Human Agents to AI Agents: Why AI Voice Agents Inherit, Then Compound, PCI DSS Cardholder Data Exposure and How to Solve It.

AI Can Only Scale at AI Speed by Removing Cardholder Data at the Point of Ingestion

August 24, 2026

By William Placke, Certified Information Privacy Professional-US (CIPP-US), Juris Doctorate, Diploma EU Law, HBS Certificate in Data Privacy and Data Security, President, SecurePII Americas.

This publication is informational and is not legal or professional advice. Readers should consult their own legal counsel, Qualified Security Assessor, or Internal Security Assessor for guidance specific to their facts and circumstances.

Executive Summary

Telephone-based payment handling has relied for more than a decade on a defined premise. A bounded set of people and systems constitutes the point where cardholder data becomes vulnerable, and controls built around that boundary are sufficient once those controls function correctly. AI voice agents, copilot tools, and transcription systems now sit inside that boundary, and in a growing number of environments, replace the human agent entirely. The question this paper answers is direct. Does moving the payment moment to an AI agent, or layering AI tooling onto a human agent's existing call, resolve the exposure the human agent premise was built to contain, or does it compound that exposure and expand PCI audit scope?

The answer is that an AI agent compounds the boundary problem and expands PCI audit scope. An AI system that recognizes a card number is processing that data under the same rule that already governs a human agent or a redaction engine. An AI-layered environment cannot contain a single recognition failure to one place the way a plain recording once could. Where that failure reaches a system trained on the data it touched, it cannot be undone the way a stored recording can. Removing cardholder data at the point of ingestion, before it reaches any human, any AI system, or any recording, is the only architecture that resolves the original exposure and the increased exposure AI adds. Importantly, it is the only architecture that lets an AI-driven voice environment scale without multiplying the AI compliance burden as the AI tools and output grow.

This paper establishes that answer through three claims. First, recognition of cardholder data does not depend on who or what is listening, such that an AI system's own speech recognition or language processing layer constitutes the same act of processing that brings a system within PCI DSS scope. Second, an AI-layered environment cannot contain a recognition failure to one place the way a plain recording once could, because a missed pause or an unredacted recording no longer sits in a single archive. It can propagate into a summary, a virtual agent's working context, and whatever system consumes that data next. Third, where that propagation reaches a training pipeline, the resulting exposure cannot be remediated the way a stored recording can, because a model does not have a deletion function for a fact it has already absorbed.[1] Based on the author's own professional experience, remediation costs in

[1]This claim is developed at length below, including citation to the European Data Protection Board's and the National Institute of Standards and Technology's published treatment of training-data memorization in AI models.

circumstances resembling this one have exceeded six figures on multiple occasions,[2] a magnitude consistent with the six-figure cost independently documented for the narrower and less severe problem of remediating an existing recorded-call archive.[3] In the more severe cases the author has encountered, remediation has required deletion of the entire underlying data set, consistent with the pattern of failed Pause-Resume controls discovered during a PCI audit in the human agent context that this series has documented elsewhere.[4]

This paper stands on its own. It restates in full the doctrinal foundation on which it builds. Readers of the companion paper on pause-and-resume controls will recognize that foundation. Readers encountering this analysis for the first time will find everything required to follow it here.[5]

This paper does not name or evaluate any specific vendor, platform, or product. Its claims are addressed to any voice environment, human staffed, AI staffed, or both, that allows an AI system to receive, process, or store spoken cardholder data without a solution that prevents that data from reaching the system in the first place before any cardholder data touches an AI tool.

Two Independent Rules Reach Any System That Touches a Card Number

PCI DSS 4.0.1 reaches any system that touches a card number through two separate and independently sufficient rules. A system that recognizes a card number in order to act on it is processing that data. A system that determines that a string of numbers is a credit card number, rather than, for example, a phone number, has processed that credit card number. A system that retains an intact, recoverable copy of that data is storing it. Either rule alone is enough to bring a system into scope. An architecture that satisfies one while quietly failing the other has not achieved compliance. It has only chosen which rule to violate.

