COUNTERING AI & DEEPFAKE SMISHING AND VISHING THREATS

PREPARED BY POLYGRAF.AI
IN ASSOCIATION WITH NOBLE

POLYGRAF NOBLE

This Whitepaper was put together by Polygraf AI in Response to FBI Alert Number I-051525-PSA (May 15, 2025)

Polygraf Inc. operating under Polygraf AI trade name is an Austin, TX based Artificial Intelligence solution provider to public and private organizations with critical operations.

The company has built a suite of Small Language Model AI solutions that detect AI threats, protect organizational and individual policies and mitigate (AI & Data) risks before they become problems.

Polygraf AI was established in 2023 in Austin, TX by Yagub Rahimov and Vignesh Karumbaya with 6 AI/ML PhD candidates and 3 competitive hackers to combat rising (digital) fraud cases and diminishing privacy challenges.

Distributed in partnership with NOBLE Supply & Logistics | www.noble.com/CI NOBLE
For further information, please contact Edward M. Levy | +1 (203)-6561 | elevy@noble.com

Table of Contents

Background: Escalating AI Impersonation Threats

On May 15, 2025 the Federal Bureau of Investigation (FBI) sounded the alarm on a **coordinated impersonation campaign** targeting senior U.S. officials and their contacts.

According to the announcement, since April 2025, cybercriminals have been **masquerading as high-ranking (FBI) officials** via **text messages and AI-generated** (deepfake) **voice calls** in order to gain unauthorized access to personal and official accounts. This tactic, which the FBI identifies as *smishing* (SMS phishing) and *vishing* (voice phishing), leverages **sophisticated and social engineered AI content** to deceive victims. Phishing isn't something new, but with Smishing and Vishing, attackers often initiate contact posing as a known official, move on **building rapport**, then send a malicious link or request sensitive information under the claim of moving the conversation to a secure platform, that is when things get not so secure any more. By exploiting trust in familiar identities, the adversaries aim to steal credentials or hijack accounts, which can then be used to target additional officials, colleagues, and even family members.

Smishing attacks typically involve text messages from spoofed or software-generated phone numbers, impersonating colleagues or family to appear legitimate. On the other hand **Vishing** attacks have become extra sophisticated with the use of **AI-generated voice cloning also known as deepfake voice cloning.** With these solutions on hand, criminals now create convincing voice messages that mimic the speech patterns of well-known officials or personal contacts. The FBI alert warns that malicious actors increasingly exploit AI audio to **impersonate public figures or trusted relations**. This new wave of digital fraud and threat mark a dangerous advance in social engineering, as the fraudulent messages and voices are far more realistic than the spam emails and robocalls of the past.

Federal authorities stress that **no incoming message or call should be assumed authentic** merely based on caller ID or familiar names. In fact, the FBI's Public Service Announcement explicitly cautions: if you receive unsolicited communication purportedly from a senior official, be skeptical and verify independently.

On the other hand, the challenge, as the FBI notes, is that *"AI-generated content has advanced to the point that it is often difficult to identify"* to an ordinary person. Traditional warning signs can be subtle or absent - for example, an AI phone call may *sound* nearly identical to a person, and a carefully crafted phishing text may contain no obvious errors. This puts even vigilant personnel at risk. The **implications for national security and critical infrastructure are severe**, as successful intrusions could expose sensitive communications or enable further supply-chain attacks through compromised contacts. In summary, the FBI alert underscores a clear and present danger: **AI-powered impersonation attacks are surging**, and new defenses are required to detect and stop them at the earliest possible stage.

Federal Guidance and the Need for Polygraph for Content

U.S. cyber authorities are emphasizing **heightened vigilance and multi-layered safeguards** towards these threats. The FBI's latest alert offers common-sense advice to **identify fake messages**, such as verifying the caller's identity via known contact channels, scrutinizing message details (e.g. slight misspellings or off-tone language), and watching for telltale glitches in images (e.g. odd appearances or items), audio lag or unnecessary pauses, or unnatural speech patterns. Users are urged to refrain from clicking unsolicited links or sharing sensitive information unless the sender's authenticity is confirmed through an independent source.

