Establishing Trustworthiness of AI-Driven Content on Social Media: Practices and Considerations

Eun Jeong Kang
Cornell University
New York, USA
ek646@cornell.edu

ABSTRACT

Social media platforms have faced the challenge of effectively ensuring the trustworthiness of AI-driven content (e.g., synthetic content, altered content) spread in their spaces to the public. While community principles to handle AI-driven content have been suggested, platforms' general approaches to and considerations for these emerging topics, as the authorities that implement governance, remain unclear. This study explores the procedures platforms use to ensure trustworthiness of AI-driven content and identifies the factors that enable platforms to establish AI-driven content principles. Through critical discourse analysis of content outlining policies, we identified that platforms are guided by their existing guidelines to distinguish and moderate AI-driven content from general content. Additionally, user spontaneity and the integration of stakeholder values into decision-making are essential for effectively fostering trust in content among users. We underscore the significance of designing user experiences that encourage social media users to take ownership of creating trustworthy AI-driven content.

CCS CONCEPTS

• **Social and professional topics → Computing / technology policy**.

KEYWORDS

social media platform, policy, synthetic content, Trustworthy, AI

ACM Reference Format:
Eun Jeong Kang. 2018. Establishing Trustworthiness of AI-Driven Content on Social Media: Practices and Considerations. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 7 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Generative AI (GenAI) has become more easily accessible, enabling laypeople to create content using AI on social media. At the same time, it raises concerns about the proliferation of misinformation and violations of privacy and rights [21]. Recognizing that this could potentially undermine the authenticity and credibility of social media [6], social media platforms, such as TikTok, have been dedicated to identifying, moderating, and regulating user-generated content that uses AI through policy [26]. In addition, they have encouraged users to responsibly clarify the use of AI in creating their content.

Scholars have proposed methods to communicate the creation of responsible, accountable, and explainable AI-driven content with users [9, 14, 30]. For instance, watermarks or visual markers might verify the content creator who owns AI-generated images [30]. However, from a platform's perspective, social media platforms should consider how such designs can be technically embedded in their platform infrastructure to enact governance [11]. Platforms play a central role in content distribution, cultural space creation, and economic value generation [15, 27]. Therefore, determining methods to establish the credibility of AI-driven content, such as content moderation rules, can be a complex process due to the large-scale interaction and the various stakeholders benefiting from these platforms [29]. For instance, some platforms might need to decide what to label as 'AI-generated content' to disclose AI usage, taking into account content creators who earn revenue from it. The practices of authorities to ensure the trustworthiness of AI-driven content and their role as content intermediaries are unclear.

Understanding policy can provide insights for navigating the trajectory of platforms' decisions for moderating content. Policy, defined as rules, principles, and guidelines, forms the basis for decision-making [13, 32, 37]. It not only sparks discussion on creating a regulatory framework but also impacts user well-being and stakeholder decisions, thereby fostering community growth [17, 18]. In this context, examining how policies influence human-system interaction can open up new avenues in HCI research. This can also help to understand which socio-technical values are publicly recognized and how evidence from HCI knowledge can contribute to the design of user-centered policies [20, 34, 37].

We investigated the principles social media platforms adhere to in establishing the trustworthiness of AI-driven content on their platforms, and what factors have been considered in relation to stakeholders and users of their platforms in constructing the principles. Through critical discourse analysis [12] of content outlining their community principles and policies (e.g., community guidelines), we traced how they have expanded their policies to solidify their stance and sought to understand the gray areas that platforms still confront in moderating all stakeholders' values as neutral intermediaries.

We found that platforms typically adhere to community guidelines when establishing management rules for AI-driven content creation. They distinguish AI-generated content through disclosure, relying on content creators' voluntary disclosure and visual labels. In their decision-making process, they have considered how

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

to balance the values of stakeholders and users, considering who could benefit from AI-driven content moderation. Based on the preliminary study, we highlight the importance of considering the stance of platforms as content intermediaries and the diverse social contexts of users in designing responsible AI systems.