The processing rule attaches to the act of recognition, not to its outcome

PCI DSS 4.0.1 Requirement 3.3.1 provides that sensitive authentication data is not stored after authorization, even if encrypted, and that all sensitive authentication data received is rendered unrecoverable upon completion of the authorization process.[6] Specifically, the CVV number of a credit card may not be stored after authorization, under any circumstances. Table 3 of the standard reinforces

---

[2]Based on the author's own professional experience advising entities on PCI DSS exposure in AI-enabled voice environments. The author is not aware of a published, independently verified statistic isolating the cost of remediating a model that has absorbed cardholder data during training, and this figure should be read as a professional estimate rather than a sourced statistic.

[3]See Paytia, "Call Center PCI Compliance: A Practical 2026 US Guide," cited and discussed in William Placke, "Pause-Resume Is Dead in the Age of AI," supra note 1, Section VI.C (describing remediation of an existing recorded-call archive through audio-redaction tooling as a six-figure undertaking for a mid-sized archive). That figure concerns a narrower problem, cleaning an existing archive, than the retraining or discarding of a model that has already absorbed cardholder data during training.

[4]See William Placke, "Pause-Resume Is Dead in the Age of AI," supra note 1, Section III (documenting, through independent research on comparable manual tasks and the author's own professional observation, expected and observed failure rates for Pause-Resume controls in the human agent context).

[5]William Placke, "Pause-Resume Is Dead in the Age of AI: Why Manual Call Recording Controls No Longer Satisfy PCI DSS 4.0.1 in AI-Enabled Voice Environments" (SecurePII working paper, 2026). That paper develops the redaction-engine and manual-versus-automated pause-and-resume analysis referenced but not repeated here.

[6]Payment Card Industry Data Security Standard: Requirements and Testing Procedures, v4.0.1 (June 2024), Requirement 3.3.1, at p. 79.

this with no exception for encryption and no exception for the sophistication of the system involved.[7] Neither provision asks how a system came to hold the data. Both ask only whether the data was stored, and under what condition it was rendered unrecoverable.

A system that masks, redacts, transcribes around, or diverts a card number must first locate where that number occurs. That location step happens before the masking, the redaction, or the diversion, and it happens regardless of what the system does next. Locating a card number inside an audio stream or a block of text is an act of processing that data.[8] The same logic applies with no modification to any AI system whose first operational step, before summarizing, redacting, or routing, is recognizing what it is looking at.

The storage rule attaches independent of any question about processing

A separate rule reaches the same architecture on its own terms, sufficient by itself to bring that architecture into scope. PCI DSS defines its scope by reference to the cardholder data environment, comprised of the system components, people, and processes that store, process, or transmit cardholder data or sensitive authentication data, together with any system that could impact the security of that data.[9] Storage is the plainest of the conditions named in that definition. It does not require an assessment of intent, of what a system was designed to do, or of whether recognition occurred as a byproduct of some other function. It asks only whether an intact, recoverable copy of the data exists somewhere in the system.

The PCI Security Standards Council's own supplemental guidance illustrates this with a call flow it uses as a standard reference example. In that illustration, once a call is recorded, the account data spoken during the call is described as stored in the call recording storage, and the guidance notes plainly that at that point, the data can be queried.[10] The Council did not need to resolve any question about processing to reach that conclusion. Storage, once established, is sufficient by itself. Thus, these are two stand-alone requirements that will expand the scope of the PCI audit unless the cardholder data is removed at ingestion before it reaches the AI tools.

Neither rule is keyed to whether a human or a machine is the one recognizing or storing the data

The text of both rules is worth reading closely for what it does not say. Requirement 3.3.1 does not distinguish between a human agent who writes down a card number and a machine that transcribes one. The scope definition in Section 4 does not distinguish between a system component operated by a person and a system component that operates on its own.[11] Both rules attach to the data and to the system

---

[7]PCI DSS v4.0.1, Table 3, Account Data Element Storage Requirements, at pp. 6–7.

[8]This point is developed at length, with supporting citations to post-call redaction architecture, in William Placke, "Pause-Resume Is Dead in the Age of AI," supra, Section V.

[9]PCI DSS v4.0.1, Section 4, Scope of PCI DSS Requirements, at p. 9; PCI DSS v4.0.1 Glossary, definition of Cardholder Data Environment (CDE), at p. 381.