The Cybersecurity and Infrastructure Security Agency (CISA) has also published guidance to **"stop the attack cycle at phase one,"** reinforcing the importance of blocking phishing attempts before they escalate. CISA and FBI both recommend measures like continuous security awareness training (*"Teach Employees to Avoid Phishing"*), strict multi-factor authentication, and prompt reporting of any suspected scam to organizational security teams and law enforcement.

These recommendations are critical, yet **they place a heavy burden on individuals** to discern truth from deception in real time. As the FBI acknowledged, today's AI-driven spoofs can be **extremely convincing** - effectively *outsmarting human perception*. Even well-trained staff could struggle to notice the subtle imperfections of a deepfake voice or the slight anomalies in a spoofed text. Traditional tools are increasingly falling behind with the rise of AI. Legacy phishing filters and caller ID systems were not built to analyze the content of a voice or the linguistic nuances of a one-on-one text from a spoofed number. In short, **the sophistication of AI-generated impersonation demands an equally sophisticated detection capability**.

Federal cybersecurity leaders and CISOs in critical infrastructure are thus seeking **advanced solutions that can automatically authenticate communications** with speed and accuracy, providing a safety net beyond what human recipients alone can offer. The ideal solution must operate in **real time**, given the rapid nature of social engineering attacks, and must deliver **extremely precise judgments** (false alarms or missed detections are costly in high-stakes environments). Moreover, any detection system should provide **explainable outputs** that analysts can audit. Explainability is an especially crucial requirement in government settings, where trust and accountability in AI decisions are mandated. And as a final point, these solutions are mandated to operate in local architecture, not leaking data outside the necessary premises. It is against this backdrop that **Polygraf AI** has developed its response, aligning with federal guidance and filling critical gaps in traditional defenses.

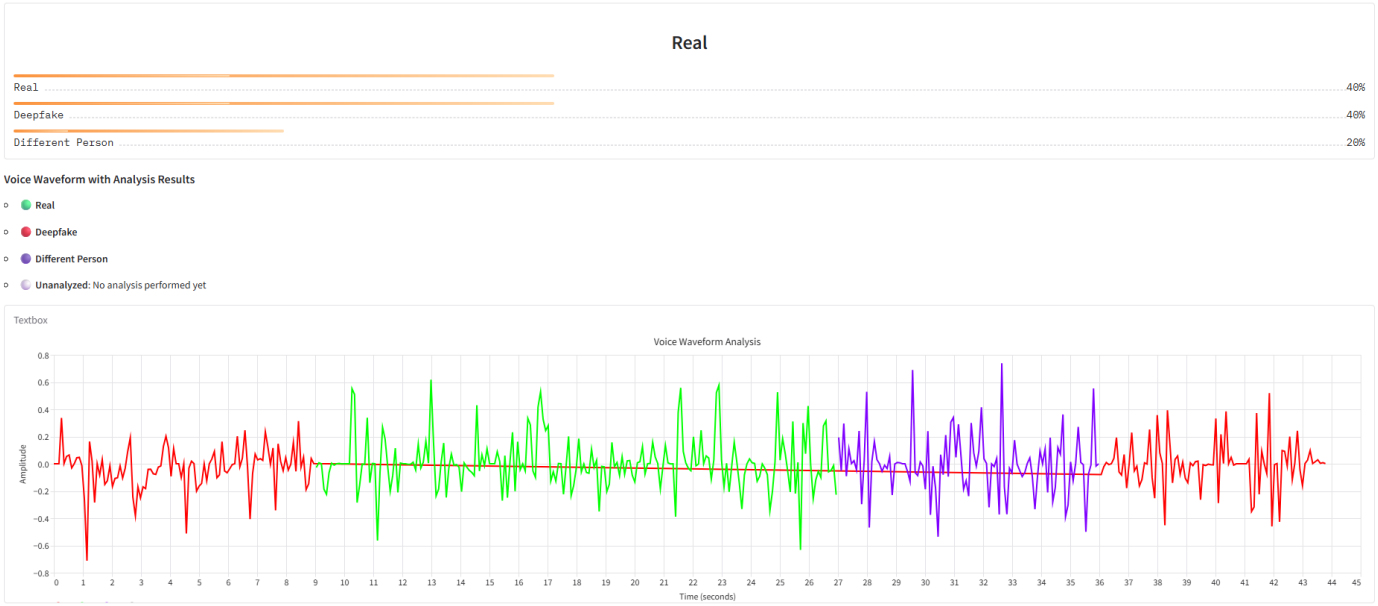
Polygraf AI - Locally deployed, explainable, real-time AI Groundtruth

