Voice-Cloning as Crisis Threat: AI Detection Limits, Defender Latency and Escalation Risk in the 2024 Marcos Audio Deepfake Crisis in the Philippines

Panthea Zagrusvon

A Dissertation submitted to

Auckland University of Technology

in partial fulfilment of the requirements for the degree of

Master of Communication Studies

Faculty of Design and Creative Technologies

School of Communication Studies

2025

Abstract

Deepfakes are increasingly used in reputational attacks, fraud and political disinformation, creating risks for public trust, crisis communication and institutional decision-making. This dissertation examines the April 2024 Philippines incident in which a fabricated audio recording, overlaid on South China Sea imagery, appeared to present President Ferdinand "Bongbong" Marcos Jr. directing the Armed Forces of the Philippines and related task groups to act against China. Philippine authorities treated the recording as a national-security concern. Because the original YouTube upload and associated DAPAT BALITA channel were later removed, the incident provides a bounded empirical case for examining how visible limits of AI-based anomaly detection and crisis response can be reconstructed when primary platform artefacts are no longer public. Using a single-case open-source intelligence design, the study integrates qualitative crisis reconstruction, digital trace analysis and bounded timing measures. It draws on lawfully obtainable online traces, including official and media reports, surviving platform traces and third-party YouTube analytics. These materials are used to reconstruct the earliest plausible public-availability window of the fabricated recording (T_0), measure defender latency (Δ) to first visible authoritative responses, and analyse the pre-crisis and trigger conditions that shaped escalation risk. The findings indicate that, under OSINT-only constraints, no AI-based anomaly-detection labels, warning banners, manipulated-media indicators or propagation-level analytical signals were publicly visible at the user-interface level. The Presidential Communications Office issued the first authoritative public advisory approximately three to four days after the reconstructed T_0 window. Major domestic news organisations amplified the denial and warning within hours, whereas platforms, cyber agencies and political or legislative actors became visible later through reported investigations, account deactivations, cross-platform takedowns and terrorism-related framing. The analysis identifies two layers of advance warning: a broader strategic environment shaped by alliance signalling, South China Sea legal and jurisdictional contestation, foreign-influence narratives and domestic tension; and a compressed 11-22 April period marked by high-level diplomacy, Balikatan-related announcements and intensified military exercise activity. In this escalation-prone context, a fabricated presidential voice apparently directing military action in disputed waters was both plausible and potentially consequential. The dissertation contributes a public-trace timing protocol for analysing deepfake crises where original content and platform records are unavailable, combining OSINT-based T_0 reconstruction with comparative defender-latency analysis. It advances interdisciplinary scholarship by integrating AI anomaly-detection research, audio-spoofing studies, crisis and issue management, platform governance, digital trace analysis and legal-policy perspectives. It treats timing, visibility and contextual escalation risk as observable indicators of how AI-supported detection systems and crisis-communication practices perform during a real-world national-security crisis. The study also offers practical implications for defenders and policymakers by showing why synthetic-voice incidents require earlier detection visibility, rapid authoritative intervention, evidence preservation and accountable governance of AI-based anomaly-detection outputs.

Table of Contents

List of Figures:

List of Tables

Attestation of Authorship

"I hereby declare that this submission is my own work and that, to the best of my knowledge and belief,

it contains no material previously published or written by another person

(except where explicitly defined in the acknowledgements),

nor used artificial intelligence tools or generative artificial intelligence tools

(Unless it is clearly stated, and referenced, along with the purpose of use),

nor material which to a substantial extent has been submitted

for the award of any other degree or diploma of a university or other institution of higher learning."

Signed by *Panthea Zagrusvon*

Acknowledgements

I would like to express my appreciation to Associate *Professor Sarah Barker* and *Dr.Christina Batistich Vogels,* my lovely supervisors, for their support and guidance throughout the development of this dissertation.

Their constructive feedback, thoughtful suggestions, and their rigorous guidance in refining the structure and presentation of this work have been greatly appreciated. I am particularly grateful for their unwavering commitment to academic excellence and their insistence on maintaining the highest scholarly standards. Their insightful comments and critical perspectives continually challenged me to strengthen my arguments, deepen my analysis, and improve the overall quality of this dissertation. It has been a privilege to learn from and work under their supervision.

I also wish to thank *Dr. Gudrun Frommherz* for their direct and candid feedback. While some of the comments were challenging, they were crucial in helping me grow and refine my approach to leadership. The experience has been an invaluable learning opportunity, and I appreciate the guidance that has pushed me to develop greater self-awareness and resilience.

With a heart full of gratitude, I wish to dedicate this moment to my beautiful family, whose unwavering love and encouragement have been the guiding light of this journey. To my extraordinary family, whose brilliance, wisdom, and insight have been an endless source of inspiration, I am profoundly thankful. You have not only offered your intellectual insights, but also an emotional strength that has carried me through every challenge along the way. Your trust in me, even in the most difficult moments, has given me the courage to push forward, knowing that your wisdom and love are always with me. From the first spark of my academic ambitions to the completion of this study, your presence and support have been the constant fuel behind my perseverance.

Thank you for sharing your brilliance and your unwavering emotional strength. Without you, this journey would not have been possible.

Chapter1: Introduction

1.1. Background/ Context

The accelerating use of AI-generated deepfakes is reshaping crisis and issue management by eroding trust and increasing the risk of time-sensitive misinterpretation. Scholars and surveys link these capabilities to reputational harm, large-scale fraud, and political or electoral disinformation that can erode public trust and complicate governance responses (Balogun et al., 2025; Khanjani et al., 2023; Patel et al., 2023). Even where regulatory approaches are developing, enforcement remains difficult as synthetic media becomes increasingly realistic and harder to differentiate through conventional means (Blancaflor et al., 2023).

This dissertation examines a single high-stakes case that allows these broader concerns to be tested against publicly observable crisis signals: the April 2024 Philippines crisis in which a fabricated audio clip was circulated to sound like President Ferdinand "*Bongbong*" *Marcos Jr*. directing the Armed Forces of the Philippines and related task groups to act against China (ANC 24/7, 2024b; NDTV, 2024; Maitem, 2024; Olander, 2024). The Presidential Communications Office publicly rejected the authenticity of the recording and warned that the audio had been manipulated. Its response involved coordination with the Department of Information and Communications Technology, the National Security Council, the National Cybersecurity Inter-Agency Committee, and relevant private-sector stakeholders, illustrating the cross-sector governance challenges posed by synthetic media and information-integrity threats (Presidential Communications Office, 2024; 2024a, 2024b; Locus, 2024a).

This incident unfolded within an already contested maritime theatre*shaped by the South China Sea dispute, maritime legal contestation, alliance signalling, and heightened diplomatic and security activity (FORUM Staff, 2024; Law of the People's Republic of China on the Exclusive Economic Zone and the Continental

*Maritime Theatre is a military term, which refers to a large ocean or sea area in which naval operations, and sometimes joint air or ground operations, may occur (Vego, 2007, p.105).

Shelf, 1998; Lema, 2024; Ministry of Foreign Affairs of the People's Republic of China, 2024; News Agencies, 2024; Permanent Court of Arbitration, 2016; U.S. Embassy Manila, 2024).

This case provides an empirically bounded context through which to examine the practical limitations of AI-based detection systems and the temporal dynamics of authoritative response under OSINT*-only conditions. It is situated within a broader crisis communication framework, where the speed and effectiveness of institutional responses can significantly influence public interpretation and information diffusion. Synthetic voice incidents generate acute uncertainty because audiences may be unable to immediately determine whether a message attributed to an authoritative figure is authentic, manipulated, or officially sanctioned. Within such environments, corrective communication extends beyond retrospective denial and becomes a central mechanism of crisis containment.

Delays in detection, classification, verification, or authoritative clarification prolong the presence of fabricated content within the information ecosystem, thereby increasing opportunities for misinformation to circulate and shape public perceptions. This temporal challenge intersects directly with both the capabilities and constraints of AI-based anomaly detection systems. Although such technologies are frequently promoted for their capacity to facilitate rapid identification of manipulated content, their practical deployment remains constrained by issues of accuracy, latency, transparency, and performance under real-world crisis conditions. Accordingly, the Marcos voice-cloning incident is not just examined as an isolated instance of manipulated media. Rather, it serves as an interdisciplinary case study at the intersection of synthetic voice technologies, crisis communication, platform governance, and AI-enabled detection mechanisms and national security risk. From a crisis communication perspective, the case raises critical questions regarding the capacity of authoritative actors to verify, contextualise, and correct fabricated claims before uncertainty becomes entrenched in public discourse and influences collective interpretation.

* OSINT refers to open-source intelligence: analysis based on lawfully accessible online traces, third-party analytics, official records, news material, and observable proxies, rather than internal platform logs or proprietary detection outputs. Observable proxies refer to publicly visible indicators used to infer timing or response patterns when direct internal data are unavailable.

From an audio-spoofing perspective, the Marcos case illustrates a comparatively underexamined form of synthetic-media risk, particularly when set against the greater research attention given to video and image deepfakes (Khanjani et al., 2023). Its significance lies in showing how fabricated authority-voice content may become consequential within a live national-security environment, where detection limitations, defender latency, platform visibility, and contextual escalation risk interact before authenticity is resolved.

From an anomaly-detection perspective, the case enables analysis of observable online traces related to detection, labelling, takedown and enforcement when platform-internal systems are inaccessible. Its national security relevance further underscores the importance of situating the incident within its broader pre-crisis context. The significance and potential effects of fabricated authority-voice content do not arise in isolation; rather, they are shaped by the political, diplomatic, and security conditions that structure how such claims are interpreted, disseminated, and acted upon within the information environment.

The Marcos case makes the relationship between detection visibility, defender-response latency, and contextual escalation risk empirically examinable through verifiable evidence. It therefore provides an opportunity to investigate the dynamic relationship between detection visibility, defender response latency, and contextual escalation risk following the emergence of a fabricated voice claim within a nationally sensitive information environment. In doing so, the case illuminates how technical detection capabilities and institutional response mechanisms jointly shape the management of uncertainty during synthetic-media crises.

The contribution of this dissertation lies in using the Marcos case to connect audio-spoofing and anomaly-detection research with crisis communication, OSINT, platform governance and national-security analysis, thereby examining how technical limitations, defender latency and contextual escalation risk interact in a real-world synthetic-voice crisis. Research on AI-based anomaly detection provides insight into technical limitations including latency, opacity, false positives, and sensitivity to contextual variation. Scholarship on audio spoofing and synthetic media demonstrates how AI-generated or impersonated voices create distinctive challenges for identity verification, trust, and perceptions of authority.

The fields of crisis communication and issue management further illuminate the importance of timely intervention, authoritative correction, and the identification of pre-crisis indicators in situations where uncertainty has the potential to escalate into public, institutional, or geopolitical harm. Methodologically, the study draws on OSINT, digital trace analysis, platform analytics, and web-based reconstruction to establish an evidentiary framework for investigating such incidents when access to internal platform records is unavailable.

At the contextual level, perspectives from international relations, security studies, and legal-policy analysis situate the fabricated audio within the broader dynamics of the South China Sea dispute, alliance signalling, maritime legal contestation, and the national-security environment that shaped its potential for escalation. The policy and governance dimension links the empirical findings to recommendations concerning synthetic-voice risk controls, platform accountability, evidence preservation, crisis-readiness reporting, and the governance of uncertain AI outputs. These perspectives establish the Marcos incident as a high-consequence case through which the technical, communicative, institutional and security dimensions of synthetic-voice crises can be examined within a single evidentiary setting.

1.1.2. Content of the fabricated clip

News and verification outlets described the clip as a slideshow video in which a synthetic voice impersonating President Ferdinand “Bongbong” Marcos Jr. was mixed over still imagery of Chinese vessels and imagery of Philippines Coast Guard resources in the South China Sea and related maritime scenes (AIAAIC, 2024; ANC 24/7, 2024b; Maitem, 2024; NDTV, 2024; Rappler, 2024b; Rappler, 2024f; VERAfiles, 2024). Several reports converge on the point that the fabricated Marcos audio referred to Chinese actions harming Filipinos and presented the situation as demanding a tougher response; one outlet, for example, noted that he was “heard saying he can no longer allow Filipinos to get hurt by China’s actions” (Maitem, 2024).

Other outlets recorded or paraphrased a call for the “armed forces and special task groups” to respond if attacked, presenting the clip as if Marcos was ordering or preparing military retaliation (FORUM Staff, 2024; Olander, 2024; Philippine Star, 2024a; Rappler, 2024b; Rappler, 2024f).

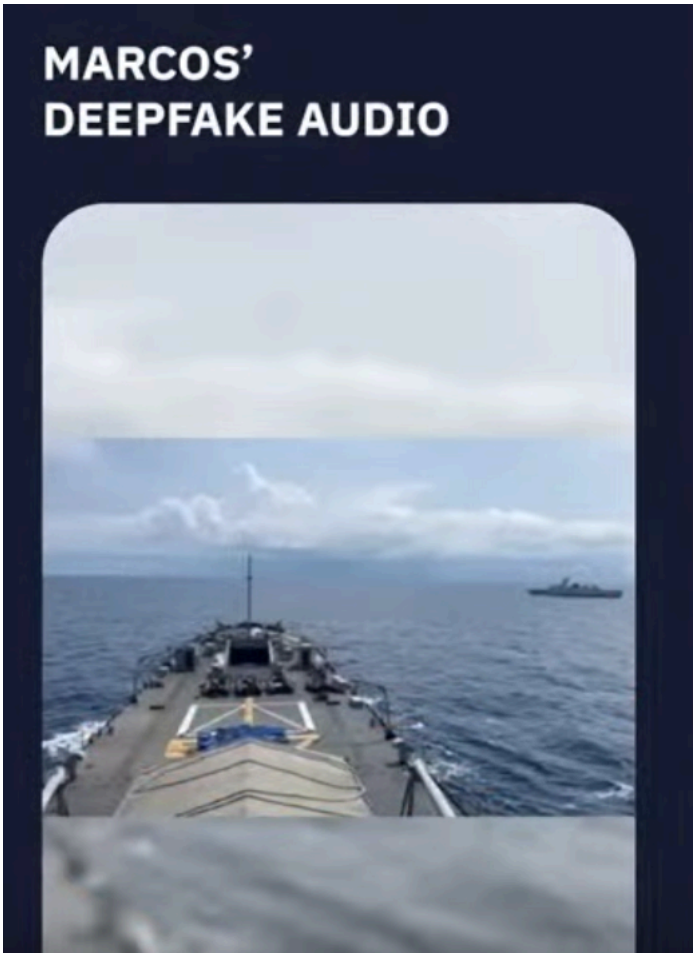


Figure 1.1. Rappler news screenshot reporting on the fabricated Marcos audio, including South China Sea imagery and a military vessel associated with the circulated video.

Source: Screenshot captured by author from Rappler (2024b), in August 2025.

The deepfake voice also linked the alleged response to the contested “gentlemen’s agreement” between former President Duterte and Chinese President Xi Jinping, framing it as a betrayal of national interests.

In a line preserved by VERA Files, the fabricated Marcos voice reportedly stated:

“I told our brave soldiers, retaliate against China if necessary. We must already stand up and use our forces to stop this bullying. As for China’s agreement with former president Duterte, which they call a gentleman’s agreement, I have assigned people who will officially conduct an investigation on what really happened. This situation must not be taken lightly. This is a clear form of treason against our motherland.” (VERAfiles, 2024).

Other segments highlighted in coverage reinforced a sovereignty & sacrifice framing. NDTV cited the line: “We cannot compromise even a single individual just to protect what rightfully belongs to us” (NDTV, 2024), while *VERA Files* recorded a closely related formulation: “We cannot further compromise the lives of our countrymen just to protect our rightfully own sea and maritime domains” (VERAfiles, 2024).

These reports document a fabricated presidential voice that invokes retaliation against China, frames an alleged Duterte-Xi understanding as “treason”, and ties potential sacrifice of *Filipino* lives to the defence of “our… sea and maritime domains” (VERAfiles, 2024; NDTV, 2024; FORUM Staff, 2024; Philippine Star, 2024a; Rappler, 2024b; Rappler, 2024f). On 23 April 2024 at 20:31 PHT, the *Presidential Communications Office (PCO)* issued the first public statement about the clip via its official Facebook page.

1.2. The Problem

Deepfakes are increasingly treated in scholarship and policy discussions as an evolving and scaling risk to public trust, governance, and security decision environments (Blancaflor et al., 2023), rather than a single isolated incident type mentioned by (Secretariat, 2024; Patel et al., 2023; Seitz-Wald, 2024). As synthetic media, deepfakes can be used to spread misinformation, manipulate public opinion, damage reputations, and facilitate coercive harms such as extortion and identity abuse, which makes these systems a cross-sector concern for public authorities, platforms, and crisis managers (Balogun et al., 2025; Khanjani et al., 2023; Patel et al., 2023).

Voice-cloning deepfakes intensify this risk because they can simulate the authority of a recognised speaker without requiring visual evidence. In political and national-security contexts, spoofed or synthetic speech can therefore circulate as if it came directly from a president, military leader, or other high-trust public actor, creating a crisis risk before technical verification or official correction can catch up (Khanjani et al., 2023; Tan et al., 2021).

Within crisis-management research, AI-driven anomaly detection on social media is frequently framed as a practical tool for monitoring public sentiment shifts and identifying high-stakes misinformation contexts including elections and terrorism-related threats (Bakkialakshmi & Sudalaimuthu, 2022; Guo et al., 2020; Khamparia et al., 2020; Mohamed et al., 2018). In technical terms, anomaly detection is commonly defined as the recognition of deviant or unexpected behaviour within data streams (Younas, 2020). However, social-media-oriented implementations that rely heavily on text or emotion classification can struggle in fast-moving and multimodal settings, where manipulated audio and cross-platform circulation complicate signal interpretation (Bakkialakshmi & Sudalaimuthu, 2022).

Recent studies of ML-based detection systems also highlight that misclassification and missed signals can compromise operational performance and public confidence, while high false-positive rates and context-blind flagging can distort what counts as a meaningful threat indicator (Alzaabi & Mehmood, 2024; Rahman et al., 2021; Visave, 2024).

Recent deepfake-risks surveys by Balogun et al. (2025) and Khanjani et al. (2023) also emphasise the persistent gaps in detection capabilities. These gaps remain a significant challenge even as the quality of synthetic media improves and its dissemination becomes increasingly widespread (Balogun et al., 2025; Khanjani et al., 2023). This gap is especially significant because audio deepfakes have received less scholarly attention than video deepfakes, even though voice impersonation can be politically disruptive and operationally difficult to verify in real time (Khanjani et al., 2023).

This limitation is particularly important for audio-based deepfakes because audio spoofing scholars have identified replay, text-to-speech, and voice conversion as practical routes for voice impersonation, while speaker-verification systems remain vulnerable to spoofing and presentation attacks (Gupta et al., 2024; Khanjani et al., 2023; Tan et al., 2021).

These technical limits matter because the harm of a leader-voice deepfake is not confined to whether the audio is eventually proven false. Disinformation events can erode public trust, create uncertainty around official messages, and weaken collective decision-making (Balogun et al., 2025), while audio-deepfake detection methods are progressing but still not keeping pace with the rapid growth of deepfake content (Khanjani et al., 2023). The problem is therefore also communicative: when detection, labelling or verification is delayed, the public information space may remain open to fabricated speech before authoritative correction appears.

Crisis-communication literature explains why timing is central to this problem. Reynolds and Seeger (2005) frame crisis communication around accuracy, credibility, timeliness and public correction, especially when uncertainty and public concern are high.

Kim and Lim (2023) further show that crisis misinformation can interfere with evidence-based messaging, and that delayed correction may allow false claims to become embedded in public understanding before authoritative correction appears. Lu and Jin (2024) extend this timing problem to social media environments, where crisis communication is no longer controlled by a single organisational voice; users, media actors and political commentators can all participate in transmitting and shaping crisis information.

In a leader-voice deepfake crisis, the timing of authoritative correction constitutes more than a procedural aspect of crisis communication; it serves as an indicator of the duration for which the public information environment remains exposed to fabricated audio before it is contextualised, verified, challenged, or formally refuted. During this interval, uncertainty may persist regarding the authenticity of the message, creating opportunities for manipulated content to influence public interpretation, media narratives, and political discourse

What remains underexplored is how anomaly detection limits appear in an OSINT-only public record during a politically sensitive, leader-voice crisis, and how "performance" can be operationalised when the original fabricated artefact and original platform records are missing or removed. In this dissertation OSINT refers to analysis based on lawfully obtainable online traces of artefacts, including surviving platform remnants and third-party analytics, official statements, news reports, and broadcast videos, rather than internal platform logs, moderation queues or model outputs for the now-removed *YouTube* channel.

This study addresses that gap by treating the timing of authoritative response as a public-facing indicator of a crisis-response performance, and by comparing this timing across defender classes in the 2024 Marcos voice-cloning case. Defender classes are defined as institutional actors possessing the authority, capability, or responsibility to respond to the incident. In this study, they comprise platforms and cybersecurity agencies, the executive centre, national media organisations, and political or legislative actors

1.3. Research Aims and Research Questions

This dissertation examines the April 2024 Marcos voice-cloning incident as a single high-stakes national-security crisis at the intersection of AI anomaly detection, audio spoofing, crisis communication and issue management. Existing research identifies important limitations in AI-based anomaly detection and deepfake detection, including latency, false positives, opacity, context sensitivity and difficulties in detecting realistic or unfamiliar synthetic audio. Crisis communication research also emphasises the importance of timely, credible, and corrective public response during fast-moving misinformation events.

The study therefore treats the incident not only as a technical detection problem, but as a staged crisis process involving pre-crisis monitoring conditions, public recognition, uncertainty reduction, and corrective communication. To examine this process, the study examines what the reconstructed OSINT record reveals about anomaly-detection limits, reconstructs when the fabricated audio first became publicly available, compares how different competent actors responded, and, as part of a crisis-management and issue-management approach, analyses how pre-crisis signals and contextual conditions shaped escalation risk.

For this study, T zero or T_0 refers to the earliest public availability of the fabricated clip that can be established from the OSINT record. Defender latency (Δ) refers to the interval between T_0 and the first visible authoritative response. On this basis, this study addresses three linked questions:

Q1. What does the OSINT record reveal about AI anomaly-detection limitations in practice during the Marcos audio deepfake incident?

Q2. What is the defender latency (Δ), from first public availability to first authoritative response, and how does this latency vary by defender class?

Q3. How did pre-crisis signals and contextual conditions shape the escalation risk of the Marcos voice-cloning incident, and what do they indicate about the need for rapid detection and authoritative response?

1.4. Scope and Limitations

This research focuses only on a retrospectively documented incident with available open-source material. It does not include platform-internal data, classified government responses, or proprietary detection logs. Consequently, the study may under-represent certain detection efforts or behind-the-scenes mitigations. This analysis draws on only publicly available indicators such as archived responses, third-party analytics, news reports, and official advisories which served as observable proxies. The original YouTube watch page and related platform metadata were no longer publicly accessible, which meant that the earliest public availability of the fabricated clip had to be reconstructed from surviving public traces rather than directly observed. Accordingly, T_0 or Tzero is treated as a bounded public-availability window rather than an exact verified upload timestamp.

This study analyses the Marcos deepfake as a national-security crisis in a sensitive context. However, the absence of the original audio and the inability to recover even a minimum sample of propagation posts intensify the platform-opacity problem, because the study cannot independently reconstruct how the fabricated clip circulated, who encountered it, or whether platform-side detection and enforcement shaped its visibility. The removal of key pages and missing archival records prevented recovering a 10 to 100 post sample to measure propagation dynamics. Therefore, it is impossible to show whether delay in defender response corresponded with rapid or widespread circulation of the fabricated audio clip across accounts. As a result, Δ can be assessed within the risk context, but it cannot be directly tied to public exposure or the impact of anomaly-detection systems.

The lack of recoverable propagation curves limits both the technical evaluation of anomaly detection and the ability to connect Δ to stakeholder experience. This study can assess defenders' response timing and its risk significance, but we cannot empirically demonstrate how many or which stakeholders encountered the fabricated message before official responses came.

The principal limitation concerns the missing circulation record, not the measurement of observable defender latency (Δ).

Language and timing presented additional methodological limitations. Some source material required translation from Filipino or Tagalog into English, while publication and analytics evidence appeared across different time zones. These limitations are noted here because they affect the interpretation of the OSINT record, while Chapter 3.7 explains how translation, time-zone normalisation, and timing verification were handled methodologically.

This study evaluated the responsiveness of the defenders to the presidential voice impersonation risk. It did not adjudicate the authenticity of the removed audio. Causal inferences about internal anomaly detection systems remain beyond the scope of this study.

1.5. Structure of The Dissertation

This study is organised into eight chapters that move from the case context and research problem to the literature base, methodological design, empirical reconstruction, analysis, conclusion and recommendations. The structure is designed to show how the Marcos audio deepfake incident is examined as both a technical detection problem and a crisis-management problem, with particular attention to OSINT evidence, defender latency and escalation risks.

Chapter 1 introduces the research context and the Marcos voice-cloning incident. It outlines the background to the case, including the reported content of the fabricated clip, before defining the research problem, aim and three research questions. The chapter also clarifies the scope and limitations of the study, including the OSINT only evidence boundary, and explains the structure of the dissertation.

Chapter 2 reviews the scholarship that frames the study. It examines literature on deepfakes, audio spoofing, voice cloning, AI-based anomaly detection, crisis communication, issue management and OSINT. This chapter establishes the technical, communicative and evidentiary foundations for the study by showing why synthetic or impersonated voice can create detection challenges, why authoritative response timing

matters in crisis settings, and why OSINT provides the basis for analysing public-facing traces when internal platform records are unavailable.

Chapter 3 sets out the methodology and operationalisation of the research. It explains the OSINT-based single-case design and defines the qualitatively dominant mixed-methods approach, drawing on Denscombe's discussion of mixed-methods research (Denscombe, 2014). The chapter shows how qualitative OSINT-based reconstruction, quantitative timing calculations and qualitative crisis interpretation are combined to analyse T_0, defender latency (Δ), defender classes, crisis phases and escalation risk. It also outlines the data collection process, timing reconstruction procedures, time-zone normalisation, coding strategy, methodological limitations and ethical position.

Chapter 4 provides the structural conditions of the case study and identifies the pre-crisis triggers that shaped escalation risk. It examines the constitutional command authority of the Philippine presidency, the legal background of the Philippines/China dispute, alliance and defence settings, joint military exercises, the deepfake regulatory environment, and international media amplification as pre-crisis signalling. The chapter then reconstructs the case across three phases: the trigger and initial public recognition of the Marcos audio deepfake, early domestic media amplification and framing, and the provenance, takedowns and cross-platform enforcement context. This chapter primarily supports the analysis of pre-crisis conditions and escalation risk in Q3, while also grounding the later examination of defender latency and visible detection limits.

Chapter 5 presents the empirical findings. It first identifies the escalation risks visible in the case, then examines defender behaviour across four defender classes. The chapter also presents the quantitative timing findings, including the reconstructed T_0 window and defender latency (Δ). The line chart and chronological timing table show how visible defender responses varied across actors, using verified timestamps, Philippine Time conversion and minimum and maximum latency estimates derived from the methodology described in Chapter 3.

Chapter 6 develops the analysis. It examines enforcement patterns relevant to T_0 reconstruction, including the *DAPAT BALITA* channel discontinuity and related public traces. It then uses semantic triangulation to assess alternative upload candidates and evaluate how the public record supports the reconstructed timing. The chapter analyses defender latency (Δ) as an observable crisis-communication indicator and considers how response timing varied across defender classes. It closes with a synthesis organised around the three research questions, drawing together what the OSINT record reveals about AI anomaly-detection limitations, defender latency and escalation risk.

Chapter 7 concludes the dissertation by answering the three research questions and synthesising the overall contribution of the study. It explains what the Marcos case shows about the practical limits of AI-based detection under OSINT conditions, the significance of defender latency as a public-facing crisis-response measure, and the role of pre-crisis signals and contextual conditions in shaping escalation risk. This chapter also reflects on the evidentiary limits of the study and the value of OSINT-based reconstruction for analysing platform opacity and crisis response.

Chapter 8 provides practical guidance for defenders, including platforms, crisis communicators, cyber agencies, organisations and regulators, on mitigating risks arising from the limitations, errors and uncertainty of AI-driven anomaly detection, particularly in synthetic-voice crises. The chapter is organised around synthetic-voice governance, audio-deepfake research and evidentiary capacity, anomaly-output and insider-threat safeguards, exposed-voice protection, policy support, and implementation capacity.

Chapter 2: Literature review

This literature review consolidates scholarship across deepfakes, audio spoofing, AI-based anomaly detection, crisis communication, issue management, OSINT, and platform governance to position the Marcos audio deepfake incident as a crisis at the intersection of technical detection limits and public-facing response. Rather than treating the case only as a technical problem, the chapter establishes why synthetic or impersonated voice can become a security-sensitive communication threat when attributed to a political leader, and why the visibility, timing, and adequacy of defender response are central to analysis.

Research on AI-based anomaly detection in social and information systems explains how machine-learning (ML) and deep-learning (DL) models are used to identify deviant behaviour, misinformation, malicious activity and abnormal patterns. However, this literature also identifies structural and operational limitations, including latency, false positives, recall-precision trade-offs, generalisation gaps, opacity, context-blind baselines, behavioural drift and biased or incomplete data, and difficulty interpreting context.

These debates inform Q1 by clarifying what detection systems are expected to do and why their limits matter in practice; they also inform Q2 and Q3 by showing how delayed, uncertain or non-visible detection can shift the burden of crisis response onto public-facing defenders. Alongside this, literature on voice biometrics*, spoofing and deepfake speech explains the threat surface of synthetic or impersonated audio, including replay attacks, text-to-speech synthesis, voice conversion, Logical Access and Physical Access attack routes, presentation attacks, speaker-verification vulnerabilities and countermeasure limits.

This body of work establishes why a voice-cloning incident involving the President of the Philippines creates not only a detection challenge, but also a crisis-risk problem involving authority, trust and security-sensitive interpretation.

*Biometrics are classified into physiological categories such as DNA, fingerprint and iris, and behavioural categories, such as typing rhythm, voice and motion (Tan et al., 2021, p. 32726). Voice biometrics refers to the use of vocal characteristics to identify or verify a speaker.

The review then connects these technical concerns to crisis communication and issue-management literature, which explains why timeliness, credibility, correction, public reassurance and pre-crisis monitoring matter when misinformation circulates under uncertainty. This strand supports the study's focus on defender latency, authoritative correction and escalation risk, particularly where fabricated audio may circulate before platform intervention, official clarification or public correction becomes visible. Finally, OSINT literature provides the evidentiary context for analysing removed or inaccessible platform content through online traces, official statements, news reports, screenshots, third-party analytics, and observable proxies.

2.1: Deepfake and Audio Spoofing attacks

"A spoofing attack refers to a malicious party launching an attack to impersonate an authorised individual in the voice recognition system to bypass and get access to the system." (Tan et al., 2021, p. 32726).

In fast-moving political or national-security crises, such spoofing allows attackers to inject convincing fake audio into critical decision and communication channels.

"Currently, deepfake attacks are known to be the most successful types of attacks to detect. Replay attacks are the easiest to mount, most difficult to detect, and do not even require the attacker to be technically knowledgeable" (Gupta et al., 2024, p. 4)

Balogun et al. (2025) explain that advances in generative AI have enabled the creation of highly realistic synthetic audio, video, and text commonly known as deepfakes. Deepfake technology leverages sophisticated generative adversarial networks (GANs), a type of a deep learning system, to produce synthetic media capable of bypassing conventional detection measures (Balogun et al., 2025). Blancaflor et al. (2023) state that deepfakes can be considered a form of cyber stalking and cyber terrorism-motivated violence against a particular person through technology. (Blancaflor et al., 2023, p. 392).

