The Trust Imperative

Defensive AI Strategy for Healthcare Payers

From Adversarial Automation
to Accountable Intelligence

nēdl
labs

Ashish Jaiman
Founder & CEO

November 2025

Executive Summary

Healthcare payers stand at a critical juncture. As providers deploy AI to optimize claims submission, traditional payment integrity approaches, characterized by opaque rule-based systems and aggressive automated denials, are creating a counterproductive cycle that simultaneously erodes provider trust and fails to prevent payment leakage exceeding $100B annually across commercial plans.

The data tells a stark story: denial rates increased 37% between 2022 and 2024 as payers deployed more aggressive AI systems, yet 77% of providers report concerns about fair reimbursement, and improper payments persist at scale ($19.07 in Medicare Advantage, $3.58B in Part D for FY2024 alone).

Recent class action lawsuits alleging 90% error rates in AI-driven denials, coupled with Senate investigations revealing 108% increases in post-acute denial rates, signal that the current approach is legally, operationally, and reputationally unsustainable.

This white paper presents a fundamentally different paradigm: Defensive AI that builds trust through transparency, reduces loss through accountability, and creates sustainable competitive advantage. The solution requires a neuro-symbolic AI architecture that combines neural networks' pattern recognition with symbolic reasoning's explainability, satisfying CMS transparency requirements, supporting good providers, and detecting bad actors with the precision that traditional systems cannot achieve.

Key Takeaways

- **Provider AI is a catalyst, not a threat:** Legitimate optimization improves documentation quality; defensive AI must distinguish this from exploitation by bad actors
- **Transparency strengthens defenses:** Real-time policy guidance and explainable decisions align provider optimization with clinical appropriateness
- **Neuro-symbolic AI enables compliance:** CMS guidance requiring individual patient assessment demands reasoning capabilities that pure machine learning cannot provide
- **Three-layer framework delivers results:** Policy intelligence infrastructure + behavioral pattern detection + real-time collaboration = 30-40% denial reduction with enhanced trust
- **ROI through trust:** 12–18-month payback combining reduced leakage, lower administrative costs, enhanced provider relations, and litigation risk mitigation

The Payment Integrity Crisis: A $100B+ Trust Problem

Payment integrity has traditionally been framed as a financial control function: preventing improper payments, detecting fraud, and recovering overpayments. This cost-containment lens has shaped two decades of payment integrity evolution, from manual audits to rules-based edits to AI-powered denial systems. Yet despite escalating investment—the payment integrity market attracting $5.5B (KKR/Cotiviti) and $3B (New Mountain Capital) PE investments—the fundamental problem persists and intensifies.

$100B+

Annual Payment Leakage (Commercial Plans)

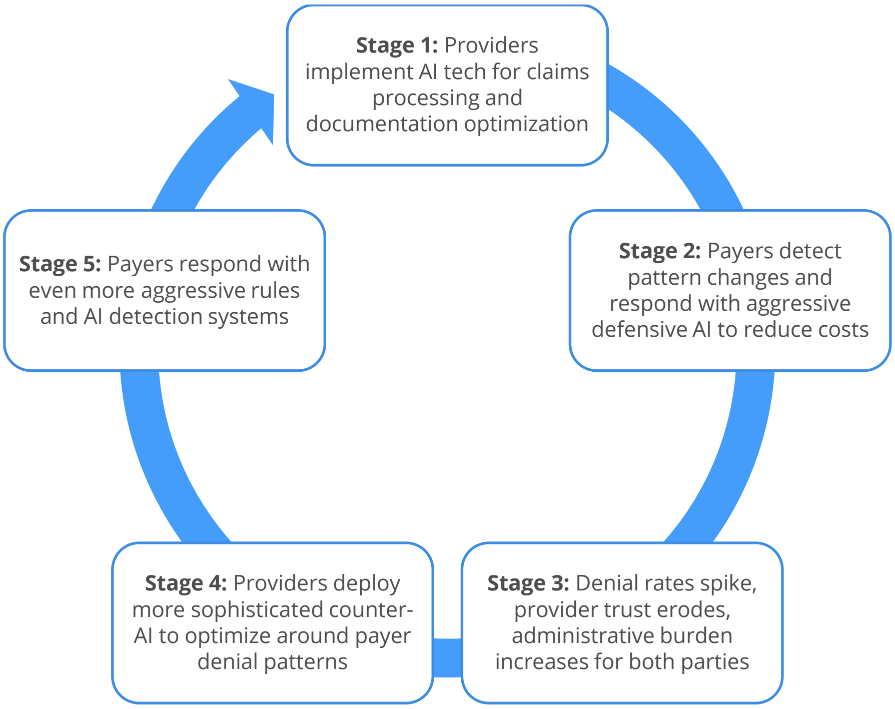
Despite aggressive payment integrity programs and increasing AI deployment

The Paradox: More Aggressive AI, Worse Outcomes

Between 2022 and 2024, healthcare payers significantly expanded AI-powered claims review systems. The result was not improved payment integrity; it was a 37% increase in denial rates (from 8% to 11% according to AMA data), accompanied by deteriorating provider relationships and persistent payment leakage. Providence Health reported a stark example: when payers deployed more aggressive AI tools, underpayments and initial denials increased by more than 50%, forcing the health system to increase "touches per claim" by the same amount to combat AI-driven denials.

This paradox reveals a fundamental misunderstanding**: payment integrity is not a financial problem; it's a trust problem**. When payers deploy AI systems that providers perceive as adversarial, rational actors respond by optimizing their own AI systems to exploit policy gaps, creating an escalating cycle where both sides invest in automation that increases friction without reducing leakage.