Polygraf AI is an award-winning AI governance and data integrity solution designed to **verify the authenticity of digital content across text, and audio,** while protecting sensitive data locally. In March 2025, Polygraf's technology earned broad recognition by sweeping multiple categories (including Best Enterprise Solution, Smart Data, Future of Work, and the coveted Best in Show) at SXSW 2025 - further validating its innovation and effectiveness in countering AI-driven and deepfake threats. Polygraf's proprietary multilayer AI engine has been **independently validated as the most accurate in the world** for detecting AI-generated and manipulated content. In fact, when it comes to AI content, just accuracy isn't enough, false positives are equally damaging, Polygraf AI Groundtruth during its latest independent analysis showed highest accuracy and lowest false positive rate in distinguishing human vs. AI-generated material, outperforming every other product in the market. Not only this, Polygraf AI Groundtruth also identifies the AI models used in text content and the context of the deepfake audio calls. This level of precision and depth is critical when dealing with nation-state-level impersonation tactics and the extra depth analysis enables the law enforcement to build guardrails and identify impersonators. Equally important, Polygraf's system is built with **transparency and explainability** at its core - every detection comes with contextual evidence and confidence scoring, so security teams can **understand and trust the decision rationale**.

Two core modules of the Polygraf AI suite directly address the smishing/vishing threat highlighted by the FBI: **VeXon** and **Groundtruth**. Together, these modules deliver unmatched speed, precision, and insight in identifying impersonation attempts:

- **Polygraf VeXon - Real-Time Voice Echo Analysis:** VeXon is Polygraf's cutting-edge voice authentication module, purpose-built to **detect AI-generated or manipulated voices** in phone calls, voice messages, and audio recordings. VeXon is a P2P2AI model, analyzing the voice, matching it to the original voice, then comparing it to synthetic solutions. Using advanced voice biometrics and spectral analysis, VeXon determines within **just a few seconds** whether an incoming voice matches the true identity or is an AI-generated *deepfake*. In a live demonstration, VeXon identified a deepfake voice impersonating a senior official within **5-10 seconds**. This speed is crucial as most of the solutions in the market require 45-90 second minimum input. Therefore, VeXon enables organizations to automatically **screen voice calls in near real time** and disrupt vishing attempts before any trust is gained or information is exchanged. VeXon analyzes subtle features of speech (tone, cadence, acoustic artifacts) that are imperceptible to human ears yet often present in synthesized audio. If an imposter is using an AI voice model of a well-known figure, VeXon will flag the anomaly with a high-confidence alert, allowing the call to be safely terminated or further verified through alternate channels. By providing instantaneous "caller authenticity" scoring, VeXon essentially acts as a **voice polygraph** for high-stakes communications - a vital tool for law enforcement cyber units and executives who are frequent targets of voice spoofing.

Figure 1: Polygraf AI VeXon real-time analytics capabilities



- **Polygraf Groundtruth - Authenticity Analysis for Text & Multimedia:** Groundtruth is the module responsible for establishing the **veracity of content** in digital communications. It operates across text content and the context of other media to detect signs of AI generation, manipulation, or inconsistency with known reality. In the context of smishing, Polygraf AI Groundtruth analyzes incoming SMS or messaging content to determine if the language and context are likely genuine. For example, if a text claims to be from a certain official, Groundtruth can cross-verify known communication patterns (writing style, communication metadata) and use AI detection algorithms to flag language that appears machine-generated or suspiciously "off." Polygraf's underlying AI detectors - which are continually updated via meta-learning - can identify the telltale statistical patterns of AI-written text with world-leading accuracy. Groundtruth also integrates contextual threat intelligence: malicious links or known phishing phrases embedded in a message are recognized and highlighted. If an attacker includes a photo or profile image, Groundtruth can check for deepfake indicators (such as facial irregularities or digital artifacts). With this approach Polygraf AI brings **instant authenticity assessment** to any communication. Groundtruth provides a clear visual report (e.g., highlighting suspect portions of text or image anomalies) alongside an overall trust score, giving security personnel an **explainable verdict** on the content's authenticity. In essence, this module serves as the "content truth verifier," ensuring that *what you see (or read) is actually what it purports to be*. As an example, we are using two AI solutions to write a section of this paragraph and let's see if Polygraf will be able to detect it.

