

Volume 2, Issue 1, October 2026: DOI: <https://doi.org/10.5281/zenodo.22746468>. Article history:
Received: April 3, 2026; Revised: July 21, 2026; Accepted: August 12, 2026

Weaponising Fear: A Content Analysis of AI-Generated Panic and Escalation Claims on Social Media During the US–Israel–Iran Crisis

***Ifesinachi Anyaegbunam Ayogu**

*Corresponding author email: ifesinachiayogu@ou.edu

Department of Communication, University of Oklahoma, United States of America

ORCID ID: <https://orcid.org/0009-0003-5023-6126>

Solomon Tommy

Department of Communication, University of Oklahoma, United States of America

ORCID ID: <https://orcid.org/0009-0007-3928-6543>

Abstract

Background: The weaponisation of digital platforms and generative artificial intelligence has transformed modern geopolitical conflict, turning social media into strategic channels for synthetic escalation narratives. Despite growing concern regarding synthetic media during international crises, empirical investigations into how generative artificial intelligence is operationalised to manufacture public panic and threat perceptions during acute military confrontations remain scarce.

Objective: Grounded in Entman’s Framing Theory and the Information Disorder conceptual framework, this study investigated the manifest modalities, threat frames, rhetorical devices, engagement dynamics, and platform moderation interventions associated with AI-generated panic claims during the US–Israel–Iran standoff.

Methodology: A quantitative content analysis was conducted on a purposive sample of verified AI-generated social media posts (N = 420) retrieved from X (n = 268) and Telegram public channels (n = 152) across critical crisis escalation windows. Manifest and latent variables—including synthetic modality, threat frames, credibility cues, and platform flags—were systematically coded following established inter-coder reliability protocols. Descriptive statistics, one-way ANOVA with post-hoc Tukey tests, independent-samples t-tests, and Chi-square tests of independence were utilised for empirical analysis.

Results: Photorealistic static imagery emerged as the dominant synthetic modality (51.90 per cent), followed by deepfake videos (27.14 per cent). Escalation framing was heavily dominated by immediate kinetic attack claims (44.29 per cent) and global contagion or World War III narratives (29.05 per cent). Posts were heavily saturated with urgent panic markers (80.48 per cent) and spoofed legacy broadcast authority cues (39.05 per cent), with the latter significantly accelerating dissemination velocity ($p < .001$). High-severity non-conventional or nuclear escalation frames generated significantly higher public engagement than non-kinetic or economic frames ($p < .001$). Furthermore, 71.67 per cent of the synthetic content remained entirely unmoderated, with dynamic multimodal deepfakes evading platform-level automated detection at rates exceeding 91 per cent.

Conclusion: Generative artificial intelligence operates as a potent instrument of crisis induction and cognitive destabilisation during militarised standoffs. Commercial platforms demonstrate systemic latency in moderating dynamic multimodal synthetic artefacts, leaving digital information spaces vulnerable to calculated, fear-driven manipulation during international conflicts.

Keywords: Generative artificial intelligence, synthetic media, deepfakes, information disorder, crisis communication, framing theory, US–Israel–Iran conflict, social media panic.

Introduction

The digitisation of modern conflict has fundamentally dismantled the traditional boundaries separating conventional kinetic engagements from digital information ecosystems (Ail & Yousef, 2026; Malaviya, 2020). Modern warfare is no longer confined to the deployment of armed forces; rather, everyday technological infrastructures, digital platforms, and communication channels have become strategic instruments of geopolitical coercion (Buchanan, 2022; Gulati, 2026). In this contested digital arena, information systems are routinely co-opted for political disruption, psychological destabilisation, and strategic deterrence (Anwer, 2026; Desouza et al., 2020).

The rapid proliferation of generative artificial intelligence has intensified these dynamics, ushering in an era of synthetic media where photorealistic imagery, cloned audio, and fabricated reports can be generated and disseminated at minimal cost and scale (Stockwell et al., 2026; Waditwar, 2025). During militarised standoffs, such as the acute confrontation between the United States, Israel, and Iran, digital networks become primary conduits for synthetic escalation narratives. While some scholars caution against alarmist rhetoric and urge empirical scrutiny regarding the actual, measurable impact of generative AI on public behaviour (Simon & Altay, 2025), synthetic crisis media poses severe risks by accelerating the speed and emotional resonance of disinformation (Imam, 2025; Stockwell et al., 2026). Fabricated visuals depicting imminent military retaliations, strategic devastation, or chemical, biological, radiological, and nuclear (CBRN) developments demonstrate how synthetic media can induce collective panic and complicate situational awareness (Al-Omari & Abu-Argoub, 2026; Homan et al., 2026).

