A R C L I G H T I N S I G H T S ™

Design Before Deploy

Objective Function Governance for AI-Assisted Medicare Review

How WISeR, Private-Payer AI Denials, and Medicare Audit Appeals Reveal Why High-Stakes Public AI Systems Must Be Co-Created Before Procurement

Lance McNeill, MBA, MPAff

April 2026

For correspondence: www.ArclightAction.com

Abstract

Public agencies are beginning to use AI-assisted tools in high-stakes medical review, payment, and benefit determination systems. The central governance question is not whether AI can reduce administrative cost or improper spending. It is whether these systems are designed to optimize the correct public-purpose objective before they are procured and deployed. Medicare's WISeR model — a six-year CMS Innovation Center initiative using AI and machine learning alongside human clinical review in Original Medicare — provides an important test case. WISeR is not an automated denial system, and CMS has included safeguards such as licensed clinician review and financial penalties for inaccurate non-affirmations. But the model raises an objective function governance problem: vendor compensation is tied in part to averted expenditures, while public materials do not show a structured pre-procurement co-creation process through which affected clinicians, providers, beneficiaries, and adjudicators helped define the objective function, edge cases, rationale requirements, and access safeguards.

This paper argues that co-creation is not stakeholder consultation after a model is designed; it is the design process itself. Borrowing from Danielle Citron's concept of technological due process and the Biden White House Blueprint for an AI Bill of Rights, it proposes a governance framework that goes further: nothing should be designed for us without us. Before any Medicare AI review system is procured, a stakeholder map must be completed, a co-creation process must define the objective function, and a governance specification must bind the RFP. The paper proposes six governance conditions: stakeholder mapping and co-created objective functions; co-creation before procurement; training on independent adjudication rather than legacy denial behavior; a persistent rationale ledger; incentives tied to accuracy, access, and fairness; and an independent AI assurance monitor separate from the frontline system. It concludes with three implementation tracks: a WISeR oversight retrofit, a governance protocol for future AI review models, and a rationale-ledger pilot for Medicare audit appeals.

Executive Summary

The paper's core claim is simple: **co-creation and stakeholder engagement are not compliance activities to be scheduled after an AI model is procured, trained, or deployed. They are the design process itself.**

The danger is not that AI will review Medicare claims. The danger is that AI will be optimized around the wrong institutional reward before anyone who bears the consequences of that optimization had a seat at the design table. AI governance scholars call this objective function misspecification. In Medicare review, 'averted spending' may be evidence of improved accuracy when inappropriate care is correctly identified. But if averted spending becomes the optimization target, the system may learn to produce non-authorization rather than correctness. The distinction between 'savings as evidence of accuracy' and 'savings as the optimization target' is the difference between a system that learns to be right and a system that learns to say no.

The governance framework this paper proposes extends the AI Bill of Rights from post-deployment protections to pre-procurement co-creation. The Bill of Rights asks whether people receive safe systems, notice, explanation, and human fallback. This paper asks an earlier question: who got to define what the system was built to optimize in the first place? The answer, for both the private-payer AI denial systems now in federal litigation and for WISeR's six-state deployment, is: not the people who bear the consequences.

One-Page Governance Map

THE PROBLEM	WHY IT MATTERS	GOVERNANCE FIX
Objective function mis-specified around savings	If averted spending is the optimization target, the system learns to say no rather than to be right. Legitimate waste reduction requires accurate determinations, not high denial rates.	Co-create the objective function with affected stakeholders before procurement. Define what the system must optimize, what it must not optimize, and which proxy metrics are unsafe if used alone.
Stakeholders who bear consequences had no design role	Clinical edge cases, beneficiary access risks, and workflow failures are invisible to system designers who have not engaged the people who encounter them daily.	Require a stakeholder map as the first deliverable of any AI review initiative. Conduct facilitated co-creation sessions with proactively recruited domain experts before any vendor solicitation.
Human clinical review may become rubber-stamping	A physician signature is not meaningful review if the workflow is designed for rapid ratification of algorithmic output. The two-wound edge case illustrates how reviewers may not scrutinize decisions closely enough.	Define meaningful human review standards: override authority, clinical information, logged disagreement, and productivity metrics that do not reward speed. Add an independent AI assurance monitor to flag discrepancies for closer scrutiny.
Rationale changes across review stages without accountability	Providers cannot respond to a moving target. The denial theory shifts across UPIC, MAC, QIC, and ALJ levels with no public	Require a persistent rationale ledger tracking the operative theory, evidence basis, and any modification with logged authority across all review stages.

	mechanism for tracking whether it changed.	
AI trained on contractor behavior reproduces contractor errors	A model trained on legacy denial patterns learns what contractors historically denied — not what independent adjudicators determine is correct.	Train on ALJ outcomes, clinician-labeled edge cases, and clinical literature. Evaluate against ALJ concordance, not contractor agreement.
No independent check on frontline AI accuracy	The same vendor with savings-based incentives operates the model that produces determinations. There is no independent, real-time view of whether the system is producing accurate, equitable results.	Establish an independent AI assurance monitor operated by a party with no savings-based compensation. Publish near-real-time accuracy, access, and burden metrics on open data portals.

The specific ask: CMS and CMMI should not expand AI-assisted Medicare review models until a formal co-creation governance phase has been completed, a rationale-ledger requirement has been published, and vendor performance is tied to accuracy, access, and appeal outcomes rather than savings alone. An independent AI assurance monitor should be required before any AI-assisted review system is operational at scale.

I. Introduction: The Objective Function Problem in Public AI

In 2014, philosopher Nick Bostrom posed a now-famous thought experiment. Imagine an AI system given a single objective: maximize the production of paperclips. Given sufficient capability, the system would convert all available resources — including humanity — into paperclip-production infrastructure. The point is not that AI will literally optimize for paperclips. The point is that a system given a mis-specified objective will pursue it with indifference to every value the designers assumed were obvious but never stated. The failure is not in the AI. It is in the governance process that specified the objective. [18]

Medicare's WISeR model does not involve paperclips. But it illustrates the same governance risk at a smaller and more consequential scale. When a public AI review system is designed around 'averted expenditures' as its primary vendor compensation metric, the objective function is not 'accurately identify clinically inappropriate services.' It is, instrumentally, 'generate non-authorizations.' Those are not the same thing. The first requires clinical judgment, evidence evaluation, and correctness. The second requires denial volume. If the governance structure does not constrain the second with equal force to the first, the system will drift toward what it is rewarded to produce.

