

A Privacy Impact Assessment Framework for Agentic Voice Assistants: The Voice Contextual Integrity (V-CI) Protocol

Aditi Balaji, Michael Stoneman, and Elijah Zhao
Carnegie Mellon University – Privacy Engineering Capstone

Faculty Advisor: Dr. Hana Habib

Meta Sponsors: Tom Wei & Nihan Brunton

Executive Summary

In an agentic voice system, a single underspecified spoken request can trigger an extended sequence of autonomous operations: interpreting intent, resolving speaker identity, retrieving sensitive records, dispatching payloads to third-party endpoints, executing irreversible external transactions, and broadcasting acoustic confirmation in a shared physical environment. Within such architectures, the point where a privacy vulnerability manifests is rarely where it originated. While established Privacy Impact Assessment (PIA) and threat-modeling frameworks can map static data flows, they fail to systematically foreground voice-native risks—such as speaker ambiguity, bystander capture, uncertainty in acoustic inference, or autonomous actions executed before a human can intervene.

To address this gap, this capstone introduced the **Voice Contextual Integrity (V-CI) Protocol**: a structured, PIA-aligned review methodology that makes the end-to-end path from spoken input to system action explicit and auditable. V-CI is designed to supplement—rather than replace—an organization’s broader PIA processes, privacy threat modeling, security reviews, and legal analyses. It does not issue automated compliance certifications or guarantee absolute system safety.

System Architecture and Artifacts. This project delivered four interconnected contributions:

1. **A 43-record Voice Risk Taxonomy (VR-1–VR-43):** Reconciles voice assistant research with agentic AI vulnerability catalogs, explicitly separating 15 *Voice-Structural* risks (unique to acoustic interaction) from 14 *Voice-Amplified* and 14 *Agentic* risks (hosted or intensified by voice). Each entry includes annotations for underlying mechanism, severity, voice-relevance, LINDDUN categories, GDPR mapping, and mitigation families (M1–M14).
2. **A Voice- and Action-Aware Threat-Modeling Framework:** Integrates a five-tier capability ladder (C0–C4),

a nine-stage voice-to-action pipeline, six Contextual Integrity parameters, and five analytical review lenses around atomic information flows, while leaving baseline LINDDUN threat classes intact.

3. **The V-CI Protocol and Interactive Prototype:** Operationalizes the framework into nine sequential reviewer steps—spanning scenario framing, capability classification, flow-level analysis, risk surfacing, priority scoring, mitigation design, and evidence export—instantiated as a local, configuration-driven web application without external model dependencies.
4. **A Transcript-First Validation Package:** Documents empirical reviewer reasoning when using a bounded instantiation of the protocol.

Contextual Integrity provided the theoretical baseline for evaluating contextual fit; LINDDUN supplied the core threat vocabulary and data-flow elicitation methodology; and frameworks from NIST, CNIL, and Hoepman informed evidence organization and control classification. Taxonomy mappings serve as traceability aids rather than legal compliance determinations.

Methodological Design. Human-subjects research proceeded across two stages under approved protocol STUDY2026_00000264. First, four moderated co-design sessions with privacy practitioners evaluated a manual review workflow against a scripted Wizard-of-Oz assisted drafting path. Practitioners strongly favored editable suggestions and human accountability while warning against rubber-stamping and opaque aggregate risk scoring, prompting us to preserve reviewer-authored findings and defer automated generation. Second, ten privacy reviewers (six industry, four academic) completed moderated think-aloud evaluations across two agentic voice scenarios (Prescription Refill and Travel Booking), anchored to specific focal interactions. Six participants used a bounded V-CI workflow and four used an individual workflow derived from LINDDUN PRO,

resulting in substantive analysis across 18 of 20 assigned tasks. Data triangulation combined structured JSON logs, full transcripts, and non-intrusive telemetry, applying 47 qualitative codes across 45 reconstructed task-level issue clusters. All empirical findings are explicitly tagged as *core*, *descriptive*, *exploratory*, or *sensitivity* to delineate the analytical scope of each claim.

Key Empirical Findings. High perceived usability did not automatically translate into complete audit artifacts. While post-study questionnaire responses were overwhelmingly positive (most item medians reached 6/7, with expected reuse and practice utility scoring 7/7), documentation ease received the lowest rating (5/7). Notably, one participant assigned a 7/7 rating to documentation ease while submitting 15 risk tag selections completely lacking written findings. The strongest empirical result revealed a significant **fidelity gap**: in **all ten sessions**, reviewers verbalized substantially richer privacy reasoning than was preserved in their exported JSON logs, and routinely relied on unstated premises without logging them as scenario facts, assumptions, or open questions. This gap reflects a combination of study procedure, time pressure, documentation overhead, and prototype friction.

Furthermore, reviewers naturally evaluated multi-step **causal paths rather than isolated interaction edges**. Nine out of ten participants linked their assigned interaction edge to upstream origins or downstream consequences; seven expanded their analysis beyond assigned task boundaries; and six struggled to connect scenario facts, DFD elements, system actors, and taxonomy categories into a unified analysis. Of the 45 reconstructed issue clusters, 26 could not even be articulated without referencing the broader interaction path, and the combined scenario/DFD context view was reopened across all 18 analyzed tasks (accumulating 114 total opens). All six V-CI participants conflated baseline scenario facts with recommended future controls. While the two workflows highlighted complementary concerns—V-CI foregrounded speaker ambiguity, acoustic output privacy, and delegated authority, whereas LINDDUN foregrounded data linking, retention, and recipient scope—the small, non-randomized sample supports no claims of statistical superiority, precision, or recall. Crucially, the qualitative data demonstrates that a reasoned no-finding conclusion and a task abandoned due to time constraints represent distinct analytical outcomes that must not be conflated with incomplete analysis.

Actionable Recommendations. Future iterations of V-CI must remain **interaction-anchored yet path-aware**. Issue reporting should anchor to a mandatory focal manifestation while providing optional fields to map upstream origins, intermediate propagation stages, downstream impacts, and evidence provenance. Workflow complexity should be managed through progressive disclosure, issue clustering, capability-based filtering, and persistent system diagram contexts. Out-

put artifacts must explicitly tag entries as *scenario fact*, *assumption*, *uncertainty*, or *desired control*, preserving status states (*retained*, *rejected*, *unresolved*, *deferred*, *not analyzed*) so second reviewers can reconstruct rationale without replaying session audio. Additionally, a concise, reviewer-authored manifestation statement should be required before any taxonomy label is logged as a finding. Organizations should evaluate V-CI as a specialized voice/agent module within existing PIA or architectural review workflows, triggering re-reviews upon material architectural changes while maintaining human accountability.

Scope and Limitations. This formative evaluation does not measure absolute analytical accuracy, establish comparative precision/recall against LINDDUN, or validate the unbounded protocol (nine-dimension risk scoring, automated P1–P4 assignment, multi-row flow authoring, and generative mitigations were excluded from the validation workspace). The two-item benchmark serves as a construct alignment audit rather than a ground-truth testing suite. The taxonomy represents an evolving synthesis rather than an exhaustive catalog, and no portion of the V-CI framework constitutes formal legal advice or compliance certification.

Artifact Availability. The participant-safe prototype codebase and technical documentation are publicly accessible in the public V-CI repository, with a live deployment hosted at <https://vci-prototype.vercel.app>. The formal report and associated research transcripts remain private.