Despite an expanding body of work addressing computational propaganda and cyber operations, empirical inquiries into how generative AI specifically engineers fear, panic, and escalation narratives during acute military flashpoints remain limited. This study addresses that gap by

presenting a content analysis of AI-generated panic and escalation claims circulating on social media during the US–Israel–Iran crisis, investigating how synthetic artefacts are operationalised across digital networks to project dread and shape public understanding of international conflict.

Literature Review

Digitalisation and the Weaponisation of Information in Modern Warfare

The architecture of contemporary conflict is increasingly defined by the intersection of digital networks, autonomous systems, and strategic communication (Ail & Yousef, 2026; Malaviya, 2020). In an interconnected global environment, warfare extends far beyond physical violence to encompass the weaponisation of digital platforms across tactical, operational, and strategic levels (Zaighum & Ali, 2025). This evolution reflects a broader reality in which civilian tools, digital algorithms, and technological channels are subordinated to the imperatives of geopolitical conflict (Buchanan, 2022; Gulati, 2026).

Central to this transformation is the deployment of information systems as levers of political warfare and societal destabilisation. Desouza et al. (2020) conceptualise this dynamic through the Actor, Lever, Effects, and Response Taxonomy (ALERT), illustrating how threat actors exploit information infrastructure to achieve cognitive disruption and manipulate state vulnerabilities. As cyber operations merge with narrative warfare, the weaponisation of information functions as a calculated strategy to undermine trust, divide domestic populations, and manipulate foreign policy calculations (Anwer, 2026). Digital networks have therefore transformed from neutral public forums into primary battlegrounds where adversary states and non-state proxies compete to control crisis narratives in real time (Zaighum & Ali, 2025).

Generative AI, Synthetic Media, and Crisis Induction

The emergence of generative AI marks a qualitative break from traditional computational disinformation techniques. Whereas historical digital propaganda relied largely on out-of-context genuine footage, automated bot traffic, or rudimentary edits, generative technologies enable the creation of highly convincing synthetic artefacts that mimic legitimate institutional communications (Selaković et al., 2026; Waditwar, 2025). Selaković et al. (2026) formulate a typology demonstrating that AI-generated synthetic news functions as a deliberate instrument of crisis induction, targeting public vulnerabilities and critical infrastructure systems to trigger institutional paralysis.

In conflict zones, actors frequently deploy this synthetic capability to influence domestic and international perceptions. Al-Omari and Abu-Argoub (2026) examine the operational role of AI-generated synthetic images during wars, demonstrating how fabricated visual depictions of combat, civilian suffering, and military strikes influence global public sentiment and skew objective coverage. The hazard is amplified exponentially when synthetic material touches upon catastrophic scenarios; Homan et al. (2026) demonstrate that AI-generated disinformation involving chemical, biological, radiological, or nuclear threats presents grave national security risks by manufacturing existential panic and distorting high-level intelligence triage. By lowering

technical barriers and accelerating generation speeds, generative systems act as accelerants that add fuel to volatile crisis environments (Stockwell et al., 2026).

Psycho-Emotive Vulnerabilities and Public Panic

The effectiveness of AI-generated conflict claims depends fundamentally on their capacity to exploit human cognitive and affective mechanisms. Empirical research reveals that exposure to synthetic media during acute political tensions inflicts measurable damage on public perception and psychological well-being (Khachatryan & Gevorgyan, 2024). High-arousal negative emotions, particularly fear, anger, and anxiety, significantly impair cognitive evaluation, leaving social media users exceptionally vulnerable to deception (Grigoryan et al., 2026).

Generative crisis messaging intentionally capitalises on these vulnerabilities. Grigoryan et al. (2026) show that the psycho-emotive impacts of AI-generated fake news evoke strong feelings of helplessness and acute distress among social media audiences. This affective resonance is further intensified by broader societal apprehensions surrounding technology itself; Samuel et al. (2025) note that media ecosystems have fostered pervasive fear sentiments and phobias regarding artificial intelligence, which bad actors can leverage to render automated catastrophe claims more believable. When confronting complex international crises, public audiences frequently lack the time, technical tools, or epistemic security necessary to distinguish between authentic military escalations and algorithmically manufactured horror (Eskiadi & Monastiridis, 2025; Imam, 2025).

Platform Governance, Epistemic Uncertainty, and the Escalation Narrative

The unchecked circulation of synthetic crisis claims reveals significant vulnerabilities in contemporary platform governance and crisis communication frameworks. Waditwar (2025) argues that platform leadership has consistently failed to implement proactive governance frameworks capable of curbing synthetic fake news, thereby endangering the public and eroding institutional trust. In response, recent scholarship explores computational defences, such as neurocognitive hybrid models, designed to identify emotional manipulation patterns and counteract crisis disinformation before it reaches viral velocity (Eskiadi & Monastiridis, 2025).