Figure 2: Polygraf AI Detecting Claude and OpenAI ChatGPT within the paragraph above

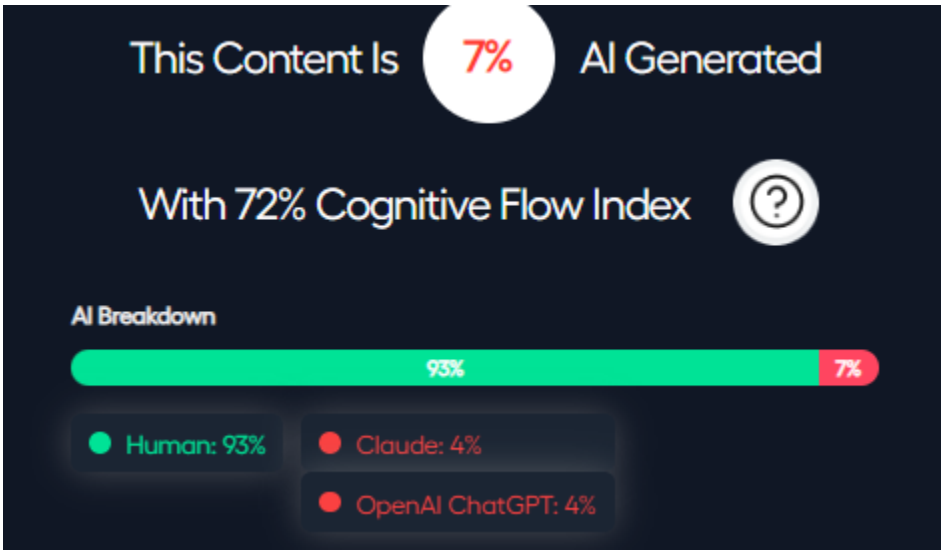
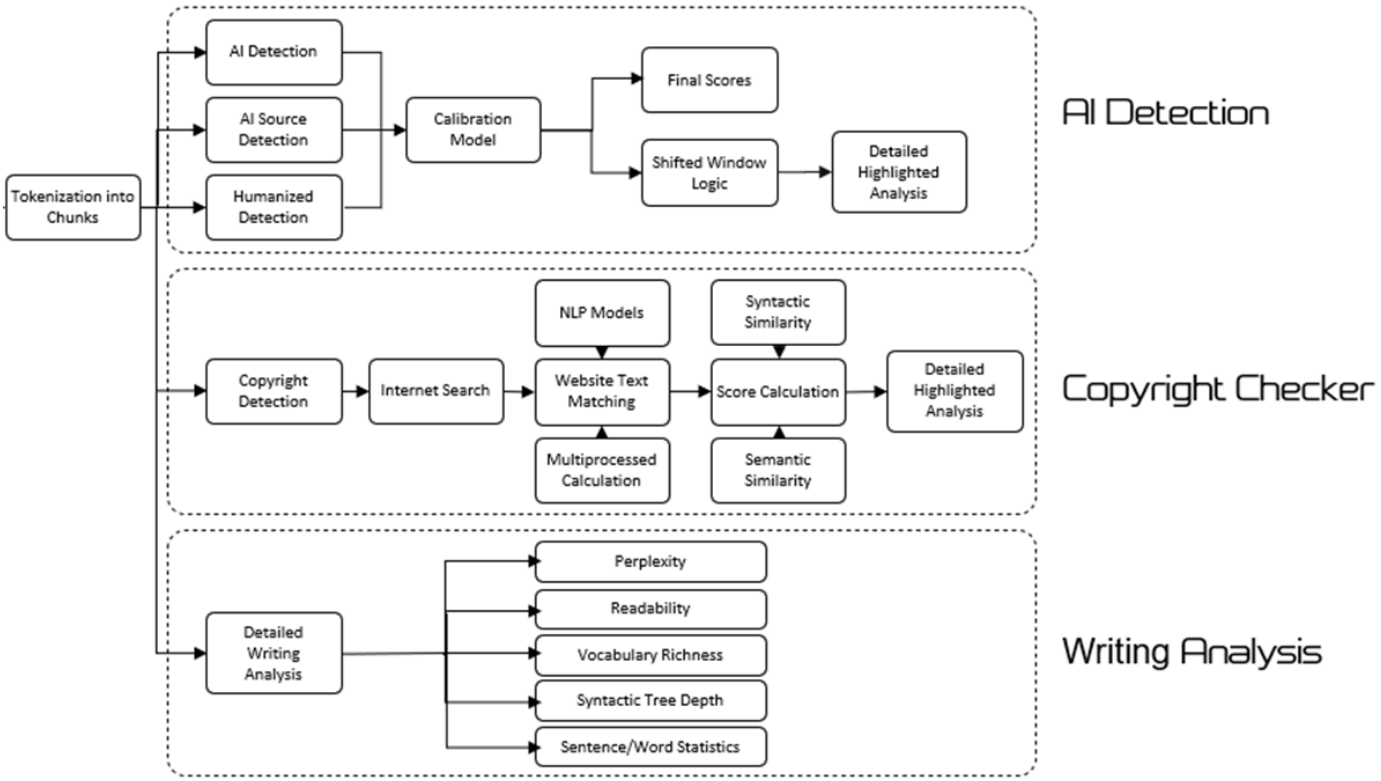


