Galileo

Technical Reference

Architecture, Algorithms, Evidence, and Limitations

Anjan Goswami

SmartInfer Inc.

April 2026

For data scientists, analytics engineers, ML scientists, and technical buyers

Contents

Executive Summary

Galileo is a marketing mix modeling system designed around three technical principles:

1. **Estimation–exploration coupling**: Exploration policy quality is bounded by the structural estimator's ability to track response curve parameters online. Advanced exploration paired with a basic estimator degrades performance by 4.19× (540-run benchmark).
2. **Trust-aware outputs**: Every user-facing number carries a computed confidence score derived from model quality, data density, scenario magnitude, and credibility gate results. Low-confidence outputs are gated: dollar amounts suppressed, directional guidance preserved.
3. **Honest limitations**: Confidence intervals propagate coefficient uncertainty but not curve-shape uncertainty. Counterfactuals are model-implied re-evaluations, not causal estimates. These boundaries are disclosed in every output.

This paper documents the system architecture, core algorithms, experimental evidence, trust mechanisms, and known limitations.

Evidence base: 3,240-run controlled benchmark (18 synthetic markets), 540-run estimator–explorer coupling study, 600-run grounded validation (siMMMulator true parameters), and cross-framework comparison against Meta Robyn and Google Meridian.

System Architecture

Galileo consists of seven layers, each independently testable:

Layer	Function
Data ingestion	CSV validation, time-series structure checks, confounder residualization
Structural estimator	Online Hill curve fitting via NLS, OLS standard errors, Fisher information matrix
Optimizer	Budget-constrained allocation via SLSQP (multi-start) with DE+SLSQP fallback for $S \geq 4$
Exploration layer	5 decision policies (ETO, Thompson, IDASat, DensityBayes, GaussianBED)
Controller	Rule-based policy selector: data profile → estimator–explorer pairing, decision trace
Trust / adjudication	Dynamic confidence scoring, credibility gating, CI honesty disclosure
Grounded chat	Structured retrieval over pipeline artifacts, regex-routed Q&A, trust-gated responses

Controller design

The controller operates in two modes:

Mode	Trigger	Behavior
Conservative (ETO)	Channels well-identified: low Fisher gaps, stable K/S, ≥20 obs/channel	Optimize directly. No exploration cost.
Adaptive (DensityBayes)	Underidentified: high Fisher eigenvalue ratio, unstable estimates, short history	Directed exploration. Requires `adaptive_exploration_safe` flag.

The controller *blocks* DensityBayes when the estimator lacks structural curve-shape updates, citing the EX3a benchmark result (4.19× regret) as evidence. This gating is deterministic and traceable.

Core Response Model

Hill saturation model

Channel c revenue response to weekly spend x_c:

$$h(x_c; K_c, S_c) = \frac{x_c^{S_c}}{K_c^{S_c} + x_c^{S_c}} \tag{1}$$

where K_c is the half-saturation point (spend at 50% of maximum response) and S_c controls curve steepness. Total media-driven revenue:

$$\hat{y}_t = \alpha + \sum_{c=1}^{C} \beta_c \cdot h(x_{c,t}; K_c, S_c) + \epsilon_t \tag{2}$$

Parameter estimation: Multi-start NLS via `scipy.optimize.least_squares` (L-BFGS-B, 5 restarts). Ridge regularization with OLS-proportional penalty: 70% OLS-proportional + 30% equal-share blend, information-scaled. Standard errors from OLS (not ridge) to preserve CI honesty—ridge SEs underestimate uncertainty, while OLS SEs naturally incorporate variance inflation from multicollinearity.

K initialization: $K_{\text{init}} = \text{median}(x_c^{\text{raw}})$, using raw spend (not adstocked). For high-adstock channels (e.g., TV with $\lambda = 0.82$), adstocked median is 6× raw median, which biases K upward.

Adstock

Geometric adstock with decay parameter $\lambda_c \in [0, 1)$:

$$\tilde{x}_{c,t} = x_{c,t} + \lambda_c \cdot \tilde{x}_{c,t-1} \tag{3}$$

When adstock is modeled, λ_c is estimated jointly with K_c and S_c via logit-transformed NLS. Adstock is a known identification challenge: all three benchmark systems (Galileo, Robyn, Meridian) show 77–93% mean error on λ recovery with aggregate weekly data.