[10]PCI Security Standards Council, Information Supplement: Protecting Telephone-Based Payment Card Data, v3.0 (Nov. 2018), Table 4, at p. 11.

[11]PCI DSS v4.0.1, Section 4, Scope of PCI DSS Requirements, at p. 9 (defining scope by reference to "system components, people, and processes," with no qualification tied to whether a given system component is human-operated or autonomous).

component holding or acting on it. Neither rule attaches to the nature of the entity performing the recognition or the storage.

This is not a gap in the standard that AI just happens to expose. It is the standard functioning exactly as written, applied to a category of system its drafters described in terms broad enough to reach whatever technology an entity chooses to deploy. An AI voice agent, a transcription engine, a sentiment analysis engine, and a large language model are each, under this framework, a system component like any other. What follows in this paper examines what that means once such a system component is the one doing the recognizing.

Recognition Does Not Require a Human Ear

An AI system's own speech recognition and language understanding layer performs the same recognition event the preceding section already established as processing. This is so regardless of where the resulting audio or text is subsequently routed, stored, or discarded, and regardless of what an architecture's designers describe that routing as accomplishing.

Nothing in Requirement 3.3.1 or the cardholder data environment's scope language exempts an automated recognition layer. An AI voice agent's initial processing step, converting spoken audio into text or into a machine-readable representation of what was said, is functionally identical to the recognition step the prior section already described. A speech-to-text engine that transcribes a spoken card number has performed exactly the recognition act this paper's opening rule addresses. A natural language understanding layer that classifies a spoken utterance as containing payment information has performed the same act by a different technical method.

Consider an architecture that markets itself as diverting the payment segment of a call away from any human agent or AI copilot, isolating that segment inside its own flow so that neither a person nor an AI tool ever sees it. The architecture's compliance claim rests entirely on where the audio is routed. Routing, however, answers only one question. *It does not by itself answer where recognition occurred*. If the platform's own speech recognition layer operates upstream of, or in parallel to, the diversion point, processing all audio, including the payment segment, as a matter of the platform's baseline architecture rather than as a configurable option, the diversion has changed where the output of recognition goes without changing whether recognition happened. A description of where a call is routed describes routing. It does not, by itself, establish where in the signal path recognition took place.

This is the same distinction the preceding section drew between recognizing a number and deciding what to do with it afterward. An architecture can divert, redact, or discard the output of recognition and still have performed the recognition itself, at the moment recognition occurred, regardless of what happens next. A compliance review that examines only where audio is routed, without separately tracing where recognition occurs, has examined half the architecture.

The same reasoning that reaches an AI system's recognition layer reaches the recording that layer operates on. What that recording contains, and whether it persists anywhere an architecture's own description does not account for, is the subject of the next section.

A Persisting Recording Is Storage, and a Persisting CVV Is a Violation With No Defense

Independent of every argument in the preceding sections, an original, unredacted recording that persists anywhere in an AI-layered environment is stored account data under the rule already established above

And, where that recording contains a spoken card verification value (the “CVV”) the violation admits no defense of any kind.[12]

The storage rule established earlier is sufficient on its own to bring a system into scope, without regard to any question about processing. That rule holds regardless of whether an AI system is also present in the environment. An original, unredacted recording that a diversion or a redaction pass was designed to avoid capturing, if it still exists anywhere, in an intermediate buffer, a backup, or a location an architecture's own documentation does not describe, is stored account data.

PCI DSS 4.0.1 treats sensitive authentication data, such as a CVV, as a stricter category than cardholder data generally. The primary account number may be stored, provided it is kept to a minimum and rendered unreadable according to the standard's requirements.[13] Sensitive authentication data, which includes the card verification value (the “CVV”), has no equivalent path. It may not be stored after authorization, under any circumstances, encrypted or not. Requirement 3.3.1 carries an express exclusion from the customized approach that otherwise allows an assessor to accept a different but equally effective compensating control in its place.[14]

A customer who reads a card number aloud to complete a purchase plausibly reads the card verification value in the same conversation, since both are typically requested together during the same portion of a call. Where the recognition analysis in the preceding section shows that an AI-layered architecture may not have prevented the underlying recording from capturing that portion of the call at all, the same recording that captures the card number plausibly captures the card verification value alongside it.

