Information Security of Territorial Communities Facing Deepfake Threats

Yaryna Mulska

Institute of Computer Science and Information Technologies,
UKRAINE, Lviv, 28a Bandery Street, E-mail: yaryna.mulska@gmail.com

Abstract – This paper develops a framework for protecting municipal crisis communication against deepfake impersonation. It combines source verification, independent confirmation, accessible correction and incident review. Its contribution is a decision protocol linking message authenticity to service continuity, supported by auditable readiness indicators and a proposed evaluation design.

Keywords – information security, deepfakes, territorial communities, crisis communication, information integrity, institutional trust, service continuity.

Introduction

The use of digital communication by local authorities creates a practical information security problem: residents must be able to distinguish an authorised municipal instruction from a convincing imitation. Deepfake audio and video can reproduce the apparent identity of a public official, allowing fabricated messages to draw credibility from a familiar voice, appearance or institutional role. A false evacuation announcement, an imitation request for an urgent payment or a fabricated notice about humanitarian assistance could influence decisions before its origin is verified. The resulting risks concern physical safety, personal data, financial resources and access to essential services.

For Ukrainian territorial communities operating under wartime conditions, this problem deserves particular attention. Urgent decisions, disrupted connectivity and reliance on several communication platforms can complicate verification. An instruction may reach residents as a forwarded recording or screenshot, detached from its original publication context. Under these circumstances, the perceived authenticity of the speaker may substitute for confirmation that the responsible authority issued the message. Delayed clarification can leave residents exposed to conflicting instructions, while an unsupported denial can undermine confidence in a genuine warning.

The OECD's information integrity framework connects reliable information environments with transparency, societal resilience and institutional capacity [1]. NIST publications provide complementary approaches to generative AI risk management and synthetic content transparency [2-3]. Applying these approaches to municipal communication requires an explicit connection between technical evidence, administrative responsibility and public action. A detection result must be interpreted, a disputed instruction must be checked by an authorised office, and residents must retain access to reliable guidance while verification proceeds. The analytical problem addressed in this paper is how to organise that connection within the staffing and resource constraints of local government. Content detection alone does not determine who may confirm a municipal instruction, which communication route remains trustworthy after an account compromise, or how services should continue during uncertainty. Conversely, publication through an official account cannot by itself resolve doubts if that account has been compromised. Verification therefore requires procedures that distinguish the authenticity of the media, the integrity of the distribution channel and the authority behind the instruction.

I. Analytical approach and scope

NIST's Generative AI Profile provides a voluntary basis for considering trustworthiness throughout the AI lifecycle [2]. Its relevance here is to frame synthetic media as an organisational risk requiring assigned responsibility. The proposed municipal unit of analysis is a consequential

message: a communication that can alter physical movement, access to a service, disclosure of personal data or a financial decision. Ordinary political disagreement and criticism fall outside the incident definition unless accompanied by demonstrable impersonation or falsification.

Three working propositions guide the framework. First, independent confirmation of a message's origin should reduce successful impersonation relative to reliance on audiovisual appearance. Second, a correction accompanied by clear service instructions should reduce disruption more effectively than a denial alone. Third, redundant communication channels should improve access to verified instructions during a platform or power outage. These propositions remain untested in this paper and define the direction of subsequent empirical research.

II. Threat mechanisms and verification architecture

The proposed mechanism links impersonation to perceived authority, behavioural compliance and possible service disruption. Urgency and difficulty of verification may strengthen this sequence. A manipulated video falsifies the speaker's appearance, whereas a compromised account distributes instructions through a normally legitimate channel. These mechanisms require different checks.

NIST's synthesis examines provenance, watermarking and detection as approaches to digital content transparency [3]. Building on those categories, this paper proposes using technical checks as supporting evidence within an institutional procedure. A detector's output should not itself authorise an evacuation, payment or disclosure. Likewise, a familiar logo, voice or account name provides insufficient institutional evidence when the account or its recovery process may be compromised. Four functions define the architecture. Identification establishes the authorised issuing office. Confirmation contacts it through a separately maintained route. Continuity maintains access to valid instructions during verification. Review records the incident and addresses procedural weaknesses. Each function requires a responsible employee and a designated deputy.

Before an incident, the municipality should publish a directory of official channels and explain how consequential instructions can be checked. A canonical notice should contain an issue time, responsible office, reference number where appropriate, status and next update time [4]. Table 1 links hypothetical scenarios to consequences, controls and accountable functions. It does not report observed attacks. Functions may be combined within existing posts, provided that independent confirmation remains possible.

Channel redundancy requires distinct failure paths. Reposting an unverified message across several accounts does not provide independent evidence. Confirmation must use a contact route or authorisation process that does not share the suspected point of compromise.

III. Decision protocol and readiness indicators

On receiving a disputed message, the duty officer should record its location, time, claimed issuer and requested action. The immediate priority is to determine whether the instruction could cause imminent harm. Verification should then contact the designated responsible office through a pre-existing directory, rather than a telephone number or link supplied by the disputed message. Relevant evidence should be retained with access limited to staff who need it.