Fisher information matrix

The Fisher information for channel c at observation x_c:

$$\mathcal{I}(\theta_c; x_c) = \left(\frac{\partial h}{\partial \theta_c}\right)^{\top} \left(\frac{\partial h}{\partial \theta_c}\right) / \sigma^2 \tag{4}$$

where $\theta_c = (\beta_c, K_c, S_c)$. The Fisher matrix quantifies how much each observation contributes to parameter precision. Eigenvalue analysis of $\mathcal{I}$ identifies which parameters are well-determined and which are uncertain—this drives exploration targeting.

Decision Policies

Galileo implements five budget allocation policies:

Policy	Type	Mechanism
ETO	Exploit-only	Estimate-Then-Optimize: fit model, optimize allocation, repeat. No exploration.
Thompson	Stochastic	Sample parameters from posterior, optimize against sample. Explores in all directions uniformly.
IDASat	Targeted	Information-Directed Allocation with Saturation awareness. Balances regret reduction against information gain.
DensityBayes	Targeted	Density-matrix Bayesian updates. Tracks per-parameter directional uncertainty via Fisher-weighted density matrices.
GaussianBED	Targeted	Gaussian Bayesian Experimental Design. Selects allocations that maximize expected information gain under Gaussian posterior.

DensityBayes exploration objective

DensityBayes selects the allocation $\mathbf{x}$ that maximizes a combined exploitation–exploration objective:

$$\mathbf{x}^* = \arg\max_{\mathbf{x}\in\mathcal{B}} \left[\hat{R}(\mathbf{x}) + \eta \cdot \mathrm{tr}\left(\rho(\mathbf{x}) \cdot \mathcal{I}^{-1}(\hat{\theta})\right)\right] \tag{5}$$

where $\hat{R}(\mathbf{x})$ is the predicted revenue, $\rho(\mathbf{x})$ is the density matrix encoding directional uncertainty at allocation $\mathbf{x}$, $\mathcal{I}^{-1}$ is the inverse Fisher information (parameter uncertainty), η is the exploration–exploitation trade-off parameter, and $\mathcal{B}$ is the budget constraint set.

The density matrix ρ concentrates exploration toward the directions of highest parameter uncertainty, as measured by the Fisher information spectrum. This is more targeted than Thompson sampling (which explores uniformly) but requires the estimator to maintain accurate K and S estimates for the Fisher computation to be meaningful.

GaussianBED exploration objective

GaussianBED selects allocations maximizing expected information gain under a Gaussian posterior approximation:

$$\mathbf{x}^* = \arg\max_{\mathbf{x}\in\mathcal{B}} \left[\hat{R}(\mathbf{x}) + \eta \cdot \frac{1}{2}\log\det\left(I + \sigma^{-2} J(\mathbf{x})^\top \Sigma J(\mathbf{x})\right)\right] \tag{6}$$

where $J(\mathbf{x})$ is the Jacobian of the Hill model at $\mathbf{x}$ and Σ is the current posterior covariance. Under correct model specification, DensityBayes and GaussianBED are statistically indistinguishable ($p > 0.24$ across 2,700 runs). They diverge under misspecification (Appendix H of the companion paper), where DensityBayes retains advantage while GaussianBED loses it.

Main Technical Finding: Exploration Is Bounded by Estimation

The central result from Galileo's benchmark program is:

> **Exploration quality is bounded by the structural estimator's quality.** An advanced exploration policy paired with an estimator that does not track response curve shape online will degrade performance, not improve it.

EX3a experiment (540 runs)

The EX3a benchmark tested all 5 exploration policies paired with 3 estimator types across 18 worlds:

Pairing	Mean alloc. regret	Safe?
DensityBayes + structural estimator	\$12,400	Yes
DensityBayes + basic estimator (no K/S updates)	\$51,900 (4.19×)	**No**
Thompson + any estimator	\$18,200	Yes
ETO + any estimator	\$22,800	Yes

Table 1: Estimator–explorer coupling. DensityBayes with a basic estimator is 4.19× worse than with a structural estimator. Thompson and ETO are estimator-agnostic.

