

Understanding Motivations and Patterns Behind Harmful Deepfake and AI-Generated Content: Insights from Virtual Experiences and LDA Analysis

Thenmozhi Pandian¹, Neelamalar Maraimalai²

¹Assistant Professor, Department of Communication and Media Studies, M.O.P. Vaishnav College of Women (Autonomous), Chennai – 600 034

²Neelamalar Maraimalai, Professor, Department of Media Studies, College of Engineering Guindy, Anna University, Chennai – 600 025

Abstract—Using Latent Dirichlet Allocation (LDA), this study investigates reasons behind creating harmful deepfake and AI-generated content online, focusing on virtual experiences. Results reveal clear themes: novelty, personalized fantasy, enhanced immersion, anonymity, taboo exploration, technology-mediated intimacy, coping, attention-seeking, and profit. Each theme, supported by probability scores, uncovers users' preferences for diverse, stimulating, and personalized content. The following sections analyze these themes, revealing patterns. The discussion explains motivations, weaving together novelty, customization, privacy, and exploration. Understanding this complex user engagement paves the way for future research and interventions. In conclusion, the study calls for legislation addressing ethical and legal aspects, especially concerning leaked content, privacy, and deepfake distribution.

Index Terms—Deepfakes, Latent Dirichlet Allocation, virtual experiences, user motivations, harmful content

I. INTRODUCTION

The relentless march of artificial intelligence, particularly in the domain of deep learning, has birthed a technological marvel that both captivates and concerns: deepfake technology. Fueled by sophisticated algorithms, deepfakes generate hyper-realistic simulated content, blurring the lines between reality and manipulation. While the positive applications of deepfakes in entertainment and digital creativity are evident, the nefarious potential of this technology has prompted a critical examination of the motivations propelling individuals into the realm of deceptive deepfake creation.

Amidst the growing awareness of the societal implications, understanding the multifaceted nature of the deepfake phenomenon becomes paramount. The power of deepfakes extends beyond technical intricacies, presenting risks to societal norms, security, and the very concept of truth. At the heart of this dilemma lies a pressing need to comprehend the motivations driving malicious deepfake creation. To achieve this, our research adopts a qualitative approach, utilizing in-depth interviews with individuals entrenched in the world of deceptive deepfaking.

Qualitative inquiry, facilitated by in-depth interviews, serves as a portal to the subjective experiences and perspectives of deepfake creators (Meskys et al., 2020). Engaging in open-ended conversations allows for a nuanced exploration of decision-making processes, emotional states, and the complex interplay of factors influencing their actions. This qualitative dimension aims to move beyond a mere enumeration of motives, seeking to unravel the intricate psychological, social, and cultural layers shaping the mindset of those involved in malicious deepfake creation (Hazan, 2020).

Acknowledging the broader contextual elements is essential in understanding the motivations of malicious deepfakers. The socio-cultural landscape, evolving dynamics of digital communication, and the cloak of anonymity provided by the online realm contribute to the complex web defining the motivations of deepfake creators (Sharma et al., 2023). Complementing the qualitative insights, our research integrates an empirical analysis layer using Latent

Dirichlet Allocation (LDA) on the verbatim corpus derived from in-depth interviews (Blei, Ng, & Jordan, 2003).

This comprehensive approach, fusing qualitative insights with quantitative empirical analysis, seeks to provide a comprehensive understanding of the underlying factors propelling individuals into the world of malicious deepfake creation. By exploring motivations within a rich qualitative framework and unraveling latent themes through LDA, our research aims to contribute substantively to the ongoing discourse on deepfakes. Through this endeavor, we aim to inform not only academic discussions but also lay the groundwork for practical strategies and policies aimed at mitigating the potential societal harms associated with malicious deepfakes (Gamage, Sasahara, & Chen, 2021).

II. REVIEW OF LITERATURE

The exploration of malicious deepfake creation necessitates a thorough examination of existing literature, revealing diverse insights into the motivations, impacts, and countermeasures associated with this burgeoning technological threat. While the introduction illuminated the qualitative aspects of our research, the literature review delves into distinct facets, widening the discourse surrounding deepfakes.