Balogun et al. (2025) also examine social, political and informational dimensions including accuracy, sources of information, dissemination channels and public perception of information.

Their findings report a significant decline in public trust following disinformation events, with average trust level dropping from almost 70 to 59 and continuing a downward trend of 0.25 points.

The study's visual analysis confirms this deterioration, highlighting a sharp post-event fall in public confidence described as “Truth Decay” effect that erodes societal cohesion and democratic integrity. The journal describes how disinformation-driven confusion and scepticism fracture societal cohesion, weaken collective decision-making, and undermine democratic institutions, resulting in the erosion of democratic integrity. Additionally, this research argues that the employed AI-based algorithms not only help disseminate disinformation but also amplify this issue by precisely targeting specific demographics and tailoring disinformation in ways that exploit biases and preferences, while detection tools struggle to keep pace (Balogun et al., 2025).

These patterns of “Truth Decay” are directly relevant to deepfake-driven political crises, where fabricated audio or video can rapidly undermine confidence in official messages. Patel et al. (2023) explain that deepfakes are typically created using methods such as deep neural networks, convolutional neural networks (CNNs), autoencoders, and generative adversarial networks (GANs). CNNs are a type of neural network often used in image and video processing. They work by automatically identifying patterns such as shapes, edges, and facial features.

GANs involve two competing AI models: a generator, which creates fake content, and a discriminator, which evaluates that content and tries to tell whether it is real or fake. Through this back-and-forth process, the generator becomes better at producing highly realistic synthetic images, audio, or video (Patel et al., 2023). Khanjani et al. (2023) report that over the past ten years, deepfakes have evolved from technological novelties to integral elements of a larger media ecosystem, impacting politics, corporate fraud, social media, and entertainment. The survey identifies itself as the first English-language survey to comprehensively examine both the generation and detection of audio deepfakes and notes that audio deepfakes have been largely neglected in prior studies, which predominantly focus on video-based deepfakes.

The authors highlight the use of manipulated political videos, altered images, and synthetic voices to influence public opinion, citing a case where a deepfake voice, impersonating a German executive, was used to authorise a large fraudulent financial transaction.

The same survey reports that the volume of research on deepfakes has extended significantly, with the number of related publications increasing from approximately 60 to over 300 in a single year, and over 1,300 papers published by 2020. Audits by Deep trace revealed that the number of deepfake videos nearly doubled from 2018 to 2019. Despite this rapid growth, the authors note that while detection methods are progressing, they are still not keeping up with the rapid rise of deepfake content. Survey data also show that most adults do not consistently fact-check what they read or share on social media, raising significant concerns about the potential impact on future elections as AI-generated disinformation becomes increasingly sophisticated (Khanjani et al., 2023).

Considering these factors, the Marcos incident sits at the sharp end of a rapidly expanding but still detection-limited ecosystem, where audio deepfakes remain underexplored and operationally difficult to contain. In this broader context, Khanjani et al. (2023) define audio deepfakes as AI-generated or edited speech that mimics natural sound, distinguishing between three primary types of manipulated audio: replay attacks, text-to-speech (TTS) synthesis, and voice conversion, including impersonation. Replay attacks involve reusing recorded speech from a target speaker, either through far-field recordings or cut-and-paste methods and are already recognised as low-cost threats to systems like automatic speaker verification and home voice assistants.

These attacks, which can be easily carried out with basic phones or recording devices, have resulted in numerous cases of exploitation. Speech synthesis, a core mechanism used in audio-deepfake generation, has been implemented in commercial systems such as Lyrebird-Descript, which uses deep-learning models to generate speech rapidly, replicate voices from recorded samples, and operate across languages. While these technologies have legitimate uses in areas like radio broadcasting and news reporting, they also have the potential for malicious applications, such as creating fake identities or inciting political and societal disruption.

The third subtype, voice conversion, involves mapping one speaker's voice onto another's identity, with applications in speech rehabilitation, personalised assistive technologies, speech-to-speech translation, language learning, and entertainment. However, it also enables impersonation, whether for fraud or entertainment, with tools like Overdub capable of mimicking "any voice" from just one minute of sample audio (Khanjani et al., 2023).

These three modes of manipulation- replay, synthetic speech, and voice conversion- describe exactly the families of techniques that can be used to fabricate a leader's voice in a crisis setting such as the Marcos case.

Tan et al. (2021) explain that voice recognition, commonly known as speaker recognition, refers to recognising the speaking person, whereas speech recognition refers to recognising the words from speech. Speaker verification is a process where the claimed identity of the owner of the voice, the target voice, is verified by comparing the target voice with the registered voice in the database (Tan et al., 2021).

In a crisis such as the Marcos deepfake incident, these mechanisms are precisely what defenders and platforms implicitly rely on when judging whether a voice clip is authentic. However, the technical literature on spoofing shows that these very verification mechanisms are systematically vulnerable to targeted attacks.

Gupta et al. (2024) explains that there are two main types of attacks: direct and indirect. Indirect attacks are those occurring at system-levels, being feasible whenever the attacker has access to the internal subsystems as a white-box attacker, or when the attacker has partial grey-box knowledge. Indirect vectors are acknowledged as system-level threats and worst-case evaluation settings rather than primary fake audio production paths. Here, the focus is on the direct spoofing routes, such as Logical Access (LA) and Physical Access (PA) (Gupta et al., 2024).

Tan et al. (2021) shows that speech synthesis or text-to-speech (TTS), voice conversion (VC) and replay were the three main spoofing attack types where LA and PA were used as evaluation scenarios (Tan et al., 2021). Gupta et al. (2024) classify LA attacks as digitally delivered attacks including TTS and voice conversion (VC) (Gupta et al., 2024).

Tan et al. (2021) further explain that VC involves converting the voice of a spoof attacker into the voice of the target speaker to deceive ASV systems (Tan et al., 2021). In this taxonomy, deepfake speech is placed under the LA category and is recognised as one of the most effective attack techniques today. It leverages TTS/VC methods and is often detected using tools like Whisper.

Gupta et al. (2024) further explain that Physical Access (PA) attacks involve injecting signals into the system via the real-world acoustic path. The most common example is the replay attack, where a pre-recorded authentic voice is played back into the system using a speaker (Gupta et al., 2024).

Li et al. (2025) and Shaaban et al. (2023) extend this concern by showing that recent advances in text-to-speech (TTS) and voice-conversion (VC) systems have made high-quality, realistic human-like audio easier to generate, while detection systems have not kept pace (Li et al., 2025; Shaaban et al.,2023). These attacks are easy to execute, require no technical expertise, and are difficult to detect (Gupta et al., 2024; Tan et al., 2021).

Tan et al. (2021) highlight the risk of sensor level* presentation attacks, which occur at the microphone and are much harder to defend against than transmission-level attacks. These attacks encompass impersonation, replay, voice conversion (VC), and speech synthesis (SS/TTS). Rather than targeting only internal system components, these presentation attacks directly target the “front door” of voice systems, bypassing voice recognition systems through spoofed voice input. Sensor-level threats are particularly concerning in high-stakes situations, such as crises or politically sensitive environments, as they bypass traditional network defences and can easily exploit publicly accessible audio from social media platforms like Facebook, Instagram and WhatsApp. Given that biometric data like voice is readily available, robust countermeasures are essential to strengthen the security of these systems, which are increasingly vulnerable to spoofing attempts.

*Sensors refer to microphone, camera or any other sensors that collect biometric information about a person and the presentation attacks in cybersecurity refers to fake biometric input.

Shaaban et al. (2023) similarly describe fake and synthetic audio as a growing challenge to vocal biometrics, noting that improved audio-synthesis methods, VC and TTS, are extremely dangerous for having the ability to generate audio that resembles target speech and create risks for vocal biometric devices.

Tan et al. (2021) warned that voice recognition systems are exposing people to spoofing attempts and remain a security risk and even the best systems have failed us. While advanced machine-learning-based automatic speaker verification ASV systems can verify speakers with low error rates and high accuracy but remain vulnerable to presentation attacks because biometric data can be obtained easily.

Among presentation attacks, impersonation involves an attacker attempting to replicate a target speaker's voice through vocal mimicry alone, without employing any synthetic or assistive technology. Although generally less effective against modern ASV systems, such attacks can occasionally succeed, as demonstrated in a notable case where a BBC reporter gained access to his twin's HSBC account using voice recognition (Tan et al., 2021).

2.2: Anomaly Detection Limitations:

In the context of crisis management, the utilisation of AI-driven anomaly detection systems on social media platforms has emerged as an essential instrument for monitoring and mitigating of public sentiment shifts (Khamparia et al., 2020) and misinformation during high-stakes events such as elections, political crises or terrorism-related activity on social media (Bakkialakshmi & Sudalaimuthu, 2022; Guo et al., 2020; Mohamed et al., 2018). In this study, the focus is on how such systems perform under fast-moving political crises and voice-cloned disinformation incidents, where timing and accuracy are critical.

Younas (2020) defines anomaly detection as the process of identifying deviance or unexpected behaviour in a certain dataset. According to Bakkialakshmi and Sudalaimuthu (2022), in social-media practice, anomaly detection has often relied on text and emotion classification (Bakkialakshmi & Sudalaimuthu, 2022). However, these approaches break down in fast-moving, multimodal crisis situations.

The current AI anomaly detection models include machine learning (ML) models, deep learning (DL) models, and hybrid models that combine ML and DL techniques. Several scholars in computer science and cybersecurity have highlighted that these models struggle to operate properly, addressing issues such as accuracy, latency, the overwhelming presence of false positives that compromise analysis, opacity in decision-making processes, and the failure to explain why something is considered a threat in addition to ethical and privacy concerns (Alzaabi et al., 2024; Guo et al., 2020; Rahman et al., 2021, p.10; Patel et al., 2023, p. 143298; Jaiswal et al., 2024; Visave, 2024). Guo et al. (2020), state that, traditional ML-based txt detection no longer accommodates multi model posts such as image, video and audio, delaying the identification of real crucial anomalies, disinformation, and public-attitude shifts. This study explains that in social media, verification and labelling are time-consuming and labour-intensive processes, and it is easier to skip the authoritative benchmark and prioritise recall over precision. This is important because higher precision means fewer false positives or false flags.

Additionally, their study clarifies that in political content monitoring, maximising the detection of false statements takes precedence over avoiding the flagging of genuine posts (Guo et al., 2020). Rahman et al (2021) provides evidence that delay, and processing time are one of the crucial limitations of the current anomaly detection models.

First, data preparation and filtering take significant processing time which contributes to end-to-end latency meaning the total time from data input to model output even before the model starts scoring. Secondly, using the current Support Vector Machine (SVM)*, at the processing stage, for large datasets takes too long and SVM becomes a computational bottleneck as data volume increases (Rahman et al., 2021). Taken together, these factors hinder real-time detection and responsiveness during crisis where minutes of delay can allow fabricated content to spread widely before any intervention.

Another common challenge in the social anomaly-detecting systems is high false positive or false-alarm rate, when the monitoring system incorrectly flags legitimate users or content as anomalous especially under high -dimensional feature spaces and with the biased-training data (Rahman et al., 2021; Alzaabi & Mehmood, 2024). For instance, a substantial number of users with over 3,000 friends were wrongfully deleted (Rahman et al., 2021).

Rahman et al. (2021) underscores that converting continuous features into fixed, biased categories for the sake of model simplification often to reduce computational cost, results in the loss of vital details, particularly near decision boundaries where data points are less clearly classified. This loss of information affects the model's ability to distinguish between relevant and irrelevant data leading to an imbalance between recall and precision. As a result, the model flags too many irrelevant instances as anomalies which increases the number of false positives or false flags (Rahman et al., 2021), cluttering the alert space that crisis managers need to monitor.

*SVM is a common supervised learning algorithm for classification tasks such as detecting anomalies or categorising different data

When millions of individuals are flagged, diverting the anomaly detection system away from identifying genuine concerns (Alzaabi & Mehmood, 2024; Guo et al., 2020; Rahman et al., 2021), the consequences of over-flagging legitimate content directly erode public trust, a key component in effective crisis management and as a result, creating unnecessary fear in public that escalates the crisis.

These challenges are especially significant in the context of social media due to the massive and diverse nature of online content.

Beyond social media platforms, similar limitations appear in organisational anomaly detection systems, especially in malicious insider-threat detection, a concern raised by several scholars (Alzaabi & Mehmood, 2024; Pandey et al., 2025).

Pandey et al. (2025) clarify that anomaly-based systems are designed to detect deviations from normal traffic patterns, enabling them to identify new, previously unseen attacks. However, they often generate a high number of false positives, overwhelming security analysts and reducing the system's effectiveness. They highlight the systems' difficulty distinguishing between normal and abnormal behavior as well as rule-based security controls failing to detect credential misuse by legitimate users. Additionally, behavioural drift, where user behavior evolves over time, causes models trained on past data to miss new threats. A further complication was stated as the static nature of most datasets, which makes it difficult to update models to account for these changing behaviours (Pandey et al., 2025).

In crisis management terms, these ML-based insider-threat detectors function as mitigation measures, designed to prevent emerging risks from turning into possible crises. However, when they misclassify benign behaviour, they expose the limits of prevention strategies that rely heavily on automated anomaly detection to distinguish genuinely consequential threats from ordinary variation.

Drawing on Alzaabi and Mehmood (2024), each person's behaviour is expected to fit within fixed biased categories, and the anomaly detection system has difficulty distinguishing routine tasks from suspicious activities. Legitimate behaviour changes are flagged when based on line or context are insufficient, wasting effort. Ordinary conversations over encrypted channels are misidentified as malicious, when context is

lacking, producing false alarms in the system (Alzaabi & Mehmood, 2024). As Alzaabi and Mehmood (2024) explain, the system does not adequately consider time, location, device or role of the individuals when evaluating actions, which leads to false positives or false flags, when contextually normal variations appear anomalous.

The study reveals that legitimate employees, students, contractors and other authorised individuals who are already part of the organisation and legitimately have access to files, computers, networks and so on, while conducting their regular work, are flagged as anomalies just because the system relies on expected patterns.

Any deviation such as working on different project or missing the task or delaying completion of a task or even using a different device, (for example when an employee logs in to her account from her home), are misclassified as suspicious activity, triggering organisational cybersecurity alerts without understanding the context or circumstances (Alzaabi & Mehmood, 2024).

Writing from a US police department emergency-management context, Visave (2024) raises ethical concerns around using AI-automated systems in the workplace particularly highlighting the unethical displacements of workers or the avoidance of hiring specific genders, such as women (Visave, 2024). Therefore, evidence that these anomaly detection systems in organisational settings routinely suffer from class imbalance, context-blind flagging, and high false-positive rates points to deeper structural limits in how behaviour is modelled, contextualised, and surfaced as “anomalous” in current systems. This illustrates how rigid behavioural baselines can systematically misread normal variation as threat, a pattern that also appears in social-media anomaly detection. These limitations matter in deepfake-driven political crises because context-blind or delayed anomaly detection can leave fabricated audio or video circulating long enough to undermine confidence in official messages.

Many scholars and practitioners have highlighted the lack of transparency in the deep learning models, the "black-box" nature which has created difficulties for operational transparency and trust among the public. (Alzaabi & Mehmood, 2024; Guo et al., 2020; Visave, 2024). For crisis managers and platform operators, this “Black box "nature makes it difficult to understand why a deepfake was or was not flagged, and to correct detection failures in real time.These technical vulnerabilities translate into concrete operational risks.

According to Fearn-Banks & Kawamoto (2024), after the COVID-19 pandemic, when online meetings were made as a frequent practice, the existence of AI-based anomalies, such as intruders or uninvited characters entering virtual meetings, became a significant concern. These unauthorised participants not only sometimes disrupt communication but also pose serious risks to privacy by silently recording the meetings, invading the company's privacy, and stealing personal and confidential information. As AI voice technologies have advanced to a point where hackers can easily duplicate the voices of legitimate participants, including authorities, and use this synthetic voice to issue convincing audio messages or instructions. Attackers can easily impersonate key personnel, potentially sabotaging the company by issuing fraudulent directives, misrepresenting decisions, or spreading misinformation, playing with their reputations etc, all while maintaining the appearance of authenticity (Fearn-Banks & Kawamoto, 2024).

The same operational risk extends beyond communication disruption to the long-term exposure of biometric identity data. Even highly secured institutions are not exempt from these risks. Congressional Research Service (2015) reports that the reality is that not even the most advanced secure offices, such as the highly secured federal government computers of the United States, were immune to such vulnerabilities and have been breached through various means, including social engineering and impersonation attacks, where over a million biometric data records of the personnel were stolen and the attackers were able to deceive personnel and exploit system weaknesses to gain unauthorised access, leading to questions about whether the existing provisions of law give agencies the legislative authority and resources they need to adequately address the risks of future intrusions especially considering that biometric data cannot be reissued and when they are hacked, there is no way to change them like passwords (Finklea et al., 2015). The limits of detection also point back to the pre-crisis conditions that make voice-cloning attacks possible in the first place. Gupta et al. (2024) identify voice privacy (VP) as a protective approach that hides a speaker's identity while retaining linguistic content and naturalness. By weakening the link between speech data and speaker identity, VP can make target selection more difficult for attackers and reduce the likelihood of successful attacks using published voice data (Gupta et al., 2024).

From a structural viewpoint, inadequate or missing voice protection means that identifiable voice data can remain accessible long before a crisis occurs. This increases the pre-crisis attack surface and heightens escalation risk when a voice-cloning incident targets a public figure, as in the Marcos case. However, this prevention-oriented approach also introduces broader ethical questions about biometric voice data, consent, privacy protection and the responsibilities of platforms or institutions that store, circulate, or expose identifiable speech.

Several studies have raised ethical concerns regarding the use of AI-driven systems in crisis management. The increasing need to preserve privacy in AI models that can identify anomalies without violating user privacy is also highlighted in these studies. The current ML algorithms cannot ensure privacy of users' data used for training the model since it relies on platform-level profile or content features , while privacy requirements or API* constraints restrict feature availability or necessitate costly privacy-preserving adaptations (Guo et al., 2020; Mohamed et al., 2018; Alzaabi & Mehmood, 2024; Visave, 2024; Rahman et al., 2021).

At the same time, the reliability of these systems is further compromised by their high false- positive rates or false alarm, which overflow crisis managers with irrelevant or misleading alerts, since it does not even show the reason why it was flagged (Guo et al., 2020). Beyond technical performance, several studies highlight serious fairness and discrimination risks. Visave (2024) highlights that not only the algorithm is biased and discriminate against communities, it also wrongfully predicts the risk of reoccurring the offences without any explanations. Sensitive personal data such as NHI numbers or locations of the survivors collected by AI during an emergency and compromising the privacy of over 147 million people, highlights the importance of establishing a strict privacy and data protection protocol.

* API stands for Application Programming Interface. Here, it refers to the technical interface through which authorised software can request structured data from a platform, subject to the platform’s access rules, permissions, quotas and rate limits

It is imperative that developers, operators, and organisations that deploy AI systems in emergencies are held accountable for potential biases in data or technical mistakes, which could result in unequal or incorrect actions, potentially causing severe consequences for individuals or communities (Visave, 2024). These concerns related to privacy, bias, and accountability are closely linked with fundamental operational challenges.

Lack of transparency is especially problematic in dynamic and rapidly evolving situations, where it is difficult for crisis managers to comprehend and defend AI-generated judgements (Alzaabi & Mehmood, 2024; Guo et al., 2020; Visave, 2024).

Current systems remain limited in real-time detection and fine-tuning* algorithms to reduce false positives has been suggested as one possible response (Rahman et al., 2021).

While understanding the limitations and failures in the current AI-based anomaly detection models is crucial, it is equally important to investigate the technical mechanism of voice spoofing in order to understand how fake or unauthorised audio can get past the detection systems through different methods and different consequences, particularly given real-world challenges such as latency, generalisation gaps and the rapid emergence of novel spoofing methods. The following studies examine how spoofing attacks succeed easily and how detection performance degrades, explaining why apparent model accuracy often fails to translate into timely, public warnings.

These general limitations become especially visible in audio deepfake detection, where spoofing systems must distinguish between genuine and manipulated speech under changing attack types, recording conditions and platform environments. This taxonomy of LA/PA and presentation attacks underscores that a *Marcos*-style leader-voice deepfake is not an exotic outlier but a textbook example of a Logical Access spoofing threat that current systems are known to find difficult. Deepfake detection systems, including deep learning models (DL), require substantial amounts of both real and fabricated data to learn how to identify manipulated content.

*Fine-tuning changes or adapts the model using additional data. It can reduce false positives by teaching the model what normal behaviour looks like in a specific setting, but it requires representative data, validation and ongoing monitoring.

However, deepfake techniques are advancing quickly, making it harder for detection tools to keep up (Patel et al., 2023). In relation to audio spoofing detection, Patel et al. (2023) point out that CNN-based models which are part of deep learning systems, are used to analyse spectrograms and classify audio as fake or real. The system's reliance on spectrograms which can be easily manipulated by the ongoing evolution of deepfake technologies presents a significant failure in the CNN-based system (Patel et al., 2023).

Tan et al. (2021) explain that speech synthesis attacks (SS) use text-to-speech (TTS) technology and that several synthetic speech detectors have been introduced to protect speaker verification systems from these attacks. Since many synthetic speeches are generated through parametric vocoders, phase-based synthetic speech detectors can be effective; however, their performance is limited when attacks use mixed-phase vocoders (Tan et al., 2021). For replay attacks, Gupta et al. (2024) show that reverberation, transmission-channel conditions, and acoustic-environment information are useful for replay-attack detection. They also identify pop noise as a voice-liveness cue: because pop noise is generated when a live speaker is close to the microphone, it may be faintly captured or absent in replayed speech (Gupta et al., 2024).

A further detection cue comes from Local Binary Patterns (LBP) features extracted from images generated from speech signals, such as spectrograms. These patterns capture subtle inconsistencies in the fine-grained spectral details of vocoded or voice-converted speech, revealing micro artifacts that differentiate fake from real signals. In addition, Constant Q Cepstral Coefficients (CQCC), identified as one of the best performing features widely used for detecting voice spoofing. Tan et al. (2021) also report their use in replay-attack detection systems (Tan et al, 2021).Tan et al., (2021) further show that detection methods rely on several types of cues to identify synthetic or manipulated speech and distinguish it from genuine speech. One important cue is phase information, which is used in synthetic speech detection because many synthetic speech attacks are generated through parametric vocoders* often exhibit unnatural phase distortions that are not present in genuine recordings (Tan et al., 2021).

*A vocoder is a speech-processing system used to analyse and synthesise speech signals, often forming part of text-to-speech and voice-conversion systems

However, Tan et al. (2021) also note that phase-based synthetic speech detectors are mainly effective against parametric vocoders that use minimum-phase filters, meaning that their performance can weaken when speech synthesis uses mixed-phase vocoders. For voice conversion attacks, the absence of natural speech phase information can also be extracted as a feature to distinguish converted speech from genuine speech (Tan et al., 2021).

From an anomaly detection perspective, Tan et al. (2021) show that replay attacks are especially important because they are straightforward to execute and expose the difficulty of building spoofing-detection systems that generalise across attack types and recording conditions. Their detection becomes particularly difficult under unseen conditions, especially channel-mismatch conditions.

A critical concern lies in the dependency of detection systems on specific attack types, as systems optimised to detect synthetic speech or voice conversion may not perform effectively against replay attacks, revealing a generalisation problem that must be addressed in anomaly-detection strategies (Tan et al., 2021). Gupta et al. (2024) add that transmission-channel conditions and over-the-air acoustic constraints can also affect spoofed-speech detection, particularly where perturbations or replayed signals must survive real acoustic transmission. Alongside spoofing attacks such as LA and PA, Gupta et al. (2024) identify adversarial attacks as another form of direct attack against ASV systems. These attacks involve small, often imperceptible perturbations that can cause machine-learning models to misclassify speech inputs, although their effectiveness may weaken under over-the-air transmission conditions (Gupta et al., 2024).

Li et al. (2025) extend this generalisation concern to current audio deepfake detection, showing that AI-synthesised audio detection techniques have not kept pace with rapidly advancing TTS models and often generalise poorly across datasets. Their results also show that many detectors can overfit to a specific dataset on which they are trained, while fine tuning on newer samples create catastrophic forgetting and reduce performance on other datasets. This creates a false sense of security when strong performance on one benchmark is treated as evidence of reliable detection across newer or unfamiliar tools.

Their benchmark results show that several detection models struggle when tested outside the datasets on which they were trained, and that no single model consistently outperforms across all datasets.

This suggests that the practical challenge is not simply detecting synthetic audio in controlled benchmark settings, but maintaining reliable performance when the audio is realistic, unfamiliar, and produced by newer generation systems (Li et al., 2025).

This gap between rapidly improving attack quality and slower, data set limited detection advances is a core structural limitation for crisis anomaly detection. Shaaban et al. (2023) reinforce this weakness that audio-deepfake detection remains operationally limited: although detection performance has improved from a scholarly perspective, existing algorithms still contain significant gaps and remain insufficient for real-world applications (Shaaban et al. 2023).

2.3: Crisis Communication and Defender responsiveness

Reynolds and Seeger's (2005) Crisis and Emergency Risk Communication (CERC) model provide a crisis communication foundation for situating fast-moving information crises within questions of accuracy, credibility, timeliness and public correction. Their work emphasises that emergent threats require communication that is accurate, credible, timely and reassuring, particularly when uncertainty and public concern are high. Crisis communication is therefore not limited to organisational image management; it also requires communicators to explain what has happened, identify likely consequences and provide harm-reducing information in an honest, prompt, accurate and complete manner. Their staged model further frames crisis communication as a process that begins before a crisis through monitoring and recognition of emerging risks, continues during the initial event through uncertainty reduction, and later involves correcting misunderstandings and rumours (Reynolds & Seeger, 2005).

This staged logic aligns with the dissertation's focus on pre-crisis conditions, public recognition and early response in the Marcos voice-cloning case. Kim and Lim (2023) position crisis misinformation as a communication threat because it can interfere with evidence-based crisis messaging and produce incorrect

beliefs among affected publics or internal stakeholders, with consequences for organisational reputation, trust and crisis-response effectiveness.

In crisis communication, misinformation is not limited to deliberately fabricated claims; it can also include rumours, inaccurate claims, insufficient information and outdated information circulating during an unfolding crisis. Correction therefore becomes a central crisis communication task.

Kim and Lim (2023) emphasise that effective debunking should not simply reject a false claim but should explain the actual situation and provide evidence to support the correction. They distinguish between simple rebuttal and factual elaboration: a simple rebuttal offers a brief denial, whereas factual elaboration reinforces accurate information through a more detailed and evidence supported. Timing is equally important because proactive correction can influence whether debunking succeeds, while delayed correction can allow misinformation to become embedded in public understanding before authoritative correction appears (Kim & Lim, 2023).

Measuring time-to-detect and time-to-respond (MTTD and MTTR) is a standard practice among security solution vendors, cybersecurity teams, and Security Operations Centres (SOCs), as shown in recent Detection and Response surveys that treat these timings as core performance indicators (Lemon, 2024). In deepfake driven crises, the timing and evidential substance of authoritative correction are therefore central to limiting the period in which fabricated audio can shape public interpretation before it is publicly corrected, contextualised or refuted.

In crisis management and public relations, proactive issue management is important because emerging information risks can escalate quickly if they are not identified, monitored and addressed before they develop into full crises. Gordon (2011), drawing on Regester and Larkin, argues that issue management, which focuses on anticipating crises, is a more proactive strategy than traditional crisis management, emphasising the need to identify and address potential crises before they fully develop.

Averill Gordon, referring to Reynolds & Seeger (2005), also explains that the pre-crisis stage involves research, monitoring, identifying and communicating about emerging issues while shaping public

perceptions to mitigate potential threats, whereas the post-crisis stage focuses on analysing the events that occurred and can lead to new practices such as legislative change or technical development .Understanding which issues need to be addressed, and when, is crucial in public relations, because the timing of responses depends on the nature and severity of the issue (Gordon, 2011).

Lu and Jin (2024) extend correction-focused crisis communication literature into the social media environment, where crisis misinformation can become not only a feature of an ongoing crisis but also a source of crisis escalation. They argue that social media has shifted crisis communication from a single organisational voice to a multivocal environment in which users participate in crisis information transmission and may shape the development of the crisis itself.

Their study identifies correction placement timing, refutation detail and narrative type as key factors in organisational crisis misinformation management, finding that pre-bunking, particularly when combined with factual elaboration, is more effective in correcting misperceptions of crisis responsibility and limiting reputational damage (Lu & Jin, 2024).

In the Marcos case, the multi-actor nature of the crisis environment renders the timing of authorised responses analytically important. Accordingly, defender latency is employed as a measure of the speed with which corrective communication emerged across relevant actors, as well as the duration for which the public information environment remained without authoritative contextualisation, correction, or refutation.

2.4. Open-Source Intelligence (OSINT)

Open-source intelligence (OSINT) is the legal and ethical gathering, processing and correlation of information from publicly available sources, conducted in ways that respect privacy and applicable regulations (Szymoniak & Foks, 2024). According to William & Blum (2018), OSINT refers to intelligence or analysis produced from publicly available information that is collected, processed, exploited, and disseminated to answer a specific information requirement (Williams & Blum, 2018). OSINT can also include unclassified information with limited public distribution or access, provided it can be used in an unclassified context without compromising national security or intelligence sources and methods (Jardines, 2002). OSINT should be distinguished from open-source information (OSIF). OSIF plays a critical role in the intelligence cycle. It refers to the legally obtained, publicly or commercially available raw or semi-processed material that feeds into the OSINT process, including print and broadcast items, web pages, social media posts, imagery and grey literature. OSINT, by contrast, is the analysed, decision-oriented product created after collection, processing, contextual evaluation and integration with other sources to address a specific information requirement (Williams & Blum, 2018).

The rise of the Internet, social media, and big-data analytics has made open-source work more complex in both sources and methods. In digital-platform contexts, OSINT may involve public webpages, news reports, social-media traces, platform-adjacent analytics, and commercial off-the-shelf tools. However, commercial tools are often designed for business purposes rather than intelligence or research needs, and their providers and capabilities change rapidly (Williams & Blum, 2018).

OSINT is employed across multiple communities. Within NATO's intelligence enterprise, the NATO OSINT Reader frames OSINT as relevant to commanders and their staffs, NATO commands, task forces, member nations, civil-military committees, working groups, and other organisations planning or engaged in combined joint operations ('t Hart, 2002).

In the United States, OSINT is used across the Intelligence Community and the wider defence enterprise for routine situational awareness, decision support and operational analysis. The Defense Open Source Council (DOSC) is the Department of Defense's primary body that organises and directs the use of OSINT by setting policy and standards while the Open Source Centre / Open Source Enterprise at CIA function as a central body for building capabilities and training those practitioners who then apply OSINT in their operational work (Williams & Blum, 2018).

The use of Open-source intelligence (OSINT) in Journalism and monitoring social media are a customary practice. In journalism and crisis response, newsrooms and journalists use OSINT for verification and fact-checking, while emergency responders, citizen reporters, and law enforcement units leverage social media monitoring to separate rumour from fact during fast-moving events such as during crisis and elections (Silverman, 2014). In security and law-enforcement contexts, OSINT can support investigations by connecting publicly available traces to persons, events or locations, as illustrated through an FBI investigation in which OSINT methods were used to identify a person connected to the burning of police cars during a protest (Szymoniak & Foks, 2024).

