International Journal
on
Management Education & Emerging Technology

GOOD OR BAD? STABILITY ON CONTENT MODERATION POLICIES AND USER RESPONSIBILITY IN SOCIAL MEDIA: AN APPROACH TO SOCIAL CONTRACT THEORY

Mohammad Jaoshannee Harris[1], Lennon Marcus Jacob Magno[2], Sherwin Villar[3], Ryne Ken Velasco[4], Rexson Acebedo[5], Angielyn Soliven[6]
[1]*Polytechnic University of the Philippines – Parañaque Campus.*

Abstract

This research explores the applicability of social contract theory in the digital realm, with a focus on Facebook and certain countries. Our research, which examines user behavior and the critical role of a digital social contract, focuses on a small number of yet representative cases from Facebook, emphasizing the platform's dedication to user safety. Our investigation, which is in line with previously published studies, highlights the continued necessity of policy adaptation, user education, and technical improvements for responsible social media dynamics. Considering the limits of our restricted approach, we propose for further study that includes a larger range of platforms and cases.

Article Info

Keywords:
Online Interactions, Social Media Safety, Policy Adaptation, User Education, Community Guidelines.

Co-Author Affiliation
[2,3,4,5,6,7]*Polytechnic University of the Philippines – Parañaque Campus.*

Date of Submission: 17-02-2024
Date of Acceptance: 03-03-2024
Date of Publication: 13-03-2024
Ijmeet/Volume2, Issue1 (Jan-March)2024

INTRODUCTION

In the expansive realm of social media, a contemporary rendition of the social contract unfolds through often-overlooked terms, agreements, and community standards. This idea, however, is a product of the digital age that has been borrowed from classical political philosophy by Michael J. Quinn in Ethics for the Information Age and is based on an agreement or contract among members of a society or community. When users join digital communities or social media, they, by implication, agree to follow the regulations that have been established by the platform. As a digital social contract, it regulates user behavior in accordance with common standards with limitations to what is acceptable and unacceptable behavior in the virtual world. However, despite this apparent harmony, a considerable task in the digital social contract lies ahead – the search for stability in the content moderation policies. The agreement seems simple; however, the quest for a consistent pattern in the ever-changing nature of online conversation presents a significant obstacle. Challenges posed by the unrelenting tide of user-generated content, the cultural dimension, and the need for constant scrutiny put pressure on the stability of such guidelines. Hence, the dilemma to be addressed surfaces: The question is, how can content moderation policies walk the fine line between expression freedom and harm-mitigation that ensures stability within the flexible constitution of the digital social contract?

Facebook is a platform where a study of this nature would be ideally conducted since it has the largest user base making it a reliable foundation to study the digital social dynamics. With over 3 billion users worldwide, Facebook emerges as the most popular social media platform according to Statista (2023). With a global user base in billions, Facebook is a leader in setting the boundaries of the digital social contract, dictating norms and expectations that guide various online interactions. This huge body of users provides a complex and wholly different background for analyzing the sustainability of content regulation guidelines and user obligations under the concept of a digital social agreement. Thus, the digital social contract serves as a moral guide, building a theoretical model and basis for interpreting both the ethicality of user interaction and platform regulation. It lays out the groundwork for accountable and respectful online socializing, pointing to the shared responsibility of both individuals and sites in conducting a beneficial and safe social media framework. In this social contract inside of virtuality, the mutuality of obligations surfaced clearly. Digital connectivity and expression bring advantages to users while they promise to be consistent with the existing rules by not doing anything that might contravene them. On the other hand, platforms shoulder their responsibility to protect user safety, privacy, and freedom of expression within appropriate limits.

Therefore, the study aims to apply the theory of social contracts to the digital world to find out if people's ideas about a social contract can explain how they behave to stay safe online and to establish the need for a digital social contract based on the digital society's connections, contributions, and sense of community. It also wants to look into the set of clear and implicit rules that govern how people and organizations interact with each other online, especially when it comes to social media policies for managing content.