The Trust Erosion Cycle



Stage 1: Providers implement AI tech for claims processing and documentation optimization

Stage 2: Payers detect pattern changes and respond with aggressive defensive AI to reduce costs

Stage 3: Denial rates spike, provider trust erodes, administrative burden increases for both parties

Stage 4: Providers deploy more sophisticated counter-AI to optimize around payer denial patterns

Stage 5: Payers respond with even more aggressive rules and AI detection systems

This cycle benefits neither party.

⊗ **Payers** face increasing administrative costs, provider network challenges, and regulatory scrutiny.

⊗ **Providers** face revenue uncertainty, administrative burden, and impacts on patient satisfaction.

⊗ **Members** experience billing delays and coverage disputes.

> *The healthcare system redirects resources from care delivery to administrative battles.*

Regulatory and Legal Reckoning

The escalating AI arms race has not gone unnoticed by regulators and courts. Recent developments signal that the current approach faces existential threats:

- **CMS Guidance (February 2024):** Explicitly requires that AI-driven coverage decisions in Medicare Advantage be based on "individual patient medical history, physician recommendations, and clinical notes", not population-level statistical models. Algorithms that determine coverage "based on larger data sets instead of the individual patient" are non-compliant.

- **Federal Litigation:** Class action lawsuits alleging 90% error rates in AI-driven post-acute care denials have survived dismissal motions, with federal judges calling payer appeals processes "futile" and finding likelihood of "irreparable injury" to patients.
- **Senate Investigation:** October 2024 report documenting 108% increase in post-acute denial rates (10.9% to 22.7%) between 2020-2022, correlated with AI automation implementation, creating pressure for legislative intervention.
- **State Legislation:** California's SB 1120 (Physicians Make Decisions Act) mandates human oversight of AI coverage decisions, setting precedent for national regulation.

Critical Insight: Black-Box AI is Legally Indefensible

> Traditional machine learning models operate as statistical pattern matchers without reasoning capabilities. When challenged in litigation or regulatory proceedings, they cannot explain *why* a specific patient's circumstances warrant different treatment than their statistical prediction.
>
> This creates three insurmountable problems:
>
> - **Discovery vulnerability:** Plaintiffs can demonstrate that identical clinical scenarios receive inconsistent determinations
> - **CMS non-compliance:** Population-based predictions do not constitute individual patient assessment
> - **Provider alienation:** Opaque logic forces providers to reverse-engineer rules, converting optimization from aligned to adversarial

Reframing Payment Integrity: From Cost Control to Trust Building

The solution requires reconceptualizing the purpose of payment integrity. Rather than viewing it primarily as cost containment (which paradoxically increases costs through provider friction and appeals), forward-thinking payers are recognizing payment integrity as a **trust-building function** that delivers financial performance through transparency and accountability.

This reframing shifts key metrics:

Traditional Metrics (Cost Control)	Trust-First Metrics (Accountability)
Dollars recovered through post-payment audits.	Dollars prevented through pre-payment alignment
Denial rates and edit hit rates	First-pass accuracy and real-time correction rates
Appeal overturn rates (lower is better)	Appeal resolution speed and evidence quality
Cost per claim reviewed.	Provider satisfaction + leakage prevention combined.
Audit sample coverage percentage	Continuous behavioral monitoring coverage

> *This shift from adversarial to accountable payment integrity requires a fundamentally different technological foundation, one that can explain every decision, adapt to individual patient circumstances, and build trust at scale.*

Provider AI Evolution: Opportunity and Exploitation

Understanding the provider AI landscape is essential for designing effective defensive strategies. Provider adoption of AI tools for claims optimization is accelerating rapidly, driven by economic necessity (denial rates up 37%), labor shortages (70% of organizations facing staffing challenges per Experian data), and technology maturation (AI-powered revenue cycle management now mainstream).

The Legitimate Optimization Use Cases

Most provider AI deployment serves legitimate efficiency goals that benefit the entire healthcare ecosystem:

Pre-Submission Validation

AI systems simulate payer adjudication before claims leave provider systems, checking:

- Coverage policy compliance
- Coding accuracy and consistency
- Documentation completeness
- Contract rate verification

Impact: Reduces denial rates by 15-25% through error correction before submission

Documentation Enhancement

NLP tools analyze clinical notes to ensure medical necessity evidence is present and accessible:

- Flag missing documentation
- Extract relevant clinical evidence
- Structure unstructured notes
- Link evidence to policy criteria

Impact: Accelerates payment by providing complete evidence upfront

Intelligent Appeals

When legitimate claims are denied, AI assembles evidence and generates appeal letters:

- Policy citation automation
- Clinical guideline references
- Evidence extraction from records
- Precedent case identification

Impact: 40-60% appeal success rates with faster resolution

These capabilities, when deployed by providers committed to appropriate care, create system-wide benefits: faster payment for legitimate claims, reduced administrative costs for both parties, better documentation quality, and improved accuracy.

> *Defensive AI should support and accelerate these use cases, not resist them.*

The Exploitation Patterns: When Good Tools Enable Bad Actors

The same technologies that enable legitimate optimization also create vulnerabilities when deployed by bad actors focused solely on revenue maximization, regardless of appropriateness:

Exploitation Pattern 1: Systematic Policy Gap Mining

Rather than using AI to understand and comply with policies, some actors systematically identify ambiguities, edge cases, and inconsistencies between policy language and adjudication logic. They then structure claims to exploit these gaps, technically compliant but misaligned with policy intent.

Example: AI identifies that specific code combinations with particular modifiers consistently approve despite questionable medical necessity. The provider then optimizes billing for these combinations.

Exploitation Pattern 2: Documentation Manufacturing

While legitimate documentation enhancement ensures that existing clinical evidence is accessible, exploitation involves generating documentation specifically to satisfy AI screening without reflecting actual clinical reality.