According to William and Blum (2018), OSINT is used across intelligence and defence settings, but its methods are also relevant to journalism, crisis monitoring, verification, and public-facing investigations where analysts must work from available digital traces rather than privileged internal records. Traditional and digital news formats such as TV/radio, newspapers, journals and online outlets are essential in providing verified timestamped and attributable reports of events. These sources are useful for tracking events chronology and serve as a reliable reference point for analysis, fact checking and intelligence work. Beyond headline media, grey literature forms an important category of OSINT material because it captures non-media institutional content, including conference papers, corporate documents, dissertations, government reports, newsletters, trade literature, trip reports, and outputs from research bodies, governments, corporations, think tanks, and academia.

William and Blum organise the OSINT process into collection, processing, exploitation, and production: acquiring open-source information, validating it, identifying its value, assessing whether the information is

what it appears to be, including authentication, credibility evaluation and contextualisation and turning it into a usable analytical product (Williams & Blum, 2018). OSINT relies on diverse source categories. Search engines and operators, reverse image search, maps including weather overlays for webcams or checking images, metadata extractors, network lookups (DNS, SSL, WHOIS, TLS), SocMINT for social profile correlation and automation tools are some of the tools used in OSINT methods (Szymoniak & Foks, 2024).

Williams and Blum argue that, because OSINT tools and providers change rapidly, analysis should focus on underlying methods such as lexical analysis, network analysis, geospatial analysis, aggregation, authentication, credibility evaluation, and contextualisation (Williams & Blum, 2018). In this dissertation, those principles inform the triangulation of public traces across source type, timing, and institutional position. Williams and Blum define grey literature as content from public and private non-media institutions, while also noting that its earlier access problem has changed as much of this material has moved online (Williams & Blum, 2018). Soule and Ryan add that grey literature may sit outside normal publication, distribution, bibliographic, or acquisition channels, but can provide information unavailable in published open sources, corroborate claims from other sources, and offer concise, focused, and detailed content (Soule & Ryan, 2002).

Automated AI-based OSINT has been increasingly used as a monitoring tool as part of the investigation process (Szymoniak & Foks, 2024), but this development raises accuracy and ethical concerns that should be addressed carefully.

In this research, OSINT provides the evidentiary basis for examining what was publicly visible during the Marcos voice-cloning crisis. Because the study cannot access internal platform systems, proprietary detection outputs, classified responses, or a complete propagation record, the analysis depends on public traces that can be attributed, time-checked, compared, and contextualised. Official statements, news reports, platform remnants, archived material, screenshots, and third-party analytics are therefore treated not as isolated proof, but as fragments of a wider public record. When triangulated, these materials allow the study to reconstruct T_0, measure defender latency (Δ), and assess visible crisis response while remaining within the limits of OSINT-based evidence.

Chapter 3: Methodology

3.1: Research Design and Operationalisation

This chapter explains how the dissertation moves from the conceptual problem identified in the literature to a measurable research design. The study is grounded in crisis-communication concerns with timeliness, accuracy, visibility, and the correction of harmful fabricated content, while also recognising that platform-based anomaly detection cannot be fully evaluated from outside proprietary systems. Methodologically, this dissertation uses an OSINT-based single-case design supported by digital trace analysis, platform analytics and web-based reconstruction. These computer-science-adjacent methods are used to examine how public platform remnants, third-party analytics, timestamps, archived online traces, official statements, news reports and broadcast material can be combined when internal platform logs, proprietary detection outputs and original platform artefacts are unavailable. This approach allows the study to reconstruct the earliest plausible public-availability window of the fabricated clip (T_0), calculate defender latency (Δ), and analyse visible detection, enforcement and response signals under conditions of platform opacity.

For that reason, the methodology focuses on what can be observed and verified through OSINT: the reconstruction of T_0 the measurement of defender latency shown as (Δ), and the interpretation of visible institutional responses during the Marcos voice-impersonation crisis. The study therefore sits at the intersection of technical detection limits, voice spoofing, crisis timing, and open-source intelligence evidence, platform analytics, and digital trace analysis.

Following the Denscombe framework of research (Denscombe, 2014), this research employs a mixed-methods, single-case design, focusing on the Philippines presidential audio incident.

The design is QUAL- dominant (Qualitative dominant), with quantitative measures supporting the analysis. It follows QUAL to QUAN to QUAL sequence, it begins with qualitative exploration to identify key patterns, followed by quantitative timing calculations, and concludes with qualitative interpretation to assess crisis-response implications.

Initial qualitative insights inform the construction of a timing model, which is then compared with alternative bounds Defender latency (Δ) and interpreted to understand institutional responses to leader-voice impersonation risks.

This mixed-methods strategy combines qualitative and quantitative approaches within a single research project, with each method used to address different dimensions of the same research problem, making an explicit link between them through triangulation, and adopts a pragmatist, problem-driven stance focused on producing practically useful findings.

Based on Denscombe's framework the single-case design in this study is supported when the case is clearly bounded and deliberately selected for attributes that are relevant to the research problem, rather than treated as a random sample (Denscombe, 2014).

Yin's case-study framework further supports the design of this study when a contemporary phenomenon needs to be examined in depth, within its real-world context, and through multiple sources of evidence that can be triangulated. Yin also distinguishes single-case and multiple-case designs, noting that multiple-case studies may offer broader comparative evidence, but that they do not always satisfy the rationale for a strategically selected single case (Yin, 2018).

Flyvbjerg similarly challenges the assumption that a single case cannot contribute to knowledge, arguing that the value of case-study research depends on the strategic relationship between the case, the research problem, and the knowledge being produced rather than on case number alone (Flyvbjerg, 2006). Applied to this dissertation, the value of a comparative design would depend on whether an additional case could strengthen the analysis of timing, visibility, and crisis context rather than simply increase the number of cases.

A multi-case design would require more than the identification of another voice-cloning incident; to be methodologically comparable, an additional case would need to involve a similar political or national-security context, preferably within the same national context or a closely comparable jurisdictional context, platform-policy environment, and within a comparable time period. This is especially important because OSINT-based reconstruction depends on the survival, accessibility, and comparability of public traces.

Older or less comparable cases are more likely to involve more deleted posts, unavailable watch pages, changed platform interfaces, expired archives, or policy environments that no longer match the case being analysed.

Considering the scope of this study, the Marcos incident offers the most appropriate single case because it is the most recent national-security voice-cloning crisis involving leader-voice impersonation, public-facing defender responses, crisis timing, and contextual conditions relevant to escalation risk. Given the restricted completion time frame, adding a weaker or less comparable case would risk reducing analytic precision rather than strengthening the study.

The study therefore adopts a single-case design to preserve depth, timing precision, and contextual consistency within one bounded crisis episode, prioritising analytic generalisation rather than statistical generalisation. The case is therefore bounded to the incident and its immediate pre-crisis and crisis windows, and the study aims for analytic generalisation rather than statistical population estimates.

3.2. OSINT methodological framework and data collection

This study relies on open-source intelligence (OSINT), using publicly accessible materials gathered and analysed to answer specific research questions. It does not have access to privileged platform metadata, internal platform logs or classified materials. Here OSINT functions as both a methodological framework and a constraint. Commercially Available Information (CAI), including purchased datasets, brokered social-media data, or audio files obtained from commercial sources, was excluded from the data collection strategy because it would not necessarily provide stronger or more verifiable evidence of the clip's first public availability or circulation history.

Without the original watch page, channel, or documented provenance, an audio file obtained from commercial third parties, could not be treated as the original artefact. Establishing its evidential value would require separate authenticity procedures, which sit beyond the scope of this study and is not the focus of this research questions. Brokered or platform-derived circulation data posed a separate methodological risk. it could create a misleading impression of completeness if its sampling method, collection window, exclusions, or processing steps were not transparent (Ramirez et al., 2014; Pfeffer et al., 2018).

For these reasons, the study relied on publicly accessible, reliable, and triangulated open-source intelligence evidence rather than paid or brokered datasets.

The OSINT process used in this study follows four steps: collection, processing, exploitation, and production. Collection involved identifying relevant public material; processing involved validating and organising it; exploitation involved assessing authenticity, credibility, timing, and context; and production involved turning verified material into the reconstructed T_0 window, defender-latency comparisons, and chronological event log.

Methodologically, this study distinguishes between OSIF as raw or semi-processed open-source material and OSINT as the analytical product generated through verification, contextualisation, and triangulation. On this basis, individual posts, screenshots, articles, platform remnants, archived responses, and analytics pages were not treated as self-standing proof. They were assessed according to source attribution, timestamp visibility, institutional authority, contextual relevance, and cross-source consistency before being used to support the reconstruction of T_0, the calculation of defender latency or (Δ), and the chronological mapping of the Marcos voice-cloning crisis.

The observed channel-level views drop was treated as a contextual timing indicator rather than direct proof of upload time, take-down time, or platform enforcement. It was interpreted alongside official advisories, news reports, surviving platform traces, and other timestamped public evidence.

This study's use of screenshots and retained copies is consistent with OSINT guidance on the volatility of online material. Jardines notes that Internet sites can disappear without notice, and that effective open-source exploitation may require archiving rather than just bookmarking web resources (Jardines, 2002). This was important in the Marcos case because the original watch page, channel traces, and propagation material were removed and later unavailable.

Sources are predominantly in English except for a limited number in Filipino, other Philippine languages, Chinese and Japanese which were translated with *Google Translate* which is AI-based.

In this study, the OSINT process is organised around four key steps: collection, processing, exploitation, and production.

Applied to the Marcos case, this cycle guided the collection of official statements, news reports, platform remnants, third-party analytics and other public traces; the verification and time-normalisation of those materials; and the production of a reconstructed T_0 window, defender-latency comparisons and a chronological event log.

The sampling strategy combines purposive selection of high-salience artefacts such as statements from the Presidency, the *Presidential Communications Office (PCO)* and the *Armed Forces of the Philippines (AFP)*, official communication channels, wire services, leading national and international news outlets and platform-adjacent analytics pages with theoretical sampling used to follow emergent leads, including specific outlets or officials not based on a fixed list of sources and snowballing across directly linked sources. This strategy is supported by *NATO* OSINT doctrine, where Allen warns that the volume of public information and the imprecision of retrieval create barriers to effective OSINT use (Allen, 2002).

The data gathering process was cumulative and exploratory, new materials were added until they no longer improved the estimate of the T_0 window or changed the Defender latency (Δ) calculation, in a meaningful way.

3.3. T_0 reconstruction procedure

Unlike studies that analyse deepfakes using pre-existing datasets or fully preserved platform archives, this research could not rely on a ready-made case file. When the project started, the original *YouTube* upload and the DAPAT BALITA channel were already unavailable and there was no existing research dataset or official reconstruction of the incident. As a result, assembling the case itself became part of the method: I had to identify the channel, recover residual references to the removed content, and use third-party channel analytics and surviving media traces to build an approximate publication window and response timeline before any analysis could begin.

Because the fabricated clip and its circulating versions were no longer publicly accessible at the time of investigation, reconstructing T_0 required an OSINT-based approach that could operate without the original watch-page URL. To determine the exact timing of publication, I would need access to the watch-page URL of the fabricated clip.

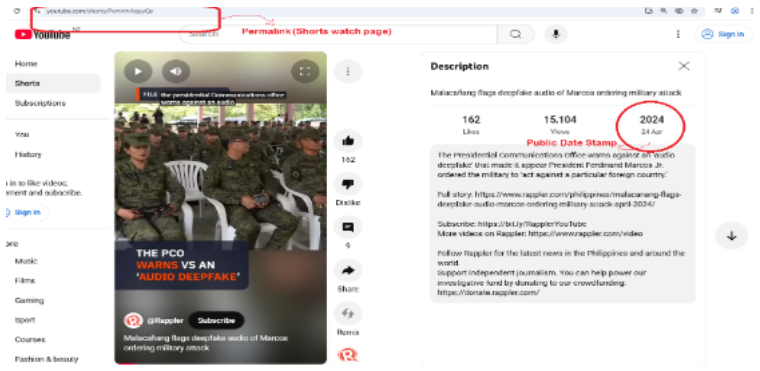


Figure 3.1. Illustrative screenshot showing the layout of a YouTube watch page.
Source: Screenshot captured by author from YouTube, Augest 2025.

What is a watch page and why is it important?

On YouTube, the watch page is the canonical single-video URL in the form https://www.youtube.com/watch?v=VIDEO_ID

It is the place where the platform displays the authoritative metadata for that specific video: the title, channel name, publish date/time (or "Premiered/Streamed live" time), the description, comments, and any policy labels (Google, 2025, May9).

This is the best anchor for establishing the T_0 or the first public appearance, because it combines the public timestamp with a stable permalink. An example of this is displayed here:

Note: It is different from a website URL which could be anything (homepage, category page, article) and may not be dedicated to a single video or contain authoritative video metadata.

As part of adaptive decision rule during early stage of the process, due to the lack of watch-page , internal logs data and API limits on verifying watch page history/engagement, which necessitate conservative OSINT and following pre-planned contingencies in a pragmatic mixed-methods design of Denscombe guideline, I estimated T_0 or the earliest publication date, as a time window from third-party channel analytics, then verified that window against the earliest dated official and news records. T_0 reconstruction constraints in OSINT and platform data hinder fact finding because based on Kohne et al., (2024) YouTube's API restricts access to private analytics and enforces quotas and rate limits, making key provenance and performance signals unavailable or delayed and limiting retrospective detection and verification. (Kohne et al., 2024, pp.7-8)

After conducting searches across *Facebook, YouTube, and X*, the primary social media platforms using multiple keyword combinations and cues derived from news coverage, I did not identify a single publicly accessible post containing the fabricated audio clip itself. No platform labels, fact-checker notices, or "video unavailable" markers were present at the time of investigation, a finding consistent with the reports cited below.

However, I was able to locate a third-party *YouTube* analytics service called *vidIQ* that, in this case, draws on channel data from *DAPAT BALITA* and generates temporal graphs and related metadata, enabling indirect observation of the clip's circulation patterns over time.

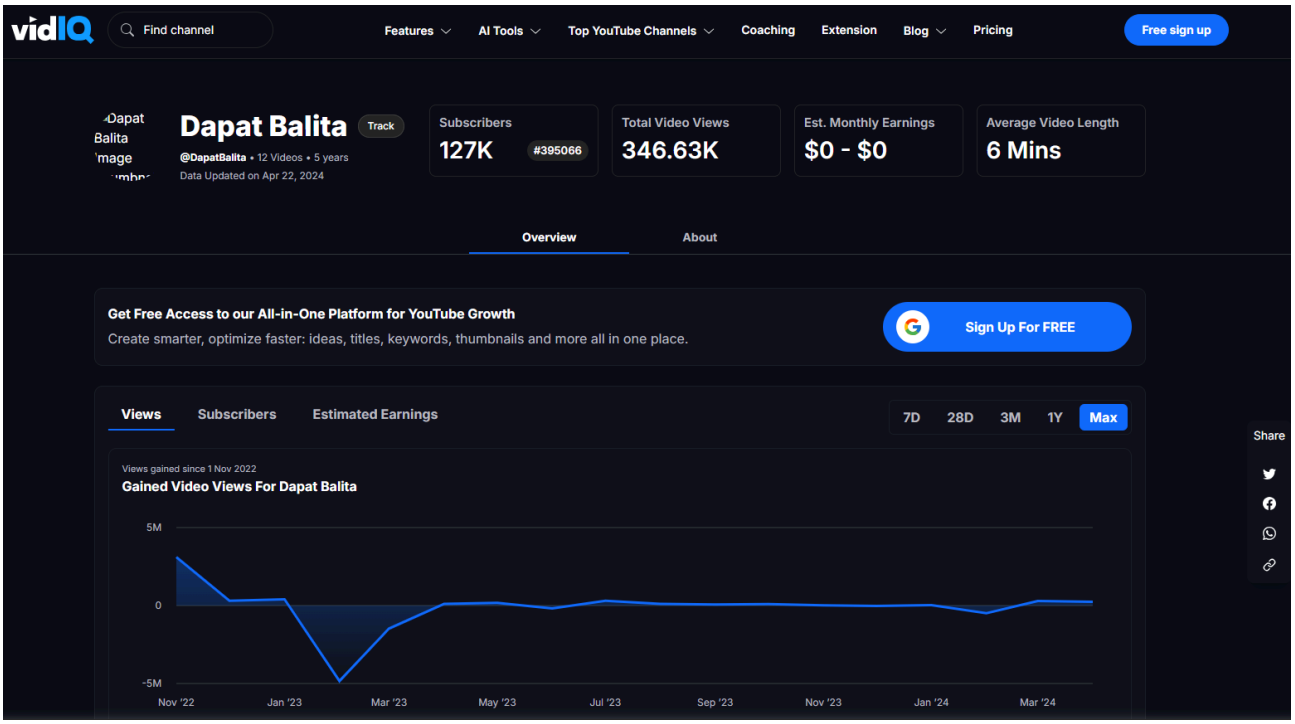
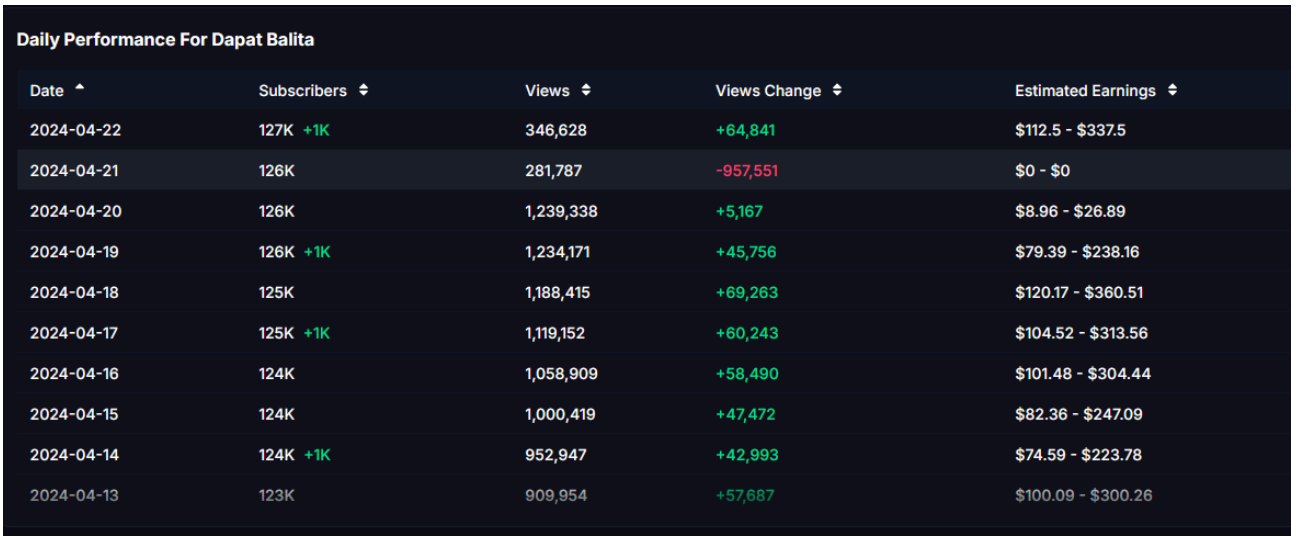


Figure 3.2. *Dapat Balita* channel analysis showing channel activity around the Marcos fabricated-audio incident period. Source: Screenshot captured by author from vidIQ, (05.09.2025). Data shown as updated on 22 April 2024.

This allows me to inspect the relative patterns of surges and drops in views, earning, subscribers and related channel metrics around the incident period, identify the channel's activation period and analyse its fluctuating data to pinpoint or estimate a window of possible publication window, but not the exact uploading time when the fabricated video could have been published.

Therefore, timing and interpretation rely on third-party channel-level time-series analytics of view counts and engagement collected by vidIQ. Without having the original watch-page URL, it is impossible to pinpoint the exact time of the publication and analysis from *vidIQ* is used as an observable proxy only. *vidIQ* is a third-party YouTube toolkit that analyses YouTube channel behaviour over time.

NOTE: These are observational signals, because they are external proxies, not internal authoritative logs from the platform itself. In other words, they allow us to observe patterns, but they cannot be treated as definitive evidence of exact timing or counts. They are valuable data because they still enable bounded and comparative measurement. For example, we can document that no labels were observed across a number of audited pages or that an official signal lagged behind the earliest credible appearance by X hours.



Daily Performance For Dapat Balita

Date	Subscribers	Views	Views Change	Estimated Earnings
2024-04-22	127K +1K	346,628	+64,841	$112.5 - $337.5
2024-04-21	126K	281,787	-957,551	$0 - $0
2024-04-20	126K	1,239,338	+5,167	$8.96 - $26.89
2024-04-19	126K +1K	1,234,171	+45,756	$79.39 - $238.16
2024-04-18	125K	1,188,415	+69,263	$120.17 - $360.51
2024-04-17	125K +1K	1,119,152	+60,243	$104.52 - $313.56
2024-04-16	124K	1,058,909	+58,490	$101.48 - $304.44
2024-04-15	124K	1,000,419	+47,472	$82.36 - $247.09
2024-04-14	124K +1K	952,947	+42,993	$74.59 - $223.78
2024-04-13	123K	909,954	+57,687	$100.09 - $300.26

Figure 3.3. Daily performance data for the Dapat Balita channel on 22 April 2024. Source: Screenshot captured by author from vidIQ, (25.09.2025).

After analysing the data retrieved from *vidIQ,* I can confirm that the *DAPAT BALITA* channel ID is (UClLHnDxskpxaR1ydIRdOJOA), retrieved on 5th of August 2025.

The channel ID does not change even if the owner changes the name of the channel, and I can confirm that the channel does not exist anymore as shown below.

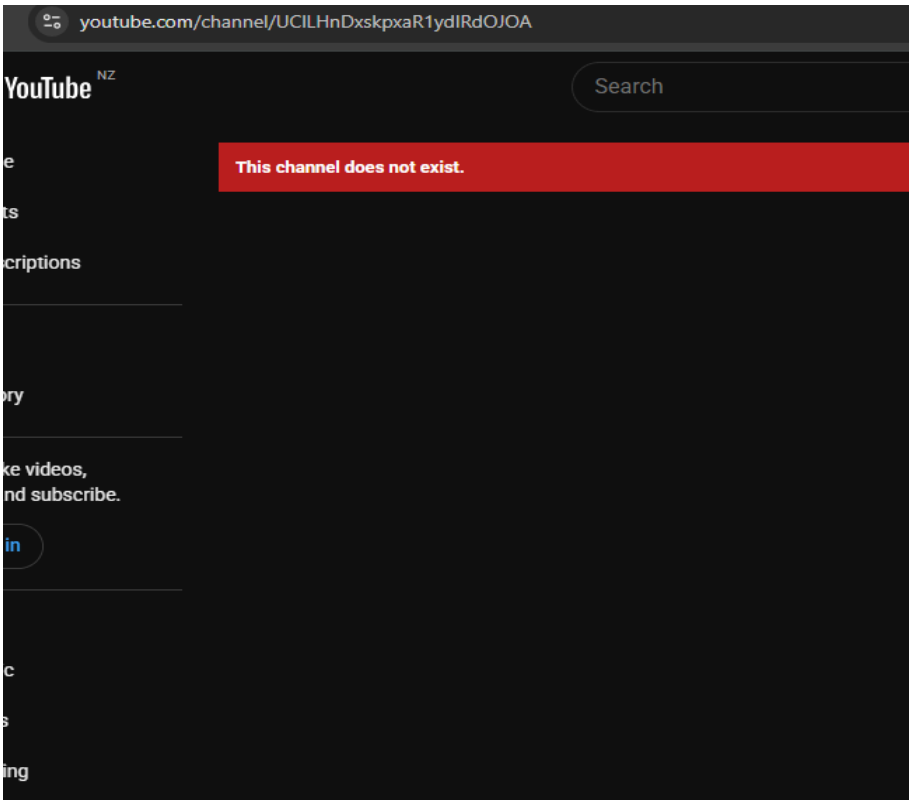


Figure 3.4. Unavailable *DAPAT BALITA YouTube* channel page, showing that the identified channel was no longer accessible at the time of investigation. Source: Screenshot captured by author from *YouTube*, on (25.09.2025).

In order to interpret the anomaly recorded by *vidIQ* on 21 April 2024 (a single-day -957,551 views), I examine official platform documentation and peer-reviewed research. I examined how YouTube's enforcement system operates and what specific platform processes can contextualise sudden drops in views and revenue, including hybrid machine-learning detection combined with human review and rapid, early removals that can precede widespread viewing, as well as broader systematic enforcement effects at the channel level. I also draw on research that examines view deduction and the limits of public and API metrics. These materials are used to contextualise the plausibility of abrupt step-changes in channel-level metrics and to guide cautious interpretation of the *vidIQ* signal. YouTube's mechanism descriptions and *Google*'s contemporary scale figures are cited only to show that abrupt step-changes are a known and routine outcome of platform enforcement or integrity corrections as contextual evidence.

Accordingly, I treat the date of the daily bin in which the anomaly appears as a temporary reference point for the likely timing of the underlying event, pending access to the original watch-page URL or an authoritative archival capture.

vidIQ channel analytics, revealed a co-occurrence on 21 April (UI time zone which is NZST, UTC+12) NZST daily bin 00:00-23:59 ~ 20 Apr 20:00-21 Apr 19:59 PHT, UTC+08) and of an Estimated earnings range of $0-$0 together with a Views Change of -957,551 prompted the anomaly review.

NOTE: A daily bin is a 24-hour aggregation window in which platform collects metrics such as views, interactions, reports... by calendar day in a UI time zone (user time zone), for example, the 21 April 2024 bin in NZST runs from 00:00 to 23:59 NZST and corresponds to 20:00 on 20 April to 19:59 on 21 April in Philippine Time.

Without access to the exact Youtube log, this one-day anomaly is used as an observational timing mark to anchor a provisional channel disruption window, not as definitive evidence of a specific enforcement action.

Because third-party tools such as *vidIQ* lack access to *YouTube*'s internal moderation/enforcement logs, and the YouTube Data API returns only point-in-time totals under authentication, quota and rate-limit controls, causal attribution is not possible. The API does not expose historical evolution or removed-view counts. Conclusions are therefore framed as consistency assessments in public signals, not as single-cause determinations.

YouTube's official also notes that metric counts may be temporarily slowed, frozen, or changed while systems verify legitimacy, and that low quality playbacks are discarded (YouTube Help, n.d.-a). YouTube further states that some invalid traffic is only detected after the fact, leading to subsequent adjustments including observable drops in views in YouTube Analytics (YouTube Help, n.d.-b).

Therefore, interpretation relies on pattern recognition from public signals cross referenced with platform documentation (Castaldo et al., 2024; Google, n.d.; YouTube, 2018). For example, an API return of 900 views at a given timestamp indicates only the total at that moment. It cannot show whether the count earlier was higher or later removals occurred.

Castaldo et al (2024), explains that this is only a snapshot of the moment, it does not contain history or how the number got there. It cannot show whether the count earlier was higher or later removals occurred (Castaldo et al., 2024). Therefore, a value of 900 views at 12:00 shows only the total at 12:00, not whether it was 1200 at 11:00 or fell further after 12:00. My goal here was only to date the removed original clip by using the one-day anomaly as a timestamped reference point. Please bear in mind that this estimate should be revised once the watch page timestamp or archival capture become available.

The third-party website (channel), vidIQ (web), was accessed on 5th of August 2025 to retrieve channel-level historical metrics and identify single-day spikes/drops as observational timing signals. These anomalies may reflect platform or creator actions but cannot support causal attribution (ex: bulk deletion, privatization, invalid-traffic corrections) because third-party tools lack access to internal enforcement logs and the YouTube Data API provides point-in-time totals only (Kohne et al, 2024; Castaldo et al., 2024).

Accordingly, these anomalies provide the strongest publicly observable timing signals available for this case and their interpretation is contextualised using platform documentation and public reporting.

Estimated earnings are modelled, meaning they are calculated using algorithms rather than actual payments, and even YouTube’s official revenue figures remain provisional until month-end (Google, 2025).

An additional step in estimating the earliest time of publication (Tzero) or T_0 from the third party vidIQ data, is a cluster of titles about China/Armed Forces of the Philippines (AFP), listed under the channel's "Latest Published Videos" for 20-21 April 2024 NZST (~ 19-21 Apr PHT). I used title semantic alignment with the reported allegations to prioritise candidate uploads potentially associated with the fabricated audio.

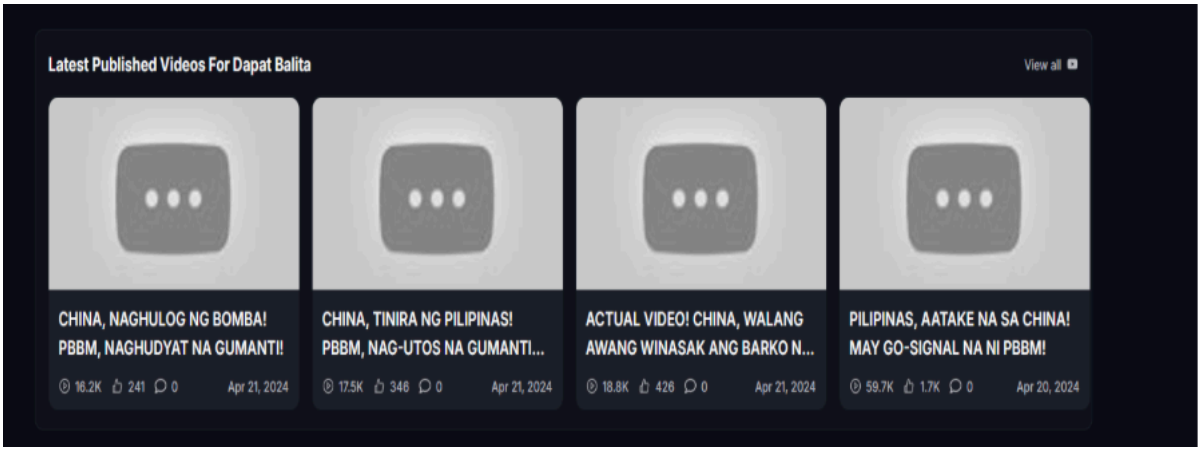


Figure 3.5. Latest published video entries for the *DAPAT BALITA* channel on 22 April 2024, including titles referencing China and the Armed Forces of the Philippines. Source: Screenshot captured by author from vidIQ, on (27.08.2025).

Since monetisation swings and aggregate deltas (overall changes), are not reliable timing signals, From the "Latest Published Videos for Dapat Balita" block, the channel shows four recent uploads clustered on April 20-21, that list is used to narrow the window by identifying the earliest published listing that matches the allegation. In this step, I use title alignment with the reported allegation as a pragmatic OSINT screening rule to prioritise candidate uploads while the original watch-page remains unavailable.

Next step, by using the per-video URLs (the unique YouTube watch-page link for each listed video above) listed in *vidIQ's* "Latest Published Videos" for *Dapat Balita*, I attempted to preserve or retrieve the original YouTube watch pages via multiple archival and cache services.

First, I consulted *archive.today* (archive.today, n.d.), an on-demand web archiving service selected for its ability to generate time-stamped, however, the tool consistently returned "video unavailable / no snapshots," indicating that the relevant watch pages were already inaccessible at the time of capture.