Social Media and Safe Space online

A safe online space – this term can be used to mean an environment cultivated within the virtual landscape where people feel protected to speak up, engage with other users, and consume information without the possibility of cyberstalking, discrimination or exposure to violent content The importance of safe setting in the social media does not come second as it affects the physical and psychological well-being of users besides becoming a much debated issue surrounding its credibility that plays an important role in taking public decision. While this idea triggers crucial possibilities of the improvement of welfare for users and severe safety outcomes, it is by far not significant enough to speak about social media and online credibility as well. Works like Marwick and Caplin's (2018) research on online users behavior regarding privacy, harassment and safety helped understand complexities of relations between individuals while portraying the dominance of risk perception in their relationships. Furthermore Gonzalez & Hancock (2020) point out that the platform design impacts user perceptions of their safety drawing attention to the role of interface elements in influencing interactions and perception.

Active user strategies have a center stage in building secure spaces, as Marwick and Boyd explain by unveiling how people shape their cyberspace presence through every-day tactics of privacy management (Marwick & Boyd, 2017). Citron's (2014) study on the legal and social mechanisms adopted by users to combat abusive online behavior sheds light on enabling user processes in the legacy of safe interactions through social

media realms. Studies done in recent times highlight a rising level of concern with regard to harmful content which includes misinformation as well as hate speech and dangerous behavior that creates the challenge of ensuring safe web practices. The emergence of deepfakes and algorithmically supercharged misinformation adds new levels of difficulty that must be met by strident content moderation practices along with greater resourcefulness on the part of users.

Changing users' behavior, characterized by the growing popularity of private or brief communication functionality, brings forth new issues related to transparency and accountability in protecting safe online spaces.Niche communities and content sharing add layers of complexity to content moderation and safety enforcement, demanding a nuanced approach to address evolving user behaviors and needs.

Social Contract Theory in the Digital Society

Similar to the traditional social contract in physical societies, the digital social contract outlines the terms, conditions, and expectations that shape the nature of online engagement (Gillespie, 2018). When people make accounts and join online groups, they kind of agree to follow the rules of the social media platforms. This agreement creates a common understanding and helps us know what's acceptable and not acceptable in how people act on the internet. Think of it as a moral guide, helping us know what's right and wrong in how we interact and how the platforms are managed This establishes a groundwork, for practicing manners and courteous behavior on the internet. It demonstrates that all individuals, whether users or website operators bear the obligation of ensuring a constructive environment. In this agreement users pledge to refrain from engaging in activities that violate the established guidelines while platforms undertake to safeguard users' safety, privacy and freedom of expression within the boundaries of norms.

The theory of social contract, which is based on the ideas of philosophers, like Hobbes, Locke and Rousseau examines the theoretical agreements that people make to create a sense of community and establish a political framework for common interests. This study applies social contract theory to the world focusing on the agreement, between internet users and online entities regarding the collection and control of sensitive information. When users participate in activities they expect their private data to be handled securely and respectfully. This forms the foundation of an internet based contract. This digital social contract significantly influences users' privacy-related behavior, trust in how internet firms manage data, and perceptions of related risks. Kruikemeier et al. (2019) classify users into categories—worried, carefree, and cautious—using Latent Class Analysis, shedding light on diverse perspectives on the legitimacy of the digital social contract. The study reveals that a significant portion of users consider the digital social contract less dependable, impacting their online behavior, particularly in safeguarding their privacy (Kruikemeier, S., et al., 2019).

A sense of community in the digital society

Exploring the intricate dynamics within online communities reveals the significant influence of content moderation policies on cultivating a shared sense of community. In this complex environment, major platforms, exemplified by Facebook, grapple with the challenging task of balancing user interactions while strategically addressing the widespread issue of disinformation. It becomes apparent how important user-generated content is to actively form these online communities' shared identities. The investigation conducted by Marwick and Lewis (2017) delves into the impact of user-generated content on shared identity, underscoring its profound significance. Gillespie's (2018) study, "Custodians of the Internet," further highlights the importance of transparent communication in content moderation policies to inform and engage the user base. This emphasis on openness becomes increasingly vital as it fosters trust and active participation among users. Moreover, the cross-cultural dimensions observed in research by van Dijck and Poell (2013) illuminate how cultural subtleties influence expectations and communication methods, emphasizing the necessity for culturally sensitive content moderation approaches. Recognizing and navigating these intricate dynamics not only aids in addressing disinformation but also in fostering a resilient and inclusive digital community.

