Regulatory
Studies Center
THE GEORGE WASHINGTON UNIVERSITY

EVIDENCE-FIRST REGULATION: A PROCEDURAL ARCHITECTURE FOR GOVERNANCE UNDER DEEP UNCERTAINTY

Steven M. Wegener

SEPTEMBER 2026

© 2026 by author. All rights reserved.

The GW Regulatory Studies Center is an academic center in the George Washington University and its Trachtenberg School of Public Policy and Public Administration. This working paper is part of a series on Regulatory Approaches to Emerging Technologies supported by the John Templeton Foundation. Any views expressed in this working paper are those of the author and not necessarily those of the John Templeton Foundation, the GW Regulatory Studies Center, the Drug Enforcement Administration, the U.S. Department of Justice, or the United States Government.

https://regulatorystudies.columbian.gwu.edu/

ABSTRACT

Federal rulemaking has grown increasingly sophisticated at ex ante analysis while remaining comparatively underdeveloped at ex post learning. This imbalance is manageable when regulatory impacts are reasonably estimable and value premises broadly shared. It becomes costly in domains marked by rapid change and contested normative premises—artificial intelligence, autonomous systems, advanced biotechnology—where assumptions drift quickly and feedback arrives unevenly.

This paper proposes *Evidence-First Regulation*, a procedural architecture for organizing regulatory reasoning under deep uncertainty. The Three Decision States distinguish what can be reasonably estimated at promulgation (State I), what cannot be predicted ex ante but can be learned through implementation (State II), and what rests on normative judgment rather than empirical resolution (State III). The Four Pillars operationalize that structure: *Adequacy over Precision*, a stop-rule for modeling grounded in marginal informational value; the *Adaptive Statement*, a *Learning Summary* outlining the rule's learning agenda—the State II uncertainties whose resolution would inform future decisions; the *Decision-Rule Table*, which makes normative premises explicit rather than embedding them in technical parameters; and the *Review Spine*, nonbinding scheduled and condition-based review points that connect ex ante assumptions to ex post evidence.

The architecture operates within existing notice-and-comment authority: no statutory change, no new enforceable procedural rights, and *State Farm* review intact. It complements benefit–cost analysis rather than replacing it, redirecting diminishing-returns modeling effort toward structured learning and transparent normative reasoning. A worked example applies the framework to a hypothetical AI auditing rule for high-risk systems. A review of current federal practice (EPA, FDA, NHTSA, SEC, FWS, and the banking agencies) finds that agencies facing uncertainty-dominant conditions have independently developed framework-like elements, suggesting the architecture names and makes portable a pattern already emerging under modern regulatory conditions rather than imposing a new one.

AUTHOR

Steven M. Wegener[*] is an Economist at the Drug Enforcement Administration.

[*] Contact: smw204@gmail.com.

ACKNOWLEDGEMENTS

This publication was made possible through the support of Grant 63641 from the John Templeton Foundation. The opinions expressed in this publication are those of the author and presented for academic discussion. They do not necessarily reflect the views of the John Templeton Foundation, the GW Regulatory Studies Center, the Drug Enforcement Administration, or the U.S. Department of Justice.

This paper was prepared for the George Washington University Regulatory Studies Center's Research Roundtable on Regulatory Approaches to Emerging Technologies, and I gratefully acknowledge the Center's support. I thank Susan Dudley and William Yeatman for organizing the roundtable and for their guidance throughout; Nick Hart, the paper's discussant, for his generous comments; and Caroline Cecot, Alta Charo, Brian Mannix, Margaret Taylor, Jonathan Wiener, and the other roundtable participants for helpful comments and discussion. The views expressed are my own and do not necessarily reflect those of the Drug Enforcement Administration, the U.S. Department of Justice, or the United States Government. All errors are mine.

ABOUT THE GEORGE WASHINGTON UNIVERSITY REGULATORY STUDIES CENTER

Established in 2009, the GW Regulatory Studies Center is an academic center of the George Washington University and its Trachtenberg School of Public Policy and Public Administration. The Center's mission is to improve regulatory policy through research, education, and outreach. The Center is a leading source for applied scholarship in regulatory issues, and a training ground for anyone who wants to understand the effects of regulation and ensure that regulatory policies are designed in the public interest.

I. Introduction — The Learning Gap in Modern Rulemaking

Federal rulemaking today reflects an impressive commitment to analytic rigor. Major regulations are accompanied by detailed Regulatory Impact Analyses (RIAs), extensive interagency review under Executive Order 12866, and formal notice-and-comment procedures designed to surface relevant evidence and arguments. Agencies routinely estimate compliance costs, model behavioral responses, and test sensitivity assumptions. In many domains, they do so with increasing methodological sophistication.[1]

Yet alongside this rigor lies a structural imbalance. Rulemaking remains heavily oriented toward ex ante prediction, while institutionalized mechanisms for ex post learning are comparatively underdeveloped. Agencies devote substantial effort to forecasting costs and benefits at promulgation, but far less to systematically measuring whether predicted outcomes occur,[2] whether assumptions hold under real-world conditions, or whether emerging evidence warrants reconsideration. Retrospective review is formally required, but not systematically embedded within the analytic architecture of rulemaking itself.

This imbalance is not the result of indifference or institutional failure. It is a predictable consequence of how the regulatory process concentrates analytic effort. Executive Order 12866 directs agencies to adopt regulations only upon "a reasoned determination that the benefits of the intended regulation justify its costs," and to quantify impacts "to the fullest extent that these can be usefully estimated."[3] Circular A-4 reinforces that estimates should not create a "false sense of precision"[4] and cautions that analysis must reflect the limits of available evidence. These directives emphasize reasoned judgment, not numerical perfection. At the same time, they place understandable weight on ex ante modeling because that is the portion of analysis most readily documented, reviewed, and audited at the time of promulgation.

Retrospective review, by contrast, appears elsewhere in the executive framework. Executive Order 12866 instructs agencies to "periodically review" significant regulations to determine whether they should be modified or eliminated. Executive Order 13563 reaffirmed this commitment.[5] In parallel, the Foundations for Evidence-Based Policymaking Act has

---

[1] Susan E. Dudley & Jerry Brito, Regulation: A Primer (2d ed., George Washington Univ. Regulatory Studies Ctr. 2012).

[2] Art Fraas, Elizabeth Kopits & Ann Wolverton, A Retrospective Review of Retrospective Cost Analyses, EPA Nat'l Ctr. for Envtl. Econ. Working Paper No. 21-04 (Sept. 2021), published as Retrospective Cost Analysis, 17 Rev. Envtl. Econ. & Pol'y 22 (2023). The authors found that firms adopted different compliance strategies than anticipated for nearly 70 percent of thirteen EPA regulations studied and recommended ex ante retrospective study planning.

[3] Exec. Order No. 12866, 58 Fed. Reg. 51,735, §§ 1(a), 1(b)(6) (Sept. 30, 1993).

[4] Office of Mgmt. & Budget, Circular A-4, Regulatory Analysis (Sept. 17, 2003). The 2023 revision (Nov. 9, 2023) was rescinded in 2025 and the 2003 Circular reinstated. Both versions carry the caution against a "false sense of precision" quoted in the text (2003 at 40; 2023 at 70).

[5] Exec. Order No. 12866, supra note 3, § 5(a); Exec. Order No. 13563, 76 Fed. Reg. 3821 (Jan. 21, 2011).

strengthened evaluation capacity within agencies.[6] What remains less developed is a procedural architecture that connects ex ante analysis with ex post evidence in a systematic and transparent way.[7] The Evidence Act builds evaluation capacity across agencies; this framework integrates that capacity into rule-level analytic design and opens the resulting learning agenda to researchers, industry, and the public—who can see what the agency intends to measure, refine it through notice-and-comment, and contribute evidence of their own during the rulemaking or in anticipation of the review.

