The Value of What the Model Knows: A working note on governed knowledge as a variable in the economics of AI tasks

Witold Reichhart and Arnaud Gelas

Working Paper

Working note v0.3 · October 2026

Abstract

Cordier's per-task value equation has no knowledge variable. It derives the value of a token from task success, oversight, and the worst outcomes, and reads task success off a benchmark as a property of the chosen model, while the benchmark record shows that a part of it moves with what the model is given to act from. We restate the equation and write it with a knowledge state K, typed as a pair: the content served to the agent, and the governance metadata on each claim, its review status, permitted use, and stratum. Task success, detection, review cost, and token spend become functions of K, each tied to the coordinate it responds to. The share of acts the check refuses and the price of taking a refused act come from the companion note on the mandate, which prints the same display at a fixed knowledge state, and the upkeep of the knowledge enters as a stock cost, with a condition for when the stock pays for itself. Cordier's equation is the case with no refusal and no upkeep. We state the set of knowledge states at which a model clears break-even, which can be empty, shows that reusing one reviewed claim across many tasks lowers its upkeep and concentrates the tail in the same step, and places the classical value of information on both sides of the workflow: the firm keeps its valuation identity, the reviewer is tested against the optimal use of what he is shown, and the theory restricts no sign for the model's response to more knowledge. A paper follows, with a factorial that separates actual review, the visibility of the review label, and enforcement.

Keywords: value of information, language agents, knowledge layer, governed knowledge, context quality, task economics, review cost, Blackwell ordering, experimental design, financial services

The problem

In Cordier's model the success rate of a task is a property of the chosen model, read off a benchmark. In enterprise work it also moves with what the model is given to act from. Holding the model fixed and supplying a short semantic description of the data raised pass rates on an analytics benchmark by seventeen to twenty-three points across three frontier models (Rumiantsau and Fokeev 2026). The study is vendor-authored, on one dataset of a hundred questions, and its finding that the three models were then indistinguishable is a failed rejection at that size and no equivalence result, so we rely on the ordering alone. A semantic description is content, and the study says nothing about review. But it shows that a part of what a benchmark reports as model performance responds to the content served, and Cordier's equation has no term for it.

Each neighbouring literature has a knowledge variable of its own. Value-of-information theory prices information structures for a decision-maker and for an organisation (Marschak and Radner 1972; Hilton 1981), the economics of prediction machines prices the prediction and the judgement around it (Agrawal, Gans, and Goldfarb 2018), and retrieval-augmented generation has measured since Lewis et al. (2020) how what a model retrieves moves its accuracy. The retrieval line stops at accuracy points: none of it attaches a value to the lift, couples it to the cost of reviewing the agent's output, or constrains it by the worst outcomes. And the knowledge-management literature of the 1990s evaluated its projects by a scatter of measures with no settled model (Loermans and Fink 2005), with no record of which item a given piece of work had used. Garicano (2000) wrote the organisation of knowledge as a trade between learning and communication, with the exceptions referred to someone who knows more, and the referral is the part of his model we use.

Agent logs record, for each task, which claims were served and in which state, which claims the agent cited when it asked to act, and what the task was worth. In this note consumption is the claims served, and declaration is the claims the agent cites to the check. Reliance is counterfactual dependence, whether the outcome would have differed without the claim, and no log shows it. So the logs give an exposure map, consumption and declaration per task, and reliance is identified only by an experiment, which the last section designs.