Governing rule on content moderation policies used in Facebook

Facebook is one of the largest social media platforms today. It has billions of users worldwide, including minors, and as a result, Facebook faces a significant challenge in maintaining a safe and respectful environment. To address this, the platform has implemented several governing rules to ensure the safety of

each user, especially for minors, and maintain a respectful online community. According to Henry et al. (2021) that sharing of illegal content, such as hate speech, pornographic images, or films that are uploaded online, is one of the main issues with social media in the digital age, which has over a billion users. The platform introduced content moderation standards that outline what types of content are permitted and prohibited in response to the intricate issues brought up by harmful online content. These standards also aim to safeguard user privacy and prevent the spread of harmful, inaccurate content that could have an adverse effect on specific individuals or groups of people. One of Facebook's guiding principles is its community standards policy. It specifies what kinds of content are permitted and prohibited on the platform. Some instances of content that is prohibited include hate speech, which is defined as the use of abusive language to attack or incite violence against specific individuals or groups of individuals, violent and criminal activity, which includes sharing graphic videos of violence, harassment, and bullying individuals by posting malicious or false information about them, spam, and misleading information, as well as anything that promotes or violates the law by engaging in illicit activities like child exploitation, drug trafficking, or terrorism.

According to research by Ari Kusyanti et al. (2017), users' primary fear is that their personal information may be viewed, shared, and altered without their consent. This suggests that internet users' personal information should be treated with consideration for their privacy. Another important rule in reaction to this is Facebook's data policy, which governs the collection, use, and distribution of user data. By outlining the security measures taken to protect personal data, it allows users to take charge of their privacy settings and manage their data. Facebook has implemented measures to prevent the spread of inaccurate data and fake news. The impact of these governing rules shows mixed results. Some studies suggest that these policies have been effective in reducing the spread of harmful content and improving user safety. However, there are also concerns about the enforcement of these rules, with instances of inconsistent moderation and potential bias. Acknowledging internet users' privacy concerns is crucial for guiding protective measures.

DATA AND METHODOLOGY

The study will utilize qualitative research methods and a case study approach to investigate the application of social contract theory in the digital realm, with Facebook, the leading social network with over 3 billion users globally (Statista, 2023), as a focal point. The data collection process involves a brief review of Facebook's official documentation, including community standards and transparency center reports, which can be accessed at the Facebook Transparency Center. This serves as the foundation for understanding the platform's content moderation policies.

Utilizing a case study methodology, we select diverse instances from individual user experiences to content removals to exemplify the challenges within the digital social contract. Our analytical approach involves a systematic categorization of cases, allowing for an adequate exploration of content moderation stability and user responsibilities. The qualitative data analysis focuses on presenting these cases in a structured manner, offering a sufficient view of the dynamics within Facebook's digital social contract. The findings contribute to a comprehensive understanding of the dynamics surrounding content moderation policies and user responsibilities on Facebook, enriching the discourse on the digital social contract in the realm of social media.

RESULTS AND DISCUSSION

The table below provides a reference guide to understand the policies within Facebook's community standards and guidelines.

Table 1. Overview of Facebook's Community Standards

Policy category	**Sections**	
Violence and criminal behavior	• Violence and incitement •Restricted Goods and Services • Dangerous organizations and individuals	• Fraud and Deception • Coordinating Harm and Promoting Crime

Safety	• Suicide, Self-Injury, and Eating Disorders • Child Sexual Exploitation, Abuse, and Nudity • Privacy Violations	• Human Exploitation • Adult Sexual Exploitation • Bullying and Harassment
Objectionable content	• Hate Speech Violent and Graphic Content •Adult Nudity and Sexual Activity	• Language • Adult Sexual Solicitation and Sexually Explicit
Integrity and authenticity	• Account Integrity and Authenticity Identity Inauthentic Behavior • Spam	• Misinformation • Cybersecurity • Memorialization
Respecting Intelectual Property	• Copyright and Trademark Infringement	• Piracy and Counterfeit Goods
Content- Related Requests and Decisions	• Appeals Process for Content Removed in Error	• Requesting Access to Personal Data