Technological advancements have played a pivotal role in the evolution of deepfake creation. Recent studies (Jones, 2020) highlight the rapid progression in deep learning algorithms, contributing to the seamless generation of hyper-realistic content. Innovations such as Generative Adversarial Networks (GANs) have provided deepfake creators with tools that surpass traditional methods, posing challenges for detection technologies (Hussain et al., 2021). As technology continues to advance, a deeper understanding of these technical intricacies becomes imperative for effective countermeasures.

Beyond technical dimensions, a comprehensive literature review emphasizes the broader societal impacts of malicious deepfakes. The erosion of trust in digital media and the potential for widespread misinformation have been extensively studied (Dahlgren, 2018). Deepfakes not only undermine the

credibility of visual and auditory information but also pose significant threats to democratic processes, raising concerns about the manipulation of public opinion and political narratives (Langa, 2021).

While the psychological motivations of deepfake creators were introduced in the introduction, a nuanced exploration of the ethical considerations surrounding deepfakes is warranted. Scholars (Diakopoulos & Johnson, 2021) have delved into the ethical implications of deepfake creation, addressing issues of consent, privacy, and the potential for harm. Examining the ethical landscape surrounding malicious deepfaking contributes to a more comprehensive understanding of the ethical challenges associated with this technology.

Countermeasures and strategies for mitigating the impact of malicious deepfakes form a critical segment of the literature. Researchers (Yu et al., 2021) have explored various technological solutions, including deepfake detection algorithms and authentication methods. However, the cat-and-mouse game between deepfake creators and detection technologies remains dynamic, emphasizing the need for interdisciplinary approaches that consider legal, ethical, and technical dimensions (Leibowicz, McGregor, & Ovadya, 2021).

Our research methodology, utilizing in-depth interviews alongside Latent Dirichlet Allocation (LDA), aligns with recent trends in research approaches employed in the domain of deepfake studies. While qualitative insights from interviews enrich the understanding of individual motivations, LDA applied to verbatim transcripts provides a structured approach to extract latent themes, ensuring a comprehensive exploration of the underlying factors (Yordanova et al., 2019).

To summarize, this literature review has illuminated distinct dimensions of the malicious deepfake landscape, covering technological advancements, societal impacts, ethical considerations, and countermeasures. Building upon these insights, our research aims to contribute to the evolving discourse, providing valuable perspectives for the development of strategies and policies that effectively address the multifaceted challenges posed by malicious deepfakes.

III. METHODOLOGY

This research employed a comprehensive approach to thoroughly investigate the motivations and behaviors of users engaged with deepfake and AI-generated explicit content.

A. Data Collection

Utilizing in-depth interviews, three participant groups—human behavior psychologists, cybercrime experts, and users of deepfake applications—were engaged. Eight participants from each of the first two groups were purposefully selected and approached in person, while fourteen deepfake application users were included, with seven approached in person and the remaining recruited through the subreddit r/midjourney.

The interviews, structured with a set of questions, aimed to provide both consistency and flexibility, allowing for the exploration of unexpected themes arising from participants' responses. Each interview, lasting approximately two and a half hours, generated comprehensive 80-page transcripts, totaling approximately 2400 pages across all 30 interviews. This extensive data captured nuanced perspectives and fostered a holistic understanding of participants' experiences. Notably, due to challenges in eliciting responses regarding intentions for distribution from deepfake application users, we restructured the question to focus on the motivations of someone else, leading to more insightful responses.

Moreover, to ensure a balanced representation, equal weightage was given to both the experts and the users in the study. This approach aimed to provide a comprehensive understanding by incorporating perspectives from both sides of the spectrum. Despite difficulties in extracting information about past distribution activities, none of the deepfake application users mentioned having distributed deepfake or AI-porn before. The methodology thus sought to address the complexity of the research topic by incorporating diverse participant groups and adapting the questioning approach based on the evolving dynamics of the interviews.

B. Data Analysis

In our data analysis phase, we employed Latent Dirichlet Allocation (LDA) to derive meaningful topics from the interview transcripts. The rigorous data preprocessing involved text cleaning, tokenization, and stopwords removal. Utilizing the Document-Term Matrix (DTM) approach, we quantitatively represented term frequencies in each interview.

