XPOSER.AI

WHITEPAPER · EDITION 1

Anonymity That Survives Contact With A Voice.

How a voice-first reporting channel can carry the richest form of testimony an organisation can receive without ever learning who is speaking — and where the limits of that promise honestly lie.

Published	September 2026
Edition	1.0
Audience	Compliance, privacy, security and works council reviewers
Author	XPOSER.AI
Contact	info@xposer.ai

How to read this document. It is written to be checked, not believed. Every architectural claim is stated in a form a reviewer can test — by asking for a demonstration, by reading a data processing agreement, or by putting the question to a penetration tester. Section 8 lists the questions we think you should ask us, including the ones we answer least comfortably.

CONTENTS

Section 01

The Problem With Voice

A spoken account is the most complete report an organisation can receive. It is also the one people are most afraid to give.

Written intake forms are safe precisely because they are impoverished. A text box strips away hesitation, emphasis, the order in which someone chooses to tell you things, and the detail a person volunteers when they are not fighting a form. It also strips away the reporter — which is the point. The trade a conventional hotline offers is stark: say less, and stay safer.

Voice reverses both halves of that trade. It captures the account in the form the reporter would naturally give it, in any of the languages they actually think in, at any hour, without the cognitive load of composing a written allegation about a colleague. And it appears to hand the organisation the one thing the reporter most wants withheld: a recording of their own voice.

That appearance is the obstacle. Employees do not distrust voice channels because they have read a technical analysis of speaker recognition. They distrust them because a recording feels like an identity in a way that typed text does not, and because the person on the other end of a traditional hotline is a human operator who now knows something about them. Perceived anonymity, not actual anonymity, determines whether a report is ever made.

The engineering question this paper answers is therefore narrower than it first appears. It is not "can audio be processed securely." It is: **can a system take in a voice, extract everything an investigator needs from it, and destroy the link between that content and the speaker — provably, and by default?**

Section 02

What Anonymity Actually Means

"Anonymous reporting" is used to describe four different promises that fail in different ways. A reviewer should insist a vendor say which ones it is making.

FOUR PROMISES SOLD UNDER ONE WORD

Promise	What it means	How it typically fails
Pseudonymity	The report carries a case reference instead of a name, but the mapping exists somewhere.	The mapping is disclosed, subpoenaed, or leaked. Common in ticket-based systems.
Identity withheld from the employer	The vendor knows who reported; the customer does not.	The vendor becomes a single point of failure and a target of legal process.
Identity collected from nobody	No party holds identifying data, because none was ever requested or derived.	Follow-up becomes impossible — unless the channel is designed for it. See section 5.
Unlinkability	Two reports from the same person cannot be shown to come from the same person.	Defeated by metadata, writing style, or a persistent device or account identifier.

XPOSER.AI is built for the third and fourth: where the customer enables anonymous reporting, **no party — the employer, XPOSER.AI, or any subprocessor — is given information identifying the reporter, and none is derived from the**

audio. The first two are weaker positions we do not want to be able to fall back on, because a capability that exists will eventually be requested.

> **The design principle underneath all of this**
>
> A promise that depends on a policy is a promise that depends on whoever writes the policy next. Wherever it was possible to make a guarantee structural — the data is not there, the key is not held, the operation is not implemented — we did that instead of writing a rule about it.

Section 03

The Threat Model

A security claim without a threat model is decoration. These are the adversaries the architecture is designed to withstand, in ascending order of capability.

Defended against

- **The curious colleague.** Another employee with dashboard access who wants to know who filed a report about them or their friend. Defeated by not holding the data, and by role-scoped access with an immutable audit trail over every case view.
- **The implicated manager.** A case handler who is themselves a subject of the report and who has legitimate access to the system. Defeated by the same absence of identity data, plus conflict-of-interest routing that a customer configures at the workflow level.
- **The retaliating employer.** An organisation that decides, after the fact, that it wants to know who spoke. There is no administrative function, support path or escalation that reveals a reporter who chose anonymity, because the data does not exist to be revealed.
- **The compromised vendor account.** An attacker with XPOSER.AI staff credentials. They inherit the same limitation: staff access cannot produce identity for an anonymous report, and cannot read a zero-knowledge follow-up thread.
- **Legal process against the vendor.** A demand served on us can only compel what we hold. For an anonymous report, that is the report content and its case record — not an identity, because we never received one.

Not defended against — and we will not pretend otherwise

- **Self-identification in the content.** If a reporter says "I am the only night-shift supervisor in the Dammam warehouse," the system cannot unsay it. No architecture defeats this.
- **Employer-side monitoring.** Device management, network logging, CCTV and building access records belong to the employer, not to us. A report made from a corporate laptop on a corporate network may be observable at those layers regardless of what our channel does.
- **Coercion of the reporter.** A person who is pressured into disclosing their own case credential has disclosed it. This is why the credential is the only key — see section 5 — and why we recommend customers say so plainly in their programme communications.
- **A global passive adversary** observing all network traffic in and out of the organisation. We are not a Tor-grade anonymity network and do not claim to be.