The equation

Cordier's "The Value of a Token" (Capgemini Invent, HAL hal-05680441, July 2026) prices one task done by an agent. A task is worth B when it comes out right, and it comes out right with probability s. A caught failure costs C_r to put right and an escaped one costs C_e. Review is a policy. The firm reviews a share κ of tasks at a cost c_r per task reviewed, review catches a failure with probability d, and tokens cost τ per task. Then

$E[V] = s \cdot B - (1 - s) \cdot A(\kappa) - \kappa \cdot c_r - \tau$, with $A(\kappa) = C_e - \kappa \cdot d \cdot (C_e - C_r)$,

maximised subject to a bound on the bad days: the average outcome on the worst five tasks in a hundred, in expectation, may not fall below V_min, which is conditional value at risk at the five percent level. Cordier claims his results as signs and orderings only, which is his own discipline. The break-even success rate at a given review share is $s^* = (A(\kappa) + \kappa \cdot c_r + \tau)/(B + A(\kappa))$, a ratio of two affine functions of κ and so monotone on [0, 1]. Review moves it. In his worked example the break-even with everything reviewed lies below the break-even with nothing reviewed, a task whose success rate lies between the two is rescued

by review alone, and the task that no oversight rescues is the one below the break-even at whichever corner minimises it. The value per token has the sign of E[V]. Under his linear review technology untargeted review is all or nothing, and the bound on the bad days changes that: where full review does not pay and no review breaches V_min, the cheapest review share that clears it lies in between. Where loss data are thin he bounds the bad days with two moments: for any distribution with mean μ and standard deviation σ, the average outcome on the worst five in a hundred is at least $\mu - \sigma \cdot \sqrt{19}$. In his sensitivity analysis the success rate explains most of the variance in a token's value, oversight comes second, and the token price last.

The cost of quality as a concept is Juran's (1951), and its partition into prevention, appraisal, and failure costs is Feigenbaum's (1956). Principle 11 of the Agentic Engineering Manifesto (Gelas and Reichhart 2026) wrote a cost of correctness for agentic work as inference spend plus verification overhead plus incident remediation, and Cordier's τ, $\kappa \cdot c_r$, and $(1 - s) \cdot A(\kappa)$ are those three. And in Feigenbaum's partition review is appraisal and remediation is failure. Inference spend is a production cost and belongs to none of them. But prevention, the cost of making failure less likely before the work is done, is absent from the equation. The upkeep of knowledge is a prevention cost, and we add it.

The knowledge state

The knowledge state K for a task is a pair (X, M). X is the content served to the agent: the claims, each a proposition with the observations it rests on. M is the governance metadata on each claim: whether it has been reviewed, its permitted use, which is explore, draft, or act, and its stratum, public, firm, or client. The stratum is an access class. It says who may see a claim and ranks nothing by reliability. Curation changes X: it supplies content the model lacks and removes the stale and the contradicted. Certification changes M: a reviewer marks a claim reviewed and permitted for a use. But actual review in a firm usually does both at once, and the design separates them. K is logged per task with its consumption and declaration, and that log makes the extension estimable. The Knowledge Layer (Reichhart and Gelas 2026) is a reference architecture for such an object; the definition here is all we use.

The success rate s(K) responds to X. It rises with curation, through content the model did not have and through the removal of what is wrong. It falls as true but irrelevant context is added, which is one of the two findings of Liu et al. (2024), the other being the drop in accuracy when the relevant passage lies in the middle of a long context, and a vendor report finds the same decline with length across eighteen models (Chroma 2025). s responds to M only through the label as a cue, where the model is shown it, and the sign of that response is unknown. Detection d(K) is Cordier's d, the probability that a reviewer who reads an output catches a failure. It rises with X: a reviewer who compares the output with a stated claim has more to check against than one who compares it with what he remembers. And it responds to M in the reviewer through an opposing mechanism: a review badge can make his check less (Parasuraman and Manzey 2010). The review cost c_r(K) falls where a cited claim answers the question the reviewer would otherwise re-derive, and rises where the content served exposes contradictions he could not see before. Token spend τ(K) rises with the volume of X.