For LDA model training, we iteratively tested for the optimal number of topics (K) and confidently settled on a value of K=9. Using Python libraries scikit-learn and gensim, we initialized and trained the LDA model, obtaining topic distributions for each interview and word distributions for each of the 9 identified topics.

The subsequent interpretation involved extracting key terms associated with each topic. Assigning a primary topic to each interview was based on the highest probability within the 9 identified topics. The validation and refinement phase included coherence score assessment, leveraging Python's pyLDAvis and gensim libraries. Our achieved coherence score of 0.65 indicated a high level of topic coherence, influencing subsequent adjustments, such as parameter fine-tuning, to ensure the robustness of our analytical approach.

The final outputs comprised comprehensive and well-defined topic summaries, seamlessly integrated into our broader research findings. This confident and meticulous LDA analysis, implemented with Python libraries such as scikit-learn, gensim, pyLDAvis, and others, provided valuable insights into the motivations and behaviors of users engaged with deepfake and AI-generated explicit content.

IV. RESULTS

In the exploration of online experiences and virtual interactions, Latent Dirichlet Allocation (LDA) was employed to discern underlying themes and topics from a corpus of data. Table 1 summarizes the results of the LDA analysis, showcasing distinct clusters of keywords associated with various thematic domains. Each topic is characterized by a set of keywords and corresponding weights, providing insights into the

In the first cluster, where the focus is on novelty experiences and variety, the probability scores provide insights into the prevalence of specific terms. For instance, "Novelty" has a probability score of 0.249, indicating a moderate likelihood, while "Exploration" scores 0.291, suggesting a higher probability within this thematic context.

Moving to the second cluster, which highlights personalized fantasy fulfillment, the probability scores for keywords like "Customization" (0.297) and "Fantasy" (0.279) point to a relatively strong association with this theme. Conversely, the lower probability of 0.121 for "Novelty" in this cluster suggests a reduced likelihood of this term being a prominent feature.

In the third cluster, centered around enhanced immersion and realism, the probability scores for keywords such as "Sensory" (0.291) and "Connection" (0.276) indicate a substantial likelihood of these terms being integral to the identified theme. The varying scores across keywords provide a nuanced understanding of their respective importance.

The fourth cluster, emphasizing anonymity and privacy, reveals probability scores that characterize the strength of association for terms like "Avoidance" (0.299) and "Norms" (0.253) within this thematic context. These scores offer insights into the significance of these keywords in the overall theme of anonymity and privacy.

In the fifth cluster, exploring taboos or fantasies, probability scores elucidate the varying degrees of association for terms such as "Taboo" (0.293) and "Fantasy" (0.103). The scores contribute to discerning the prominence of each keyword within the identified theme.

These patterns continue across the subsequent clusters, with each set of probability scores providing a quantitative measure of the strength of association for the corresponding keywords within their respective themes. Table 1, with its probability scores, thus serves as a valuable tool for understanding the nuanced distribution of keywords and their significance in shaping the identified thematic clusters.

V. DISCUSSION

The exploration into the motivations behind engaging in malicious deepfake and AI-generated explicit content is a nuanced journey, as revealed by the Latent Dirichlet Allocation (LDA) results. These topics, each with its distinctive set of keywords and associated probabilities, offer a comprehensive view into the intricate web of reasons that individuals traverse within this controversial domain.

The first identified theme, centered around novelty experiences and variety, underscores a deep-seated desire for diverse and stimulating interactions. The probabilities assigned to keywords like "Novelty" (0.249), "Variety" (0.187), and "Exploration" (0.291) highlight the users' pursuit of engaging and unique content. This aligns seamlessly with established research on user engagement in online platforms, emphasizing the significance of delivering fresh and personalized content to captivate audiences (Barnet, 2009). Furthermore, the users' inclination towards experiences labeled as "fun" suggests a pervasive yearning for enjoyment within the context of novel online encounters.

