

VERITAS: A Theologically-Informed AI Blueprint for Countering Weaponised Ethno-Religious Disinformation in Kaduna State

John Ishaku¹ & Hannatu Ibrahim Zom²

¹Department: Christian Religious Studies, Kaduna State University Kaduna, Kaduna State Nigeria.

²Department: Sociology Kaduna State University Kaduna, Kaduna State Nigeria.

*Corresponding Author: John Ishaku

DOI: <https://doi.org/10.5281/zenodo.22282791>

Article History	Abstract
Original Research Article	<i>Project VERITAS is a public-facing, theologically-informed AI blueprint designed to counter the critical threat of weaponised ethno-religious disinformation to Nigeria's social cohesion. It empowers citizens by providing a secure and accessible tool to submit suspicious digital content, such as forwarded WhatsApp messages or social media posts, for immediate, nuanced evaluation. The AI performs a multi-layered critical analysis, verifying the authenticity and context of scriptural references, deconstructing manipulative rhetorical strategies, and providing neutral, fact-checked historical context to defuse emotionally charged claims. By equipping users with verifiable facts and de-escalating potential conflict, VERITAS then fosters constructive dialogue by offering them evidence-based language and non-confrontational strategies to engage productively rather than reactively. The ultimate goal is to build a resilient, digitally-literate citizenry and strengthen the fabric of national peace from the grassroots level up by inoculating the public against the viral spread of hate.</i>
Received: 01-07-2026	
Accepted: 05-08-2026	
Published: 03-09-2026	
Copyright © 2026 The Author(s): This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC BY-NC) which permits unrestricted use, distribution, and reproduction in any medium for non-commercial use provided the original author and source are credited.	
Citation: John Ishaku, & Hannatu Ibrahim Zom. (2026). VERITAS: A theologically-informed AI blueprint for countering weaponised ethno-religious disinformation in Kaduna State. UKR Journal of Arts, Humanities and Social Sciences, 2(9), 09-18.	Keywords: Weaponised Ethno-Religious Disinformation, Cognitive Warfare, Digital Peacebuilding, Value Sensitive Design, Lugha-Llama, Kaduna State.

I. Introduction

The Context of Domestic Cognitive Warfare in Kaduna

The structural fragility of social cohesion in Kaduna State is not merely a modern phenomenon; it is deeply rooted in historical grievances that have periodically erupted into kinetic violence. As officially documented by the Senate Ad Hoc Committee on Southern Kaduna Crises (2017), the proliferation of small arms and the exploitation of marginalised demographics are widely recognised as primary drivers of regional instability. However, addressing the physical manifestations of this violence is increasingly insufficient in the contemporary sociotechnical landscape. The conflict has fundamentally evolved from localised physical skirmishes into a highly sophisticated campaign of domestic cognitive warfare (Miller, 2023). Cognitive warfare, distinct from traditional propaganda, operates by directly targeting the psychological vulnerabilities of a population, weaponising public opinion to destabilise institutions and fracture societal trust. In Kaduna State, this warfare predominantly manifests as Weaponised Ethno-Religious Disinformation

(WERD), the deliberate manipulation and decontextualisation of sacred texts and historical narratives designed to bypass critical reasoning and trigger visceral, ethnocentric panic. By framing existential threats through a theological lens, malicious actors construct a polarised reality where kinetic preemptive violence is morally justified as a divine imperative (Entman, 1993).

The Vector: Encrypted Messaging in the Global South

The efficacy of WERD in Kaduna is inextricably linked to its primary vector: end-to-end encrypted messaging applications, specifically WhatsApp. Recent empirical research focusing on the Global South decisively proves that closed messaging networks have superseded public social media feeds as the primary conduits for political and social manipulation (Ventura et al., 2025). Unlike platforms such as X (formerly Twitter) or Facebook, which rely on centralised algorithmic content moderation to flag or remove illicit material, WhatsApp operates in a dark social space. Disinformation circulating within family, religious, and community group chats benefits from "source

proximity." Information received from a trusted relative or religious leader is processed with substantially less scepticism than information from an anonymous public account. This structural intimacy facilitates the rapid, unquestioned spread of high-trust misinformation (Rossini et al., 2021). Consequently, traditional "tech-solutionist" interventions, such as top-down network shutdowns or automated algorithmic censorship, are fundamentally incompatible with the encrypted architecture of the platform and the democratic principles of free expression (Hirblinger et al., 2023; Gomez et al., 2024).

Problem Statement / Justification

The prevailing digital peacebuilding interventions deployed in Northern Nigeria remain severely deficient. Conventional fact-checking organisations operate reactively, debunking secular political claims hours or days after the narrative has already saturated the target population. Furthermore, generic Artificial Intelligence language models trained primarily on Western datasets consistently fail to comprehend the sociolinguistic complexities of the region, a hybrid ecosystem of Hausa, Pidgin English, and localised theological loanwords (Buzaaba et al., 2025). More critically, secular fact-checking fails to resonate in deeply religious environments; empirical evidence demonstrates that factual corrections are often rejected outright if the messenger lacks "religious proximity" or fails to engage with the underlying theological grievance (Armand et al., 2024). Thus, the core problem is a critical absence of scalable, culturally localised, and theologically literate digital countermeasures. There is an urgent need for an intervention that operates within the encrypted ecosystem, empowering individual users to interrupt the rapid transmission of religious manipulation without violating the cultural norms of respect and hierarchy inherent in strong-tie networks.