The share of acts refused, ρ(K), responds to M, deterministically, given the claims an act declares. Before every act a status check reads whether each declared claim is reviewed, permitted for an act, and in the stratum the act requires, and refuses the act if any is not. The lookup is near-free per act. Its mistakes are not free: a correct act on an unreviewed claim is refused, and the cost of that false refusal, with the latency,

is priced through ρ and through the person who takes the referred act. ρ falls with coverage and review in the usual case, and review can raise it, where a reviewer finds that acts the check permitted should be refused. The status check catches no failed output. But an act that declares a reviewed, permitted claim in the right stratum and is wrong passes it. So the check belongs in ρ and in the mix of acts the agent performs. Placed inside d it would be multiplied by κ, and a firm that relied on the check and reviewed nothing would get nothing from it in the algebra. A semantic check, whether the act follows from the claim, is a model call with its own cost, miss rate, and false-refusal rate, and we do not claim it.

Knowledge enters the mandate at a second place. Where an act's class is a judgement, advice, or a counterparty that may be relevant to sanctions, the classification the check reads is itself a claim with a review state, and an act whose class rests on an unreviewed classification is referred.

The extended objective

The companion note on the mandate, A Risk Appetite Where the Act Happens, writes Cordier's equation with refusal, per rung r of the mandate's ladder: ρ_r is the share of acts the check refuses, π_r the share of refused acts a person performs, κ_r the review share, and N the acts in a period. With K in it, the value per arriving act is

$$E[V_r \mid K] = (1 - \rho_r(K))\cdot[s_r(K)\cdot B - (1 - s_r(K))\cdot A(\kappa_r, K) - \kappa_r\cdot c_r(K)] + \rho_r(K)\cdot\pi_r\cdot[s_{h,r}\cdot B - (1 - s_{h,r})\cdot A_{h,r} - c_h(\rho_r(K)\cdot\pi_r\cdot N)] - \rho_r(K)\cdot(1 - \pi_r)\cdot C_{n,r} - \tau(K),$$

with $A(\kappa_r, K) = C_e - \kappa_r\cdot d(K)\cdot(C_e - C_r)$.

Each symbol is a measurement in the experiment of the last section. s_r(K) is the success rate of the acts the check permits at rung r, against ground truth adjudicated by people who did not write the claims. d(K) is the share of seeded failures the reviewer catches. c_r(K) is reviewer time per reviewed act at its loaded rate. ρ_r(K) is the share of arriving acts the check refuses, read from its log, and τ(K) the token spend per act, read from the same log. s_h,r, A_h,r, and c_h are the person's success rate on referred acts, the expected cost of his failures, and his average cost per act at the referred volume, which rises past the capacity of the escalation path; C_n,r is the expected cost of an act nobody performs, zero where stopping is the right outcome. The mandate note defines these terms and prices them. The display charges τ on every arriving act, which assumes the agent generates the act before the check reads it. But a check that runs before generation multiplies τ(K) by (1 − ρ_r(K)).

The value is per arriving act. A narrower M raises ρ_r(K), and where the check refuses the acts most likely to fail it raises s_r(K) on the acts it permits. So a rise in s on permitted acts can be the refusal of the hard cases, and only the value per arriving act shows whether the workflow gained. And the referred acts are a selected set, harder than the average act, and s_h,r moves with ρ_r(K) as well. The upkeep of K is a stock cost, c_K per act, and the value of a knowledge state is $E[V_r \mid K] - c_K$. Cordier's equation is the case ρ_r = 0 and c_K = 0, and the mandate note prints this display at a fixed K.

For a given model, workflow, and rung, the knowledge states at which the workflow pays form the set $\{K : E[V_r \mid K] - c_K \geq 0$, with the bound on the bad days holding at rung r$\}$. The set can be empty. K is a structured set of claims with no natural order, and its terms do not move together: s_r(K) can fall with volume while c_r(K) rises with exposed contradictions and c_K rises with every claim governed, so a better knowledge state need not cross zero anywhere. A model whose attainable success rate on a workflow never exceeds 0.4, with B = 1, C_r = C_e = 1, and τ = 0.1, loses at least 0.3 per task with

no review and no upkeep, whatever it is given. The set is non-empty when the gap between the model and the workflow is one of missing content, so that some attainable K lifts s_r(K) above break-even, and when the best attainable value after c_K is positive with the bound holding. Then whether a cheaper model with a governed claim set can replace a stronger one is an inequality between the two models' net values, decided by data. And the context literature is at least as consistent with the opposite, a stronger model using the same claim set better.