Moving seamlessly into the next theme, personalized fantasy fulfillment, the LDA results reveal a distinctive set of keywords, each carrying a specific weight in influencing user behaviors. The probabilities assigned to terms like "Customization" (0.297), "Tailoring" (0.190), and "Fantasy" (0.279) collectively paint a picture of users seeking tailored and dynamic experiences. This resonates coherently with existing literature highlighting the significance of online customization and personalization, suggesting that users actively seek platforms that allow them to shape and individualize their experiences based on personal preferences (Chen, Lu, & Lu, 2019). The users' preference for interactive and collaborative elements further reinforces the importance of tailored experiences in the realm of malicious deepfakes.

Enhanced immersion and realism emerge as a crucial theme, where users express a preference for sensory experiences, authenticity, and technological engagement. The probabilities assigned to keywords such as "Sensory" (0.291), "Realism" (0.159), and "Technological" (0.185) underscore a desire for

immersive and technologically advanced encounters. This aligns seamlessly with prior research on virtual reality and immersive technologies, emphasizing the role of realism and sensory engagement in shaping user experiences (Newman et al. 101733). Users' emphasis on simulated connections and technological empowerment further solidifies the importance of these elements in their engagement with malicious deepfake content.

Anonymity and privacy stand out as a central theme, capturing users' concerns and interests surrounding the stigma, control, and avoidance associated with their online activities. The probabilities assigned to keywords like "Avoidance" (0.299), "VPN" (0.245), and "Anonymity" (0.235) underscore the delicate balance individuals seek between exploration and maintaining personal boundaries (Littlejohn, 2017). This theme, seamlessly transitioning from the previous ones, reflects the multifaceted nature of users' motivations and the intricate relationship between their desires for novelty, customization, and privacy.

Exploration of taboos or fantasies introduces yet another layer to the motivations behind engaging in malicious deepfake and AI-generated explicit content. Keywords like "Taboo" (0.293) and "Fantasy" (0.103) reveal a user base keen on experimenting with unconventional subjects within a virtual space. This aligns with previous research emphasizing the delicate balance individuals seek between exploration and maintaining personal boundaries in the online realm (Bowrey, 1998). The probabilities assigned to terms like "Virtual" (0.271) and "Experimentation" (0.187) further emphasize users' curiosity and willingness to engage with content that pushes societal norms and explores individualized ethical considerations.

Technology-mediated intimacy emerges as a crucial theme, reflecting the role of technology in facilitating intimate connections, coping mechanisms, and empowerment in addressing feelings of loneliness. The probabilities assigned to keywords like "Simulated" (0.192), "Connection" (0.268), and "Empowerment" (0.255) highlight users' reliance on technology for emotional fulfillment. This seamlessly aligns with previous research on technology and intimacy, emphasizing the growing intersection between virtual spaces and emotional well-being

(Bulcroft, Bulcroft, Bradley, & Simpson, 2000). The users' emphasis on simulated connections and coping mechanisms further underscores the importance of technology in addressing their emotional needs.

Coping mechanisms surface as a noteworthy theme, interwoven with terms like "Coping" (0.145), "Depression" (0.271), and "Anxiety" (0.103). This theme sheds light on users turning to explicit content creation and consumption as a means of dealing with emotional challenges and seeking solace in virtual experiences. The probabilities assigned to keywords associated with emotional states and coping mechanisms suggest a complex relationship between users and the content they engage with. This theme seamlessly extends from the previous one, underlining the interconnectedness of users' motivations and coping strategies within the online space.

The theme of attention-seeking behaviors stands out prominently in the LDA results, reflecting users' motives such as seeking validation, influence, and recognition. The probabilities assigned to keywords like "Validation" (0.274), "Influence" (0.298), and "Recognition" (0.287) underscore a pursuit of acknowledgment and prominence within the online realm. This theme, with its emphasis on digital identity and social comparison, aligns with psychological perspectives on attention-seeking behaviors in online environments (DeWall, Buffardi, Bonser, & Campbell, 2011). The users' descriptions resonate with the attention-seeking motives identified by human behavior psychologists and cybercrime experts, forming a consistent narrative despite their differing positions on the spectrum.