This paper is about that governance structure. Specifically, it is about the design process that must precede any public AI system affecting payment, coverage, or access to care: a structured, proactive, multi-stakeholder co-creation process that defines the objective function, the evidence standards, the edge cases, the rationale requirements, and the accountability mechanisms before a vendor is selected or a model is trained.

The central governance question is not how to reduce waste with AI. It is: **who defines what the system is built to optimize, who gets to participate in that definition, and what accountability structures ensure the deployed system stays faithful to the public-purpose objective rather than the proxy metric that drives vendor revenue?**

The federal government is beginning to deploy AI-assisted tools in Medicare review before fully answering those questions. Medicare's WISeR model is the clearest current test case, but it will not be the last. If the same governance sequence — internally designed model, limited comment period, multi-state deployment, post-launch problem discovery — is applied to AI-assisted Medicare audit and appeals review, the structural failures visible in WISeR will be replicated in a system with even higher stakes for providers who face recoupment demands and patients whose care depends on the outcome.

This paper proposes a governance framework grounded in three foundational principles. The first draws on Danielle Citron's concept of technological due process: when automated systems materially influence individual rights or benefits, affected parties require notice, explanation, human review, and accountable appeal. [19] The second comes from the Biden White House Blueprint for an AI Bill of Rights, which establishes five protections — safe and effective systems, discrimination protections, data privacy, notice and explanation, and human alternatives with fallback — as the minimum floor for rights-impacting AI. [20] This paper goes further, because floors are not enough. The third principle is the one that post-deployment protections cannot provide: nothing designed for us without us. The people who bear the consequences of an AI system's outputs must have genuine influence over what that system is built to optimize before it is built.

THE THROUGH-LINE

The distinction between 'savings as evidence of accuracy' and 'savings as the optimization target' is the difference between a system that learns to be right and a system that learns to say no. Specifying which objective applies — and designing the governance infrastructure to keep the system faithful to it — is not a technical problem. It is a governance problem. It must be solved before procurement, not after deployment.

II. What Private-Payer AI Denial Cases Teach

The AI denial controversy in American health care is often framed as a technology story: automation outpacing regulation. That framing is incomplete. The problem is not that AI entered claims review. The problem is that AI entered claims review inside incentive and workflow structures that rewarded speed, savings, and throughput without equally enforceable safeguards for individualized review, clinical context, and appealable reasoning, and without meaningful participation by the people those systems would affect.

Cigna PxDx: Speed Without Review

ProPublica reported in 2023 that Cigna's PxDx system allegedly allowed company physicians to deny large volumes of claims using an automated procedure-to-diagnosis matching process, with reported average review times of approximately 1.2 seconds per case and more than 300,000 claims denied over two months. [6] Cigna disputed aspects of the reporting, and litigation followed. The governance lesson does not require accepting every allegation as proven.

A physician signature is not the same as clinical judgment. A model-generated denial is not procedurally fair merely because a licensed professional was technically present in the workflow. If the workflow is designed for rapid ratification of algorithmic output — if productivity metrics reward speed over scrutiny, if the AI output is treated as presumptively correct — then 'human in the loop' becomes a legal formality rather than a substantive safeguard. Meaningful review requires override authority, clinical information, logged reasoning, and institutional protection against rubber-stamping.

UnitedHealth nH Predict: The Training Data Problem

Litigation involving UnitedHealth's nH Predict system has alleged that a post-acute care prediction tool was used in ways that constrained Medicare Advantage coverage determinations, with plaintiffs claiming high appeal reversal rates for some denials. [8] UnitedHealth has denied that the algorithm made coverage decisions. A federal magistrate judge ordered broad discovery into the model's development, training, validation, and deployment in March 2026. The allegations remain unproven.

The governance lesson is about the training data question that the litigation is now forcing into court. A model trained primarily on incumbent denial behavior may reproduce incumbent denial behavior. What the model was trained to optimize is a governance decision that must be answered before procurement, not extracted through federal court orders after deployment.

The Shared Governance Failure

In both cases, the people who bear the consequences of the system's outputs had no meaningful role in designing it. No stakeholder map was created before the system was specified. No co-creation process surfaced edge cases, defined evidence standards, or established the objective

function in consultation with those affected. No independent assurance layer monitored whether the deployed system was producing accurate results. The problems emerged through litigation, patient harm, and congressional attention. The governance failure preceded and caused the operational failure. Public systems should understand this sequence before reproducing it.

THE GOVERNANCE SEQUENCE THAT FAILED

Design by vendor → Deploy to affected parties → Discover problems through harm → Reconstruct governance through litigation. This sequence is not inevitable. It is a choice. The alternative sequence is what this paper proposes: stakeholder map → co-created objective function → governance specification → procurement → controlled deployment → independent assurance → continuous measurement.

III. WISeR: The Public-Sector Test Case

CMS describes the WISeR (Wasteful and Inappropriate Service Reduction) model as a voluntary Innovation Center initiative using AI and machine learning alongside human clinical review to ensure timely and appropriate Medicare payment for selected services in Original Medicare. [1] WISeR runs from January 1, 2026 through December 31, 2031 in six states: New Jersey, Ohio, Oklahoma, Texas, Arizona, and Washington. [2] CMS requires all non-payment recommendations to be made by appropriately licensed clinicians using evidence-based procedures, and the model includes quality adjustments and financial penalties if denied claims are later overturned. [1] These are meaningful safeguards. WISeR is not the same as the private-payer cases described above.

Still, WISeR creates an objective function governance problem. According to CMS's own Request for Applications, model participants receive a percentage of the reduction in expenditures attributable to non-affirmed requests that do not later become paid claims. [1] KFF has described this as a model where technology companies are eligible to receive a share of savings associated with services determined to be wasteful, while noting that prior authorization can create delays, denials of needed care, uncertainty for patients, and administrative burden for providers. [3] When the primary revenue driver for a WISeR vendor is net savings generated through non-authorization, the objective function is not neutral. Quality adjustments and clawbacks modify it at the margins; they do not change what the model's economics are built around.