This architecture is not without precedent. EPA's review of the National Ambient Air Quality Standards, refined through five decades of Clean Air Act implementation and judicial review, embodies much of this structure—grading evidence by confidence level, connecting promulgation to mandatory review, and recognizing the Administrator's judgment as normative discretion. Yet even this mature process has gaps: its normative judgments remain embedded in the scientific record rather than surfaced as separable premises; its review cycle is time-based rather than condition-based, and it operates under unique statutory authority most agencies lack.

Each component of the framework proposed below has precedent in existing federal practice; what no current rulemaking does is implement them together as a coherent architecture. The contribution of this paper is to make the underlying architecture's logic explicit, portable, and complete.

The need for such an architecture is most evident in domains marked by rapid change and uncertainty. In slower-moving sectors, gaps between predicted and realized outcomes tend to emerge gradually, and institutions can absorb modest forecasting error. In faster-moving areas—such as artificial intelligence, autonomous systems, digital platforms, and advanced biotechnology—uncertainty appears more quickly and more visibly. Model drift, context-dependent performance, and feedback effects can undermine static assumptions. Under these conditions, an approach focused primarily on one-time prediction risks entrenching estimates that may diverge from real-world outcomes.

The underlying issue is not that prediction is misguided. Forecasting remains a critical directional tool, and benefit–cost analysis remains the analytic engine for estimable effects. Nothing in this framework alters the welfare objective embedded in Executive Order 12866. But BCA alone does not exhaust the analytic demands of rulemaking under deep uncertainty.[8] When

[6] Foundations for Evidence-Based Policymaking Act of 2018, Pub. L. No. 115-435, 132 Stat. 5529 (2019) (codified in scattered sections of 5 U.S.C.).
[7] U.S. Gov't Accountability Office, GAO-14-268, Reexamining Regulations: Agencies Often Made Regulatory Changes, but Could Strengthen Linkages to Performance Goals (Apr. 2014). GAO found that agencies did not consistently link retrospective analyses to performance goals or embed evaluation planning within rule-level analytic structure.
[8] Deep uncertainty is a term of art from the decision-analysis literature, describing conditions in which analysts do not know, or the parties to a decision cannot agree on, the appropriate system model, the probability distributions over key parameters, and/or the valuation of alternative outcomes. See Robert J. Lempert, Steven W. Popper &

key effects are learnable only through post-implementation observation, and when some elements of a decision rest on contested normative judgments, the analytic architecture needs to extend beyond estimation—to structured learning and transparent normative reasoning. The framework provides that extension: BCA handles what can be estimated; the *Learning Summary* plans for what must be measured over time; and the *Decision-Rule Table* surfaces the value judgments that no amount of data will resolve.

The challenges addressed here have received sustained scholarly attention. Adaptive governance theorists have developed frameworks for managing regulatory problems under conditions of deep uncertainty, emphasizing institutional flexibility, iterative decision-making, and feedback as alternatives to static ex ante optimization.[9] Experimentalist approaches to governance have explored how regulatory commitments can be structured as provisional and subject to deliberate revision based on performance data, embedding learning mechanisms into the architecture of governance itself.[10] Evidence Act scholarship has examined how federal evaluation infrastructure can be strengthened to support more systematic learning across agencies.[11]

The contribution of this paper is to integrate existing analytic practices into a single procedural architecture that operates at the level of individual rule design—actionable structure instead of aspirational theory. Existing approaches address modeling, evaluation, or adaptive governance separately; this framework embeds all three within the administrative record at promulgation. It does so by introducing an explicit epistemic grammar—the Three Decision States—that distinguishes what can be reasonably estimated, what must be learned through observation, and what reflects normative judgment.

Building on this structure, the paper develops four procedural tools—the Four Pillars of Evidence-First Regulation. *Adequacy over Precision* formalizes a stop-rule for modeling

---

Steven C. Bankes, Shaping the Next One Hundred Years: New Methods for Quantitative, Long-Term Policy Analysis (RAND 2003); Warren E. Walker, Robert J. Lempert & Jan H. Kwakkel, Deep Uncertainty, in Encyclopedia of Operations Research and Management Science 395 (Saul I. Gass & Michael C. Fu eds., 3d ed. 2013); see generally Vincent A. W. J. Marchau, Warren E. Walker, Pieter J. T. M. Bloemen & Steven W. Popper eds., Decision Making under Deep Uncertainty: From Theory to Practice (Springer 2019).

[9] Robin Kundis Craig & J.B. Ruhl, Designing Administrative Law for Adaptive Management, 67 Vand. L. Rev. 1 (2014). See also Lori S. Bennear & Jonathan B. Wiener, Built to Learn: From Static to Adaptive Environmental Policy, in A Better Planet: Forty Big Ideas for a Sustainable Future 490 (Daniel C. Esty ed., 2019) (arguing that policies be "built to learn," with data collection, review, and revision processes incorporated in the initial regulation); Bradley C. Karkkainen, Adaptive Ecosystem Management and Regulatory Penalty Defaults, 87 Minn. L. Rev. 943 (2003).

[10] Charles F. Sabel & Jonathan Zeitlin, Experimentalist Governance, in The Oxford Handbook of Governance 169 (David Levi-Faur ed., 2012); Charles F. Sabel & William H. Simon, Minimalism and Experimentalism in the Administrative State, 100 Geo. L.J. 53 (2011).

[11] Foundations for Evidence-Based Policymaking Act, supra note 6. See also Advisory Committee on Data for Evidence Building, Year 1 Report (Oct. 2021); Office of Mgmt. & Budget, M-19-23, Phase 1 Implementation of the Foundations for Evidence-Based Policymaking Act of 2018 (July 10, 2019). See generally Kathryn E. Newcomer & Nicholas Hart, Evidence-Building and Evaluation in Government (SAGE 2022) (examining the Act's learning-agenda and evaluation architecture).

grounded in marginal informational value, in line with Circular A-4's guidance on useful estimation.[12] The *Adaptive Statement* organizes retrospective review into a learning plan. The *Decision-Rule Table* makes explicit the normative commitments that shape welfare calculations. The *Review Spine* connects ex ante assumptions to ex post evaluation through scheduled, nonbinding review points. Although the architecture is domain-general, it is especially salient for emerging technologies—artificial intelligence, autonomous systems, advanced biotechnology—where uncertainty is structural and the gap between prediction and reality materializes quickly.

The architecture extends the temporal horizon of benefit–cost analysis. By pairing ex ante modeling with structured ex post evaluation, it allows welfare analysis to be progressively refined rather than fixed at promulgation. Courts retain their traditional role under *State Farm*; the framework organizes the agency's explanation more transparently.

The sections that follow proceed in six stages. Section II establishes the value-of-information logic that supports reallocating analytic effort toward observation. Section III introduces the Three Decision States. Section IV develops the Four Pillars. Section V applies the framework to a hypothetical AI auditing rule for high-risk systems, demonstrating how the architecture operates in an emerging technology domain where uncertainty is structural and feedback cycles are compressed. Sections VI and VII address legal compatibility and institutional feasibility. Section VIII examines precedent in existing federal practice—illustrating both statutory and voluntary adoption of framework-like elements, and identifying where the architecture adds value. The paper concludes by returning to the core proposition: that structured learning, embedded within existing authorities, can enhance both the rigor and legitimacy of federal regulation in a changing technological environment.