Root cause: DensityBayes computes exploration directions from the Fisher information matrix, which depends on the partial derivatives of the Hill function with respect to K and S. If the estimator holds K and S fixed (basic mode), the Fisher matrix reflects stale parameters, and exploration is directed to the wrong parts of the response curve.

Controller gating

This result directly motivates the controller's gating logic: DensityBayes and GaussianBED are blocked unless the estimator carries the `adaptive_exploration_safe` flag, indicating it performs online K/S updates. The controller falls back to Thompson (which is estimator-agnostic) when the structural estimator is not available.

Estimator	DensityBayes	Thompson	ETO
Structural (online K/S)	Safe	Safe	Safe
Basic (β-only)	**Unsafe (4.19×)**	Safe	Safe

Table 2: Safe and unsafe estimator–explorer pairings. The controller enforces this matrix.

Experimental Evidence

Controlled benchmark (3,240 runs)

18 synthetic worlds: 3 time horizons × 3 collinearity levels × 2 shape regimes. 5 channels per world, log-normal noise ($\sigma = 0.05$). 30 seeds per (world, policy) pair.

Allocation regret (how close to optimal by the end):

- DensityBayes vs. ETO: 8.6% improvement in hard-identification worlds ($p = 0.023$, Wilcoxon)
- GaussianBED vs. ETO: similar improvement ($p < 0.05$)
- Thompson vs. ETO: modest improvement, not always significant

Cumulative regret (total revenue lost during learning):

- DensityBayes vs. ETO: $p = 0.007$ (significant)
- Exploration costs less than it saves: 4.7% higher cumulative revenue

All p-values from paired Wilcoxon signed-rank tests (non-parametric, no normality assumption).

Grounded validation (600 runs)

Parameters derived analytically from Meta's siMMMulator true DGP: $S \in [1.3, 2.5]$, 5 channels, \$880K weekly budget. 12 worlds (10 well-specified + 1 adstock misspecification + 1 interaction misspecification). 10 seeds per (world, policy).

Policy	Mean	Median	P25	P75	P90
ETO	\$2,951	\$2,765	\$1,493	\$4,308	\$5,448
Thompson	\$2,538	\$2,110	\$1,133	\$3,806	\$5,296
DensityBayes	\$2,566	\$2,411	\$1,224	\$3,699	\$4,613
GaussianBED	\$2,824	\$2,473	\$1,493	\$3,966	\$4,768
IDASat	\$2,815	\$2,456	\$1,470	\$3,773	\$4,826

Table 3: Allocation regret (\$) on siMMMulator-grounded parameters. All Wilcoxon $p > 0.05$. Regret is $< 0.3\%$ of weekly budget for all policies.

Interpretation: On realistic parameters with smooth Hill curves ($S < 2.5$), the online fitter converges within ∼20 observations per channel, and all policies allocate near-optimally. Policy differentiation is regime-dependent: it emerges in hard-identification conditions (high collinearity, steep saturation, short horizons) and vanishes when identification is easy.

Cross-framework comparison

Galileo, Robyn (Meta), and Meridian (Google) benchmarked on the same 156-week, 5-channel dataset with known ground truth (siMMMulator).

Dimension	Galileo	Robyn	Meridian
Method	NLS (Hill)	Ridge + evolution	Bayesian MCMC
Runtime	0.2s	∼5 min	∼1 min
Top channel	Meta (correct)	Meta (correct)	Not extractable
Saturation recovery	Different scale*	21.9% mean error	Fixed at 1.0
Adstock recovery	Not modeled	92.5% error	77.4% error
Uncertainty	Fisher CIs	None	Full posterior
Online updating	Yes	No	No
Exploration-ready	Yes	No	No

Table 4: Cross-framework comparison. *Galileo fits Hill on raw spend (no adstock transform), so K/S are on a different scale.

Channel *ranking* is robust across all three systems. Adstock decay is poorly recovered by all (77–93% mean error)—a fundamental identification limitation with aggregate weekly data.