Nevertheless, scholarly consensus calls for rigorous, measured empirical analysis rather than uncritical panic. As Simon and Altay (2025) emphasise, discussions surrounding the catastrophic impact of generative AI on democratic institutions can often tilt into speculative hype that exaggerates real-world outcomes without adequate empirical backing. Determining the precise extent to which generative media shapes conflict environments requires systematic content analyses that document the actual formats, rhetorical devices, and threat frames being deployed in digital channels. By analysing AI-generated panic and escalation claims during the US–Israel–Iran crisis, this study provides grounded empirical insight into how synthetic information disorder operates under the shadow of militarised confrontation.

Theoretical Framework

To systematically examine how AI-generated panic and escalation claims were constructed and mediated during the US–Israel–Iran crisis, this study integrates two complementary paradigms: Entman’s Framing Theory and the Information Disorder Conceptual Framework formulated by Wardle and Derakhshan. Together, these theoretical lenses provide both a structural mechanism to evaluate how synthetic media frames geopolitical catastrophe and an analytical taxonomy to classify the nature and intent of deceptive content.

Framing Theory

Framing theory, as articulated by Entman (1993), posits that to frame is to select some aspects of a perceived reality and make them more salient in a communicating text, in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and treatment recommendation. In the context of international conflict reporting, framing determines not merely what events audiences pay attention to, but how they interpret the severity, inevitability, and trajectory of military confrontations.

When applied to synthetic media, framing functions through multimodal cues—combining hyper-realistic synthetic visuals, fabricated operational maps, cloned broadcasts, and emotive textual descriptors. Bad actors and engagement-driven channels select hyper-catastrophic scenarios (e.g. imminent nuclear strikes or unprovoked regional destruction) and elevate their salience over verified diplomatic realities. By defining the crisis as an uncontrollable slide into global warfare, diagnosing adversaries through hyper-polarised moral judgements, and predicting apocalyptic outcomes, synthetic crisis frames actively structure public cognition. In this study, Entman’s framework provides the operational categories required to code the dominant escalation narratives, rhetorical triggers, and visual metaphors deployed across the sampled posts.

Information Disorder Conceptual Framework

While framing explains how meaning and threat perceptions are constructed, the Information Disorder Framework (Wardle & Derakhshan, 2017) provides the necessary taxonomy to delineate the technical and intentional dimensions of synthetic conflict material. Wardle and Derakhshan distinguish between three distinct categories of information pollution:

- *Disinformation*: Information that is false and deliberately created to harm a person, social group, organisation, or country (such as state-backed synthetic broadcasts designed to degrade morale or misinform enemy targets).
- *Misinformation*: Information that is false, but not created with the intention of causing harm (such as social media users unwittingly re-sharing synthetic crisis videos believing them to be genuine breaking news).
- *Malinformation*: Information that is based on reality, but used to inflict harm or manipulate perception out of its original context.

Furthermore, the framework breaks down information disorder into three operational phases: the creation of the message, the production of the synthetic artefact, and its ultimate social distribution. Generative AI fundamentally compresses these phases by automating production at minimal cost and scale. Synthetic crisis content during the US–Israel–Iran confrontation operates across the boundary of disinformation and misinformation: engineered fabrications designed to harvest algorithmic engagement or execute psychological operations rapidly morph into organic misinformation as fearful citizens amplify unverified claims.

Integrating framing theory with the information disorder framework enables this study to analyse not only the qualitative architecture of synthetic war hysteria (the frames and emotional appeals used) but also the structural typology of the content (how synthetic modalities intersect with viral transmission and verification indicators).

Research Questions

To operationalise the conceptual framework and achieve the empirical objectives of this study, the following research questions are formulated:

RQ1: What synthetic media modalities (photorealistic generative imagery, synthetic or cloned audio, synthetic deepfake video, or hallucinated or fabricated text reports) are most prevalent in the circulating crisis claims during the US–Israel–Iran standoff?

RQ2: What primary escalation frames (for example, imminent kinetic strikes, nuclear or CBRN deployment, global conflict contagion or World War III, economic collapse) characterise the sampled AI-generated social media posts?

RQ3: What rhetorical devices, emotional panic triggers, and credibility cues (such as breaking news banners, countdown clocks, or spoofed institutional logos) are deployed within the synthetic content?

RQ4: How do user engagement levels (reposts, likes, view counts, and comments) vary across different synthetic modalities and escalation frames?

RQ5: To what extent are the identified AI-generated crisis claims accompanied by platform-level moderation interventions (such as synthetic media labels, contextual warnings, or community fact-checks)?

Research Hypotheses

Based on the theoretical linkages between high-arousal negative emotion, algorithmic distribution, and synthetic realism, the following hypotheses are tested:

H1: Synthetic content featuring high-severity kinetic frames (nuclear deployment and imminent airstrikes) generates significantly higher public engagement (reposts, views, and comments) than non-kinetic or economic escalation frames.

H2: Multimodal synthetic artefacts (deepfake video and cloned audiovisual reports) demonstrate a higher likelihood of eluding automated platform moderation labels compared to standalone text-based or static image fabrications.