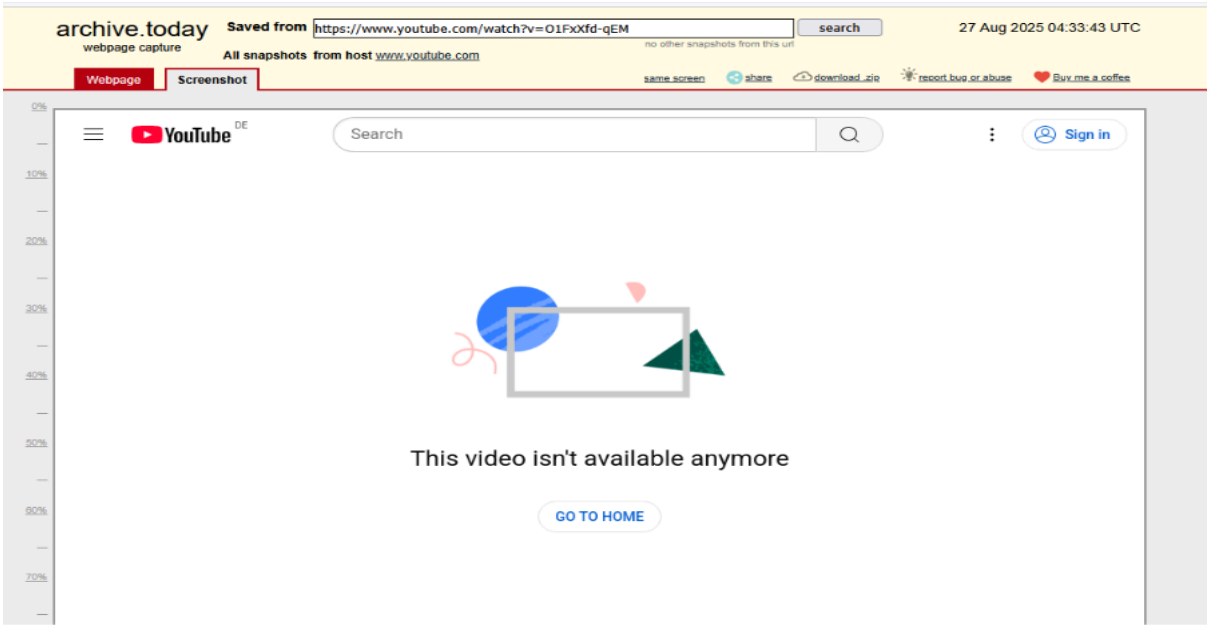


Figure 3.6. archive.today result showing no available snapshot for the relevant YouTube watch page.
Source: Screenshot captured by author from archive. On 27.08.2025.

To cross-check, I also used *Google Cache* (CachedView, n.d.), which sometimes holds short lived cached HTML from recent crawls, and *Wayback Machine* (Internet Archive, n.d.), the standard source for historical snapshots in academic work; neither had a prior capture of the target watch URLs. Crawling is another word for bot and is an automated program that systematically visits web pages, downloads their content, and follows links to discover more pages.

I also tested *WebCite* (WebCite, n.d.) which is a legacy academic archiving service long used in scholarly publishing to create citable, fixed snapshots of web pages and *Megalodon* (ウェブ魚拓 [Megalodon], n.d.) which is an independent Japanese snapshot service with different capture policies and regional crawling that can sometimes preserve pages missed elsewhere to add independent coverage. and both likewise yielded no copies. This was done only to confirm absence and to freeze secondary evidence which are the vidIQ listing; and the official advisory.

As of 27.8.2025, I found no publicly accessible snapshots of the target watch-page URLs across five archives This may reflect removal before capture, restricted access (login/age/region), robots/rate-limit policies, dynamic rendering issues, URL/parameter changes, or post-capture takedown of archived copies. I therefore treat the original publish time as currently unrecoverable from public source, subject to revision if the watch page timestamp or archival capture is later recovered. Based on all gathered data so far, I adopt a T_0 as a window, from 19 April 20:00 to 20 of April 19:59(PHT).

I have estimated the incident window for the deepfake to be between April 19, 2024, 20:00 PHT and April 20, 2024, 19:59 PHT, and all Maximum and minimum latencies below are calculated against this window.

3.4 Response-Timing Dataset and Defender Latency

Defender latency is measured (shown by a triangle sign), by deducting the time when the first time the fabricated clip had appeared online from the first time an official agency responded as follow:

Δ latency = T(official) -T(zero). For instance, the fabricated clip appeared at 10:30am Tzero or T_0, and the first visible authoritative response or defender stakeholder was at 14:00 the same day, so the defender latency shown as Δ is 3 hours 30 minutes. I report defender latency bounds Δ (min-max) normalised to PHT and visualise them in a line chart (Figure X) to show dispersion across defender classes..

I measure response latency for two reasons. One is that a single defender response does not determine when the earliest response was. Secondly, a single response timestamp does not show how long the public was exposed to the fabricated post without an authoritative response. It indicates a period when misinformation could shape the narrative and, as a result shape public perception of the incident and influence decisions.

It can indicate the points where platform anomaly detection mechanisms should have triggered but did not in a high stake, real world political crisis. By comparing these intervals, I can assess whether detection and mitigation occurred in a timely manner which are central to accessing anomaly detection limitations in practice and as a crucial factor in Crisis management at the Pro-Crisis management Stage.

Since anomaly detection pipelines exhibit opacity, latency and fragility under adversarial bursty loads, as established in Literature review chapter, in this study I measure the first public availability T_0 as a point when determinable or as a bounded interval when not. Defender (stakeholder) latency is also assessed for public facing responsiveness. In this way the limitations identified in all the anomaly detection literature will be tested against what the public actually experienced. This operationalises the evidence used to address research questions one and two of the study under explicit OSINT only constraints.

Defender responses posted online are timestamped from their public platforms and when exact times were not visible, coding is used from page metadata or structured UI residues (ex: Twitter/X permalinks, newsroom CMS content management system) timestamps and logged for audit. UI residues are small, persistent traces left after a post is changed/removed that still let you anchor time of the events (ex: “Video unavailable” page still showing the stable video ID, visible fact-check banner with its own timestamp, missing slot in a channel’s “Latest” grid, embed fallback thumbnail showing former title, or permalink entry with publish date time)

Figure 3.7(a). Code-inspection screenshot used to verify the exact publication timestamps from accessible news reports and public webpages.
Source: Screenshot captured by author from the web page code/metadata inspection, On August 2025.

Note. The screenshot documents how visible page metadata or structured web residues were inspected where displayed publication times were incomplete or unavailable.

Figure 3.7(b). Close-up of the timestamp metadata field identified through webpage metadata inspection.
Source: Screenshot captured by author from visible webpage metadata / structured data, in August 2025.

3.5 Time-Zone Normalisation and Latency Calculation

UTC stands for Coordinated Universal Time (no daylight saving)

NZST is (Pacific/Auckland) which = UTC+12

PHT(Philippines) = UTC+8

The Time zone figures were captured in the respective tool/UI (User interface) time zones which is my time zone in Auckland/Pacific (website UIs vary). For analysis, all timestamps from vidIQ, YouTube UI pages, web-archive snapshots, and news outlets were converted to Philippine Time (PHT; UTC+08:00). Because the source day boundary may be ahead of or behind PHT, an event near midnight in the source UI (user interface) time zone can move to the previous or next calendar day after conversion. For instance, when UI based in Auckland shows the time at 22nd of April, 00:15, PHT (UTC+08:00) shifts it to 21st of April, 20:15 and therefore a "daily" count that tool shows under April 22, would be analysed under April 21 in PHT. Maps/figures serve only to anchor locations and time-zones. They are not used to assess veracity or legal claims.

3.6 Coding, Defender Classes and Crisis-Phase Mapping

After establishing a bounded window for the earliest public appearance (T_0), I compile a chronological event log to identify pre-crisis indicators of escalation risks as part of crisis management analysis. I then coded contemporaneous records, documents and news reports to identify post triggers developments, early stage of crisis impacts and immediate consequences.

The case analysis first establishes the structural conditions, and pre-crisis triggers that shaped escalation risk, including constitutional command authority, the Philippines-China legal dispute, international alliances and joint exercises, the deepfake regulatory environment, and international media amplification as pre-crisis signalling.

The analysis then proceeds across three phases that map onto the incident's escalation logic: Phase 1, trigger and initial public recognition; Phase 2, early domestic media amplification and framing; and Phase 3, provenance, takedowns and cross-platform enforcement.

Across this structure, Q1 is addressed across Phases 1-3 because visible anomaly-detection failure is assessed through the absence or presence of public-facing labels, warnings, enforcement traces and platform signals across the incident. Q2 is addressed primarily in Phases 2-3 because defender latency is measured through the timing of visible responses from competent actors after the fabricated audio entered public view.

Q3 is addressed through the structural/pre-crisis section and Phases 1-2, with supporting evidence from Phase 3 where defender behaviour and platform enforcement intersect with early-risk conditions. For Q3, "contextual conditions" refers to the institutional, security/geopolitical and information-system conditions shaping escalation risk, while "timing" refers to the OSINT-reconstructed sequence across the pre-circulation, trigger and early-crisis windows, operationalised through bounded timestamps and latency measures.

Defender classes are examined within Phase 2 to align role-based response patterns with the earliest period of domestic amplification. This allows the study to compare how different competent actors entered the public information space and whether their visible responses appeared during, or only after, the period in which the fabricated audio was being publicly framed.

3.7 Methodological Limitations and Ethical Position

The limitations outlined in this section are not treated as incidental obstacles, but as methodological conditions that shaped what could be reconstructed from the OSINT record. Tyrrell's observation that relevant open-source information may be non-digital, non-English, or unpublished supports this framing: language, access, and missing records are therefore treated as methodological limitations rather than minor practical inconveniences (Tyrrell, 2002). In this study, those limitations were evident in the need to translate Filipino, other Philippine-languages, Chinese, and Japanese resources into English with the assistance of the AI-powered translation tool Google Translate. Although the broad meaning of the source material was generally clear, machine translation may not fully capture nuances, idiomatic expressions, tone, or political context. Additional limitations arose from the absence of the original watch page and audio file, as well as the lack of recoverable propagation posts.

The single-case design also limits the breadth of comparison and the extent to which findings can be generalised across all voice-cloning or synthetic-audio crises. This limitation was addressed by treating the Marcos incident as an analytically bounded case rather than a statistically representative sample. As explained in Section 3.1, adding a comparative case would have required more than identifying another political voice-cloning incident; it would have required a case with comparable political or national-security stakes, defender-response visibility, public traceability, platform-policy conditions, and a closely comparable time frame.

Most available examples were either older, election-focused, non-military in framing, or not contextually compatible with the Marcos case. The restricted completion time frame further limited the feasibility of developing a robust comparative design, particularly because each additional case would have required its own OSINT reconstruction, timing verification, source triangulation, policy-context review, and limitations assessment. Including a weaker or less comparable case would therefore have risked reducing analytic precision rather than strengthening the study.

This dissertation therefore does not claim to generalise statistically across all voice-cloning crises; instead, it uses the Marcos case to examine how anomaly-detection limitations, defender latency, and escalation risk can be analysed within one politically sensitive national-security voice-cloning crisis.

In anomaly-detection and misinformation research, propagation curves help assess whether delayed detection allowed harmful content to spread widely. In this study, access to propagation dynamics was not possible. Propagation dynamics refers to the volume, speed, and breadth of reposts or re-shares across accounts, which would be needed to show how widely and quickly the fabricated audio circulated before defenders responded. Measuring propagation dynamics would have linked defender latency (Δ) to the practical impact of the incident.

Mapping propagation dynamics is also critical for crisis communication and stakeholder analysis. The severity of a communication failure depends not only on the existence of harmful content, but also on when and how it reaches key publics. Without knowing how quickly or widely the fabricated audio circulated, this study cannot determine whether stakeholders such as journalists, opposition figures, or international audiences were exposed before official responses arrived, or whether circulation stayed limited. OSINT research also emphasises that understanding who amplifies content across networks is crucial for assessing political significance and the opportunities defenders had to detect and verify harmful content.

Given the absence of the original audio, the removed watch page, and a recoverable sample of propagation posts, a propagation-dynamics investigation was not possible. Although third-party channel-level analytics helped identify the reported original YouTube channel name, channel ID, and later unavailability of the channel/page, these traces did not provide the original upload metadata, the audio file, comments, repost chains, or user-level propagation history needed for propagation-dynamics analysis.

The removal and deactivation of key monitoring pages, together with missing archival records, prevented recovery of the minimum observable sample typically required to assess propagation dynamics across 10 to 100 distinct posts from independent accounts. As a result, propagation dynamics was effectively impossible to reconstruct from the available OSINT record.

Tyrrell argues that open-source analysis is not a substitute for non-public intelligence work but can provide context and identify knowledge gaps (Tyrrell, 2002). In this dissertation, that distinction matters because OSINT can reconstruct visible timing, public-facing responses, and available platform traces, but it cannot reveal internal platform-detection outputs, classified government responses, proprietary moderation logs, or complete propagation dynamics.

The absence of the original watch page and recoverable propagation samples therefore creates an evidence boundary, not a failure of method. Although the exact original upload timestamp could not be recovered, the available evidence allowed the study to establish a substantiated bounded T_0 window, from surviving public traces. Because the defender-response timestamps could be recoverable and documented precisely, defender latency (Δ) was calculated against that T_0 window as minimum and maximum latency values.

The absence of recoverable propagation evidence creates both a methodological limitation and an ethical boundary. The analysis can report verifiable defender latency (Δ), but it cannot infer propagation dynamics, audience exposure, stakeholder encounter, public impact, or the causal effect of platform detection beyond the available evidence. Therefore, this study avoids making unsupported claims about circulation, exposure, or platform-side intervention.

The missing circulation record also creates evidence of asymmetry: visible defender responses can be documented more reliably than platform-side detection activity, user-level exposure, or cross-platform circulation. Consequently, the available data tend to favour defender-side evidence.

The metrics used to reconstruct the publication timing of the fabricated clip were derived from third-party tools and interpreted as contextual references rather than platform-internal ground truth. Because some timing evidence was recorded across different time zones, time normalisation was required. Such conversion can shift daily data bins and affect the interpretation of the OSINT record, although visible defender publication times were documented where public timestamps were available.

Platform opacity was another limiting factor in this time-bounded study. With no publicly timestamped enforcement logs and extensive removals, the study could not verify whether platform-side detection, labelling, or enforcement occurred before public-facing defenders responded.

Ethically, and in line with anomaly-detection scholarship that emphasises latency, opacity, and adversarial sensitivity, the study therefore refrains from making propagation claims under limited evidence. Instead, it focuses on reporting verifiable defender latency (Δ) and underscores the need for transparent, timestamped platform telemetry to enable independent auditing.

From a governance standpoint, labelling policies were also in transition during the case period. Meta announced new "Made with AI" labels for AI-generated video, images, and audio beginning in May, broadening its earlier doctored-video policy, which applied only to videos that had been manually manipulated or edited in misleading ways.

At the same time, the Meta Oversight Board described the existing guidelines as "incoherent" and recommended expanding them to include non-AI and audio-only content. Reuters corroborates that, before May 2024, Meta's approach covered only a narrow subset of manipulated media and was under active revision, with specific calls to include audio and non-AI edits (Paul, 2024). These constraints help explain both the coverage gap for voice-based media and the absence of user interface (UI)* labels observed in April. Therefore, during the critical first 24-36 hours, the information environment was shaped more by public-facing defenders than by verifiable platform intervention.

* User-interface (UI) layer refers to the public-facing platform surface visible to ordinary users, including labels, warnings, content notices, metadata displays, unavailable-page messages, and other visible platform signals

Chapter4: Structural conditions of the case study and pre-crisis triggers

4.1. Constitutional Command Authority

Since the deepfake clip appeared to show President Ferdinand R. Marcos Jr. issuing military instructions to attack China, in the context of tensions with China, it is necessary to clarify how command authority over the armed forces is structured in the Philippines.

Under the 1987 Constitution of the Philippines, the President holds the position as Commander-in-Chief of all armed forces of the Philippines. The President exercises direct control over the military, including calling out forces to prevent or suppress lawless violence, invasion, or rebellion. that command authority is limited but regulated. The powers include deploying the armed forces, declaring martial law, and suspending the writ of habeas corpus with varying degrees of authority and constraints (Respicio & Co., 2024).

In the Philippines, operational control is centralised to ensure coherence and prevent fragmentation of authority. According to Executive Order No. 308 (1950), “All the five major commands shall be under the direct command of the Commanding General of the Armed Forces of the Philippines” (Government of the Philippines, 1950). As the president holds the highest position in a centralised, constitutionally defined chain of command, any communication that seems to involve presidential military orders, is taken very seriously and must be understood within the legal and geopolitical framework in which those forces function.

Therefore, a forged apex “order” purportedly, coming from President Marcos of Philippines, commanding the military to be prepared to attack China was a dangerous national security news.

It is also important to first consider the strategic locations and maritime zones involved within the South China Sea, which was displayed in the video where the synthetic audio of the president Marco was used. The area is claimed by multiple countries in the region for economic purposes.

The maritime domain encompasses distinct economic, diplomatic, military, and legal dimensions. Diplomatic and political tensions in this sphere intensify as states seek to expand their jurisdiction over offshore resources. Such expansionist claims frequently generate disputes concerning the precise delineation of maritime boundaries and exclusive economic zones (EEZs). Notably, the construction of artificial islands does not confer, nor does it extend, entitlements to territorial seas, EEZs, or continental shelf areas (Joint Chiefs of Staff, 2021b).

4.1.2. Legal background of Philippines-China Dispute (The Philippines arbitration against China (2013-2016)

Filipino fishermen traditionally had the right to fish at *Scarborough Shoal,* a rich fishing ground in the South China Sea also frequented by Chinese fishermen. However, the Philippines' claim was that China began restricting Filipino access, promoting to initiate arbitration proceedings against China, under Annex VII of the *United Nations Convention on the Law of the Sea (UNCLOS).* In its application to the *Permanent Court of Arbitration (PCA),* an independent intergovernmental body that administered but did not adjudicate the case. The Philippines sought a ruling on several key issues including:

The legality of China's "nine-dash line" and its associated claim to historic rights within the South China Sea, the status of maritime features and what maritime zones they generate under *UNCLOS,* and whether China's activities such as land reclamation, interference with fishing and petroleum exploration, and failure to prevent Chinese fishing violated *UNCLOS* (Permanent court of arbitration, 2016, p. 1-3).

China refused to participate, arguing that the tribunal lacked jurisdiction and asserting that the dispute concerned sovereignty and maritime boundary delimitation and claimed that Philippines has breached the international law by dismissing the previous arrangements (Ministry of Foreign Affairs of the People's

Republic of China, 2014) which are excluded from compulsory arbitration under Article 298 and China's 2006 declaration (Permanent court of arbitration, 2016, p. 6-7).

The tribunal was composed of five arbitrators from Ghana, France, Poland, the Netherlands, and Germany (Permanent court of arbitration, 2016, p. 3). Vietnam submitted a statement to the Tribunal and other observer states that attended the hearing were Malaysia, and Indonesia (Permanent court of arbitration, 2016, p. 7). The tribunal did not decide sovereignty over any land territory or questioned it (which is part of China's historical sovereignty claim) (Permanent court of arbitration, 2016).

"Because Scarborough Shoal is above water at high tide, it generates an entitlement to a territorial sea, its surrounding waters do not form part of the exclusive economic zone (EEZ)*, and traditional fishing rights were not extinguished by the Convention." (Permanent court of arbitration, 2016, p.10).

Note: EEZ is not sovereignty over the shoal but only over the resources such as fishing (Permanent court of attribution, 2016).

However, the tribunal emphasised that both Chinese and Filipino fishermen had traditional fishing rights at Scarborough Shoal, and that China had unlawfully interfered with those rights by restricting Filipino access (Permanent court of arbitration, 2016, p. 2).

"The Convention gives other States only a limited right of access to fisheries in the exclusive economic zone (EEZ), (in the event the coastal State cannot harvest the full allowable catch) and no rights to petroleum or mineral resources" (Permanent court of arbitration, 2016, p. 9).

It concluded that China constructed artificial islands, and failed to prevent its fishermen from fishing in the Philippine EEZ, in addition the Chinese fishermen knowingly have caused severe harm to coral reefs and large-scale harvesting of endangered sea turtles, coral, and giant clams, with a failure by the Chinese authorities to stop it. The tribunal had also "held that Chinese law enforcement vessels had unlawfully created a serious risk of collision" when they physically obstructed Philippine vessels (Permanent court of attribution, 2016, p. 2).

*EEZ or "exclusive economic zone," is an area of the ocean, generally extending 200 nautical miles (230 miles) beyond a nation's territorial sea, within which a coastal nation has jurisdiction over both living and non-living resources (Ocean Exploration, 2023).

The tribunal expressly rejected China's asserted historic rights to resources where they conflict with UNCLOS's maritime zone system and extinguished to the extent of incompatibility (Permanent court of attribution, 2016, p. 9).

"Since the Chinese fisher men knowingly have engaged in the harvesting of endangered sea turtles, coral, and giant clams on a substantial scale, and the Chinese authorities had not fulfilled their obligations to stop such activities" (Permanent court of arbitration, 2016, p. 11).

According to Chinese sources, the maritime feature at issue, Ren'ai Jiao in Chinese and Ayungin in Filipino, lies within the "U zone" which Beijing presents as part of its territorial sovereignty and interests over islands and reefs in the South China Sea (Chubb, 2016 ; Wang, 2024; Zong, 2024;). It is important to note that this dash-line claim was not adjudicated in the 2016 proceedings before the Permanent Court of Arbitration (Permanent Court of Arbitration, 2016). At the same time, the area lies close to, and overlaps with, the Philippines' Exclusive Economic Zone (EEZ), whose EEZ rights were recognised in the 12 July 2016 award (Permanent Court of Arbitration, 2016, July 12).

Under the Exclusive Economic Zone and the Continental Shelf law of the People's Republic of China, Articles 3, 4, 8 and 14 grant China sovereign rights over the exploration, exploitation, conservation and management of natural resources both living and non-living in its EEZ and continental shelf, and assert jurisdiction over construction, operation and management of artificial islands, installations and structures, maritime scientific research, environmental protection, drilling operations, safety zones and related customs, financial, public-health and entry/ exit controls in those maritime zones (National People's Congress of the People's Republic of China, 1998, June 26).

The implications of a fabricated clip suggesting Philippine retaliation against China become more significant when viewed considering existing alliance arrangements, where defence commitments with partners mean that any perceived escalation in the South China Sea can carry consequences beyond the bilateral Philippines - China relationship.

4.1.3. International Alliance and Joint Exercises

“An alliance is a relationship that results from a formal agreement (e.g., treaty) between two or more nations for broad, long-term objectives that further the common interests of the members. Many contingency plans to deter or counter threats are prepared within the context of a treaty or alliance framework” (Joint Chiefs of Staff, 2017a, Ch. II, p. II-21)

Bilateral Defence Policy:

On May 3, 2023, a joint policy framework was issued by the U.S. Department of Defence and the Philippine Department of National Defence, reaffirming Mutual Defense Treaty (MDT) coverage (including any attack anywhere in the South China Sea on either side’s public vessels, aircraft, armed forces, or coast guards) and setting multi-domain (land, sea, air, space, cyber) lines of effort to modernise the alliance (U.S. Department of Defense, 2023).

Balikatan Joint Military Training:

Initially the primary participant of the Balikatan exercise was the AFP (Armed Forces of the Philippines) and the U.S. military with over 16000 personnel from both sides in addition to 14 other countries. On 17th of April the scale and scope of the Balikatan 2024 exercise was highlighted by announcing Australia and France as the new direct participants in addition to 14 other nations including France, Germany, Canada, Australia, New Zealand, Brunei, Great Britain, India, Indonesia, Japan, Malaysia, Republic of Korea, Singapore, Thailand, and Vietnam as the AFP-hosted international observer in this program. The inclusion of the Australian Defence Force and, for the first time, the French Navy as direct participants marked an expansion of the exercise's international involvement. It was stated that the exercise will encompass a variety of complex missions, such as maritime security, sensing and targeting, air and missile defense, dynamic missile strikes, cyber defense, and information operations (Flores,2024; U.S. Embassy Manila, 2024).

Lt. Cmdr. Maria Christina Roxas, Chief of External Affairs Branch of the Philippine Navy's Public Affairs Division, briefs her team's information Operation plan to participants during the Information Warfighter Exercise ahead of Exercise Balikatan 24 at Camp Aguinaldo on April 4.

Figure 4.1. Lt. Cmdr. Maria Christina Roxas, Philippine Navy, briefing her team's information operations plan for the upcoming Exercise Balikatan 24. Source: Screenshot captured by author in August 2025, from U.S. Embassy Manila (2024). Original post-dated 6 April 2024; reshared on 16 April 2024.

Within this legal and command framework, a fabricated presidential "order" referring to military action against China carried escalation risk because it appeared within a disputed maritime domain already shaped by contested maritime claims, command authority and defence obligations. The issue was therefore not only that the audio was false, but that its content connected misinformation to an existing national-security setting.

4.1.4. Deepfake Regulatory Environment

Based on A Literature Review of the Legislation and Regulation of Deepfakes in the Philippines, Blancaflor et al. (2023) stated that the United States has already established several targeted measures such as the *2018 Malicious Deepfake Prohibition Act, California's Deceptive Recordings Act* and Elections. *Deceptive Audio or Visual Media Act, Texas Senate Bill 751 (2019),* and *the federal IOGAN Act of 2020* which address issues ranging from criminal penalties in election-related contexts to disclosure requirements, research mandates, and national-security directives involving DHS, NIST, and NSF (Blancaflor et al., 2023, pp. 394-395).

By contrast, the Philippines still lacked a mature regulatory framework, resulting in a comparatively underdeveloped system. Instead, legislators had only "begun to move toward controlling deepfake concerns by filing a resolution," while the state continued to rely on existing data privacy and cybercrime provisions to protect institutions and citizens (Blancaflor et al., 2023, p. 392). Since there is no new regulation, in practice, this means the regulation around synthetic media as of April 2024, was still indirect and incomplete, with no specific statute written for deepfakes themselves. This regulatory gap justifies the absence of timestamped platform labels, takedown traces, or other visible enforcement artifacts during the incident window reflecting a governance environment with limited public-facing regulatory signals.

4.1.5. International Media Amplification as pre-crisis signalling

International outlets contributed to the escalation context across three windows by signalling alliance visibility, amplifying the official framing once the incident was publicly acknowledged, and documenting continued real-world friction in the maritime environment.

In the pre-crisis window, coverage of Balikatan 2024 highlighted heightened alliance visibility and regional sensitivity, including reporting that the drills would extend beyond Philippine territorial waters and would involve expanded international participation and complex missions, framing these developments as likely to be read by Beijing as provocative signalling in a contested theatre (Lema, 2024; Flores, 2024; Reuters, 2024; Al Arabiya English, 2024a).

This pre-circulation attention functions as an external indicator that the broader security environment was already salient internationally before the fabricated Marcos clip entered public view.

4.2. Phase 1: Trigger and initial public recognition of the Marcos deepfake

Right from the beginning, the content of the clip and the way it was officially described positioned the incident above ordinary misinformation. The reported lines calling on “armed forces and special task groups” to retaliate against China, labelling the alleged Duterte- Xi “gentlemen’s agreement” as “a clear form of treason against our motherland” and asserting that Filipinos’ lives could not be “compromised” to defend “our… sea and maritime domains” presented a fabricated presidential voice apparently authorising or justifying military action in a disputed maritime theatre (VERAfiles, 2024; NDTV, 2024; FORUM Staff, 2024; Philippine Star, 2024a; Rappler, 2024b; Rappler, 2024f). In the context established in 4.1 section, centralised Commander in Chief authority, contested South China Sea zones, and heightened pre-crisis signalling such a message is structurally escalation prone even before any verification or response.

Additionally, the first visible response in the public record is an authority intervention not a platform machine. The PCO’s Facebook advisory describing the clip as an AI-generated deepfake, denying the directive, and announcing coordination with Department of Information and Communications Technology (DICT), National Security Council (NSC), National Cybersecurity Inter Agency Committee (NCIAC) and private-sector partners (Presidential Communications Office, 2024, April 23).

PNA’s follow-up the next morning reinforces this framing with definitional language on deepfakes and calls for public vigilance but likewise takes the form of a traditional written advisory rather than a platform-native label or warning (Esguerra, 2024). From an OSINT-only perspective, no publicly observable evidence in this initial trigger window of automated anomaly-detection or labelling systems acting on the clip, such as fact-check banners, manipulated-media labels, or other user interface (UI) markers such as “Made with AI” Label, was found in this study.

Instead, the early defence visible in the record consists of the executive centre’s written denial and deepfake framing (AIAAIC, 2024; Presidential Communications Office, 2024b, Rappler, 2024d, Rappler, 2024b), and subsequent news reports that repeat and contextualise that message.

The combination of a highly charged presidential voice clip and an already tense security environment describe how easily the fabricated message could have escalated the risk before any initial authoritative response appeared.

4.3. Phase 2: Early media amplification and framing

Following the *Presidential Communications Office*'s *Facebook* warning on 23 April 2024, domestic defenders across print, broadcast and online outlets began relaying and contextualising the official denial. (AIAAIC, 2024; Presidential Communications Office, 2024b, Rappler, 2024d, Rappler, 2024b) *Manila Bulletin* highlighted the Palace warning alongside the government's Media and Information Literacy (MIL) campaign and coordination with the *Department of Information and Communications Technology (DICT), the National Security Council (NSC) and the National Cybersecurity Inter-Agency Committee* (NCIAC) (Geducos, 2024).

GMA Integrated News carried an initial straight news item on the PCO advisory, followed by an explainer that explicitly described the audio as a deepfake, labelled the content as false and cited PCO Assistant Secretary *Dale de Vera* (Locus, 2024a; Locus, 2024b). *ABS-CBN News* similarly reported the government's denial in the early hours of 24 April (ABS-CBN, 2024b).

State outlet Philippine News Agency (PNA) republished the warning, stating that the clip falsely portrayed President Marcos directing the Armed Forces of the Philippines to act against another nation and adding definitional language on deepfakes as AI-generated manipulated media (Esguerra, 2024). *RADYO Inquirer 990AM* broadcast a Filipino-language report that referenced and translated the PCO's English statement (Escosio, 2024), while *Philstar Global* reported that Malacañang had warned the public about a manipulated audio clip depicting President Marcos as ordering military action (Flores, 2024b).

On platforms, *ANC 24/7 (ABS-CBN's YouTube* news channel) carried coverage that included screenshots of the removed *YouTube* video and reported coordination with major platforms over takedown efforts (ANC 24/7, 2024b).

Business Mirror framed the situation primarily through the government's response, emphasising the removal of the video from YouTube, the activation of the Media and Information Literacy (MIL) campaign, and coordination with *DICT, NSC and NCIAC* (Medenilla, 2024).

Rappler documented the case across multiple formats: a web article and YouTube Shorts that showed how the fabricated Marcos audio was overlaid on slides featuring *South China Sea* and Armed Forces imagery and quoted key lines from the clip, preserving public-trace evidence of both message content and visual presentation (Rappler, 2024b; 2024c; 2024d; 2024f).

At the same time, Assistant Secretary Dale de Vera appeared on camera in a *PTV* Bagong Pilipinas Ngayon segment to reiterate the government's position and describe ongoing coordination with other agencies (PTV Philippines, 2024), while political figure Jolo Revilla amplified the PCO's warning by reposting it to his own audience (Revilla, 2024).

