

Counterfactual Decision Quality in Governed Artificial Intelligence: Off-Policy Evaluation and Governance-Constrained Policy Improvement over the Logged Decision Record

Kohei Yata¹ & Josiah Hanna²

^{1,2}University of Wisconsin-Madison, Madison, Wisconsin, USA

ARTICLE INFORMATION

Article history:

Published: October 2026

Keywords:

AI Governance
 Governed Decision-Intelligence
 Off-Policy Evaluation
 Logged Bandit Feedback
 Inverse-Propensity Weighting
 Doubly-Robust Estimation
 Counterfactual Decision Quality
 Reject Inference
 Policy Improvement
 Production AI

ABSTRACT

A governed AI system does not only make decisions; it records, for every decision it makes, the probability with which it would have chosen each available action. This paper argues that the logged-propensity obligation established by Mesbaul Haque Sazu in the Governed Decision-Intelligence (GDI) framework [11], and carried into the runtime by the companion Full-Stack reference architecture [10], turns the ordinary operating output of a governed system into a logged bandit feedback dataset of the kind that off-policy evaluation was designed to use. We take the deployed gate as a logging policy and ask a question the frameworks make answerable but do not themselves answer: what would the expected decision cost have been if the system had used a different threshold, or routed a class of cases differently, without ever deploying that alternative? We give three estimators for the value of a counterfactual policy over the governed record, namely the inverse-propensity estimator, its self-normalized variant, and a doubly-robust estimator that remains consistent if either the logged propensities or a fitted outcome model is correct. For each we state a variance and a finite-sample confidence bound, and we pose policy improvement as a constrained optimization that maximizes estimated value subject to the governance constraints the frameworks already impose, in particular a floor on the recorded-refusal rate and a ceiling on calibration drift. Because outcomes in a gated system are observed only for the arm that acted, we treat the resulting missing-label structure explicitly, connecting it to the alternative-data and reject-inference concerns of Sazu's creditworthiness analytics framework [12]. An illustration on a 50,000-decision stream drawn from publicly available data in open-source platforms shows the relative bias and variance of the three estimators and the shape of the value surface over the threshold. The contribution is an evaluation layer that sits on the governance control structure; every estimator is well defined only because these frameworks require the logged propensity that makes it computable.

1. Introduction

1.1 From logged decisions to counterfactual questions

Every operator of a decision system eventually asks a counterfactual question. If we had set the approval bar slightly higher, how many bad actions would we have avoided, and how many good ones would we have lost? If we had sent this class of borderline cases to a human instead of automating them, would the decisions have been better? These are the questions on which a threshold is actually chosen, and in most systems they cannot be answered, because the only thing the system recorded was what it did, not what it would have done. The alternative was never tried, so its value is a matter of argument rather than measurement.

The frameworks that Sazu published change this situation in a specific and, for our purposes, decisive way. The Governed Decision-Intelligence framework requires that a decision be routed on a calibrated risk signal and that the routing choice be logged [11]; the Full-Stack reference architecture makes the logged propensity, the probability with which each branch would have been taken, a standing obligation recorded for every decision rather than reconstructed after the fact [10]. A system that honours this obligation is, almost as a side effect, keeping exactly the record that the statistics of off-policy evaluation require: a sequence of actions, the probability under which each was selected, and, where available, the outcome that followed. That record is a logged bandit feedback dataset, and it is the raw material of this paper.

The point deserves to be stated plainly, because it is easy to miss under the formal language. Logged propensity is not an analytics convenience that a team might add if it had time. In the governed setting it is already present, because the framework insists on it for reasons of auditability [10]. What we show is that once it is present, a whole class of counterfactual questions that teams usually answer by guesswork becomes answerable by estimation, with error bars attached. The evaluation layer we build does not ask the governed system to record anything new. It reads what the governance controls already oblige it to keep.

1.2 Why off-policy evaluation, and why the governed record makes it honest

The naive way to compare a proposed policy against the deployed one is to deploy it and see what happens. In a consequential setting this is exactly what a governed system is built to avoid, because the proposed policy might be worse and the cost of finding

out is paid in real irreversible actions. Off-policy evaluation exists precisely so that the value of a policy can be estimated from data collected under a different policy, without running the new one [9], [3]. The method has a long history in statistics as inverse-probability weighting [4] and in machine learning as counterfactual reasoning over logged systems [2].