II. The Analytical Case for Structured Learning

The gap between prediction and learning described above is not merely institutional—it has a precise analytical consequence. When key regulatory uncertainties are empirical and learnable only through implementation, the question of how to allocate analytic effort between further ex ante modeling and structured ex post observation becomes amenable to value-of-information reasoning.[13]

When the marginal informational value of additional modeling is low relative to the expected informational value of structured observation, reallocating analytic effort toward measurement can improve long-run decision quality. Under such conditions, a sequence of act–observe–revise may approximate expected welfare more closely than a single static optimization performed

[12] Circular A-4, supra note 4, at 3 (directing that analytic emphasis vary with "the sensitivity of the benefit and cost estimates to the key assumptions"). The 2023 revision carried the same directive with updated wording ("key modeling choices," at 4) and expressly adopted Executive Order 12866's "usefully estimated" standard (at 2).
[13] Ronald A. Howard, Information Value Theory, 2 IEEE Transactions on Systems Science and Cybernetics 22 (1966); Jonathan B. Wiener, Disregard and Due Regard, 29 N.Y.U. Envtl. L.J. 437 (2021).

under incomplete information. The welfare objective remains unchanged; the question is how best to pursue it when uncertainty cannot be eliminated ex ante.

This argument supports action under uncertainty, not delay. Waiting for perfect information may itself impose costs or foreclose beneficial outcomes. The claim is procedural: when key uncertainties are learnable only through implementation, the analytic architecture of rulemaking can acknowledge that fact and specify how learning is expected to occur.

III. The Three Decision States: An Epistemic Structure for Rulemaking

Modern rulemaking typically presents itself as a unified analytic exercise. A regulatory preamble often moves fluidly from empirical estimates to projections under uncertainty to statements of policy judgment, all within a single narrative of justification. Costs are quantified, benefits are modeled, uncertainties are described, and tradeoffs are explained. This integrated presentation reflects the complexity of policymaking, but it can also obscure an important structural distinction: not all components of a regulatory decision rest on the same epistemic footing.

Empirical estimation, empirical uncertainty, and normative judgment are analytically distinct activities. They answer different kinds of questions and require different forms of reasoning. When these activities are blended into a single undifferentiated analysis, it becomes difficult to determine what is known, what is conjectural but learnable, and what is ultimately a matter of policy choice. That blending can produce over-precise modeling in some areas, insufficient planning for evidence generation in others, and value judgments embedded implicitly within technical parameters.[14]

To clarify this structure, this paper proposes a new epistemic grammar for regulatory analysis: the Three Decision States.[15] These States classify components of a regulatory decision according to the status of their evidence and the type of reasoning they require. Because the unit of classification is the component, a single parameter may span more than one state: its empirically estimable elements belong to State I or II, while any irreducible value choice it carries belongs to

[14] This dynamic—policy and value judgments embedded in, and presented as, technical or scientific findings, where they escape both public scrutiny and reasoned-explanation review—is the "science charade" identified in Wendy E. Wagner, The Science Charade in Toxic Risk Regulation, 95 Colum. L. Rev. 1613 (1995) (urging that decisionmakers be "forthright in distinguishing policy judgments from scientific facts"); see also Susan E. Dudley & Marcus Peacock, Improving Regulatory Science: A Case Study of the National Ambient Air Quality Standards, 24 Sup. Ct. Econ. Rev. 49 (2018) (documenting the pattern in EPA's NAAQS process).

[15] Frank H. Knight, Risk, Uncertainty, and Profit (Houghton Mifflin 1921); Daniel Ellsberg, Risk, Ambiguity, and the Savage Axioms, 75 Q.J. Econ. 643 (1961). Knight's foundational distinction between measurable risk and unmeasurable uncertainty informs the paper's State I/State II taxonomy. Ellsberg's complementary analysis demonstrates that unmeasurable uncertainty is behaviorally distinct from risk. State III, addressing irreducible normative judgment rather than epistemic uncertainty, lies beyond Knight's and Ellsberg's purely epistemic distinctions, corresponding instead to the valuation dimension of deep uncertainty—disagreement over how to value outcomes (supra note 8).

State III. The task is to decompose a parameter into its components rather than to label the parameter as a whole.

The Three Decision States answer three distinct questions:

What can be reasonably estimated at the time of promulgation?
What cannot be reliably predicted ex ante but can be learned through implementation?
What rests on normative judgment rather than empirical resolution?

The move itself is familiar; what the framework adds here is the routing rule that sends each uncertain component to its state—learnable if better evidence could resolve it, normative if it survives complete information.

Separating these domains improves analytic clarity, strengthens the administrative record, and supports more disciplined learning under uncertainty.

a. State I — What Is Estimable

State I includes empirical components of a regulatory decision that can be reasonably estimated using available data, established methods, and well-understood relationships. These are the elements for which modeling is informative and decision-relevant at the time the rule is issued.

State I does not imply certainty. It reflects adequacy. An estimate belongs in State I when the available evidence allows the agency to characterize direction, magnitude, and plausible ranges with sufficient confidence to inform the decision rule. The question is not whether uncertainty remains, but whether additional refinement would materially change the decision.

Examples of State I components include:

Compliance costs in mature industries with observable data.
Health impacts tied to well-studied exposure–response relationships.
Behavioral responses grounded in stable elasticities or historical patterns.
Administrative burdens that can be estimated from reporting requirements and labor inputs.

In these contexts, benefit–cost analysis performs its core function. Modeling clarifies tradeoffs among alternatives, supports reasoned comparison, and approximates expected welfare at promulgation. Modeling effort should be calibrated to the decision context, stopping when additional refinement would no longer improve the decision, consistent with the principle that estimates should be "usefully estimated" rather than perfected.

State I aligns directly with existing regulatory practice. It covers the portion of a decision grounded in reasonably estimable empirical relationships and preserves the welfare objective and central role of benefit–cost analysis.

b. State II — What Is Uncertain but Learnable

Not all regulatory effects are estimable at promulgation. Some depend on conditions that will only become observable after a rule is implemented. These include effects shaped by dynamic interaction, context-specific behavior, technological evolution, or system feedback.

State II captures this domain of empirical uncertainty. A component belongs in State II when it cannot be reliably predicted ex ante but can be measured through real-world observation after implementation. Disagreements in State II concern empirical questions about how the world behaves and how systems perform under real conditions.

This is not necessarily ignorance. It can also be a recognition of how certain systems behave. In domains characterized by drift, nonstationarity, or emergent behavior, relationships that appear stable in controlled settings may shift once exposed to operational conditions. Performance may vary across populations, environments, or time horizons. New failure modes may appear only in deployment. In such cases, further speculative modeling at promulgation may provide diminishing marginal informational value, while measurement over time can generate more reliable evidence.

Examples of State II components include:

Real-world performance variation of complex systems across demographic or geographic contexts.
Behavioral adaptation in response to new regulatory requirements.
Technological drift following software updates or environmental change.
Incident patterns that emerge only at scale.

State II is defined by learnability. If measurement is feasible after implementation, then the appropriate analytic response is to plan for that measurement rather than to embed assumptions indefinitely within ex ante models. The key move is not to guess more precisely, but to specify what could be observed, how it could be evaluated, and when reconsideration may be warranted.

If additional empirical information could, in principle, resolve the disagreement, the issue belongs in State II; if disagreement persists even under full information, it belongs in State III.

State II extends the commitment to welfare analysis temporally. Where ex ante modeling cannot resolve uncertainty, structured learning may improve the alignment between regulatory action and long-run welfare. The objective remains the same; the pathway incorporates measurement over time.

c. State III — What Is Normative

Some components of a regulatory decision cannot be resolved through empirical evidence, even in principle. These are matters of value judgment or policy priority. They concern what should be prioritized, what level of risk is acceptable, or how competing interests should be balanced.