These responses show how, in the days immediately after the PCO statement, four defender classes became visible in the public record: the executive centre (*PCO* and Malacañang), national news media such as *Manila Bulletin, GMA, ABS-CBN, PNA, Philstar, Rappler*, platforms and cyber/security agencies as described in coverage *DICT, NSC, NCIAC, PNP* Anti-Cybercrime Group, and political influencers.

During and after the trigger window, international reporting primarily acted as secondary amplification of Philippine authorities' reframing of the clip (Calonzo,2024; Maitem, 2024; Reuters 2024a, 2024b; NDTV, 2024) and as continued observation of the underlying geopolitical dispute. *NDTV*'s coverage described the fabricated audio as urging military action against China and underscored the authorities' concern over foreign-policy implications, indicating that the case was intelligible to external audiences once the incident was publicly surfaced through domestic channels (NDTV, 2024).

In the pro-crisis environment reports documented an intensifying South China Sea dispute, including Chinese water-cannon incidents, dangerous manoeuvres, obstruction of Philippine patrols, vessel damage, access restrictions around contested shoals, competing claims over fishing grounds and maritime zones and

later reports of injuries to Filipino personnel (GMA Integrated News, 2024, 2024a; GT Staff Reporters, 2024; Rappler, 2024e; Rita, 2024; VERA Files, 2024) international outlets intensified renewed and ongoing South China Sea confrontations and competing state narratives, including reports on water-cannon incidents near Scarborough/Panatag Shoal, Ayungin/ Second Thomas Shoal and other contested maritime zones, as well as subsequent diplomatic-security framing (Al Jazeera, 2024; Forum staff, 2024; Wang & Fan, 2024; Magramo et al., 2024; Morales, 2024; Reuters Staff, 2024, Reuters, 2024b; Sina News, 2024; Stewart, 2024;The Straits Times, 2024).

This wider reporting context highlights the significance of the fabricated audio not just as misinformation, but as a destabilising information problem that inserted a false presidential voice and apparent military-action claim into an already active maritime-security confrontation.

4.4. Phase 3: Provenance, takedowns and cross-platform enforcement

Publicly visible provenance becomes traceable on 25 April 2024, when the Philippine National Police Anti-Cybercrime Group (PNP-ACG) is reported to have located the uploader and linked the deepfake to specific accounts (Caliwan, 2024; Murcia & Kabagani, 2024). *Philippine News Agency* reports that PNP-ACG cyber patrols acting on a citizen tip "found deepfake videos uploaded by "Dapat Balita" on video platform *YouTube*" and that an associated *Facebook* page named "Dapat Balita" created on 19th of April 2024, had page managers whose primary country locations were listed as the Philippines (Business World, 2024) and Pakistan (Murcia & Kabagani, 2024).

The unit requested preservation of the *YouTube* channel's data and deactivation of the Facebook page (Caliwan, 2024). A *Daily Tribune* report states that PNP-ACG found deepfake videos uploaded by "Dapat Balita" on YouTube, had begun an investigation and sought preservation of all video data video posted on the channel, adding that "Dapat Balita" was deactivated, but another YouTube account with the same name reposted the videos and again describing a Facebook page named "Dapat Balita" created on 19th of April 2024 with managers based on Philippines and Pakistan (Murcia & Kabagani, 2024).

Inquirer.net names DAPAT BALITA as the *YouTube* uploader, reports that the original channel was deactivated, notes that another *YouTube* account with the same name reposted the video, and records that a *Facebook* account using the same branding existed, reinforcing the cross-platform provenance picture (Argosino, 2024).

On 26 April 2024, the *Presidential Communications Office* states that the deepfake video using President Marcos' face and voice has been taken down and that the government intends to pursue legal action against those who created and spread it, moving the response from simple warning to explicit enforcement (Presidential Communications Office, 2024a).

In *PNA* report on the same day, PCO Assistant Secretary Patricia Kayle Martin reiterates that *DICT* and the *National Security Council (NSC)* are actively investigating the issue, that deepfake videos using the President's face and voice have been taken down, and that the objective is not only to identify the culprits but also to file appropriate legal actions against them (Cervantes, 2024).

On 27 April 2024, *PCO* issued a further update stating that a "possible source" of the deepfake audio has been identified, while emphasising that the extent of this actor's involvement remains under investigation, indicating that attribution is still incomplete (Presidential Communications Office, 2024b).

The *Department of Justice* separately directs the *National Bureau of Investigation* to investigate the deepfake audio and instructs the *NBI* to "file the necessary legal action, if warranted, against those behind this fake news," explicitly describing the material as social-media content manipulated using AI deepfake technology and linking the directive to the earlier *PCO-DICT* warning (Pulta, 2024).

PNP- ACG's operational posture is described as intensified cyber patrols targeting deepfake videos of the President, including the request to preserve all video data posted on the *Dapat Balita* channel and to deactivate the associated *Facebook* page, and reference to possible charges under Article 154 of the Revised Penal Code in connection with the Cybercrime Prevention Act (Caliwan, 2024; Murcia & Kabagani, 2024).

By the end of April, a later Daily Tribune report notes that PNP, working with DICT, has successfully removed the deepfake audio recording created and altered using AI, and that a possible source of the deepfake has been identified although the degree of involvement is still being investigated; the same report records House leaders calling for prosecution and a deputy speaker speculating that the bogus material likely originated "somewhere in the south of the country" (Business World Online, 2024).

In parallel with this domestic attribution line, PCO Assistant Secretary Martin states that *DICT* suspects a foreign actor behind the deepfake videos and warns that such content could harm foreign relations and national security, framing the incident as more than ordinary misinformation (Cervantes, 2024).

Inquirer.net reports that “foreign entities’ eyes” are on the Marcos deepfake clip, quoting government concerns about foreign involvement (Mangaluz, 2024). *Philstar Global* similarly reports PCO’s position that a foreign actor is suspected in the incident (Presidential Communications Office, 2024a). *TIME* then links the “foreign actor” framing to a broader Microsoft assessment that China is using AI to sow disinformation and discord across Asia and the U.S., placing the Marcos clip within a wider pattern of AI-assisted operations (Calonzo, 2024).

Chapter 5: Findings

5.1. Escalation risk

The structural and temporal conditions outlined earlier indicate that any message that appears to come from the Philippine President and that speaks in the language of military response operates inside an already heightened risk environment. Constitutionally, the President is Commander-in-Chief of all armed forces of the Philippines and exercises direct control over the military, including calling out forces to prevent or suppress lawless violence, invasion, or rebellion (Respicio & Co., 2024). Executive Order No. 308 further centralises operational control by placing all five major commands under the direct command of the AFP Commanding General (Government of the Philippines, 1950).

In parallel, the 2016 arbitral award clarified specific Philippine EEZ entitlements under UNCLOS, while Chinese legal and policy materials continue to describe features such as Ren'ai Jiao/ Ayungin and some surrounding waters in terms of China's maritime rights and jurisdiction, including resource exploitation and enforcement powers (Permanent Court of Arbitration, 2016; Chubb, 2016; Wang, 2024; Zong, 2024; Law of the People's Republic of China on the Exclusive Economic Zone and the Continental Shelf, 1998). Within this framework, any perceived presidential directive regarding retaliation at sea involves both command authority and disputed maritime areas, making it inherently sensitive to escalation.

The pre-crisis events amplified that sensitivity further. In late March, *Al Jazeera* reported Marcos promising a "countermeasure package that is proportionate, deliberate, and reasonable" in response to what he described as "open, unabating, and illegal, coercive, aggressive, and dangerous attacks" by Chinese Coast Guard and maritime militia agents, adding that "Filipinos do not yield" (News Agencies, 2024).

In early April, domestic reporting framed the Philippines as "prepared to respond to China's attempts to interfere with re-supply missions," referred to China as "pretending to man their water cannon," cited earlier water-cannon incidents and "multi-dimensional counter measures," and warned about "foreign malign

influence" and "amplifiers" spreading "Chinese narratives which run counter to the truth" (Lema, 2024).

Around the same time, Marcos publicly raised the possibility that an undisclosed "gentleman's agreement" by his predecessor might have compromised Philippine sovereign rights, saying he was "horrified by the idea" that territory or sovereign rights had been compromised "through a secret agreement with China," emphasising the absence of documentation or briefing when he took office (ANC 24/7, 2024).

Former officials responded by denying any such compromising deal and warning against misleading narratives, indicating that debates over China policy and misinformation were already politically salient inside the Philippines (ANC 24/7, 2024).

External signalling added another layer. On 12 April 2024, China's Foreign Ministry rejected outcomes of the U.S.-Japan-Philippines summit, described U.S.-Japan-Philippines security cooperation as "bloc politics" that "escalate tensions," reiterated China's position of "indisputable sovereignty" over *Nanhai Zhudao* including *Ren'ai Jiao*, and framed Manila's alignment with U.S. and Japan as a source of regional tension while stating that China would "firmly guard" its territorial and maritime rights (Ministry of Foreign Affairs of the People's Republic of China, 2024).

Around the same period, *Rappler* reported *Marcos* refusing to cooperate with the International Criminal Court (ICC) over the *Duterte* "war on drugs" and later quoted him describing China as "the source of tension in the region," in coverage that also visually situated him alongside U.S. President *Joe Biden* and his inclusion in *Time*'s list of influential leaders (Rappler, 2024; Rappler, 2024a). In parallel, the *Balikatan 2024 exercises* were being expanded and internationalised, with new partners and complex multi-domain drills scheduled in the same April window (U.S. Embassy Manila, 2024).

These structural (constitutional Commander-in-Chief role, centralised AFP command, EEZ entitlements and Chinese maritime claims, alliance frameworks and exercises, and the lack a mature regulatory framework around Deepfakes in Philippines) and temporal (recent coercive-incident reporting, "countermeasure package"

rhetoric, domestic controversy over a “gentleman’s agreement,” Chinese “bloc politics” framing, Marcos’s public characterisation of China as a source of tension) conditions created a pre-crisis scenario where a Marcos fabricated voice, instructing the AFP and "special task groups" to retaliate against China, while calling a

previous agreement “a clear form of treason” is more than just another misleading post (VERAfiles, 2024; Maitem, 2024; NDTV, 2024; Philippine Star, 2024a).

In this setting, Marcos' fabricated voice risks being taken as a serious signal about intent and use of force, the moment it enters the public space, increasing the potential for miscalculation and escalation before any verification or official rebuttal occurs.

The following section examines how each of these defender classes behaved, and how their timing and framing relate to Questions 1-3.

5.2. Stakeholder defender behaviour

Class A: Executive centre

The Executive Centre, represented by the *Presidential Communications Office (PCO)*, was the first to take visible and official action in response to the fabricated audio deepfake.

On 23 April 2024, at 20:31 PHT, the *PCO* issued its first public warning on Facebook, addressing the deepfake video that falsely depicted President Ferdinand Marcos Jr. directing the Armed Forces of the Philippines (AFP) to act against a foreign country. In this post, the *PCO* clarified that “no such directive exists nor has been made” and identified the video as a deepfake created through generative Artificial Intelligence (AI) (Presidential Communications Office, 2024, April 23). This was the first official government acknowledgement of the incident, marking the beginning of the public response. The *PCO’*s response was the first publicly visible, strategic step in the official reaction to the deepfake, but it was not just about issuing a public statement.

It also highlighted the government's ongoing *Media and Information Literacy (MIL)* Campaign, aimed at combating misinformation and disinformation. PCO made it clear that it was working in collaboration with government agencies such as the *Department of Information and Communications Technology (DICT), the National Security Council (NSC), and the National Cybersecurity Inter-Agency Committee (NCIAC)*, along with private sector partners, to address the proliferation of deepfakes and other malicious digital content (Presidential Communications Office, 2024, April 23).

Regarding Q2, from a defender latency perspective, the PCO's actions define the first observable authoritative response. Using the T_0 window, which spans from 19 April 20:00 PHT to 20 April 19:59 PHT, the T(official) response is set at the 23 April 20:31 advisory issued by the PCO. The latency periods between these points are as follows:

Max latency: From T_0 (19 April 20:00) to T(official) (23 April 20:31) = 96 hours 31 minutes (4 days 31 minutes)

min latency: From T_0 (20 April 19:59) to T(official) (23 April 20:31) = 72 hours 32 minutes (3 days 32 minutes).

These latencies define the period during which the fabricated content circulated without an official response. This max latency represents the longest gap between the first public circulation of the deepfake (T_0) and the first official government acknowledgment (T(official)).

Regarding escalation of the Issue (Q3), on 26 April 2024, the PCO escalated its response, framing the issue as a legal and security concern. The PCO announced that the deepfake video had been removed from YouTube and that the government would pursue legal action against those responsible for creating and distributing the deepfake content (Presidential Communications Office, 2024a). This marked a shift from handling misinformation as a public information issue to treating it as a security and legal threat.

On 27 April 2024, the PCO provided an update, confirming that a “possible source” of the deepfake video had been identified, but the full extent of the individual’s involvement remained under investigation (Presidential Communications Office, 2024b). This update signalled that the investigation was ongoing and that enforcement actions were in progress.

By taking legal and investigative action, the PCO framed the deepfake incident within the broader context of national security. The PCO’s escalation to involve agencies like DICT, NSC, and NCIAC, as well as private platforms like YouTube, underscored the gravity of the situation and its potential impact on both national security and the Philippines' international relations. The PCO’s responses over the course of several days show a late yet strategic, escalating approach.

The PCO’s first public response, issued on 23 April, served as the initial defense in combating the spread of the deepfake. In the following days, the PCO’s messages evolved from providing information and clarifications to engaging legal and investigative mechanisms. The shift in messaging from informing the public about misinformation to taking active legal and enforcement actions demonstrates how the PCO responded to the crisis.

The PNA’s reports, which echoed the PCO’s statements, reinforced the PCO's central role in responding to the crisis and amplifying the government’s position on the issue (Esguerra, 2024; Caliwan, 2024). These reports tracked the PCO’s actions, confirming that the deepfake video had been removed and that legal action was being pursued.

As the PCO escalated the situation to a security and legal issue, its involvement in coordinating with other agencies and monitoring the platforms for removal efforts reflects a coordinated national response. The PCO’s public messages, alongside PNA’s coverage, played a crucial role in shaping public understanding of the crisis, framing it as a serious issue that involved not just the manipulation of content, but also the need for legal and cybersecurity measures.

The PCO's responses were crucial in shifting the conversation from misinformation to national security, with the PCO leading the way in terms of both public messaging and government action.

Class B: Media outlets

Class B comprises early domestic national outlets that relayed and contextualised the executive-centre warning across print, broadcast, and online formats, including *Manila Bulletin, GMA Integrated News, ABS-CBN News, Philippine News Agency (PNA), RADYO Inquirer 990AM, Philstar Global, OneNews, Business Mirror* and *ANC 24/7, Daily Tribune* (Geducos, 2024; Locus, 2024a; ABS-CBN, 2024b ; Esguerra, 2024; Escosio, 2024; Medenilla, 2024; Flores, 2024a, 2024b; ANC 24/7, 2024a; Murcia, 2024).

Their earliest reports in this phase primarily functioned as amplification of the Palace position rather than independent discovery of the original upload, with reports appearing only after the *Presidential Communications Office (PCO)* issued its public warning, not after the initial upload from the now-removed *DAPAT BALITA* channel (Esguerra, 2024; Geducos, 2024; Locus, 2024a).

Manila Bulletin highlighted the Palace warning and underscored the government's Media and Information Literacy (MIL) campaign and inter-agency coordination described in the official response (Geducos, 2024). *GMA Integrated News* published an initial report on the PCO advisory and subsequently added explanatory framing that described the manipulated audio as false and attributed the clarification to PCO Assistant Secretary Dale de Vera (Locus, 2024a; Locus, 2024b). The government denial also appeared in *ABS-CBN* news coverage during the early hours of 24 April (ABS-CBN, 2024b).

State outlet *PNA* repeated the PCO warning and added definitional language explaining deepfakes as AI-made manipulated media while reiterating that no such presidential directive existed (Esguerra, 2024). *RADYO Inquirer 990AM* extended domestic reach by translating and broadcasting the warning in Filipino-language radio form (Escosio, 2024). *Philstar Global* similarly reported the Palace warning about the manipulated audio (Flores, 2024a).

In relation to the questions, this class provides the principal downstream comparison group for latency measurement once Class A has spoken: the timestamps on these national items constitute the first observable Class B artefacts against which their minimum-Maximum latency ranges are calculated in Chapter 4.6 using the established T(zero) window.

Regarding Q1, the pattern of Class B reliance on the PCO advisory and descriptive news framing rather than on visible in-product labels, warning messages, or platform-supplied spread indicators supports the OSINT-bounded finding that publicly observable early defence in this phase was driven by official statements and journalistic relay rather than by detectable, anomaly-detection or labelling interventions that would be visible to the users (Esguerra, 2024; Locus, 2024b; Geducos, 2024).

In response to Q3, these outlets' early descriptions of the clip's alleged content portraying a fabricated presidential voice ordering the Armed Forces of the Philippines and "special task groups" to act against another country in a South China Sea setting (Maitem, 2024; NDTV, 2024; Rappler, 2024b; Rappler, 2024f; VERAfiles, 2024; AIAAIC, 2024), reinforced why the message was treated as escalation-sensitive once the executive centre framed it as an AI-generated national-security concern.

Class C: Platforms & Cyber Agencies

Several reports identify the *YouTube* channel *DAPAT BALITA* as the uploader of the fabricated Marcos clip. *Philippine News Agency, INQUIRER.NET* and *Politiko* each name *DAPAT BALITA* as the source channel, even though the original URL and watch-page are no longer available (Caliwan, 2024; Argosino, 2024; Golez, 2024). *INQUIRER.NET*, reporting on 25 April, states that the *DAPAT BALITA YouTube* channel had been deactivated, that another *YouTube* account under the same name had reposted the video, and that a *Facebook* page called DAPAT BALITA was created on 19 April 2024 (Argosino, 2024). later the same day, the same news outlet records a legal-escalation response, noting that Justice Secretary Jesus Crispin Remulla ordered the *NBI* to conduct a swift and comprehensive investigation into the fabricated Marcos audio and pursue accountability for those behind the deceptive act (Torres Tupas, 2024). *Philippine News Agency* reports that the DAPAT BALITA *Facebook* page managers' primary countries or regions were listed as the

Philippines (Business World, 2024) and Pakistan (Murcia & Kabagani, 2024), based on *PNP-ACG* findings (Caliwan, 2024). *Politiko* later notes that the YouTube audio clip and similar material from other sources had already been removed (Golez, 2024). Taken together, these platform reports show that, by the time authorities reacted, the clip was linked to a specific branded uploader with cross-platform presence on YouTube and Facebook, and that some accounts had already been deactivated, recreated or had their content removed (Caliwan, 2024; Argosino, 2024; Golez, 2024).

Subsequent executive and cyber-agency communications shifted from simple warning to enforcement and attribution. On 26 April 2024, the *Presidential Communications Office (PCO)* announced that the deepfake video using President Marcos' face and voice had been taken down and that the government would pursue legal action against those responsible for creating and spreading it (Presidential Communications Office, 2024a). On 27 April 2024, the *PCO* stated that a "possible source" of the deepfake audio had been identified but that the extent of this actor's involvement remained under investigation, noting that the individual's role was still the "subject of our investigation" (Presidential Communications Office, 2024b). On 28 April, Politiko reported that the audio clip from YouTube, as well as similar material from other sources, had already been taken down (Golez,2024).

On 28 April, *Business World Online* reported that a potential source of the deepfake audio had been identified, but that the extent of their involvement remained under investigation by the Department of Information and Communications Technology and stated that the AI-generated audio of the President had been removed. In the same report, Deputy Speaker David Suarez suggested that the origin was likely from "somewhere in the southern part of the country" rather than from a foreign entity (Business World Online, 2024). This domestic-source framing sits alongside and partly qualifies earlier statements from the previous week that had raised the possibility of a foreign actor behind the spread of the deepfake audio (Flores, 2024b; Presidential Communications Office, 2024a).

On 30 April the *Philippine News Agency* reported that the Department of Information and Communications Technology and the National Security Council had worked to take down circulating videos in cooperation

with platforms such as *Google*, *TikTok, X* and Meta (including *Facebook* and *Instagram*) as part of the enforcement response (Cervantes,2024).

Cyber agencies and inter-agency mechanisms were repeatedly named as part of this response. Early reports note that the incident prompted investigations by the *Philippine National Police Anti-Cybercrime Group (PNP-ACG)* and *DICT,* and that these entities were coordinating with other agencies and social media platforms to address the deepfake (Medenilla, 2024; Presidential Communications Office, 2024, April 23; Presidential Communications Office, 2024a; Presidential Communications Office,2024b).

In response to Q1 and visibility limits, the platforms *(YouTube, Facebook* and other major services) played an instrumental role in the public-facing response as the main channels through which the *PCO* warning, news coverage and later removals became visible, with the initial advisory issued on *Facebook* and subsequent reports describing coordination with *Google, Meta, TikTok* and *X* to remove the content (Presidential Communications Office, 2024, April 23; Cervantes, 2024; Medenilla, 2024; Rappler, 2024b; ANC 24/7, 2024c).

However, across the *PCO* advisory, news reports and surviving platform traces examined for this case, there was no mention of any AI-based anomaly detection or automatic platform responses such as visible labels, warning banners or in-product markers, nor any public-facing propagation analytics on *Facebook, YouTube* or X that exposed how the clip spread over time in a way that could be audited as an anomaly-detection signal (Presidential Communications Office, 2024, April 23; Argosino, 2024; Rappler, 2024b).

The platforms appear in this record through deactivations and takedowns described by *PCO,* state agencies and news outlets, rather than through UI-visible anomaly-detection interventions.

Regarding Q2, the first defender actions came from the *Presidential Communications Office (PCO)* on 23 April 2024 at 20:31 PHT, with a public *Facebook* advisory (Presidential Communications Office, 2024, April 23). In terms of defender latency, the PCO's response occurred after a four-day maximum latency (96 hours, 31 minutes), calculated from the earliest estimated publication of the deepfake (19 April 2024, 20:00

PHT), and a minimum latency of 72 hours, 32 minutes from the upper bound of the T(zero) window on 20 April 2024, 19:59 PHT. Various national outlets such as *Manila Bulletin, GMA Integrated News* and *PNA* picked up the story and amplified the *PCO*'s message later that night and into the following morning (Geducos, 2024; Locus, 2024a; Esguerra, 2024).

Because internal logs and enforcement dashboards are not accessible, and because platforms do not expose per-video propagation metrics or anomaly-detection views for this clip to ordinary users, this study cannot calculate an independent Δ for platform-side anomaly detection; instead, Class C timing is inferred from when *PCO* and media reports (Presidential Communications Office, 2024a; Caliwan, 2024; ANC 24/7, 2024a, 2024b, 2024c), first state that the video has been taken down or channels deactivated (Rappler, 2024b; Flores, 2024a, 2024b) and these timings are compared with other classes in 5.3 section.

In response to Q3, the fabricated deepfake audio clip emerged at a highly sensitive moment in the geopolitical landscape of the South China Sea. It coincided with ongoing military exercises such as Balikatan 2024 (ABS-CBN. 2025a; U.S. Embassy Manila, 2024), and escalating tensions between the Philippines and China, heightening the potential for miscalculation (Rappler, 2024b; Chi & Esteban, 2024).

When the *PCO* and other agencies framed the incident within a national-security context, the perceived urgency of the response increased: the deepfake was treated not only as misinformation but also as a potential trigger for diplomatic or security consequences. The *Department of Information and Communications Technology (DICT)* and *National Security Council (NSC)* were actively involved in the investigation, emphasising that the deepfake had implications for diplomatic relations and national security (Presidential Communications Office, 2024a; Cervantes, 2024). This shows how the response was shaped not only by the need to counter false content but also by broader national-security and foreign-policy considerations tied to the Philippines' ongoing tensions with China.

Class D: Political Influencers

Class D is represented by actor-politician Jolo Revilla, who reshared the *Presidential Communications Office (PCO)* advisory about the Marcos deepfake on his *Facebook* page on 24 April 2024 at 2:34 Pacific time, corresponding to 17:34 PHT. His post relayed the same core claims as the original *PCO* statement that the circulating audio was a deepfake created with generative AI and that "no such directive exists nor has been made" instructing the Armed Forces of the Philippines to act against another country thus extending the official rebuttal into his influencer network on Facebook (Revilla, 2024). From a timing perspective (Q2), *Revilla*'s post appears several hours after the initial *PCO Facebook* advisory on 23 April and is therefore treated as a Class D artefact anchored to the same T_0 window reconstructed through the procedure set out in Section 3.3 and applied in the defender-latency calculations used in Chapter 5.3.

Using the established T_0 interval of 19 April 20:00 to 20 April 19:59 (PHT), his repost corresponds to an estimated maximum defender latency of 117 hours 34 minutes (from 19 April 20:00) and a minimum defender latency of 93 hours 34 minutes (from 20 April 19:59), placing Class D clearly after the executive centre and early national outlets in the comparative latency chart (Revilla, 2024).

In relation to visible anomaly detection (Q1), *Revilla'*s contribution does not introduce any additional platform-originated labels, warning banners, or automated deepfake markers beyond those already present in the PCO advisory; instead, it reproduces the government's framing through a personal/political account on Facebook (Revilla, 2024; Presidential Communications Office, 2024, April 23).

Under an OSINT restriction only lens, this confirms the previous findings that early defence still appears as human amplification of official statements rather than as an AI-based anomaly-detection or labelling pipeline. For escalation risk (Q3), the timing and content of this Class D intervention show that by the evening of 24 April 2024, the government's framing of the incident as a deepfake with potential consequences for foreign relations and national security had been taken up and circulated by domestic political figures, but still in terms that mirrored the Palace line rather than reframing it (Revilla, 2024; Presidential Communications Office, 2024, April 23).

He functions as a later political-influencer amplification layer. He leaves the executive-centre message unchanged but carries it into audiences reached through *Revilla's* combined roles as politician and public figure and remains dependent on prior Class A and Class B actions for both content and timing (Revilla, 2024).

These records show two competing attribution narratives: one emphasising possible foreign actor (Calonzo, 2024; Flores, 2024b; Piatos&Oliquino, 2024; Presidential Communications Office, 2024a) and regional information operations, and another implying a domestic origin "somewhere in the south of the country" (Business World Online, 2024) even as PNP- DICT continue to pursue accountability.

In response to Question 1 (visible anomaly signalling), Phase 3 demonstrates that what becomes observable at the UI level is the defenders enforcement: identification of the uploader cluster, references to deactivated channels and pages, and statements about completed removals and investigations (Caliwan, 2024; Argosino, 2024; Golez, 2024; Piatos & Oliquino, 2024). None of these sources describe automated, user-visible AI flags or platform-level warning labels on the original clip and "deepfake" appears only in government and news language, not as a UI badge. Regarding Question 2 (defender latency), the timestamps associated with *DAPAT BALITA* identification, deactivation/preservation requests, the DOJ- NBI directive, and the "possible source" announcements occur several days after both the T(zero) window and the 23 April *PCO* advisory that defines T(official) for Class A (Presidential Communications Office, 2024, April 23; Caliwan, 2024; Pulta, 2024; Piatos & Oliquino, 2024). Phase 3 therefore provides later upper-bound markers for cyber-agency and legal responses, without moving the earliest observable executive response point. Addressing escalation risk and context, Phase 3 shows that once provenance is partially reconstructed and takedowns are underway, the incident is treated explicitly as both a cybercrime and national-security problem: a branded uploader cluster is associated with the content (Caliwan, 2024; Argosino, 2024), deepfake manipulation is described as capable of harming foreign relations and national security (Cervantes, 2024), and attribution is debated between foreign and domestic sources while *PNP, DICT, NBI* and the legislature discuss prosecution and legal reforms (Pulta, 2024; Piatos & Oliquino, 2024).

5.3. Quantitative timing: T_0 window and defender latency (Δ)

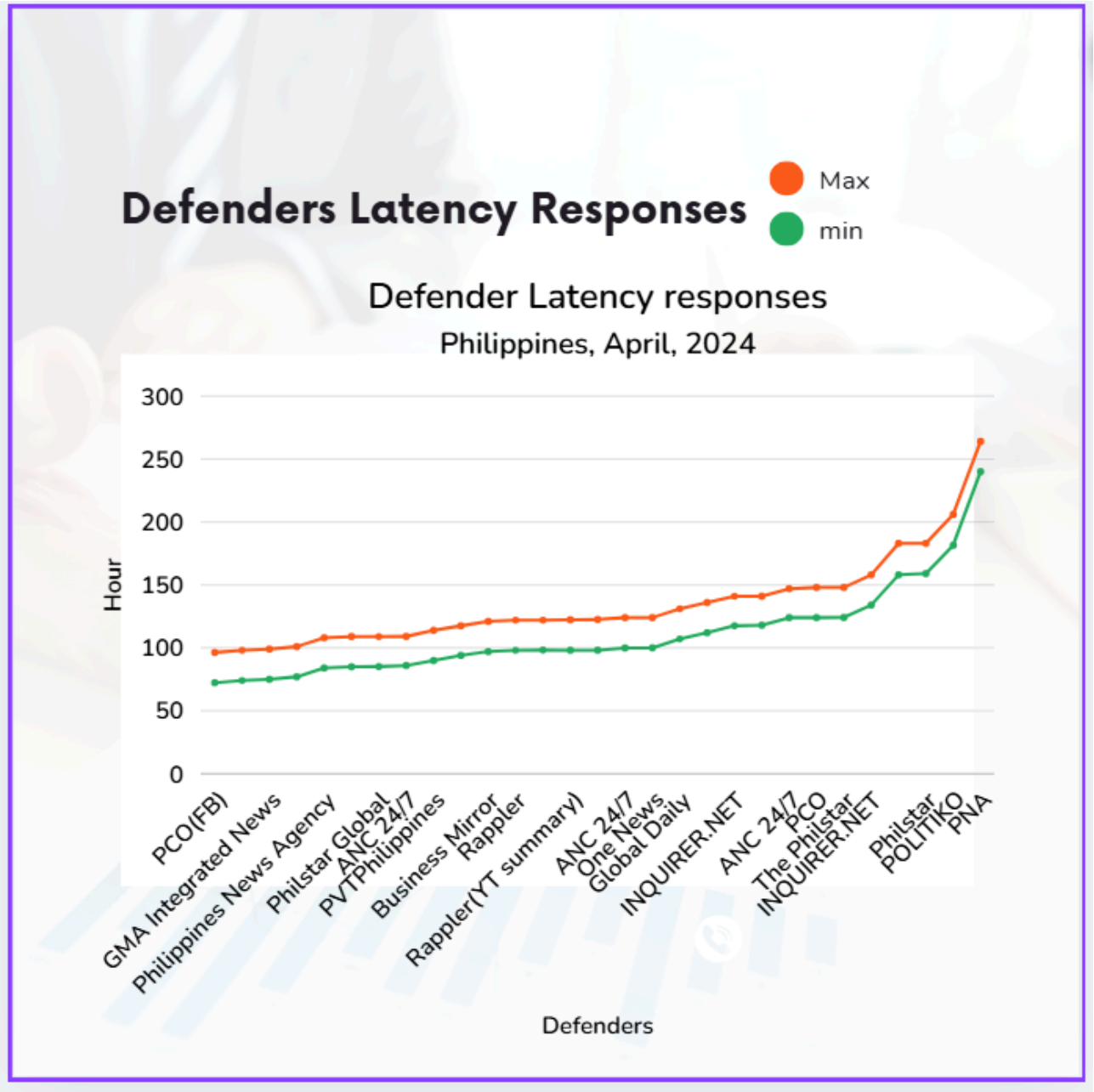


Figure 5.1. Line chart showing minimum and Maximum defender latency (Δ) for visible Philippine defenders. Figure created by the author using latency calculations derived from verified timestamps.