source: https://transparency.fb.com/policies/community-standards/

The study of current issues involving prominent individuals and political leaders on social media platforms is critical for understanding the changing dynamics of online behavior, ethical concerns, and the integration of digital spaces with real-world effects. This study examines six separate Facebook cases, focusing on the Rendon Labador Raid Video Incident, Hun Sen's Cambodia, the acts taken against exiled Chinese businessman Guo Wengui, Goh Meng Seng's COVID-19 related posts, the Fruit Juice Diet video, and the Video Discussing Corruption of Law Enforcement in Indonesia. Each case provides a distinct combination of issues and ethical dilemmas, giving insight on how human acts affect privacy, political debate, public health, and community norms in the digital domain. As we continue through these cases, the results and discussions plan to provide clarity regarding the various aspects of social media, highlighting the necessity of ethical behaviour, adherence to platform standards, and the consequences that arise when powerful persons interact with these platforms.

Case 1: Rendon Labador Raid Video Incident

Rendon Labador, a self-proclaimed motivational speaker and internet influencer, is presently under investigation by the Philippine National Police (PNP) due to a suspected breach of privacy during a police operation against an online lending company in Makati City. Labador extensively documented the operation on his Facebook Live, eliciting adverse reactions from the relatives of the company's employees, who allege that their faces were exposed in Labador's video. Labador's live coverage of the police operation and his subsequent interview with PNP Anti-Cybercrime Group spokesperson Capt. Michelle Sabino is currently under intense scrutiny. Despite the PNP ACG initially granting media access to the operation as a gesture of transparency, concerns regarding a potential breach of privacy have prompted a thorough investigation. The severity of the situation has resulted in the suspension and termination of Labador's media access. This underscores the significance of adhering to ethical standards and established protocols, particularly when dealing with sensitive situations such as police operations. The investigation is not only examining immediate privacy concerns but is also evaluating the broader implications of Labador's actions.

To Facebook policies, it is essential to note that Facebook has strict guidelines regarding the dissemination of sensitive content, including content that invades privacy or violates the rights of individuals. Labador's actions may have breached these policies, contributing to the severity of the consequences he is facing. The outcome of the investigation may have lasting consequences for Labador's role as an influencer and motivational speaker, especially if his actions are deemed in violation of established policies and ethical standards.*(excerpt from:PNP probes vlogger for possible privacy breach during police op (cnnphilippines.com))*

Case 2: HUN SEN'S CAMBODIA Cambodian PM Hun Sen deletes Facebook page after criticism

Cambodian Prime Minister Hun Sen deleted his Facebook account following a six-month ban imposed by the Meta's oversight board. This decision came after the board ruled against him due to a video post where he threatened to have his opponents beaten. Hun Sen, who had over 14 million Facebook followers, announced on his Telegram account that he was discontinuing Facebook usage, citing confusion caused by account impersonation. Despite the board's findings of community standards violation and a history of intimidation against opponents, he asserted his departure was unrelated and preempted Facebook's decision-making period.

The banned video, in which Hun Sen targeted critics accusing his ruling party of electoral fraud, was reported to Meta for inciting violence. While Facebook left the video online due to its "newsworthiness," the oversight board found the decision incorrect and demanded content removal. The board's recommendations also included taking down the prime minister's Facebook and Instagram accounts, citing concerns about human rights violations, political intimidation, and the misuse of social media. The move away from Facebook, weeks before a national election where Hun Sen's party faces no competition, marks a significant setback for the prime minister's extensive use of the platform.

Human Rights Watch Deputy Asia Director Phil Robertson welcomed the oversight board's decision, highlighting Hun Sen being held accountable for using social media to incite violence. Robertson noted the prime minister's shift to Telegram, emphasizing the platform's association with authoritarian leaders and suggesting a departure from Facebook due to being held accountable to community standards. Government spokesperson Phay Siphan, however, rejected claims of incitement, asserting that Telegram was a more popular and effective platform based on government studies.(excerpt from:https://asia.nikkei.com/Spotlight/Hun-Sen-s-Cambodia/Cambodian-PM-Hun-Sen-deletes-Facebook-page-after-criticism)

Case 3: SOCIAL MEDIA Facebook pulls page, limits posting for exiled Chinese tycoon Guo

Following heightened tensions between exiled millionaire Guo Wengui and Chinese authorities, Facebook has taken severe action against the prominent persona. Facebook has deleted a page associated with Guo Wengui and temporarily restricted his ability to post on his profile, citing violations of the platform's community guidelines. These steps were purportedly prompted by the exposure of "personal identifier information" on Guo's profile, a violation that violated Facebook's commitment to user safety and privacy.