2 RELATED WORK

2.1 Social Media Platforms that Ensure Trustworthiness of AI-driven content

The use of AI to generate images, text, and audio has become pervasive on social media, as Generative AI becomes more accessible to the general public [23]. This involves creating synthetic media content, such as creating realistic videos that impersonate public figures and enhancing or altering parts of the content (AI-altered content) [28]. In this paper, we refer to any type of content for which a content producer can potentially utilize AI in their workflow as 'AI-driven content'. While AI produces opportunities for social media users to be innovative in creative expression [1, 25], it also presumably increases fraudulent activities and societal issues that cause ethical problems [4, 7, 35]. Platforms and relevant organizations have established community standards for regulating AI-driven content and developed an appropriate AI infrastructure [26]. For example, YouTube announced a guideline indicating how they responsibly manage potential AI problems [3].

HCI communities have identified problems that might be caused by AI, and a number of studies have proposed designing AI systems responsibly to mitigate the problems [5, 24]. This ranges from increasing awareness of datasets [14] to involving multiple stakeholders' viewpoints to identify ambiguous problems [36]. While these strategies may help tackle anticipated issues, platforms make decisions and implement governance by considering various associated values [19]. They function as both computational systems and repositories for cultural production and economic resources [15, 29]. As content intermediaries, their practices and the considerations behind them can provide insights into the challenges platforms face in regulating AI-driven content and what should be effectively designed for responsible AI tools. Thus, we investigated social media platforms' current practices and their considerations in constructing principles for AI-driven content.

2.2 Policy Research in HCI and Its Implication on Responsible Platforms

Policies are often established as principles, guidelines, and rules by governments, organizations, or institutions to align with the values of a broad audience or society [37]. At first glance, HCI and policy may seem unrelated, but scholars have emphasized the potential synergy HCI research and policy-making can create in designing responsible systems that address societal problems [16, 18, 31, 37]. The design knowledge created by HCI research can inform the policy-making process so that it resonates with users, which allows policy-makers to consider the viewpoints of potential stakeholders when formulating policies [22, 34]. On the other hand, policy can also provide insights for HCI researchers and platform designers to consider what regulatory authorities and conflicts exist in design decisions [18]. HCI researchers have studied how policies have been constructed across platforms for content moderation. Jiang et al. [20] conducted an analysis of the rules outlined in the community guidelines of social media platforms, highlighting the common types of misbehavior across these platforms. Schaffner et al. [31] critically analyzed platform policy texts to understand how these platforms define, describe, and implement content moderation. These researchers implied that community standards to moderate AI-driven content have recently been established [31, 37], but the practices of platforms to establish principles for trustworthy AI-driven content remain unexplored. Therefore, we examined the policies of these platforms regarding AI-driven content.

3 METHODOLOGY

3.1 Data Collection

We gathered data from social media platforms regarding their guidelines, principles, and any official posts about how they manage AI-driven content. We defined the range of AI-driven content as any type of content (e.g., sound, image) that uses AI in production. First, we selected keywords related to AI, such as 'Artificial Intelligence', 'AI', 'Synthetic', and 'DeepFake'. We then used the keywords to search for web pages detailing policies related to AI-driven content. We followed the guidance of Schaffner et al. [31] to locate relevant policy web pages. These include community guidelines, standards, and official posts explaining their methods for processing synthetic content. We selected the top 10 globally popular social media platforms [2] and intentionally included two additional platforms to ensure a balanced representation of case characteristics (Appendix A). Additionally, since policies and design components can be frequently updated [27], we gathered previous records of the web pages from the last three years (May 2021–May 2024), using the Wayback Machine[1]. We excluded content related to AI features of the platforms. After filtering data that was not relevant to the research, we finally collected 38 pieces of web content[2].

3.2 Data Analysis

We used the critical discourse analysis method [10], an analysis that examines the intricate relationships between communication and other factors. This method aids in understanding data narratives deeply by connecting them to societal discourses referenced by external data sources [12]. We chose this method because the policy-making process is constructive and discursive, and platforms have often updated their policies [38]; social media platforms' internal directions can be changed based on external events and the reactions of users. To interpret high-level discourse and narratives of data points, we utilized media press articles related to the data points.