Shared claims and the tail

Many acts consume the same claims, and here K changes Cordier's problem. Conditional value at risk is subadditive on loss, so where each class carries a valid bound on the same loss over the same horizon, the class bounds add up to a ceiling on that part of the firm's book whatever the dependence between classes. The ceiling is reached only when the acts fail together. Shared claims do not make that certain, but they make failures correlated, and they are the dependence that pushes the book towards it. Reuse acts on both sides of the ledger at once. A reviewed claim consumed by N tasks costs each of them a share of its review, and the share falls as N grows. And a reviewed claim that is wrong can fail many of the N at once: in the limit where every task that consumed it relies on it, their outcomes move as one and the variance of the period's total loss on them grows with N squared, against N for independent failures; between the two limits the covariance the shared claim induces is what the logs measure. Economies of scale in upkeep and the concentration of the tail are one trade, made claim by claim. The validity window after which a claim is reviewed again is therefore a tail control before it is an accounting convention, and so is the exposure limit per claim that the mandate note sets on the acts not yet adjudicated. The logs show the exposure, for each claim the acts that consumed it and the acts that declared it. Which of them would have gone differently without it is the reliance the experiment identifies.

The upkeep of knowledge

Cordier's equation has review labour per task and no term for the labour that produces a reviewed claim. A claim is reviewed once inside a validity window Δt_j and consumed by many tasks. Its cost is quasi-fixed and non-rival within the window, which is Arrow's (1962) observation about information, and its cost per task is an allocation. The simplest allocation, on consumption, is

$$c_K = \Sigma_{j \text{ in } K} \; C_j / N_j(\Delta t_j),$$

with C_j the cost of reviewing claim j for its window and N_j the number of tasks forecast to consume it in that window. N_j is a forecast under the deployment. It falls when ρ_r rises and when the firm runs the agent on fewer workflows, so the per-task figure moves with the decisions it is meant to inform. The key can run on declaration in place of consumption, and either way the allocation is activity-based costing (Cooper and Kaplan 1988). But no allocation of a joint cost is uniquely right (Thomas 1969), and nobody runs a factorial to produce an amortisation schedule. So c_K is an average cost, fit for reporting the net value per act, and the decision to maintain the layer runs on the stock. With C_K the total cost of reviewing and re-reviewing the claims in K over a window, the layer pays for itself when

$$\Sigma_i \left(E[V_{r,i} \mid K] - E[V_{r,i} \mid \text{none}]\right) \geq C_K,$$

summed over the tasks i in the window, where none is the same workflow served with no governed claims. A claim that serves one task only, a client fact used once, has N_j = 1 and no amortisation, and its whole

cost falls on that task. A claim that serves tasks in several mandate classes is allocated across them by consumption, and its tail enters every class it serves. Review labour is booked once. The labour that reviews a claim is in C_j and the labour that reviews an output is in κ_r·c_r(K), and where reviewed claims make output review cheaper, the saving shows as a lower c_r(K) and is not booked again as a surplus of the layer. Garicano's referral to someone who knows more is the second bracket of the display, ρ_r(K)·π_r, the acts the check sends to a person.

Where the theory holds, and where it breaks

Value-of-information theory has priced information since 1961. A signal is worth the difference in expected value between deciding with it and without it, net of what it costs (Raiffa and Schlaifer 1961; Howard 1966), and an information structure more informative in Blackwell's sense weakly raises the value an optimising decision rule can extract, for every decision problem (Blackwell 1953). The theory also prices structures an organisation uses again and again (Marschak and Radner 1972; Arrow 1974). So a stock amortised over a stream of tasks is within it, and we add an operationalisation, a per-task record and an estimand, to a valuation the theory already has. Economics knows one break in the theory already, information whose private value exceeds its social value and which a market can be worse off for having (Hirshleifer 1971). The break that concerns us lies between the optimal and the implemented use of information.

