Vol. 2 No. 4 Edisi Oktober 2025 E. ISSN. 3032-2472
Multidisciplinary Indonesian Center Journal (MICJO)
Journal page is available to
https://e-jurnal.jurnalcenter.com/index.php/micjo
Email: admin@jurnalcenter.com

PSYCHOLOGICAL MECHANISMS BEHIND THE ACCEPTANCE OF DEEPFAKE-BASED HUMOR AND DIGITAL HARASSMENT

MEKANISME PSIKOLOGIS DI BALIK PENERIMAAN HUMOR BERBASIS DEEPFAKE DAN PELECEHAN DIGITAL

Kurrota Aini[1*], Vidya Nindhita[2], Hapsari Puspita Rini[3]

[1*]Universitas Trunodjoyo Madura, Email: kurrota.aini@trunojoyo.ac.id
[2]Universitas Trunodjoyo Madura
[3]Universitas Trunodjoyo Madura

*email koresponden: kurrota.aini@trunojoyo.ac.id

DOI: https://doi.org/10.62567/micjo.v2i4.2093

Abstract

The objective of this study was to examine how deepfake-based humor becomes socially acceptable despite its potential to function as digital harassment. This study focused on psychological mechanisms that explain audience tolerance and normalization of harmful, identity-based humorous content in online environments. This study used a scoping review design to map and synthesize existing research across psychology, media studies, and cyberpsychology. The sources were identified through searches in major academic databases and were selected based on their relevance to deepfake technology, digital humor, online harassment, and psychological processes such as moral disengagement, online disinhibition, empathy reduction, and social norm reinforcement. The results indicate that acceptance of deepfake-based humor is commonly supported by four interrelated mechanisms, namely normalization through participatory digital culture, psychological distancing that weakens empathy, moral ambiguity created by humorous framing, and reduced accountability through diffusion of responsibility in online spaces. In addition, the literature conceptualizes deepfake humor as a hybrid phenomenon situated between remix-based entertainment and identity-targeting harm, shaped by platform visibility and engagement dynamics. This review highlights that deepfake-based humor may be tolerated not because it is harmless, but because it is routinely framed as "just a joke," making its harm easier to minimize and socially overlook. Therefore, this study emphasizes the need for more direct empirical research and stronger interventions to prevent deepfake-based humor from becoming a normalized form of digital harassment in increasingly synthetic digital environments.

Keywords : Deepfake, oral disengagement, online disinhibition, digital harassment.

Abstrak

Tujuan penelitian ini adalah untuk menguji bagaimana humor berbasis deepfake menjadi dapat diterima secara sosial meskipun berpotensi berfungsi sebagai pelecehan digital. Studi ini berfokus pada mekanisme psikologis yang menjelaskan toleransi audiens dan normalisasi konten humor berbahaya berbasis identitas di lingkungan online. Studi ini menggunakan desain tinjauan lingkup untuk

Multidisciplinary Indonesian Center Journal (MICJO)
Journal page is available to
https://e-jurnal.jurnalcenter.com/index.php/micjo
Email: admin@jurnalcenter.com

memetakan dan mensintesis penelitian yang ada di berbagai bidang psikologi, studi media, dan psikologi siber. Sumber-sumber tersebut diidentifikasi melalui pencarian di basis data akademik utama dan dipilih berdasarkan relevansinya dengan teknologi deepfake, humor digital, pelecehan online, dan proses psikologis seperti pelepasan moral, disinhibisi online, pengurangan empati, dan penguatan norma sosial. Hasil penelitian menunjukkan bahwa penerimaan humor berbasis deepfake umumnya didukung oleh empat mekanisme yang saling terkait, yaitu normalisasi melalui budaya digital partisipatif, jarak psikologis yang melemahkan empati, ambiguitas moral yang diciptakan oleh kerangka humor, dan berkurangnya akuntabilitas melalui penyebaran tanggung jawab di ruang online. Selain itu, literatur mengkonseptualisasikan humor deepfake sebagai fenomena hibrida yang terletak di antara hiburan berbasis remix dan bahaya yang menargetkan identitas, dibentuk oleh visibilitas platform dan dinamika keterlibatan. Ulasan ini menyoroti bahwa humor berbasis deepfake mungkin ditoleransi bukan karena tidak berbahaya, tetapi karena secara rutin dibingkai sebagai "hanya lelucon," sehingga lebih mudah untuk meminimalkan kerusakannya dan diabaikan secara sosial. Oleh karena itu, penelitian ini menekankan perlunya penelitian empiris yang lebih langsung dan intervensi yang lebih kuat untuk mencegah humor berbasis deepfake menjadi bentuk pelecehan digital yang dinormalisasi di lingkungan digital yang semakin sintetis..