Duration Is Not Design

WISeR's six-state, six-year deployment is described as a pilot. But a true pilot is a controlled experiment with a narrow scope designed to test a hypothesis before broader deployment. Six states at once — affecting potentially hundreds of provider organizations and thousands of beneficiaries from day one — is not an alpha test. It is a phased national launch.

For AI systems affecting access to medically necessary care, the design sequence should have included: shadow-mode review in which AI outputs are observed but not acted on; limited beta deployment with a small number of volunteer providers in one or two service lines; edge-case stress testing with clinical panels; workflow validation with providers who will experience the authorization process; and an independent assurance layer before vendor compensation is tied to savings. None of these stages is visible in WISeR's public design documentation.

CMS states that pilot states were selected for comparison feasibility, service volume, geographic diversity, and MAC/LCD/NCD considerations — a defensible evaluation rationale. [4] A model

can be statistically evaluable and still be operationally under-designed. The question is not whether WISeR can be measured. The question is whether it was appropriately tested before it began making determinations that delay or deny care to beneficiaries.

> **FIELD NOTE: A Multi-Wound Prior Authorization Edge Case**
>
> In one wound care case observed firsthand by the author, a Medicare beneficiary presented with two distinct wounds requiring separate treatment episodes. A prior authorization determination associated with one wound created downstream barriers when the provider attempted to treat the second wound. When the provider contacted the Medicare contractor to clarify that these were two distinct clinical sites, the response was that the prior authorization outcome could not be corrected administratively and the provider would need to proceed through the formal appeals process. A wound care clinician examining this case would immediately recognize two distinct wounds as two distinct clinical problems. But the review workflow, built around episode-level authorization logic, processed them as a single unit. This is exactly the class of edge case a structured co-creation process — with wound care clinicians in the room before design — would have identified and addressed. It is also exactly the class of case where a second, independent AI assurance monitor would flag a discrepancy for closer human scrutiny: the frontline system applies one authorization logic; the assurance system recognizes a clinical distinction; the discrepancy triggers a genuine clinical review rather than a rubber-stamp ratification.

The early warning signals from WISeR's first months are consistent with the objective function concern. The American Hospital Association raised concerns before launch about payment structure, appeal rights, AI oversight, and timeline. [4] In April 2026, Senator Maria Cantwell reported Washington hospital accounts of authorization timelines expanding from roughly two weeks to four to eight weeks for some procedures, along with concerns about inconsistent denials and unclear reasoning. [5] CMS delayed implementation of two WISeR services in April 2026. These signals do not prove failure. They show that the governance questions the co-creation process should have resolved before launch are now being resolved through political friction after it.

IV. From Prior Authorization to Audit Appeals: Why the Governance Problem Travels

WISeR is a pre-payment, prior authorization model. Medicare audit and appeals review — the UPIC/MAC/QIC/ALJ chain — is a post-payment review and appeals system. They are different workflows, different legal authorities, and different points in the claims lifecycle. But they share the same governance vulnerability: high-stakes determinations made by systems whose objective functions, stakeholder inputs, and accountability structures have not been co-designed with the people who bear the consequences.

If CMS or CMMI were to deploy AI into the audit and appeals process using the same design sequence applied to WISeR — internal model development, limited stakeholder comment, multi-contractor deployment, post-launch problem discovery — the result would be predictable and worse. The audit appeals chain involves recoupment demands, provider financial distress, and retrospective denial of claims that were already paid, sometimes months earlier. The harm from a wrong determination compounds over time rather than being visible immediately. The feedback signal that something is wrong is slower, more filtered, and harder to connect to the upstream decision that caused it.

Understanding the specific fragility of the audit appeals architecture requires a brief orientation. When Medicare denies or recoups a claim, the provider can request a redetermination from the MAC, then a reconsideration from a QIC, then an ALJ hearing at OMHA, and finally a review by the Medicare Appeals Council. UPICs investigate fraud, waste, and abuse and refer cases to MACs. Each actor in this chain operates under different authority, different performance metrics, and different incentives. No actor in the chain is publicly evaluated on whether its determinations agree with what an ALJ independently concludes about the same cases. No closed feedback loop carries ALJ corrections back to upstream contractors.

Rationale Drift: The Visible Symptom

Rationale drift — the shift in the operative denial theory across successive appeal levels — is the most visible procedural symptom of this deeper structural failure, not the root cause. Under 42 CFR §405.968(b)(5), QICs may raise and develop new issues relevant to the claims in a case. That authority serves legitimate completeness goals. But it also means the theory a provider prepared to contest at Level 2 may not be the theory on which the denial is sustained. No public metric tracks whether the operative denial theory remained stable across stages.

The Q2A data makes the measurement failure visible. In Q3 2025, 85.0 percent of favorable Part A QIC reversals, 92.6 percent of Part B reversals, and 98.9 percent of DME reversals were coded as ‘found new documentation or evidence persuasive.’ That single code does most of the explanatory work for favorable reversals in the current system — and it cannot distinguish late evidence submission from changed legal reasoning. Prior Arclight analysis estimates hidden annual systemic costs from reversed or likely reversible UPIC-related determinations at $49 million to $250 million, with a conservative zero-drift floor of $29 million to $65 million. [13] These are not just dollar figures. They represent the provider burden, patient disruption, and OMHA caseload generated by a system with no accountability loop between what ALJs correct and what upstream contractors are trained to do.

AI deployed into this architecture without rationale governance will not solve the fragmentation problem. It will accelerate it. When a human QIC reviewer shifts the legal basis for a denial, there is at least a human accountable for that shift. When an AI system generates a reconsideration output resting on a different theory than the initial denial, the shift may be structurally invisible without a persistent rationale ledger specifically designed to surface it. The two governance failures: WISeR’s objective function misalignment and the audit appeals chain’s rationale accountability gap are the same problem expressed in different workflows. The governance framework required to address both is the same framework.