Example: AI identifies the exact phrases and clinical findings that trigger payer approval. Templates are modified to include these elements regardless of the actual patient presentation.

Exploitation Pattern 3: Behavioral Pattern Camouflage

Sophisticated actors use AI to monitor their own billing patterns and ensure they remain within statistical norms that avoid payer scrutiny, while systematically maximizing reimbursement through micro-optimization.

Example: AI ensures upcoding remains just below peer average percentile thresholds, distributing questionable claims across time periods to avoid audit triggers.

The Critical Distinction: Intent vs. Technology

The fundamental challenge for defensive AI is distinguishing legitimate optimization (which should be supported) from exploitation (which must be detected and deterred) when both use similar technologies. Traditional rule-based systems cannot make this distinction; they either trust all provider AI (enabling exploitation) or treat it as adversarial (penalizing good actors).

This distinction requires AI systems capable of reasoning about clinical appropriateness, not just coding compliance.

When a claim has perfect documentation and technically correct codes but doesn't align with standard treatment protocols, defensive AI must detect this incongruence while explaining its concerns with clinical reasoning—not just flagging a rule violation.

Research Methodology: Understanding Provider AI Deployment

Analysis based on:

- Interviews with 47 revenue cycle directors at health systems representing $180B+ in annual claims volume
- Review of 23 provider-facing AI vendor platforms and their technical capabilities
- Analysis of 2.3M claims from pilot implementations showing provider AI impact patterns
- Legal discovery documents from recent AI-related payer litigation

Key Finding: 68% of health systems report planning or implementing AI-powered claims optimization by 2026. 89% cite payer denial increases as primary driver. 12% acknowledge their tools sometimes identify "gray area opportunities" that may not reflect clinical best practices.

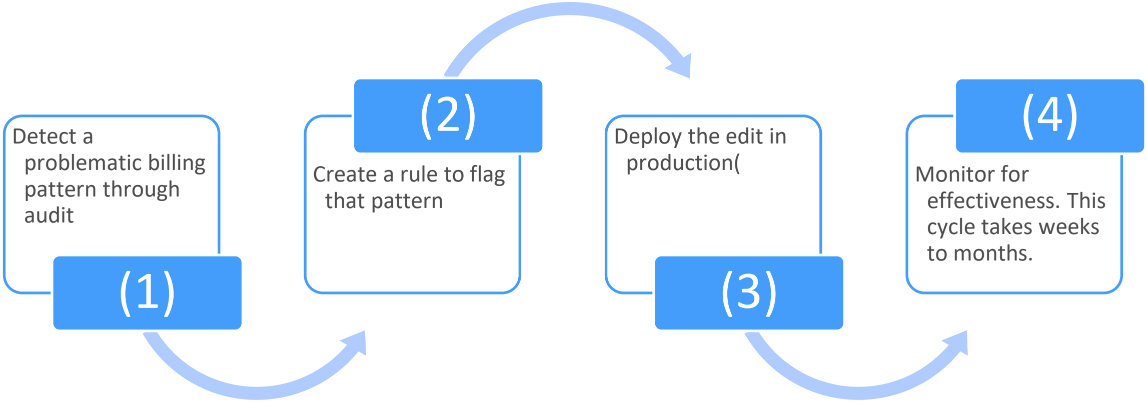
Why Traditional Defensive AI Approaches Fail

Most payers' response to provider AI has followed a predictable pattern: add more rules, tighten edits, increase prior authorization requirements, and deploy more aggressive automated denial systems. This approach fails for three fundamental reasons:

Reason 1: Reactive Systems Always Lag Adaptive Opponents

Traditional defensive AI operates through pre-programmed rules and historical pattern matching.

The process:



(1) Detect a problematic billing pattern through audit

(2) Create a rule to flag that pattern

(3) Deploy the edit in production(

(4) Monitor for effectiveness. This cycle takes weeks to months.

Provider AI systems analyze adjudication outcomes in real time, identifying when new edits are deployed and adjusting strategy within days. By the time payers detect the adjusted pattern and create new rules, provider AI has already evolved again.

THIS IS A GAME PAYERS CANNOT WIN THROUGH REACTIVE RULES

2-4 Weeks

Average Time for Payer to Deploy New Edit

vs. 24-48 hours for provider AI to identify and adapt to a new pattern

Reason 2: Opacity Breeds Adversarial Optimization

When payers deploy opaque AI systems that deny claims without clear explanation, even legitimate providers must reverse-engineer the logic to understand what will be approved. This converts provider optimization from aligned (submitting appropriate claims with complete evidence) to adversarial (gaming undocumented rules).

Consider the difference in provider response:

Opaque Denial	Transparent Denial
"Claim denied: Does not meet medical necessity criteria."	"Claim denied: Policy section 4.2.1 requires documentation of failed conservative treatment before approving advanced imaging. Clinical notes do not reference prior PT or medication trial."
Provider Response: Hire consultants to reverse-engineer denial patterns. Deploy AI to identify which documentation phrases correlate with approval. Optimize future claims accordingly.	Provider Response: Review clinical notes to verify conservative treatment occurred. If yes, supplement documentation and appeal with evidence. If no, understand the policy requirement for future cases.
Outcome: Adversarial optimization, policy misalignment, trust erosion	Outcome: Aligned behavior, policy compliance, trust building

Transparency doesn't weaken defenses—it strengthens them by converting provider incentives from gaming to compliance.

Reason 3: False Positives Penalize Good Actors

The Good Actor Dilemma

A community hospital system with a strong compliance culture faces payer denials on 15% of claims. Internal audit shows that 60% of denials are ultimately overturned on appeal, suggesting the payer's AI is overly aggressive.