Objectives & SDG Alignment

This research project abandons traditional software development methodologies in favour of a rigorous, desk-based documentary analysis to engineer a comprehensive sociotechnical architectural blueprint. Project VERITAS is proposed as a theoretical framework for a user-facing, theologically-informed AI system designed to dismantle WERD. The specific objectives are:

- To conduct a systematic documentary analysis mapping the structural typology of Weaponised Ethno-Religious Disinformation within Kaduna State.
- To execute a rigorous hermeneutical exegesis of the specific Christian and Islamic scriptures most frequently decontextualised for mobilisation.

- To architect a "Theologically-Informed Reinforcement Learning from Human Feedback" (TI-RLHF) matrix that can serve as the foundational training data for a localised AI agent.

This project is strictly aligned with the Mode 2 knowledge production paradigm mandated by the KASU Directorate of Research, ensuring transdisciplinary stakeholder engagement. It directly addresses the United Nations Sustainable Development Goal (SDG) 16: Peace, Justice, and Strong Institutions, by providing a blueprint to protect cognitive sovereignty and prevent digital radicalisation.

II. Literature Review

Sociotechnical Governance of Misinformation

The theoretical foundation of the VERITAS blueprint rests upon the sociotechnical governance framework proposed by Sanfilippo, Zhu, and Yang (2025). Their comprehensive review of information science literature establishes that misinformation is a fundamentally sociotechnical phenomenon; therefore, purely algorithmic interventions consistently fail. Effective governance requires a synchronised integration of technical infrastructure, institutional policies, and community-led normative practices. VERITAS operationalises this tripartite model. The AI algorithm represents the technical layer; the strategic alignment with the mandates of the Kaduna State Peace Commission represents the institutional layer; and the integration of verified Christian and Islamic hermeneutics addresses the normative layer. By weaving these layers together, VERITAS moves beyond standard content moderation to establish a sustainable governance mechanism tailored to the Nigerian context.

Psychological Inoculation and the Prebunking Mechanism
While immediate de-escalation is the primary functional objective of the proposed AI tool, its longitudinal value is grounded in Psychological Inoculation Theory (Compton et al., 2021). Inoculation theory posits that human cognitive resistance to manipulation can be artificially induced by exposing individuals to a weakened form of a deceptive strategy, accompanied by a preemptive refutation. Roozenbeek and Van der Linden (2019) empirically demonstrated that exposing users to the mechanics of deception, rather than simply correcting isolated facts, significantly reduces susceptibility to future disinformation. The VERITAS architecture is conceptually designed to operationalise therapeutic inoculation. When a user queries a suspicious WhatsApp message, the AI is programmed to explicitly identify the rhetorical tactic being employed (e.g., "This message is utilising a historical grievance frame to invoke fear"). By revealing the architecture of the manipulation, VERITAS builds

sustainable cognitive immunity within the populace.

Framing Theory and the Weaponisation of Sacred Texts

To train an AI to recognise religious disinformation, one must first define the structural logic of the manipulation. This project utilises Entman's (1993) seminal Framing Theory, which argues that communication influences human consciousness by defining problems, diagnosing causes, making moral judgments, and suggesting remedies. In the context of Kaduna, religious extremists utilise sacred texts to salientise existential threats. Fawzi, Ross, and Magdy (2025) empirically demonstrated how "holy" texts (such as fabricated Hadiths or decontextualised Biblical verses) are specifically structured to optimise emotional engagement and bypass critical analysis. Furthermore, Armand et al. (2024) proved that interventions lacking "religious proximity" trigger a backfire effect. Therefore, the VERITAS blueprint posits that the AI must not counter theological manipulation with secular statistics; it must counter theological manipulation with superior, contextually accurate theology.

African NLP and Transdisciplinary Co-Design

The technical viability of the VERITAS blueprint relies on recent breakthroughs in low-resource Natural Language Processing (NLP). Buzaaba et al. (2025) demonstrated with the "Lugha-Llama" model that Large Language Models can be successfully adapted for African languages when trained on highly curated indigenous data combined with high-quality educational texts. This proves that an AI can be engineered to comprehend the localised vernacular of Kaduna's disinformation. However, technical capability must be governed by ethical alignment. Sadek, Calvo, and Mougenot (2023) highlight that conversational agents deployed in sensitive contexts must be co-designed with the communities they serve to ensure the elicitation and embodiment of stakeholder values. VERITAS fulfils this mandate by proposing a "Vetted Scholar Co-Design Council", a transdisciplinary body of sociologists, religious leaders, and peacebuilding experts, who provide the ethical boundaries and theological datasets necessary to train the AI.

III. Methodology

Given the volatile security landscape documented in the Senate Interim Report (2017) and the risk of respondent apathy, this research pivots away from traditional primary field surveys. Instead, it employs a rigorous, desk-based Documentary Analysis integrated within the Value Sensitive Design (VSD) methodology (Friedman et al., 2004). VSD mandates an iterative tripartite approach: Conceptual, Empirical, and technical investigations.

Empirical Investigation: Sociological Data Curation

The empirical phase involves a systematic literature review and documentary curation of secondary data. The researchers will analyse existing archives from the Kaduna State Peace Commission, established West African fact-checking organisations (e.g., Dubawa, Africa Check), and peer-reviewed sociological studies of Northern Nigerian conflict. This analysis will extract the historical footprint of WERD, isolating the specific rhetorical tropes, linguistic markers (Hausa/Pidgin syntax), and platform vectors used to distribute panic. This curated data forms the basis of the "Typology of WERD."