THE ROOT PROBLEM

The actors who make initial Medicare payment determinations are evaluated on volume and speed, not on accuracy as measured by independent adjudication. No closed feedback loop connects what ALJs decide back to what upstream contractors do. AI deployed into this architecture without governance reform will not improve accuracy, it will reproduce existing errors faster, at higher volume, with less visibility into where they originated.

V. Nothing Designed for Us Without Us: Co-Creation as the Design Process

Co-creation is not a survey. It is not a public comment page. It is not a listening session after the RFP has already been written. Those mechanisms gather input from whoever chooses to

respond. Co-creation requires that affected stakeholders be proactively identified, directly invited, and genuinely engaged before the design is established, not consulted after it is fixed.

The principle 'nothing designed for us without us' is older than AI governance. It comes from disability rights advocacy, from community organizing, from participatory design. Applied to public AI systems, it means: before any Medicare AI review system is procured, a complete stakeholder map must be developed, the affected groups must be meaningfully engaged, and the objective function must be co-defined rather than handed down. This is not stakeholder courtesy. It is the governance mechanism that prevents objective function misspecification, surfaces clinical edge cases before they harm patients, and builds the political and clinical legitimacy that makes the system defensible when it is challenged.

The AI Bill of Rights asks whether people affected by AI systems receive safe systems, notice, explanation, and human fallback. [20] Those are minimum post-deployment protections. This paper asks an earlier question: who got to define what the system was built to optimize in the first place? For WISeR and the private-payer cases before it, the answer is: not the people who bear the consequences. That is the governance gap this framework closes.

The Stakeholder Map: First Deliverable of Every Initiative

Before any facilitated session, before any RFP, the first required artifact of a Medicare AI review initiative should be a stakeholder map: an explicit identification of every group affected by the system, what they know, what they risk, what incentives they face, and how they must be engaged. The table below shows what a stakeholder map for a wound care AI review initiative should include:

STAKEHOLDER	WHAT THEY KNOW	WHAT THEY RISK	HOW TO ENGAGE
Wound care clinicians	Clinical standards, multi-wound complexity, documentation logic, treatment sequences	Wrong determinations harm patients; workflow disruption; defensive documentation burden	Direct outreach to leading voices; structured clinical panels; prototype testing with real case scenarios
Medicare beneficiaries / caregivers	Access delays, continuity of care disruption, lived experience of denials and appeals	Delayed or foregone care; financial exposure; inability to navigate appeals	Patient advisory sessions; caregiver interviews; benefit-claims journey mapping
Independent wound care providers	Mobile practice realities, documentation constraints, prior auth friction, billing patterns	Recoupment demands; cash flow disruption from delayed payment; audit burden	Provider workflow sessions; mobile practice focus groups; documentation burden surveys
QIC / ALJ-experienced practitioners	What makes a rationale legally sufficient; how denial theories shift across levels; what first-pass	If AI encodes legally insufficient rationales, appeal volume rises systemically	Structured expert interviews; rationale schema co-design sessions; edge-case labeling workshops

	notices fail to communicate		
Program integrity professionals	Fraud and abuse risk signals; documentation red flags; historical abuse patterns by service category	If AI misidentifies fraud patterns, legitimate claims are caught; actual fraud may evade detection	Risk-scenario workshops; fraud pattern review against AI training data
Data scientists (no vendor relationship)	Model feasibility; training data quality; proxy variable risks; bias detection	Poorly specified models encode proxy variables that produce inequitable results	Technical co-design sessions; independent model-risk review before procurement
Health equity experts	Subgroup error patterns; rural and underserved population effects; algorithmic bias identification	Disparate impact on vulnerable populations goes undetected without explicit monitoring	Equity review of proposed metrics; subgroup analysis design; disparate-impact monitoring specification

The Co-Creation Process

The Co-Creation Process (such as the Strategic Co-creation process developed by Humantific or the Human-Centered Design framework developed by IDEO and the Stanford d.school) provides the operational methodology for translating the stakeholder map into a governance specification. It is an eight-step framework built around three core behaviors applied across all phases: active divergence (generating options without applying judgment), deferral of judgment (withholding evaluation to allow open thinking to flourish), and active convergence (selecting key alternatives from a broad range of possibilities). The table below shows how each step applies to Medicare AI governance:

CO-CREATION STEP	APPLIED TO MEDICARE AI GOVERNANCE
1. SCOUT Diverge on challenges, risks, and stakeholder pain points	Identify failure modes. What keeps wound care clinicians up at night about AI-assisted review? What are the edge cases that existing logic mishandles? What does an ALJ see that contractors miss? Gather these without filtering.
2. CLARIFY Gather facts; converge on what matters most	Collect the evidence base: QIC reversal data, UPIC determination patterns, ALJ outcome distributions, clinical literature, documentation burden. Distinguish what is known from what must be found out.
3. FRAME Use altitude laddering to find the right challenge level	Ask Why Up (the reasons we want to address this) and Why Down (what is preventing us). Is the real challenge accuracy? Access? Rationale stability? Incentive alignment? Frame the problem at the right altitude before building a solution, or risk solving the wrong thing precisely.
4. CONCEIVE Generate governance design options with deferred judgment	Diverge on multiple models for the rationale ledger, the co-creation council structure, incentive alignment approaches, and assurance

	monitoring designs. The governance innovations worth pursuing are often inside the options that initially seem impractical.
5. PROTOTYPE Test governance concepts; iterate with real scenarios	Build draft rationale schemas. Test denial notice templates with provider panels. Apply the multi-wound edge case to the proposed review logic. Prototype the independent assurance monitor's discrepancy flagging criteria. Iterate until the design holds up to real clinical scenarios.
6. PLAN Build the implementation plan with clear milestones	Draft the governance specification document. Map each governance condition to its implementation vehicle. Define the council's charter and deliverable accountability. Never end a session without clear next steps.
7. BUILD Deploy the controlled pilot with governance specification as the binding contract	Launch with the governance specification in place, not alongside it. The specification is the contract the technology must satisfy before live determinations begin.
8. MEASURE Monitor; feed learnings back as input to next cycle	Track rationale concordance, ALJ affirmance rates, provider burden, beneficiary access delay, independent assurance discrepancy rates. Feed learnings back to the council as the Scout input for the next improvement cycle.