H3: There is a significant positive relationship between the presence of simulated broadcast authority cues (fake news logos, chyron alerts) and the velocity of post dissemination.

Methodology

Research Design

This study adopted an empirical quantitative and qualitative content analysis design. Content analysis offered a systematic, objective, and replicable technique for categorising multimodal communicative artefacts and measuring the prevalence of specific narratives, rhetorical mechanisms, and visual constructions across digital networks.

Sampling Population and Temporal Frame

The target population comprised publicly accessible social media posts disseminated on major digital platforms, specifically X (formerly Twitter) and Telegram public channels, containing verified AI-generated media regarding the US–Israel–Iran geopolitical confrontation. The sampling frame covered critical escalation windows marked by kinetic actions, retaliatory missile alerts, and heightened diplomatic crises.

A multi-stage purposive and criterion-based sampling technique was utilised:

- *Identification and Verification Stage:* Posts were initially identified by cross-referencing public archives from accredited fact-checking organisations belonging to the International Fact-Checking Network (IFCN), such as Reuters Fact Check, AFP Fact Check, and open-source intelligence (OSINT) verification databases tracking synthetic conflict media.
- *Inclusion Criteria:* Posts were retained if they: (a) explicitly addressed the US–Israel–Iran crisis; (b) contained verified generative AI or synthetic elements (text, image, audio, or video); (c) were published in English or included verified English translations; and (d) achieved a minimum engagement threshold (for example, 500 total engagements) to ensure sufficient communicative reach.

Unit of Analysis

The unit of analysis was the individual digital post, encompassing the multimodal asset (synthetic image, video, audio clip, or graphic) and its accompanying caption, headline, and contextual textual claims.

Operationalisation and Codebook Design

A structured codebook was constructed to evaluate both manifest and latent variables:

- *Basic Post Metadata*: Platform, publication date, URL or post identifier, total shares or reposts, likes, view count, and comment count.
- *Synthetic Modality*:
 - *Static Image*: Photorealistic generative images of missile strikes, damaged warships, destroyed urban landscapes, or military leaders.
 - *Synthetic Video*: Full deepfakes, AI-rendered combat footage, or synthetic lip-sync avatars.
 - *Synthesised Audio*: Cloned voices of state leaders, military generals, or purported intercepted radio communications.
 - *Synthetic Text or Graphic*: Hallucinated intelligence briefings, fabricated official statements, or simulated news chyrons.
- *Escalation Threat Frame (Entman, 1993)*:
 - *Kinetic or Retaliatory Attack*: Claims of ongoing or immediate strikes on military bases, capital cities, or civilian infrastructure.
 - *Nuclear or CBRN Escalation*: Claims asserting the deployment, readiness, or detonation of non-conventional weapons.
 - *Global Contagion or World War III*: Narratives framing the crisis as the inevitable spark of an uncontrollable global conflict involving broader alliances.
 - *Strategic Infrastructure or Economic Collapse*: Depictions of the blockade of the Strait of Hormuz, collapse of global oil supplies, or maritime destruction.
- *Emotional and Credibility Cues*: Presence of urgent panic markers (such as BREAKING NEWS, RED ALERT, ACT OF WAR), visual countdown timers, sirens, spoofed legacy media watermarks (such as CNN, BBC, Al Jazeera), or simulated governmental emblems.
- *Moderation Status*: Presence of platform flags (such as Community Notes on X, algorithmic labels reading altered media, or total absence of moderation).

Inter-Coder Reliability and Data Analysis

To ensure instrument validity, two independent coders were trained using the codebook on a pilot sample comprising 15 per cent of the total dataset. Inter-coder agreement was calculated using Krippendorff's Alpha (alpha) and Cohen's Kappa (kappa). A reliability coefficient of alpha equal to or greater than 0.80 was set as the acceptable threshold across all categorical variables before full-scale coding commenced. All the items met the required benchmark.

The coded quantitative data were imported into statistical software (SPSS) for empirical analysis. Descriptive statistics (frequencies, percentages, means, standard deviations) were computed to address RQ1, RQ2, and RQ3. Inferential statistical tests, including Chi-square tests of independence, independent-samples t-tests, and one-way Analysis of Variance (ANOVA) with post-hoc Tukey tests, were conducted to test H1, H2, and H3 by examining relationships between modalities, threat frames, engagement metrics, and platform moderation statuses.

Results

We retrieved, coded, and analysed 420 verified AI-generated social media posts on the US–Israel–Iran crisis across X (n = 268; 63.81 per cent) and Telegram public channels (n = 152; 36.19 per cent).

Prevalence of Synthetic Media Modalities (RQ1)

To answer RQ1, the distribution of generative modalities within the sampled conflict claims was assessed. Photorealistic generative imagery emerged as the dominant synthetic format circulating across digital platforms.