State III includes such irreducible normative premises and choices. Unlike State II, which concerns empirical uncertainty that can be clarified through measurement, State III concerns judgments that remain even when empirical information is complete. Disagreements in State III concern how agencies should respond—how risks should be weighed, how distributional tradeoffs should be resolved, and how statutory objectives should be prioritized.

State III encompasses a range of normative choices—from premises that are broadly shared and already embedded in standard welfare analysis without controversy, to choices that are genuinely contested across political or ethical perspectives. The procedural response is the same in either case: make the premise explicit instead of embedding it within technical parameters.

Examples include:

Acceptable levels of residual risk.
Distributional tradeoffs across populations.
Relative weighting of competing statutory objectives.
Transparency baselines or precautionary thresholds.

Empirical evidence can inform these judgments. Data can reveal who is affected, by how much, and under what conditions. But evidence alone cannot determine what level of disparity is acceptable, what degree of risk tolerance is appropriate, or how to balance safety against cost in a given statutory context. These choices ultimately reflect policy judgment.

Separating State III from empirical domains clarifies welfare analysis. Benefit–cost analysis necessarily incorporates parameters that combine empirical and normative components—the discount rate, the value of a statistical life, or distributional assumptions—but treating these as purely technical inputs can make value disagreements appear as factual disputes.[16] Even a perfect, universal measurement of willingness to pay would not resolve the value choices at stake: it would reveal what people prefer, but not how those preferences should be aggregated, whose should count, or whether preference satisfaction is the right measure at all—judgments

[16] That such parameters carry a normative dimension alongside their empirical content is long recognized. See F. P. Ramsey, A Mathematical Theory of Saving, 38 Econ. J. 543 (1928) (the pure rate of time preference as an ethical choice); Cass R. Sunstein, Lives, Life-Years, and Willingness to Pay, 104 Colum. L. Rev. 205 (2004) (the value of a statistical life as embedding contestable value judgments).

that survive any amount of evidence.[17] By identifying normative components explicitly, the analytic record becomes more transparent about the basis of decision.

State III therefore emphasizes transparency rather than optimization. It asks the agency to articulate the value commitments that shape the decision, strengthening democratic accountability.[18]

d. Why Separation Matters

The distinction among State I, State II, and State III is not merely a descriptive convenience. It is a structural claim about what regulatory analysis is doing at different points in a decision—and getting the classification right has direct consequences for institutional design. The separation serves four functions.

1. Analytical clarity.
Separating estimable components from learnable uncertainties and normative judgments reduces the risk of false precision. It prevents empirical uncertainty from being masked as stable estimation and reduces the likelihood that value judgments are embedded invisibly within technical parameters. By clarifying what is known, what must be learned, and what must be chosen, the agency can allocate analytic effort more effectively.

2. Legal clarity.

Under the Administrative Procedure Act, agencies must provide a reasoned explanation for their decisions.[19] Clear separation among empirical findings, acknowledged uncertainties, and policy judgments supports this obligation. When the administrative record makes explicit which components are empirical and which are normative, courts reviewing the rule under *State Farm* can more readily evaluate whether the agency has considered relevant factors and articulated a coherent rationale.

3. Institutional clarity.
Within agencies and across interagency review, conflating empirical uncertainty with normative disagreement can generate confusion. Analysts may debate model parameters when the

[17] On the value choices that persist even under complete preference data: on aggregation, see Kenneth J. Arrow, Social Choice and Individual Values (1951); on standing (whose preferences count), see Dale Whittington & Duncan MacRae, Jr., The Issue of Standing in Cost-Benefit Analysis, 5 J. Pol'y Analysis & Mgmt. 665 (1986); and on whether preference satisfaction is the appropriate measure of welfare, see Amartya Sen, Utilitarianism and Welfarism, 76 J. Phil. 463 (1979), and Matthew D. Adler, Well-Being and Fair Distribution: Beyond Cost-Benefit Analysis (2012).
[18] See Eric A. Posner & Cass R. Sunstein, Moral Commitments in Cost-Benefit Analysis, 103 Va. L. Rev. 1359 (2017) (arguing that willingness-to-pay can incorporate normative commitments within BCA). This paper proceeds differently: normative choices function as premises to benefit–cost analysis rather than inputs within it. Monetizing contested premises through WTP relocates rather than resolves the underlying judgment; the Decision-Rule Table makes them transparent without requiring monetization.
[19] 5 U.S.C. § 706(2)(A).

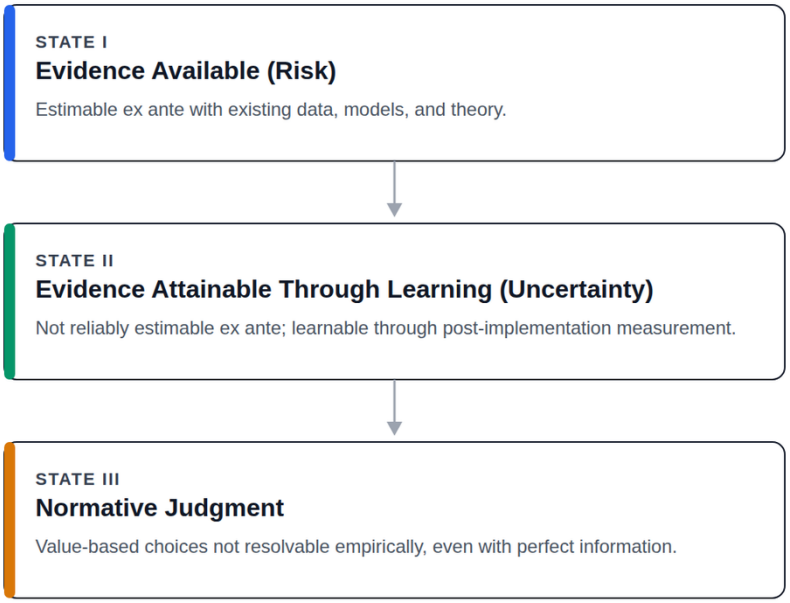
underlying disagreement concerns acceptable risk. Policymakers may appear to challenge empirical findings when they are expressing policy preferences. In interagency review, this conflation can be particularly costly: modeling disputes that are actually disagreements about values produce extended clearance timelines and analytic rework without resolving the underlying question. A shared epistemic grammar—one that identifies whether a given dispute concerns estimation, measurement design, or normative priority—reduces this internal friction and allows deliberation to focus on the actual source of disagreement.

4. Cross-administration durability.
When normative premises are embedded within the empirical analysis rather than identified separately, policy reversal requires reinterpreting the factual record itself—because the values were never distinguishable from the evidence. The result is analytic erosion: the same data reinterpreted to support opposing positions across administrations, with each reinterpretation degrading the evidentiary base that future rulemaking depends on. On this view, embedding the values is the less stable arrangement, not the more stable one—it converts every shift in priorities into a fight over the facts themselves. The three-state separation offers a different pathway. Empirical findings documented in State I and evidence gathered through State II persist across administrations; they are facts, not preferences. Normative premises identified in State III can be openly examined and, where open to agency revision, reassessed when priorities change, without corrupting the evidentiary base. The separation thus enables policy evolution without analytic destruction.

These clarifications address a practical problem: when these domains are conflated, normative disagreements present as empirical disputes, methodological debate becomes a proxy for policy disagreement, and regulatory action stalls. The adequacy standard of Pillar 1 is designed to break this pattern by establishing that sufficient empirical analysis supports promulgation without resolving every contested value premise, allowing decisions to proceed while State III choices are made explicit.