Note. The horizontal axis lists defender actors. The vertical axis shows elapsed time from the reconstructed T_0 window. The lower green line shows minimum latency calculated from the latest T_0; the upper orange line shows maximum latency calculated from the earliest T_0.

As it was confirmed by several scholars in the spoofing section, Voice-cloning has made it much easier for attackers to deceive the detecting system and public while increasing the effort for authorities to counter it.

The observed Δ ranges show how long a believable presidential voice persisted without challenge in a security sensitive period, which many consistently flagged as high-risk, regardless of whether it was real or fake. The first response is seen from the *PCO*.

Figure 5.1 allows the timing of first observable responses to be compared across defender categories. The first verified defender is the *Presidential Communications Office (PCO)*, where both minimum and maximum latency lines are far above zero, clearly highlighting the significant delay in their response. *Presidential Communications Office,* warning the public about the deepfake on Facebook, dating on 24th of April at 00:13 pacific time, which is equivalent to 23rd of April at 20:31 PHT) (PCO, 2024, April 23).

The lines for the news and broadcast outlets sit higher on the same vertical scale and cluster after the *PCO,* showing that their first traceable reports follow the *Palace advisory* rather than approaching the first publication T_0.

In other words, the chart makes two patterns visible at a glance: a substantial delay between T_0 and the first official response, and a sequence in which domestic outlets largely enter the record after the *PCO* and reinforce its framing, rather than independently detecting or challenging the fabricated clip closer to the time it first appeared.

Table 5.2 presents the chronological sequence of verified defender responses, entries 1-31, including the original discovered timestamps, converted Philippine Time (PHT), and the minimum and maximum latency estimates calculated from the bounded T_0 window. Table created by the author using verified timestamps and the latency calculation protocol described in Chapter 3.

Table 5.2 (a). Chronological sequence of verified defender responses, entries 1-12:

	Defender / outlet	Date & time (PHT)	Min latency (your value)	Max latency (your value)
1	Presidential Communications Office (PCO) - FB advisory	23 Apr 2024, 20:31	3 days / 72 hours 32 minutes	96 hours 31 minutes (4 days, 31 minutes)
2	Manila Bulletin (Geducos)	23 Apr 2024, 22:41	74 hours 42 minutes	98 hours 41 minutes
3	GMA Integrated News (Locus, 2024a)	23 Apr 2024, 23:39	75 hours 40 minutes	99 hours 39 minutes
4	ABS-CBN News	24 Apr 2024, 01:01	77 hours 2 minutes (3 days, 5 hours, 2 min)	101 hours 1 minute (4 days, 5 hours, 1 min)
5	Philippines News Agency (Esguerra)	24 Apr 2024, 08:33	3 days, 12 hours, 34 minutes	4 days, 12 hours, 33 minutes
6	RADYO Inquirer 990AM (Escosio)	24 Apr 2024, 09:03	85 hours 4 minutes	109 hours 3 minutes
7	Philstar Global - early piece (The Philippine Star, 2024)	24 Apr 2024, 09:17	85 hours 18 minutes	109 hours 17 minutes
8	ANC 24/7 - first YouTube report (ANC 24/7, 2024a)	24 Apr 2024, 09:46:04	85 hours 47 minutes 4 seconds	109 hours 46 minutes 4 seconds
9	PTV (PTV Philippines, 2024)	24 Apr 2024, 13:52:02	89 hours 53 minutes 2 seconds	113 hours 52 minutes 2 seconds
10	Jolo Revilla (Revilla, 2024)	24 Apr 2024, 17:34	93 hours 34 minutes	117 hours 34 minutes
11	Business Mirror	24 Apr 2024, 21:12:33	97 hours 13 minutes 33 seconds	121 hours 12 minutes 33 seconds
12	Rappler - website story with video (Rappler, 2024c)	24 Apr 2024, 22:05	98 hours 6 minutes	122 hours 5 minutes

Table 5.2(b). Chronological sequence of verified defender responses, entries 13-31:

13	Rappler YouTube Short (Rappler, 2024d)	24 Apr 2024, 22:05:01	98 hours 6 minutes 1 second	122 hours 5 minutes 1 second
14	Rappler YouTube summary clip (Rappler, 2024f)	24 Apr 2024, 22:07:57	98 hours 8 minutes 57 seconds	122 hours 7 minutes 57 seconds
15	Rappler - additional article (Rappler, 2024b)	24 Apr 2024, 22:11:19	98 hours 12 minutes 19 seconds	122 hours 11 minutes 19 seconds
16	ANC 24/7- second clip (ANC 24/7, 2024b)	24 Apr 2024, 23:50:55	99 hours 51 minutes 55 seconds	123 hours 50 minutes 55 seconds
17	OneNews (Flores, 2024)	25 Apr 2024, 00:00	100 hours 1 minute	124 hours
18	Global Daily Mirror - FB post	25 Apr 2024, 07:13	107 hours 14 minutes	131 hours 13 minutes
19	PNA (Caliwan, 2024)	25 Apr 2024, 12:10	4 days, 16 hours, 11 minutes (from 20 Apr 19:59)	5 days, 16 hours, 10 minutes (from 19 Apr 20:00)
20	INQUIRER.NET (Argosino, 2024)	25 Apr 2024, 17:23	4 days, 21 hours, 24 minutes	5 days, 21 hours, 23 minutes
21	GMA News (Locus, 2024b)	25 Apr 2024, 17:43	117 hours 44 minutes	141 hours 43 minutes
22	ANC Digital News (ANC 24/7)	25 Apr 2024, 23:54:32	123 hours 55 minutes 32 seconds	147 hours 54 minutes 32 seconds
23	Presidential Communications Office (PCO, 2024a)	26 Apr 2024 (time not specified)	**Min latency range:** 5 days, 4 minutes → 6 days, 4 hours	**Max latency range:** 6 days, 4 hours → 7 days, 3 hours, 59 minutes
24	The Philippine Star	26 Apr 2024, 00:00	124 hours 1 minute	148 hours
25	INQUIRER.NET (Mangaluz, 2024)	26 Apr 2024, 10:21:16	5 days, 14 hours, 21 minutes, 17 seconds	6 days, 14 hours, 21 minutes, 6 seconds
26	The Philippine Star	27 Apr 2024, 11:00	159 hours 1 minute	183 hours
27	Presidential Communications Office (PCO, 2024b)	27 Apr 2024 (time not specified)	**Min range:** 6 days, 4 minutes → 7 days, 4 hours	**Max range:** 7 days, 4 hours → 8 days, 3 hours, 59 minutes
28	POLITIKO (Golez, 2024)	28 Apr 2024, 09:08:05	7 days, 13 hours, 9 minutes, 5 seconds	8 days, 13 hours, 8 minutes, 5 seconds
29	PNA (Cervantes, 2024)	30 Apr 2024, 20:01	10 days, 0 hours, 2 minutes	11 days, 0 hours, 1 minute
30	GMA Integrated News	14 May 2024, 16:03	23 days, 20 hours, 4 minutes	24 days, 20 hours, 3 minutes
31	Philstar (Chi & Esteban, 2024)	8 Jun 2024 (time not specified)	47 days, 13 hours, 36 minutes	48 days, 13 hours, 35 minutes

Chapter 6: Analysis

6.1. Analysis of the enforcement patterns for T_0

YouTube (2018) reports a hybrid enforcement model (machine-learning detection plus human review) and documents rapid, early removals in Q3 2018-7.8 million videos removed, 81% machine detected, 74.5% with zero views at removal indicating action before widespread viewing.

The Google Transparency Report provides a contemporary baseline for enforcement scale; for Apr-Jun 2025, it reports 2,105,778 channels and 11,408,261 videos removed, with a large share detected via automated flagging and a substantial proportion removed at extremely low view counts, consistent with removal before significant viewing (Google, n.d.).

YouTube's mechanism description and Google's contemporary scale figures are cited only to show that abrupt step-changes are a known and routine outcome of platform enforcement or integrity corrections as the contextual evidence. They do not identify the exact cause in this instance. I have only used these signals to show that the change is real and sharp enough that an intervention such as response action would sensibly treat it as meaningful breakpoint, therefore it is reasonable to use the date of the bucket/bin which is the time window where the jump appears as a temporary reference date for when the original event occurred until the original watch page URL become accessible.

The first detectable platform disruption appears approximately 0-24 hours after Tzero or T_0, which is the earliest publication time, as indicated by the daily bin revisions on 20-21 April (PHT). Without access to the exact Youtube log, this is only an observational timing indicator rather than definitive evidence of the enforcement action. This opacity is itself a relevant finding, consistent with known delays and limited transparency in anomaly-detection pipelines. Because of this, measuring competent stakeholder's defender

latency as a publicly auditable KPI* becomes necessary for this study, as no other measurable enforcement signal is available.

On 22 April 2024 NZST ~ 21 Apr 20:00 -22 Apr 19:59 PHT, reported traffic and estimated earnings resume, suggesting the 21 April zero revenue state was temporary (such as policy/eligibility review or cleanup).

To identify the probable cause of the observed single-day drop, it is necessary to examine whether the decline resulted from bulk deletion, which may occur either through platform-initiated actions, such as the removal of content due to community guideline violations, or through creator-initiated actions, including the mass deletion of a video archive (Google, n.d.; The YouTube Team, 2018).

The drop could result from bulk deletion, which can be either platform-initiated (e.g., removal due to community guideline violations) or creator-initiated (e.g., mass deletion of a video archive). When violations are found, YouTube removes the video, applies a strike to the channel, and can terminate entire channels for egregious cases (The YouTube Team, 2018). When channels are terminated, all of their videos are removed, and for Apr-Jun 2025 this accounted for 55,015,741 video removals (Google, n.d.).

Secondly, privatisation of previously public videos by the channel owner also leads to public view count reductions, as these views are subtracted from visible channel metrics (Schwartz, 2013).

According to Castaldo et al. (2024) *YouTube* performs retroactive invalid traffic removals, this is how the platform adjusts view counts to eliminate non-genuine traffic such as views generated by bots, replay loops, incentivized clicks, or other forms of manipulated engagement. These patterns are robust to sampling frequency; a 5-minute dataset reproduces the same timing and popularity effects, ruling out artifacts from hourly sampling (Castaldo et al., 2024, p. 10).

*KPI or key performance indicator is a measurable indicator that can be tracked over time and in this case used to assess how well organizations respond during crisis communication events

This rules out a binning artefact. As Castaldo et al. (2024) explain these corrections are not applied in real time, they occur in batches with daily peaks around 4-5 p.m. and predominantly late in a video's lifecycle, with peaks observed on day 2 and a further rise on day 6 after publication. Practically, the platform executes enforcement in batch runs rather than as a steady stream and consequently, revisions often cluster on one calendar day. This empirical pattern indicates structured batch enforcement of invalid traffic "fake views" removals rather than continuous adjustment (Castaldo et al., 2024, p. 1-10).

Therefore, a one-day negative revision is consistent with batch-based invalid traffic enforcement rather than a gradual decay process. Batch application helps explain why discontinuities can appear confined to a single daily

Comment enforcement is similarly largescale -1,448,003,664 comments removed in April to Jun 2025, with 99.7% first detected by automated flagging reflecting ongoing, high volume integrity operations (Google, n.d.). At these volumes, public metrics can shift abruptly even when no (UI) user visible notice or label accompanies the change.

Enforcement volumes can also spike when policies are tightened, so removals and related public metrics may fluctuate by category and period (YouTube, 2018). These volumes indicate ongoing, highcapacity enforcement that can register as abrupt step-changes in public metrics. Considered together, the batch timing and the enforcement scale explain why an abrupt, single-day drop can appear in channel totals.

The zero-revenue state reflects a temporary monetisation hold or eligibility review, while the large drop in views may result from video removals or corrections for invalid traffic. Estimated earnings are modelled, meaning they are calculated using algorithms rather than actual payments, and even YouTube's official revenue figures remain provisional until month-end (Google, 2025). On 20 April 2024 (NZST), corresponding roughly to 19 April 20:00-20 April 19:59 (PHT), the video received remarkably high views (~1.24 million) but low estimated earnings (~$9-$27), a pattern consistent with limited ads, demonetisation, or pending ad review. These figures provide contextual insight but do not confirm policy status, monetisation, or exact upload timing

Considering these publicly high views signals with low estimated earning on 20 April (NZST, a negative Views Change with $0 - $0 estimated earnings on 21 April NZST, and a return to normal traffic and earnings on 22 April NZST, the 21 April NZST bin is the strongest anchor for the removed original. However if the video had first been made public on 21st(NZST), then th earliest bin(17- 20) would be expected to show around zero views, instead on 20th (NZST), (19th PHT 20:00- 20th April 19:59), shows very high views and on 21st of April (NZST), bin shows a large negative revision with zero earning.

Since I can't prove that the video was privately shared before 20th PHT, and made public on 21st(NZST), I cannot claim that 21st was the first day or T_0 of the publication of the fake clip (video clip with the fake audio), so the window of the appearance would be close to 19 April 20:00 to 20 April at 19:59(PHT).

6.2. Semantic triangulation analysis

From the vidIQ "Latest Published Videos" list, analysed from the third-party channel, the channel shows four recent uploads clustered on 20-21 April 2024 NZST (~19-21 Apr PHT). The titles below were translated using Google Translate for interpretive alignment with the reported allegation:

CHINA, NAGHULOG NG BOMBA! PBBM, NAGHUDYAT NA GUMANI! (Apr 21, 2024). Translation: "China dropped bombs! PBBM signalled retaliation!"

CHINA, TINIRA NG PILIPINAS! PBBM, NAG-UTOS NA GUMANTI SA CHINA (Apr 21, 2024). Translation: "China was hit/attacked by the Philippines! PBBM ordered retaliation against China."

ACTUAL VIDEO! CHINA, WALANG AWANG WINASAK ANG BARKO NG PILIPINAS!

(Apr 21, 2024).

Translation: Actual video! China mercilessly destroyed a Philippine ship!"

PILIPINAS, AATAKE NA SA CHINA! MAY GO-SIGNAL NA NI PBBM (Apr 20, 2024).

Translation: “The Philippines will attack China! PBBM has given the go-signal!”

Based on the list above, and by considering “Audio ng ‘attack order’ ni PBBM, ‘deepfake’ - Palasyo” published by *RADYO990AM* on 24 April (Escosio, 2024) and the *VERAFiles* report stating that the AI-generated reel was posted on TikTok as early as April 20 and that it attempted to make it appear the President had directed the *AFP* to “fight back” against China (VERAFiles, 2024), the second video is the leading candidate to be the one containing the fabricated audio because of the phrase (“…NAG-UTOS NA GUMANTI…”). The phrase “nag-utos” literally means “ordered,” which matches the allegation that the audio made it seem the President had ordered the AFP to act/retaliate (Escosio, 2024; VERAFiles, 2024). A second candidate could be video 4 (“…MAY GO-SIGNAL NA…”), because “go-signal” similarly implies authorisation to attack, aligning with the same order/authorisation narrative reported in these accounts (Escosio, 2024; VERAFiles).

Since *DAPAT BALITA* on YouTube was identified by several reputable news outlets as the reported origin of the fabricated clip featuring the fabricated audio of President *Marcos* (Argosino, 2024; Caliwan, 2024; Golez, 2024), the *YouTube* upload likely preceded or at minimum coincided with the earliest *TikTok* circulation noted on 20 April. Therefore, unless internal platform logs or a recoverable watch-page timestamp becomes available, the exact upload time cannot be confirmed. Because internal platform logs are not available and the original watch-page timestamp has not been recovered, pinpointing the exact upload time is not possible; however, triangulating this title-semantic alignment with *VERAFiles*’ earliest-post cue supports retaining a bounded publication window rather than a definitive time, which, when normalised to Philippine Time (PHT), remains 19 - 20 April 2024.

6.3. Defenders Latency Measurement shown as Δ, as an observable crisis-communication KPI

Given the limitations of AI anomaly-detection systems and the opacity of platform enforcement, defender latency became the primary observable metric for assessing the responsiveness of key stakeholders in this crisis. With no publicly visible platform labelling or takedown timestamps available and considering that platform-side anomaly detection was either ineffective or unavailable, measuring the interval between the first public availability of the deepfake and the first authoritative response provided a more accurate reflection of the crisis communication effectiveness. This approach was necessary due to the brittleness of platform monitoring systems, which are prone to latency, opacity, and adversarial manipulation, as outlined in the literature review. By focusing on latency (Δ), this study prioritises observable public responsiveness over internal platform actions that were either non-transparent or absent, making it a more reliable KPI for evaluating crisis communication performance during this geopolitical incident.

On April 23, 2024, the *Presidential Communications Office (PCO),* acting as the government communications office, issued its first warning to the public regarding a deepfake circulating on Facebook, falsely attributed to President Ferdinand Marcos Jr. The warning was officially released at 00:13 Pacific Time (PCT; converted to April 23, 2024, 20:31 Philippine Time, PHT) (PCO, 2024, April 23). For this initial official warning, the maximum latency was 96 hours and 31 minutes (from April 19, 20:00 PHT to April 23, 20:31 PHT), and the minimum latency was 72 hours and 32 minutes (from April 20, 19:59 PHT to April 23, 20:31 PHT). From a crisis communication perspective, this marked the first formal alert from the central authority, defining the content as a deepfake on Facebook and distinguishing it from the President's genuine communications.

This was amplified by the national media outlet on the same night at 10:41pm PHT, which is still within 24hours, by the *Manila Bulletin on 23rd*. They published a warning about deepfake incident including the PCO's warning (Geducos, 2024). This brought the story into the online news cycle, with a maximum latency of 98 hours and 41 minutes and a minimum latency of 74 hours and 42 minutes. Shortly afterwards, GMA

Integrated News, a television news network and digital news outlet, released its own report at 11:39 PM PHT, explicitly quoting the PCO warning about the fake video (Locus, 2024a).

For this broadcast, the maximum latency was 99 hours and 39 minutes (from April 19, 20:00 PHT to April 23, 11:39 PM PHT) and the minimum latency was 75 hours and 40 minutes (from April 20, 19:59 PHT to April 23, 11:39 PM PHT). Taken together, these national outlets began the amplification phase, relaying the PCO's language and framing to broad domestic audiences.

In the early hours of April 24, 2024, the incident entered a new news cycle, with more broadcasters and agencies reinforcing the PCO line. *ABS-CBN*, a major television network and online news outlet, posted a report at 01:01 AM PHT (ABS-CBN, 2024b). This report covered the deepfake incident, with a maximum latency of 101 hours and 1 minute (from April 19, 20:00 PHT) and a minimum latency of 77 hours and 2 minutes (from April 20, 19:59 PHT).

Later that morning, the *Philippine News Agency (PNA)*, as the state news agency, issued a report about the PCO warning at 8:33 AM PHT (Esguerra, 2024). This official wire-style report focused on the PCO warning about the deepfake and registered a maximum latency of 4 days, 12 hours, and 33 minutes and a minimum latency of 3 days, 12 hours, and 34 minutes, embedding the government message in state news distribution.

Radio and online outlets then carried the message into a broader audience. At 9:03 AM PHT on April 24, 2024, *RADYO Inquirer 990 AM*, a leading news radio station in the Philippines, published a report in the original language that included a reference to the PCO statement (Escosio, 2024).

This broadcast and online report had a maximum latency of 109 hours and 3 minutes and a minimum latency of 85 hours and 4 minutes, extending the official warning into radio audiences in the vernacular. At 9:17 AM PHT, *Philstar Global,* the online arm of *The Philippine Star,* followed with a similar report, explicitly referencing the Malacañang warning, stating that the authorities must combat the AI generated misinformation with full force of the law. This deepfake audio of the president should spur the government toward more resolute action against deepfake especially as we approach the coming 2025 election

(Philippine Star, 2024a). This *Philstar Global* piece logged a maximum latency of 109 hours and 17 minutes and a minimum latency of 85 hours and 18 minutes, reinforcing the Palace's position in a prominent online broadsheet.

Video-based news channels then translated the response into visual crisis coverage. At 9:46 AM PHT on April 24, 2024, *ANC 24/7,* the *ABS-CBN*-owned *YouTube* news channel, published its response to the incident (ANC 24/7, 2024a). The published time was derived from the Pacific Time Zone and converted to April 24, 2024, 9:46 AM PHT, with a maximum latency of 109 hours and 46 minutes and a minimum latency of 85 hours and 47 minutes. That afternoon, April 24, 2024, As the state broadcaster, *PTV Philippines* increased the visibility and authority of the response by interviewing *PCO* Assistant Secretary Dale de Vera on camera, signalling the seriousness and escalating tension around the incident.

They conducted an interview with *PCO* Assistant Secretary Dale de Vera conducted by PTV Philippines on YouTube at 1:52 PM PHT (PTV Philippines, 2024). As a state broadcaster, *PTV Philippines* used this interview to present a *PCO* official on camera during the unfolding incident, with a maximum latency of 113 hours and 52 minutes and a minimum latency of 89 hours, 53 minutes, and 2 seconds.

As part of the business coverage, Political figures and business-focused media then widened the perspective of the crisis. In the afternoon of April 24, 2024, *Jolo Revilla*, a Filipino politician and celebrity, shared the *PCO* announcement via *Twitter* at 2:34 PM PHT (Revilla, 2024). This political amplification from an individual actor, functioning similarly to a political news / commentary outlet, had a maximum latency of 117 hours and 34 minutes and a minimum latency of 93 hours and 34 minutes, signalling elite endorsement of the *PCO* warning.

Beijing's response on the same day on 7:19pm, 24 Apr 2024, could be described in a way that avoids direct backlash while still benefiting China from the narrative. The headline from the South China Morning Post (SCMP) went as follows:

“Audio deepfake” of Marcos Jnr ordering military action against China prompts Manila to debunk clip (Maitem, 2024).

By using the quotation marks around the audio deepfake, China created scepticism, prompting people even in the international scope to question it. By using quotes from authorities of Philippines such as *“no such directive exists”,* they were reminding the opponent and the other international allies of what has been said in a precise way to defuse the risk.

They used a tone that does not directly contradict Beijing’s strategic narratives to describe themselves as the peace seeker of the region. While reporting on the manipulated audio clip allegedly featuring President Marcos, the report references that he ordered *“military action against China”* despite the fact that on April 23, Philippines officials did not officially name any country in their denial. This way they amplified the tension, to escalate the perceived stakes and subtly frame the Philippines as a provocateur considering what had been depicted in the video.

Additionally, they used condescending language such as the deepfake clip causing *“serious concern”* among Philippines authorities regarding their foreign policy, implying that even if the clip was real, it would be Manila’s problem, not China's, and ultimately backfires on the Philippines. They also included some quotations such as phrases like *“Take action If China attacks “,* and *“They no longer want to be hurt by China", China* preserved their power while painting the Philippines as the weaker, overly reactive, or paranoid country.

Later that evening, at 9:12 PM PHT, *Business Mirro*r, a business newspaper and online news outlet, issued a report that discussed the Philippine government's response, specifically highlighting coordination with the *Department of Information and Communications Technology (DICT)* and the *National Cybersecurity Inter-Agency Committee (NCIAC)* (Medenilla, 2024). This business-focused coverage documented coordination between the central government, *DICT*, and *NCIAC* as part of the official response, with a maximum latency of 121 hours, 12 minutes, and 33 seconds and a minimum latency of 97 hours, 13 minutes,

and 33 seconds, shifting the emphasis from simple denials to institutional coordination in cyber and information security.

Independent digital media then layered in multimedia formats. On April 24, 2024, *Rappler* published its coverage of the deepfake incident, pairing an article with a video at 10:05 PM PHT (Rappler, 2024c). This combined article & video publication recorded a maximum latency of 122 hours and 5 minutes and a minimum latency of 98 hours and 6 minutes. Rappler also issued two *YouTube* Shorts on the same date. The first Short was published at 7:05 PM Pacific Daylight Time (PDT; converted to April 24, 2024, 10:05 PM PHT) (Rappler, 2024d), and the second at 7:07 PM PDT (Rappler, 2024f), with conversion to Philippine Time applied in the timing. For these Shorts, the maximum latency was 122 hours, 5 minutes, and 1 second, and the minimum latency was 98 hours, 6 minutes, and 1 second.

By April 25, 2024, some outlets were explicitly attributing motives. At 12:00 AM PHT, *OneNews*, a news channel and digital outlet, reported that the deepfake video was part of a broader effort to discredit the President (Flores, 2024a). This report directly stated that the deepfake was part of a broader effort to discredit President *Marcos Jr.*, with a maximum latency of 124 hours and a minimum latency of 100 hours and 1 minute, marking a move from neutral reporting to intent-focused framing. Later that morning, at 7:13 AM PHT, the *Global Daily Mirror*, a news outlet and online newspaper, published its coverage as a response to the incident (Global Daily Mirror, 2024). This added another general news account, with a maximum latency of 131 hours and 13 minutes and a minimum latency of 107 hours and 14 minutes, keeping the incident visible in the morning news agenda.

On the same day, operational details of the enforcement response became clearer. At 12:10 PM PHT on April 25, 2024, the *Philippine News Agency (PNA)* published a detailed account of the Anti-Cybercrime Group (ACG) investigation, describing how the *ACG* continued tracking the source of the deepfake (Caliwan, 2024). As the state news agency, this detailed update emphasised that the *ACG* was actively working to identify the source, with a maximum latency of 5 days, 16 hours, and 10 minutes and a minimum latency of 4 days, 16 hours, and 11 minutes, signalling a concrete investigation phase within the crisis response. At 5:23

PM PHT the same day, *INQUIRER.NET* (Philippine Inquirer's online platform) reported that the *DAPAT BALITA YouTube* account, which initially spread the video of fabricated audio, had been deactivated, and that another account with the same name had reposted the fake video (Argosino, 2024). The report also noted that a *Facebook* account under the name *DAPAT BALITA* had been created on April 19, 2024. This coverage was documented with a maximum latency of 5 days, 21 hours, and 23 minutes and a minimum latency of 4 days, 21 hours, and 24 minutes, illustrating the persistence and migration of the deepfake across channels and platforms.

Another Chinese news outlet, questioned whether President Marcos had truly ordered an attack on China or not, whilst referencing the official denial issued by the *Presidential Communications Office* with the from Dale De Vera, Assistant Secretary for the *PCO*, who noted that while this was not the first deepfake involving Marcos, this case was more serious due to its implications on their foreign policy. The report also mentioned that the video was originally uploaded to a *YouTube* channel, which had thousands of subscribers as were mentioned by the Indian news outlet (Sina News, 2024). This was reported on 25th of April and translated by Google.

That evening, attention turned explicitly to the nature of the audio itself. At 5:43 PM PHT on April 25, 2024, *GMA News* published a follow-up piece explaining deepfake audios (Locus, 2024b). This report included a quotation from *PCO* Assistant Secretary *Dale de Vera*, who clarified that the audio used in the deepfake video was fabricated and emphasised that the video did not represent the President's actual statements. In crisis-management terms, this was a crucial clarification: it explicitly stated that the audio track was fabricated audio, not just edited or taken out of context. The maximum latency was 141 hours and 43 minutes and the minimum latency was 117 hours and 44 minutes.

Later that night, at 11:54 PM PHT on April 25, 2024, ANC 24/7 / ANC Digital News, operating under *ABS-CBN*, reported further on the deepfake incident (ANC 24/7, 2024b). This report highlighted screenshots of the fake video and mentioned ongoing cooperation between the authorities and major social media platforms, including *Google and Facebook,* in removing the video and investigating the source. For this

coverage, the maximum latency was 147 hours and 54 minutes and the minimum latency was 123 hours, 55 minutes, and 32 seconds, documenting a phase in which the response relied on joint action between national authorities and major platforms such as *Google and Facebook.*

On April 26, 2024, the response moved into a more formal post-incident crisis management phase. The *Presidential Communications Office (PCO)* issued a post-incident crisis management statement, underscoring key elements of post-incident crisis management and stressing that legal actions would be pursued against those responsible for creating and distributing the deepfake video. Although the exact publication time was not provided even after the systematic investigation realising that it is deliberating not included, the maximum latency was calculated as a bound 148 hours (6 days and 4 hours) and the minimum latency as 124 hours and 1 minute, marking a clear shift from warnings and explanations to a legal enforcement posture.

On the same day, *Philippine Star*, a national news outlet, published an editorial article titled “The Deepfake Menace" which focused on the government’s push for stronger law enforcement, particularly as the 2025 election approached (Philippine Star, 2024).The maximum latency was 148 hours and the minimum latency was 124 hours and 1 minute, linking the incident to law enforcement debates and electoral risk in the run-up to the 2025 election.

The investigation narrative advanced the next day. On April 27, 2024, the *Philippine News Agency (PNA)*, as the state news agency, reported that the *Philippine National Police (PNP)* had identified the possible source of the deepfake audio (Presidential Communications Office, 2024b). This state agency report marked a significant investigative step, with a maximum latency of 7 days, 4 hours and a minimum latency of 6 days, 4 minutes, indicating that the *PNP* had moved from general investigation to identifying a possible source of the deepfake audio. The following day, the story took on a more international tone. *Politiko,* a political news /commentary outlet, reported that foreign entities might be behind the creation of the fabricated audio (Golez, 2024). Published on April 28, 2024, this piece carried a maximum latency of 8 days, 13 hours, 8 minutes, and 5 seconds and a minimum latency of 7 days, 13 hours, 9 minutes, and 5 seconds, introducing an

explicit foreign-actor dimension by stating that foreign entities might be behind the creation of the fabricated audio.

By the end of April, the response had developed into legislative and complaint-driven responses. On April 30, 2024, the *Philippine News Agency (PNA)* published a report titled *"Lawmaker Suggests Classifying Use of Deepfakes as Terrorism*" at 8:01 PM PHT. Rep. *Mohamad Khalid Dimaporo* suggested to consider the deepfake contents classified as terrorism activity. "If there's a need to amend our laws and declare this type of behavior as a form of terrorism, well then we will refer to our leader here in Congress, our Speaker Martin Romualdez". Rep. Rodge Gutierrez stated that this as a national security matter that could become an international security concern (Cervantes, 2024). This response to the incident was with a maximum latency of 11 days and 1 minute and a minimum latency of 10 days, and 2 minutes.

On May 14, 2024, *GMA Integrated News* released further coverage, highlighting that broadcasters and the public were working with the government to file complaints about channels spreading deepfake content (Rita, 2024). This coverage showed broadcasters and public filing complaints as a practical extension of the legislative and complaint-driven responses, with a maximum latency of 24 days, 20 hours, and 3 minutes and a minimum latency of 23 days, 20 hours, and 4 minutes, documenting coordinated complaint activity against channels disseminating deepfake content.

On June 8, 2024, *Philstar* published a report by Chi and Esteban that discussed the deeper geopolitical implications of the deepfake, specifically stating that China may be using this as part of its broader strategy to discredit the Philippines (Chi & Esteban, 2024). This report focused on geopolitical implications and China, situating the deepfake incident within a broader strategy to discredit the Philippines, with a maximum latency of 48 days, 13 hours, and 35 minutes and a minimum latency of 47 days, 13 hours, and 36 minutes, and marking the transition from immediate crisis response to strategic geopolitical interpretation.

6.4. Synthesis to research questions

The preceding chapters set out the empirical findings by phase (pre-crisis priming, trigger and initial recognition, early media amplification, provenance and takedowns) and then analysed enforcement patterns, semantic triangulation of the likely upload, and defender latency (Δ) against the bounded T_0 window. This section does not add new evidence; instead, it organises the results around the three research questions. It first draws together what can be seen, from an Open-source intelligent perspective, about the visible limits of AI-based anomaly detection during the Marcos deepfake incident (Referring to Q1). It then synthesises how response timing varies by defender class once T_0 is bounded and Δ is calculated (Referring to Q2).

