Net Systemic Value: Measuring the Downstream Value AI Agents Create and Destroy

Ildar Fazulyanov · Independent Researcher · dar@iqvaluations.com

Working paper · v0.9 · July 2026.

Abstract

An AI agent inside a firm optimizes a local objective it can measure and report, and improves it. Whether the firm is better off is a different question, and a structural one: the cost of a local improvement often lands downstream, on the units the agent serves, where the agent's own measurement cannot see it. We define **Net Systemic Value (NSV)** of a part as the value it creates for the parts it serves minus the value it destroys within them, over a period, net of its own cost, scored from the realized outcomes of those served parts rather than from the part's self-report. NSV is the successor to Economic Value Added for a networked, AI-augmented firm: where EVA charges profit for the capital a unit consumes, NSV charges a part for the value it destroys in the units it serves. NSV consolidates without double counting only under a no-interaction/modularity condition on shared external receivers; without that condition, group NSV must be measured directly at the group boundary. We place a **value-observability frontier** on how much systemic value is visible at what measurement cost. The downstream effect is a treatment effect whose counterfactual is not identifiable from observational logs alone, so the defensible deliverable is a bounded estimate with a sensitivity band and periodic holdout calibration, not a single audited figure. We frame the **AI Suboptimal Optimum**, local metric up while NSV is negative, as a measurable dollar object that producer-side attribution (process mining, agent-ROI) and marginal-contribution estimators structurally miss, and distinguish the contribution from each.

Keywords: AI value measurement, economic value added, causal inference, treatment effect, management accounting, responsibility accounting, performance measurement, internal chargebacks, controllability, transfer pricing, AI governance, Goodhart's law, Net Systemic Value.

1. The problem

A firm is a system of parts that serve one another, and value is created and destroyed at the interfaces between them. A part that hands off cheaper, faster, or cleaner output raises the value its downstream parts create; a part that reduces its own cost by degrading what it hands off destroys value elsewhere that its own scorecard never shows. The gap between a part's local value and its effect on the whole is the object of interest, and for AI it is now a documented, board-level problem: a large share of enterprise AI pilots show no measurable profit impact (MIT NANDA 2025), nearly 40% of AI time savings are lost to downstream rework, a finding Workday reports as the “AI Productivity Paradox” (Workday 2026), and a recurring cause of cancelled AI programs is “unclear business value” (Gartner 2025). The accountable executives report they cannot measure it.

They cannot measure it because they measure the wrong thing. Any signal a part reports about itself is a function of the part’s own activity, and a measurement downstream of the thing it measures can be shaped by that thing; the producer cannot certify from its own records that it did not push a cost onto another unit. The information that reveals systemic value lives in the parts that were served. Producer-side instruments, process-mining attribution of KPI gains to agents and agent-ROI comparisons of business value against token cost (Celonis 2026), credit the local win and never debit the downstream loss, because the loss is not in the producer’s record. To measure the net, you measure from the served.