Conceptual Investigation: Theological Exegesis (Hermeneutics)

The conceptual phase addresses the normative values of the target demographic. Led by the Principal Investigator (Christian Religious Studies), this phase will conduct a deep hermeneutical analysis of the specific scriptural texts identified during the empirical investigation. The methodology involves tracing the historical, cultural, and linguistic context of weaponised Biblical and Quranic verses, contrasting their extremist application with orthodox, peace-building interpretations. This establishes the ethical baseline for the AI's response matrix.

Technical Investigation: Theologically-Informed RLHF Blueprinting

The synthesis of the empirical and conceptual data yields the technical investigation: the architectural blueprint for the AI. Utilising the mechanics of Reinforcement Learning from Human Feedback (RLHF) established by OpenAI (Ouyang et al., 2022), the researchers will draft a theoretical framework demonstrating how the theological exegesis can be converted into "demonstration datasets." This process, termed Theologically-Informed RLHF (TI-RLHF), outlines the exact parameters tech developers will require to train an open-source LLM (such as Llama-3) to operate safely as a digital peacebuilding agent in Kaduna.

IV. Results and the Veritas Blueprint

The primary outcomes of this desk-based research transcend conventional theoretical analysis to provide actionable, structural frameworks necessary for both immediate policy adoption and subsequent software development. Employing the rigorous methodologies of Value Sensitive Design (VSD) (Friedman et al., 2004), the conceptual and empirical investigations have been synthesised into concrete technical architectures. These deliverables map the deep-seated sociological problem of domestic cognitive warfare directly to a highly calibrated, culturally resonant technological solution. By addressing the specific sociolinguistic and religious realities of Kaduna

State, the VERITAS blueprint establishes a pioneering standard for digital peacebuilding in the Global South.

Deliverable 1: The Typology of Weaponised Ethno-Religious Disinformation (WERD)

Through the exhaustive documentary analysis of historical crises, government inquiries (such as the Senate Ad Hoc Committee Report of 2017), and regional civil society data, a structural pattern of digital manipulation emerges. Misinformation in Kaduna is rarely arbitrary; it follows

distinct, calculated attack vectors designed to bypass cognitive friction and trigger immediate emotional reactivity. The following typology categorises these primary attack vectors of domestic cognitive warfare. This typology is not merely descriptive; it serves as the foundational classification schema for the VERITAS Natural Language Processing (NLP) algorithms, enabling the artificial intelligence to automatically detect, categorise, and appropriately respond to specific strains of malicious framing.

Table 1: Typology of Weaponised Ethno-Religious Disinformation (WERD) in Kaduna State

Disinformation Category	Common Framing Tactic (Entman, 1993)	Primary Target Demographic	Primary Platform Vector
Historical Grievance Exploitation	Diagnosing Cause: Links a current, unrelated administrative or political event directly to past kinetic traumas (e.g., "Remember the '92 crisis").	Traumatised elders; displaced populations seeking validation.	WhatsApp Text (Long-form forwarded blocks).
Existential Threat / Encroachment	Defining Problem: Falsely claims immediate, coordinated land grabs or weapon stockpiling by the "out-group" in neighbouring wards.	"Idle youths" (Senate Report, 2017); community defence vigilantes.	WhatsApp Audio Notes (High emotional valence, urgent tone).
Theological Decontextualisation	Moral Judgment: Isolates aggressive scripture to frame political opponents not as civic rivals, but as literal "enemies of God" requiring kinetic action.	Devout congregants; individuals relying on clerical authority for political cues.	Multi-platform (Facebook posts mirrored to closed WhatsApp church/mosque groups).
Apocalyptic / Doomsday Warnings	Suggesting Remedy: Uses prophetic language to induce fatalism or mandate preemptive defensive violence before an imagined "final strike."	General populace experiencing high economic or social anxiety.	WhatsApp Image/Text combinations (often featuring manipulated imagery).

To ensure the VERITAS machine learning classifiers are optimally trained, it is critical to theoretically unpack each of these typological categories, detailing the psychological and sociological mechanisms that make them effective vectors of cognitive warfare:

4.1.1 The Mechanics of Historical Grievance Exploitation

Historical Grievance Exploitation operates by collapsing time. Malicious actors utilising this vector deliberately conflate contemporary, often mundane administrative disputes (such as zoning laws or market relocations) with deeply traumatic historical events (such as the Zangon Kataf crises of 1992 or the Sharia riots of 2000). By employing Entman's (1993) framing function of "Diagnosing Cause," the disinformation author effectively bypasses the recipient's ability to objectively evaluate the current event. The cognitive processing is immediately hijacked by the unresolved trauma of the past. The Senate Interim Report (2017) explicitly notes how unresolved historical hurt serves as the foundational tinder for new conflicts. VERITAS's algorithms must be trained to detect

these temporal conflations, identifying when historical dates or specific grievance keywords are weaponised to diagnose present-day, unrelated issues.