Trust Layer

Dynamic confidence scoring

Scenario confidence is computed as a weighted average of measurable inputs:

$$\text{conf}(\mathbf{s}) = w_1 \cdot f(R^2) + w_2 \cdot g(\text{CI\_cov}) + w_3 \cdot (1 - \text{shift\_mag}) + w_4 \cdot d(\text{density}) + w_5 \cdot \mathbb{1}[\text{cred} = \text{PASS}] \quad (7)$$

where:

- $w_1 = 0.30$ (model fit), $w_2 = 0.20$ (CI coverage), $w_3 = 0.25$ (spend shift), $w_4 = 0.15$ (data density), $w_5 = 0.10$ (credibility)
- $f(R^2)$ maps model fit to $[0, 1]$
- $g(\text{CI\_cov})$ measures whether CIs actually contain true values (when testable)
- $\text{shift\_mag} = \max_c |x'_c - x_c|/x_c$ captures the largest relative spend change
- $d(\text{density})$ reflects per-channel data sufficiency

Confidence classes and gating

Class	Range	Chat behavior	Counterfactual behavior
HIGH	≥ 0.70	Full answer with dollar amounts and CIs	Full scenario with dollar impact and factors
CAUTION	$[0.40, 0.70)$	Dollar estimates shown with caveat: "model-implied, not predictive"	Scenario shown with extrapolation warning
LOW	< 0.40	Dollar amounts suppressed. Directional guidance only.	Trust warning: "Treat as directional only"

Table 5: Trust gating behavior by confidence class.

CI honesty

Confidence intervals propagate β uncertainty only (substituting β_{lo} and β_{hi} from OLS standard errors). K and S are treated as fixed. This produces CIs that are narrower than full-parameter uncertainty would warrant. Every CI-bearing output includes the disclosure: "beta-only uncertainty; K/S parameters treated as fixed."

Full-parameter propagation (delta-method or parametric bootstrap through the Hill parameter vector) is on the roadmap but not yet implemented.

Credibility certificate

40+ deterministic quality gates organized into categories:

- **Model fit**: R^2, MAPE, Durbin-Watson, residual normality
- **Parameters**: β sign consistency, K/S bounds, CI width reasonableness
- **Saturation**: monotonicity, coverage of spend range
- **Interactions**: false positive rate, consistency with priors
- **Optimization**: convergence, budget constraint satisfaction

Verdicts: PASS (all gates clear), FLAG (advisory issues), FAIL (material concerns). The verdict feeds into the confidence score via w_5.

Interview Priors

Galileo supports a guided interview that captures domain knowledge from marketing practitioners. This section documents where that knowledge enters the system, what is mathematically sound, and what is heuristic.

Current integration points

Integration	Status	Mathematical basis
Interaction cascade voting	Implemented	Heuristic: prior adds +1 vote to 3-method cascade
Prior-only detection (Mode 3)	Implemented	Heuristic: conf $\geq$ 4, strong evidence $\rightarrow$ PRIOR_SUPPORTED
Calibration divergence check	Implemented	Sound: |model $-$ calibration|/calibration, thresholds at 30%/50%
Prior validation gate	Implemented	Sound: 5 deterministic checks (impossible, direction, magnitude, lag, evidence)
Prior quantifier	Implemented	Sound: $\sigma = \min(\sigma_0 / \max(w, 0.1), 0.40)$

PriorQuantifier

Evidence type maps to prior standard deviation via inverse-weight regularization:

$$\sigma_{\text{prior}} = \min\left(\frac{\sigma_0}{\max(w_{\text{evidence}}, 0.1)}, 0.40\right) \tag{8}$$

where evidence weights are: experiment = 1.0, natural experiment = 0.8, tracked = 0.5, strong belief = 0.3, vague impression = 0.15, speculation = 0.05. Strong evidence yields low σ (confident prior); weak evidence yields high σ (skeptical prior). The cap at 0.40 prevents completely uninformative priors.

Cascade voting mechanism

The interaction detection cascade runs three statistical methods (partial F-test, residual correlation, block bootstrap). A prior adds +1 to the vote count if prior confidence ≥ 3. Detection threshold: $\geq$2 total votes. This is a tie-breaker heuristic—it can flip a 1-vote pair to detected, but cannot rescue a 0-vote pair (unless Mode 3 applies).