2. Net Systemic Value

For a part A over a period,

$$\mathrm{NSV}(A) = \sum_{B:\, A \to B} \Delta V(A \to B) - C(A),$$

where $\Delta V(A \to B)$ is the realized outcome of B with A’s output bundle minus the counterfactual outcome of B under a specified replacement/absence regime, valued in money, and $C(A)$ is A’s own fully-loaded cost, including for an AI its compute, capex and human oversight. The unit of account is the supplier-receiver pair over the chosen period: if A sends multiple deliverable streams to B, those streams are aggregated into the A $\rightarrow$ B output bundle before the counterfactual is evaluated. Four conventions are fixed at the opening of each interface account, before any estimation. (i) Counterfactual regime. The default regime is status-quo replacement (R1): in the counterfactual, A’s output bundle to B is replaced by the incumbent process the firm ran, or would run, for that work absent A, named explicitly on the account card; in the worked example below, the human-served process. Pure absence (R0), in which the bundle is withheld and nothing substitutes, is admissible only where no incumbent exists and must be declared. R0 and R1 answer different questions and their ΔVs are not comparable across regimes. (ii) Outcome functional. O_B is a pre-registered list of B’s realized, receiver-side outcome variables (for example retention, rework hours, cycle time), read over the account period plus a declared maturation lag, with true-ups posted to the account as outcome data mature. (iii) Monetization. m(·) is a pre-registered map from outcome units to money using the firm’s realized economics (remaining lifetime value for churn, fully loaded labor rates for hours), marked to realization at true-up. (iv) Aggregation. There is exactly one interface account per ordered supplier-receiver pair per period; all deliverable streams from A to B are aggregated into the single A $\rightarrow$ B bundle before the counterfactual is evaluated, and the counterfactual operates on the whole bundle. Stream-level ΔV within a pair is not defined in this version. The test for all four conventions is operational: two analysts holding the same account card and the same data must compute the same ΔV; anything that fails this test is fixed here by convention, not left to the estimator. ΔV is positive where A lifts B and negative where A degrades it. NSV is a single money figure, the systemic profit-and-loss of a part.

NSV extends Economic Value Added (Stewart 1991): where EVA charges a unit's profit for the capital it consumes, NSV charges a part for the value it destroys in the units it serves. Four principles fix the measure. It is **receiver-scored**: A is credited by B's realized outcomes, read from data, not by A's report or opinion, which is the controllability principle of responsibility accounting (Solomons 1965) applied at the interface. It is measured against a **counterfactual baseline**, with attribution (which unit moved) kept separate from baseline (what it would have done anyway). It is **money-denominated**, against the firm's realized economics. And it is **net of the part's own consumption**.

Worked example. A customer-service AI frees twenty agents of labor (\$400k per quarter) at a compute-capex-and-oversight cost of \$100k, and its dashboard is excellent. Measured from retention, the served outcome, its handled cohort churns at 5.0% against a 3.0% human-served holdout; the excess of 2.0 percentage points on 50,000 higher-value customers at \$2,000 remaining lifetime value is \$2.0M of value destroyed downstream. Its receiver-scored interface account, Support-AI -> Retention, records: +\$400k operating-capacity credit for service labor released to the receiver-side service process, −\$2.0M retention debit, −\$100k own cost, **NSV = −\$1.7M**. A star on the dashboard, a \$1.7M-per-quarter loss on the systemic ledger, worse the more it scales. This is the number no producer-side tool computes, because the \$2.0M lives in retention's record, not the agent's.

3. The account: unit of analysis and consolidation

Accounting must balance only after the counterfactual unit is fixed. A **part** is any unit with a measurable interface to the parts it serves (agent, desk, function, person). An **interface account** is the supplier-receiver pair A $\rightarrow$ B over the chosen period, with all deliverable streams bundled per the Section 2 aggregation convention; a part's NSV sums its ΔV across the receiver pairs where it is supplier, minus its cost. Operationally, the interface account is a management-accounting subledger: it is opened for a period, posts a receiver-scored point estimate plus sensitivity band, carries a true-up line when holdout or outcome data mature, and closes into the supplier's NSV view and the receiver's disclosed exposure view. On **consolidation**, internal interface accounts are eliminated from the external group view rather than netted as offsetting cash entries; the elimination removes the internal edge and then recomputes or imports the group-boundary counterfactual. The sum of member NSVs equals the group's NSV to the rest of the firm only when shared external receivers are modular: removing one member's output does not change the counterfactual marginal value of another member's output except through the internal interface being eliminated. Under complements, substitutes, or shared downstream interactions, member removal counterfactuals are not additive; the group NSV must be measured directly by removing or replacing the group boundary as a whole. Cyclic dependencies are handled by Convention EC (Section 3.1) under its equilibration assumption, and by the condensation fallback otherwise.