After reviewing the content, the author focused on its narratives, considering: 1) 'the target audience', 2) 'the reason for announcing this guideline or introducing design features', 3) 'the topics covered in the content', and 4) 'the terminologies most frequently used in the content'. Then, the author inductively developed codes from the articles while reading the raw data. At the same time, we aligned

[1]https://archive.org/
[2]The list of web content utilized for the analysis can be accessed at https://bit.ly/3KtByDM

Themes	Definition	Platforms associated with the themes
Defining the scope of non-permissible AI-Driven content guided by community standards	Designate what AI-driven content not to create upon existing policies (community guidelines, standards)	X, YT, TT, IG, FB, TWTCH, LD, SNP, WHT, QUR, RDD
Disclosing AI use in publishing content	Encourage content creators to disclose if they used AI for content creation when they publish content in the platforms	YT, TT, IG, FB
Displaying Content with identifiable labels	Content identified as 'AI-driven', either through self-disclosure or automatic detection, displays with identifiable labels designated by platforms	YT, TT, IG, FB, WHT

Table 1: Platforms' Practices to Establish Trustworthiness of AI-driven Content

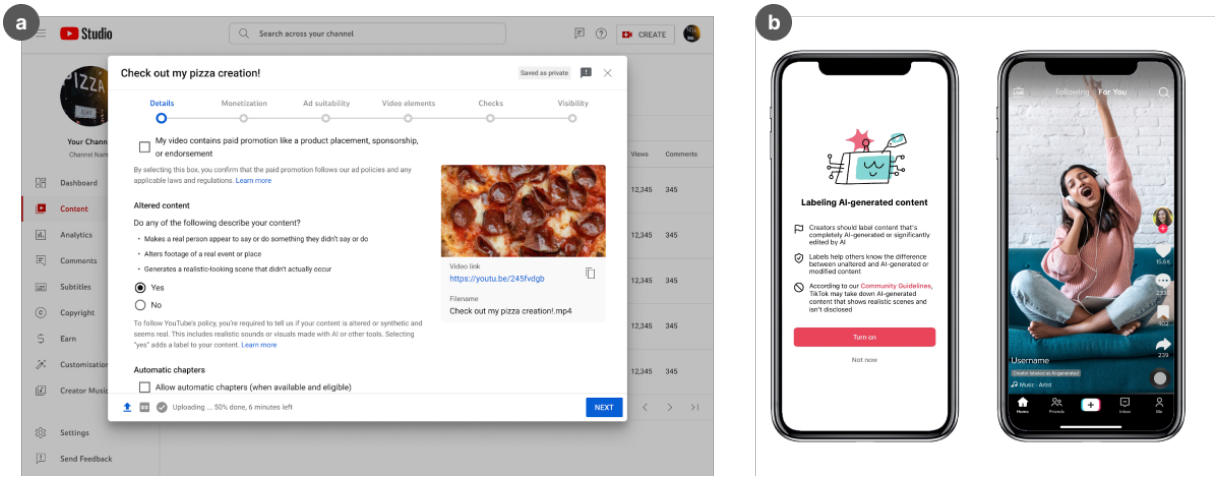

Figure 1: a) YouTube's interface where creators can disclose AI use when publishing their content; b) TikTok's interface displaying a label that discloses AI-generated content

other data published before and after each data point to understand the context of the content and any differences or similarities between the content. After several iterations, the author selected codes, then synthesized codes to find overarching themes, guided by the RQs.

4 FINDINGS

4.1 RQ1. Platforms' Practices to Establish Trustworthiness of AI-driven Content

4.1.1 Defining the Scope of AI-Driven Content Guided by Community Standards. All AI-driven content should adhere to community standards established by platforms, and if these rules are violated, strict regulations are enforced. These include privacy infringement, harassment, and incitement that can deceive others and violate others' rights. The range of AI-driven content regulated is not limited to content fully generated by AI ('AI-generated content', 'synthetic content' and 'deepfake'). Platforms also use terms like 'altered', 'manipulated', 'AI-assisted' and 'misleading media' to indicate that any level of AI used for content creation can be subject to regulation of potential issues. Synthetic content, which is fully created with AI, is strictly removed if it impersonates real people or manipulates events in ways that risk misinformation, particularly when creators deliberately make it realistic. All platforms strongly emphasized not creating such content through their guidelines and official announcements, specifying what is not to be posted.