Mode 3 (prior-only): If confidence ≥ 4 and evidence type is "experiment" or "natural_experiment," the interaction is marked PRIOR_SUPPORTED even with 0 data votes. Coefficient set to μ_{prior}, p-value = 1.0 (placeholder). This is a business decision to trust strong domain knowledge over insufficient statistical power.

Not yet integrated

- Fitter K/S initialization from saturation beliefs (Priority 1 on roadmap)
- Interview-driven custom counterfactual scenarios (Priority 2)
- Controller risk-stance from interview (Priority 3)
- Bayesian prior combination in cascade (future work—current heuristic is adequate)
- Exploration geometry injection (tested and parked: correct priors hurt performance by 11–12% due to rank-1 phantom confidence in Fisher matrix)

Limitations

Synthetic-heavy evidence

The primary benchmark is fully synthetic (18 worlds with controlled difficulty). The grounded validation uses siMMMulator true parameters but the data-generating process is still simulated. The cross-framework comparison is a single-dataset sanity check. No experiments use real advertising data. Policy differentiation should be validated on real-world pilots.

Model misspecification

The Hill saturation model assumes a specific functional form. If the true response is not Hill-shaped (e.g., threshold effects, non-monotone responses), the fitted parameters and derived quantities (saturation levels, optimal allocations) may be systematically biased. The credibility certificate catches some misspecification symptoms (R^2, residual patterns) but cannot detect all forms.

Uncertainty propagation

CIs propagate β uncertainty only. K and S are treated as known. For scenarios involving channels near their half-saturation point (where the Hill curve is steepest), K uncertainty has outsized impact on revenue predictions. Full-parameter delta-method or bootstrap CIs would be wider and more honest.

Counterfactuals are model-implied

All counterfactual estimates answer: "Under the fitted model, what revenue would we predict under an alternative spend plan?" They do not answer: "What would actually happen?" The distinction matters for large reallocations, market regime changes, or scenarios involving channels with thin data.

Interaction detection power

With $N = 52$ weekly observations and typical model R^2, the minimum detectable interaction effect (MDE) is approximately 0.40 across all three cascade methods. True interactions at 0.12–0.30 (common in practice) are below the detection threshold. Absence of detected interactions does not mean channels are independent.

No formal regret bound

The estimation–exploration coupling principle is established empirically. The Fisher information argument is structural and geometric, not a formal proof. No regret bound or convergence rate is provided.

Adstock identification

Adstock decay (λ) is poorly recovered by all benchmark systems (77–93% mean error). This is a fundamental identification challenge with aggregate weekly data: when spend patterns are temporally correlated, separating "this week's effect" from "last week's carryover" is mathematically underdetermined.

Priority	Item	Impact
1	Full-parameter CI propagation	Honest confidence intervals (delta-method through β, K, S)
2	Real-world pilot validation	Evidence on real advertising data, not synthetic
3	Fitter K initialization from interview beliefs	Better convergence in early periods (warm-start)
4	Formal Bayesian prior combination	Replace cascade voting heuristic with posterior update
5	Broader external benchmark	Multiple real datasets, diverse verticals
6	Structural model extensions	Confounder adjustment, regime-switching, non-Hill response forms

Roadmap

Full-parameter uncertainty

The highest-priority improvement is replacing β-only CIs with delta-method or parametric bootstrap propagation through the full Hill parameter vector (β, K, S). This produces wider, more honest intervals—especially for scenarios near the half-saturation point where K uncertainty dominates.

Real-world validation

The synthetic benchmark establishes controlled comparisons. Real-world pilots would validate whether the policy differentiation observed in hard-identification synthetic worlds transfers to actual advertising environments.

Bayesian prior integration

The current +1 vote heuristic in the interaction cascade is pragmatic but not formally Bayesian. Proper posterior combination (combining prior $N(\mu_{\text{prior}}, \sigma^2_{\text{prior}})$ with data likelihood) would be more principled. Parked until a customer dataset demonstrates the heuristic's failure mode.