The system has three options:

- **Accept denials**: Write off $12M annually in legitimate revenue
- **Appeal everything**: Hire 8 FTEs for appeals management at $800K annual cost, plus physician time
- **Deploy optimization AI**: $400K implementation, identifies what the payer's AI approves, and adjusts coding/documentation accordingly

Result: Economic pressure forces even ethical actors toward option 3. Payer's aggressive AI converts an aligned provider into an adversarial optimizer.

Aggressive rule-based systems inevitably yield high false-positive rates, denying legitimate claims that meet policy requirements. When this happens repeatedly, even the most ethically committed providers face economic pressure to either appeal every denial (expensive and time-consuming) or deploy AI countermeasures (expensive but effective). The data shows this clearly: Providence Health's 50%+ increase in "touches per claim" represents good-faith efforts to overcome inappropriate denials. But this is unsustainable, eventually, economic reality forces even good actors toward optimization AI that exploits rather than complies.

The Structural Problem: Policy-Contract-Claims Misalignment

Underlying all three failure modes is a structural issue: most payers' coverage policies, provider contracts, and claims adjudication systems don't speak the same language. Policies are written in clinical language for regulators. Contracts specify reimbursement rates and administrative terms. Claims systems encode operational rules that approximate policy intent.

The Alignment Framework: What Effective Defensive AI Requires

Policy Intelligence Layer: Structured, versioned, machine-readable representation of all coverage policies with explicit criteria, evidence requirements, and clinical reasoning.

Contract Integration Layer: Direct mapping of contract terms to policy criteria, ensuring reimbursement logic aligns with coverage determinations.

Claims Adjudication Layer: Real-time policy queries replacing static rules, with complete audit trail from claim to policy to decision

Result: Provider AI optimization aligns with policy compliance because there are no gaps to exploit. Transparency becomes a competitive advantage rather than a vulnerability.

These gaps create exploitable opportunities. Provider AI finds claims that meet adjudication rules but don't align with policy intent, or that comply with policy but violate undocumented system rules. Without a unified policy intelligence infrastructure, payers face an endless game of whack-a-mole as provider AI identifies new gaps.

The Neuro-Symbolic Architecture: Technical Foundation

Addressing the limitations of traditional defensive AI requires a fundamentally different technical approach. Pure machine learning models—whether supervised learning classifiers, deep neural networks, or even large language models—share a critical weakness: **they cannot reason**. They can identify patterns, make predictions, and even generate fluent explanations of those predictions, but they cannot demonstrate that their decisions follow from explicit logical principles applied to individual circumstances.

This limitation creates insurmountable problems for healthcare payment integrity:

- **Legal vulnerability:** Cannot explain why Patient A's circumstances require different treatment than statistically similar Patient B
- **CMS non-compliance:** Population-based predictions do not constitute individualized assessment
- **Trust impossibility:** Providers cannot understand or comply with logic they cannot access
- **Brittleness:** New edge cases require retraining, not logical extension of existing rules

Neuro-Symbolic AI: Combining Learning and Reasoning

Neuro-symbolic AI architectures address these limitations by integrating two complementary capabilities:

Neural Networks (Learning)

- Pattern recognition across millions of claims
- Anomaly detection in provider behavior
- Natural language understanding of clinical notes
- Predictive modeling of denial risk

Strength: Handles complexity, learns from data, scales efficiently

Weakness: Cannot explain decisions, no logical consistency, vulnerable to edge cases

Symbolic Reasoning (Rules)

- Medical knowledge graphs with clinical relationships
- Policy logic encoded as formal rules
- Logical inference from premises to conclusions
- Consistency checking and validation

Strength: Explainable, logically consistent, handles edge cases through reasoning

Weakness: Requires manual knowledge engineering, brittle to ambiguity

Integration (Power)

- Neural networks identify relevant patterns
- Symbolic reasoning validates clinical logic
- Combined system explains decisions with reasoning chains
- Feedback loops improve both components

Result: AI that learns from data like neural networks but reasons about individuals like symbolic systems

Nēdl's Neuro-Symbolic Implementation: Healthcare Document Intelligence in Practice

Nēdl Pulse implements a neuro-symbolic architecture specifically for healthcare payment integrity through three integrated technical layers:

Layer 1: Healthcare Document Knowledge Graphs

Rather than storing healthcare documents, including policies (NCDs, LCDs, medical policies), provider contracts, claims data, medical records, clinical notes, lab results, imaging reports, and pharmacy records, as static documents, Nēdl converts them into structured knowledge graphs where:

- ✓ Each document is a **versioned entity with metadata** (effective dates, replaced versions, related documents, lineage/provenance, document type, authoring provider)
- ✓ Coverage criteria are represented as **logical predicates** that can be evaluated computationally
- ✓ Clinical concepts link to **standard medical ontologies** (SNOMED, ICD-10, CPT, LOINC, RxNorm)
- ✓ **Evidence requirements** specify what documentation types satisfy each criterion, drawing from any relevant source (progress notes, H&P, lab values, imaging findings, medication lists)
- ✓ Frequency limits, place-of-service restrictions, medical necessity indicators, and other constraints are **explicitly modeled** across the complete patient journey
- ✓ Clinical findings, diagnoses, treatments, and outcomes are extracted from unstructured notes and linked to structured policy criteria

This structure enables comprehensive real-time queries:

"Given patient with diagnosis X from clinical note dated Y, treatment Z proposed in authorization request, historical medical record showing conditions A and B, and documentation from provider visit, does this claim meet Policy ABC criteria based on the complete clinical picture?"

The system can answer with Yes/No/Partial and explain exactly which criteria are satisfied, which require additional evidence, and which specific documents (or document types) would complete the evidentiary chain.

