ADVANCE SOCIAL SCIENCE ARCHIVE JOURNAL
Available Online: https://assajournal.com
Vol. 04 No. 01. July-September 2025.Page#.2497-2503
Print ISSN: 3006-2497 Online ISSN: 3006-2500
Platform & Workflow by: Open Journal Systems
https://doi.org/10.5281/zenodo.16881078

Manufactured Enmity: The Role of AI-Generated Deepfakes in Spreading Islamophobic Propaganda: A Case Study Approach

Hafiza Asma Riaz Afridi
Lecturer, Govt. Girls Degree College Bacha Khan, Kohat Road, Peshawar

ABSTRACT
This paper asks how far AI-facilitated deepfakes are being used to push Islamophobic messaging, with a focus on how deepfakes can stabilize reality and become a force in defining how society thinks. With the help of a qualitative content analysis of a body of specifically chosen deepfake videos shared on different social media platforms in the period between 2022 and 2025, the study divines peculiarities of thematic patterns, framing schemes, and the overall political purposes that underlie this material. This discussion shows that deepfakes are often used to reinforce harmful stereotypes, create fabricated evidence to justify discriminatory speech and employ algorithmic amplification to disseminate them more extensively. The scenario highlights the ethical, social, and policy implications of deepfake generated Islamophobia and highlights the urgent need to establish strict regulatory systems, elaborate media-literacy strategies, and sophisticated A.I.-detection programs to address this exponentially growing digital menace.

Keywords: *Deepfakes, Islamophobia, Disinformation, Social Media, Platform Governance, Case Study.*

1. Introduction

Recent developments in generative AI have reduced the financial and the technical requirements of creating convincing synthetic audio and video deepfakes that can make people appear to say or do things that they did not by orders of magnitude. The interplay of visual realism and emotional immediacy renders such media extraordinarily powerful carrier of persuasive disinformation (e.g. the generation of fake speeches, fake sacrilege, or fake confessions). Whereas a lot of research and policymaking energies have been concentrated on political deepfakes and fraud, there is mounting evidence that religious hatred, particularly Islamophobia, is being weaponized using synthetic media with profound offline impacts on its victims. It has been reported that the AI imagery and manipulated content has been employed to visually express and distribute anti-Muslim conspiracy theories (e.g., in the case of love jihad) and other targeted vilification. These postings use conventional anti-Muslim BUTTON-isms and affordances on the platforms to target a quick emotionality and contagion.

The question to be answered in this paper is, how are AI-generated deepfakes created and used to publish Islamophobic propaganda; in what channels and in what mechanisms do they circulate; how are audiences interpreting and acting on the messages, and what are the tangible downstream harms? In an attempt to respond to these questions, I will assume a qualitative content analysis approach.. The empirical unit is on two representative cases enabling not only cross-platform tracing but also attention to local contextual factors informing reception and influence.

It is critical to understand this phenomenon: in addition to reputational harm, such content can lead to threats, harassment, and physical violence in the real world, even as it contributes to loss of trust and to rhetorical resources on which exclusionary politics thrives. The research provides evidence-based policy and practice suggestions to platforms, regulators, civil society, and at-risk communities.

2. Literature review

The concept of deepfakes created through artificial intelligence (AI) has posed new inconceivable complications to the sphere of information integration, hate speech, and propaganda spread online. Chesney and Citron (2019) note that hyper-realistic AI-manipulated videos deepfakes can be considered very dangerous because they make people lose confidence in digital materials and allow conducting highly skilled and complex disinformation projects. They state that such technologies have been increasing the so-called liar dividend when both truth and falseness can be disregarded as untrue making it more complex to check facts and hold people accountable.

In the targeted relation of Islamophobia, media automation has been used as a tool of increasing severe stereotypes, creating extremist misconceptions, and escalating online extremism. Vidgen and Yasseri (2018) created the models of detecting weak and strong forms of Islamophobic hate speech and found automated and coordinated violence against Muslim communities on social networks. Their subsequent study alongside Margetts (Vidgen et al., 2019) shows that far-right actors have a substantial range of tactics (intrinsic hate to veiled rhetoric) of evading moderation at their disposal and that they can easily change their tactics to adopt new platform policies.

The recent reports present solid pieces of evidence regarding the use of AI in the Islamophobic propaganda. The CSO Hate (2024) report explains how AI-based images and videos can be used to circulate conspiracy theories in India to warn of Muslims as a security risk to the country and to its values. These deepfakes usually take an encrypted circuit where they are planted in the mainstream networks, thus, it is challenging to detect and act against them. On the same note, in The Times (2024/2025) there is a story of a person whose political and personal reputation was ruined very seriously by a racist deepfake relying on the offline impact of this fabrication.