Kata Kunci : Deepfake, disenggagement oral, disinhibisi online, pelecehan digital.

1. INTRODUCTION

Advances in artificial intelligence have profoundly reshaped the contemporary media landscape, particularly through the emergence of deepfake technology. Deepfakes refer to synthetic media generated using machine learning techniques that enable the realistic manipulation of faces, voices, and bodily expressions of real individuals (Citron & Chesney, 2019). While early discussions surrounding deepfakes focused on political misinformation and security threats, a growing body of scholarship has highlighted their increasing presence in everyday digital culture, especially in the form of memes, parody videos, and humorous content circulated on social media platforms (Vaccari & Chadwick, 2020).

Humor has long been recognized as a central feature of online interaction. Digital humor facilitates social bonding, reinforces group norms, and provides emotional relief in highly mediated environments (Meyer, 2000). However, humor also possesses an ambivalent character. Research in social psychology has demonstrated that jokes can function as vehicles for prejudice, aggression, and symbolic violence, particularly when they target marginalized individuals or exploit power asymmetries (Billig, 2001; Bourdieu, 2001). In online media, this ambivalence is intensified by the speed, scale, and anonymity of content dissemination.

Deepfake-based humor occupies a particularly problematic position within this landscape. Unlike conventional jokes or edited images, deepfakes appropriate an individual's likeness and recontextualize it without consent, often stripping the subject of agency over their own representation. Empirical studies on non-consensual synthetic media suggest that such practices can elicit psychological distress, including shame, anxiety, loss of control, and identity disruption among victims (Kira, 2024). Yet, when deepfake content is framed as "just a joke," its harmful potential is frequently minimized or dismissed by audiences.

Multidisciplinary Indonesian Center Journal (MICJO)
Journal page is available to
https://e-jurnal.jurnalcenter.com/index.php/micjo
Email: admin@jurnalcenter.com

The acceptance of harmful humor cannot be understood solely as a matter of individual intent. Psychological theories provide important insights into why audiences tolerate or even endorse digital practices knowing they may cause harm. Bandura's theory of moral disengagement explains how individuals cognitively restructure harmful actions so they appear socially acceptable, particularly when responsibility is diffused or harm is trivialized (Albert Bandura, 1999). Humor serves as a powerful disengagement mechanism by reframing harm as playfulness, thereby weakening empathic responses and moral accountability (Woodzicka et al., 2015).

In digital environments, these processes are further amplified by online disinhibition. Suler (2004) argues that anonymity, invisibility, and reduced social cues in online spaces lower self-regulation and moral restraint, making users more likely to engage in or endorse behavior they would otherwise reject offline. When deepfake-based harassment is collectively validated through likes, shares, and humorous commentary, social norms shift toward normalization, reinforcing the perception that such content is acceptable or inconsequential.

Research on digital harassment consistently demonstrates that psychological harm is not limited to explicit abuse. Subtle, normalized forms of aggression, such as ridicule, humiliation, or identity distortion, can be equally damaging, particularly when victims are denied recognition as legitimate targets of harm (Jane, 2017). Humor-driven harassment often places victims in a paradoxical position: expressing distress risks being perceived as humorless or overly sensitive, while silence reinforces the acceptability of the behavior. This dynamic mirrors what scholars describe as symbolic violence, wherein domination is maintained through social practices that appear natural or benign (Bourdieu, 2001). Despite these insights, existing literature remains fragmented. Studies on deepfakes tend to focus on detection technologies, legal regulation, or ethical risks, while research on online humor and harassment often examines memes, trolling, or cyberbullying without addressing synthetic media specifically. Consequently, there is limited integrative understanding of the psychological mechanisms that allow deepfake-based humor to be socially accepted despite its potential to function as digital harassment.