Figure 3: Polygraf AI Detection infrastructure



Mitigating Each Phase of the Attack Cycle

Polygraf AI's integrated approach with VeXon and Groundtruth is designed to **mitigate each phase of a smishing/vishing attack**, aligning with federal "left of boom" strategies (i.e., stopping threats at the earliest point). Below is an overview of how Polygraf addresses the typical stages of an AI-enabled impersonation campaign:

1. **First Contact Detection:** When a text or voice call comes in, Polygraf AI starts monitoring. **Groundtruth scans inbound messages** in real time to identify if an SMS, email, or chat has any indicators of spoofing or AI-generated content. Simultaneously, **VeXon listens to incoming voice calls or voicemails,** instantly comparing the audio to the expected speaker's voice characteristics or to other known human speech patterns, **locally without transmitting any private data outside**. VeXon also transcribes the speech context and looks for signs of AI generated speech context. This proactive content screening acts as the first tripwire. For example, if an impostor posing as "Director X" sends a text, Groundtruth might flag anomalies (e.g. an odd greeting or syntax that doesn't match Director X's usual style) and detect the hallmarks of AI-written text. Similarly, if "Director X" calls with an urgent request, VeXon will analyze the voice; if it's an AI clone rather than the real person, VeXon will detect the synthetic audio artifacts and trigger an alert. **Polygraf AI thus spots the impersonation at the very outset** - before the attacker can gain a foothold. This fulfills CISA's guidance to disrupt phishing at phase one, but with automation and AI rigor that far exceed human ability.