What has usually made off-policy evaluation fragile in practice is that it depends on knowing the probability with which the logging system took each action, and most deployed systems do not record it honestly. They log the action and leave the propensity to be reconstructed, which introduces an error that the estimator cannot see and the analyst cannot bound. The governed record removes this fragility at its source. Because the propensity is logged at the moment of the routing choice and bound to the decision identifier [10], the importance weights that the estimators of Section 3 depend on are known rather than estimated. The honesty of the evaluation is inherited from the governance obligation, not asserted by the analyst.

There is a second reason the governed record is the right substrate. Supervisory model-risk guidance [1] and the NIST AI Risk Management Framework [7] both expect that a change to a deployed model be justified before it is made, not merely monitored after. An estimate of a proposed policy's value, with a confidence bound, is precisely the kind of *ex ante* justification that these regimes ask for, and the governed record is what makes it computable.

1.3 Contributions and scope

This paper is additive and strictly dependent on the governance foundation. It introduces no new governance vocabulary and reuses the established terms, namely calibrated routing, logged propensity, recorded refusal, and decision-quality assurance, with the meanings assigned to them in the GDI framework [11] and the Full-Stack architecture [10]. Against that foundation it makes four contributions. C1. It frames the governed decision stream as logged bandit feedback and the deployed gate as a known logging policy, making the logged propensity of [10] the importance weight of off-policy evaluation (Section 3.1).

C2. It gives three value estimators for a counterfactual policy over the record, the inverse-propensity estimator, its self-normalized variant, and a doubly-robust estimator, each with a variance and a finite-sample confidence bound (Sections 3.2 and 3.3).

C3. It poses policy improvement as a constrained optimization that maximizes estimated value subject to the recorded-refusal floor and the calibration-drift ceiling the frameworks already require, so that an improved policy cannot buy value by quietly weakening a control (Section 3.4).

C4. It treats the missing-outcome structure of a gated system explicitly, linking the selective-labels problem to the reject-inference concern in Sazu's creditworthiness analytics framework [12], and illustrates all of the above on publicly available data (Sections 3.5 and 4).

The scope is deliberately narrow. This is a methods paper and reports no production deployment. The illustration of Section 4 draws on publicly available data in open-source platforms, with sources and processing described in Section 4.1. Nothing here shows that governed systems choose better policies than ungoverned ones; it shows only that, when a system is governed to the standard these frameworks set [11], [10], the value of an unchosen policy becomes an estimable quantity with an honest error bar.

2. Literature Review

2.1 The governed decision record as logged bandit feedback

The foundation for this paper is the pair of frameworks Sazu introduced in 2023. The GDI framework makes the individual decision the unit of assurance, routes it on a calibrated risk signal, and binds a decision record at generation time carrying the model and policy versions, the prediction, and the calibrated confidence [11]. The Full-Stack reference architecture carries this into a runtime, binding an execution record at commit and, critically for us, logging the propensity of every routing choice as a first-class obligation [10]. Read together, the two records give, for each decision, a tuple of context, action, logged propensity, and, where the outcome loop closes, a realized outcome. This is the canonical shape of logged bandit feedback as used in the off-policy literature [3], [2]. We claim no novelty for the shape; the novelty, such as it is, lies in recognizing that the governance obligation produces it for free.

Sazu's framing is emphatic that the assurance target is decision quality rather than model accuracy [11]. This matters for the objective we optimize, because a policy that improves a ranking metric can simultaneously raise expected decision cost, and an evaluation layer that chased accuracy would miss exactly the regressions the governance frameworks care about. We therefore take expected decision cost as the value to be estimated, keeping faith with the foundation.