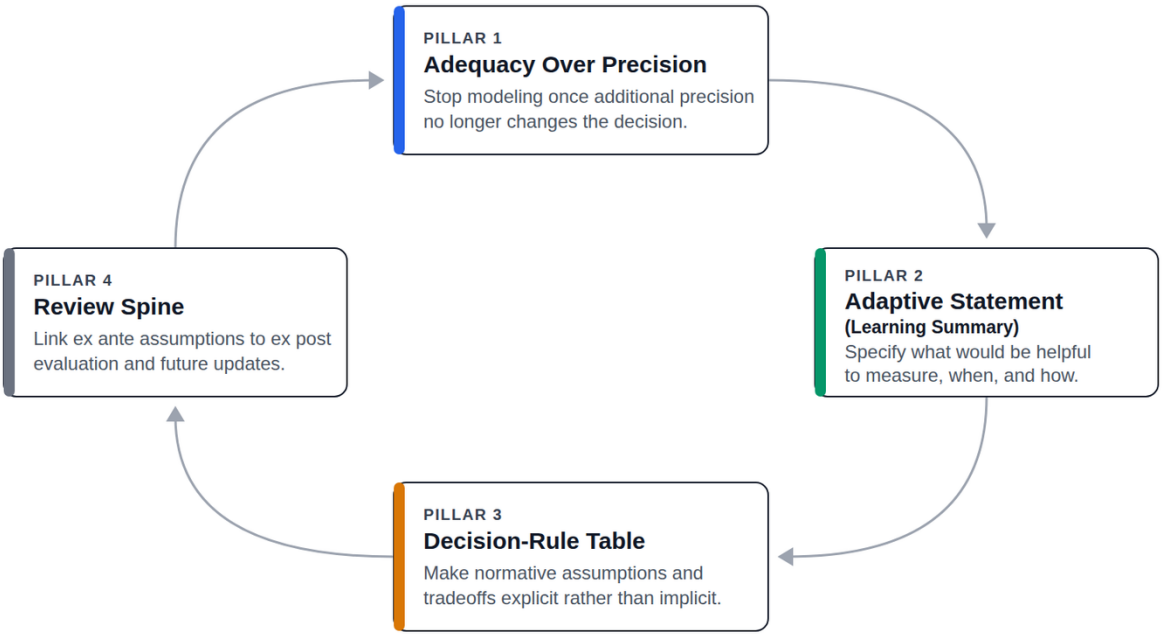
The Three Decision States organize regulatory reasoning without altering its substantive objectives. They describe the status of evidence and the type of reasoning involved in different components of a rule. Figure 1 summarizes the classification. The next section introduces the procedural tools that operationalize this structure.

Figure 1: The Three Decision States



IV. The Four Pillars of Evidence-First Regulation

Figure 2: The Four Pillars of Evidence-First Regulation



The Three Decision States provide an epistemic structure for distinguishing what is estimable, what is learnable, and what is normative within a regulatory decision. The Four Pillars translate that structure into practical rulemaking outputs that fit within existing analytic and legal frameworks. Each Pillar corresponds to a distinct need in the administrative record: disciplined estimation where estimation is feasible; structured learning where impacts are uncertain but measurable; transparency where judgments are normative; and follow-through mechanisms that connect promulgation to evaluation, as can be seen in Figure 2. Critically, the Pillars form a feedback cycle rather than a linear sequence: adequacy at promulgation defines what remains uncertain; the Learning Summary specifies what measurement would be helpful and when; the Decision-Rule Table records the normative premises against which evidence will be evaluated; and the Review Spine connects observed outcomes back to the original analytic logic, informing whether assumptions hold or require reconsideration. All four are designed to operate within current authorities and to preserve agency discretion; their legal conservatism is addressed in Section VI.

a. Pillar 1 — Adequacy Over Precision

Pillar 1 addresses the State I portion of a regulatory decision: components that can be reasonably estimated at the time of promulgation using available evidence, established methods, and stable

empirical relationships. In this domain, modeling remains essential and benefit–cost analysis continues to provide the core discipline for comparing alternatives and approximating expected welfare. The question Pillar 1 addresses is not whether to model, but when additional modeling stops improving the decision.

Adequacy is a stop-rule grounded in marginal informational value. At some point, additional specification—more parameterization, more scenario permutations, more decimal precision—does not materially change the decision rule or reduce uncertainty in a decision-relevant way. This logic is reflected in Executive Order 12866's directive to quantify impacts "to the fullest extent that these can be usefully estimated" and in OMB Circular A-4's caution against creating a "false sense of precision." Pillar 1 operationalizes this principle through a brief adequacy statement in the preamble: a concise explanation of why the agency's ex ante analysis is sufficient for decision-making and why further refinement would not materially alter the decision rule.

Adequacy redirects analytic effort rather than reducing it. Resources not spent pursuing marginal precision in State I estimation are reinvested in Pillars 2 through 4: designing monitoring architecture, articulating decision logic, and establishing review mechanisms. The efficiency case rests on this rebalancing: adequacy shifts analytic investment from diminishing returns in ex ante modeling to high-return investments in learning and accountability.

This rebalancing addresses a specific distortion in current practice. Regulatory impact analyses routinely model compliance costs to a high degree of numerical precision—detailed line items, multiple scenarios, sensitivity tables—while the benefits that motivate the regulation often remain qualitatively described or estimated with wide uncertainty bands. The asymmetry is structural: cost data are more available at promulgation, cost estimates are more auditable during interagency review, and numerical precision on the cost side is easier to defend than equivalent precision on the benefit side. The result is an analytic record that can appear rigorous on costs and thin on benefits, even when the benefits are the primary justification for action. Adequacy corrects this tilt. By capping diminishing-returns modeling where additional precision no longer changes the decision, it frees analytic capacity for the less-developed side of the ledger: measuring benefits, characterizing uncertainty in outcome pathways, and designing the post-promulgation evidence that Pillars 2 through 4 require.

b. Pillar 2 — Adaptive Statement (Learning Summary)

Pillar 2 addresses the State II portion of a regulatory decision: components that cannot be reliably predicted ex ante but can be learned through observation after implementation. When key uncertainties are empirical and learnable, the appropriate response is to identify how evidence could be generated and used, not to embed speculative assumptions indefinitely in ex ante models.

The core output of Pillar 2 is an Adaptive Statement, presented as a concise Learning Summary in the preamble. The Learning Summary identifies the rule's principal State II uncertainties and the agency's plan for learning about them over time. It specifies: what the agency expects to remain uncertain at issuance; what the agency plans to measure; when measurement is expected to occur; what data sources are relevant; and how evidence is expected to inform potential reconsideration. Because the Learning Summary appears in the preamble, the learning agenda is open to notice-and-comment: regulated entities, researchers, and other agencies can see what the agency intends to measure, contest or refine it, and contribute evidence on their own terms—data they already hold or choose to generate, including suitable substitutes for what the agency specified—by submitting it through the comment process or otherwise making it public, without a government-mandated collection.

Two institutional anchors make Pillar 2 administratively conservative. First, the Learning Summary aligns with the Evidence Act's emphasis on evaluation planning: [20] it provides a rule-level bridge connecting uncertainties identified at promulgation to the evaluation planning obligations under § 312, without duplicating statutory requirements. Second, it aligns with retrospective review requirements under Executive Order 12866 by providing, at promulgation, a structured basis for later review—anchoring any reconsideration in the original analytic logic of the rule.

Not all State II uncertainties are equally measurable in practice. Attribution challenges, data limitations, privacy constraints, and long time horizons may limit the precision of post-implementation evaluation, particularly in complex or emerging domains. The framework does not assume that all uncertainties can be resolved cleanly, but rather that agencies can use professional judgment to identify which uncertainties are most decision-relevant and feasibly informed by available evidence over time, in line with the same adequacy principle that bounds ex ante modeling. Where direct measurement is difficult, agencies may rely on proxies, partial indicators, or staged evidence accumulation, while making those limitations explicit in the record.