3.1 Convention EC (surgical edge cut)

Model the firm as a service graph: parts as nodes, one directed edge ($i \to j$) per supplier-receiver pair, and, at period grain, each part's output bundles determined by its input bundles and its own conditions,

$$Y_{i \to j} = g_{i \to j}\left((Y_{k \to i})_{k:(k,i) \in E}, \epsilon_i\right),$$

with B's outcome O_B likewise a function of B's inputs and conditions. The counterfactual defining $\Delta V(A \to B)$ is an intervention on the edge, not on the node: pin the value carried by the single edge ($A \to B$) to the regime value r (the R1 replacement bundle, or the null bundle under a declared R0); hold every structural function fixed; leave every other edge free; and take as the counterfactual the solution of the pinned system, in which all other edge values are recomputed. A remains present throughout: it keeps serving its other receivers, and it keeps receiving its inputs, including any return flow from B, at their recomputed counterfactual values. Then $\Delta V(A \to B) = m(O_B^{\text{factual}}) - m(O_B^{\text{pinned}})$.

Three consequences. First, on an acyclic service graph the pinned system solves in a single forward pass from the cut edge and the solution is unique; there, and only where the direct edge is the sole directed path from A to B, edge-cut and agent-removal semantics coincide. Second, on a cyclic graph the two semantics diverge by exactly the reverse-interface term: removing the agent also severs B's return flow into A and A's edges to its other receivers, so an agent-removal holdout measures a node effect, not the interface effect the ledger posts. Convention EC declares the edge semantics to be the ledger's estimand, and Section 5's calibration is redesigned to estimate it. Third, indirect A-to-B effects that travel through an intermediary (A $\to$ C $\to$ B) are deliberately not charged to the A $\to$ B account; they post on the interfaces where they physically occur, which is what keeps the ledger from double counting one effect across accounts.

Equilibration assumption. On a cyclic graph the pinned system is a fixed-point problem, and Convention EC is well defined only if the period-level pinned system has a unique solution. A sufficient condition is that the composed response around any cycle through the cut edge is a contraction at period grain (loop gain below one). This is an economic assumption about the firm's process dynamics, not a theorem; it fails, for example, where a two-part loop amplifies within the period rather than damping.

Fallback (acyclic restriction by condensation). Where the equilibration assumption cannot be certified, the convention retreats to the graph's acyclic skeleton: contract each strongly connected component of the service graph into a single part and measure NSV at the contracted part's boundary. The condensation of a directed graph is acyclic, so every interface account in the contracted graph falls under the one-pass case above. Directional attribution within a cyclic component, how much of the pair's joint effect belongs to $A \to B$ rather than $B \to A$, is then explicitly unidentified and is carried as an open item on the account, in the same way Section 3 treats non-modular groups: measured at the boundary, not decomposed. This version claims cycle-safe accounting only under Convention EC with equilibration, or under the condensation fallback; it makes no directional within-cycle claims otherwise.

4. Value observability

Not all systemic value is observable. **Value observability** is the share of a firm's systemic value creation and destruction an instrument captures as a function of measurement spend. Under the passive-frontier hypotheses formalized in the Welfare Observability companion (continuously divisible measurement, concave per-interface capture, and positive unobservable value mass), it is a concave frontier: early measurement, aimed at the highest-value, most-detectable interfaces, buys the most sight; later spend buys less; and the frontier asymptotes below one, the gap being value that lives where no instrument can validly report it. With indivisible instrumentation or fixed integration costs, the frontier can be non-concave and must be treated as a knapsack. The discipline is to spend measurement where value at risk per dollar is highest and to acknowledge the "invisible" interfaces, turning "we do not know what our AI is doing to the business" into a budgeted line.