At the same time, Chinese prosecutors have stepped up legal proceedings against people linked to companies associated with Guo Wengui. They filed lawsuits against employees of Beijing Pangu Investment and Henan Yuda Real Estate, accusing them of misusing funds, engaging in loan fraud, and accepting bills improperly. These legal actions make the ongoing dispute between Guo Wengui and the ruling Communist Party more complicated. Guo, living in exile, is already facing corruption allegations against top officials, and these recent legal moves add another layer of complexity to the situation.

In conclusion, Facebook has acted decisively against the exiled Chinese millionaire Guo Wengui's profile, claiming personal information disclosure violations, privacy violations and other community standards violations. Guo, who is well-known for accusing the Chinese government of corruption. Facebook's response shows its dedication to user privacy and safety, and it invites users to report any violations to it.(excerpt from:https://jakartaglobe.id/news/china-turns-legal-pressure-exiled-tycoon-guo-xinhua)

Case 4: Facebook's removal of Goh Meng Seng's posts due to violation of its policies on COVID-19 claims.

Facebook has announced that it will and does not allow false claims on its platform that may contribute to the unrest in which may lead to the rejection of COVID-19 vaccines, adamant on its statement by removing such posts. The statement was made in response to the questions from Channel NewsAsia (CNA) after People's Power Party politician Goh Meng Seng stated that that Facebook had taken down "several" of Facebook posts, suggesting government response and intervention in the process. Goh Meng Seng is a political figure in Singapore and he had stated earlier that a video from his People's Power Party page was removed by Facebook without some any sort of justification. A Facebook spokesperson clarified that the company enforces policies against false claims in regards to addressing the removal of Goh's posts, identified by global health experts that could contribute to vaccine rejection or hesitancy.

While Facebook did not specify which posts were removed, Goh Meng Seng had been issued a correction direction from Singapore's Protection from Online Falsehoods and Manipulation Act (POFMA) for

posting content on two Facebook pages connecting COVID-19 vaccination to a doctor's stroke and the cause of death of an 81-year-old man. These claims have been refuted by The Ministry of Health (MOH), stating that there is no reliable evidence supporting a correlation in risk of heart attack or stroke associated with the Pfizer-BioNTech and Moderna COVID-19 vaccines. The ministry clarified that the 81year-old man's cause of death was ischaemic heart disease, which is not connected to vaccination. (excerpt from: https://www.channelnewsasia.com/singapore/facebook-goh-meng-seng-pofma-covid-19-posts-removed-policies-222286)

Case 5: Fruit Juice Diet video

The Oversight Board is currently addressing two connected cases involving Meta's content decisions on a Facebook page in Thailand. The videos in question feature a woman advocating for a fruit juice-only diet, sparking discussions on potential endorsements of unhealthy lifestyles, particularly related to eating disorders.

Posted in 2022, the first video highlights the woman's health journey with the fruit juice-only diet, while a 2023 follow-up discusses her intention to embrace a more extreme diet—living solely on energy. Both videos have garnered attention and concern, leading to multiple reports for violating Meta's Suicide and Self Injury Community Standard.