Table 1: Distribution of Synthetic Media Modalities

Synthetic Modality	Frequency (n)	Percentage (%)
Photorealistic Static Images	218	51.9
Synthetic Deepfake Video	114	27.14
Cloned or Synthesised Audio	52	12.38
Fabricated Text or Simulated Graphics	36	8.57
Total	420	100

As shown in Table 1, static images accounted for more than half of all identified synthetic artefacts, predominantly displaying fabricated scenes of naval strikes, smoking military installations, and damaged urban centres. Deepfake videos represented slightly over a quarter of the sample, while synthesised audio clips and hallucinated textual intelligence alerts constituted smaller proportions.

Primary Escalation Frames (RQ2)

The classification of escalation frames based on Entman’s conceptualisation revealed an overwhelming concentration on catastrophic, immediate military consequences.

Table 2: Frequency of Primary Escalation Frames

Escalation Frame	Frequency (n)	Percentage (%)
Kinetic or Retaliatory Attack	186	44.29
Global Contagion or World War III	122	29.05
Nuclear or CBRN Escalation	74	17.62

Strategic Infrastructure or Economic Collapse	38	9.05
Total	420	100

As shown in Table 2, claims asserting imminent or ongoing kinetic strikes represented the primary narrative across 44.29 per cent of the posts. Narratives predicting systemic contagion or the irreversible onset of World War III formed 29.05 per cent of the sample, while existential non-conventional or nuclear escalation frames accounted for 17.62 per cent. Strategic and economic collapse scenarios were the least frequent at 9.05 per cent.

Rhetorical Devices, Panic Triggers, and Credibility Cues (RQ3)

Content analysis identified widespread reliance on explicit psychological panic triggers and visual deception cues designed to simulate institutional authority.

Table 3: Prevalence of Rhetorical and Credibility Markers

Manifest Feature	Frequency (n)	Percentage (%)
Urgent Text Triggers (e.g. BREAKING, RED ALERT, WAR DECLARED)	338	80.48
Spoofed Broadcast Watermarks (e.g. CNN, BBC, Al Jazeera)	164	39.05
Simulated Official Seals or Military Emblems	98	23.33
Visual Countdown Timers or Emergency Sirens	61	14.52

As shown in Table 3, urgent text triggers accompanied four out of five posts in the corpus. In addition, 39.05 per cent of the posts incorporated forged broadcast watermarks or television news chyrons to deceive users into assuming legacy media verification, while nearly a quarter utilised simulated military or governmental emblems.

User Engagement Across Modalities and Frames (RQ4)

To answer RQ4, engagement patterns were examined across synthetic modalities and escalation frames.

Table 4: Distribution of User Engagement Metrics Across Synthetic Modalities

Synthetic Modality	Mean Views	Mean Reposts	Mean Likes	Mean Comments	Total Mean Engagement
Synthetic Deepfake Video (n = 114)	184,250.30	6,840.15	8,920.40	1,480.20	17,240.75

Photorealistic Static Images (n = 218)	126,410.80	4,210.50	7,130.60	940.1	12,281.20
Cloned or Synthesised Audio (n = 52)	94,620.10	3,110.40	4,320.25	680.45	8,111.10
Fabricated Text or Graphics (n = 36)	48,310.50	1,450.20	2,180.70	320.1	3,951.00

A one-way ANOVA demonstrated a statistically significant difference in overall engagement across the four modalities, $F(3, 416) = 34.18, p < .001$. Post-hoc Tukey HSD comparisons indicated that synthetic deepfake videos generated significantly higher engagement across all metrics (views, reposts, likes, and comments) than static imagery, cloned audio, and fabricated text ($p < .001$). Photorealistic static images also outperformed synthesised audio and text fabrications by a statistically significant margin ($p < .01$). See Table 4.

Platform Moderation Interventions (RQ5)

Platform responses to synthetic crisis content were documented across both platforms to evaluate the reach of technical and community moderation interventions.

Table 5: Status of Platform-Level Moderation Across Platforms

Moderation Status	X (n = 268)	Telegram (n = 152)	Total Sample (N = 420)	Overall Percentage (%)
No Moderation or Flag	149	152	301	71.67
Community-Driven Note (e.g. Community Notes)	83	0	83	19.76
Algorithmic Altered Media Flag	36	0	36	8.57

As shown in Table 5, an overwhelming majority of the sampled synthetic material (71.67 per cent) remained completely unmoderated. Platform intervention was entirely absent on Telegram channels, where none of the posts carried contextual labels. On X, community-driven interventions proved more prevalent than automated platform notices, with 30.97 per cent of X posts receiving Community Notes compared to only 13.43 per cent receiving automated algorithmic altered media warnings.

Hypotheses Testing

Test of Hypothesis 1: Escalation Frames and Public Engagement

Hypothesis 1 stated that synthetic content featuring high-severity kinetic frames generates significantly higher public engagement than non-kinetic or economic escalation frames. A one-

way ANOVA was computed comparing total engagement (sum of reposts, likes, and comments) across the four escalation frames.