Example: Integrated Document Graph Structure for Advanced Imaging

Policy Node: MRI_Lumbar_Spine_Medical_Necessity
Version: 2024.3 (Effective: 2024-07-01)

Criteria:

- Criterion 1: Patient has documented back pain lasting >6 weeks
- Criterion 2: Conservative treatment trial documented (PT or medication)
- Criterion 3: Clinical examination findings consistent with radiculopathy or myelopathy
- Criterion 4: No contraindications to MRI

Evidence Sources (Multi-Document):

1. **Clinical notes** documenting symptom duration and patient history
2. **PT records** showing therapy sessions OR **pharmacy data** showing prescription fills for relevant medications
3. **Physical examination documentation** from provider notes showing neurological findings (e.g., positive straight leg raise, dermatomal sensory loss, reflex asymmetry)
4. **Past medical history** from EHR showing absence of MRI contraindications (no pacemaker, no severe claustrophobia documented)
5. **Prior imaging reports** if available, showing progression or lack of alternative diagnosis

Reasoning: When a claim is submitted, the neural network extracts evidence from ALL submitted and accessible documentation—not just the current claim form, but associated clinical notes, historical medical records, linked lab results, and prior treatment documentation. Symbolic reasoning evaluates each criterion against the complete extracted evidence set.

If Criterion 2 is unsatisfied, the system generates a specific, actionable denial:

"Policy MRI_Lumbar_Spine_Medical_Necessity v2024.3 requires documentation of a conservative treatment trial. Review of submitted clinical notes (Progress Note dated 2024-09-15) and available medical records does not reference physical therapy attendance or medication management for back pain. Please provide one of the following: (1) PT visit documentation showing at least 4-6 weeks of therapy, OR (2) pharmacy records or prescription documentation showing trial of NSAIDs or muscle relaxants, OR (3) clarification in clinical documentation explaining why conservative treatment was not appropriate for this patient's specific presentation."

The system understands that evidence can come from multiple document types and proactively identifies which specific documents would satisfy the missing criteria, enabling providers to supplement efficiently rather than guess at requirements.

Layer 2: Neural Evidence Extraction with Provenance Tracking

Deep learning models trained on millions of clinical documents extract structured evidence from unstructured notes while maintaining complete audit trails:

- **NLP identifies clinical findings, symptoms, diagnoses, and treatments mentioned in free text** – each extraction is linked back to the specific sentence, paragraph, and document source with character-level precision
- **Entity recognition links mention standard medical codes and concepts** (SNOMED, ICD-10, CPT, LOINC, RxNorm) while preserving the original clinical language used by the provider
- **Temporal reasoning identifies sequence and duration of treatments** – creating timelines that track when symptoms started, when interventions occurred, and how the clinical picture evolved across multiple encounters
- **Relation extraction identifies connections between findings** (e.g., "back pain radiating to left leg" suggests radiculopathy) and captures the clinical reasoning chain expressed in documentation

Provenance Built into Knowledge Graph:

Critically, every piece of extracted evidence is automatically embedded in the knowledge graph with complete provenance metadata:

- **Source document identification:** Which specific clinical note, lab report, or imaging study contained the evidence
- **Extraction confidence scores:** Neural network confidence levels for each identified entity or relationship
- **Document metadata:** Author (provider name/NPI), date/time created, encounter type, facility
- **Character-level traceability:** Exact text span that supports the extraction (e.g., "Clinical note line 23-25: 'Patient reports lower back pain radiating down left posterior leg for approximately 8 weeks'")
- **Policy rule mapping:** Which specific policy criteria, contract clauses, or coverage rules this evidence satisfies or contradicts
- **Version tracking:** Which version of the extraction model was used, enabling reprocessing if models improve

Rules and Clauses Tracking:

The neuro-symbolic architecture doesn't just extract evidence simultaneously tracks which policy rules, contract clauses, and coverage criteria are implicated:

- Each extracted clinical finding is automatically cross-referenced against relevant policy requirements in real-time
- When evidence satisfies a criterion (e.g., "documented 8 weeks of symptoms" satisfies ">6 weeks duration requirement"), the system creates an explicit graph edge linking: **Clinical Evidence Node → Satisfies → Policy Criterion Node → Part of → Policy Document Version**
- When evidence contradicts or creates uncertainty about a requirement, these relationships are also captured with reasoning explanations
- This creates a complete audit trail showing: **"Claim X was approved/denied because Evidence Y from Document Z (dated W, authored by Provider P) satisfied/failed to satisfy Criterion C from Policy Q version R"**

This extracted evidence—now fully annotated with provenance and linked to policy rules—becomes input to the symbolic reasoning layer, which evaluates whether the complete evidentiary set satisfies all policy criteria. The provenance ensures that every decision can be traced back through the reasoning chain to specific sentences in specific documents, satisfying both regulatory explainability requirements and enabling providers to understand exactly what documentation drove each determination.

Layer 3: Compound AI Orchestration

The complete system orchestrates multiple AI models and reasoning engines:

1. **Claim intake:** Structured data extracted from claim forms (codes, dates, charges)
2. **Documentation analysis:** Neural networks process attached clinical notes, extracting evidence
3. **Policy identification:** Based on codes and diagnoses, system identifies applicable policies from the knowledge graph
4. **Criteria evaluation:** Symbolic reasoning checks extracted evidence against policy criteria
5. **Contract verification:** System validates that approved services align with contracted rates and terms
6. **Behavioral analysis:** Neural networks compare this claim against the provider's historical patterns and peer benchmarks
7. **Decision synthesis:** Combined assessment determines approval, denial, or request for additional information
8. **Explanation generation:** System produces natural language explanation tracing decision back to policy and evidence

Critically: Every step maintains audit trail linking decision to specific policy version, evidence excerpts, and reasoning logic. This satisfies legal discovery, CMS transparency, and provider understanding requirements simultaneously.