Users, worried about the potential impact on vulnerable individuals, especially teenagers, have appealed to the Oversight Board. They argue that the content could normalize unhealthy behaviors related to eating disorders. This scrutiny prompts the Oversight Board to assess Meta's content policies and enforcement practices regarding diet, fitness, and content tied to eating disorders on Facebook. Additionally, Facebook's response to these concerns becomes crucial in this analysis. The platform's actions, or lack thereof, in addressing reported violations and user appeals will be examined. This analysis aims to navigate the complexities of moderating content, considering both personal expression and the potential propagation of harmful narratives. It underscores the need for a balanced approach, recognizing the freedom of expression while emphasizing the responsibility of social media platforms to protect user well-being.(excerpt from: https://oversightboard.com/news/238140992424411-oversight-board-announces-fruit-juice-diet-cases-and-a-case-about-violence-in-the-indian-state-of-odisha/?ref=shareable)

Case 6: Video Discussing Corruption of Law Enforcement in Indonesia

In a recent case, Meta's Oversight Board interfered with the company's decision to banish a Facebook post that had addressed corruption inside Indonesia's National Police. The user's video in the form of Bahasa Indonesia criticized the policing force's corrupt practices, emphasizing the promotion of particular people engaged in the corruption. The post was first taken down by Meta under its Violence and Incitement policy, but Meta soon reversed the decision following the intervention of the Oversight Board. The Board pointed to an inconsistency in Meta's implementation of the policy with regards to political metaphorical statements, stressing the need to work on context-dependent moderation systems. The case highlights the need for safeguarding political speech and encourages Meta to improve measures in determining the content within its relevant environment. The fact that the Oversight Board chose to make an exception to its previously established policy in a case of critically discussing governmental power highlights the determination to fairness and understanding the details of content moderation.(excerpt from:Criticism of Indonesian Law Enforcement (fb.com))

Table 2. Case Analysis

Case 1: Rendon Labador Raid Video Incident	
Violation under Facebook Community Standards	Safety under privacy violations

Facebook Action	Suspension that eventually led to the termination of the account.
Author's Point of View	Facebook action demonstrates the platform's dedication to upholding its rules and guidelines for the community. By taking this step, Facebook would demonstrate its commitment to holding users accountable for future breaches and highlight its position on privacy and content standards. It would also demonstrate the platform's dedication to enforcing rules consistently and making
Case 2: HUN SEN'S CAMBODIA Cambodian PM Hun Sen deletes Facebook page after criticism	
Violation under Facebook Community Standards	Violence and criminal behavior under Violence And Incitement
Facebook Action	They gave the Prime Minister a 6 months ban but left the video for its "Newsworthiness".
Author's Point of View	Facebook's response to the issue was in line with their community standards. But Meta's oversight board stated that instead of letting the content as is due to its "Newsworthiness", It is a content in which incites violence and must be taken down and recommended to take down the Prime Minister's account on both Facebook and Instagram.
Case 3: SOCIAL MEDIA Facebook pulls page, limits posting for exiled Chinese tycoon Guo	
Violation under Facebook Community Standards	Safety under privacy violations
Facebook Action	Deletion of Guo Wengui's affiliated page; temporary restriction on his posting privileges.
Author's Point of View	Facebook's firm response shows that they take user safety and privacy seriously. Meanwhile, the legal proceedings make Guo Wengui's situation with the Chinese authorities more complicated, adding complexity to the existing corruption allegations against him.
Case 4: Facebook's removal of Goh Meng Seng's posts due to violation of its policies on COVID-19 claims.	
Violation under Facebook Community Standards	Integrity and authenticity under Misinformation/ Misinformation and Harm
Facebook Action	Facebook had actively removed Goh Meng Seng's posts that could lead to the rejection of COVID-19 vaccines, and removed/deleted several posts of similar nature.
Author's Point of View	Facebook has remover over 12 million contents about COVID-19 in which violates the company's misinformation and harm policy since the start of the pandemic. They are also increasing their prevention against the posts that are of "Sensationalist or alarmist" regardless of whether they violate the policies or not.