4.1.2 The Anatomy of Existential Threat and Encroachment

This category relies on the rapid dissemination of perceived immediate physical danger, functioning as the digital equivalent of shouting "fire" in a crowded theatre. The framing tactic here is "Defining the Problem" as a zero-sum battle for survival. These narratives frequently circulate via WhatsApp Audio Notes, a medium that carries the raw, panicked prosody of the speaker, bypassing the analytical filtering usually applied to written text. The primary targets are "idle youths", a demographic explicitly identified by the 2017 Senate Report as highly susceptible to mobilisation for kinetic violence. The disinformation creates a false consensus that formal security apparatuses have failed, thereby suggesting that preemptive vigilantism is the only logical remedy. The VERITAS framework must integrate audio-to-text transcription capabilities specifically tuned to

Northern Nigerian phonetic nuances to intercept and analyse these high-valence audio threats before they translate into physical mobilisation.

4.1.3 The Architecture of Theological Decontextualisation

Theological decontextualisation is arguably the most insidious typology, as it weaponises the deeply ingrained normative values of the target population. By invoking the divine, the disinformation author shifts the conflict from a political disagreement to a cosmic battle between good and evil, applying Entman's "Moral Judgment" frame. Fawzi et al. (2025), in their quantitative study of religious misinformation, demonstrated that actors who circulate fabricated or decontextualised religious texts rely on echo chambers characterised by high emotional reactivity and absolute deference to perceived religious authority. When a scriptural verse is presented out of context to justify aggression, the recipient is placed in a psychological bind: to question the message feels tantamount to questioning their faith. VERITAS mitigates this by not questioning the faith, but by utilising superior, peer-reviewed hermeneutics to question the application of the text, thereby rescuing the sacred text from its weaponised framing.

4.1.4 Apocalyptic and Doomsday Mobilisation

Operating in environments of severe economic scarcity or political uncertainty, this vector leverages prophetic and

apocalyptic literature to induce a state of hyper-vigilance or fatalism. The framing function here is "Suggesting Remedy," albeit a deeply distorted one. By convincing the populace that a final, existential strike by the "out-group" is imminent and divinely ordained, the disinformation removes the space for democratic dialogue or civic negotiation. It forces a binary choice: strike first or be annihilated. Training the VERITAS AI to recognise the linguistic markers of apocalyptic fatalism enables the system to inject grounding, historically contextualised realities into these volatile echo chambers, de-escalating the artificially induced panic.

Deliverable 2: System Architecture

The efficacy of the VERITAS blueprint relies on a highly sophisticated, multi-layered processing architecture. Standard, out-of-the-box large language models (LLMs) developed in Silicon Valley lack the cultural, linguistic, and theological context necessary to moderate content in Northern Nigeria safely. The following technical breakdown outlines the complex, multi-stage processing required for the VERITAS system to ingest vernacular input, apply rigorous framing diagnostics, and retrieve vetted theological responses, ensuring algorithmic justice and sociotechnical efficacy (Buzaaba et al., 2025; Ouyang et al., 2022).