Creators are guided by general guidelines to disclose if they used AI for the content. The level of detail specifying what creators are permitted to create differs across platforms. A video-sharing platform, YT, has provided specific examples of modified or synthetic content that necessitates disclosure. In this case, AI-driven content that does not require disclosure includes content that uses AI for minor purposes (such as ideation) or caption creation. On the other hand, some platforms, such as LD, provide guidelines that creators

can more flexibly interpret, such as providing best practices for using AI responsibly in content.

4.1.2 Disclosing AI Use in Publishing Content. Under the guideline for AI-driven content (Section 4.1.1), creators should spontaneously disclose AI-driven content, following the disclosure process when they publish content. Platforms, especially video and image-sharing platforms (YT, IG, FB, TT), have provided a feature that creators can use to disclose during the upload procedure, encouraging creators to voluntarily disclose using this option. Creators can set an option to declare whether their content is altered or synthetic content made by AI when they upload reels, shorts, or posts. This also applies to advertisers who promote products using AI-driven content. Apart from disclosure provided by platforms, in TT, creators are also allowed to disclose the fact in any way on their own (e.g., caption). To assist understanding of what to disclose, a platform (YT) provides general rules alongside the feature to help creators judge when their AI-driven content needs to be disclosed.

4.1.3 Displaying Content with Identifiable Labels. Content identified as 'AI-driven', either through self-disclosure or automatic detection, is displayed with visible labels designed by platforms to enable viewers to identify the fact that AI is used for the content. The design of the labels, which content receives them, and what information is introduced with the labels vary across platforms. FB and IG use 'Made with AI' while TT uses 'AI-generated' if the content is identified as 'AI-driven content'. YT uses more descriptive wording, 'How this content was made – altered or synthetic content', to indicate AI-driven content. TT distinguishes between content automatically detected as AI-generated and content that the creator has disclosed as AI-generated. In some cases, when platforms have content utilizing their own AI features (e.g., AI-generated sticker), they use distinct visual markers to differentiate it from user-generated content. FB and IG distinguish AI-driven content made with their own AI features using a different label, 'Imagined with AI', while AI-driven content from other sources uses the label 'Made with AI'.

4.2 RQ2. Factors Social Media Platforms Consider in Establishing Trustworthiness of AI-driven Content in Community Standards

4.2.1 Being Responsible for Stakeholder Values in Breadth. To establish universally accepted criteria as community standards, platforms collaborate with other social media platforms, relevant third-party application companies, and media companies. By continuing to collaborate with other social media companies and AI companies, platforms seek consistency with other platforms in detecting the metadata of AI-driven content. They share information like metadata and criteria that guide algorithms in detecting AI. This ensures that AI-driven content is uniformly regulated, even when users share it across different platforms. A prime example is the Coalition for Content Provenance and Authenticity (C2PA) project[3], which constructs technical standards for media content certification. Media companies promote AI literacy to enhance user awareness of how AI content and visual labels function.

[3]https://c2pa.org/

Platforms also incrementally partner with non-expert industries to incorporate their values, working with them to gain feedback when establishing moderation rules. Many domain experts also work on social media platforms; music artists, in particular, utilize platforms to promote their songs and interact with fans.

4.2.2 Encouraging Users to be Responsible. Automatically moderating AI-driven content using technology is still in development. While platforms aim to automate moderation technically, the process of creating AI-driven content is so complex that the technologies may not always accurately understand users' contexts. For instance, algorithms are designed to automatically flag incorrect content, but they may yield false positives due to insufficient data, potentially leading to the unjust punishment of some individuals. To correct this, platforms encourage users to audit by giving feedback such as reporting misidentified content. When outlining identifiable label or flag guidelines, platforms warn that the algorithms used in automatic technologies may be prone to errors and are not yet comprehensive. They emphasize the need for user involvement.