The firm is the classical consumer on both sides of the workflow. It chooses whether to serve K, with which metadata, and under which mandate, and it can always decline to serve it. So the value to the firm of the option to serve K is non-negative, and $\Delta E[V]$ is the classical estimand. Inside the workflow two users of K are guaranteed nothing. The reviewer is a bounded person shown reviewed claims. His benchmark is the optimal attainable value under the structure he is shown, and the hypothesis is that he approximates it. The alternatives are automation bias, where the label lowers his checking (Parasuraman and Manzey 2010), and the appropriate-reliance account, in which his reliance follows the label's actual reliability only when the label is informative about it (Lee and See 2004). The one study that manipulated a verification label alone, on account badges for tweets, found no significant effect on credibility judgements (Vaidya et al. 2019). Account verification is a different construct from claim-level review, and the experiment tests the reviewer directly. But the model is the instrument the firm deploys. It has no decision rule the theory describes, so the benchmark says nothing about it: added true context can lower its success rate, and the sign of s(K) is an estimate. On the model's side the contribution is the estimator.

The paper

The Value of What the Model Knows takes this note to its formal statement and its experiment. Its hypotheses are the net orderings of the mechanisms above. At a fixed model, the governed workflow raises value per arriving act over the same workflow with raw context; actual review raises s through content; the reviewer's detection moves with the label in the direction the optimal benchmark predicts, against automation bias; and enforcement lowers escaped errors and raises escalation cost. The design crosses three treatments that a firm's review usually bundles.

Treatment	What it changes	Arms
Actual review	X: the content, curated and corrected by a reviewer	the reviewed version of the claim set against the unreviewed version
Label visibility	M as the model and the reviewer see it, at fixed X	label shown, hidden, or shuffled across claims
Enforcement	whether the status check runs before the act	check on, check off

A context-volume arm adds true but irrelevant claims at fixed relevant content, so that curation is told apart from volume. The design estimates the intention-to-treat value per arriving act of the whole governed workflow against raw context, the label effect at fixed content on the model and on the reviewer, the enforcement effect on coverage, escaped errors, and escalation cost, and the content effect of actual review. Each act's record separates task error, policy violation, refusal, detection by the reviewer, and remediation. Because a wrong reviewed claim that causes an error and then persuades the reviewer is the common failure the design exists to expose, errors are seeded in reviewed claims as well as in outputs. Ground truth is adjudicated by people who did not write the claims. The unit of randomisation is the task, blocked by the claims it consumes. Tasks that share a claim are not independent observations, so standard errors are clustered on the claim. A pilot's discordant rate sizes the golden set for one primary endpoint, success per arriving act under the governed workflow against raw context, paired on the task; interactions, review time, and rare losses are secondary and reported as such. The experiment runs on two instrumented implementations and on a public benchmark arm, and the upkeep is in the equation before any value is claimed.

A search of the retrieval, knowledge-management, and human-factors literatures to October 2026 found no study that varies review status at fixed content and measures task success, detection, review cost, and refusals together; the nearest work conditions retrieval on source credibility, on small corpora and with no costs. The permitted uses, the strata, and V_min come from the Knowledge Layer and from the mandate note, and we take them as given.

Provenance

The knowledge layer as the governed object an agent acts from is the published programme of the authors: Build the Medium (February 2026), Governed Intelligence Architecture (May 2026), and The Knowledge Layer (June 2026), with total cost of correctness in the Agentic Engineering Manifesto (March 2026). The mandate under which the agent acts, the share of acts it refuses, the price of the person who takes a refused act, and the bound on the worst outcomes are the subject of a companion working note by the first author, which prints the extended objective at a fixed knowledge state. Philippe Cordier and the first author work for the same firm; this note extends his published paper and has had no access to unpublished material of his. The authors used large language models as drafting, editing, and adversarial-review tools, and take full responsibility for the content.