Given the rapid diffusion of generative AI technologies and their embedding in everyday online humor, there is a pressing need to synthesize existing psychological research to clarify how acceptance of deepfake-based harm is produced and sustained. Examining mechanisms such as moral disengagement, social norm internalization, online disinhibition, and empathy suppression is crucial for understanding why audiences participate in or tolerate practices that undermine individual dignity and psychological well-being. Addressing this gap will not only advance theoretical discussions in psychology and media studies but also inform interventions, digital literacy efforts, and policy responses aimed at mitigating harm in increasingly synthetic digital environments.

Multidisciplinary Indonesian Center Journal (MICJO)
Journal page is available to
https://e-jurnal.jurnalcenter.com/index.php/micjo
Email: admin@jurnalcenter.com

2. RESEARCH METHOD

a. Design

This study adopts a scoping review of literature to synthesize existing research on the psychological mechanisms underlying the acceptance of deepfake-based humor and its relationship to digital harassment. Scoping reviews are particularly suitable for examining emerging, complex, and multidisciplinary phenomena where conceptual boundaries are still evolving and empirical findings remain fragmented. Unlike systematic reviews, which typically focus on narrowly defined intervention questions, scoping reviews aim to map the breadth of existing knowledge, identify dominant theoretical perspectives, and highlight research gaps (Grant & Booth, 2009; Peters et al., 2015).

Given the rapid diffusion of generative artificial intelligence technologies and the relatively recent emergence of deepfake-based humor as a social phenomenon, a scoping review approach allows for a comprehensive examination of diverse theoretical, empirical, and conceptual contributions across psychology, media studies, and cyberpsychology. The present review follows general principles outlined in the Joanna Briggs Institute (JBI) Manual for Evidence Synthesis, emphasizing transparency, replicability, and methodological rigor, while maintaining flexibility appropriate to exploratory synthesis.

b. Review Questions

The scoping review is guided by the following research questions:

1. What psychological mechanisms have been identified in the literature to explain the acceptance of harmful or identity-based digital humor?
2. How do existing studies conceptualize deepfake-based humor in relation to digital harassment and psychological harm?
3. Which theoretical frameworks have been most frequently employed to explain the normalization of harmful digital practices framed as humor?

These questions are designed to capture both theoretical developments and empirical insights relevant to understanding acceptance processes rather than prevalence or causal effects.

c. Eligibility Criteria

Inclusion and exclusion criteria were established based on the research questions and adapted from a structured framework similar to the DS-CPC format (Documents, Studies, Constructs, Participants, Contexts), as applied in prior scoping reviews in psychology

1) Type of Documents

Included documents comprised peer-reviewed journal articles, scholarly books, and book chapters that addressed deepfake technology, digital humor, online harassment, or relevant psychological mechanisms. Grey literature such as policy reports and doctoral dissertations was included selectively to reduce publication bias and capture emerging theoretical discussions. Excluded were editorials, opinion pieces, newspaper articles, technical reports without psychological analysis, and non-scholarly sources.Type of Studies

Multidisciplinary Indonesian Center Journal (MICJO)
Journal page is available to
https://e-jurnal.jurnalcenter.com/index.php/micjo
Email: admin@jurnalcenter.com

2) Type of Studies
Both empirical studies (quantitative, qualitative, and mixed-methods) and theoretical or conceptual papers were included, provided they explicitly addressed psychological processes related to humor, moral judgment, online behavior, or digital harassment. Studies focusing solely on technical detection of deepfakes or legal regulation without psychological analysis were excluded.

3) Constructs
The review focused on constructs related to psychological acceptance and normalization, including but not limited to moral disengagement, humor perception, online disinhibition, empathy, social norms, symbolic violence, and digital harassment. Studies examining deepfakes exclusively as political misinformation or cybersecurity threats were excluded unless psychological mechanisms were explicitly discussed.

4) Particpants
Studies involving adolescents and adults as media users, audiences, or victims were included. Research exclusively focused on children or clinical populations without relevance to digital media contexts was excluded.