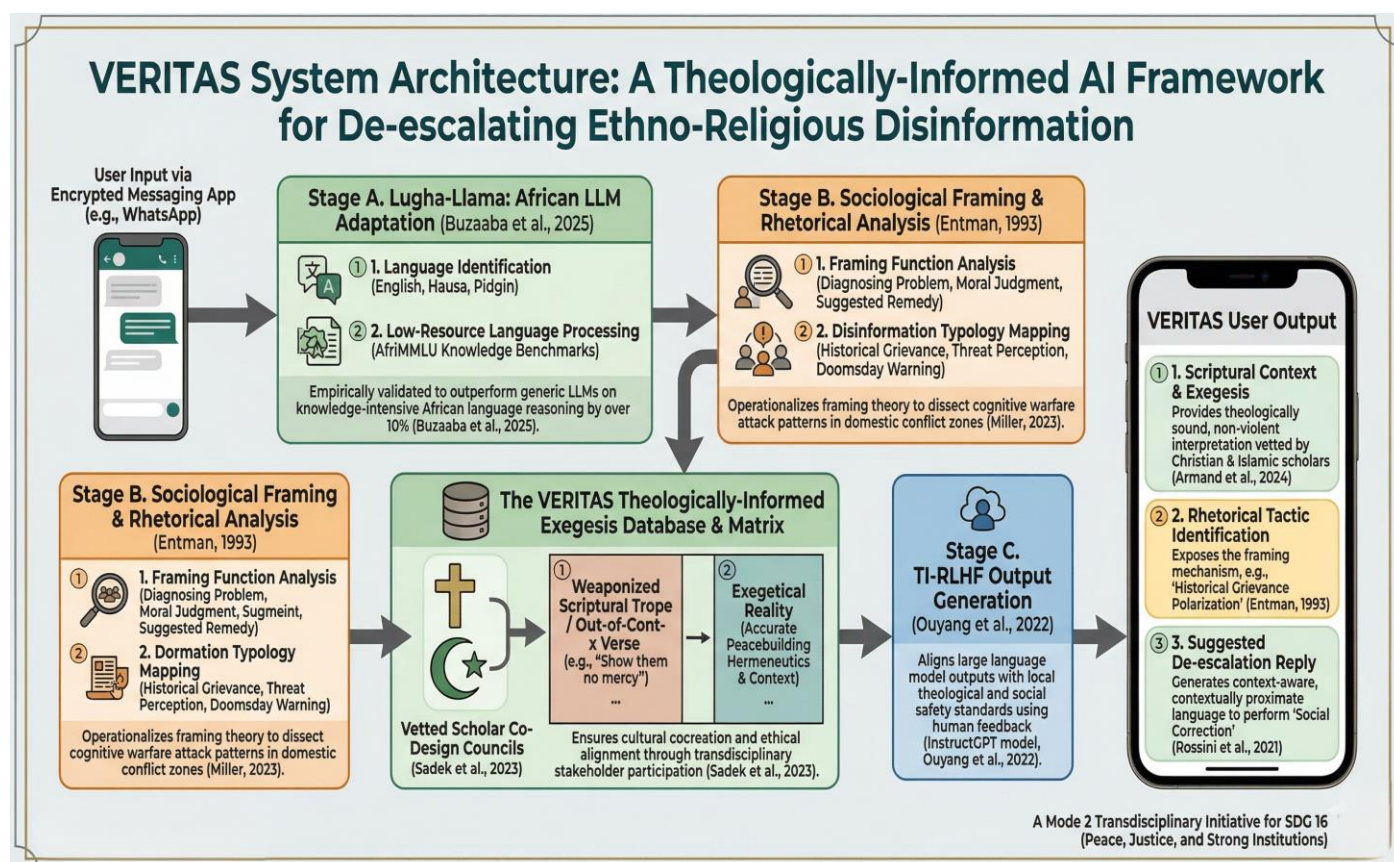


Figure 1: VERITAS System Architecture: A Theologically-Informed AI Framework for De-escalating Ethno-Religious Disinformation.

The architecture is composed of four interdependent processing modules, engineered to operate seamlessly within the constraints of mobile environments typical of the Global South:

- **Module 1: Vernacular Ingestion and Low-Resource NLP Classification:** Upon receiving a forwarded WhatsApp message, the system first passes the text (or transcribed audio) through an adapted LLM framework. Relying on the empirical breakthroughs of the Lughu-Llama model (Buzaaba et al., 2025), this module is specifically fine-tuned on the AfriMMLU benchmarks. It accurately parses the complex sociolinguistic blend of Hausa, Pidgin English, and standard English, ensuring that localised idioms and cultural colloquialisms are not misinterpreted by the machine.
- **Module 2: Rhetorical Framing Diagnostics:** Once the linguistic content is parsed, the AI applies a diagnostic layer based on Entman’s (1993) framing theory. The system scans the text to identify the active manipulation strategy: Is it diagnosing a false cause? Is it prescribing violent remediation? By classifying the message into one of the WERD typologies (Table 1), the AI prepares to expose the specific psychological trick being played on the user.
- **Module 3: Theological Exegesis Database Cross-Referencing:** If the message contains scriptural references, theological terminology, or religious justifications for conflict, the system queries the VERITAS Vetted Exegesis Database. This database, populated through the transdisciplinary efforts of Christian and Islamic scholars, houses the accurate historical, cultural, and theological

context of commonly weaponised verses.

- **Module 4: TI-RLHF Synthesis and Output Generation:** Utilising Theologically-Informed Reinforcement Learning from Human Feedback (Ouyang et al., 2022), the AI synthesises the data from the previous modules into a coherent, culturally sensitive, and empathetic response. The system ensures the output tone is non-condescending, respectful of religious devotion, and strictly aligned with the peacebuilding values embedded during the co-design process.

Deliverable 3: The Theological Counter-Narrative Matrix

A critical failure point of traditional digital governance is the over-reliance on secular fact-checking mechanisms in deeply religious societies. Empirical research by Armand et al. (2024) demonstrates that when individuals are presented with corrective information during periods of ethno-religious tension, the cognitive acceptance of that correction is heavily mediated by the "religious proximity" of the source. If a secular algorithm dismisses a religiously framed grievance as "False Data," it often triggers a backfire effect, hardening the user's cognitive resistance and reinforcing the narrative that the secular state (or Western tech company) is allied against their faith.

To circumvent this, VERITAS relies entirely on religious proximity. The intervention does not attack the religion; it champions orthodox, peace-oriented theology over extremist distortion. This approach is empirically validated by Bélanger et al. (2023), whose randomised control trials in the MENA region proved that highly tailored, culturally resonant counter-narratives significantly reduce the cognitive attractiveness of extremist ideologies compared to generic counter-messaging.

Table 2: The Theological Counter-Narrative Matrix (Conceptual VSD Output)

Weaponised Scriptural Trope / Out-of-Context Verse	Exegetical Reality (Accurate Context & Peacebuilding Hermeneutics)	VERITAS De-escalation Output (AI Generated Response Blueprint)
"Show them no mercy, for they are the enemies of God." (Often generalised to target contemporary ethnic neighbours).	Context: Historically situated in ancient, specific geopolitical warfare and not a universal, perpetual mandate for inter-communal relations in a modern republic.	Scriptural Analysis: "This verse refers to a specific historical conflict over 3,000 years ago. Mainstream Christian and Islamic scholars agree this is not a mandate for violence against neighbours today."
"They plan to deceive you, strike them first." (Used to justify preemptive violence against perceived encroachment).	Context: Both major religions in Nigeria prioritise principles of justice, due process, and the sanctity of life. Preemptive violence based on rumour strictly violates orthodox theological jurisprudence.	Rhetorical Tactic: "This message uses a 'Threat Perception' frame to create panic. Spreading this unverified rumour goes against scriptural mandates for truthfulness and peace."