Two process principles deserve emphasis. The first is altitude framing: before settling on a specific design question, ask why we want to address it and what is preventing us. This surfaces the real challenge and prevents solving the wrong problem with great precision. The second is the prohibition on killer phrases: premature convergence kills the superior but risky options that contain the governance innovations worth pursuing. Protected divergence time is not inefficiency. It is the mechanism for finding answers that a conventional procurement process cannot produce.

VI. Six Governance Conditions for AI-Assisted Medicare Review

The following conditions are implementable design requirements, not aspirational principles. A preliminary note: several — especially the rationale ledger — have value regardless of AI deployment. They are the minimum governance infrastructure Medicare payment review needs whether the review is conducted by human contractors, a machine learning model, or a combination. AI deployment makes them urgent. It does not make them novel.

Condition 1: Stakeholder Mapping and Co-Created Objective Function

Before any Medicare AI review system is specified, a complete stakeholder map must be developed as the first formal deliverable. The map identifies every group affected by the system, what they know, what they risk, what incentives they face, and how they must be engaged. It is not an administrative checkbox. It is the governance mechanism that determines whether the system's objective function is defined by the people who will bear its consequences.

The objective function itself must be co-created, not handed down. For Medicare review, the correct objective is not 'reduce wasteful spending.' It is: accurately apply Medicare coverage, documentation, and medical necessity rules in a way that protects beneficiary access, supports clinically appropriate care, reduces genuinely inappropriate utilization, and uses Medicare funds

prudently. Savings are a byproduct of accuracy, not its proxy. The governance specification must state explicitly what the system is allowed to optimize, what it must not optimize, and which proxy metrics are unsafe if used alone.

Condition 2: Co-Creation Before Procurement

CMS or CMMI should constitute a co-creation council before a vendor solicitation is finalized. The council should not merely advise on a completed design. It should produce the governance specification that binds the RFP, the vendor evaluation, the model design, the pilot metrics, and the public reporting obligations. CMS would need to structure the council consistent with Federal Advisory Committee Act requirements or through an appropriate technical expert panel, CMMI stakeholder mechanism, or other lawful convening authority. The governance form matters less than the governance substance: genuine stakeholder influence over the objective function before procurement, with public accountability for how the council's outputs were incorporated.

Condition 3: Train on Independent Adjudication, Not Only Legacy Denial Behavior

An AI-assisted review model should not be trained primarily on incumbent contractor decisions. Those decisions may contain the very errors the system is meant to correct. The training and validation stack should include: ALJ favorable, partially favorable, and unfavorable decisions disaggregated by reason; clinician-labeled edge cases developed through co-creation; peer-reviewed clinical evidence and specialty society guidance; National Coverage Determinations, Local Coverage Determinations, billing articles, and CMS manuals; and random samples of non-appealed denials and paid claims to reduce selection bias.

ALJ outcomes are not perfect ground truth. They represent a selected set of disputed cases. But ALJ adjudicators are the first genuinely independent reviewers in the Medicare appeals chain — conducting de novo review without contractor incentives — and their outcomes are the closest available independent benchmark for determination accuracy. They should be central to model validation, not ignored.

Condition 4: Require a Persistent Rationale Ledger

Every AI-assisted determination should generate a rationale ledger that persists across review stages and is accessible to the provider at each stage without requiring escalation. The ledger records: the governing authority relied upon; the operative denial or non-affirmation theory in plain language; the evidence reviewed and the specific evidence missing; the explicit overturn criteria; the model's confidence level if AI-assisted; the human reviewer's role and attestation with logged reasoning; any later rationale modification and its authority; and the appeal outcome and reason for reversal or affirmance. The full field list appears in Appendix B.

The rationale ledger is valuable before and independent of any AI deployment. It is the mechanism that makes rationale drift visible and measurable. Without it, Medicare cannot distinguish late-submitted evidence from a changed denial theory. Implementing it through QIC contract modifications and MPIM guidance updates requires no rulemaking.

Condition 5: Tie Incentives to Accuracy, Access, and Fairness — Not Savings Alone

Vendor performance should be measured through a balanced scorecard: clinical accuracy; ALJ or independent review concordance; rationale stability; beneficiary access delay; provider burden; explanation quality; reversal and resubmission rates; clinician override rates; and disparate-

impact metrics. Averted spending should be treated as evidence of value only when paired with evidence that medically necessary care was not delayed or denied.

Meaningful human review is a condition within this condition. For a human review safeguard to count, the reviewer must have override authority; sufficient time and clinical information to exercise it; an obligation to log agreement or disagreement with reasons; and institutional protection against productivity metrics that reward ratification speed. Override rates and affirmance rates should be subject to periodic public audit.

Condition 6: Independent AI Assurance Monitor

This paper proposes a governance condition not present in current federal AI frameworks for Medicare review: an independent AI assurance monitor that reviews the outputs of the frontline AI-assisted review system, operated by a party with no savings-based compensation and no financial relationship with the frontline vendor.

The architecture works as follows. The frontline AI-assisted review system evaluates the prior authorization request or claim. The licensed human clinical reviewer reviews non-affirmations and edge cases. The independent AI assurance monitor — a separate model with separate incentives — reviews determinations for inconsistency, missing clinical context, rationale instability, anomalous denial patterns, and divergence from Medicare rules. When the assurance monitor identifies a discrepancy between its assessment and the frontline system's output, the case is flagged for enhanced human clinical review rather than routine ratification. Aggregate discrepancy rates, denial patterns, and access delay metrics are published near real-time on open government data portals.

The two-wound edge case illustrates why this matters. A wound care clinician examining two distinct wounds would immediately recognize two distinct clinical problems. But if the workflow is designed for rapid ratification, the human reviewer may not scrutinize the determination closely enough to catch the distinction. An independent assurance monitor, applying different pattern logic, is more likely to flag the discrepancy and trigger the deeper review that the human reviewer should have provided. The assurance monitor does not make the determination. It protects the integrity of the determination process.