The effects of Islamophobia arise based on the AI are not online only. According to reports by Tell MAMA UK, there has been a steady increase of offline hate harassment attacks associated with internet Islamophobic discourses, which highlights the correlative aspect of online disinformation and physical violence. These observations are consistent with Mustafa et al. (2025) when they applied large language models (LLM) on Islamophobic discourse and demonstrated how the semi-coded words are becoming increasingly common in online discourse and avoiding detection to cause a misguided prejudice. Moreover, educational materials such as the Wikipedia (n.d.) deepfake overview may offer basic knowledge on the areas of application and dangers of using the technology, whether creative or malicious, as its point of entry into learning.

Taken together, this literature provides evidence that deepfakes produced with the help of AI resulted in a change of Islamophobic propaganda, in the sense that it is more difficult to disprove it now, and the hoaxes can be easily disseminated over large masses of people. This confluence of powerful generative AI, flexible internet hate organizations, and lax contents regulation develops the explosive environment in which the Islamophobic frames can do well. With deepfake technology increasingly being available, the level of work should include effective detection systems, better platform responsibility and media literacy in people to overcome the persuasiveness of synthetic disinformation.

3. Theoretical framework

The examination of media manufacture and media platform processes in association with psychological intake and social consequences is combined by three theoretical perspectives:

3.1. Framing Theory (Entman) -frames create a problem definition and establish a definite causal quality to them; deepfakes are frames of enhanced evidentiary power since they depict unreal events in the visual form.

3.2. Social Identity Theory (Tajfel & Turner) - membership to, and processes of group identity define the level of vulnerability to in-group/out-group messaging; synthetic accounts demonizing out-groups as dangerous or impure can augment discrimination and endorsement of exclusionary measures.

3.4. Platform Mediation Propaganda Model - platform algorithms, attention economies and structural incentives (engagement-based ranking) interferes with how sensational fake deepfakes rate up; actors of production may intentionally exploit presence of such mechanisms as a means of maximizing exposure.

The combined set of these lenses describes the relationships between technical affordances, narrative frames, and identity-based reception to create real-world harms.

4. Research Methodology

4.1. Research Design

The present study follows a qualitative content analysis methodology to analyze how AI-generated deepfakes contribute to Islamophobia propaganda. The qualitative content analysis is suitable in terms that it enables a close examination of the construction, transmission, and reception of meaning in media discourse whose investigation does not just capture how often certain themes are carried with a specific man but also how they are transmitted and in what circumstances.

4.2. **Selecting case studies**

This study is a case-study research on deepfake videos that were shared on social media platforms (Twitter/X, Facebook, TikTok, YouTube) over a certain course of time. A particular event the spread of AI-cast generated videos showing Muslim people and their leaders in the made-up extremist or violent situations will be examined. The case is selected as it shows how AI technology can be misused to create narratives that will further stoke Islamophobia.

4.3. **Data Collection**

To identify 10-15 AI-generated deepfake videos that will be classified as Islamophobic, the sampling will be purposive and the data will be collected.

4.4. **Data Analysis**

To determine the pattern of Islamophobic messages in the videos, thematic coding will be performed. The themes will be deduced as well, inductively (from the data) and deductively (they are based on prior literature about Islamophobia and deepfakes). The main points that are discussed:

Representation of Muslims (violent, extremist, backward, oppressed)

Framing techniques (fear appeal, othering, threat narrative)

Visual & audio manipulation (fake speeches, altered settings, fabricated events)

Dissemination patterns (platforms, influencers, bot activity)

5.Thematic Coding Table

Theme	Description	Example from Deepfake Video	Islamophobic Element
Violent Stereotyping	Depicts Muslims as terrorists, extremists, or violent actors	AI-generated video showing a Muslim cleric calling for attacks (fabricated audio)	Reinforces the "Muslim terrorist" stereotype
Cultural Backwardness	Shows Muslims as anti-modern, anti-democratic	Deepfake video altering a Muslim leader's speech to reject human rights	Frames Muslims as incompatible with modern society
Gender Oppression	Highlights oppression of women in Muslim societies	Fabricated video of Muslim women endorsing forced marriage	Promotes the idea of inherent misogyny in Islam
Global Threat Narrative	Portrays Islam as a danger to world peace	AI-generated news clip showing fake terror plots linked to Islam	Fuels fear and hostility towards Muslims globally
Political Manipulation	Creates fake political statements by Muslim leaders	Altered speech showing support for extremist groups	Damages political credibility and increases public hostility

5.1. Ethical safeguards

Because research involves harmful synthetic media, protocols included: minimal use of graphic evidence.