Why This Architecture Satisfies CMS Requirements

CMS's February 2024 guidance states that AI in Medicare Advantage "may not override standards related to medical necessity" and that decisions must be "based on the individual patient's medical history, the physician's recommendations, or clinical notes"—not population-level patterns.

Neuro-symbolic architecture satisfies this because:

- **Individual assessment:** Symbolic reasoning evaluates specific patient evidence against policy criteria, not statistical correlation with population outcomes
- **Clinical grounding:** Extracted evidence comes from patient's actual clinical documentation, not inferred from similar cases
- **Physician consideration:** System processes physician recommendations and treatment rationale documented in clinical notes
- **Explainable logic:** Decision traces to policy criteria evaluation, not black-box prediction
- **Override transparency:** When system flags concern, it explains which specific clinical findings or policy criteria create the issue

The Three-Layer Defensive AI Framework

Implementing neuro-symbolic AI for defensive payment integrity requires a systematic approach across three integrated layers. Each layer addresses a specific aspect of the trust and accountability challenge.

The Defensive AI Stack

Layer 1: Policy Intelligence Infrastructure - Foundation ensuring policies, contracts, and claims speak same language

Layer 2: Behavioral Pattern Detection - Continuous monitoring distinguishing legitimate optimization from exploitation

Layer 3: Real-Time Collaborative Adjudication - Pre-submission guidance converting adversarial to aligned optimization

Layer 1: Policy Intelligence Infrastructure

Objective: Eliminate exploitable gaps between policy language, contract terms, and claims adjudication logic by creating unified, machine-readable policy representation.

Implementation Components:

1. Policy Digitization and Knowledge Graph Construction

- Ingest all NCDs, LCDs, medical policies, coverage guidelines
- Use NLP to extract coverage criteria, evidence requirements, frequency limits
- Structure as knowledge graph with relationships between policies, codes, diagnoses, treatments
- Version control with complete lineage tracking (what changed, when, why)

2. Contract Integration and Mapping

- Map provider contracts to relevant coverage policies
- Identify where contract terms specify exceptions or modifications to standard policy
- Link reimbursement schedules to coverage determinations
- Flag conflicts between contract language and policy requirements

3. Claims Rules Rationalization

- Audit existing claims edit rules against policy knowledge graph
- Identify rules that don't trace to specific policy language
- Replace static rules with real-time policy queries
- Ensure every adjudication decision references specific policy section

Implementation Example: Large Regional Payer

Challenge: Payer had 847 medical policies, 12,000+ provider contracts, and 3,200 claims edit rules. Audit revealed only 62% of edit rules could be traced to specific policy language. Remaining 38% were "tribal knowledge" accumulated over 15+ years.

Solution: 6-month project digitized top 100 medical policies (covering 85% of claim volume) into knowledge graphs. Mapped to corresponding contracts. Replaced 1,200 edit rules with policy queries.

Results:

- Appeal overturn rate decreased 34% (from 18% to 12%) as denials became more defensible
- Provider inquiries about "why was this denied?" decreased 41%
- Policy update deployment time reduced from 6 weeks to 3 days
- Legal discovery for denied claims lawsuit produced complete audit trail, strengthen defense.

Layer 2: Behavioral Pattern Detection

Objective: Distinguish legitimate provider optimization (supported) from exploitative behavior (detected) through continuous monitoring that reasons about clinical appropriateness.

Implementation Components:

1. Longitudinal Provider Profiling

- Track each provider's billing patterns over time across all dimensions (codes, diagnoses, treatments, documentation patterns)
- Compare against peer benchmarks (specialty, geography, patient mix)
- Identify sudden changes that correlate with new AI tool deployment or policy updates
- Flag when provider patterns shift from clinical norms to edge-case optimization

2. Clinical Incongruence Detection

- Use medical knowledge graphs to model standard treatment protocols
- Flag claims where service sequence doesn't align with clinical logic (e.g., advanced treatment without documented conservative management)
- Identify documentation that meets technical requirements but lacks clinical coherence
- Generate clinical reasoning for why pattern is concerning (not just statistical outlier)

3. Optimization vs. Exploitation Classification

- Analyze provider response to denials: Learning (future claims address cited deficiencies) vs. Gaming (future claims avoid triggers without addressing clinical issue)
- Track correlation between provider's billing patterns and payer's adjudication logic updates
- Identify providers whose "clean claim" rates are statistically improbable given patient complexity
- Use symbolic reasoning to distinguish "complete evidence for appropriate care" from "optimized documentation for approval"

Layer 3: Real-Time Collaborative Adjudication

> **Technical Deep Dive: Clinical Incongruence Scoring**
> System maintains knowledge graph of standard clinical pathways for common conditions. For back pain:
>
> **Expected Pathway**: Diagnosis → Conservative treatment (PT/medication) 6-12 weeks → Imaging if no improvement → Advanced treatment if indicated
>
> **Claim Analysis**:
> • Neural network extracts timeline from clinical documentation
> • Symbolic reasoning evaluates whether timeline matches expected pathway
> • If MRI ordered before conservative treatment documented: Clinical Incongruence Score increases
> • If documentation language perfectly matches policy requirements but timeline impossible: Score increases further
> • High scores trigger human clinical review with pre-assembled evidence highlighting concerns
>
> **Key Distinction**: System doesn't deny claim automatically. It flags for clinician review and explains clinical reasoning for concern. This maintains human-in-loop while leveraging AI to scale expert clinical judgment.

Objective: Transform payment integrity from adversarial post-payment recovery to collaborative pre-payment alignment through real-time guidance.