A finding of this kind carries no available defense. An entity can argue in the context of a credit card number found in recordings, with varying success, that a given control was reasonable, that a compensating control achieved an equivalent outcome, or that a missed pause was an isolated human error. Each of those arguments has a path to some result. A stored card verification value has none. The rule admits no compensating control and no customized approach. Discovery of the fact is the entire inquiry and a consequential PCI audit failure requiring remediation.

A Single Miss No Longer Stays in One Place

In a plain voice environment, a single Pause-Resume miss or a single unredacted recording is a bounded, discoverable, correctable event confined to one file in one storage system. In an AI-layered environment, the same miss is no longer bounded this way, and an entity that treats it as though it still were will grossly underestimate its own exposure.

Once a recording, or the raw audio or text stream containing a recognition failure, feeds a transcription engine, a summarization tool, or a virtual agent's working context, the content of that failure can appear in multiple, independent locations. Each location may carry its own retention policy, its own access controls, and its own audit trail, or the absence of one. A single missed pause that once lived in a single

---

[12] A Card Verification Value, or CVV, is the three or four digit number printed on the card. Sometimes the CVV is also referred to as the CVC, CSC, CVN, or CID depending upon the card issuer. For purposes of this paper, these have been referred to in the collective as CVV. See, “What is a CVV?” June 16, 2025 by American Express. Available at https://www.americanexpress.com/en-us/credit-cards/credit-intel/what-is-cvv/ PCI DSS glossary also uses CAV2, CVC2, CVN2, CVV2, or CID. If it seems the naming convention is like a naval fleet, there is no argument here.

[13] PCI DSS v4.0.1, Table 3, Account Data Element Storage Requirements, at pp. 6–7 (previously cited above).

[14] PCI DSS v4.0.1, Requirement 3.3.1, at p. 79 (“This requirement is not eligible for the customized approach.”).

call-recording archive can now live simultaneously in that archive, in a generated summary stored in a separate system, and in the working memory of a virtual agent handling a later, unrelated call.

This propagation carries three distinct characteristics, and an entity's exposure is a function of all three together. The propagation is unbounded, since nothing in a typical AI tooling stack limits how many downstream systems may consume a given recording or transcript. The resulting copies are frequently unknown to the entity itself, since visibility into where a recording's contents have traveled after the fact requires instrumentation many voice environments were not built with. The receiving systems are often uncertified for this purpose, built and evaluated as general-purpose productivity or analytics tools rather than as components of a cardholder data environment.

A companion body of work by the author's colleague Jason Thals describes this phenomenon as anomalous cardholder data, defined as account data stored, transmitted, accessed, or retained in ways inconsistent with an entity's defined PCI scope, documented control intent, or established operational baselines, and identifies cardholder data appearing inside an AI transcript as a direct example.[15] Cardholder data surfacing inside an AI transcript is, under that framing, exactly this kind of anomaly. It should not be there, and its presence brings the AI system, the transcript, and the handling of both within PCI DSS scope precisely because the anomaly exists, independent of how it got there.

An entity that has only ever audited its call-recording archive, the way a Pause-Resume review would, has audited one of what may be several locations where a single failure now lives. The audit becomes incomplete the moment an AI layer is introduced, regardless of how thoroughly that audit examines the recording archive itself.

A Trained Model Cannot Be Redacted

Where a propagated instance of cardholder data reaches a system's training or fine-tuning pipeline, the resulting exposure cannot be remediated the way a stored recording or a transcript can, because removing a file is a different act than removing a fact a model has already learned.

Deleting a recording removes the recording. Deleting a transcript removes the transcript. Neither act has an equivalent when the data in question has already been used to adjust a model's parameters during training. A model does not store the training examples it learned from in a form that can be located and deleted the way a file can. It stores the effect those examples had on a very large number of internal values, and reversing that effect for one specific example, without retraining the model from a clean dataset, is not a capability generally available AI systems currently possess.