5. The causal core

Measuring ΔV is hard. Outcome-with-agent minus outcome-without-agent, of which only one branch is ever observed. The counterfactual is not identifiable from observational logs alone, and three difficulties compound it: linking an agent action to a downstream outcome across systems and a long, variable time lag; reconstructing the baseline; and selection, since agents are not randomly assigned to work. The mature method stack applies, causal forests (Wager and Athey 2018), difference-in-differences, and synthetic control (Abadie, Diamond and Hainmueller 2010) and its doubly-robust variants, all on panel data firms already hold, but each rests on assumptions (parallel trends, a valid donor pool, no unobserved confounding) that can fail silently.

The consequence is that NSV cannot be delivered as a single audited figure from logs. The defensible deliverable is a **point estimate with a sensitivity band**, reporting how much unobserved confounding it would take to flip a part from net-positive to net-negative, and it is anchored by **periodic calibration** designed to the edge-cut estimand of Convention EC. The naive design, switching the agent off for a holdout, implements node removal and, on any graph with a return flow or a second path, estimates a different quantity from the one the ledger posts. The matching design runs as follows. (1) At account opening, fix the account card per Section 2: regime, outcome list, monetization map, period and lag. (2) Partition the receiver's work into clusters at the grain where the loop closes (customer, case, region), chosen so that within-period feedback from any single cluster back into A is negligible. (3) Assign a random fraction of clusters to the cut arm: for those clusters only, A's deliverable bundle is withheld and the regime replacement is served instead (under R1, the incumbent process handles them). A is not deactivated: it continues to serve all control clusters and all its other receivers, and continues to receive its input flows. The manipulated object is the edge, matching the estimand. (4) Estimate the per-cluster outcome contrast, monetize per the account card, scale to the interface, and post the point estimate with its sensitivity band. (5) Run the leakage diagnostic: compare the flows into A (volume and composition of B's return stream and of upstream inputs) between calibration and non-calibration periods. A detectable perturbation means the cut arm is feeding back into A's production for control clusters, the no-leakage condition has failed, and the estimate is contaminated by the reverse-interface term; in that case the account widens its band or the pair retreats to the condensation fallback and boundary measurement. Equating this cluster-randomized contrast with ΔV under Convention EC requires, beyond the diagnostic: no interference across clusters (cluster-level SUTVA), transportability of the contrast from the calibration fraction to the served population, and stability of the parts' structural responses under the substitution. These are assumptions of the design, not results of it; the sensitivity band bounds unobserved confounding only, and none of the three is testable from the same data that produced the estimate. This is the lesson of multi-touch marketing attribution, which discredited itself by asserting causation from correlation and was replaced by incrementality testing plus modeling. A log-only point estimate is attribution theater; a bounded estimate with a stated robustness test, no-interference/transportability conditions, and holdout calibration is what a risk officer can defend. The framework treats this as a feature of the instrument, not an apology for it.

Scope note. Convention EC repairs one thing: the divergence between the ledger's edge-level definition of ΔV and the estimator's design, which coincide on acyclic graphs and differ on cycles by the reverse-interface term. It does not repair, and does not claim to repair, anything else. Consolidation still requires the Section 3 modularity/no-interaction condition; edge-cut ΔVs remain non-additive under complements or shared-receiver interactions exactly as before. The equilibration assumption on cycles, and the no-leakage, cluster-SUTVA, transportability, and stability conditions on calibration, are identification assumptions. They are standard in causal-design language but are not certified by the same data that estimate NSV. Accounts computed under Convention EC should therefore be represented as bounded estimates, not audited figures, until the identification design has been reviewed for the specific deployment.

6. The AI Suboptimal Optimum, and positioning

The **AI Suboptimal Optimum** is the measurable state where an agent's local metric rises while its NSV is negative. It is the NSV-accounting instance of the Suboptimal Optimum cell in the System Adjusted Payoff Model: the local private channel rises while the relevant system channel falls. It is not exotic misalignment; the agent optimizes what it was measured on, the Goodhart dynamic (Goodhart 1975), the folly of rewarding A while hoping for B (Kerr 1975), and the cost lands where its measure cannot see, and under optimization pressure it drifts there because the local metric is what it can shape. NSV detects it by charging the downstream loss to the agent's account.