The Adaptive Statement also has consequences that extend beyond a single rulemaking cycle. In domains characterized by genuine normative disagreement and limited ex ante evidence, policy change across administrations is inevitable—and not necessarily pathological. When successive administrations revisit regulatory choices, they explore different points in a contested policy space. That exploration becomes informative when each iteration inherits evidence generated under the prior regime. A rule promulgated with State II learning infrastructure—specified metrics, data collection pathways, measurable propositions linked to regulatory assumptions—produces evidence that accumulates during implementation regardless of subsequent political

[20] Foundations for Evidence-Based Policymaking Act, supra note 6, § 101(a)(2) (codified at 5 U.S.C. § 312) (requiring agencies to develop learning agendas and annual evaluation plans connecting agency priorities to evidence needs).

direction. The next administration inherits not just a rule but a factual record of what happened under it. Without such infrastructure, reversal is as analytically thin as the original action. By embedding structured learning at promulgation, the Adaptive Statement helps convert policy oscillation from a random walk into a convergent process—one in which each regulatory cycle is more informed than the last.

c. Pillar 3 — Decision-Rule Table

Pillar 3 addresses the State III portion of a regulatory decision: the normative premises that structure regulatory judgment. Benefit–cost analysis necessarily incorporates such premises—discount rates, value-of-statistical-life estimates, distributional assumptions, and safety thresholds all reflect policy judgments about acceptable tradeoffs. These premises are not flaws in welfare reasoning; they are unavoidable features of it. Difficulties arise when they are embedded implicitly within empirical modeling, allowing disagreements about values to appear as disagreements about facts and obscuring the basis for the agency's choice.

The Decision-Rule Table is an explanatory device included in the preamble that organizes normative reasoning in a structured format. In practice, the table identifies: the outcome or risk of concern; the associated cost, burden, or tradeoff; the benchmark, statutory value, or policy concept being applied; and a narrative explanation of how that value concept informs the agency's decision. The purpose is clarification, not quantification. The table does not require numeric thresholds, prescribe welfare weights, or convert normative judgments into enforceable criteria. An agency remains free to state a premise quantitatively; when it does, the table records that judgment as transparency, not as a binding trigger. Surfacing a value premise does not necessarily mean decentralizing it: agencies could apply OMB's standardized default while noting the value choice it embeds—transparency about the premise, not a different number for each rule.

By making value premises visible as discrete analytic inputs rather than embedded assumptions, the Decision-Rule Table serves three practical functions. It reduces ambiguity in the administrative record, enabling stakeholders to identify whether disagreements concern empirical claims or policy values. It supports clearer engagement during notice-and-comment, as commenters can address stated value concepts directly. And it reinforces reasoned explanation under the APA: when normative premises are explicit, the connection between facts found and choices made is more legible to courts reviewing under *State Farm*.

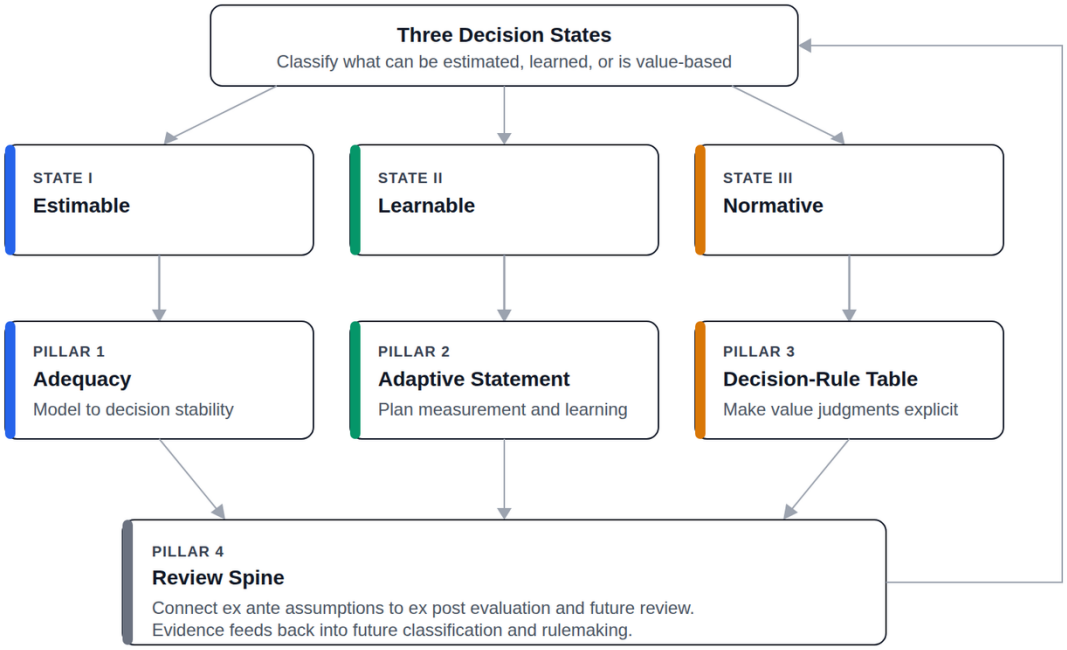
d. Pillar 4 — Review Spine

Pillar 4 provides the connection between ex ante analysis and ex post evaluation. Where Pillar 1 clarifies modeling boundaries and Pillar 2 identifies learnable uncertainties, Pillar 4 organizes how initial assumptions may later be compared with observed outcomes.

The Review Spine has three components. Scheduled review: rules may identify anticipated intervals at which relevant evidence is expected to be compared against assumptions recorded at issuance. These are planning signals, not binding timetables—they help evaluation offices, researchers, and industry anticipate resource needs while preserving agency discretion to reassess earlier, later, or not at all. Condition-based review: rules may identify informational signals—such as sustained divergence between predicted and observed outcomes, clustering of adverse events, or significant technological change—that may warrant internal deliberation. These conditions are analytic markers, not self-executing triggers, and create no enforceable duties. Evaluation memo: this document compares ex ante assumptions with ex post evidence, identifies where predictions aligned with or diverged from outcomes, and records whether initial reasoning remains reasonable. It does not mandate a policy response; it documents the agency's reasoning as it assesses evolving evidence, preserving full discretion over whether and how to act.

Although nonbinding, the Review Spine offers practical institutional benefits. When agencies revisit a rule, the relationship between new evidence and prior reasoning is transparent, reducing the risk of opaque reversal and strengthening the administrative record against arbitrary-and-capricious challenge. Regulated parties can better understand how reconsideration would be reasoned if it occurs. Section VI addresses the full legal framework governing the Review Spine and the other Pillars. Figure 3 shows the entire architecture incorporating the three States and four Pillars.

Figure 3: States → Pillars → Review Spine (Integration)



V. Worked Example — Hypothetical AI Auditing Rule for High-Risk Systems

Scenario Overview

Consider a hypothetical federal rule addressing auditing for high-risk artificial intelligence systems deployed in consequential settings. Covered systems might include hiring algorithms used in employment screening, clinical decision-support tools influencing medical judgments, or safety-critical automated systems operating in industrial or transportation contexts.

Such systems exhibit performance variation across populations, model drift over time, and emergent effects from human-AI interaction—characteristics that traditional ex ante modeling can identify directionally but cannot fully characterize before deployment.

Decision State Classification

The first step under the Evidence-First architecture is to classify elements of the rule according to their evidentiary status.

State I — Estimable Components

Certain elements are reasonably estimable at promulgation:

- Common classes of AI model architectures used in high-risk settings.
- Known categories of potential harms, such as disparate error rates, calibration failures, or unsafe edge-case outputs.
- Feasibility of maintaining audit logs, documentation artifacts, and system metadata.
- Administrative costs associated with reporting, documentation, and internal audit processes.