The European Data Protection Board reached a closely related conclusion in a different legal context. Addressing when an AI model may be considered to hold personal data, the Board found that information from a training dataset may remain absorbed in the parameters of a model, represented as mathematical values, even where the model was never designed to reproduce that information on request.[16] The Board's opinion interprets the General Data Protection Regulation, not PCI DSS, and this paper does not

---

[15] Jason Thals, "PCI DSS v4.0.1 Data Security Standards: Anomalous Cardholder Data, Scope Drift, and the End of Pause/Resume in PCI Compliance" (SecurePII, v.1.0.1, Apr. 10, 2026), colloquially the "Platypus Problem" paper, summarized at https://www.securepii.cloud/the-platypus-problem/.

[16] European Data Protection Board, Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models (adopted 17 Dec. 2024), ¶ 31. This Opinion interprets the General Data Protection Regulation and is cited here for its technical reasoning, not as PCI DSS authority.

suggest otherwise. Its technical reasoning is directly analogous nonetheless. Whether training data persists inside a model in a form that resists simple deletion is a technical question with one answer, independent of which legal framework is asking it.

The National Institute of Standards and Technology reached the same technical conclusion independently, in a framework built for a different purpose. NIST's Generative AI Profile names Data Privacy as one of twelve defined risk categories for generative AI systems, and identifies training-data memorization, a model reproducing personal information it was trained on, as a specific risk within that category.[17] A government framework built to help organizations manage AI risk and a data protection authority interpreting a privacy statute have reached the same finding from different directions. A model can retain what it was trained on in a way an ordinary deletion process does not reach.

PCI DSS's own requirement that sensitive authentication data be rendered unrecoverable sets a bar this technology may not be able to meet through any process short of retraining or discarding the model entirely.[18] A recording can be rendered unrecoverable by deleting it. Whether a trained model can be rendered unrecoverable of a single absorbed fact, without discarding everything else it learned alongside it, is a question this paper does not attempt to answer technically. What this paper establishes is that the standard's own remediation bar does not lower itself to accommodate the answer, whatever that answer turns out to be.

Five Questions Determine Whether an AI Voice Environment Is Defensible

A Qualified Security Assessor or Internal Security Assessor examining an AI-layered voice environment needs a review structure built for that environment specifically, not a checklist inherited from a plain recording review. Five questions, addressed in sequence, determine whether such an environment is defensible.

Who owns AI and PCI Compliance Governance?

AI governance is increasingly recognized as its own discipline requiring a defined owner rather than an unassigned responsibility spread across existing teams. The International Association of Privacy Professionals' own 2025 survey of the AI governance profession found organizations placing this ownership predominantly with privacy, compliance, and legal functions rather than leaving it undefined.[19] An entity that cannot identify who owns AI-related PCI exposure specifically has not yet answered this question, regardless of how mature its general privacy or security program is.

Where does recognition happen, and where does audio go?

A reviewer needs a complete map of every point in the signal path where speech recognition or language understanding occurs including redaction engines. Too often, auditors only map where the resulting audio or text is routed afterward. As the earlier hypothetical illustrated, these are different questions, and an architecture's own description typically answers only the second one.

---

[17] National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, NIST AI 600-1 (July 2024).

[18] PCI DSS v4.0.1, Requirement 3.3.1, at p. 79 (previously cited above).

[19] International Association of Privacy Professionals, AI Governance Profession Report 2025. Cited as evidence of an emerging professional consensus on ownership, not as a technical or legal standard.

Does any production call data reach a training or fine-tuning pipeline, and how?

An entity must be able to state affirmatively whether call recordings, transcripts, or derived summaries are used to train, fine-tune, or improve any AI system, whether operated internally or by a vendor, and must be able to describe what sanitization, if any, occurs before that data is used.

Do third-party service provider agreements name the AI vendor specifically?

Requirement 12.8.1 requires an entity to maintain a list of all third-party service providers with whom account data is shared or that could affect the security of account data.[20] An AI vendor that receives call audio or transcripts for transcription, uses voice recognition, engages a redaction engine, provides summarization, or analytics is such a provider. Requirement 12.9.1 separately requires that provider to furnish a written acknowledgment of its responsibility for the data it processes.[21] A general technology vendor agreement that predates an AI feature's deployment does not necessarily satisfy either requirement for that feature specifically.

Can the entity demonstrate that data has actually been rendered unrecoverable, including from a trained model?