5) Contexts
The review focused on digital and online environments, particularly social media platforms, meme culture, and user-generated content spaces. Offline harassment contexts were excluded unless explicitly connected to online or digitally mediated practices.

d. Search Strategy

The literature search was conducted across multiple electronic databases commonly used in psychology and behavioral sciences, including PsycINFO, Scopus, Web of Science, PubMed, and Google Scholar. To ensure comprehensive coverage, multidisciplinary databases were also consulted.

Search strings were constructed using combinations of keywords and Boolean operators, for example:

(“deepfake” OR “synthetic media”) AND (“humor” OR “joke” OR “meme”) AND (“psychological mechanisms” OR “moral disengagement” OR “online disinhibition” OR “digital harassment”)

Manual searches were additionally performed by screening reference lists of key articles and reviews to identify relevant studies that may not have been captured through database searches. This multi-step approach mirrors established scoping review practices in psychological research.

e. Study Selection

All retrieved records were screened in two stages. In the first stage, titles and abstracts were independently reviewed to assess relevance based on the inclusion criteria. In the second stage, full texts of potentially eligible articles were examined to determine final inclusion.

Multidisciplinary Indonesian Center Journal (MICJO)
Journal page is available to
https://e-jurnal.jurnalcenter.com/index.php/micjo
Email: admin@jurnalcenter.com

Discrepancies in study selection were resolved through discussion and re-examination of the eligibility criteria to ensure consistency and transparency.

f. Data Extraction and Analysis

Data extraction focused on key descriptive and conceptual elements, including publication year, disciplinary background, study design, theoretical framework, and core psychological mechanisms discussed. Rather than aggregating effect sizes or outcomes, the analysis emphasized thematic synthesis to identify recurring patterns in how acceptance of deepfake-based humor is explained across studies.

The extracted material was coded iteratively, allowing themes to emerge inductively while remaining informed by established psychological theories. Particular attention was paid to how humor was framed in relation to harm, responsibility, and moral evaluation. This approach aligns with the exploratory and mapping-oriented goals of scoping reviews in psychology.

g. Methodological Rigor

Although scoping reviews do not typically involve formal quality appraisal in the same manner as systematic reviews, methodological transparency was prioritized throughout the review process. Search strategies, inclusion criteria, and analytic procedures were explicitly documented to enhance replicability and reliability. Where applicable, methodological limitations of included studies were noted to contextualize findings and inform future research directions.

3. RESULT AND DISCUSSION

a. Scope and Characteristics of the Reviewed Studies

The scoping review identified a heterogeneous body of literature addressing the psychological and social dynamics of humor, digital manipulation, and online harassment. While empirical studies explicitly examining deepfake-based humor remain scarce, a growing number of publications investigate adjacent phenomena such as synthetic media practices, meme-based ridicule, online norm formation, and the social acceptance of digitally mediated harm. These studies span diverse disciplinary backgrounds, including social psychology, communication studies, digital sociology, and media ethics, reflecting the complex and interdisciplinary nature of the topic.

The majority of reviewed studies do not conceptualize deepfake humor as an isolated category. Instead, deepfake-based practices are frequently discussed within broader analyses of visual manipulation, participatory culture, and the normalization of harmful online behaviors. Research on meme culture and remix practices highlights how digitally altered representations of individuals are often detached from their original contexts and reinserted into humorous or satirical frames, reducing perceived accountability for potential harm (Milner, 2016; Shifman, 2014).

Multidisciplinary Indonesian Center Journal (MICJO)
Journal page is available to
https://e-jurnal.jurnalcenter.com/index.php/micjo
Email: admin@jurnalcenter.com

b. Normalization of Harm Through Participatory Digital Culture

A consistent theme across the literature concerns the role of participatory digital culture in normalizing harmful humor. Studies on online communities demonstrate that humor functions as a key mechanism for establishing in-group norms and reinforcing shared values (Dynel, 2016). Within such environments, repeated exposure to humor that targets individuals or groups contributes to the gradual erosion of moral boundaries, particularly when these practices are framed as culturally acceptable or creatively justified.