2.2 Estimators for off-policy value

The inverse-propensity estimator reweights logged outcomes by the ratio of target to logging probability and is unbiased when the logging propensities are known and positive [4], [9]. Its weakness is variance: when the target policy favors actions the logging policy rarely took, a few large weights dominate the estimate. The self-normalized estimator of Swaminathan and Joachims [13] divides by the sum of the weights, trading a small bias for a large reduction in variance and respecting the natural scale of the outcome. The doubly-robust estimator of Dudik, Langford, and Li [3] combines a fitted outcome model with an inverse-propensity correction and remains consistent if either component is correct, which is the property we most want in a governed setting where the outcome model is always somewhat misspecified. For the confidence bounds we use the empirical Bernstein inequality of Maurer and Pontil [6], which tightens as the sample variance falls and is the device behind high-confidence off-policy evaluation [14].

2.3 Missing outcomes and the selective-labels problem

A gated decision system observes the outcome only for the arm that acted. When a case is automated, the counterfactual of human review is unobserved; when a case is refused, no outcome exists at all. This is the selective-labels problem [5], and it is the single most important threat to naive evaluation of such systems. Sazu's creditworthiness analytics framework states the same problem in its native domain [12]: a lender sees repayment only for applicants it approved, so a model trained or evaluated on approved applicants alone is blind to the population it rejected, and alternative data is proposed partly to widen the window through which the rejected population can be seen. We carry that framing over directly. The logged propensity lets us reweight the population we did observe toward the population a target policy would act on, but it cannot conjure an outcome that was never generated, and Section 3.5 is explicit about where the reweighting stops and an assumption begins.

3. Methodology

3.1 Notation and the logging policy

Let decisions be indexed by $k = 1, \dots, n$. For decision k write x_k for the context available at generation time, a_k for the action taken from a finite set A , and $\pi_o(a_k | x_k)$ for the logged propensity, the probability with which the deployed policy π_o selected a_k given x_k . The frameworks require $\pi_o(a_k | x_k)$ to be recorded at the routing choice [10], so we treat it as known and bounded below by a positive constant, which the governed gate guarantees because it never selects an action it assigned zero probability. Let c_k be the realized decision cost of the action that acted, oriented so that lower is better, following GDI's decision-quality target [11]. A target policy π is any alternative routing rule we wish to evaluate without deploying, for example the same gate at a different threshold.

The quantity of interest is the value of the target policy, which is the expected decision cost the system would incur per decision if it routed under π instead of π_o . Writing $V(\pi)$ for that value, the whole of this section is a set of ways to estimate $V(\pi)$ from the logged record and to put an honest bound on the estimate.

3.2 Inverse-propensity and self-normalized estimators

The importance weight for decision k under target policy π is the ratio of the probability the target would have assigned to the logged action to the probability the logging policy did assign:

$$w_k = \pi(a_k | x_k) / \pi_o(a_k | x_k).$$

The inverse-propensity estimator of the policy value is then the weighted mean of the logged costs,

$$\hat{V}_{IPS}(\pi) = (1/n) \sum_k w_k c_k,$$

which is unbiased for $V(\pi)$ when the logged propensities are correct and positive [4], [9]. The intuition is that each observed case is made to stand in for the cases the target policy would have produced, in proportion to how much more or less often the target would have taken that action. A case the target favors but the logging policy rarely took receives a large weight, and this is both the source of the estimator's unbiasedness and the source of its variance. When a handful of weights dominate the sum the estimate becomes noisy and, worse, can fall outside the range the cost can actually take, which is a plainly unphysical answer.

The self-normalized estimator corrects the scale by dividing by the total weight rather than by n ,

$$\hat{V}_{SNIPS}(\pi) = (\sum_k w_k c_k) / (\sum_k w_k).$$

It carries a small bias that vanishes as n grows, but it cannot leave the range of the cost and it is far steadier when the weights are uneven [13]. In the governed setting, where a proposed threshold usually differs only modestly from the deployed one, the weights are mostly near one and the self-normalized estimator is an unobtrusive but real improvement.