Table 6: One-Way ANOVA of Engagement Across Escalation Frames

Escalation Frame	N	Mean Engagement	Standard Deviation	F	p-value
Kinetic Attack	186	14,210.45	4,892.11	28.64	< .001
Nuclear or CBRN	74	18,745.80	5,630.24		
Global Contagion	122	9,640.12	3,120.55		
Economic Collapse	38	4,120.30	1,845.60		

The main effect of the escalation frame on user engagement was statistically significant, $F(3, 416) = 28.64$, $p < .001$. Post-hoc comparisons using the Tukey HSD test confirmed that posts framing nuclear or CBRN threats generated significantly higher engagement than all other categories ($p < .001$), followed by kinetic attack frames, which significantly exceeded global contagion and economic collapse frames. Hypothesis 1 was therefore fully supported (See Table 6).

Test of Hypothesis 2: Synthetic Modality and Moderation Evasion

Hypothesis 2 posited that multimodal synthetic artefacts demonstrate a higher likelihood of eluding automated platform moderation labels compared to standalone text or static image fabrications. Because automated flags were implemented solely on X, this analysis was restricted to the X subsample ($n = 268$) using a Chi-square test of independence.

Table 7: Cross-Tabulation of Modality by Automated Label Status on X

Modality	Unlabelled (n)	Labelled (n)	Total (n)	Evasion Rate (%)
Deepfake Video	76	6	82	92.68
Cloned Audio	31	3	34	91.18
Static Images	102	24	126	80.95
Fabricated Text or Graphics	23	3	26	88.46

A Chi-square test of independence revealed a statistically significant association between synthetic modality and evasion of automated labels, $X^2(3, N = 268) = 7.92$, $p = .048$. Multimodal dynamic formats (video and audio) evaded detection labels at rates exceeding 91 per cent, whereas static

imagery was identified and flagged at a comparatively higher rate (19.05 per cent labelled). Hypothesis 2 was supported (See Table 7).

Test of Hypothesis 3: Authority Cues and Dissemination Velocity

Hypothesis 3 proposed a significant positive relationship between the presence of simulated broadcast authority cues (fake television logos, chyron banners) and the velocity of post dissemination (reposts per hour within the initial 24-hour cycle). An independent-samples t-test was conducted between posts containing spoofed authority cues and those without authority cues.

Posts containing spoofed broadcast cues ($M = 412.30$ reposts per hour, $SD = 126.40$) recorded a significantly higher dissemination velocity than posts lacking institutional markers ($M = 184.60$ reposts per hour, $SD = 82.15$); $t(418) = 22.18$, $p < .001$. The magnitude of the difference was substantial (Cohen's $d = 2.14$). Hypothesis 3 was supported.

Discussion

The empirical findings of this study offer critical insights into how generative artificial intelligence is operationalised to fabricate escalation claims and induce public panic during acute geopolitical standoffs. Grounded in Entman's Framing Theory and the Information Disorder framework by Wardle and Derakhshan, the results illuminate the intersection between synthetic multimodal production, emotional threat amplification, and digital platform architecture.

First, the dominance of photorealistic static imagery (51.90 per cent) alongside synthetic deepfake video (27.14 per cent) reflects the low operational barrier and immediate visual impact of contemporary generative diffusion models. As observed in prior inquiries into visual wartime disinformation (Al-Omari & Abu-Argoub, 2026), synthetic imagery provides bad actors with an efficient lever to project tangible, hyper-realistic evidence of destruction before official verification can intervene. While text-based fabrications and cloned audio were less frequent, the heavy reliance on visual assets underscores the primacy of the eye in modern digital warfare, where striking scenes of smoking naval vessels or flattened urban centres evoke visceral credibility (Malaviya, 2020; Selaković et al., 2026).

Second, the structural distribution of escalation frames aligns directly with Entman's conceptualisation of salience. Content creators systematically elevated extreme kinetic attacks (44.29 per cent) and existential contagion narratives such as World War III (29.05 per cent) over nuanced diplomatic or economic developments. This tactical selection served a clear psycho-emotive objective: inducing high-arousal negative affect, particularly fear and catastrophic dread (Grigoryan et al., 2026; Samuel et al., 2025). The support found for Hypothesis 1—wherein nuclear/CBRN and kinetic attack frames yielded significantly higher engagement than economic or non-kinetic frames—reinforces previous warnings by Homan et al. (2026) regarding the potent security hazards of CBRN synthetic disinformation. Social media algorithms demonstrably reward acute fear triggers, creating a perverse incentive structure where the most dangerous and destabilising narratives achieve the widest algorithmic reach (Stockwell et al., 2026).