Empirical research on online harassment indicates that harmful behaviors are more likely to persist when embedded within everyday communicative practices rather than framed as explicit aggression. For instance, Matamoros-Fernández (2017) demonstrates how platform affordances can enable the circulation of harmful content under the guise of humor or irony, thereby obscuring its discriminatory or abusive dimensions. Similarly, (Phillips & Milner, 2021) argue that digital humor often operates through ambiguity, allowing users to deny harmful intent while still producing negative social effects. Within this context, deepfake-based humor emerges as an extension of existing participatory practices rather than a radical departure. The reviewed literature suggests that audiences often interpret manipulated media through culturally learned frameworks of remix and play, which diminishes sensitivity to the ethical implications of identity manipulation.

c. Psychological Distance and the Attenuation of Empathic Response

Another prominent finding relates to the attenuation of empathic responses in digitally mediated interactions. Research in social and media psychology consistently shows that visual and emotional distance reduces empathic concern, particularly when individuals are encountered as images rather than embodied persons (Ahn et al., 2014). Studies examining online ridicule and visual shaming further indicate that audiences are less likely to perceive harm when victims are represented through altered or stylized media forms (Udris, 2014).

In the case of deepfake-based humor, the synthetic nature of the content itself contributes to psychological distancing. The manipulated image or video is often perceived as an artifact rather than as a representation of a real person, which weakens emotional engagement with the subject's experience. This distancing effect is reinforced by the absence of direct feedback from those depicted, allowing audiences to engage with content without confronting its personal consequences (Weller & Kinder-Kurlanda, 2016).

d. Moral Ambiguity and Audience Interpretation

The reviewed literature also highlights the role of moral ambiguity in shaping audience responses to harmful humor. Rather than evaluating content based on fixed ethical standards, users often rely on contextual cues, peer reactions, and platform norms to guide interpretation. Research on moral judgment in online environments shows that individuals are more likely to suspend moral evaluation when content is framed as ironic, satirical, or humorous (LaCroix & Pratto, 2015).

This interpretive flexibility allows deepfake-based humor to occupy a morally ambiguous space in which responsibility is diffused and ethical evaluation becomes negotiable.

Multidisciplinary Indonesian Center Journal (MICJO)
Journal page is available to
https://e-jurnal.jurnalcenter.com/index.php/micjo
Email: admin@jurnalcenter.com

As a result, audiences may simultaneously recognize the potential for harm while continuing to endorse or circulate the content, reflecting a broader pattern of moral ambivalence documented in studies of online behavior (Fiesler & Proferes, 2018). Across the reviewed corpus, four recurring psychological mechanisms appear to organize how harmful, identity-targeting digital humor becomes acceptable to audiences. First, participatory digital culture provides the normative infrastructure through which humor circulates, gains legitimacy, and becomes embedded in routine interaction, particularly through group-based norm reinforcement and identity processes (Reicher et al., 1995). Second, psychological distance. supported by mediated visibility, abstraction, and the aestheticization of altered representations, attenuates empathic engagement and weakens sensitivity to harm. Third, moral ambiguity operates as an interpretive condition that enables audiences to shift evaluative criteria away from ethical appraisal toward contextual and social cues, particularly when content is framed as ironic or playful; this ambiguity has been increasingly recognized as a key mechanism through which harmful humour can evade governance and be normalized within platformed communication (Matamoros-Fernández et al., 2023). Fourth, online interactional dynamics such as disinhibition and diffusion of responsibility function as enabling conditions that reduce perceived accountability, especially when engagement is distributed across large and loosely coordinated publics, consistent with evidence on bystander processes in cyberbullying environments (You & Lee, 2019). Taken together, these mechanisms suggest a sequential reinforcement process in which platformed circulation and participatory norms facilitate moral ambiguity, which in turn sustains empathic attenuation and reduced accountability, thereby stabilizing acceptance even when harm is recognizable in principle.

In terms of conceptualization, the literature tends to position deepfake-based humor in four overlapping ways. A first cluster frames it primarily as a form of remix practice rooted in participatory culture, emphasizing creativity, circulation, and meme logic (Milner, 2016). A second cluster treats it as a case of ambiguous harm, highlighting its borderline status between entertainment and harassment and the interpretive variability that follows from that ambiguity (Matamoros-Fernández et al., 2023; Phillips & Milner, 2021). A third conceptualization foregrounds identity-based violation, in which the appropriation of likeness is treated as a boundary transgression that undermines agency and personhood (Paris, 2021; Romero-Moreno, 2024). Finally, a fourth strand situates deepfake humor as platform-mediated harassment, emphasizing how visibility regimes, engagement incentives, and community norms can transform isolated acts into normalized practices (Im et al., 2022; Matamoros-Fernández et al., 2023).