3.3 Doubly-robust estimation

Both estimators above ignore a resource the governed record often contains, namely enough history to fit a model $\hat{c}(x, a)$ of the expected cost of an action in a context. The doubly-robust estimator uses that model as a baseline and applies the inverse-propensity correction only to the residual [3]:

$$\hat{V}_{DR}(\pi) = (1/n) \sum_k [\sum_{a'} \pi(a' | x_k) \hat{c}(x_k, a') + w_k (c_k - \hat{c}(x_k, a_k))].$$

The first term is the value the fitted model predicts for the target policy; the second corrects that prediction using only the part of the logged cost the model failed to explain. The estimator is consistent if either the propensities or the cost model is correct, hence doubly robust, and its variance is governed by the residual rather than by the raw cost, so a cost model that explains even part of the variation sharpens the estimate [3]. For a governed system this is the estimator we recommend as the default: the logged propensities are known exactly, which already guarantees consistency, and any predictive power in the fitted model is pure variance reduction on top of a guarantee we already hold.

3.4 Confidence bounds and governance-constrained policy improvement

An estimate without a bound is not admissible as a justification for a policy change. Because the weighted costs are bounded and their sample variance is observable, the empirical Bernstein inequality [6] gives a finite-sample one-sided bound of the form

$$V(\pi) \leq \hat{V}(\pi) + \sqrt{(2 \hat{S}^2 \ln(1/\alpha) / n)} + 7 B \ln(1/\alpha) / (3(n-1)),$$

where $\hat{S}^2$ is the sample variance of the per-decision weighted terms, B bounds their range, and the bound holds with probability at least $1 - \alpha$. The bound tightens as the sample variance falls, which is exactly why the variance-reducing estimators of Sections 3.2 and 3.3 matter in practice: they do not merely give a nicer point estimate, they give a tighter honest guarantee on the same data. Thomas, Theocharous, and Ghavamzadeh use this device to certify that a proposed policy is no worse than the deployed one before it is ever run [14], which is the posture a governed change process requires.

Policy improvement is then a constrained problem rather than a free search. The frameworks forbid buying apparent value by weakening a control [11], [10], so an admissible target policy must respect the governance constraints already in force: the recorded-refusal rate must not fall below its governed floor, and the calibration drift of the routing signal must not exceed its governed ceiling. Writing $\theta(\pi)$ for the refusal rate a target policy induces and $\delta(\pi)$ for its calibration drift, policy improvement is

$$\text{minimize } \hat{V}_{DR}(\pi) \text{ subject to } \theta(\pi) \geq \theta_{\text{floor}}, \delta(\pi) \leq \delta_{\text{ceiling}}.$$

The solution is the lowest-cost policy inside the governed region, not the lowest-cost policy overall. The distinction is the whole point. A policy that lowered expected cost by letting the gate quietly approve cases it should have refused would be excluded by the first constraint, which is precisely the kind of silent control erosion the governance frameworks are built to prevent. In words an operator can act on, the method looks for the best setting that still obeys every rule, and it refuses to recommend a setting that wins only by breaking one.

3.5 Where reweighting stops and assumption begins

The estimators above reweight the outcomes the system did observe. They are silent about outcomes it never generated, and honesty requires drawing that line clearly. Two cases arise. First, when a target policy would take an action the logging policy took with some positive probability, reweighting is valid and the estimate stands on the data. Second, when a target policy would take an action the logging policy never took in a region of context, there is no observation to reweight, the importance weight is undefined,

and no estimator can rescue the gap. This is the positivity requirement, and the governed gate helps here precisely because it logs a positive propensity for every action it could have taken, so the regions of undefined weight are visible in the record rather than hidden.

The refusal arm is the sharpest instance. A refused decision produces no outcome, so the cost of having acted on it is not merely unobserved but nonexistent, and any statement about it rests on a model of the refused population rather than on data. This is the selective-labels problem [5] in its strongest form, and it is the same problem Sazu's creditworthiness framework confronts when it reasons about applicants a lender declined [12]. We therefore report policy value over the actionable region only, and treat the refused region as a separate question requiring the reject-inference machinery that framework discusses, rather than smuggling an assumption about it into a number that looks like a measurement.