In disclosing AI-driven content, creators' spontaneity is important, as disclosures are later used to strengthen automatic detection technologies. Some creators may neglect this process or disregard the importance of AI disclosure. Creators may be reluctant to actively participate in AI disclosure due to potential algorithmic repercussions of AI disclosure. Platforms emphasize the 'responsibility of creators', implying that creators should cooperate to maintain a safe and innovative platform space. Otherwise, creators who fail to disclose can face penalties. In addition, platforms openly share the potential impacts of AI disclosure, about which creators may otherwise form folk theories.

4.2.3 Balancing Trade-offs between Users' Benefits. Determining the permissible boundary for AI-driven content, considering its potential benefits for users, can be a delicate task. For some users, like creators, expanding the criteria of AI-driven content might be helpful in expressing their creativity. On the other hand, opening the criteria freely could potentially harm other users. For example, although allowing content with artificial nudity might incentivize some video streamers, the platform (TWTCH) withdrew this policy due to concerns raised about the potential creation of pornographic content.

In addition, identifiable label designs can also affect how users perceive AI-driven content [33]. The design of labels can reduce trustworthiness and perceived ownership of content creators, whereas nuanced descriptions in them may fail to convey their meaning to audiences. In the case of TT, the platform chose to explicitly state 'AI' in any content relevant to AI so users clearly understand the source of the content.

4.2.4 Sustaining Responsive Communication with Users. Platforms keep users updated about which features have been applied through their official press channels. As the types of AI-driven content expand and associated issues continue to arise, platforms are demonstrating how they responsibly address these problems. Platforms often have their own official channels (safety center, creator studio) introducing new features or recent principles that reflect their current situation.

Themes	Definition	Platforms associated with the themes
Being responsible for stakeholder values in breadth	Collaborate with other social media platforms, relevant third-party application companies, and media companies to incorporate their values into principles	YT, TT, IG, FB
Encouraging users to be responsible	encourage users to audit by giving feedback such as reporting misidentified content	YT, TT, LD, SNP
Balancing trade-offs between users' benefits	Determine the permissible boundary for AI-driven content, considering its potential benefits for users	TWTCH
Sustaining Responsive Communication with Users	Maintain their updates about what features were applied through their official media press	YT, TT, LD, TWTCH, IG, FB, X, SNP

Table 2: Factors social media platforms consider in establishing trustworthiness of AI-driven content in community standard

5 DISCUSSION

Our goal is to comprehend the strategies and considerations platforms have in establishing trustworthiness of AI-driven content on their platforms. We do this by examining community guidelines and relevant posts that establish trustworthiness of AI-driven content. Our analysis highlights that platforms, as content intermediaries, construct ways to practice their policies with their users and commit to scoping the moderation of AI-driven content. Here, we underscore potential implications for future exploration.

First, our findings underscore the importance of communication methods that effectively translate policies into forms users can understand and engage with. If users fail to comprehend the need for platforms to moderate AI-driven content, such as the importance of disclosing AI usage, the methods used to implement these policies may not be as effective. Researchers have suggested ways to make AI-driven content accountable and trustworthy [9, 14]. To execute these methods for social media platforms, it is crucial to understand whether users comprehend the process, what the potential impacts might be across different user groups, and how users can be engaged in these methods. The trustworthiness of AI-driven content may not be established solely through user experience design elements such as AI-generated labels or documentation. Instead, it may require a more interactive process that encourages users to actively engage with the activities as well as users' agreement. As shown in Section 4.1, to display AI-generated content, platforms design how they explain the principle to content creators to encourage engagement.

Considering the unique characteristics of various platforms and the social context of the users on platforms is also integral to communicating the trustworthiness of AI-driven content. For instance, the management of AI-driven content may differ between video-sharing and chat-based platforms. Additionally, the cues that instill trust in the content may vary among users of each platform [24]. While platforms have adopted a generic approach allowing creators to disclose AI-driven content by providing an option during upload, we believe a more engaging and standardized design could be proposed to facilitate appropriate content disclosure.