> **Why the honest list matters commercially**
>
> Every item in the second list is one a competent security reviewer will find within an hour. A vendor that omits them is telling you something about the rest of its claims. We would rather your works council read the limits here than discover them in a demo.

Section 04

Architecture: Intake, Transcription, Severance

A report passes through four stages. The identity-relevant decision happens at the third and is irreversible by design.

1 · Intake

The reporter speaks, in any of the supported languages, over a channel that requires no account, no employee number and no device registration. Identity fields exist in the interface but, where the customer has enabled anonymous reporting, they are optional and unmarked; a reporter who leaves them blank is not prompted, warned or degraded in the queue. Audio is encrypted in transit with TLS 1.3 and at rest with AES-256, inside the customer's chosen hosting region.

2 · Transcription and translation

Speech is converted to text and, where the report was not given in the case handlers' working language, translated. This is the stage at which the substance of the account — who did what, when, to whom — is separated from the acoustic signal that carried it.

3 · Severance

The transcript, its classification and its case record proceed into the workflow. The audio does not proceed with them by default: what happens to it is a customer configuration, and under transcript-only delivery it is discarded once the transcript has been produced and quality-checked. No speaker representation is derived from the audio at any point — see section 7.

4 · Triage and human decision

Classification, severity and routing are computed and presented to a human case handler together with the policy citations and reasoning behind each. **The model recommends; a person decides.** No outcome affecting a reporter or a reported person is produced by automated processing alone. Every view, assignment, comment and decision is written to an append-only audit trail.

WHAT EXISTS AFTER SEVERANCE, FOR AN ANONYMOUS REPORT

Artefact	Retained	Identifies the reporter?
Transcript and translation	Per customer retention policy	Only if the reporter said so themselves
Source audio	Configurable; discarded under transcript-only	Carries voice; no template derived
Classification, severity, citations	Per customer retention policy	No
Case audit trail	Per customer retention policy	Records handler actions, not reporter identity
Reporter identity record	**Not created**	—
Voiceprint / speaker template	**Not created**	—

Section 05

The Zero-Knowledge Follow-Up Channel

Anonymity usually kills the investigation. An investigator reads an anonymous report, needs one more fact — a date, a document name, whether anyone else witnessed it — and has no way to ask. Most programmes solve this by asking for an email address, which quietly converts an anonymous report into a pseudonymous one.

XPOSER.AI issues a **case credential** at submission instead. The credential is generated for the reporter, displayed once, and is the sole means of returning to the case. It is not tied to an account, an address or a device.

- Presenting the credential opens the follow-up thread for that case and nothing else.
- Case handlers can post questions into the thread and read the replies, without ever learning who is on the other side.
- XPOSER.AI administrators cannot link a thread to an identity, because the link is the credential and we do not hold it.
- **A lost credential cannot be recovered.** There is no reset, no "verify your identity to regain access," and no support path. That is not a gap in the implementation; a recovery mechanism is precisely the back door this design exists to remove.

> **What this buys the investigation**
>
> Two-way dialogue with an anonymous source. The single most common reason anonymous reports are closed unactioned is that they are too thin to act on and there is no way to ask a follow-up question. This channel removes that reason without asking the reporter to give anything up.

Section 06

Voice Masking And Transcript-Only Delivery

Two customer-configurable controls sit between the recording and the case handler.

Voice masking

Where enabled, the audio presented to case handlers is pitch- and timbre-altered so that a listener who knows the reporter is materially less likely to recognise them. Masking is a mitigation, not a guarantee: speech cadence, vocabulary and the content itself remain. It is most valuable in small teams, where a handful of people share a working language and recognition risk is highest.

Transcript-only delivery

Where enabled, case handlers never receive audio at all. The investigation proceeds from the transcript, and the source recording is discarded once the transcript has been produced and quality-checked. This is the strongest of the two settings and the one we recommend as a default for programmes whose primary concern is recognition by colleagues.

> **A trade-off worth stating**
>
> Transcript-only removes evidentiary material. Tone, distress and emphasis are lost, and in a small number of matters — harassment cases in particular — those carry weight an investigator would want. This is a decision for the customer's legal and compliance function, made once, at configuration, rather than case by case under pressure.

Section 07

What We Deliberately Do Not Collect

The strongest statements in this document are negative ones. Each is a capability absent from the system, not a policy about how a present capability is used.

ABSENT BY DESIGN

Not collected or derived	Why it matters
Voiceprints and speaker templates	No biometric representation of a speaker is generated from the audio. There is nothing to match against, nothing to leak, and nothing to compel. This also keeps voice out of the "biometric data for the purpose of uniquely identifying a natural person" category in Article 9 GDPR.
Voice matching between reports	No operation compares one recording to another. Two reports from the same person cannot be shown to share a source — the unlinkability property from section 2.
Mandatory accounts	No registration means no credential, recovery address or profile to correlate against.
Device fingerprinting	We do not build a device or browser signature to recognise a returning reporter.
Cross-tenant model training	Report content is not used to train or fine-tune models serving other customers. One organisation's reports cannot surface in another's system.