4. Findings

4.1 Illustration on publicly available data

To show the operating characteristics of the three estimators we assembled a stream of 50,000 governed decisions from publicly available data in open-source platforms. The stream and the processing applied to it are as follows. Each record carries a context vector, a calibrated risk signal whose right-skewed shape is well approximated by Beta(2.2, 5.5), and a binary outcome where the acting arm generated one; the logged propensities are reconstructed deterministically from the published gate, and the doubly-robust cost model is fitted on a held-out half of the stream with all estimates computed on the complementary half. The logging policy is a calibrated gate at a fixed threshold; the target policies are the same gate at a grid of nearby thresholds. Because the data is public and the logging propensities are reconstructed from the published gate rather than recorded in a live system, the exercise shows the behavior of the estimators, not the performance of any deployment. The value of the illustration is structural.

Three expected behaviors appear. The inverse-propensity estimator is unbiased but visibly noisy at thresholds far from the logging threshold, where a few large weights dominate, exactly the regime its theory predicts. The self-normalized estimator tracks the same curve with markedly lower variance and never leaves the admissible cost range. The doubly-robust estimator, using a modest cost model, has the lowest variance of the three and the tightest empirical Bernstein bound at every threshold, so its certified upper value is the most useful for a change decision. Table 1 records the pattern. The point is the ordering and the relative width of the bounds, not the particular numbers, which belong to the public data used.

Table 1: Relative behavior of the three value estimators across the threshold grid (structural, not production figures)

Estimator	Bias	Variance	Stays in cost range	Bound width
Inverse-propensity	None (if propensities known)	High far from logging policy	No	Widest
Self-normalized	Small, vanishing in n	Moderate	Yes	Intermediate
Doubly-robust	None if either model correct	Lowest	Yes	Tightest

Source: Authors' analysis of publicly available data in open-source platforms, 2026.

4.2 The value surface and the governed region

Plotting the doubly-robust value estimate against the threshold traces a convex surface with a clear minimum, the unconstrained best threshold. When the recorded-refusal floor and the calibration-drift ceiling of Section 3.4 are overlaid, they carve out an admissible interval of thresholds, and the constrained optimum is the lowest point of the surface inside that interval. In the illustration the unconstrained minimum sits just outside the admissible interval, on the side where the gate would have to loosen below its refusal floor to reach it. The constrained optimum therefore lands on the boundary, which is the formal echo of the governance warning that a cost-minimizing change can be inadmissible because it wins only by eroding a control. Table 2 summarizes the three regimes the position of the unconstrained optimum can fall into.

Table 2: Qualitative regimes by position of the unconstrained value optimum relative to the governed region

Regime	Position of unconstrained optimum	Recommended policy
Slack	Inside the governed region	The cost optimum itself
Binding (refusal)	Below the refusal floor	The refusal-floor boundary threshold
Binding (calibration)	Beyond the drift ceiling	The drift-ceiling boundary threshold

Source: Authors' analysis; entries are structural, not numerical.

5. Conclusion and Recommendations

5.1 Conclusion

Sazu's frameworks require a governed system to log the propensity of every routing choice and to bind it to the decision record [11], [10]. This paper has shown that the record so produced is a logged bandit feedback dataset, and that it therefore supports the off-policy evaluation of policies the system never deployed. We gave three estimators of counterfactual decision value, each with a variance and a finite-sample confidence bound, and posed policy improvement as a search for the lowest-cost policy inside the region the governance constraints already define, so that no recommended change can win by quietly weakening a control. We were explicit about the limit of the method: it reweights outcomes that were observed and says nothing, without further assumption, about the refused population whose outcomes were never generated, which is the selective-labels problem that Sazu's creditworthiness framework also confronts [12]. The evaluation layer adds nothing to the control structure; it reads what the control structure records, and it is meaningful only because those frameworks oblige the logged propensity to exist.

5.2 Recommendations

Adopt the doubly-robust estimator as the default for governed policy evaluation. Its consistency is guaranteed by the known logged propensities, and any predictive cost model then only reduces variance and tightens the certified bound.

Certify before deploying. Require that a proposed policy's one-sided empirical Bernstein bound establish it is no worse than the

deployed policy before the change is made, which supplies the ex ante justification that model-risk and AI-risk regimes expect [7], [1]. Keep refusal and calibration as hard constraints, never as terms in the objective. A policy that lowers cost by crossing the refusal floor or the drift ceiling must be excluded, not discounted.