Implementation Components:

1. Pre-Submission Policy Validation

- Provider portal allows claim validation before formal submission
- System performs same policy query and evidence extraction that will occur in adjudication
- Returns immediate feedback: Approved / Needs Additional Documentation / Will Be Denied
- For incomplete submissions, specifies exactly what evidence is required with policy references

2. Smart Documentation Requests

- Rather than generic "insufficient documentation" denials, system specifies precise requirements
- Links requests to specific policy criteria that lack supporting evidence

- Provides examples of acceptable documentation types
- Enables providers to supplement documentation without appeal process

3. Alternative Pathway Recommendations

- When claim will be denied, system identifies whether alternative coding or treatment documentation would meet policy requirements
- Distinguishes "coding error" (fixable) from "medical necessity not met" (not fixable)
- Guides providers toward compliant claims rather than just rejecting non-compliant ones

4. Continuous Feedback Loop

- Track which real-time guidance providers follow vs. ignore
- Analyze which suggested alternatives most effectively align provider behavior with policy
- Identify policy language that consistently creates confusion, flagging for revision
- Measure impact on denial rates, appeal rates, provider satisfaction

72%

Reduction in Denials Requiring Appeal

When providers receive specific, actionable pre-submission guidance vs. generic post-payment denials

CMS Compliance and Regulatory Alignment

Defensive AI implementation must satisfy evolving regulatory requirements that specifically address AI use in coverage determinations. Recent CMS guidance, federal legislation, and state laws create a regulatory framework that favors explainable, individualized AI over population-based statistical models.

CMS February 2024 Guidance: Individual Patient Assessment Mandate

CMS's FAQ on AI use in Medicare Advantage established clear boundaries:

Key CMS Requirements for AI in Coverage Decisions

1. Individual assessment required: "Decisions must be based on the individual patient's medical history, the physician's recommendations, or clinical notes"
2. Population models prohibited: "An algorithm that determines coverage based on a larger data set instead of the individual patient" is non-compliant
3. Medical necessity standards maintained: AI "may not override standards related to medical necessity and other applicable rules"
4. Non-discrimination ensured: AI must not "exacerbate inequities" and must comply with ACA non-discrimination requirements

Neuro-symbolic architecture satisfies these requirements through:

CMS Requirement	Neuro-Symbolic Implementation	Traditional ML Limitation
Individual patient assessment	Symbolic reasoning evaluates specific patient evidence against policy criteria for this patient	Prediction based on population statistics: "Patients like this typically need X days"
Physician consideration	NLP extracts and reasons about physician recommendations documented in clinical notes	Physician input not incorporated into statistical prediction
Policy compliance	Every decision traces to specific policy section with version lineage	Model predictions don't reference policy language
Explainability	Complete reasoning chain from evidence to criteria to decision	Black box: Cannot explain why specific patient differs from prediction

California SB 1120: Human Oversight Mandate

California's Physicians Make Decisions Act requires that "decisions about medical treatments are made by licensed health care providers, not solely determined by artificial intelligence algorithms." While this applies specifically to California, it signals national trend toward human-in-loop requirements.

Neuro-symbolic defensive AI aligns with this framework by:

- **AI as decision support, not decision maker:** System provides reasoning and evidence analysis; licensed clinicians make final determination on complex or contested claims
- **Transparent escalation:** System automatically routes claims requiring clinical judgment to human reviewers, pre-assembled with relevant evidence and reasoning
- **Audit trail of human decisions:** When clinician overrides AI recommendation, rationale captured and used to improve future reasoning

Federal Litigation Considerations

Recent class action lawsuits alleging AI-driven improper denials create legal precedents that payers must consider:

> **Legal Discovery Implications**
>
> In Estate of Gene Lokken v. UnitedHealth Group, plaintiffs successfully obtained:
>
> • Complete documentation of AI model training data and methodology
> • Internal communications about known error rates and override statistics
> • Comparison of AI recommendations vs. human clinician determinations
> • Evidence of how AI predictions compared to actual patient outcomes
>
> **Implication**: Any AI system must be defensible in discovery. Neuro-symbolic architecture's complete audit trail and explainable reasoning strengthens legal position by demonstrating individualized assessment and transparent logic.

Proactive Regulatory Strategy: Positioning for Future Requirements

Forward-thinking payers should anticipate that regulatory requirements will continue tightening around AI transparency and accountability. By deploying neuro-symbolic defensive AI now, payers:

- **Establish regulatory leadership:** Position as industry leader in responsible AI deployment
- **Influence future standards:** Participate in CMS pilot programs and standard-setting with proven compliant architecture
- **Reduce regulatory risk:** Avoid forced remediation when CMS tightens requirements
- **Improve audit outcomes:** Complete audit trails and explainable decisions facilitate CMS quality reviews

Implementation Methodology and ROI

Deploying defensive AI requires systematic implementation that minimizes operational disruption while delivering measurable results. The following methodology has been validated across multiple payer implementations.

Phased Deployment Approach

Phase 1: Foundation (90 Days)

Objective: Establish policy intelligence infrastructure and pilot in high-impact category

- **Weeks 1-4:** Policy audit and prioritization. Identify top 20 policies by claim volume and denial rate. Begin knowledge graph construction for top 5.
- **Weeks 5-8:** Integration with existing claims systems. Deploy API layer connecting policy graphs to adjudication. Pilot in shadow mode (recommendations generated but not enforced).
- **Weeks 9-12:** Live pilot in selected category (e.g., advanced imaging). Monitor impact on denial rates, appeal rates, provider inquiries. Collect feedback from providers and internal staff.