CMS, OIG, a CMMI-designated public-interest evaluator, or a separately contracted third party with no savings-based compensation should operate the assurance monitor. Its findings should be published without trade-secret barriers and should be accessible to OIG, GAO, and academic evaluators.

VII. Three Implementation Tracks

WISeR-like prior authorization and UPIC/QIC-style audit appeals are distinct workflows requiring distinct governance approaches. A third track — addressing WISeR's existing deployment — is also needed, because the governance questions that should have been resolved before launch must now be addressed while the model is operational. This paper separates these three tracks rather than treating them as a unified pilot framework.

Track 1: WISeR Oversight Retrofit (Immediate)

Because WISeR is already operational across six states, the immediate governance priority is not redesign but oversight augmentation. The objective function problem cannot be fully corrected mid-pilot, but its risks can be monitored and constrained.

- Establish an independent AI assurance monitor before the next service line or state is added to WISeR. The monitor should be operated by a party with no savings-based compensation.
- Publish near-real-time metrics on authorization turnaround time, non-affirmation rates by service category, resubmission rates, reversal rates on appeal, clinician override rates, and beneficiary access delay. These should live on existing CMS open data portals.
- Constitute a clinical edge-case review panel for each WISeR service category. The panel should include specialty clinicians who did not participate in model design and should have authority to flag systematic errors for model review.
- Define and publish the meaningful human review standard for WISeR clinical reviewers: override authority, minimum review time, logging requirements, and productivity metric constraints.
- Require discrepancy-triggered enhanced review for any case where the independent assurance monitor identifies a divergence from the frontline system's determination.

Track 1 evaluation metrics, each followed by the governance question it answers:

- **Authorization turnaround time:** Is AI reducing or increasing access delay compared to pre-WISeR baseline?
- **Beneficiary access delay rate:** Are patients waiting longer for medically necessary care?
- **Non-affirmation rate by service line:** Is utilization being reduced in a clinically defensible way?
- **Clinician override rate:** Are human reviewers exercising independent judgment or rubber-stamping?
- **Independent assurance discrepancy rate:** How often does the assurance monitor diverge from the frontline system?
- **Reversal rate after appeal:** Are initial non-affirmations surviving independent review?
- **Provider burden per request:** Is administrative burden being reduced or shifted onto providers?
- **Disparate impact by geography and population:** Is the model producing inequitable results across provider types or patient populations?

Track 2: Future Prior Authorization Model Governance Protocol

For any future AI-assisted prior authorization model in Original Medicare, the governance protocol should require the following stages before any live determinations begin:

- Complete a stakeholder map as the first deliverable, before any internal model scoping begins.

- Conduct the co-creation process (Scout through Plan) to produce a governance specification document that defines the objective function, evidence standards, edge-case rules, and evaluation metrics.
- Publish the governance specification before the vendor RFP is issued. The specification becomes a binding constraint on vendor selection.
- Require shadow-mode testing — in which AI outputs are observed but not acted on — before any live determinations.
- Require a limited beta deployment with a small number of volunteer providers in one or two service lines before multi-state expansion.
- Require the independent AI assurance monitor to be operational before the model goes live.
- Require clinical edge-case validation with specialty panels before any service line is added.

Track 3: Audit Appeals Rationale Governance Pilot

This track applies to UPIC/MAC/QIC/ALJ-style audit and appeals review. The pilot should not deploy AI as a decision-making tool at this stage. It should first implement the governance infrastructure that would be required before any AI is deployed — and use that infrastructure to generate the data that would be needed to evaluate whether AI-assisted review is warranted and under what conditions.

Recommended service categories for Track 3: DME and SNF services, where Q2A data shows the highest variance between UPIC determination rates and ALJ outcomes; wound care and skin substitute products, where high per-claim cost, complex clinical judgment, and well-documented determination variability make rationale instability especially consequential.

- **Rationale concordance rate:** Is the operative denial theory consistent across UPIC, MAC, QIC, and ALJ stages absent new evidence or logged authority?
- **Rationale modification rate:** When rationale changes, is the modification logged with authority and reason?
- **ALJ affirmance/reversal rate:** Are determinations in the pilot arm more accurate than the control arm?
- **QIC-to-ALJ escalation rate:** Are disputes being resolved earlier in the chain?
- **Reversal reason decomposition:** Can CMS distinguish new evidence from changed rationale from policy interpretation from clinical judgment?
- **Provider documentation burden:** Is the rationale ledger reducing the number of additional submissions required?
- **Time to final resolution:** Is the overall duration from initial determination to final outcome shorter?

WHAT WOULD FALSIFY THIS PROPOSAL?

A pilot should be considered unsuccessful if it reduces spending while increasing beneficiary access delay; if it produces lower ALJ concordance than the control arm; if it generates more provider resubmissions rather than fewer; if the independent assurance monitor produces discrepancy rates above a pre-specified threshold without triggering meaningful clinical review; or if savings appear

primarily through unappealed denials rather than demonstrably more accurate determinations. This is a testable governance proposal.

VIII. Implementation Pathway

The governance framework does not require a single nationwide leap. But it does require a different design sequence from the one WISeR used. The comparison below makes the required change explicit:

CURRENT SEQUENCE (WISeR Pattern)	REQUIRED CO-CREATION SEQUENCE
CMS defines objective function internally	Develop stakeholder map as first deliverable; identify all affected groups before any internal scoping
RFP published to select technology vendor	Co-create the objective function through facilitated stakeholder sessions; publish governance specification before RFP
Comment period opens after model structure is fixed	Run Scout through Plan phases; produce governance specification that binds procurement
Critical feedback received but architecture unchanged	Prototype governance tools with clinical panels; iterate before vendor selection
Model launches across six states simultaneously	Shadow-mode testing; limited beta deployment; independent assurance monitor operational before live determinations
Providers experience delays and denials; concerns surface in Senate inquiry	Independent assurance monitor publishes near-real-time metrics; discrepancy-triggered enhanced review catches edge cases
CMS delays services; adjusts model mid-pilot	Measure phase: rationale concordance, ALJ outcomes, access, burden fed back to council for next improvement cycle

Near-Term Actions (No Rulemaking Required)

- Implement Track 1 WISeR oversight retrofit immediately: independent AI assurance monitor, near-real-time metric publication, clinical edge-case panels, meaningful human review standards.
- Require rationale ledger fields in WISeR non-affirmations and in selected QIC reconsiderations through contract modifications and MPIM guidance.
- Publish a Medicare AI Review Governance Report quarterly: authorization turnaround times, non-affirmation rates, independent assurance discrepancy rates, provider burden, and access delay metrics.
- Separate averted-spending metrics from accuracy, access, and fairness metrics in WISeR vendor performance evaluation.