e. Dominant Theoretical Lenses Used in the Literature

A further pattern concerns the theoretical lenses most frequently mobilized to explain normalization processes in digitally mediated humor and harassment. Across the reviewed studies, humor scholarship is commonly used to account for how audiences interpret norm violations as acceptable or trivial, particularly in contexts marked by irony and play (Billig, 2001). Alongside this, theories of online behavior and social norms are frequently invoked to

Multidisciplinary Indonesian Center Journal (MICJO)
Journal page is available to
https://e-jurnal.jurnalcenter.com/index.php/micjo
Email: admin@jurnalcenter.com

explain how platformed interaction reduces accountability and encourages norm convergence through observable engagement, especially in environments characterized by anonymity and attenuated interpersonal cues (Suler, 2004). A smaller but influential subset draws on moral psychological perspectives to explain cognitive rationalizations that neutralize perceived harm and enable endorsement of practices that would otherwise conflict with personal moral standards (A Bandura, 2016), while critical sociological traditions are employed to interpret the normalization of identity-based harm as a function of symbolic domination and the misrecognition of violence. Finally, platform and HCI-oriented approaches are used to connect psychological acceptance to infrastructural conditions, algorithmic amplification, affordances, and governance regimes, that shape what becomes visible, shareable, and socially rewarded in everyday online participation (Bucher, 2018). This distribution of theoretical approaches indicates that the literature most often explains acceptance as a multi-level phenomenon spanning individual cognition, social interaction, and platform-mediated cultural normalization. Where frequency could not be quantified due to heterogeneity in study designs and reporting, the mapping nevertheless indicates consistent reliance on humor interpretation, social norm formation, moral psychological rationalization, and platform-mediated visibility as the primary explanatory families.

Discussion

a. Deepfake-Based Humor as an Extension of Normalized Digital Practices

The findings of this scoping review suggest that acceptance of deepfake-based humor is best understood as an extension of existing digital practices rather than as a phenomenon driven solely by technological novelty. Consistent with prior research on participatory culture, humor-based manipulation of images and identities is embedded within long-standing traditions of remix, parody, and meme production (Shifman, 2014). However, the introduction of deepfake technology intensifies these practices by increasing realism and reducing the visibility of manipulation, thereby amplifying their potential psychological impact.

The reviewed studies indicate that normalization occurs through cumulative exposure and social reinforcement. When manipulated content is repeatedly encountered in humorous contexts and met with positive engagement, it becomes integrated into everyday digital interaction. This process mirrors patterns observed in other domains of normalized harassment, where subtle and ambiguous behaviors are more difficult to contest than overt aggression (Udris, 2014).

b. The Role of Platforms and Social Norm Formation

Importantly, the acceptance of deepfake-based humor cannot be separated from the structural conditions of digital platforms. Research on platform governance highlights how algorithmic amplification and engagement-driven design prioritize content that provokes emotional responses, including humor and shock (Bucher, 2018). As a result, humorous deepfake content may receive disproportionate visibility, reinforcing its perceived legitimacy.

Studies on online norm formation demonstrate that users often infer acceptable behavior from observable patterns of engagement rather than from formal rules (Fiesler & Proferes,

Multidisciplinary Indonesian Center Journal (MICJO)
Journal page is available to
https://e-jurnal.jurnalcenter.com/index.php/micjo
Email: admin@jurnalcenter.com

2018). In environments where deepfake humor circulates without sanction, silence or passive endorsement functions as normative approval. This dynamic contributes to the gradual institutionalization of harmful practices within platform cultures.

c. Implications for Psychological Harm and Victim Recognition

The reviewed literature raises significant concerns regarding the psychological consequences of deepfake-based humor, particularly in relation to victim recognition. Research on online victimization consistently shows that harm is exacerbated when victims' experiences are invalidated or dismissed (Im et al., 2022). When deepfake content is framed as humorous, individuals who experience distress may struggle to articulate their suffering in socially acceptable terms, increasing the risk of internalized blame and withdrawal.