Our findings also suggest that platforms' decisions in moderating AI-driven content can create new experiences for user groups, particularly those incentivized by the platform's situation. Creators might consider not only content production but also the use of AI technologies and their appropriateness within platform-defined principles to maintain their benefits. Viewers, on the other hand, might distinguish facts from synthetic AI-driven content using cues provided by the platforms. We anticipate that a new workflow to optimize creators' content with the rules and folk theories of how AI-driven content affects their monetization can be created, as they try to understand how platform algorithms work [8]. In addition, understanding how viewers perceive trustworthiness from visual cues, such as certification labels, could offer insights for platforms on what to present to enhance viewer awareness [33].

Finally, we noticed that social media platforms have varying levels of detail in community guidelines describing how they manage AI-driven content. Some platforms, which are mostly video-sharing platforms, provide detailed criteria for which content is permissible, while other platforms only guide users to follow community guidelines, not even adopting explicit visual labels on AI-driven content. Although there has been organizational commitment to uniformly set guidelines across platforms to moderate AI-driven content (C2PA), platforms may have inherent strategies or mechanisms in deciding principles, which reduce transparency [31].

5.1 Limitations

We recognize the limitations of our research. First, our content sources are predominantly from specific platforms such as YouTube, TikTok, and Meta, which may introduce a bias towards video-sharing platforms. This could be attributed to these platforms having the largest user base and being primarily video-centric social media channels.

6 CONCLUSION

In conclusion, this study explored the procedures adopted by social media platforms to communicate the trustworthiness of AI-driven content, such as synthetic or altered content, to the public effectively. Through critical discourse analysis of policies, it was found that platforms are guided by their existing guidelines to distinguish and moderate AI-driven content from general content. However, fostering trust in AI-driven content among users requires not only platforms' efforts but also user spontaneity and the integration of stakeholder values into decision-making processes. The study underscores the significance of designing user experiences that encourage social media users to take ownership of creating trustworthy AI-driven content. Effective communication of trustworthiness in AI-driven content necessitates a collaborative approach involving platforms, users, and stakeholders.

REFERENCES

[1] Meta [n. d.]. *Introducing New AI Experiences Across Our Family of Apps and Devices.* Meta. https://about.fb.com/news/2023/09/introducing-ai-powered-assistants-characters-and-creative-tools/

[2] [n. d.]. List of Social Platforms with at Least 100 Million Active Users. https://en.wikipedia.org/w/index.php?title=List_of_social_platforms_with_at_least_100_million_active_users&oldid=1224784689

[3] [n. d.]. *New Disclosures and Labels for Generative AI Content on YouTube - YouTube Community.* https://support.google.com/youtube/thread/264550152/new-disclosures-and-labels-for-generative-ai-content-on-youtube?hl=en

[4] Esma Aïmeur, Sabrine Amri, and Gilles Brassard. 2023. Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining* 13, 1 (2023), 30.

[5] Glen Berman, Nitesh Goyal, and Michael Madaio. 2024. A Scoping Study of Evaluation Practices for Responsible AI Tools: Steps Towards Effectiveness Evaluations. In *Proceedings of the CHI Conference on Human Factors in Computing Systems.* 1–24.

[6] Jasper David Brüns and Martin Meißner. [n. d.]. Do You Create Your Content Yourself? Using Generative Artificial Intelligence for Social Media Content Creation Diminishes Perceived Brand Authenticity. 79 ([n. d.]), 103790. https://doi.org/10.1016/j.jretconser.2024.103790

[7] Canyu Chen and Kai Shu. 2023. Combating Misinformation in the Age of LLMs: Opportunities and Challenges. *arXiv preprint arXiv: 2311.05656* (2023).

[8] Yoonseo Choi, Eun Jeong Kang, Min Kyung Lee, and Juho Kim. 2023. Creator-friendly Algorithms: Behaviors, Challenges, and Design Opportunities in Algorithmic Platforms. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems.* 1–22.