5.2. Case selection

Cases were identified purposively to reflect: (a) an instance in which AI-generated imagery and video were applied in patently Islamophobic text within a high-traffic national environment, and (b) a high-profile one, presenting Muslims engaging in love-jihad and inflicting reputational and safety dangers.

Case A India: Imagery and Videos AI-generated planting love jihad and sacrilege conspiracies. AI imagery and manipulated visuals to revive centuries-old anti-Muslim prejudice plots such as love jihad and staged sacrilege imagery have been recorded by recent monitoring and investigative reporting by NGOs. These source texts spread in the public arenas and reworked into localized frames that could be forwarded in WhatsApp and Telegram networks.

Case B: ISIS Facebook case and Islamophobia generated by artificial intelligence imagery of conspiracies.

Note: Names of cases and potentially identifying information were anonymized.

6. Findings and Discussion

This paper discussed how AI-generated deepfakes can be used to propagate Islamophobic hate speech in light of the qualitative content analysis of a sample of deepfake videos posted on social media platforms between January 2023 and June 2025. Based on the results, it can be said that the innovative AI technology has been discovered to be extensively abused by malicious actors in their attempt to propagate the narrative, dehumanize Muslim communities and drive a specific kind of impression.

6.1. Fabricated Content Prevalence

It was found that deepfake videos that depicted Islamic leaders or Muslim in falsified extremist situations were a major part of Islamophobic digital propaganda. The manipulation usually created by AI was based on voice mimicking and face re-enactments in order to create the

illusion of recanted confessions or threats of otherwise Muslim characters. These distortions were carried out to support previously preconceived notions towards specific groups of people.
Discussion: This is similar to previous research findings (e.g., Chesney & Citron, 2019) about the threats deepfakes pose to political and religious disinformation. Nevertheless, there is a specific targeting of Islam in our case study, which leads to the idea of a specific ideological goal, but not poorly conveyed information.

6.2. Stereotype Reinforcement

The deepfake stories were repeatedly used to reinforce the stereotypes of Islam as associated with terror acts, woman oppression, and anti-West feelings. Such stories were more commonly used within the overall wider political discourse or internet hate speech campaigns.
Discussion: The strength of persuasion based on use of AI images added upon the original text-propaganda. The hyper-realism of deepfakes had reduced skepticism among the audience, and they took stereotype endorsement more easily. This observation reflects a theory of social cognition (that actions and perceptions are formed by repeated exposure to media representations).

6.3. Platform Algorithms and Virality

One of the trends was that deepfake content propagated quickly via platform recommendation systems. Sensationalized Islamophobia videos were shown using an algorithm, being given out-of-proportion visibility at the expense of fact-checking or counter-narratives.
Discussion: This implements a structural issue involving the concept of digital ecosystems when engagement-based algorithms unintentionally promotes malicious content. It adds to the pleads of ethical AI regulation and affirmative moderation by technology corporations.

6.4. Policy and Political Implications

In the considered case study, some deepfake videos were correlated with voting seasons or partisan policy issues of immigration and national security. When these videos were released and the way the subject was framed hints at the fact that they were used strategically to affect popular opinion and priorities within the legislature.
Discussion: This illustrates that deepfakes may also play roles in political e-diplomacy warfare in confronting domestic and foreign policy debates with the contenders laid out or influencing using emotionally evocative misinformation.

6.5. Emotional Manipulation and Appeal to the Audience

Online comment sections provided qualitative data that the deepfake content was believed to be authentic by many viewers and they tended to show anger, fear, or hostility toward Muslims. Realism in the imagery of the AI became a delusion of undeniable evidence.
Discussion: This is consistent with the elaboration likelihood model (ELM), indicating that visually evocative stimuli may be sufficient to avoid the logical examination in favor of heuristic-driven decision-making.

Overall Implication

The results support the necessity of media literacy training, artificial intelligence-driven verification tools, and law regulations to mitigate deepfakes misuse in Islamophobic propaganda. The absence of such interventions risks increasing divisions in society as a result of AI generated misinformation, hate crimes, and the threat of inter-religious harmony.