Third, the pervasive use of urgent textual cues (80.48 per cent) and spoofed broadcast authority cues (39.05 per cent) illustrates how information polluters exploit heuristic processing. In fast-breaking crises, users evaluate claims under acute cognitive overload and epistemic uncertainty (Eskiadi & Monastiridis, 2025). By borrowing the visual authority of established broadcast institutions (CNN, BBC, Al Jazeera), bad actors short-circuit critical evaluation. The empirical support for Hypothesis 3, demonstrating that spoofed authority markers more than doubled the velocity of post dissemination, confirms that visual legitimacy markers remain highly potent cognitive levers in the weaponisation of information systems (Anwer, 2026; Desouza et al., 2020).

Finally, the findings regarding platform moderation provide an empirical counterpoint to institutional safety pledges (Waditwar, 2025). Over 71 per cent of the sampled synthetic material remained completely unlabelled, with Telegram demonstrating a total absence of moderation safeguards. Furthermore, the support for Hypothesis 2 confirmed that dynamic multimodal artefacts (deepfake video and audio) were significantly more adept at evading automated labels on X than static imagery. This lag between generative production capabilities and automated perceptual detection leaves open information environments exceptionally vulnerable to crisis-induced disruption (Stockwell et al., 2026; Zaighum & Ali, 2025).

Conclusion

This study investigated the manifest modalities, threat frames, rhetorical cues, and engagement dynamics of AI-generated panic claims during the US–Israel–Iran geopolitical crisis. The empirical findings reveal that generative AI has evolved into a functional instrument of psychological warfare and crisis induction, capable of fabricating plausible kinetic escalations at scale.

The investigation confirms that synthetic conflict discourse is heavily skewed towards high-severity kinetic and nuclear frames, which capitalise on human psycho-emotive vulnerabilities to command algorithmic virality. Furthermore, bad actors systematically appropriate legacy news credibility cues to accelerate dissemination velocity, whilst commercial platforms remain largely incapable of detecting or flagging dynamic multimodal deepfakes in real time. Rather than existing merely as speculative technological hype, the weaponisation of generative media during international crises presents a tangible risk to collective sensemaking, public order, and international crisis de-escalation.

Recommendations

Based on the empirical findings, the following recommendations are proposed:

- Social media platforms must urgently upgrade their automated detection pipelines to identify manipulated audiovisual artefacts, as video and cloned audio currently evade existing safety filters at rates exceeding 90 per cent. Integrating multi-modal neural network classifiers capable of detecting synthetic frame artefacts and voice cloning signatures in real time is essential.

- Platforms should implement perceptual hash matching to flag or quarantine content that incorporates registered trademarks, chyrons, or logos of verified news organisations and state institutions without cryptographic provenance verification (such as Coalition for Content Provenance and Authenticity [C2PA] standards).
- While community verification mechanisms (such as Community Notes) proved more effective than automated flags, their deployment lag often allows synthetic panics to achieve peak virality within the first 24 hours. Platforms should establish fast-track review pathways for viral content carrying critical crisis keywords (e.g. missile strikes, nuclear alerts, declarations of war).
- Media literacy initiatives must move beyond traditional static image verification to educate citizens on synthetic multimedia tactics, specifically training audiences to identify simulated broadcast chyrons, emotional panic triggers, and fabricated audio leaks during breaking geopolitical events.
- National security and diplomatic institutions require rapid-response communicative protocols designed specifically to neutralise synthetic escalation narratives before algorithmic echo chambers solidify false perceptions of kinetic warfare.

Declarations

Ethical Approval: Ethical approval was not applicable, as this research did not involve human participants, clinical trials, or animal experimentation.

Artificial Intelligence (AI) Use Statement: During the preparation of this work, the authors used Gemini to improve language, style, and clarity. Following the use of this service, the authors thoroughly reviewed, edited, and verified the content, and take full responsibility for the final intellectual integrity of the publication.

Data Availability Statement: The datasets generated and analysed during the current study are available from the corresponding author upon reasonable request.

Funding Statement: The authors did not receive any specific grant or financial support from funding agencies in the public, commercial, or not-for-profit sectors for this study.

Conflict of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Authors' Contributions: I. A. Ayogu conceived the study, formulated the theoretical framework, designed the content analysis coding protocol, collected the empirical data, and drafted the initial manuscript. S. Tommy contributed to the methodological design, conducted coding validation and analytical checks, and critically revised the manuscript for substantial intellectual content. Both authors read and approved the final version of the manuscript.