[9] Anamaria Crisan, Margaret Drouhard, Jesse Vig, and Nazneen Rajani. [n. d.]. Interactive Model Cards: A Human-Centered Approach to Model Documentation. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul Republic of Korea, 2022-06-21). ACM, 427–439. https://doi.org/10.1145/3531146.3533108

[10] Elaine Czech, Ewan Soubutts, Rachel Eardley, and Aisling Ann O'Kane. [n. d.]. Independence for Whom? A Critical Discourse Analysis of Onboarding a Home Health Monitoring System for Older Adult Care. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg Germany, 2023-04-19). ACM, 1–15. https://doi.org/10.1145/3544548.3580733

[11] Laura DeNardis and Andrea M Hackl. 2015. Internet governance by social media platforms. *Telecommunications Policy* 39, 9 (2015), 761–770.

[12] Norman Fairclough. [n. d.]. *Critical Discourse Analysis: The Critical Study of Language* (2. ed., [nachdr.] ed.). Routledge.

[13] Terry Flew and Rosalie Gillett. 2021. Platform policy: Evaluating different responses to the challenges of platform power. *Journal of Digital Media & Policy* 12, 2 (2021), 231–246.

[14] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. [n. d.]. Datasheets for Datasets. 64, 12 ([n. d.]), 86–92. https://doi.org/10.1145/3458723

[15] Tarleton Gillespie. [n. d.]. The Politics of 'Platforms'. 12, 3 ([n. d.]), 347–364. https://doi.org/10.1177/1461444809342738

[16] Barbara Grimpe, Mark Hartswood, and Marina Jirotka. [n. d.]. Towards a Closer Dialogue between Policy and Practice: Responsible Design in HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto Ontario Canada, 2014-04-26). ACM, 2965–2974. https://doi.org/10.1145/2556288.2557364

[17] Robert Britt Horwitz. 1989. *The irony of regulatory reform: The deregulation of American telecommunications.* Oxford University Press, USA.

[18] Steven J. Jackson, Tarleton Gillespie, and Sandy Payette. [n. d.]. The Policy Knot: Re-Integrating Policy, Practice and Design in Cscw Studies of Social Computing. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Baltimore Maryland USA, 2014-02-15). ACM, 588–602. https://doi.org/10.1145/2531602.2531674

[19] Michael G Jacobides, Carmelo Cennamo, and Annabelle Gawer. 2018. Towards a theory of ecosystems. *Strategic management journal* 39, 8 (2018), 2255–2276.

[20] Jialun 'Aaron' Jiang, Skyler Middler, Jed R. Brubaker, and Casey Fiesler. [n. d.]. Characterizing Community Guidelines on Social Media Platforms. In *Companion Publication of the 2020 Conference on Computer Supported Cooperative Work and Social Computing* (Virtual Event USA, 2020-10-17). ACM, 287–291. https://doi.org/10.1145/3406865.3418312

[21] Ignas Kalpokas and Julija Kalpokiene. 2021. Synthetic media and information warfare: Assessing potential threats. *The Russian Federation in Global Knowledge Warfare: Influence Operations in Europe and Its Neighbourhood* (2021), 33–50.

[22] Hyunsoo Lee, Yugyeong Jung, Hei Yiu Law, Seolyeong Bae, and Uichin Lee. 2024. PriviAware: Exploring Data Visualization and Dynamic Privacy Control Support for Data Collection in Mobile Sensing Research. In *Proceedings of the CHI Conference on Human Factors in Computing Systems.* 1–17.

[23] Dave Lewis and Joss Moorkens. 2020. A rights-based approach to trustworthy AI in social media. *Social Media+ Society* 6, 3 (2020), 2056305120954672.

[24] Q Vera Liao and S Shyam Sundar. 2022. Designing for responsible trust in AI systems: A communication perspective. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency.* 1257–1268.

[25] Yao Lyu, He Zhang, Shuo Niu, and Jie Cai. 2024. A Preliminary Exploration of YouTubers' Use of Generative-AI in Content Creation. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems.* 1–7.