These components are amenable to traditional regulatory analysis. Modeling can reasonably estimate compliance burdens and characterize known risk categories using available research and technical literature.

State II — Learnable Through Measurement

Other components are not reliably predictable ex ante but are learnable through observation:

- Real-world subgroup performance variation across protected characteristics.
- Drift patterns over time as models are retrained or operating conditions shift.
- Frequency and severity of incident reports or near-miss events.
- Behavioral responses of users interacting with AI outputs.
- Performance changes associated with system updates.

These uncertainties are empirical. They concern how systems perform under real conditions and can be clarified through measurement once the rule is implemented.

State III — Normative Judgments

Certain elements involve policy judgment rather than empirical resolution:

- The acceptable degree of disparity across demographic groups.
- The level of residual risk tolerable in safety-critical settings.
- Baseline transparency expectations for documentation and auditability.
- Values concerning efficiency, fairness, and precaution.

These are normative premises that shape how empirical findings are evaluated. They do not disappear with additional data and therefore belong in State III.

Pillar 1 — Adequacy in Context

Under Pillar 1, the agency explains why its ex ante modeling is sufficient to support promulgation.

The hypothetical rule might explain that existing research and technical evidence establish the plausibility of subgroup performance variation, drift risk, and documentation feasibility. Compliance costs associated with audit logging and reporting can be estimated using available industry data, with sensitivity analysis capturing variation across firm size or deployment scale.

At the same time, the agency may acknowledge that further refinement of predictive modeling would not materially alter the decision to require auditing. Because real-world subgroup performance and drift behavior depend on deployment context, additional modeling would have limited informational value. The adequacy statement would therefore clarify that ex ante analysis supports the directional conclusion that auditing is warranted, and that remaining uncertainty is appropriately addressed through measurement planning rather than further model refinement.

Pillar 2 — Learning Summary in Practice

The Learning Summary would identify the principal State II uncertainties and outline how evidence is expected to be generated.

For the AI auditing rule, the Learning Summary might identify categories of metrics such as:

- Overall and subgroup error rates.
- Calibration consistency across demographic groups.
- Indicators of model drift, such as changes in confidence distributions or performance variance over time.
- Incident reports, including system failures or unexpected outputs.
- Documentation completeness and update histories.

The Summary would describe anticipated data sources, such as system logs, audit reports, incident registries, or periodic compliance submissions. It could note that measurement is expected to occur at defined intervals or in connection with major system updates. It would clarify how observed evidence may inform internal deliberation regarding potential reconsideration.

The Learning Summary does not impose numeric thresholds, prescribe a specific fairness metric or dictate a performance target. It identifies what the agency believes would be helpful to monitor and why those metrics are relevant to the uncertainties identified at promulgation.

It clarifies the agency's understanding of where learning is most valuable without creating enforceable commitments regarding outcomes, or mandating revision in response to particular findings. It functions as a bridge between promulgation and evaluation, consistent with Evidence Act planning principles and retrospective review guidance under Executive Order 12866.

Pillar 3 — Decision-Rule Logic in Practice

The Decision-Rule Table would articulate the normative premises shaping the rule. For the hypothetical AI auditing rule, a simplified illustration of the table might look like this:

Table 1: Illustrative Decision-Rule Table

Outcome / Risk	Tradeoff	Value Concept	Decision Implication
Subgroup error rates 5% higher for protected groups	Higher compliance cost ($5k per firm)	Nondiscrimination	5% disparity warrants continued intervention despite per-firm cost; statutory nondiscrimination obligation outweighs compliance burden at this magnitude
Rare high-impact failures	Slower deployment	Precaution	Lower tolerance for uncertainty in high-risk contexts
Innovation burden on small firms	Reduced competition	Innovation / access	Scaled compliance requirements

The table surfaces the normative premises that structure the agency's decision, placing value judgments in visible relationship with the empirical estimates they act on. Where quantitative findings are available—here, subgroup disparity rates and per-firm compliance costs from State I and State II analysis—they can be included to make the tradeoff concrete. The table's function is

not to impose binding thresholds but to ensure that the normative basis for the agency's choice is transparent.

Pillar 4 — Review Spine Application

The Review Spine organizes how ex ante assumptions and ex post evidence may be compared.

The hypothetical rule could identify anticipated intervals at which accumulated audit data may be assessed in light of the original analytic premises. It could describe conditions—such as sustained divergence between predicted and observed performance, repeated incident clustering, or significant model updates—that may warrant internal deliberation.

These conditions are informational signals, not self-executing triggers—they clarify how the agency may organize its deliberation if evolving evidence warrants reconsideration.

An evaluation memorandum could document comparisons between predicted compliance burdens and observed reporting patterns, between anticipated risk categories and recorded incidents, and between stated normative premises and emerging empirical findings. The memorandum would clarify what has been learned and whether initial reasoning remains reasonable.

The Review Spine functions as a nonbinding documentation framework within existing procedural authority.

Closing the Example

Viewed together, the hypothetical AI auditing rule demonstrates how the Evidence-First architecture operates as a structured learning loop.

1. Ex ante estimation clarifies known risks and compliance burdens.
2. Measurement planning identifies learnable uncertainties.
3. Normative premises are articulated transparently.
4. Observed outcomes are compared with initial assumptions.
5. Reconsideration, if undertaken, proceeds through documented reasoning.

VI. Legal Compatibility and Administrative Conservatism

Clarification, Not Innovation

The Evidence-First framework operates entirely within the Administrative Procedure Act, existing notice-and-comment authority, and executive guidance governing regulatory analysis. The Decision States and Four Pillars are voluntary analytic tools—they organize how agencies explain their reasoning under uncertainty without creating new procedural requirements, conditioning any rule's validity on a particular analytic format, or altering judicial review.

Agencies may adopt them incrementally, with no enforceable duties attached and no procedural consequence for non-adoption.

State Farm Compatibility

Under *Motor Vehicle Manufacturers Association v. State Farm*,[21] agencies must provide a reasoned explanation for their decisions and consider relevant factors within the scope of their statutory authority. The arbitrary-and-capricious standard does not prescribe a particular analytic format; it requires coherence between facts found and choices made.

The Evidence-First architecture supports this requirement by clarifying the structure of reasoning already present in the administrative record. The Decision States separate empirical findings from learnable uncertainties and normative judgments. The Four Pillars organize modeling boundaries, measurement plans, value premises, and review expectations. This structure makes the record more legible, while leaving the standard of review unchanged.

Courts continue to assess whether the agency considered relevant factors and articulated a rational connection between the evidence and its decision. The framework organizes that presentation.

Loper Bright reinforces the value of this separation. Courts exercise independent judgment on statutory meaning, while delegated policy choices remain subject to reasoned decision-making review.[22] The framework does not resolve statutory meaning; it distinguishes evidentiary conclusions from normative values, including those supplied by governing law, without altering the governing standard.

Vermont Yankee and Nonbinding Planning

The framework also operates within the space *Vermont Yankee Nuclear Power Corp. v. NRDC* leaves agencies: courts may not impose procedural requirements beyond those specified in the APA and governing statutes.[23] The Decision States and Pillars are voluntary analytic structures, not procedural mandates—they create no judicially enforceable rights, authorize no court-

[21] *Motor Vehicle Mfrs. Ass'n v. State Farm Mut. Auto. Ins. Co.*, 463 U.S. 29 (1983). The Court held that an agency rule is arbitrary and capricious if the agency has relied on factors Congress did not intend it to consider, failed to consider an important aspect of the problem, or offered an explanation counter to the evidence.