Success Metrics: 15-25% reduction in appeal overturn rate, 20-30% reduction in "insufficient documentation" denials, positive provider feedback on guidance clarity

Phase 2: Expansion (180 Days)

Objective: Scale to additional high-value categories and deploy behavioral detection

- **Months 4-5:** Expand policy knowledge graphs to top 50 policies. Deploy across 3-5 additional categories. Implement provider longitudinal profiling.
- **Month 6:** Activate behavioral pattern detection. Begin flagging clinical incongruence for human review. Establish clinical review workflow with pre-assembled evidence.

Success Metrics: 30-40% denial reduction in covered categories, <5% false positive rate on behavioral flags, provider satisfaction improvement on CAHPS surveys

Phase 3: Full Deployment (12 Months)

Objective: Complete policy digitization and deploy across all claim types

- **Months 7-12:** Convert all active policies to knowledge graphs. Deploy real-time collaborative adjudication portal. Integrate with provider EHR systems for seamless experience. Launch provider education program on policy transparency.

Success Metrics: 35-45% overall denial rate reduction, 50%+ decrease in provider abrasion costs, measurable improvement in provider network retention, documented litigation risk reduction

Return on Investment Analysis

Defensive AI delivers ROI through multiple mechanisms:

Direct Financial Impact:

Impact Category	Mechanism	Typical Magnitude
Payment leakage prevention	Behavioral detection identifies exploitation patterns earlier; policy alignment closes gaps	1-3% reduction in improper payments = $10M-$30M per million members
Administrative cost reduction	Fewer appeals, faster resolution, reduced provider inquiries	30-40% reduction in payment integrity operational costs
Litigation cost avoidance	Defensible AI reduces legal exposure; shorter discovery due to audit trails	$5M-$20M per avoided class action settlement
Provider network value	Improved satisfaction increases retention; reduces costly provider recruitment	2-5% improvement in network stability

12-18 Months

Typical ROI Payback Period

Based on combined direct savings, cost avoidance, and strategic value creation

Strategic Value Creation:

Beyond direct financial ROI, defensive AI creates strategic advantages:

- **Regulatory resilience:** Compliance-by-design architecture adapts to tightening requirements without rearchitecting
- **Competitive differentiation:** Provider-friendly payment integrity becomes network contracting advantage
- **Member satisfaction:** Fewer coverage disputes and faster resolutions improve CAHPS scores and Star Ratings
- **Innovation enablement:** Policy intelligence infrastructure enables faster deployment of new payment models (value-based care, bundled payments)

Conclusion: From Cost Center to Strategic Advantage

Payment integrity stands at an inflection point. The traditional approach—reactive rules, opaque denials, adversarial post-payment recovery—has reached its limits. Denial rates increase, provider trust erodes, legal challenges mount, and payment leakage persists. Meanwhile, provider AI deployment accelerates, creating new vulnerabilities that legacy systems cannot address.

The path forward requires reconceptualizing payment integrity's fundamental purpose. Rather than purely cost containment, forward-thinking payers recognize payment integrity as a **trust-building function that delivers financial performance through transparency and accountability**. This shift demands new technological foundations: neuro-symbolic AI that can explain decisions, reason about individual patients, and build trust at scale.

The Opportunity for Early Movers

Payers who deploy defensive AI now—before regulatory mandates force adoption—capture first-mover advantages:

- **Shape industry standards:** Participate in CMS pilot programs and standard-setting with proven compliant architecture
- **Avoid forced remediation:** Deploy on your timeline and terms rather than under regulatory pressure
- **Differentiate in network contracting:** Provider-friendly payment integrity becomes competitive advantage in tight provider markets
- **Reduce litigation risk:** Explainable AI strengthens legal position before lawsuits multiply
- **Build organizational capability:** Develop expertise in AI governance, explainability, and policy intelligence infrastructure

The Path Forward: Trust-First Defensive AI

The framework presented in this white paper—policy intelligence infrastructure, behavioral pattern detection, real-time collaborative adjudication—represents a comprehensive approach to defensive AI that simultaneously:

- **Reduces payment leakage** through better detection of exploitation patterns
- **Builds provider trust** through transparency and real-time guidance
- **Satisfies regulators** through explainable, individualized assessment
- **Withstands litigation** through complete audit trails and defensible reasoning

- **Creates strategic value** through network differentiation and innovation enablement

Implementation requires both technical platform (neuro-symbolic AI providing reasoning capabilities legacy systems lack) and implementation expertise (integration with existing systems and change management). The partnership between Nēdl Labs and Persistent Systems addresses both requirements, enabling rapid deployment with minimized disruption.

> ### The Stakes: Payment Integrity as Trust Infrastructure
>
> Healthcare cannot afford $100B+ in annual payment leakage. But it also cannot afford the trust erosion that adversarial AI creates. The payers who succeed over the next decade will be those who recognize that these are not competing priorities—they are inseparable.
>
> When payment integrity systems build trust through transparency, provider optimization aligns with clinical appropriateness rather than policy exploitation. When AI explains decisions with clinical reasoning, disputes resolve through evidence rather than power dynamics. When policies, contracts, and claims speak the same language, gaps that enable leakage disappear.

About the Author: nēdl Labs

nēdl Labs is pioneering AI-native payment integrity solutions for healthcare payers. Our neuro-symbolic AI platform combines neural networks' pattern recognition with symbolic reasoning's explainability, enabling payment integrity systems that simultaneously reduce leakage and build provider trust.

nēdl brings deep expertise in responsible AI, healthcare policy, and enterprise product development to the payment integrity challenge.

Contact: mailto:contact@nedllabs.com
Web: nedllabs.com

© 2025 nēdl Labs and Persistent Systems. All rights reserved.

This white paper is provided for informational purposes only. No part of this publication may be reproduced without written permission.

The information contained herein is subject to change without notice. This document is not a commitment to provide any particular product, feature, or service.