Medium-Term Actions

- Pilot rationale ledgers in one or two MAC jurisdictions for Track 3 service categories; link rationale ledger data to ALJ outcomes for retrospective validation.
- Develop specialty-specific edge-case libraries through facilitated co-creation work sessions for each WISeR service line.
- Publish the governance specification template and stakeholder mapping protocol as CMS guidance documents available to any agency considering AI-assisted review deployment.
- Require QIC issue disposition tables in selected appeal categories.

Longer-Term Actions

- Incorporate the independent AI assurance monitor requirement and the balanced scorecard into any future prior authorization CMMI model through model authority.
- Use rulemaking to constrain QIC new-issue authority under 42 CFR §405.968, requiring explicit logging and provider notice when a new denial theory is raised at reconsideration.
- Establish statutory authority for ‘nothing designed for us without us’ as a governance principle: any federal AI system materially affecting payment, coverage, or access to care must complete a stakeholder map and co-creation phase before procurement. This extends the AI Bill of Rights from post-deployment protections to pre-procurement design obligations.

IX. Conclusion: Design Before Deployment

The paperclip maximizer thought experiment is extreme. But its lesson is ordinary: a system optimized for a mis-specified objective will pursue that objective effectively while producing outcomes its designers never intended. In Medicare review, the mis-specified objective is ‘averted expenditures.’ The unintended outcomes are authorization delays, rationale instability, provider burden, and beneficiary harm. The governance failure that allows this is the same in WISeR, in Cigna’s PxDx, and in UnitedHealth’s nH Predict: the people who bear the consequences of the system’s outputs had no meaningful role in defining what that system was built to optimize.

The purpose of AI in Medicare review should not be to say no faster or more efficiently. It should be to make more accurate determinations, explain them clearly, preserve meaningful appeal rights, focus public dollars on care that is clinically justified and legally payable, and reduce the burden that misaligned governance places on providers and beneficiaries. Those objectives can only be co-created. They cannot be handed down.

Three specific asks:

- **CMS and CMMI:** Implement the Track 1 WISeR oversight retrofit immediately — independent AI assurance monitor, near-real-time metric publication, meaningful human review standards — before any expansion of WISeR’s service lines or states. Do not

launch any future AI-assisted Medicare review model without first completing a stakeholder map, a co-creation phase, and a published governance specification.

- **Senate Finance Committee:** Direct GAO to study three things within 18 months: rationale stability rates across UPIC/QIC/ALJ stages; authorization delay patterns under WISeR versus pre-pilot baselines; and the adequacy of current WISeR governance documentation relative to OMB M-25-21 high-impact AI requirements. These are measurable questions with public-interest answers.
- **Provider associations and patient advocates:** Insist on a seat at the design table as a precondition of participation in any AI-assisted review initiative. A public comment period is not co-creation. Demand facilitated sessions, proactive outreach to leading clinical voices, and a published governance specification before the RFP is issued. Nothing designed for us without us should be written into the participation terms.

The governance sequence that produced WISeR's early problems is not an anomaly. It is what happens by default when AI procurement proceeds faster than governance design. The alternative sequence — stakeholder map, co-created objective function, governance specification, shadow testing, limited deployment, independent assurance, continuous measurement — is available. It is not complicated. It is just not the default. Making it the default is the work.

Appendix A: Draft Governance Specification Template

A governance specification should be completed before procurement. It is the design contract the technology must satisfy. At minimum, it must answer:

- What is the system being asked to decide or support? What specific determination is at issue?
- What is the objective function? What outcomes is the system being rewarded to produce?
- What outcomes are explicitly not acceptable, even if savings increase?
- Which proxy metrics are unsafe if used as the primary optimization target?
- What clinical evidence standards apply, by service category?
- What policy authorities govern determinations (NCD, LCD, statute, regulation, manual)?
- What edge cases are known before deployment, and how will they be handled?
- Who has authority to override the AI output, and under what conditions?
- What must be included in every denial or non-affirmation explanation?
- What data will be logged for audit and appeal review?
- What metrics will be published, and how frequently?
- Who operates the independent AI assurance monitor, and how will its findings be reported?
- What would trigger a pause, retraining, or termination of the model?
- What would count as pilot failure?

Appendix B: Rationale Ledger Fields

The following fields should appear in every AI-assisted adverse Medicare determination, accessible to the provider at the initial determination stage:

- **Case/request ID:** unique persistent identifier for tracking across all review stages.
- **Review stage:** prior authorization, pre-payment, post-payment, redetermination, reconsideration, ALJ, or council.
- **Governing authority:** NCD, LCD, billing article, statute, regulation, manual citation, or other authority.
- **Operative rationale:** the theory supporting denial or non-affirmation, stated in plain language a non-attorney provider can understand and respond to.
- **Evidence reviewed:** documents, notes, measurements, imaging, literature, or other evidence considered.
- **Evidence missing:** specific documentation absent and what its inclusion would change.
- **Overturn criteria:** specific factual or documentary showing that would reverse the determination.
- **Model output:** recommendation, confidence level, and rationale if AI-assisted.
- **Independent assurance flag:** whether the independent assurance monitor identified a discrepancy requiring enhanced review.

- **Human reviewer action:** accepted, modified, rejected, or escalated — with logged reasoning.
- **Modification reason:** new evidence, policy change, reviewer correction, clinical override, or other.
- **Appeal outcome:** affirmed, reversed, partially reversed, dismissed, or pending.
- **Reversal reason:** new evidence, rationale change, policy dispute, or clinical judgment.