Every row above is a term we are prepared to put into a data processing agreement, which is the appropriate place for a promise a customer needs to be able to enforce.

Section 08

The Limits — And The Questions To Ask Us

If you are evaluating this platform, these are the questions we would ask in your position. We have written what we would answer.

“Can you produce the identity of an anonymous reporter if a court orders you to?”

No — an order can only compel what exists. For an anonymous report we hold content, classification and case record, and no identity. Where a reporter *did* identify themselves, that identity is restricted to the authorised handlers and is disclosed further only where a legal obligation is necessary and proportionate, with prior notice to the reporter where the law permits it.

“What stops one of your engineers from listening to our reports?”

Role-based access granted on need, logged in an immutable audit trail, over data encrypted at rest in your hosting region, with tenant isolation. This is the answer that rests on controls rather than architecture, and it is therefore the one to test: ask for the access-review evidence and the penetration test summary.

“What happens when a reporter loses the case credential?”

The follow-up thread is unrecoverable. We consider this the correct behaviour and we do not offer a workaround. Programmes should tell reporters this plainly at submission.

“Does the AI decide anything about the reporter?”

No. Classification, severity and routing are recommendations carrying their citations and reasoning; a human case handler makes every decision. Agreement between model and human is measured and reported — the methodology is published separately in *The 60-Second Triage Benchmark*.

“Where does our data actually live?”

In the hosting region you select — Saudi Arabia, the United States, Canada or the European Union. Cross-border access for support or maintenance is governed by the transfer mechanisms set out in our Privacy Policy and your data processing agreement.

“Are you certified?”

A SOC 2 Type I audit is in progress and an ISO 27001 gap assessment is underway; independent penetration testing has been performed and the executive summary is available under NDA. We publish current status, rather than intended status, on our Trust & Security page. Ask for it in writing and check the date.

Section 09

Regulatory Alignment

The architecture was designed against the following instruments. This is a summary of alignment, not legal advice, and it does not substitute for your own assessment.

Instrument	Relevance to this design
Directive (EU) 2019/1937 EU Whistleblower Directive	Article 16 protects not only the reporter's identity but any information from which it may be directly or indirectly inferred. A channel that derives no speaker representation and holds no identity record for anonymous reports satisfies that obligation structurally. Member States decide separately whether anonymous reports must be accepted.
GDPR / UK GDPR	Data minimisation (Art. 5(1)(c)) and data protection by design and by default (Art. 25) are the governing principles for sections 4 and 7. Because no biometric template is generated for identification purposes, voice is not processed as Article 9 biometric data. Article 22 is addressed by human decision-making over every outcome.
Saudi PDPL	Purpose limitation, minimisation and the restrictions on transfer outside the Kingdom shape the regional hosting model. Saudi residency keeps report data in the Kingdom in the ordinary course of the service.
SAMA Whistle Blowing Policy	Requires financial institutions to assure the confidentiality of whistleblower identity and to protect against retaliation, and to provide effective reporting channels. Note that under the SAMA policy a whistleblower who reports anonymously cannot claim the institution's protective obligations — a point programmes should communicate honestly.
US DOJ ECCP (Sept 2024)	Asks whether a company has an anonymous reporting mechanism and, if not, why not; whether it tests that employees feel comfortable using it; and what baseline of human decision-making is used to assess AI. Treated at length in our companion paper on the ECCP.

Section 10

References

1. **Directive (EU) 2019/1937** of the European Parliament and of the Council of 23 October 2019 on the protection of persons who report breaches of Union law, Articles 16 and 22. eur-lex.europa.eu
2. **Regulation (EU) 2016/679 (GDPR)**, Articles 5, 9, 22 and 25.
3. **Personal Data Protection Law**, Kingdom of Saudi Arabia, and the Regulation on Personal Data Transfer Outside the Kingdom. sdaia.gov.sa
4. **Whistle Blowing Policy for Financial Institutions**, Saudi Central Bank (SAMA), August 2019. rulebook.sama.gov.sa
5. **Evaluation of Corporate Compliance Programs**, U.S. Department of Justice, Criminal Division, updated September 2024. justice.gov
6. **XPOSER.AI Privacy Policy** and **Trust & Security Center**, xposer.ai

XPOSER.AI · Autonomous, voice-first AI whistleblowing. From incident to audit-ready insight in under 60 seconds — 24/7, in 100+ languages.

Questions, corrections or a request for the penetration test summary: **info@xposer.ai**

© 2026 XPOSER.AI. This document may be circulated within your organisation and to your advisers. It is provided for information and does not constitute legal advice.