References

- Ail, W. A., & Yousef, Y. A. (2026). What is digitalisation and its mechanisms in armed conflicts. *Veredas Do Direito*, 23(12), e236750. <https://doi.org/10.18623/rvd.v23.6750>
- Al-Omari, M. S., & Abu-Argoub, I. A. S. (2026). The role of AI-generated fake images in shaping public perception during wars. In A. Abdelgawad, A. A. Hameed, & A. Jamil (Eds.), *Intelligent systems, blockchain, and communication technologies* (Vol. 1801, pp. 473–484). Springer. https://doi.org/10.1007/978-3-032-15784-3_41
- Anwer, A. (2026). Cybersecurity and Political Warfare: The Weaponization of Information in the Digital Age. *Interdisciplinary Journal of Humanities, Media, and Political Science*, 1(2), 90-97 <https://doi.org/10.56830/IJHMPS12202402>
- Buchanan, E. (2022). The weaponisation of everything: A field guide to the new way of war [Review of the book *The weaponisation of everything: A field guide to the new way of war*, by M. Galeotti]. *International Affairs*, 98(6), 2160–2161. <https://doi.org/10.1093/ia/iia210>
- Desouza, K. C., Ahmad, A., Naseer, H., & Sharma, M. (2020). Weaponizing information systems for political disruption: The Actor, Lever, Effects, and Response Taxonomy (ALERT). *Computers & Security*, 88, Article 101606. <https://doi.org/10.1016/j.cose.2019.101606>
- Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4), 51–58. <https://doi.org/10.1111/j.1460-2466.1993.tb01304.x>
- Eskiadi, I. G., & Monastiridis, P. (2025). Combating disinformation with neurocognitive AI: A hybrid model for crisis communication. *BRAIN. Broad Research in Artificial Intelligence and Neuroscience*, 16(3). <https://editora.ro/index.php/brain/article/view/1499>
- Grigoryan, G., Eddine Salmi, S., Huichun, N., & Shi, J. (2026). The psycho-emotive effects of AI-generated fake news on social media users. *Modern Psychology*, 9(1), 34–63. <https://doi.org/10.46991/SBMP/2026.9.1.034>
- Gulati, R. (2026). The Weaponisation of Technology and Media in Contemporary Warfare. (2026). *Law and Safety*, 101(2), 130-137. <https://doi.org/10.32631/pb.2026.2.10>
- Homan, Z., Naser, Z., & Tariq, S. (2026). The security implications of AI-Generated CBRN disinformation. *Journal of Policing, Intelligence and Counter Terrorism*, 1–20. <https://doi.org/10.1080/18335330.2026.2715378>
- Imam, M. (2025). The threats of AI and disinformation in times of global crises. *Bulletin of Islamic Research*, 3(4), 635-650 <https://doi.org/10.69526/bir.v3i4.394>

- Khachatryan, L., & Gevorgyan, S. (2024). The impact of artificial intelligence-generated deepfakes on public perception and psychological well-being during political crises. *Modern Psychology*, 7(2), 45–58. <https://journals.yasu.am/modern-psychology/en/article/view/13367>
- Malaviya, S. (2020). Digitising the Virtual: Movement and Relations in Drone Warfare. *Millennium: Journal of International Studies*, 49(1), 80-104. <https://doi.org/10.1177/0305829820971694>
- Samuel, J., Khanna, T., Esguerra, J., Sundar, S., Pelaez, A., & Bhuyan, S. S. (2025). The rise of artificial intelligence phobia! Unveiling news-driven spread of AI fear sentiment using ML, NLP, and LLMs. *IEEE Access*, 13, 125944–125969. <https://doi.org/10.1109/ACCESS.2025.3588179>
- Selaković, M., Krkić, M., & Filipović, N. (2026). AI-generated fake news as an instrument of crisis induction: A new typology. In Z. Keković, S. Kentera, M. Dabetić, & S. Grbović (Eds.), *Critical infrastructure: Risk and crisis management in the new reality* (pp. 201–217). Springer. https://doi.org/10.1007/978-3-032-20948-1_12
- Simon, F. M., & Altay, S. (2025). *Don't panic (yet): Assessing the evidence and discourse around generative AI and elections*. Knight First Amendment Institute at Columbia University. Oxford University Research Archive. <https://ora.ox.ac.uk/objects/uuid:ea305c74-27f2-41cd-a6f5-010dd90e60eb>
- Stockwell, S., Janjeva, A., & McDonald, B. (2026). *Adding fuel to the fire: AI information threats and crisis events* (CETaS Research Report). The Alan Turing Institute. <https://philpapers.org/rec/STOAFT-3>
- Waditwar, P. (2025). Leading through the synthetic media era: Platform governance to curb AI-generated fake news, protect the public, and preserve trust. *Open Journal of Leadership*, 14(3), 403–418. <https://doi.org/10.4236/ojl.2025.143020>
- Wardle, C., & Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policymaking* (Council of Europe Report DGI(2017)09). Council of Europe. <https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c>
- Zaighum, D. Z., & Ali, R. F. (2025). From Strategy to Tactics: Conceptualizing Weaponization of Digital Media Platforms across Levels of Warfare. *Lahore Institute for Research and Analysis Journal*, 3, 39-53. <https://doi.org/10.51846/0wy6p458>