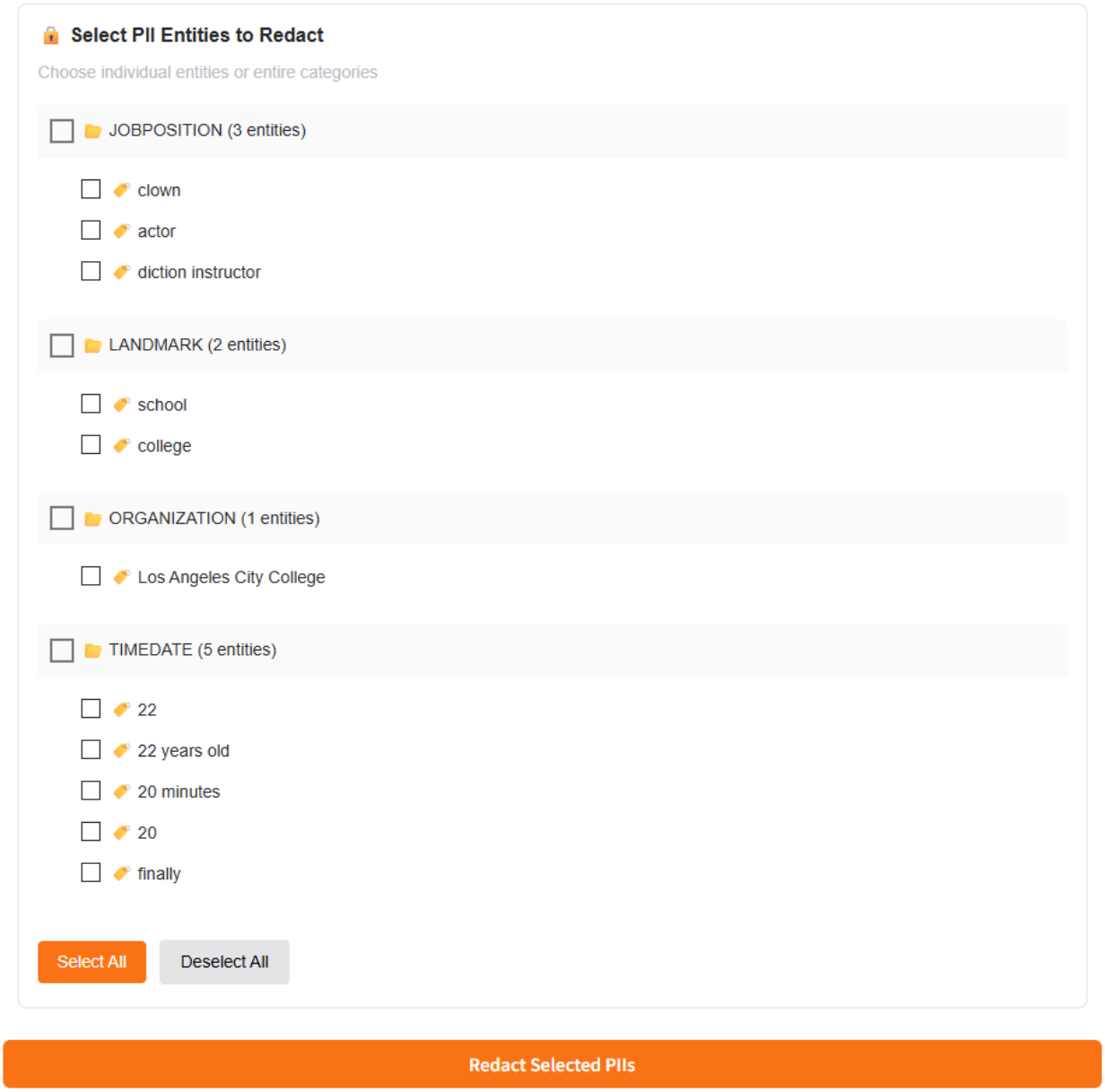
2. **Verification and Containment:** Once Polygraf's sensors detect a likely impersonation, the platform can automatically **contain and escalate** the event, if the system is set up to do so. Security teams and/or the intended victim receive an immediate alert (e.g., "Warning: This message may be fraudulent - sender identity cannot be verified"). In a high-security configuration based on the company policies, Polygraf AI will then also scan through the context of the audio and identify all the identifiers to map out the conversation to predict where the next attack is likely to appear. This prevents the victim from engaging with the attacker under false pretenses. Importantly, Polygraf provides **evidence-based explanations** with each alert. A Polygraf dashboard or report might show, for example, that *"The caller's voiceprint did not match known samples and exhibited 95% similarity to an AI-generated voice model"* or *"The text message was flagged as AI-generated with 99% confidence, using Polygraf's Groundtruth linguistic analysis"*. Such explainability not only supports swift action but also equips investigators to understand the attacker's modus operandi. The FBI recommends independently verifying a caller's identity and examining message details for subtle clues; Polygraf automates and strengthens this verification step by programmatically checking those details against trusted baselines (and doing so **within seconds**, far faster than a manual review).