Moreover, the ambiguity surrounding humor complicates efforts to challenge harmful practices. As Phillips & Milner (2021) note, ironic or playful framing often provides social cover for harmful behavior, allowing perpetrators and audiences alike to evade accountability. This suggests that psychological harm is not merely an unintended byproduct of deepfake humor but is structurally enabled by the cultural logic of digital entertainment.

d. Directions for Future Research

Taken together, these findings underscore the need for future research to move beyond technological and legal analyses of deepfakes and engage more deeply with the psychological and cultural processes that shape audience acceptance. Empirical studies examining emotional responses, moral reasoning, and social norms in relation to deepfake humor are particularly needed. Additionally, longitudinal research could illuminate how repeated exposure influences moral sensitivity and empathic engagement over time.

By situating deepfake-based humor within established literatures on digital culture, moral ambiguity, and normalized harassment, this review contributes to a more nuanced understanding of how emerging technologies intersect with enduring psychological processes. Addressing these dynamics is essential for developing interventions that promote digital environments grounded in respect, accountability, and psychological well-being.

4. CONCLUSION

This scoping review examined why deepfake-based humor can become socially acceptable even though it may function as a form of digital harassment. The review suggests that acceptance is supported by several common psychological processes. First, harmful content can become "normal" when it circulates repeatedly and is treated as part of everyday joking culture on social media. Second, because interactions happen through screens, audiences may feel less emotionally connected to the person being targeted, which can reduce empathy. Third, when content is framed as humor, people may interpret it as harmless and avoid making a serious moral judgment. Fourth, responsibility is often weakened in online spaces, where many users watch, like, and share content, creating a sense that the harm is not any single person's concern.

Multidisciplinary Indonesian Center Journal (MICJO)
Journal page is available to
https://e-jurnal.jurnalcenter.com/index.php/micjo
Email: admin@jurnalcenter.com

The reviewed studies also indicate that deepfake humor is rarely discussed as a standalone phenomenon. Instead, it is often understood as part of broader meme and remix practices. As a result, it frequently sits in a grey area between "joking" and "attacking," especially when a real person's face or identity is used without consent. Platform dynamics also matter. Engagement features such as likes, shares, and algorithmic promotion can rapidly spread deepfake humor and make it appear increasingly common and socially acceptable.

Overall, the literature draws on multiple perspectives to explain these patterns, including theories of humor, online behavior, social norms, and moral psychology. The key conclusion is that deepfake-based humor may be accepted not because it is harmless, but because it is widely framed as ordinary entertainment and its impact is often not felt immediately by audiences. Further research is therefore needed to better understand how people evaluate deepfakes as "just a joke" and how to prevent such practices from becoming a normalized form of digital harassment.

5. REFERENCES

Ahn, S. J. (Grace), Bailenson, J. N., & Park, D. (2014). Short- and long-term effects of embodied experiences in immersive virtual environments on environmental locus of control and behavior. *Computers in Human Behavior*, *39*, 235–245. https://doi.org/10.1016/j.chb.2014.07.025

Bandura, A. (2016). *Moral disengagement: How people do harm and live with themselves*. Worth Publishers.

Bandura, Albert. (1999). Moral disengagement in the perpetration of inhumanities. *Personality and Social Psychology Review*, *3*(3), 193–209. https://doi.org/10.1207/s15327957pspr0303_3

Billig, M. (2001). Humour and hatred: The racist jokes of the Ku Klux Klan. *Discourse & Society*, *12*(3), 267–289. https://doi.org/10.1177/0957926501012003001

Bourdieu, P. (2001). *Masculine domination*. Polity Press.

Bucher, T. (2018). *If...Then: Algorithmic power and politics*. Oxford University PressNew York. https://doi.org/10.1093/oso/9780190493028.001.0001

Citron, D. K., & Chesney, R. (2019, February). Deepfakes and the new disinformation war. *Foreign Affairs*.

Dynel, M. (2016). "I has seen Image Macros!" Advice animals memes as visual-verbal jokes. *International Journal of Communication*, *10*, 660–688.