[26] Filippo Menczer, David Crandall, Yong-Yeol Ahn, and Apu Kapadia. [n. d.]. Addressing the Harms of AI-generated Inauthentic Content. 5, 7 ([n. d.]), 679–680. https://doi.org/10.1038/s42256-023-00690-w

[27] David B Nieborg and Anne Helmond. [n. d.]. The Political Economy of Facebook's Platformization in the Mobile Ecosystem: Facebook Messenger as a Platform Instance. 41, 2 ([n. d.]), 196–218. https://doi.org/10.1177/0163443718818384

[28] Reza Arkan Partadiredja, Carlos Entrena Serrano, and Davor Ljubenkov. 2020. AI or human: the socio-ethical implications of AI-generated media content. In *2020 13th CMI Conference on Cybersecurity and Privacy (CMI)-Digital Transformation-Potentials and Challenges (51275).* IEEE, 1–6.

[29] Thomas Poell, David B Nieborg, and Brooke Erin Duffy. 2021. *Platforms and cultural production.* John Wiley & Sons.

[30] Francesco Regazzoni, Paolo Palmieri, Fethulah Smailbegovic, Rosario Cammarota, and Ilia Polian. 2021. Protecting artificial intelligence IPs: a survey of watermarking and fingerprinting for machine learning. *CAAI Transactions on Intelligence Technology* 6, 2 (2021), 180–191.

[31] Brennan Schaffner, Arjun Nitin Bhagoji, Siyuan Cheng, Jacqueline Mei, Jay L Shen, Grace Wang, Marshini Chetty, Nick Feamster, Genevieve Lakier, and Chenhao Tan. 2024. " Community Guidelines Make this the Best Party on the Internet": An In-Depth Study of Online Platforms' Content Moderation Policies. In *Proceedings of the CHI Conference on Human Factors in Computing Systems.* 1–16.

[32] Rebecca Scharlach, Blake Hallinan, and Limor Shifman. 2023. Governing principles: Articulating values in social media platform policies. *new media & society* (2023), 14614448231156580.

[33] Nicolas Scharowski, Michaela Benk, Swen J. Kühne, Léane Wettstein, and Florian Brühlmann. [n. d.]. Certification Labels for Trustworthy AI: Insights From an Empirical Mixed-Method Study. In *2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago IL USA, 2023-06-12). ACM, 248–260. https://doi.org/10.1145/3593013.3593994

[34] Anne Spaa, Abigail Durrant, Chris Elsden, and John Vines. [n. d.]. Understanding the Boundaries between Policymaking and HCI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow Scotland Uk, 2019-05-02). ACM, 1–15. https://doi.org/10.1145/3290605.3300314

[35] Rebecca Umbach, Nicola Henry, Gemma Faye Beard, and Colleen M. Berryessa. [n. d.]. Non-Consensual Synthetic Intimate Imagery: Prevalence, Attitudes, and Knowledge in 10 Countries. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu HI USA, 2024-05-11). ACM, 1–20. https://doi.org/10.1145/3613904.3642382

[36] Steven Umbrello. 2020. Imaginative value sensitive design: Using moral imagination theory to inform responsible technology design. *Science and Engineering Ethics* 26, 2 (2020), 575–595.

[37] Qian Yang, Richmond Y. Wong, Steven Jackson, Sabine Junginger, Margaret D. Hagan, Thomas Gilbert, and John Zimmerman. [n. d.]. The Future of HCI-Policy Collaboration. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu HI USA, 2024-05-11). ACM, 1–15. https://doi.org/10.1145/3613904.3642771

[38] Philippe Zittoun. 2009. Understanding policy change as a discursive problem. *Journal of Comparative Policy Analysis* 11, 1 (2009), 65–82.

A PLATFORM CODES

- TT: TikTok
- YT: YouTube
- IG: Instagram
- FB: Facebook
- LD: LinkedIn
- TWTCH: Twitch
- SNP: Snapchat
- TEL: Telegram
- WHT: WhatsApp
- QUR: Quora
- RDD: Reddit
- X: X (Formerly Twitter)