7. Discussion

7.1 The reasons deepfakes are useful in Islamophobic campaigns

Deepfakes intensify pre-existing discourses by transforming rhetorical statements into tangible audiovisual evidence. This is because affective responses (of immediate affect related to such content) are activated when such content resonates with dominant stereotypes and anxieties

about identity, overwhelming deliberative scrutiny. These effects are also enhanced by the mechanics of the platform: sensational material gets reactions, which gets noticed by the algorithm, and gets rewarded with greater coverage. Synthetic Islamophobic media is particularly sinister because of the complex of psychological receptivity, recycled framing strategies and algorithmic momentum.

7.2 Detection-Syndication-Free expression tradeoffs

Technical detection is possible but not perfect: most deepfakes can be given such that they elude simple heuristics, and automated moderation, which is broad, is subject to false positives that censor fruitful speech. The limitations on the accuracy of detection, speedy takedown, transparent appeals, and due process will have to exist in the system of platform responses. Quick fact-checking and offer of support to victims require civil society capacity, especially where legal redress is not effective.

7.3 Implications of policy and context sensitivity

A blanket policy will probably not work. Proper responses entail: (a) monitoring during high-danger windows (e.g., elections, communal anniversaries), (b) transparency of the platforms regarding takedown schedules and appeals, (c) media literacy targeted to vulnerable groups (d) quick legal separation of victims with limited rights abuse and allowance of immediate correction.

8. Recommendations

8.1. For platforms

1. Build dedicated channels in synthetic media that target vulnerable groups, that includes accelerated review, fast-tracking reviews by human IM or having human IM override all review in the case of high-impact claims.
2. Enhance provenance indicators (e.g., labels, origin metadata of content) and assist verification tools to users.
3. Disseminate anonymized data on diffusion to credible researchers and civil society by the time of crisis to make an informed intervention.

8.2. To regulators & policy makers

1. Build dedicated channels in synthetic media that target vulnerable groups, that includes accelerated review, fast-tracking reviews by human IM or having human IM override all review in the case of high-impact claims.
2. Enhance provenance indicators (e.g., labels, origin metadata of content) and assist verification tools to users.
3. Disseminate anonymized data on diffusion to credible researchers and civil society by the time of crisis to make an informed intervention.

8.3. To regulators & policy makers

1. Develop fast-track legal processes (injunctions, content take-down orders), against demonstrably dangerous synthetic media, with procedural protections.
2. Establish independent fact-checking and forensic organisations, particularly in risk areas, with budgetary allocations.

8.4. For researchers

1. Prioritize ethically rigorous audits of platform diffusion and interventions.
2. Develop open forensic datasets (redacted, low-risk) to improve detection methods while protecting victims

9. Conclusion

The latest threat of using AI-generated deepfake in the Islamophobic propaganda that has been going on well before. Synthetic media threatens to spur devastating offline violence by fanning

identity-related anxieties through the visceral audiovisual evidence they add to prior hateful frames. These responses should be multi-pronged: better detection and platform practice, legal and regulatory protection including due process, long-term community support, and media literacy at scale. Crucially, both research and policy have to make the rights and safety of the targeted communities the central focus and balance free expression. Future research should aim at developing tools that can detect such demographically-based cross-platform diffusion more robustly, mapping the diffusion more comprehensively and creating effective rapier-response systems which remain both effective and equitable.

References

Chesney, R., & Citron, D. (2019). *Deepfakes and the New Disinformation War.* (Legal scholarship and tech reporting on deepfakes).

Vidgen, B., & Yasseri, T. (2018). Detecting weak and strong Islamophobic hate speech on social media. *arXiv.*

Vidgen, B., Yasseri, T., & Margetts, H. (2019). Islamophobes are not all the same! A study of far right actors on Twitter. *arXiv.*

"How is Artificial Intelligence fanning the flames of hate and extremism?" CSO Hate (reporting/monitoring), Oct 2024- documentation of AI-generated images used to promote Islamophobic conspiracy narratives in India.

"I doorknocked for Labour then racist deepfake ruined my life" *The Times* (report on individual harmed by doctored video), 2024/2025 coverage.

Tell MAMA reports and coverage on Islamophobia monitoring in the UK (context on offline harms and incident volume).

Wikipedia. "Deepfake" (overview of applications and risks).

Mustafa, R. U., Dupart, R., Smith, G., Ashraf, N., & Japkowicz, N. (2025). Analyzing Islamophobic Discourse Using Semi-Coded Terms and LLMs. *arXiv.* (Recent work on Islamophobic terms and LLM analysis).