While authenticity remains unresolved, the municipality should state what is known, identify a safe verification route and specify when an update will follow. A provisional notice must distinguish uncertainty about the message from confirmed information about service availability. Where emergency action is required, the authorised emergency procedure takes precedence; residents should not be asked to wait for a forensic assessment of a recording before following independently confirmed safety instructions.

Table 1

Illustrative municipal impersonation scenarios and response controls

Scenarios	Potential consequences	Verification and continuity actions	Accountable functions
Fabricated evacuation video	Unsafe movement or refusal to follow valid guidance	Confirm through the designated emergency authority; publish current instructions on backup channels.	Emergency coordination
Cloned voice requesting a payment	Unauthorised financial transfer	Call a previously registered number; retain dual approval of payment instructions.	Finance and authorisation
False announcement of aid registration	Disclosure of personal data or missed assistance	Check the responsible service and destination domain; publish the verified application route.	Social services
False post on a compromised official account	Conflicting instructions and wider credibility loss	Suspend compromised access; confirm instructions independently and activate the backup notice route.	IT and communication

Once a message is assessed, the public update should identify the valid instruction, its issuing authority and the action residents should take. Confirmed impersonation should also trigger measures appropriate to the mechanism: account recovery, correction of a fraudulent link or notification of the relevant service. A genuine municipal error requires an explicit correction. A suspected fabrication should not be labelled as confirmed solely because an automated tool flags it.

NIST SP 800-61 Revision 3 integrates cybersecurity incident response into broader risk management [4]. The present framework extends that organisational logic to municipal communication: preparation precedes detection, and recovery includes changes to practice. This is an analytical adaptation, not a claim that the NIST publication provides a dedicated municipal deepfake protocol. A short review should examine how the message entered the information environment, which control failed and whether residents could find reliable instructions.

Readiness should be measured through observable records. A publication completeness rate is the number of sampled consequential notices containing all required verification fields divided by the number sampled. Independent confirmation compliance is the proportion of sampled high-consequence instructions checked through the designated separate route before implementation. Both denominators and exclusions must be reported. The proposed measures describe procedural performance; they do not estimate the probability that a particular message is false.

Response performance requires distinct time measures. Verification latency runs from the first recorded report to an authorised authenticity assessment. Correction latency runs from that assessment to publication of the verified update. Separating the intervals prevents a fast but unsupported denial from being scored as success. Median times should be accompanied by the slowest cases, incident counts and the proportion left unresolved. No universal target is proposed because urgency, staffing and service conditions differ.

An accessibility test should ask whether residents can obtain the same verified instruction through an online channel and an alternative route, such as a telephone service or established community intermediary. Tests should include people with limited connectivity or digital skills. Communication reach should be distinguished from understanding: publication counts alone cannot establish

that residents identified the valid instruction or knew what to do next.

IV. Evaluation design and implementation limits

A pilot should audit channel ownership, contact directories and recent notices, followed by standardised exercises before and after training. Scenarios should include fabricated media, a compromised account and a genuine urgent message to detect excessive suspicion. Exercises must remain separate from real public communication.

Outcomes should include source identification accuracy, independent confirmation, decision time and service continuity. A comparison group or staggered introduction would help distinguish training from practice effects. Analysis should consider staffing, population, connectivity and scenario difficulty. Trust effects require a separate resident survey.

Implementation can begin with verified contact lists, named deputies, consistent notice fields and rehearsed backup communication. Regional shared support could assist smaller communities with analysis and exercises. Its costs, governance and effectiveness require separate assessment.

The framework has clear limits. It has not been validated against municipal incident data, does not quantify deepfake prevalence and cannot establish causal effects on trust or resilience. Strong source authentication also cannot guarantee that an authorised statement is factually correct. Public correction, service-level checking and independent scrutiny remain necessary. Incident records should minimise personal data, while public explanations should avoid exposing sensitive operational details.

Conclusion

The framework defines municipal information security through authentication of consequential instructions and continuity of reliable services. It contributes a protocol with assigned responsibilities, independent confirmation, accessible updates and auditable indicators. Controlled exercises and municipal pilots should test whether these procedures improve decision accuracy and service continuity without delaying genuine warnings.

References

[1] OECD. (2024). *Facts not fakes: Tackling disinformation, strengthening information integrity*. OECD Publishing. https://doi.org/10.1787/d909ff7a-en

[2] Autio, C., Schwartz, R., Dunietz, J., Jain, S., Stanley, M., Tabassi, E., Hall, P., & Roberts, K. (2024). *Artificial intelligence risk management framework: Generative artificial intelligence profile* (NIST AI 600-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.600-1

[3] Chandra, B., Dunietz, J., Roberts, K., Lee, Y., Fontana, P., & Awad, G. (2024). *Reducing risks posed by synthetic content: An overview of technical approaches to digital content transparency* (NIST AI 100-4). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-4

[4] Nelson, A., Rekhi, S., Souppaya, M., & Scarfone, K. (2025). *Incident response recommendations and considerations for cybersecurity risk management: A CSF 2.0 community profile* (NIST SP 800-61r3). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-61r3