The components are cited, not claimed; the contribution is the receiver-scored net and its properties. Producer-side AI attribution (process mining with agent-ROI) measures gross value created and is, by construction, positive-outcome and producer-side. Marginal-contribution estimators (Zhu 2026) price an agent by its Shapley surplus (Shapley 1953) within a workflow, net of a production-time failure scalar. They are near neighbors for workflow surplus, accountability, and production-time failures, but they do not price receiver effects outside the workflow boundary, so they can overprice agents whose local workflow surplus carries a downstream tax; the exact "agent that games" statement is directional, not a biconditional. AI observability scores technical quality, not money. The receiver-scored mechanism descends from the controllability principle and transfer pricing (Eccles 1985), whose gaming pathologies we control by reading realized outcomes from data rather than taking a receiver's opinion; EVA is the acknowledged parent of the charge-for-what-you-consume idea, extended here to charge for what you destroy. Claimed as new: NSV as a receiver-scored, conditionally consolidating, money-denominated measure of a part's net systemic effect; the value-observability frontier over interfaces; the AI Suboptimal Optimum as a measurable dollar object; and the causal treatment that makes the estimate defensible.

7. The discipline

Net Systemic Value is the operational core of Systemic Value Accounting: the measurement and management of value creation and destruction at the level of the whole system. Its claim is narrow. Local and producer-side measurement cannot see whether a part creates or destroys value for the whole where the relevant information lives in the parts served; NSV measures that receiver side, in money, with consolidation valid under the stated modularity/no-interaction conditions. Value observability keeps track of what remains unseen, and the causal treatment keeps the number defensible rather than theatrical under its stated identification assumptions. The question to ask of an AI inside your business is not whether it is aligned in the abstract, but whether the units it serves are measurably better off, net, and Net Systemic Value is how you answer this question.

References

Abadie, A., Diamond, A., and Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, 105(490), 493–505.

Celonis (2026). *Scaling the Agentic Enterprise with Microsoft Agent 365 and Celonis*. Celonis Blog.

Eccles, R. G. (1985). *The Transfer Pricing Problem: A Theory for Practice*. Lexington, MA: Lexington Books.

Fazulyanov, I. (2026). *Systemic Value Accounting: Measuring and Managing Value Creation and Destruction in the Age of AI*. Framework document, v4, July 2026.

Fazulyanov, I. (2026). *Welfare Observability: A Cost-Coverage Frontier for AI Oversight Under Evasion*. Working paper, v1.4, July 2026.

Gartner (2025). *Gartner Predicts Over 40% of Agentic AI Projects Will Be Canceled by End of 2027*. Press release, June 25, 2025.

Goodhart, C. A. E. (1975). Problems of monetary management: The U.K. experience. In *Papers in Monetary Economics*, Vol. 1. Sydney: Reserve Bank of Australia.

Kerr, S. (1975). On the folly of rewarding A, while hoping for B. *Academy of Management Journal*, 18(4), 769–783.

MIT Project NANDA (2025). *The GenAI Divide: State of AI in Business 2025*. Massachusetts Institute of Technology.

Shapley, L. S. (1953). A value for n-person games. In H. W. Kuhn and A. W. Tucker (Eds.), *Contributions to the Theory of Games, Volume II* (pp. 307–317). Princeton: Princeton University Press.

Solomons, D. (1965). *Divisional Performance: Measurement and Control*. New York: Financial Executives Research Foundation.

Stewart, G. B., III (1991). *The Quest for Value: A Guide for Senior Managers*. New York: HarperBusiness.

Wager, S., and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523), 1228–1242.

Workday (2026). *Companies Are Leaving AI Gains on the Table.* Workday Newsroom, January 14, 2026.

Zhu, Q. (2026). *Agentomics: Economic Foundations for the Valuation, Attribution, and Pricing of AI Agents in Human-AI Workflows.* arXiv:2606.14769.