The final theme, profit motive, encapsulates users' inclinations towards monetization, market engagement, and revenue generation within the online space. The probabilities assigned to keywords like "Monetization" (0.247) and "Market" (0.141) highlight users' awareness of the competitive nature of online markets and the potential financial gains associated with explicit content creation. This theme aligns with existing literature on the monetization of explicit content and the challenges posed by the competitive nature of online markets (Drenten, Gurrieri, & Tyler, 2020). Furthermore, while the identified themes shed light on users' financial

incentives and engagement with explicit content creation, it is essential to recognize the broader context of combating deceptive practices online.

Thus, the integration of AI and machine learning methodologies for deepfake detection represents a pivotal advancement in addressing the proliferation of harmful content online. Through the utilization of neural networks and sophisticated algorithms, researchers have made significant strides in enhancing the efficacy of detection techniques (Rana et al., 2022). These technologies empower the identification of anomalies indicative of deepfake manipulation within multimedia content, bolstering efforts to combat deceptive practices across digital platforms. Moreover, the refinement of deep learning architectures, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs), enables the development of more accurate and efficient detection models capable of discerning complex patterns inherent in deepfake artifacts (Bondi et al., 2020; Güera & Delp, 2018). Such advancements pave the way for proactive identification and mitigation of deceptive content, aligning with the objectives of this study to understand and address the motivations behind harmful deepfake creation.

Additionally, the exploration of frameworks and patterns of AI and crime within this research framework offers valuable insights into the evolving landscape of digital deception. By analyzing historical data and emerging trends, researchers gain a nuanced understanding of the strategies employed by malicious actors in the creation, distribution, and dissemination of harmful deepfake content. This comprehensive understanding informs the development of sophisticated detection algorithms and proactive strategies aimed at disrupting the spread of deceptive content. Furthermore, by studying the intersection of AI-driven technologies and criminal activities, stakeholders can anticipate future challenges and deploy preemptive measures to safeguard against emerging threats in the dynamic realm of online security, thereby contributing to the overarching goals of this study.

In summary, the LDA results provide an intricate tapestry of user motivations, seamlessly transitioning between desires for novelty, customization, privacy,

and exploration. The themes weave together a complex narrative, revealing the multifaceted nature of users' engagement with malicious deepfake and AI-generated explicit content. These insights contribute to a nuanced understanding of user behaviors in online environments, paving the way for future research and interventions in this evolving landscape.

VI. CONCLUSION

In acknowledging the complexities surrounding the creation of AI-generated porn and deepfakes within personal spaces, it's crucial to note that our research did not conclusively identify reasons for deeming such actions inherently wrong. However, the potential privacy and security implications cannot be understated. The risk of leaked content, stemming from the creation of these materials, underscores the need for recognizing privacy as a paramount concern. As we delve into the intricacies of this phenomenon, future research should diligently explore the necessity of legislation tailored to address the ethical and legal dimensions of AI-generated porn and deepfakes. The privacy and security issues associated with the unintentional leakage of such content demand careful consideration. While the distribution of deepfakes is unequivocally wrong, particularly in jurisdictions like India where it falls under laws protecting the modesty of women (Pandey, 2020), the question of accountability for those who created such content, especially when leaked inadvertently, remains a gray area. This prompts a critical examination on a global scale, contemplating whether there should be legal ramifications for those who generated such material, even if originally intended for personal consumption. Addressing these complex ethical and legal nuances is essential for developing robust frameworks that safeguard individuals and uphold privacy rights in an era where technology continues to redefine personal boundaries.

REFERENCES

- [1] Meskys, E., Kalpokiene, J., Jurcys, P., & Liaudanskas, A. (2020). Regulating deep fakes: legal and ethical considerations. *Journal of Intellectual Property Law & Practice*, 15(1), 24-31.