Case 5: Fruit Juice Diet video	
Violation under Facebook Community Standards	Safety under Suicide, Self Injury, and Eating Disorders
Facebook Action	The investigation includes a look at Facebook's response to reported infractions and user complaints. The platform's actions, or lack thereof, addressing the content showcasing the fruit juice-only diet will be thoroughly investigated, revealing insight on its dedication to enforcing Community Standards and protecting user safety.
Author's Point of View	Users, particularly concerned about the potential impact on vulnerable individuals, have appealed to the Oversight Board. Their perspective underscores worries that the content could normalize unhealthy behaviors related to eating disorders. The analysis aims to navigate the complexities of content moderation, emphasizing a balanced approach that respects freedom of expression while prioritizing the responsibility of social media platforms to protect user well-being.
Case 6: Video Discussing Corruption of Law Enforcement in Indonesia	
Violation under Facebook Community Standards	Violence and criminal behavior under Violence and Incitement policy
Facebook Action	According to Meta, the company Facebook, Instagram, Threads, and WhatsApp, that after reviewing this case in the oversight board, they've determined that they removed the content in error and will reinstate the post
Author's Point of View	By having an oversight board that allows users to appeal if they disagree with the platform's decision of taking down a content and Meta will review the cases, helps to make the platform a safe place by resolving some difficult questions around freedom of expression. The platform's decision to reinstate the content because they admit that it was removed by error, shows that Facebook is enforcing a fair content moderation policies.

Discussion

The findings of the research align with the objectives of this study, which aimed to apply social contract theory to the digital environment, particularly emphasis on user behavior and the necessity for a digital social contract. Examining the said cases on Facebook reveals important insights into the dynamics of online relationships, ethical problems, and the consequences of significant persons using social media. As an overall response, these cases highlight the crucial role of social media platforms in maintaining community standards in order to ensure user safety and privacy.

The application of social contract theory to digital society involves an understanding of the implicit agreements that individuals make while engaging in online communities. Facebook, as an internationally powerful platform, plays an important role in creating these agreements through its community guidelines. The investigated incidents highlight the critical relevance of complying to these standards, as demonstrated by Facebook's responses to data breaches, encouragement to violence, and content violations. These occurrences are clear indications of the platform's dedication to provide a safe and courteous environment for its various users.

Our results support scholars such as Gillespie (2018) who discussed the digital social contract almost as if it were a sort of social contract that had been established on paper. This idea is spot on with how Facebook enforced its rules in the case studies we conducted. These policies serve as a form of a digital social contract, defining the behavior of the virtual community on the platform. The cases, including those heard at the oversight board, provide practical living cases that demonstrate how Facebook utilizes its content

moderation regulations. In particular, cases in which Facebook admits to mistakes and restores content following review by the oversight board point to the fluid nature of this digital social contract. The platform's commitment to rectify mistakes and ensure a fair content moderation process reflects its dedication to maintaining a balanced and just online environment. Another study by Kruikemeier et al. (2019) categorized users based on how much they trust this digital social contract. This supports what we saw in the cases, where people had different concerns, from privacy to calls for violence.

The cases highlight the issues that big platforms, such as Facebook, encounter when navigating content moderation regulations. Establishing the right balance between user-generated content, building online communities, and combating misinformation becomes a difficult task. Marwick and Lewis (2017) provide insights on the influence of user-generated material on shared identity, highlighting the necessity of transparency and openness in content moderation procedures.

CONCLUSION

In summary, this paper explains the application of social contract theory in the digital society, particularly in Facebook. This investigation reveals insights into user behavior while highlighting the vital role of a digital social contract theory in determining how people interact online. The analyzed cases underscore the challenges of navigating content moderation and fostering a safe online space. Gillespie's (2018) digital social contract concept aligns with our findings, emphasizing the need for agreed-upon rules. Kruikemeier et al.'s (2019) trust categorization mirrors users' varied concerns, from privacy to calls for violence.

Despite these insights, our study is limited to Facebook and specific cases. To broaden our understanding, future research should encompass diverse platforms and a wider array of cases. Recommendations include in-depth data gathering through user interviews to capture more perspectives on digital social contracts, contributing to a more comprehensive grasp of online community dynamics and content moderation challenges.

REFERENCES

Ali, M. N., Almagtome, A. H., & Hameedi, K. S. (2019). Impact of accounting earnings quality on the going-concern in the Iraqi tourism firms. African Journal of Hospitality, Tourism and Leisure, 8(5), 1-12. https://doi.org/10.26668/businessreview/2022.xxxxxxx

Breaching the contract? using social contract theory to explain ... (n.d.). https://www.tandfonline.com/doi/pdf/10.1080/15213269.2019.1598434

Cambodian PM Hun Sen deletes Facebook page after criticism … (n.d.). https://asia.nikkei.com/Spotlight/Hun-Sen-s-Cambodia/Cambodian-PM-Hun-Sen-deletes-Facebook-page-after-criticism

Castets, R. C., (2020). Algorithmic Content Moderation on Social Media in EU Law: Illusion of Perfect Enforcement http://dx.doi.org/10.2139/ssrn.3535107 Henry, N., Witt, A. (2021).