Finally, it integrates the pre-crisis signals and contextual patterns to show how they shaped escalation risk and indicate when rapid detection and response were most needed in this case (Referring to Q3).

In the case of President *Ferdinand "Bongbong" Marcos Jr.'s* fabricated voice crisis in the Philippines, publicly visible evidence formed two tiers of advance warning: a weeks-scale priming backdrop (alliance posture, legal/jurisdictional dispute salience, "foreign influence" framing, domestic contention) and a ten-day high-intensity pre-publication of the fabricated audio clip cluster (11-22 April) of diplomatic signalling, exercise expansion, and drill kick-off.

By systematically assembling and time-aligning these open sources, the analysis shows that a predictably sensitive window was created in which a leader-voice fabrication was more likely to be believed and to spread, justifying elevated anomaly triage before circulation. In this case, a fabricated voice message of the Philippines president coincided with a centralised command context, a legally contested setting, and a tight pre-circulation window of diplomacy and exercises that amplified the existing tensions and risk, pushing the incident from routine misinformation into a national-security matter. Authorities then began to treat the incident as a security and crisis-management problem rather than ordinary misinformation. Findings are case bounded to the Philippines April 2024 incident only.

The escalation from routine misinformation to a national security matter occurred when a fabricated voice of the apex leader appeared to issue or justify sovereign military action, intersecting a centralised chain of command with Commander in Chief authority and AFP operational centralisation, thereby heightening operational consequences.

This took place in a legally contested maritime theatre, with dash-line claims overlapping the Philippine EEZ (Permanent Court of Arbitration, 2016) and PCA rights and coincided with a compressed pre-circulation window for diplomacy as well as the kick-off of Balikatan expansion drills (U.S. Embassy Manila, 2024), raising treaty and ally stakes.

Immediate, close-quarters maritime friction further heightened tensions, including incidents involving water cannon use, obstruction near Scarborough/Panatag Shoal on April 30, and intentional ramming. It was reported that China obstructed a legitimate maritime patrol conducted by the Philippine Coast Guard (PCG) and the Bureau of Fisheries and Aquatic Resources (BFAR) near Bajo de Masinloc through dangerous manoeuvres by China Coast Guard and maritime militia vessels. China also installed a 380-metre floating barrier that restricted access to a traditional fishing ground. (cupin, 2024). The water-cannon and ramming actions damaged both the PCG's BRP Bagacay and BFAR's BRP Bankaw: Bagacay sustained damage to its railings, LED display and rear canopy, while Bankaw sustained damage to its HVAC, electrical, navigation and radio systems, as well as superficial hull damage (Daily motion, 2024; GMA Integrated News 2024; Nepomuceno, 2024; Reuters, 2024b).

By mapping these events through OSINT sources rather than classified feeds, the study shows how routine tension, exercise signalling and live incidents combined to make the information environment unusually escalation prone. These conditions were reinforced by threat signalling and provocation frames as well as narratives of foreign influence and culminated in overt security activation through foreign-actor designations, cybercrime investigations, and terrorism framing, formally shifting the incident into the crisis management and security management scope. In combination, these factors provide a practical framework for anticipating comparable cases and their escalation to national-security handling.

Under OSINT-only constraints, this study reconstructed the public responses timeline of the April 2024 Marcos deepfake incident to measure the earliest public availability of the fabricated clip from the removed *YouTube* channel called *DAPAT BALITA,* which no longer exists. By triangulating third-party channel analytics, cross-platform traces and media descriptions of the clip's content, the analysis bounded the first publication date (T_0) to be in the window of 19 April 20:00 to 20 April 19:59 (PHT).

Against this point, the *Presidential Communications Office (PCO)* posted the first authoritative response on 23 April 2024, 20:31 PHT on the Facebook platform, establishing the earliest measurable point of official action. Multi-day gaps between the bounded T_0 window and this first advisory suggest delayed responses, highlighting the operational limits of existing anomaly-detection mechanisms when internal logs are unavailable. Under OSINT-only observation, no AI-based anomaly-detection labels, warning banners or "manipulated media" markers appeared at the user-interface level in this case; all early defence was visible only through human and institutional responses rather than any auditable AI signal. Defender latency or (Δ) varied by defender class.

The executive centre responded first after 3-4 days, followed within hours by major domestic outlets including *Manila Bulletin, GMA* and *ABS-CBN, then PNA* as state media on the morning of 24 April, and later platform-hosted broadcast channels across 24-25 April. Later-appearing defender classes, including platforms and cyber agencies as well as political and legislative actors, only became visible several days after this initial wave through reports of investigation, deactivation, cross-platform takedowns and terrorism-framing, confirming substantially longer Δ ranges for these groups than for the executive centre and early national media.

In this high-risk context, a 3-4 day lag before the first authoritative warning shows that a convincing leader-voice deepfake was able to circulate for several days before any public rebuttal appeared at the surface.

Chapter 7: Conclusion

This dissertation set out to examine, what the April 2024 Marcos voice-cloning crisis reveals about the visible limits of AI-based anomaly detection, the timing of responses by key defender classes, and the pre-crisis signals that showed when rapid detection and response were most needed (under open-source intelligence constraints). Working within a qualitatively driven, OSINT-based mixed-methods single-case design, it reconstructed the crisis from public documentary evidence and integrated quantitative measures the first-publication window (T_0) and defender latency (Δ) to analyse anomaly-detection visibility, response timing, and escalation context.

Regarding Question 1, How do AI anomaly-detection systems fail, in practice, during the Marcos deepfake incident, under OSINT constraints only? the study does not infer the internal workings of platform-side AI; it defines "failure" strictly in terms of what appears, or fails to appear, at the public surface. Across the artefacts examined, no AI-based anomaly-detection labels, warning banners or "manipulated media" markers were visible on the fabricated audio at the user-interface level, and no propagation analytics or anomaly signals were exposed that would allow external observers to see whether the clip had been automatically flagged. Removal and deactivation only become visible indirectly, when official statements and news reports later describe the video as taken down or channels as deactivated.

Set against the technical literature on anomaly detection which documents high false-positive rates, class imbalance, context-blind flagging, and information loss near decision boundaries, this absence of any surface signal can be interpreted in two OSINT-consistent ways. One possibility is that the fabricated clip never crossed internal anomaly thresholds and was not treated as a high-priority incident. Another is that it did trigger some form of anomalous score but entered an overloaded or noise-heavy queue, where large volumes of false alarms, rigid baselines and lack of temporal/contextual sensitivity delayed escalation to visible enforcement.

From an OSINT perspective, the only demonstrable failure is therefore that a convincing leader-voice deepfake circulated for several days in a high-risk context with no visible automated warning, no exposed propagation-level detection trace, and no UI-level markers that external observers or ordinary users could see. Any internal anomaly-detection processes that may have fired left no surface-level signal before or during the period in which the clip remained uncontested.

In response to Question 2, what is the defender latency from first public availability to first authoritative response and how does Δ vary by defender class? The study bounded, and then compared, response times across stakeholders. Using third-party *YouTube* channel analytics, cross-platform traces and media descriptions of the clip's content, it bounded the first-publication window (T_0) for the now-removed *DAPAT BALITA* upload to 19 April 20:00 - 20 April 19:59 (PHT), rather than a single timestamp. Against this anchor, the *Presidential Communications Office (PCO) Facebook advisory* on 23 April 2024, 20:31 PHT is treated as the first authoritative response, giving a defender latency of approximately three to four days for the executive centre.

Within hours of that advisory, a major domestic outlet amplified the denial and warning, followed by state media and then platform-hosted broadcast channels across 24- 25 April. Platforms and cyber agencies, and political and legislative actors, only become visible several days later through reports of investigation, deactivation, cross-platform takedowns and terrorism-framing, indicating substantially longer Δ ranges than for the executive centre and early national media and showing that the most formalised security and legislative framings emerged only after the immediate crisis window had passed.

In practical terms, the timing structure that emerges is one in which the first visible response comes from the executive centre after a (3-4) day gap, early national media act as rapid amplifiers within hours, and platform and cyber/legislative responses lag by several further days.

Regarding Q3, how did pre-crisis signals and contextual conditions shape the escalation risk of the Marcos voice-cloning incident, and what do they indicate about the need for rapid detection and authoritative response? The analysis identified two layers of advance warning rather than a sudden, context-free shock.

At the broader level, a weeks-scale priming backdrop combined alliance signalling, legal and jurisdictional dispute salience in the South China Sea, intensified "foreign influence" narratives and domestic contention. Within that environment, an 11-22 April cluster of summit meetings, diplomatic signalling, *Balikatan* exercise announcements and drill kick-off created a compressed, high-intensity window immediately before the deepfake's circulation. By systematically assembling and time-aligning these open sources, the dissertation shows that this combination of legal contestation, alliance commitments and live exercises produced a predictably sensitive information environment in which a leader-voice fabrication was more likely to be believed and to spread. In this setting, a fake presidential order intersecting a centralised chain of command and a contested maritime theatre pushes the incident from routine misinformation into a national-security matter. These contextual patterns therefore indicate when anomaly-detection and crisis-response systems most urgently needed to act during a pre-circulation and early-circulation window in which escalation risk was already elevated.

These findings show that in a highly sensitive national-security context, a leader-voice deepfake was able to circulate for several days before any authoritative rebuttal appeared in public, and that when responses did come they were led by the executive centre and domestic media rather than by visible AI-based anomaly detection or platform labelling. The dissertation does not determine whether internal anomaly-detection systems fired and stalled inside opaque escalation chains or never fired at all; it shows that, from an OSINT-only perspective, if such systems were operating they produced no surface-level signals that external observers could see, and that crisis-management performance at the public layer was characterised by a multi-day delay during a period when the information environment was unusually escalation-prone.

Within these limits, the study contributes three main elements. Empirically, it provides a phase-based reconstruction of a single high-stakes deepfake incident after the original upload and channel had disappeared, showing how contextual risk, trigger timing and defender latency interact in practice. Methodologically, it demonstrates how OSINT-based reconstruction, temporal bounding of T(zero) or T_0 and Δ by defender class can be used to generate structured, auditable evidence about deepfake crisis handling when internal data are unavailable. Conceptually, it connects anomaly-detection debates in AI and

cybersecurity with crisis and issue management by treating timing and surface visibility as key indicators of how well, or how poorly, emerging AI-supported detection and crisis-communication practices function in a real national-security crisis.

In doing so, it complements technical work on anomaly detection and deepfake detection which typically evaluates model accuracy, false positives, latency and context blindness on benchmark datasets or within cybersecurity pipelines by showing how those documented limitations translate (or fail to translate) into a publicly observable response timeline in one leader-voice crisis. In particular, whereas prior studies warn that AI-supported monitoring can be slow, brittle or error-prone in principle, very few map those concerns onto an OSINT-auditable sequence of pre-crisis signals, first publication (T_0) and defender latency (Δ) in an actual voice-cloned national-security incident. This case study therefore helps bridge the gap between technical detection research and crisis-management practice, by operationalising defender latency and surface-visible signals as practical indicators of how anomaly-detection limits matter when a fabricated leader's voice enters a sensitive geopolitical environment.

Chapter 8: Recommendations

This chapter provides concise recommendations for defenders addressing the limitations of AI-driven anomaly detection, focusing on preventive platform governance, limits on the use of anomaly-detection outputs, contextual safeguards, crisis-readiness reporting, exposed-voice protection and implementation capacity

This dissertation first recommends that platforms recognise synthetic or impersonated voice as a distinct risk-management category rather than a subset of manipulated media. Audio-spoofing research identifies attacks involving replay, text-to-speech synthesis, voice conversion and presentation attacks (Gupta et al., 2024; Khanjani et al., 2023; Tan et al., 2021). Detection remains constrained by dataset dependence and limited generalisation to unfamiliar synthesis systems and attack conditions (Gupta et al., 2024; Li et al., 2025; Tan et al., 2021). To address these limitations, this dissertation proposes a recognisable-human-voice risk tier for content meeting three conditions: the voice is recognisable or represented as belonging to a real person; consent, attribution or lawful use is uncertain; and the content is monetised or situated in a high-consequence context, including political communication, crisis response, public safety, reputational harm or apparent official instruction.

The tier should operate at both upload and monetisation stages. Review should assess consent, attribution, lawful use and monetisation eligibility where voice is the principal evidentiary or persuasive element. Enhanced scrutiny should focus on audio-only recordings, static-image audio videos, slideshow-style content and material in which a voice has been cloned, replayed, converted, synthetically generated or materially edited to create new speech, music, political messaging, reputational claims or apparent official instructions. More intensive review should be reserved for recognisable-voice content combining uncertain provenance with high-consequence use.

Platforms and institutional defenders should adopt a voice-specific escalation protocol using proportionate friction, including reduced amplification, restricted monetisation, warning labels or specialist referral, while preserving disputed audio and associated platform evidence. Automated Automated detection and authoritative claims or denials should be assessed alongside technical indicators, contextual evidence and provenance information, but should not independently determine authenticity. Immediate removal may nevertheless be required where mandated by law, authorised through appropriate legal or institutional procedures, or necessary to prevent serious and imminent harm.

Where legally and technically feasible, defenders should preserve the original upload, source-account information, repost and amplification networks, interface labels, engagement metrics, monetisation status, warnings, takedown notices, corrections and archival actions. In political crisis or national-security cases,

case studies and open source investigations should commence during the incident or as soon as legally and ethically practicable because platform records may later be modified, removed or rendered inaccessible. Where safe and practicable, the represented person should be notified and allowed to hear, challenge and contextualise the disputed content. Their response should inform, but not determine, the evidentiary assessment, and review arrangements should guard against conflicts of interest arising from political, reputational, employment, regulatory or other interests in the outcome.

The central recommendation is that AI-driven anomaly outputs operate as limited-use triage signals rather than as automatic evidence. Existing scholarship supports explainability, adaptive behavioural baselines, hybrid or multimodal analysis, forensic correlation, confidence-based reporting, accountability and human oversight (Alzaabi & Mehmood, 2024; Pandey et al., 2025; Patel et al., 2023; Visave, 2024).These measures assist interpretation and review but do not define which consequential decisions an uncertain alert may independently support.

This dissertation therefore proposes a limited-use rule under which anomaly outputs initiate technical and contextual review rather than independently establish authenticity, manipulation, malicious intent, misconduct or threat. The rule applies wherever an alert may influence content restriction, account action, authoritative public denial or correction, discipline, exclusion, reputational judgement, regulatory escalation or crisis intervention. Each alert should identify the exact triggering action, event, data point or sequence; the alleged risk category; its confidence level and principal uncertainties; and the decisions it is authorised to support. This extends category-level explanation into action-specific traceability: labels such as “suspicious” or “anomalous” are insufficient unless reviewers can assess the underlying conduct, context and evidentiary weight.

Weakly substantiated, unexplained or low-confidence alerts should not independently trigger content removal, account restriction or suspension, authoritative public denial or correction, discipline, exclusion, reputational judgement, regulatory escalation or crisis intervention. Higher-consequence decisions should require stronger explanation, independent corroboration and appropriate human review. Multiple alerts should not be treated as independent corroboration where they reproduce the same incomplete data, outdated assumptions, context-insensitive baseline or model bias.

Organisational and insider-threat systems should distinguish behavioural deviation from malicious intent. Missing context concerning role, task, project changes, time, location, device, workload, access rights or evolving work practices can lead to ordinary conduct being misclassified. Behavioural drift causes historical baselines to flag newly normal activity or miss emerging threats, while biased or unrepresentative data may produce unequal or inaccurate predictions (Alzaabi & Mehmood, 2024; Pandey et al., 2025; Visave, 2024).

A functionally independent contextual review should occur before an uncertain alert is used by a decision-maker with disciplinary, regulatory or reputational authority. Reviewers should assess whether the

conduct reflects authorised activity, changing responsibilities, incomplete data, outdated baselines, shared-account activity, technical compromise or another plausible explanation. Review should also distinguish a person who presents a risk from a person who may be at risk through coercion, victimisation, harassment, retaliation, insider targeting or institutional conflict. Where these possibilities remain plausible, protective measures and independent investigation should precede punitive consequences, except where proportionate temporary action is necessary to address serious and imminent harm.

Unverified classifications should remain time-limited, separate from confirmed findings and subject to reassessment. Affected individuals should be able to challenge and contextualise alerts without bearing the burden of disproving misconduct solely because an uncertain classification exists. Where contextual information remains insufficient, the classification should remain unresolved and should not independently support adverse action.

Controlled benchmark performance should not be treated as evidence that a detector is suitable for time-sensitive crisis use. Operational usefulness depends on real-time processing constraints (Rahman et al., 2021), performance under lower-quality social-media conditions (Patel et al., 2023), generalisation across replay, voice-conversion and text-to-speech attacks and across channel and environmental variation (Tan et al., 2021), and adaptation to newer synthesis systems without overfitting or catastrophic forgetting (Li et al., 2025).

This dissertation therefore proposes crisis-readiness reporting that brings technical performance, end-to-end response latency, human-review capacity, escalation, public communication and authorised decision use into the same operational assessment. Reports should disclose latency across data intake and pre-processing, model scoring, human review, escalation, authoritative verification, platform action and public communication. Model-inference speed alone is insufficient where delay arises from evidence collection, expert review, institutional coordination or preparation of an authoritative response.

Reports should state whether testing covered replay, voice conversion, text-to-speech synthesis, compressed or reposted audio, background noise, unfamiliar devices, over-the-air transmission, microphone injection, unseen datasets and synthesis systems released after training. Cross-dataset and unseen-generator performance should be reported separately from results obtained on data resembling the training set, and unavailable testing should be disclosed rather than absorbed into an aggregate score. Each result should state its confidence level, unresolved limitations, false-positive and false-negative risks, further review required and authorised decision use. Reports should distinguish model failure from wider response-system failure, including missing evidence, inadequate preservation, delayed review, unclear authority, undefined escalation, excessive alert volume or insufficient forensic and human-review capacity.

An exposed or plausibly imitated voice should be treated as both a continuing security condition and potentially disputed evidence because biometric characteristics cannot readily be reissued or changed after

compromise in the same manner as passwords (Finklea et al., 2015). Defenders should minimise unnecessary recording, transcription, retention, access and redistribution of identifiable voice material. Sensitive meetings and recorded communications should use controlled participation, restricted recording and transcription, role-based access, auditable handling and documented chain-of-custody procedures. High-consequence instructions should be verified through an independent non-voice credential, communication channel or decision process. Reposting, engagement or repetition may demonstrate circulation but not authenticity, authorisation or correct attribution; disputed audio should therefore remain preserved and classified as disputed until sufficiently corroborated.

Regulators should clarify responsibilities concerning consent, attribution, biometric voice data, platform-metadata preservation, monetisation, accessible escalation and remedies for affected persons. Such clarification could reduce the burden on affected persons to disprove authenticity only after harmful content has circulated. Verification and labelling on social media are time-consuming and labour-intensive (Guo et al., 2020). Defenders should therefore provide dedicated operational funding for trained human review, forensic-audio support, evidence preservation, crisis-communication liaison, affected-person-accessible escalation pathways and public-interest review capacity.

Training should be role-specific. Technical and forensic personnel should be trained to interpret confidence levels, false positives and false negatives, dataset- and attack-specific limitations, and the distinction between an automated indicator and verified evidence. Managers and other consequential decision-makers should be trained to understand the unreliability and contextual limits of anomaly flags, including the fact that ordinary changes in devices, locations, working hours, assigned roles, responsibilities, tasks or time spent on particular files or documents may trigger alerts. Training should include practical examples showing how legitimate activity, changing workloads, shared accounts, new projects or altered access patterns can be misclassified as suspicious. Managers should therefore be able to assess an alert in relation to the person's role, position, authorised duties and current organisational circumstances, and to recognise when corroboration and independent contextual review are required before disciplinary, reputational, regulatory or crisis action. Crisis communicators and public-information personnel should be trained to distinguish circulation from authenticity and to explain both AI capabilities and limitations without presenting unresolved outputs as conclusive findings. These measures are consistent with existing recommendations for explainability, accountability and human oversight (Alzaabi & Mehmood, 2024; Patel et al., 2023; Visave, 2024).

This dissertation extends existing support for human-AI collaboration by treating funded human review and role-specific training as integral components of detection infrastructure rather than as external safeguards. Without timely access to trained reviewers, forensic expertise, informed decision-makers and defined escalation pathways, technical detection performance cannot become crisis-ready defensive capacity.

References:

ABS-CBN. (2025a, April 21). PH, U.S. troops kick off annual “Balikatan” exercises / ANC. YouTube. https://www.youtube.com/watch?v=_KW-k92yc1E

ABS-CBN. (2024b, April 24). PCO refutes audio ‘deep fake’ of Marcos Jr. *ABS-CBN*. https://www.abs-cbn.com/news/2024/4/23/pco-refutes-audio-deep-fake-of-marcos-jr-018

AIAAIC. (2024, May). Deepfake Philippines President urges military action against China.

https://www.aiaaic.org/aiaaic-repository/ai-algorithmic-and-automation-incidents/deepfake-philippines-president-urges-military-action-against-china

Al Arabiya English. (2024a, April 12). Beijing slams US, Japan, Philippines summit, defends actions in South China Sea. *Al Arabiya English*. https://english.alarabiya.net/News/2024/04/12/beijing-slams-us-japan-philippines-summit-defends-actions-in-south-china-sea

Al Arabiya English. (2024b, April 22). Philippines, US troops begin annual combat drills. *Al Arabiya English*. https://english.alarabiya.net/News/world/2024/04/22/philippines-us-troops-begin-annual-combat-drills

Allen, C. (2002). The role of open sources as the foundation for successful all-source collection strategies. In *NATO open-source intelligence reader* (pp. 12-16). North Atlantic Treaty Organization. https://cyberwar.nl/d/NATO%20OSINT%20Reader%20FINAL%20Oct2002.pdf

Al Jazeera. (2024, April 30). *Philippines alleges China damaged vessel in South China Sea disputed shoal*. Al Jazeera. https://www.aljazeera.com/news/2024/4/30/philippines-and-china-in-new-confrontation-at-scarborough-shoal

Alzaabi, F. R., & Mehmood, A. (2024). A Review of Recent Advances, Challenges, and Opportunities in Malicious Insider Threat Detection Using Machine Learning Methods. *IEEE Access, 12*, 30907-30927. https://doi.org/10.1109/ACCESS.2024.3369906

ANC 24/7. (2024, April 11). Marcos summons Chinese envoy to explain Duterte-Xi ‘gentleman’s agreement’ /*ANC* [Video]. *YouTube* https://www.youtube.com/watch?v=viRGXMkziwU

ANC 24/7. (2024a, April 23). Malacañang warns public vs Marcos 'deepfake' audio [Video]. *YouTube.*https://www.youtube.com/watch?v=bS_3no6ohvA

ANC 24/7. (2024b, April 25). PCO dismisses Marcos Jr.'s supposed "audio recording" as deepfake [Video]. *YouTube.* https://www.youtube.com/watch?v=DjZwAii8tes

ANC 24/7. (2024c, April 26). Authorities probe Marcos ‘deepfake’ audio / *ANC* [Video]. *YouTube.* https://www.youtube.com/watch?v=_fTORiJqCEM

Archive.today. (n.d.). *archive.today*. Retrieved August 27, 2025, from https://archive.ph/

Argosino, F. (2024, April 25). PNP tracing persons behind ‘deepfake’ video showing false clips of Marcos. *INQUIRER.net*. https://newsinfo.inquirer.net/1933654/pnp-tracing-persons-behind-deepfake-video-showing-false-clips-of-marcos

Bakkialakshmi, V.S., Sudalaimuthu, T. (2022). Anomaly Detection in Social Media Using Text-Mining and Emotion Classification with Emotion Detection. In: Guru, D.S., Y. H., S.K., K., B., Agrawal, R.K., Ichino, M. (eds) *Cognition and Recognition. ICCR 2021. Communications in Computer and Information Science, vol 1697. Springer*, Cham. https://doi.org/10.1007/978-3-031-22405-8_5

Balogun, A. Y., Alao, A. I., & Olaniy, O. O. (2025, March). Disinformation in the digital era: The role of deepfakes, artificial intelligence, and open-source intelligence in shaping public trust and policy responses. *Computer Science & IT Research Journal, 6*(2), 28-48. Retrieved from DOI: 10.51594/csitrj.v6i2.1824

Blancaflor, E., Anave, D. M., Cunanan, T. S., Frias, T., & Velarde, L. N. (2023). A literature review of the legislation and regulation of deepfakes in the Philippines. *In Proceedings of the 2023 14th International Conference on E-business, Management and Economics* (pp. 392-397). *Association for Computing Machinery*. https://dl.acm.org/doi/10.1145/3616712.3616722

Business World Online. (2024, April 28). Marcos deepfake probe sought. *Business World Online.* https://www.bworldonline.com/the-nation/2024/04/28/591407/marcos-deepfake-probe-sought/

CachedView. (n.d.). CachedView. Retrieved on August 27, 2025, from https://cachedview.com/

Castaldo, M., Frasca, P., Venturini, T., & Gargiulo, F. (2024). Fake views removal and popularity on YouTube. *Scientific Reports, 14,* 15443.https://www.nature.com/articles/s41598-024-63649-w

Caliwan, C. L. (2024, April 25). ACG boosts cyber patrols, tracks culprits behind PBBM ‘deepfake’. Philippine News Agency. https://www.pna.gov.ph/articles/1223389

Calonzo, A. (2024, April 25). Philippines says ‘foreign actor’ behind deepfake of Marcos urging combat with China. *Time*. https://time.com/6971239/philippines-marcos-deepfake-china-foreign-actor/

Cervantes, F. M. (2024, April 30). Lawmaker suggests classifying use of deepfakes as form of terrorism. *Philippine News Agency.* https://www.pna.gov.ph/articles/1223790

Chi, C., & Esteban, N. (2024, June 8). *Chinese media pushes 'Philippines as aggressor' narrative before viral Marcos deepfake*. Philstar.com. https://www.philstar.com/headlines/2024/06/08/2361039/chinese-media-pushes-philippines-aggressor-narrative-viral-marcos-deepfake

Chubb, A. (2016, July 14). Did China just clarify the nine-dash line? *East Asia Forum.* https://eastasiaforum.org/2016/07/14/did-china-just-clarify-the-nine-dash-line/

Cupin, B. (2024, April 30). *PCG reports damage after China uses water cannons in Bajo de Masinloc*. Rappler. https://www.rappler.com/philippines/coast-guard-ship-damaged-china-water-cannons-scarborough-shoal/

Daily motion (2024, May 1). DepEd wants to fast-track return to old academic calendar by March 2025/ The wRap. *Rapplerdotcom*. https://www.dailymotion.com/video/x8xp9xo

Denscombe, M. (2014). The good research guide: For small-scale social research projects (5th ed.). Open University Press.