"God has given us this land, defend it with blood." (Ethno-nationalism disguised as divine right).	Context: A fusion of modern political land disputes with ancient theocratic promises. The orthodox theological focus is on stewardship, communal harmony, and reliance on civic justice systems.	Constructive Dialogue Prompt: "Instead of forwarding, consider replying: 'I checked this claim and it lacks evidence. Let us pray for peace and wait for official community leaders to verify before we react in anger.'"
--	--	---

The psychological engine powering this matrix is Inoculation Theory (Compton et al., 2021). By preemptively (or therapeutically) exposing users to the weakened form of a deceptive strategy and explaining exactly how the trick works, the AI builds durable cognitive immunity. When VERITAS outputs the "Rhetorical Tactic" analysis, it is actively revealing the architecture of the deception. According to Roozenbeek and Van der Linden (2019), this methodology ensures that users do not just discard the single piece of fake news in front of them; they develop a generalised psychological resistance, empowering them to spot and neutralise future disinformation campaigns independently.

Deliverable 4: The Digital Firebreak Concept

The ultimate sociotechnical objective of the VERITAS architecture is to fundamentally alter network behaviour on encrypted platforms. Traditional social media research focuses heavily on public feeds, yet Ventura et al. (2025)

empirically demonstrated through their WhatsApp deactivation experiment in Brazil that in the Global South, closed, encrypted messaging applications are the primary, most potent vectors for political and social misinformation. Because end-to-end encryption renders automated, top-down algorithmic moderation impossible without violating user privacy, the only viable intervention must occur at the level of individual user agency.

By equipping individual users with the analytical output derived from the Theological Matrix (Table 2), the VERITAS system enables ordinary citizens to perform "social corrections" within their intimate networks. Rossini et al. (2021) established that because WhatsApp networks are comprised of strong ties (family, religious groups, close friends), they provide a uniquely safe environment for constructive pushback, provided the user has the rhetorical tools to do so without destroying the relationship. VERITAS provides these precise tools.

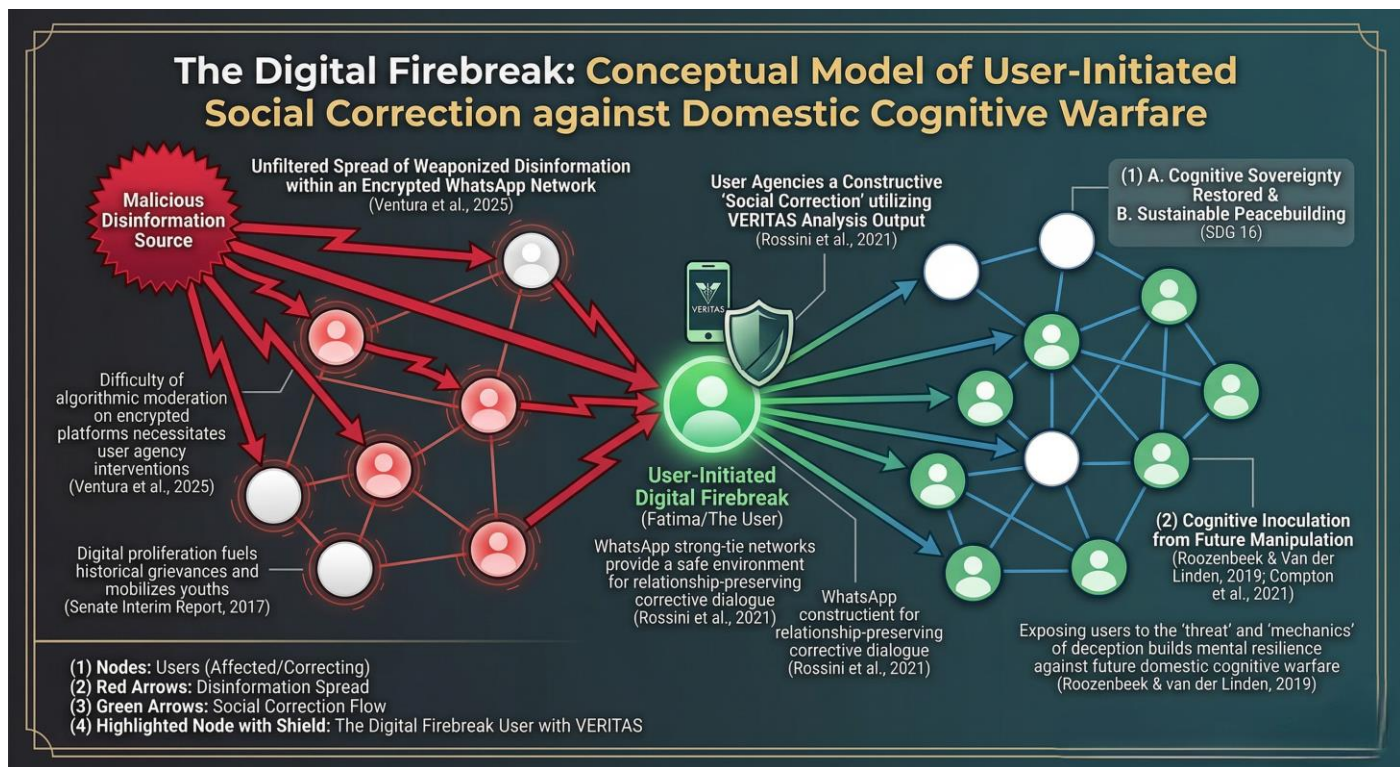


Figure 2: The Digital Firebreak: Conceptual Model of User-Initiated Social Correction against Domestic Cognitive Warfare. Demonstrates how individual user agency intercepts malicious nodes and injects de-escalating, inoculated narratives back into the encrypted network, referencing Ventura et al., 2025.