References

Agrawal, A., Gans, J., and Goldfarb, A. (2018). Prediction Machines: The Simple Economics of Artificial Intelligence. Harvard Business Review Press.

Arrow, K. J. (1962). Economic Welfare and the Allocation of Resources for Invention. In The Rate and Direction of Inventive Activity, 609-626. Princeton University Press for the NBER.

Arrow, K. J. (1974). The Limits of Organization. Norton.

Blackwell, D. (1953). Equivalent Comparisons of Experiments. Annals of Mathematical Statistics 24(2), 265-272.

Chroma (2025). Context Rot: How Increasing Input Tokens Impacts LLM Performance. Technical report, 14 July 2025.

Cooper, R., and Kaplan, R. S. (1988). Measure Costs Right: Make the Right Decisions. Harvard Business Review 66(5), 96-103.

Cordier, P. (2026). The Value of a Token: Token Cost Doesn't Matter; Model Reliability, Human Oversight, and Tail Risk Are. Capgemini Invent. HAL hal-05680441, v1, 5 July 2026.

Feigenbaum, A. V. (1956). Total Quality Control. Harvard Business Review 34(6), 93-101.

Garicano, L. (2000). Hierarchies and the Organization of Knowledge in Production. Journal of Political Economy 108(5), 874-904.

Gelas, A., and Reichhart, W. (2026). The Agentic Engineering Manifesto. March 2026. github.com/arnaudgelas/agentic-engineering-manifesto.

Gelas, A., and Reichhart, W. (2026). Governed Intelligence Architecture for Institutional AI. SSRN, doi:10.2139/ssrn.6701941.

Hilton, R. W. (1981). The Determinants of Information Value: Synthesizing Some General Results. Management Science 27(1), 57-64.

Hirshleifer, J. (1971). The Private and Social Value of Information and the Reward to Inventive Activity. American Economic Review 61(4), 561-574.

Howard, R. A. (1966). Information Value Theory. IEEE Transactions on Systems Science and Cybernetics SSC-2(1), 22-26.

Juran, J. M. (1951). Quality Control Handbook. McGraw-Hill.

Lee, J. D., and See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. Human Factors 46(1), 50-80.

Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems 33, 9459-9474.

Liu, N. F., et al. (2024). Lost in the Middle: How Language Models Use Long Contexts. Transactions of the Association for Computational Linguistics 12, 157-173.

Loermans, J., and Fink, D. (2005). How Organisations Evaluate Their Knowledge Management Projects: A Meta-Study of the Period 1992-2002. Knowledge Management Research and Practice 3(3), 125-135.

Marschak, J., and Radner, R. (1972). Economic Theory of Teams. Yale University Press.

Parasuraman, R., and Manzey, D. H. (2010). Complacency and Bias in Human Use of Automation: An Attentional Integration. Human Factors 52(3), 381-410.

Raiffa, H., and Schlaifer, R. (1961). Applied Statistical Decision Theory. Harvard Business School.

Reichhart, W., and Gelas, A. (2026). Build the Medium: Why Organizational Intelligence Is Mechanism, Not Metaphor. SSRN, doi:10.2139/ssrn.6230858.

Reichhart, W., and Gelas, A. (2026). The Knowledge Layer: A Reference Architecture for Delegated AI Action in Regulated Institutions. Zenodo, doi:10.5281/zenodo.20784662.

Rumiantsau, M., and Fokeev, A. (2026). Semantic Layers for Reliable LLM-Powered Data Analytics: A Paired Benchmark of Accuracy and Hallucination Across Three Frontier Models. arXiv:2604.25149.

Thomas, A. L. (1969). The Allocation Problem in Financial Accounting Theory. Studies in Accounting Research 3. American Accounting Association.

Vaidya, T., Votipka, D., Mazurek, M. L., and Sherr, M. (2019). Does Being Verified Make You More Credible? Account Verification's Effect on Tweet Credibility. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.