Figure 4: Polygraf AI Mapping out the speech context for the identifiers

3. **Disruption of Social Engineering Sequence:** By flagging or blocking the fake initial contact, Polygraf breaks the attacker's script before the scam can progress. The FBI observed that these actors try to **transition victims to a secondary platform or link** after initial contact, where malware or credential theft tools await. Polygraf's early intervention thwarts this transition. If a malicious link is present, Groundtruth will identify it, allowing web filters or secure browsers to block the URL proactively. If the attacker attempts to persuade the target to continue the conversation on another app (a common tactic to evade security monitoring), the delay and warning created by Polygraf's alert give the security team time to step in and advise the target. In essence, Polygraf **cuts off the "handoff" to danger**, keeping the interaction on channels that can be supervised or shutting it down entirely. This stage is crucial - it prevents what could have been a successful infiltration from ever reaching that point.

4. **Post-Incident Analysis and Adaptive Learning:** Every impersonation attempt detected by Polygraf AI feeds into a repository of threat intelligence after privileged data has been sanitized off the audio, which happens in real time. The system logs the artifacts of the attack (voice characteristics, language used, any malicious payloads) for forensic analysis. These records help organizational cyber units and law enforcement build cases against threat actors and understand emerging tactics. Because Polygraf's detections are explainable, analysts can quickly **extract the "indicators of impersonation"** and share them as needed (e.g., with other agencies via information-sharing programs). Furthermore, Polygraf's AI continuously **learns from each incident** - its Groundtruth module updates models to recognize new linguistic tricks or deepfake audio improvements that attackers deploy. This adaptive learning means Polygraf only gets more effective over time, staying ahead of attackers even as AI threats evolve. The end result is a virtuous cycle: attempted attacks are not only blocked in the moment, but they also strengthen the defensive posture going forward.

Figure 5: Polygraf AI Enabling the forensic team to clear any privileged data off the audio content before storing it locally.



Through these phased defenses, Polygraf AI provides a **comprehensive shield** against the kind of impersonation attack described in FBI Alert I-051525-PSA. It addresses both the technological challenge (detecting AI-fabricated content) and the human factor challenge (ensuring timely awareness and response) with a solution that is **automated, fast, and reliable**.

Unmatched Speed, Precision and Explainability vs. Traditional Methods

Polygraf AI's capabilities far exceed those of **traditional detection methods.** Running locally within a customer-controlled environment, Polygraf AI offers significant advantages in speed, accuracy, and clarity of analysis. Below, we contrast Polygraf's approach with conventional measures in the context of smishing/vishing threats:

- **Speed of Response:** Traditional verification - such as a user manually calling back the supposed official on a known number, or a security officer reviewing a suspicious message - takes hours, if not weeks, during which an intruder might already be exploiting the trust gained. Polygraf operates at machine speed: VeXon's 5-10-second voice analysis and Groundtruth's instant message parsing mean potential scams are flagged **near-real-time**. This real-time responsiveness can halt an attack **before** any damage is done, a critical factor when one considers how quickly a victim can be socially engineered into clicking a malicious link. In high-tempo attack scenarios, humans alone simply cannot react fast enough; Polygraf's AI augmentation ensures no critical time window is missed.

- **Detection Accuracy:** Conventional phishing filters rely on known bad indicators (blacklisted phone numbers, suspicious link domains, keywords) and can be easily bypassed by novel, targeted attacks that lack those telltales. And expecting busy executives to flawlessly spot AI-crafted deceptions is unrealistic. Polygraf, on the other hand, brings **scientific rigor** to the problem. Its detectors have demonstrated **world-leading accuracy (93-98%)** in identifying highest engineered AI-generated fraud and vishing content. Polygraf's AI protocol has consistently outperformed all competing solutions in industry tests, meaning it catches subtle anomalies that others miss. For example, the platform might notice the almost imperceptible audio glitches or unnatural inflections in a deepfake voice that a person would overlook. This precision gives security teams confidence that when Polygraf issues an alert, it is **very likely correct** - a crucial consideration for avoiding alert fatigue or, worse, a false sense of security from missed threats.