The concept of the Digital Firebreak represents the synthesis of the entire project. When a piece of highly contagious, weaponised disinformation enters a WhatsApp

group, it spreads exponentially from node to node. However, when a single node (a user) utilises VERITAS, they act as a firewall. Armed with a culturally respectful,

theologically verified response, they intercept the malicious narrative, preventing its onward transmission, and simultaneously inject a de-escalating truth back into the network, thereby "inoculating" the other members of the group.

V. Discussion and Summary

Operationalizing the Digital Firebreak via Social Corrections

The theoretical frameworks and architectural matrices constructed in this proposal underscore a critical, necessary paradigm shift in the field of digital peacebuilding: the movement away from the centralised, automated censorship models of the West toward decentralised, empowered user agency models optimised for the Global South.

As visualised in the Digital Firebreak model, the operational success of VERITAS hinges entirely on its ability to seamlessly facilitate "social corrections" within complex cultural hierarchies.

In the sociopolitical context of Northern Nigeria, directly contradicting an elder, a family patriarch, or a religious leader in a communal WhatsApp group carries significant social risk. The barrier to entry for debunking misinformation is not merely a lack of facts; it is the fear of committing a cultural faux pas that could fracture strong-tie networks. Rossini et al. (2021) noted that the intimate nature of these networks makes them ideal for correction, but the execution requires extreme tact. VERITAS bridges this gap by providing an authoritative, third-party, theologically sound script. When a user utilises the AI to analyse an inflammatory message forwarded by a respected elder, she is no longer forced to choose between passive complicity (silence, which implies consent) and aggressive, disrespectful confrontation.

The VERITAS output provides her with a culturally calibrated mechanism for pushback. By framing the correction as a shared inquiry ("I used a tool to check this scriptural quote... let us wait for official verification"), the user de-escalates the panic while preserving the social hierarchy and dignity of the original sender. The AI absorbs the burden of authoritative correction, allowing the user to act as an empathetic conduit for peace. This operationalises the abstract concept of digital peacebuilding, turning everyday citizens into active, resilient defenders of their community's cognitive sovereignty.

Ethical AI Sovereignty and Scalability

A vital policy discussion emerging from this blueprint concerns the imperative of digital sovereignty and the ethical deployment of artificial intelligence in post-colonial contexts. The uncritical deployment of Western-trained

foundation models in African conflict zones presents a profound sociotechnical risk. These models, trained predominantly on North American and European datasets, suffer from what Gomez et al. (2024) term "predictive multiplicity" and the emergence of an "algorithmic leviathan." When applied to the deeply nuanced, multi-lingual, and historically burdened context of Kaduna State, these generic algorithms frequently act with algorithmic arbitrariness: misunderstanding local dialects, misinterpreting theological expressions, and disproportionately censoring marginalised voices while allowing sophisticated local dog-whistles to pass undetected.

The VERITAS framework explicitly mitigates this risk by establishing a blueprint for true Ethical AI Sovereignty. By mandating that the core training data, the Theological Counter-Narrative Matrix, be curated strictly by a Vetted Scholar Co-Design Council comprised of local Nigerian academics, sociologists, and religious clerics, the project aligns with the highest standards of transdisciplinary co-design (Sadek et al., 2023). Maintaining the theological database and the Reinforcement Learning from Human Feedback (RLHF) loop within the jurisdiction of Nigerian institutions ensures that the AI remains a transparent, ethical tool of the state and its people, rather than an opaque, culturally detached instrument of foreign technology conglomerates.

Furthermore, this localised, decoupled architecture guarantees massive scalability. While the current blueprint specifically maps the Christian-Islamic sociopolitical dynamics of Kaduna State, the underlying TI-RLHF framework is entirely modular. It can be seamlessly adapted to address localised conflicts in the Middle Belt, the South-East, or other regions of the Global South by simply swapping the underlying theological and sociological datasets. The architecture remains constant; only the cultural contextualisation layer requires localisation.

VI. Recommendations

To transition this desk-based theoretical blueprint into a functional, deployed sociotechnical intervention, the following strategic recommendations must be explicitly adopted by policymakers, the KASU Directorate of Research, and allied peacebuilding institutions:

- **Institutionalise the Transdisciplinary Co-Design Council:** Kaduna State University, in partnership with the State Peace Commission, must formally institutionalise a permanent, vetted council of Christian and Islamic theologians, sociologists, and linguists. This body will be solely responsible for the continuous curation, review, and authorisation of the hermeneutical

datasets used to train the VERITAS model.

- **Phased Technical Development of the TI-RLHF Model:** Technology partners must strictly adhere to the Value Sensitive Design methodology mapped in this proposal. Development should proceed in distinct phases: first, adapting an open-source, low-resource LLM (such as Llama-3) utilising the AfriMMLU benchmarks; second, integrating the framing diagnostics; and third, applying the TI-RLHF training.
- **Integration with State Peacebuilding Architectures:** The VERITAS deployment should not exist in a vacuum. It must be integrated into the broader strategic initiatives of the Kaduna State Peace Commission. Anonymised, aggregate data generated by user queries should be utilised as an early-warning radar system.
- **Execute Comprehensive Digital Literacy and Inoculation Campaigns:** Simultaneous with the software development, civil society organisations must roll out educational campaigns focused on "Psychological Inoculation." Communities must be taught the basic mechanics of how cognitive warfare operates on WhatsApp.
- **Establish a Sovereign Data Enclave for Religious Text Processing:** To protect against data colonialism and algorithmic bias, all processing of sensitive ethno-religious data, user inputs, and theological matrices must be hosted within a secure, sovereign data enclave geographically located within Nigeria.