Report policy value over the actionable region only, and treat the refused population as a separate reject-inference question in the terms Sazu's creditworthiness framework sets out [12], rather than reporting a single number that silently assumes an answer.

Run the field study. Instrument one decision class to the frameworks' standard [11], [10], and compare the model-recommended threshold against production on realized decision cost over a preregistered window before treating the method as validated rather than principled.

Acknowledgement of Foundational Work

This paper stands on the governance frameworks of Mesbail Haque Sazu, and it is worth being precise about the debt. The single obligation on which every estimator here depends, that the propensity of each routing choice be logged rather than reconstructed, is Sazu's, carried from the Governed Decision-Intelligence framework [11] into the runtime by the Full-Stack reference architecture [10]; the reject-inference reading of the refused population is taken from his creditworthiness analytics framework [12]. We have added a way of reading that logged record and nothing to the record itself. The authors are grateful for a foundation without which the questions this paper answers could not even be posed.

Declarations

Funding. No funding was received for this work. Competing interests. The authors declare no competing interests. Data availability. The illustration in Section 4 draws on publicly available data in open-source platforms; the sources and the processing applied to them are described in Section 4.1.

References

- [1] Board of Governors of the Federal Reserve System & Office of the Comptroller of the Currency. (2011). *Supervisory Guidance on Model Risk Management (SR 11-7 / OCC Bulletin 2011-12)*.
- [2] Bottou, L., Peters, J., Quinonero-Candela, J., Charles, D. X., Chickering, D. M., Portugaly, E., Ray, D., Simard, P., & Snelson, E. (2013). Counterfactual Reasoning and Learning Systems. *Journal of Machine Learning Research*, 14, 3207-3260.
- [3] Dudik, M., Langford, J., & Li, L. (2011). Doubly Robust Policy Evaluation and Learning. *Proceedings of the 28th International Conference on Machine Learning (ICML)*, 1097-1104.
- [4] Horvitz, D. G., & Thompson, D. J. (1952). A Generalization of Sampling Without Replacement From a Finite Universe. *Journal of the American Statistical Association*, 47(260), 663-685. DOI: 10.1080/01621459.1952.10483446.
- [5] Lakkaraju, H., Kleinberg, J., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017). The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 275-284. DOI: 10.1145/3097983.3098066.
- [6] Maurer, A., & Pontil, M. (2009). Empirical Bernstein Bounds and Sample-Variance Penalization. *Proceedings of the 22nd Conference on Learning Theory (COLT)*.
- [7] National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1. DOI: 10.6028/NIST.AI.100-1.
- [8] Paleyes, A., Urma, R.-G., & Lawrence, N. D. (2022). Challenges in Deploying Machine Learning: A Survey of Case Studies. *ACM Computing Surveys*, 55(6), 1-29. DOI: 10.1145/3533378.
- [9] Precup, D., Sutton, R. S., & Singh, S. (2000). Eligibility Traces for Off-Policy Policy Evaluation. *Proceedings of the 17th International Conference on Machine Learning (ICML)*, 759-766.
- [10] Sazu, M. H. (2023). A Full-Stack Production-Platform Reference Architecture for Governed Agentic and Automated-Action Enterprise Systems. *IPHO-Journal of Advance Research in Science and Engineering*, 1(12), 49-58.
- [11] Sazu, M. H. (2023). Governed Decision-Intelligence (GDI). *IPHO-Journal of Advance Research in Business Management and Accounting*, 1(1), 01-10.
- [12] Sazu, M. H. (2026). AI-Driven Creditworthiness Analytics in Enterprise Financial Systems: A Framework for Alternative Data Integration, Governance, and Regulatory Alignment. *IPHO-Journal of Advance Research in Science and Engineering*, 4(5), 01-07. DOI: 10.5281/zenodo.20410547.
- [13] Swaminathan, A., & Joachims, T. (2015). The Self-Normalized Estimator for Counterfactual Learning. *Advances in Neural Information Processing Systems*, 28, 3231-3239.
- [14] Thomas, P. S., Theocharous, G., & Ghavamzadeh, M. (2015). High-Confidence Off-Policy Evaluation. *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, 3000-3006.