Where an entity has determined that cardholder data reached a training pipeline, it must be able to describe, concretely, what rendering that data unrecoverable would require. An answer that only describes deleting a file has not addressed a model that may have already absorbed the data during training.

The Cost of This Exposure Exceeds the Cost of Preventing It

The costs this paper has identified compound rather than replace the costs already documented for plain Pause-Resume environments, and the AI-specific costs are, based on the available evidence, larger than the ones already discussed.

The companion paper on Pause-Resume controls documents that a full Report on Compliance assessment can run from the tens of thousands of dollars into several hundred thousand dollars for a complex, multi-location environment.[22] An AI-layered environment increases this cost. Every consideration in the five-question review above is now a subject a QSA must examine, in addition to the underlying voice environment itself.

Where an existing recorded-call archive is found to contain unredacted cardholder data, remediating that archive through audio-redaction tooling has been documented at a six-figure cost for a mid-sized archive.[23] Where the exposure instead reaches a trained model rather than a static archive, the author's own professional experience places the remediation cost above this figure on multiple occasions, and in

---

[20]PCI DSS v4.0.1 ROC Template r3 (Jan. 2025), Requirement 12.8.1, at p. 305.

[21]Id., Requirement 12.9.1, at p. 310.

[22]See William Placke, "Pause-Resume Is Dead in the Age of AI," supra note 1, Section VI.A, citing SecurityMetrics, "How Much Does PCI Compliance Cost," by Gary Glover.

[23]See William Placke, "Pause-Resume Is Dead in the Age of AI," supra note 1, Section VI.C, citing Paytia, "Call Center PCI Compliance: A Practical 2026 US Guide."

the more severe instances the author has encountered, remediation has required discarding the entire underlying training data set and retraining the model from a sanitized version.[24]

Layered against each other, these costs describe an entity paying for the same underlying failure more than once, in ongoing assessment scope, again if an existing archive requires remediation, and again, at a materially higher figure, if the failure has already reached a trained model by the time it is discovered.

Removing Cardholder Data at Ingestion Is the Only Architecture That Scales With AI

An architecture that removes cardholder data before it reaches any human, any AI system, or any recording resolves every exposure this paper has identified. Removing credit card data at its source is the only architecture whose compliance burden does not grow as an entity adds more AI tools.

The recognition analysis above establishes that an AI system's recognition layer cannot be designed around after the fact once it has already processed a card number. The CVV analysis establishes that a persisting recording is a violation with no available defense regardless of what caused it to persist. The propagation analysis establishes that a single failure in an AI-layered environment does not stay in one place. The training-data analysis establishes that a failure reaching a model may not be correctable at all in the way a file is correctable. Each of these problems shares a common root. Cardholder data reached a system, human or machine, that did not need to receive it to complete the transaction.

An architecture that captures the card number, expiration date, and card verification value (CVV) directly from the cardholder through a channel no human agent or AI system ever receives, of the kind SecurePII's SecureCall product is designed to provide, removes the problem at its root rather than managing it after the fact. Because the data never reaches any recognition layer, human or AI, the processing rule this paper opens with never triggers. Because the data never reaches any recording, the storage rule and the CVV-specific violation it can produce never arise. Because the data never reaches any transcript, summary, or training pipeline, the propagation and training-data problems this paper describes have nothing to propagate.

This is also the only architecture whose compliance burden does not scale with AI adoption. An entity that instead relies on governance, redaction, or after-the-fact review must revisit the five-question checklist above every time it adds a new AI tool, a new integration point, or a new vendor, since each addition is a new place recognition could occur and a new place data could propagate. An entity that has removed the data at ingestion has nothing new for an additional AI tool to receive, and its compliance obligations under this analysis do not grow as its use of AI grows. Scaling AI adoption and containing PCI DSS exposure are not, on this architecture, in tension with each other.

Moving from a human agent to an AI agent does not change the rule. It changes how many places the rule can be violated, and how difficult a violation is to undo once it happens. An entity that wants to adopt AI throughout its voice channel without multiplying its own exposure alongside it has one architectural choice available under the analysis in this paper, removing cardholder data before any system, human or machine, ever receives it.

---

[24] See notes above regarding the author's professional experience with remediation costs in AI-enabled environments. This figure is likewise the author's professional estimate rather than an independently verified statistic.