VII. Conclusion

The proliferation of weaponised ethno-religious disinformation demands interventions that transcend traditional algorithmic censorship. Project VERITAS operationalises a critical paradigm shift, moving digital peacebuilding from reactive fact-checking to proactive cognitive inoculation within encrypted networks. By grounding artificial intelligence in Value Sensitive Design and local theological hermeneutics, this blueprint provides a scalable, sovereign solution to the crisis of domestic cognitive warfare in Kaduna State. The proposed system architectures, disinformation typologies, and counter-narrative matrices equip policymakers and developers with the requisite frameworks to build AI tools that respect cultural nuances and religious proximity. Ultimately, addressing the complexities of the twenty-first century requires transdisciplinary collaboration. Implementing the numbered recommendations herein will empower individual users to serve as digital firebreaks, fostering sustainable social cohesion, protecting democratic

discourse, and actively advancing the objectives of Sustainable Development Goal 16 across Nigeria and the broader Global South.

References

1. Armand, A., Augsburg, B., Bancalari, A., & Kameshwara, K.K. (2024). Religious proximity and misinformation: Experimental evidence from a mobile phone-based campaign in India. *Journal of Health Economics*, 96, 102883. <https://doi.org/10.1016/j.jhealeco.2024.102883>
2. Bélanger, J. J., Snook, D. W., Dzitac, D., & Cheppih, A. (2023). Challenging extremism: A randomized control trial examining the impact of counternarratives in the Middle East and North Africa. *Current Research in Ecological and Social Psychology*, 4, 100097. <https://doi.org/10.1016/j.cresp.2023.100097>
3. Buzaaba, H., Wettig, A., Adelani, D. I., & Fellbaum, C. (2025). Lugh-Llama: Adapting Large Language Models for African Languages. *arXiv preprint arXiv:2504.06536*.
4. Compton, J., van der Linden, S., Cook, J., & Basol, M. (2021). Inoculation theory in the post-truth era: Extant findings and new frontiers for contested science, misinformation, and conspiracy theories. *Social and Personality Psychology Compass*, 15(6), e12602. <https://doi.org/10.1111/spc3.12602>
5. Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4), 51–58. <https://doi.org/10.1111/j.1460-2466.1993.tb01304.x>
6. Fawzi, M., Ross, B., & Magdy, W. (2025). Fabricating Holiness: Characterizing Religious Misinformation Circulators on Arabic Social Media. *arXiv preprint arXiv:2508.07845*.
7. Friedman, B., Kahn, P. H. Jr., & Borning, A. (2004). Value Sensitive Design and Information Systems. In P. Zhang & D. Galletta (Eds.), *Human-Computer Interaction in Management Information Systems: Foundations*. M.E. Sharpe.
8. Gomez, J. F., Machado, C. V., Paes, L. M., & Calmon, F. P. (2024). Algorithmic Arbitrariness in Content Moderation. *arXiv preprint arXiv:2402.16979*.
9. Hirblinger, A. T., Hansen, J. M., Hoelscher, K., Kolås, Å., Lidén, K., & Martins, B. O. (2023). Digital Peacebuilding: A Framework for Critical-Reflexive Engagement. *International Studies Perspectives*, 24, 265-284. <https://doi.org/10.1093/isp/ekac015>

10. Miller, S. (2023). Cognitive warfare: an ethical analysis. *Ethics and Information Technology*, 25(46). <https://doi.org/10.1007/s10676-023-09717-7>
11. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Aspell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730-27744.
12. Roozenbeek, J., & van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Palgrave Communications*, 5(1), 65. <https://doi.org/10.1057/s41599-019-0279-9>
13. Rossini, P., Stromer-Galley, J., Baptista, E. A., & de Oliveira, V. V. (2021). Dysfunctional information sharing on WhatsApp and Facebook: The role of political talk, cross-cutting exposure and social corrections. *New Media & Society*, 23(8), 2430-2451. <https://doi.org/10.1177/1461444820928059>
14. Sadek, M., Calvo, R. A., & Mougenot, C. (2023). Co-designing conversational agents: A comprehensive review and recommendations for best practices. *Design Studies*, 89, 101230. <https://doi.org/10.1016/j.destud.2023.101230>
15. Sanfilippo, M. R., Zhu, X. A., & Yang, S. (2025). Sociotechnical governance of misinformation: An Annual Review of Information Science and Technology (ARIST) paper. *Journal of the Association for Information Science and Technology*, 76(1), 289-325. <https://doi.org/10.1002/asi.24953>
16. Senate Ad Hoc Committee. (2017). Interim Report of the Senate Ad Hoc Committee on Southern Kaduna Crises and Other Parts of the Country (Volume I). National Assembly, Federal Republic of Nigeria.
17. Ventura, T., Majumdar, R., Nagler, J., & Tucker, J. A. (2025). Misinformation Beyond Traditional Feeds: Evidence from a WhatsApp Deactivation Experiment in Brazil. *Social Science Research Network (SSRN)*.