- **Explainability and User Trust:** A hallmark of Polygraf's solution is its commitment to **explainable AI**. In contrast, legacy tools (and certainly a human gut feeling) often provide little insight into *why* something is flagged. Polygraf AI generates a transparent trail for each decision. With this multi-dimensional reporting analysts can see the specific factors that led to a content being labeled inauthentic, or even those SMS messages that were marked unsure, with lower confidence scores. As Polygraf's founder, Yagub Rahimov, has emphasized, embedding transparency and explainability in AI systems allows users to **question and scrutinize decisions**, ultimately building trust in the technology. For federal use, this is indispensable - explainability aligns with AI ethics guidelines and facilitates adoption by letting stakeholders audit the system's performance. Polygraf effectively serves as a **forensic partner** to investigators, not a black-box. Its clear, evidence-backed

reports on each incident can be used in internal investigations or even as supporting material in legal action against perpetrators.

- **Adaptability and Integration:** Traditional defenses often operate in silos (separate email filters, phone call policies, employee training programs) and may not talk to each other. Polygraf offers an integrated platform that can cover multiple communication channels and feed into existing security operations. Deployed on-premises or in a secure cloud per agency requirements, Polygraf AI solutions easily integrates with SOC workflows - for example, it can send alerts to a Security Information and Event Management (SIEM) system, trigger an incident response playbook, or simply notify the target user with a mobile alert. Its **zero-trust design** ensures that even internal communications are verified before trust is granted or controlling which data is being accessed by which type of clearance. Unlike ad-hoc manual checks, Polygraf AI consistently enforces verification policies across the board. Polygraf becomes a force multiplier for existing cybersecurity investments, transforming what used to be a largely manual, human-error-prone process into an automated, consistent shield.

In summary, Polygraf AI delivers a **military-grade solution** to the impersonation threat that law-enforcement is puzzled with. Polygraf AI also meets the stringent requirements of federal cybersecurity (speed, accuracy, explainability, and privacy) and materially improves upon the status quo. By deploying Polygraf's VeXon and Groundtruth modules, agencies and enterprises in critical sectors reduce the risk of falling victim to AI-enhanced social engineering. Instead of operating under purely reactive stance ("hope the user notices something fishy") defenders move to a proactive posture with Polygraf AI. **Instead of the attacker's AI tipping the scales in favor of deception, Polygraf's AI now tips them back in favor of truth. Polygraf AI offers Clarity within Chaos.** ®

Conclusion

The May 15, 2025 FBI alert rightfully brings public attention to the critical failure of reality deception. Threat actors are weaponizing artificial intelligence against most vulnerable operations as well as innocent people. With this whitepaper we have outlined how Polygraf AI solutions - particularly the VeXon deepfake voice detector and Groundtruth content authenticity module - directly counters this emerging threat with unparalleled efficacy.

Polygraf AI solutions **verify identities and content near-real-time** with machine precision and neutralize smishing and vishing attacks right at their inception. Polygraf AI prevents security breaches before they even start. Independent audits confirm Polygraf's highest accuracy with minimal false positives. Current users are already utilizing the platform's in-depth explainability features mitigating risks before they turn into challenges. Combining these features law enforcement cyber units and CISOs alike can now get insights that they never had before. Moreover, Polygraf AI's zero-trust, on-premises deployment model aligns with federal security and data governance standards, ensuring that sensitive information remains protected during the detection process or even beyond.

Polygraf AI offers a timely and powerful solution wherever it is needed. These locally running AI models transform how organizations defend against AI-enabled threats. It operationalizes the best practices urged by the FBI and CISA. From user verification to early-cycle threat disruption Polygraf AI reveals the threats, relieving the impossible burden on servicemen on the frontline

Polygraf AI's VeXon and Groundtruth modules represent a leap forward in protecting nation's critical institutions from impersonation fraud, deepfake attacks, human errors and more. The solution combines speed, precision, and transparency in a manner unmatched by traditional tools. Agencies must adopt Polygraf AI's advanced solutions today to answer the call of the FBI's alert and take a definitive step toward securing their operations against the parabolic growth of AI, deepfake threats.

Sources

- FBI IC3 Alert I-051525-PSA and FBI/CISA guidance;
- Cyber Press news
  Yagub Rahimov Architecture & Governance interview;
- Polygraf AI corporate site.