The Use of Buzzers by Political Parties Which Result in Black Campaign Practices in Indonesia". *JISIP (Jurnal Ilmu Sosial dan Pendidikan)*, *7*(4).

Upupa, S., Gagliardone, I., Hervik, P. (2021). *Digital Hate: the Global Conjuncture of Extreme Speech*. Bloomington: Indiana University Press.

Udupa. S. and Pohjonen, M. (2019). "Introduction: Special Issue on Global Digital Cultures and Extreme Speech." *International Journal of Communication*, Vol. 13: 3049–67

Wardle, C., and Derakshan, H. (2017). "Information disorder: Toward an interdisciplinary framework for research and policy making." *Council of Europe report DGI(2017)09*. https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html

West, H. G., & Sanders, T. (2003). *Transparency and conspiracy: Ethnographies of suspicion in the new world orde*r. Duke University Press.

Wilson, R.A. (2022). "Digital Authoritarianism and The Global Assault on Human Rights." *Human Rights Quarterly* 44(4), 704-739. https://doi.org/10.1353/hrq.2022.0043.

Workneh, T. W. (2020). "Social media, protest, & outrage communication in Ethiopia: Toward fractured publics or pluralistic polity?" *Information, Communication & Society*, 24(3), 309–328. Doi:10.1080/1369118X.2020.1811367

Yilmaz, I., Saleem, R. M. A., Pargoo, M., Shukri, S., Ismail, I., & Shakil, K. (2022). "Religious Populism, Cyberspace and Digital Authoritarianism in Asia: India, Indonesia, Malaysia, Pakistan, and Turkey (Version 1)." *Deakin University.* https://hdl.handle.net/10536/DRO/DU:30162438

Zeng, J., & Kaye, D. B. V. (2022). "From content moderation to visibility moderation: A case study of platform governance on TikTok." *Policy & Internet*, 14, 79–95.