- [2] Hazan, S. (2020). Deep fake and cultural truth-custodians of cultural heritage in the age of a digital reproduction. In *Culture and Computing: 8th International Conference, C&C 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings 22* (pp. 65-80). Springer International Publishing.
- [3] Sharma, I., Jain, K., Behl, A., Baabdullah, A., Giannakis, M., & Dwivedi, Y. (2023). Examining the motivations of sharing political deepfake videos: the role of political brand hate and moral consciousness. *Internet Research*.
- [4] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 993-1022.
- [5] Gamage, D., Sasahara, K., & Chen, J. (2021). The Emergence of Deepfakes and its Societal Implications: A Systematic Review. *TTO*, 28-39.
- [6] Jones, V. A. (2020). Artificial intelligence enabled deepfake technology: The emergence of a new threat (Doctoral dissertation, Utica College).
- [7] Hussain, S., Neekhara, P., Jere, M., Koushanfar, F., & McAuley, J. (2021). Adversarial deepfakes: Evaluating vulnerability of deepfake detectors to adversarial examples. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 3348-3357).
- [8] Dahlgren, P. (2018). Media, knowledge and trust: The deepening epistemic crisis of democracy. *Javnost-The Public*, 25(1-2), 20-27.
- [9] Langa, J. (2021). Deepfakes, real consequences: Crafting legislation to combat threats posed by deepfakes. *BUL Rev.*, 101, 761.
- [10] Diakopoulos, N., & Johnson, D. (2021). Anticipating and addressing the ethical implications of deepfakes in the context of elections. *New Media & Society*, 23(7), 2072-2098.
- [11] Yu, P., Xia, Z., Fei, J., & Lu, Y. (2021). A survey on deepfake video detection. *Iet Biometrics*, 10(6), 607-624.
- [12] Leibowicz, C. R., McGregor, S., & Ovadya, A. (2021, July). The deepfake detection dilemma: A multistakeholder exploration of adversarial dynamics in synthetic media. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 736-744).
- [13] Yordanova, K. Y., Demiray, B., Mehl, M. R., & Martin, M. (2019, March). Automatic detection of everyday social behaviours and environments from verbatim transcripts of daily conversations. In *2019 IEEE International Conference on Pervasive Computing and Communications (PerCom)* (pp. 1-10). IEEE.
- [14] Barnet, B. A. (2009). Idiomediality: The rise of personalized, aggregated content. *Continuum*, 23(1), 93-99.
- [15] Chen, Z. H., Lu, H. D., & Lu, C. H. (2019). The effects of human factors on the use of avatars in game-based learning: Customization vs. non-customization. *International Journal of Human-Computer Interaction*, 35(4-5), 384-394.
- [16] Newman, M. A. R. K., Gatersleben, B., Wyles, K. J., & Ratcliffe, E. (2022). The use of virtual reality in environment experiences and the importance of realism. *Journal of environmental psychology*, 79, 101733.
- [17] Littlejohn, W. B. (2017). Addicted to Novelty: The Vice of Curiosity in a Digital Age. *Journal of the Society of Christian Ethics*, 37(1), 179-196.
- [18] Bowrey, K. (1998). Ethical boundaries and Internet cultures. *Na*.
- [19] Bulcroft, R., Bulcroft, K., Bradley, K., & Simpson, C. (2000). The management and production of risk in romantic relationships: A postmodern paradox. *Journal of Family History*, 25(1), 63-92.
- [20] DeWall, C. N., Buffardi, L. E., Bonser, I., & Campbell, W. K. (2011). Narcissism and implicit

attention seeking: Evidence from linguistic analyses of social networking and online presentation. *Personality and Individual Differences*, 51(1), 57-62.

[21] Drenten, J., Gurrieri, L., & Tyler, M. (2020). Sexualized labour in digital culture: Instagram influencers, porn chic and the monetization of attention. *Gender, Work & Organization*, 27(1), 41-66.

[22] Pandey, V. (2020). Outraging the Modesty of Women. *Jus Corpus LJ*, 1, 7.

[23] Rana, M. S., Nobi, M. N., Murali, B., & Sung, A. H. (2022). Deepfake detection: A systematic literature review. *IEEE access*, 10, 25494-25513.

[24] Bondi, L., Cannas, E. D., Bestagini, P., & Tubaro, S. (2020, December). Training strategies and data augmentations in cnn-based deepfake video detection. In *2020 IEEE international workshop on information forensics and security (WIFS)* (pp. 1-6). IEEE.

[25] Güera, D., & Delp, E. J. (2018, November). Deepfake video detection using recurrent neural networks. In *2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS)* (pp. 1-6). IEEE.