Appendix C: Key Medicare Appeals Data (Q3 2025 / FY2025)

METRIC	VALUE AND NOTE
QIC Part A reconsiderations (Q3 2025)	15,706 processed; 19% favorable, 2% partially favorable, 79% unfavorable
QIC Part B reconsiderations (Q3 2025)	22,291 processed; 24% favorable, 1% partially favorable, 75% unfavorable
QIC DME reconsiderations (Q3 2025)	12,243 processed; 59% favorable, 2% partially favorable, 39% unfavorable
UPIC-originated reconsiderations (Q3 2025)	14% favorable, 2% partially favorable, 84% unfavorable
QIC timeliness (Q3 2025)	99.93% Part A; 99.52% Part B; 99.76% DME — timeliness is not the primary performance concern
Reversal reason coding (Q3 2025)	85.0% Part A; 92.6% Part B; 98.9% DME reversals coded as ‘new documentation/evidence persuasive’ — a code that cannot distinguish late evidence from changed reasoning
OMHA claim-level outcomes (FY2025)	16.1% favorable; 14.0% partially favorable; 30.1% total appellant relief at Level 3 — the only fully independent adjudication stage
OMHA processing time (FY2025)	74 days average; within the statutory 90-day frame

Appendix D: References

[1] Centers for Medicare & Medicaid Services. “WISeR (Wasteful and Inappropriate Service Reduction) Model.” CMS Innovation Center, 2026. https://www.cms.gov/priorities/innovation/innovation-models/wiser

[2] Centers for Medicare & Medicaid Services. “WISeR Model Frequently Asked Questions.” CMS Innovation Center, 2026. https://www.cms.gov/priorities/innovation/files/document/wiser-model-frequently-asked-questions

[3] Kaiser Family Foundation. "Examining the Potential Impact of Medicare's New WISeR Model." February 10, 2026. https://www.kff.org/medicare/examining-the-potential-impact-of-medicares-new-wiser-model/

[4] American Hospital Association. "AHA shares concerns, recommendations with CMS on WISeR model." October 23, 2025. https://www.aha.org/news/headline/2025-10-23-aha-shares-concerns-recommendations-cms-wiser-model

[5] U.S. Senator Maria Cantwell. "Cantwell Presses RFK Jr. on New AI Filter Blocking Care for Medicare Patients." April 2026. https://www.cantwell.senate.gov

[6] ProPublica. "How Cigna Saves Millions by Having Its Doctors Reject Claims Without Reading Them." March 2023. https://www.propublica.org/article/cigna-pxdx-medical-health-insurance-rejection-claims

[7] Axios. "AI lawsuits spread to health." July 25, 2023.

[8] Estate of Gene B. Lokken et al. v. UnitedHealth Group Inc. et al. Federal class action, District of Minnesota, filed 2023. See complaint and subsequent reporting on algorithmic use in post-acute care review.

[9] Office of Management and Budget. Memorandum M-25-21. April 2025.

[10] U.S. Department of Health and Human Services. "HHS Artificial Intelligence Strategy." December 2025. https://www.hhs.gov/sites/default/files/hhs-artificial-intelligence-strategy.pdf

[11] McNeill, Lance. "Rationale Drift in Medicare Audit Appeals." Arclight Action Medicare Audit Reform Series, Issue 1, 2026.

[12] McNeill, Lance. "Rationale Drift in Medicare Appeals: A Structural Source of Administrative Burden and Adjudicatory Inconsistency." Draft manuscript, 2026.

[13] McNeill, Lance. "Upstream Denials, Downstream Costs: Hidden Systemic Costs and Measurement Failure in Medicare UPIC Determinations." Arclight Action, 2026.

[14] Humantific. "Strategic Co-Creation Process Quick Guide." Humantific Complexity Navigation, 2015.

[15] McNeill, Lance. "Open Innovation: Collaboration Design." City of Austin Office of Innovation presentation.

[16] McNeill, Lance. Challenge US: Inspiring the Next Generation of Solution Builders with Lessons from the Past and a Hope for the Future. 2022.

[17] White House Office of Science and Technology Policy. "Blueprint for an AI Bill of Rights." October 2022. https://www.whitehouse.gov/ostp/ai-bill-of-rights/

[18] Bostrom, Nick. Superintelligence: Paths, Dangers, Strategies. Oxford University Press, 2014. The paperclip maximizer thought experiment illustrates how an AI system optimized for a narrow proxy objective can pursue that objective in ways that violate every value its designers assumed were obvious but never explicitly specified.

[19] Citron, Danielle Keats. "Technological Due Process." Washington University Law Review 85, no. 6 (2008): 1249–1313. Citron argues that automated decision systems can collapse adjudication and rulemaking while preserving neither meaningful individualized notice nor participatory rulemaking, and that automation can undermine hearings when decision-makers presume computational outputs are correct.

[20] The White House. "Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People." OSTP, October 2022. Establishes five protections: safe and effective systems; algorithmic discrimination protections; data privacy; notice and explanation; human alternatives and fallback.

Appendix E: About the Author

Lance McNeill is the founder of Arclight Action, a Medicare audit defense consulting and civic accountability platform. He co-founded Victory Wound Care, a mobile wound care practice, bringing direct operational experience with Medicare program integrity, UPIC/MAC/QIC/OMHA appeals, WISeR prior authorization review, and skin substitute coverage determinations — including firsthand experience with the multi-wound edge case described in this paper. He previously served in the City of Austin's Office of Innovation, where he led structured multi-stakeholder co-creation campaigns including the Idea Accelerator (which engaged 9.2 percent of the entire city workforce through facilitated diverge-converge sessions) and the Austin Resource Recovery Insights initiative (which synthesized 1,359 community responses into 11 actionable policy insights through structured facilitation). He holds a Master of Public Affairs from the LBJ School of Public Affairs at the University of Texas at Austin and an MBA from Texas State University. He is the author of The Resilient Entrepreneur (2019) and Challenge US (2022).

The co-creation framework described in this paper may be implemented by CMS, CMMI, contractor oversight bodies, provider associations, or neutral third-party facilitators with expertise in public-sector innovation and healthcare review systems.