Fiesler, C., & Proferes, N. (2018). "Participant" Perceptions of Twitter Research Ethics. *Social Media + Society*, *4*(1). https://doi.org/10.1177/2056305118763366

Im, J., Schoenebeck, S., Iriarte, M., Grill, G., Wilkinson, D., Batool, A., Alharbi, R., Funwie, A., Gankhuu, T., Gilbert, E., & Naseem, M. (2022). Women's perspectives on harm and justice after online harassment. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW2), 1–23. https://doi.org/10.1145/3555775

Jane, E. (2017). *Misogyny online: A short (and brutish) history*. SAGE Publications Ltd.

Multidisciplinary Indonesian Center Journal (MICJO)
Journal page is available to
https://e-jurnal.jurnalcenter.com/index.php/micjo
Email: admin@jurnalcenter.com

https://doi.org/10.4135/9781473916029

Kira, B. (2024). When non-consensual intimate deepfakes go viral: The insufficiency of the UK Online Safety Act. *Computer Law & Security Review, 54*, 106024. https://doi.org/10.1016/j.clsr.2024.106024

LaCroix, J. M., & Pratto, F. (2015). Instrumentality and the denial of personhood: The social psychology of objectifying others. *Revue Internationale de Psychologie Sociale*, *28*(1), 183–212.

Matamoros-Fernández, A. (2017). Platformed racism: the mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube. *Information, Communication & Society*, *20*(6), 930–946. https://doi.org/10.1080/1369118X.2017.1293130

Matamoros-Fernández, A., Bartolo, L., & Troynar, L. (2023). Humour as an online safety issue: Exploring solutions to help platforms better address this form of expression. *Internet Policy Review*, *12*(1). https://doi.org/10.14763/2023.1.1677

Meyer, J. C. (2000). Humor as a double-edged sword: Four functions of humor in communication. *Communication Theory*, *10*(3), 310–331. https://doi.org/10.1111/j.1468-2885.2000.tb00194.x

Milner, R. M. (2016). *The World Made Meme*. The MIT Press. https://doi.org/10.7551/mitpress/9780262034999.001.0001

Paris, B. (2021). Configuring fakes: Digitized bodies, the politics of evidence, and agency. *Social Media + Society*, *7*(4). https://doi.org/10.1177/20563051211062919

Phillips, W., & Milner, R. M. (2021). *You are here*. The MIT Press. https://doi.org/10.7551/mitpress/12436.001.0001

Reicher, S. D., Spears, R., & Postmes, T. (1995). A social identity model of deindividuation phenomena. *European Review of Social Psychology*, *6*(1), 161–198. https://doi.org/10.1080/14792779443000049

Romero-Moreno, F. (2024). Generative AI and deepfakes: a human rights approach to tackling harmful content. *International Review of Law, Computers & Technology*, *38*(3), 297–326. https://doi.org/10.1080/13600869.2024.2324540

Shifman, L. (2014). *Memes in digital culture*. The MIT Press.

Suler, J. (2004). The online disinhibition effect. *CyberPsychology & Behavior*, *7*(3), 321–326. https://doi.org/10.1089/1094931041291295

Udris, R. (2014). Cyberbullying among high school students in Japan: Development and validation of the Online Disinhibition Scale. *Computers in Human Behavior*, *41*, 253–261. https://doi.org/10.1016/j.chb.2014.09.036

Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society*, *6*(1). https://doi.org/10.1177/2056305120903408

Weller, K., & Kinder-Kurlanda, K. E. (2016). A manifesto for data sharing in social media research. *Proceedings of the 8th ACM Conference on Web Science*, 166–172.

Multidisciplinary Indonesian Center Journal (MICJO)
Journal page is available to
https://e-jurnal.jurnalcenter.com/index.php/micjo
Email: admin@jurnalcenter.com

https://doi.org/10.1145/2908131.2908172

Woodzicka, J. A., Mallett, R. K., Hendricks, S., & Pruitt, A. V. (2015). It's just a (sexist) joke: Comparing reactions to sexist versus racist communications. *HUMOR*, *28*(2). https://doi.org/10.1515/humor-2015-0025

You, L., & Lee, Y.-H. (2019). The bystander effect in cyberbullying on social network sites: Anonymity, group size, and intervention intentions. *Telematics and Informatics*, *45*, 101284. https://doi.org/10.1016/j.tele.2019.101284