Escosio, J. (2024, April 24). Audio ng "attack order" ni PBBM, "deepfake" -Palasyo [Audio]. *Radyo Inquirer 990AM*. https://radyo.inquirer.net/339554/video-ng-attack-order-ni-pbbm-deepfake-palasyo

Esguerra, D. J. (2024, April 24). PCO warns vs. PBBM 'deepfake' asking AFP to act against another nation. *Philippine News Agency.*

https://www.pna.gov.ph/index.php/articles/1223289

Fearn-Banks, K., & Kawamoto, K. (2024). Crisis communications: A casebook approach (6th ed.). *Routledge*. https://doi.org/10.4324/9781003019282

Finklea, K., Christensen, M. D., Fischer, E. A., Lawrence, S. V., & Theohary, C. A. (2015, July 17). Cyber intrusion into U.S. Office of Personnel Management: In brief (CRS Report No. R44111). *Congressional Research Service*. https://sgp.fas.org/crs/natsec/R44111.pdf

Flores, M. (2024, April 17). Philippines, US forces to train retaking island in joint drills. *Reuters.* *https://www.reuters.com/world/philippines-us-forces-train-retaking-island-joint-military-drills-2024-04-17/*

Flores, H. (2024a, April 24). Palace warns public versus deepfake audio of Marcos ordering military attack. *OneNews.PH*. https://www.onenews.ph/articles/palace-warns-public-versus-deepfake-audio-of-marcos-ordering-military-attack

Flores, H. (2024b, April 27). "Foreign actor" seen behind President Marcos audio deepfake. *The Philippine Star*. https://www.philstar.com/headlines/2024/04/27/2350826/foreign-actor-seen-behind-president-marcos-audio-deepfake

Flyvbjerg, B. (2006). Five misunderstandings about case-study research. *Qualitative Inquiry, 12*(2), 219-245. https://www.researchgate.net/publication/221931884

FORUM Staff. (2024, May 12). Deepfake audio falsely portrays Marcos as confrontational. *Indo-Pacific Defense FORUM*. https://ipdefenseforum.com/2024/05/deepfake-audio-falsely-portrays-marcos-as-confrontational/

Geducos, A. C. (2024, April 23). Malacañang warns public vs fake Marcos audio. *Manila Bulletin*. https://mb.com.ph/2024/4/23/malacanang-warns-public-vs-fake-marcos-audio

Global Daily Mirror. (2024, April 24). PCO warns public of deepfake video of PBBM circulating on socmed. Facebook. https://www.facebook.com/globaldailymirror/posts/pco-warns-public-of-deepfake-video-of-pbbm-circulating-on-socmedthe-presidential/748519874072439/

GMA Integrated News (2024, May 1). Nations express concern over China water cannon assault on PCG ship. *GMA News Online*. https://www.gmanetwork.com/news/topstories/nation/905327/nations-express-concern-over-china-water-cannon-assault-on-pcg-ship/story/

GMA Integrated News. (2024a, May 8). AFP on China claim: Audio can be forged with deep fakes. GMA News.https://www.gmanetwork.com/news/topstories/nation/906192/afp-on-china-claim-audio-can-be-forged-with-deep-fakes/story/

Golez, P. (2024, April 28). “Possible source” of Marcos deepfake audio identified. *Politiko.* https://politiko.com.ph/2024/04/28/possible-source-of-marcos-deepfake-audio-identified/daily-feed/

Google. YouTube removals [Transparency report]. (n.d.). *Google Transparency Report*.https://transparencyreport.google.com/youtube-policy/removals

Google. (2025, May 9). Video SEO best practices. *Google Search Central.* https://developers.google.com/search/docs/appearance/video

Gordon, A.E. (2011). Public relations. *Oxford University Press*.274-376

Government of the Philippines. (1950, March 30). Executive Order No. 308, s. 1950: Reorganization of the Armed Forces of the Philippines [Executive order]. *Lawphil*. https://lawphil.net/executive/execord/eo1950/eo_308_1950.html

GT Staff Reporters. (2024, May 8). Manila’s denial of objective facts hurts its own credibility. *Global Times.* https://www.globaltimes.cn/page/202405/1311920.shtml

Guo, B. Yasan Ding, Lina Yao, Yunji Liang, & Zhiwen Yu. (2020). *The Future of False Information Detection on Social Media: New Perspectives and Trends.* ACM Computing Surveys, 53(4), 68:12-27. https://doi.org/10.1145/3393880

Gupta, P., Patil, H.A. & Guido, R.C. Vulnerability issues in Automatic Speaker Verification (ASV) systems. *J AUDIO SPEECH MUSIC PROC. 2024, 10 (2024).* 1-14. DOI: https://doi.org/10.1186/s13636-024-00328-8

Khamparia, A., Pande, S., Gupta, D., Khanna, A., & Sangaiah, A. K. (2020). Multi-level framework for anomaly detection in social networking. *Library Hi Tech, 38(2)*, 350-366. https://doi.org/10.1108/lht-01-2019-0023

Khanjani, Z., Watson, G., & Janeja, V. P. (2023). Audio deepfakes: A survey. *Frontiers in Big Data, 5,* Article 1001063.2-15.https://doi.org/10.3389/fdata.2022.1001063

Kim, Y., & Lim, H. (2023). Debunking misinformation in times of crisis: Exploring misinformation correction strategies for effective internal crisis communication. *Journal of Contingencies and Crisis Management, 31*(3), 406-420. https://onlinelibrary.wiley.com/doi/epdf/10.1111/1468-5973.12447

Kohne, J., Breuer, J., Deubel, A., & Mohseni, M. R. (2024). How to collect data with the YouTube Data API (*GESIS Guides to Digital Behavioral Data, No. 13, Version 1.0)*. *GESIS-Leibniz Institute for the Social Sciences*. https://doi.org/10.60762/ggdbd24013.1.0

Jaiswal, A., Dwivedi, P., & Dewang, R. K. (2024). Handling imbalance dataset issue in insider threat detection using machine learning methods. *Computers & Security, Volume 120, Part B*, 109726, 1-16. https://doi.org/10.1016/j.compeleceng.2024.109726

Jardines, E. A. (2002). Understanding open sources. NATO open-source intelligence reader (pp. 9-11). *North Atlantic Treaty Organization*. Retrieved from https://cyberwar.nl/d/NATO%20OSINT%20Reader%20FINAL%20Oct2002.pdf

Joint Chiefs of Staff. (2017a, July 12). *Doctrine for the Armed Forces of the United States: Joint Publication 1* (Incorporating Change 1, Original work published 25 March 2013). U.S. Department of Defense. https://www.usna.edu/Training/_files/documents/References/3C%20MQS%20References/jp1.pdf

Joint Chiefs of Staff. (2021b). *Joint publication 3-32: Joint maritime operations* (Incorporating Change 1, Original work published 2018). U.S. Department of Defense. https://irp.fas.org/doddir/dod/jp3_32.pdf

Lema, K. (2024, April 3). Philippines prepared to respond to China’s attempts to interfere with re-supply missions. *Reuters*. https://www.reuters.com/world/asia-pacific/philippines-says-it-is-committed-maintaining-position-second-thomas-shoal-2024-04-03/

Lemon, J. (2024). SANS 2024 detection and response survey: Transforming cybersecurity operations: AI, automation, and integration in detection and response. SANS™ Institute https://21984718.fs1.hubspotusercontent-na1.net/hubfs/21984718/Content/Survey_2024-Detection-Response_Prelude%20(1).pdf

Li, X., Chen, P. Y., & Wei, W. (2025). Where are we in audio deepfake detection? A systematic analysis over generative and detection models. *ACM Transactions on Internet Technology.*22:7-15. https://doi.org/10.1145/3736765

Locus, S. (2024a, April 23). PCO slams Marcos 'deepfake' audio. *GMA Integrated News.* https://www.gmanetwork.com/news/topstories/nation/904561/pco-slams-marcos-deepfake-audio/story/

Locus, S. (2024b, April 25). EXPLAINER: What are deepfakes? *GMA Integrated News.*https://www.gmanetwork.com/news/scitech/technology/904779/deepfakes/story/

Lu, X., & Jin, Y. (2024). Optimizing organizational corrective communication: The effects of correction placement timing, refutation detail level, and corrective narrative type on combating crisis misinformation narratives. *Public Relations Review, 50,* Article 102514. https://doi.org/10.1016/j.pubrev.2024.102514

Magramo, K., Tikekar, D., & Lendon, B. (2024, April 30). Chinese water cannon damages ship in new South China Sea flare-up, Philippines says. *CNN.* https://edition.cnn.com/2024/04/30/asia/china-water-cannon-damages-philippines-ship-intl-hnk-ml

Maitem, J. (2024, April 24). Audio deepfake of Marcos Jnr ordering military action against China prompts Manila to debunk clip. *South China Morning Post.*https://www.scmp.com/week-asia/politics/article/3260229/audio-deepfake-marcos-jnr-ordering-military-action-against-china-prompts-manila-debunk-clip

Mangaluz, J. (2024, April 25). Foreign entities eyed in Marcos deepfake clip. *INQUIRER.net.* https://newsinfo.inquirer.net/1933703/foreign-elements-may-be-involved-in-bbm-deepfake

Medenilla, S. P. (2024, April 24). PCO warns public against ‘deepfake’ about Marcos. *BusinessMirror.* https://businessmirror.com.ph/2024/04/24/pco-warns-public-against-deepfake-about-marcos/

ウェブ魚拓 [Megalodon]. (n.d.). Megalodon.jp. Retrieved from https://megalodon.jp/

Ministry of Foreign Affairs of the People’s Republic of China. (2014, December 7). Position paper of the Government of the People’s Republic of China on the matter of jurisdiction in the South China Sea arbitration initiated by the Republic of the Philippines. *FMPRC.* https://www.fmprc.gov.cn/eng/wjb/zzjg_663340/tyfls_665260/tfsxw_665262/202406/t20240606_11405446.html

Ministry of Foreign Affairs of the People’s Republic of China. (2024, April 12). Foreign Ministry Spokesperson Mao Ning’s regular press conference on April 12, 2024. *FMPRC.* https://www.fmprc.gov.cn/nanhai/eng/fyrbt_1/202405/t20240530_11347735.htm

Mohamed, H., Syahaneim Marzukhi, Zuraini Zainol, Mohd, T., & Zakaria, O. (2018). Semantic-based Social Media Threats Detection. *In Proceedings of the 12th International Conference on Ubiquitous Information Management and Communication, Langkawi, Malaysia, January 2018 (IMCOM’18), 5pages.* DOI: https://doi.org/10.1145/3164541.316462

Morales, N., & Orr, B. (2024, June 16). China and Philippines quarrel over South China Sea collision. *Reuters*. https://www.reuters.com/world/asia-pacific/china-coast-guard-says-philippine-supply-ship-illegally intruded-waters-second-2024-06-16/

Murcia, A., & Kabagani, L. J. (2024, April 25). Probe Bongbong 'deepfake' creators, DoJ orders NBI. *Daily Tribune*. https://tribune.net.ph/2024/04/25/probe-bongbong-deepfake-creators-doj-orders-nbi

National People's Congress of the People's Republic of China. (1998, June 26). *Law of the People's Republic of China on the Exclusive Economic Zone and the Continental Shelf* (Order No. 6). AsianLII. https://www.asianlii.org/cn/legis/cen/laws/lotprocoteezatcs790/

NDTV. (2024, April 25). Deepfake audio of Philippine president urging military action against China sparks concerns. *NDTV.* https://www.ndtv.com/world-news/deepfake-audio-of-philippine-president-urging-military-action-against-china-sparks-concerns-5517913

Nepomuceno, P. (2024, May 1). NTF-WPS scores latest Chinese harassment of PH ships in Panatag Shoal. *Philippines News Agency*. https://www.pna.gov.ph/articles/1223792

News Agencies. (2024, March 28). Philippines's Marcos promises action after China's "dangerous attacks". *Al Jazeera.* https://www.aljazeera.com/news/2024/3/28/philippiness-marcos-promises-measures-after-chinas-dangerous-attacks

Ocean Exploration. (2023, January 6). What is an EEZ? *NOAA Ocean Exploration.* https://oceanexplorer.noaa.gov/ocean-fact/useez/

Olander, E. (2024, April 25). Philippine government disavows deepfake audio of President Marcos directing military to act against China. *The China-Global South Project.* https://chinaglobalsouth.com/2024/04/25/philippine-government-disavows-deepfake-audio-of-president-marcos-directing-military-to-act-against-china/

Pandey, A., Gogoi, A. S., Chaudhary, D., Rajput, T., & Mangala, L. (2025). Enhanced insider threat detection using machine learning and Hayabusa-based user behaviour analytics. *Journal of Emerging Technologies and Innovative Research, 12(11)*, e346-e352. https://www.jetir.org/papers/JETIR2511443.pdf

Patel, Y., Tanwar, S., Gupta, R., Bhattacharya, P., Davidson, I. E., Nyameko, R., Aluvala, S., & Vimal, V. (2023). Deepfake generation and detection: Case study and challenges. *IEEE Access, 11*, 143296-143323, https://doi.org/10.1109/ACCESS.2023.3342107

Paul, K. (2024, April 5). Meta overhauls rules on deepfakes, other altered media. *Reuters.* https://www.reuters.com/technology/cybersecurity/meta-overhauls-rules-deepfakes-other-altered-media-2024-04-05/

Permanent Court of Arbitration. (2016, July 12). *The South China Sea arbitration* (The Republic of the Philippines v. The People's Republic of China): Press release. *PCA.* https://pca-cpa.org/en/news/pca-press-release-the-south-china-sea-arbitration-the-republic-of-the-philippines-v-the-peoples-republic-of-china/

Pfeffer, J., Mayer, K. & Morstatter, F. Tampering with Twitter's Sample API. *EPJ Data Sci. 7,* 50 (2018). https://doi.org/10.1140/epjds/s13688-018-0178-0

Philippine Star. (2024, April 24). Palace warns vs Marcos deepfake audio ordering military action. *Global Philstar*. https://www.philstar.com/headlines/2024/04/24/2350113/palace-warns-vs-marcos-deepfake-audio-ordering-military-action

Philippine Star. (2024a, April 26). EDITORIAL - The deepfake menace. *Philstar.* https://www.philstar.com/opinion/2024/04/26/2350581/editorial-deepfake-menace

Piatos, T. C., & Oliquino, E. (2024, April 28). BBM audio deepfake 'source' identified. *Daily Tribune.* https://tribune.net.ph/2024/04/28/bbm-audio-deepfake-source-identified

Presidential Communications Office. (2024, April 23). It has come to the attention of the Presidential Communications Office that there… [Facebook post]. *Facebook.* https://www.facebook.com/pcogovph/posts/837267288425999/

Presidential Communications Office. (2024a, April 26). Gov't vows legal action vs deep fake video creators, spreaders. *PCO*. https://pco.gov.ph/news_releases/govt-vows-legal-action-vs-deep-fake-video-creators-spreaders/

Presidential Communications Office. (2024b, April 27). PNP identifies possible source of deepfake PBBM audio [News release]. *PCO*. https://pco.gov.ph/news_releases/pnp-identifies-possible-source-of-deepfake-pbbm-audio/

PTV Philippines. (2024, April 24). Panayam kay PCO Asec. Dale de Vera kaugnay sa deep fake video na ginamit ang imahe at boses ng Pangulo [Video]. *YouTube.* https://www.youtube.com/watch?v=dzwgkGoTNGw

Pulta, B. (2024, April 25). DOJ tasks NBI to probe 'deepfake' PBBM audio. *Philippine News Agency.* https://www.pna.gov.ph/articles/1223398

Rahman, Md. S., Halder, S., Uddin, Md. A., & Acharjee, U. K. (2021). An efficient hybrid system for anomaly detection in social networks. *Cybersecurity, 4(1)*.2-10. https://doi.org/10.1186/s42400-021-00074-w

Ramirez, E., Brill, J., Ohlhausen, M. K., Wright, J. D., & McSweeny, T. (2014, May). Data brokers: A call for transparency and accountability. *Federal Trade Commission.* https://www.ftc.gov/system/files/documents/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014/140527databrokerreport.pdf

Rappler. (2024, April 16). Marcos on getting to the bottom of Duterte-China secret agreement: "Daming palusot" [Video]. *YouTube*. https://www.youtube.com/shorts/j011FtB1Wig

Rappler. (2024a, April 18). Today's headlines: Marcos in TIME, Arnie Teves, Philippines-China relations /*The wRap* /Apr 18, 2024 [Video]. *YouTube.* https://www.youtube.com/shorts/BW76I9dC1Ic

Rappler. (2024c, April 24). Malacañang sounds alarm over Marcos deepfake audio /*The wRap* [Video]. *Rappler.* https://www.rappler.com/video/daily-wrap/april-24-2024/

Rappler. (2024d, April 24). MARCO'S DEEPFAKE AUDIO [YouTube Shorts]. *YouTube.* https://www.youtube.com/shorts/6x_a6L7l-gA

Rappler. (2024b, April 25). Malacañang sounds alarm over Marcos deepfake audio /*The wRap* [Video]. *YouTube.* https://www.youtube.com/watch?v=2kmOVLzBy7M&t=118s

Rappler. (2024f, April 24). PCO warns VS audio deepfake [YouTube Shorts]. *YouTube.* https://www.youtube.com/shorts/PwnnmAquyQs

Rappler. (2024e, May 7). Marcos says the Philippines will not use water cannon vs Chinese ships [Video]. *YouTube.* https://www.youtube.com/shorts/iXZYZYmCN6s

Respicio & Co. (2024, October 13). Commander-in-Chief powers of the President (Philippines). *Respicio & Co*.https://www.respicio.ph/bar/2025/political-law-and-public-international-law/executive-department/powers-of-the-president/commander-in-chief-powers

Reuters. (2024, April 13). Ties with US and Japan will 'change dynamic' in South China Sea, Philippines president says. *The Guardian.* https://www.theguardian.com/world/2024/apr/13/dynamic-in-south-china-sea-is-changing-through-growing-us-and-japan-ties-says-philippines-president

Reuters. (2024a, April 22). Philippines, US troops begin annual combat drills. *Reuters.* https://www.reuters.com/world/philippines-us-troops-begin-annual-combat-drills-2024-04-22/

Reuters. (2024b, April 30). Philippines accuses China of damaging its vessels in disputed South China Sea shoal. *Reuters*. https://www.reuters.com/world/asia-pacific/chinas-coast-guard-expels-philippine-vessels-scarborough-shoal-state-media-says-2024-04-30/

Reuters Staff. (2024, June 24). Philippines accuses China of using "illegal force" to deliberately disrupt resupply mission. *Reuters*. https://www.reuters.com/world/asia-pacific/philippines-continue-south-china-sea-resupply-missions-defense-sec-says-2024-06-24/

Revilla, J. [@jolorevillaIII]. (2024, April 24). From Presidential Communications Office: It has come to the attention of the PCOGO... [Facebook post]. *Facebook*. https://www.facebook.com/jolorevillaIII/posts/from-presidential-communications-officeit-has-come-to-the-attention-of-the-pcogo/1006721114142661/

Reynolds, B., & Seeger, M. W. (2005). Crisis and emergency risk communication as an integrative May. *Journal of Health Communication, 10*(1), 43-55. https://doi.org/10.1080/10810730590904571

Rita, J. (2024, May 14). Social media group files complaint over Marcos deepfakes. *GMA Integrated News.* https://www.gmanetwork.com/news/topstories/nation/906706/social-media-group-files-complaint-over-marcos-deepfakes/story/

Schwartz, B. (2013, February 22). YouTube Channel Page Now Only Showing Public Video View. *Search Engine Roundtable*. https://www.seroundtable.com/youtube-views-channel-16403.htm

Secretariat, D. (2024, July 10). Joe Biden's "Magical Pistachio" video has A.I. voice. *Deepfake Analysis Unit*. https://www.dau.mcaindia.in/blog/joe-bidens-magical-pistachio-video-has-a-i-voice

Seitz-Wald, A. (2024, May 22). Political consultant who admitted deepfaking Biden's voice is indicted. *CNBC news*. https://www.cnbc.com/2024/05/23/political-consultant-who-admitted-deepfaking-bidens-voice-is-indicted-.html

Shaaban, O. A., Yildirim, R., & Alguttar, A. A. (2023). Audio deepfake approaches. *IEEE Access, 11*, 132652-132682. DOI:10.1109/ACCESS.2023.3333866

Silverman, C. (Ed.). (2014). Verification handbook: An ultimate guideline on digital-age sourcing for emergency coverage. *European Journalism Centre* https://verificationhandbook.com/downloads/verification.handbook.pdf

Sina News. (2024, April 25). 马科斯威胁对华开战？菲律宾总统通讯办公室:音频是伪造的 [Marcos threatens to go to war with China? Presidential Communications Office: Audio is fake]. Sina News. https://news.sina.com.cn/c/2024-04-25/doc-inasziaa7695874.shtml

Soule, M. Ryan. P. (2002). Grey Literature. NATO open source intelligence reader. *North Atlantic Treaty Organization*. https://cyberwar.nl/d/NATO%20OSINT%20Reader%20FINAL%20Oct2002.pdf

Stewart, Ph. (2024, May 3). Damage, injury to Philippines in South China Sea is 'irresponsible behaviour', says US Defense Secretary. *Reuters*. https://www.reuters.com/world/asia-pacific/damage-injury-philippines-south-china-sea-is-irresponsible-behaviour-says-us-2024-05-03/

Szymoniak, S. & Foks, K. (2024). Open-Source Intelligence Opportunities and Challenges -A Review. *Advances in Science and Technology Research Journal, 18(3),* 123-139. https://doi.org/10.12913/22998624/186036

't Hart, F. M. P. (2002). *Introduction to the NATO OSINT reader*. In NATO open source intelligence reader. North Atlantic Treaty Organization. https://cyberwar.nl/d/NATO%20OSINT%20Reader%20FINAL%20Oct2002.pdf

Tan, C. B., Hijazi, M. H. A., Khamis, N., Nohuddin, P. N. E. binti, Zainol, Z., Coenen, F., & Gani, A. (2021). A survey on presentation attack detection for automatic speaker verification systems: State-of-the-art, taxonomy, issues and future direction. *Multimedia Tools and Applications, 80*(21-23), 32725–32762. https://doi.org/10.1007/s11042-021-11235-x

The Straits Times. (2024, May 1). Philippines says Chinese coast guard elevating tensions in South China Sea. *The Strait Times*. https://www.straitstimes.com/asia/philippines-says-chinese-coast-guard-elevating-tensions-in-south-china-sea

Tuor, A., Kaplan, S., Hutchinson, B., Nichols, N., & Robinson, S. (2017). Deep Learning for Unsupervised Insider Threat Detection in Structured Cybersecurity Data Streams. *ArXiv, abs/*1710.00811. https://www.semanticscholar.org/reader/bac292286c7d12df6863c7ca8dec720ed6288302

Torres Tupas, T. (2024, April 25). NBI ordered to investigate deep fake video of President. *INQUIRER.net*. https://newsinfo.inquirer.net/1933697/remulla-on-deep-fake-audio

Tyrrell, P. (2002). Open-source intelligence: The challenge for NATO. In NATO open-source intelligence reader (pp. 3-8). *North Atlantic Treaty Organization.* https://cyberwar.nl/d/NATO%20OSINT%20Reader%20FINAL%20Oct2002.pdf

U.S. Department of Defense. (2023, May 3). Fact sheet: U.S.-Philippines bilateral defense guidelines. https://media.defense.gov/2023/May/03/2003214357/-1/-1/0/THE-UNITED-STATES-AND-THE-REPUBLIC-OF-THE-PHILIPPINES-BILATERAL-DEFENSE-GUIDELINES.PDF

U.S. Embassy Manila. (2024, April 17). Philippine, U.S. troops to kick off Exercise Balikatan 2024. *U.S. Embassy in the Philippines.* https://ph.usembassy.gov/philippine-u-s-troops-to-kick-off-exercise-balikatan-2024/

Vego, M. (2007). ON MAJOR NAVAL OPERATIONS. *Naval War College Review, 60*(2), 94-126. http://www.jstor.org/stable/26396823

VERA Files. (2024, May 10). Factcheck: Audio clip of Marcos calling Duterte's 'gentlemen's agreement' with China treason FAKE. *VERAFiles*. https://verafiles.org/articles/vera-files-fact-check-audio-clip-of-marcos-calling-dutertes-gentlemens-agreement-with-china-treason-fake

vidIQ. (2024, April 22). Dapat Balita's subscriber count, stats & income- vidIQ YouTube Stats. *vidIQ.* https://vidiq.com/youtube-stats/channel/UClLHnDxskpxaR1ydIRdOJOA/

Visave, J. (2024). AI in Emergency Management: Ethical Considerations and Challenges. *Journal of Emergency Management and Disaster Communications,* 165-182. https://doi.org/10.1142/s268998092450009x

Wang, O. (2024, September 5). What is Beijing's 9-dash line in the South China Sea and what does it mean? *South China Morning Post.* https://www.scmp.com/news/china/diplomacy/article/3275865/what-beijings-9-dash-line-south-china-sea-and-what-does-it-mean

Wang, Q., & Fan, W. (2024, April 30). China Coast Guard expels two Philippine vessels that illegally intruded in waters off China's Huangyan Dao. *Global Times.* https://www.globaltimes.cn/page/202404/1311524.shtml

WebCite. (n.d.). *WebCite query interface*. Retrieved August 27, 2025, from https://www.webcitation.org/query.php

Wayback Machine. Retrieved August 27, 2025. *Internet Archive.* Available at https://web.archive.org/

Williams, H. J., & Blum, I. (2018). Defining second generation open-source intelligence (OSINT) for the defense enterprise. *RAND Corporation*.10-23. https://apps.dtic.mil/sti/html/trecms/AD1053555/

Yin, R. K. (2018). Case study research and applications: Design and methods (6th ed.). *SAGE.*

Younas, M. (2020, November). Anomaly Detection Using Data Mining Techniques: A Review. *International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887 Volume 6 Issue II*, *February 2018*.568-574. Available at https://www.ijraset.com/fileserve.php?FID=32188

YouTube Help. (n.d.-a). How engagement metrics are counted. Retrieved from https://support.google.com/youtube/answer/2991785

YouTube Help. (n.d.-b). Invalid traffic on your videos. Retrieved from https://support.google.com/youtube/answer/10285842

YouTube. (2018, December 13). Faster removals and tackling comments -An update on what we're doing to enforce YouTube's Community Guidelines. *YouTube Blog*.https://blog.youtube/news-and-events/faster-removals-and-tackling-comments/

Zong, H. (2024, March 28). The facts and truth about Ren'ai Jiao. *Global Times*. https://www.globaltimes.cn/page/202403/1309664.shtml

Appendix

The attached maps below illustrates the key Strategic locations and maritime zones involved within South China Sea, including the Second Thomas Shoal (pinned zone which is called Ren'ai Jiao in Chinese, Ayungin in Filipino) and the surrounding countries in the South China sea including China, Vietnam, Malaysia, Brunei, and Philippines., directly related to the case of the fabricated clip circulated on April, 2024.

Figure A.1. Map showing Ren'ai Jiao / Ayungin Shoal in relation to the nine-dash / eleven-dash line and the Philippine EEZ. Source: Screenshot captured by author from Wang (2024, September 5), Captured in Oct 2025.

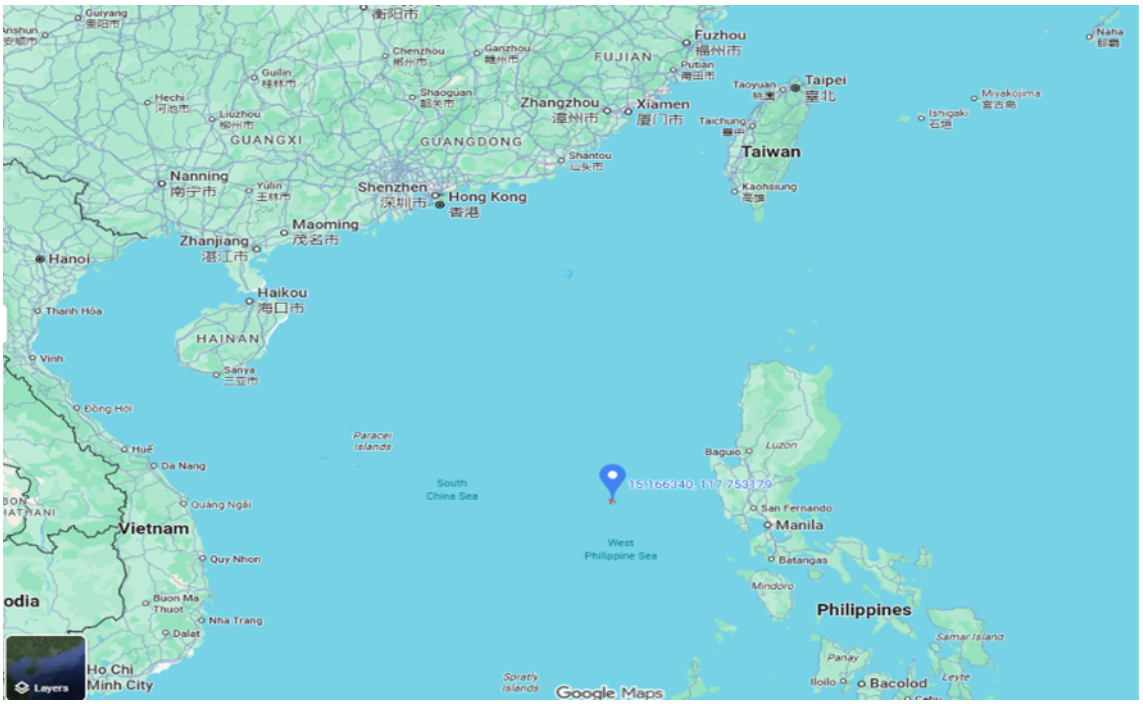


Figure A.2. Map showing Scarborough Shoal as a conflict zone.
Source: Screenshot captured by author from Google Maps, October 2025.

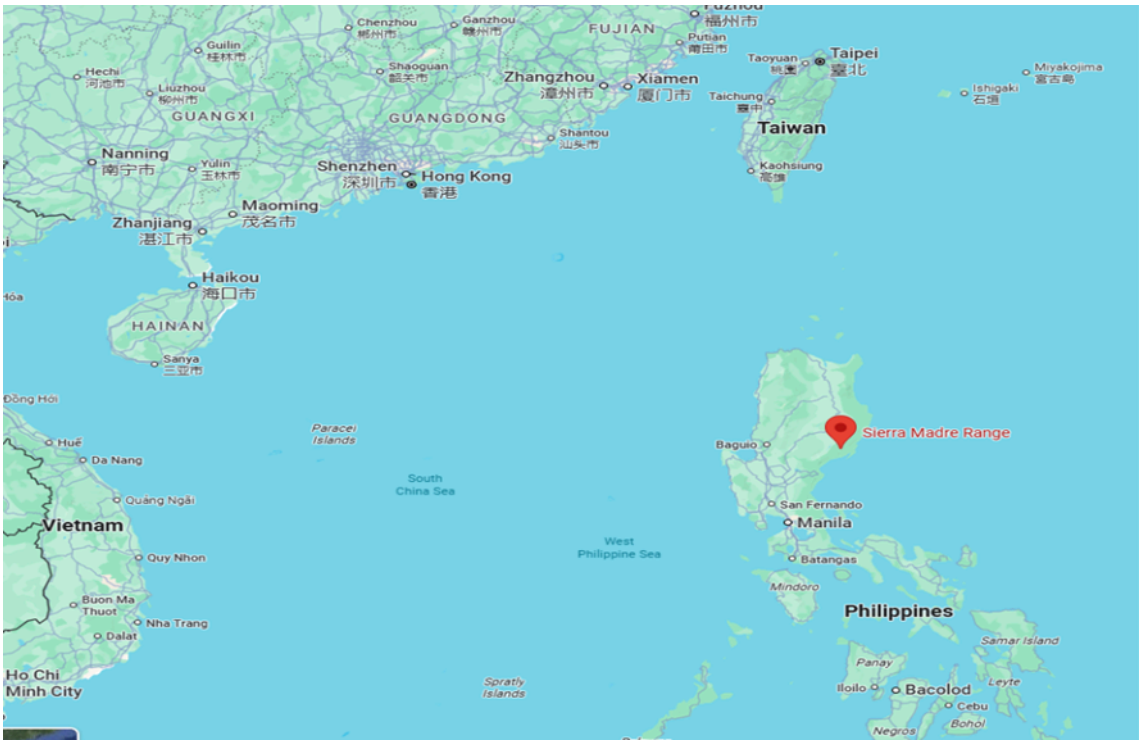


Figure A.3. Map showing the BRP Sierra Madre / Ayungin Shoal military outpost area.
Source: Screenshot captured by author from Google Maps, October 2025.

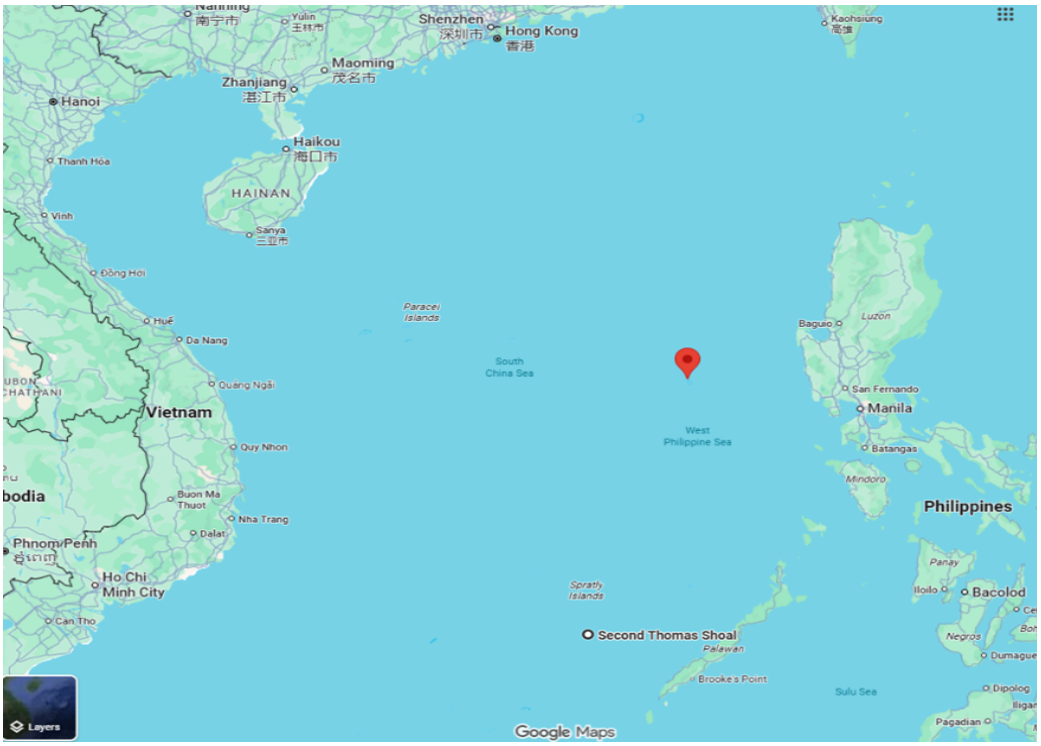


Figure A.4. Map showing The Thomas Shoal as a contested maritime zone.
Source: Screenshot captured by author from Google Maps, October 2025.

China coastguard blasts water cannon at Philippine coastguard ship

30 Apr 2024

Figure A.5. Image showing a confrontation at Scarborough Shoal between the Philippines and China following circulation of the fabricated Marcos audio. Source: Screenshot captured by author in October 2025 from *Al Jazeera* (2024, April 30th) documenting the confrontation at Scarborough Shoal

Figure A.6. Image showing the opening ceremony of Balikatan 2024 between the United States and the Philippines. Military leaders, including Philippine Generals Licudine, Brawner, Beleran, U.S. Lt. Gen. Jurney, and Chargé d'Affaires Ewing, link arms at the Balikatan 2024 opening ceremony in Quezon City.

Source: Screenshot captured by author from (Reuters. 2024a), captured in October,2025.