China turns up legal pressure on exiled tycoon Guo: Xinhua. Jakarta Globe. (n.d.). https://jakartaglobe.id/news/china-turns-legal-pressure-exiled-tycoon-guo-xinhua.

CNA. (2021, April 16). Facebook says it removed Goh Meng Seng's posts as they violated its policies on COVID-19 claims.https://www.channelnewsasia.com/singapore/facebook-goh-meng-seng-pofma-covid-19-posts-removed-policies-222286

CNBC. (2017, October 2). Facebook pulls page, limits posting for exiled Chinese tycoon Guo. CNBC.https://www.cnbc.com/2017/10/02/guo-wengui-facebook-pulls-page-limits posting for-exiled-chinese-tycoon.html

CNN Philippines. (2023, October 23). PNP probes vlogger for possible privacy breach during police op https://www.cnnphilippines.com/news/2023/10/23/PNP-probes-vlogger Rendon-Labador.html

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

Eleni, A., Christiana, V. et al. (2020) Combating misinformation online: Re-imagining social media for policy-making doi:10.14763/2020.4.1514

Facebook pulls page, limits posting for exiled Chinese tycoon Guo ... (n.d.). https://www.reuters.com/article/us-china-facebook-tycoon/facebook-pulls-page-limits-posting-for-exiled-chinese-tycoon-guo-idUSKCN1C70R0

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

Gonzalez, A. C., & Hancock, J. T. (2020). Mirror, mirror on the wall: The effects of Instagram use on selfie-related behavior and body image. Cyberpsychology, Behavior, and Social Networking, 23(2), 113-123.

Governing Image-Based Sexual Abuse: Digital Platform Policies, Tools, andPractices" The Emerald International Handbook of Technology-Facilitated Violence and Abuse (Emerald Studies In Digital Crime, Technology and Social Harms), Emerald Publishing Limited, https://www.emerald.com/insight/content/doi/10.1108/978-1-83982-848-520211054/full/html https://doi.org/10.48550/arXiv.2101.04618

Johnson, D. G. (2019). The Internet of Garbage. MIT Press. Kusyanti, A., Puspitasari, D. R., Harin, C., Yustiyana, A. L. (2017). Information Privacy Concerns on Teens as Facebook Users in Indonesia https://doi.org/10.1016/j.procs.2017.12.199

Marwick, A., & boyd, d. (2017). Networked privacy: How teenagers negotiate context in social media. New Media & Society, 20(8), 1-17.Citron, D. K. (2014). Hate crimes in cyberspace. Harvard University Press.

Marwick, A., & Caplan, R. (2018). "Trouble on the 'left'" in social media: Exploring attitudes towards online safety among feminist and LGBTQ activists. New Media & Society, 20(8), 2820-2837.

Oversight Board announces Fruit Juice Diet Cases and a case about violence in the Indian state of Odisha. Oversight Board. (n.d.). https://oversightboard.com/news/238140992424411-oversight-board-announces-fruit-juice-diet-cases-and-a-case-about-violence-in-the-indian-state-of-odisha/?ref=shareable

Thero, H., Vincent, M. (2021). Investigating Facebook's interventions against accounts that repeatedly share misinformation https://doi.org/10.1016/j.ipm.2021.102804

Transparency center. Case Bundle Featuring Videos Making Claims About a Fruit Juice-Based Diet. (n.d.). https://transparency.fb.com/oversight/oversight-board-cases/fruit- juice-diet

Tufekci, Z. (2018). Twitter and Tear Gas: The Power and Fragility of Networked Protest.

Yale University Press.van Dijck, J., & Poell, T. (2013). Understanding Social Media Logic.Media and Communication, 1(1), 2-14.

Yi Liu, T., Yildirim P., Zhang, Z. J (2021). Social Media, Content Moderation, and Technology