[22] *Loper Bright Enterprises v. Raimondo*, 603 U.S. 369, 395–96, 413 (2024). The Court overruled *Chevron U.S.A., Inc. v. Natural Resources Defense Council*, 467 U.S. 837 (1984), holding that courts must exercise independent judgment in determining whether an agency has acted within its statutory authority, rather than deferring to the agency's interpretation of an ambiguous statute. See also Cong. Rsch. Serv., R48320, Loper Bright Enterprises v. Raimondo and the Future of Agency Interpretations of Law (2024) (noting that exercises of delegated policy discretion remain subject to arbitrary-and-capricious review notwithstanding Loper Bright).

[23] *Vermont Yankee Nuclear Power Corp. v. Nat. Res. Def. Council*, 435 U.S. 519 (1978). The Court unanimously held that reviewing courts may not impose additional procedural requirements on agencies beyond those mandated by the APA or applicable statutes, absent constitutional constraints. See also Emily S. Bremer & Sharon B. Jacobs, Agency Innovation in Vermont Yankee's White Space, 32 J. Land Use & Envtl. L. 523 (2017) (agencies are free to adopt procedures of their own beyond what courts may impose—Vermont Yankee's "white space").

imposed formats or review schedules, and leave unchanged the steps required before promulgation. The Review Spine and Learning Summary are likewise designed as nonbinding planning and documentation tools, not mandatory deadlines. An agency's departure from an anticipated review interval or monitoring plan therefore would not, by itself, amount to a failure to take a legally required discrete action, consistent with the distinction drawn in *Norton v. Southern Utah Wilderness Alliance*.[24] A different question could arise if an agency codified a mandatory deadline or duty. A reviewing court would continue to apply ordinary APA standards, without treating compliance with the framework as an independent legal criterion.

EO 12866 and Circular A-4 Alignment

Executive Order 12866 requires agencies to determine that benefits justify costs and to quantify impacts to the extent they can be "usefully estimated."[25] Circular A-4 cautions against false precision and emphasizes the role of professional judgment in regulatory analysis.

These directives already contain the conceptual foundation for this framework. The "usefully estimated" standard is, in effect, an adequacy principle—it implies that modeling effort should be bounded by informational value and not extended indefinitely. Pillar 1 formalizes this principle as a stop-rule; Pillars 2 and 4 extend the welfare objective temporally by pairing ex ante modeling with structured ex post evaluation; and Pillar 3 surfaces the normative premises that A-4 recognizes as inherent in welfare calculations but does not require agencies to identify explicitly.

The framework leaves the welfare objective embedded in Executive Order 12866 unchanged, organizing how uncertainty and value premises are presented within existing decision criteria rather than providing an alternative to them.

Evidence Act Alignment

The Foundations for Evidence-Based Policymaking Act requires agencies to develop learning agendas and evaluation plans and designates evaluation officers to coordinate analytic capacity. While the Evidence Act establishes evaluation capacity, it does not systematically anchor that capacity in the logic of specific regulatory decisions. In practice, the Act's learning agendas and evaluation plans operate at the level of agency programs and spending rather than individual rulemakings. The Evidence-First framework complements this architecture by embedding evaluation-relevant questions and measurement pathways within the analytic structure of individual rules at promulgation.

---

[24] *Norton v. S. Utah Wilderness All.*, 542 U.S. 55 (2004). The Court distinguished between discrete, mandatory agency duties that courts may compel under APA § 706(1) and programmatic agency conduct not subject to judicial compulsion. See also Mariah Mastrodimos, Note, Self-Imposed Agency Deadlines, 75 Stan. L. Rev. 675 (2023) (nonbinding agency self-commitments are not judicially enforceable).

[25] Exec. Order No. 12866, supra note 3, §§ 1(a), 1(b)(6); Circular A-4, supra note 4.

The Learning Summary provides the connection point. It links rule-specific uncertainties to the evaluation planning contemplated under § 312, clarifying how evidence generated after implementation may relate to assumptions embedded at promulgation. It does not duplicate or expand the Act's requirements—it operates within existing reporting channels and preserves methodological flexibility.

The Decision States and Four Pillars clarify reasoning within existing administrative law rather than altering it. The architecture preserves agency discretion, operates within notice-and-comment procedures, and remains subject to ordinary arbitrary-and-capricious review. It creates no new legal duties, no enforceable procedural rights, and no additional axes of judicial review. Its contribution is organizational: by separating estimable components, learnable uncertainties, and normative premises, and by structuring how those elements are documented and revisited, the framework strengthens defensibility while maintaining a conservative procedural posture.

VII. Incentive Compatibility and Institutional Feasibility

A framework that operates within existing legal authority still requires a theory of adoption. The Evidence-First architecture is designed to align with institutional incentives, not against them. Its feasibility rests on whether key actors within the regulatory process find that it reduces uncertainty, clarifies responsibilities, and improves documentation without expanding legal exposure.

For line analysts and regulatory economists, the framework clarifies the scope of modeling and the treatment of uncertainty. Pillar 1's adequacy principle delineates when additional refinement ceases to affect the decision rule, aligning with Executive Order 12866's emphasis on "usefully estimated" impacts and Circular A-4's guidance against false precision. The Learning Summary distinguishes effects that can be estimated at promulgation from those learnable only through observation, allowing analytic effort to be allocated more efficiently without weakening rigor. Clear documentation of uncertainty reduces exposure to claims that assumptions were concealed—and because assumptions are recorded against measurable benchmarks, the same documentation makes retrospective review more tractable. Analysts build the evidentiary record at promulgation that future evaluation will draw on.

For agency leadership, the framework's principal value is defensive. A structured articulation of assumptions, uncertainties, and value premises strengthens the administrative record that courts review under arbitrary-and-capricious standards—reducing litigation vulnerability without creating binding commitments that constrain future discretion. The Review Spine and Learning Summary function as planning tools, not enforceable obligations. The framework also reduces the cost of acting under uncertainty: because the Adaptive Statement plans for learning where predictions are provisional, leadership need not treat the promulgation-stage analysis as a final answer—they can move forward with adequate evidence and a credible plan to improve it, rather than delaying action in pursuit of an unattainable precision. And by surfacing normative

premises as distinct from empirical findings, the record gives future leadership—including political appointees—a clearer view of where prior value judgments shaped the rule, enabling deliberate policy evolution without relitigating the empirical base.

For OIRA and interagency reviewers, the adequacy principle may reduce modeling sprawl and encourage focus on decision-relevant parameters. Because the framework preserves the welfare objective and introduces no new analytic tests, it operates within the existing review structure. For evaluation offices established under the Evidence Act, the Learning Summary connects rule-level uncertainties to potential evidence generation pathways, providing practical input for § 312 evaluation planning without creating new reporting mandates.

These incentive effects have a common structural explanation. Public choice scholarship has long demonstrated that regulatory processes are shaped by the incentives facing participants, not by assumptions about good intentions.[26] Opaque analytic environments reward strategic behavior: agencies can embed value judgments inside technical parameters, reviewers can demand additional modeling to delay action, and stakeholders can challenge empirical findings as a proxy for policy disagreement.

The Evidence-First architecture addresses each of these dynamics by increasing transparency as opposed to relying on good faith. Adequacy limits the returns to over-modeling by capping analysis at marginal informational value. State III surfaces normative premises that would otherwise hide within technical assumptions, reducing the space for strategic embedding of preferences. The Learning Summary and Review Spine create a forward-looking evidence record that increases the long-term cost of shading assumptions, because documented predictions will eventually be compared against observed outcomes.

For the same reason, the framework resists hollow adoption. Where analysis stops—and where it never starts—is otherwise implicit in the record; the framework makes both explicit and reviewable: an adequacy statement subject to comment and to *State Farm* scrutiny, and a Learning Summary visible at promulgation and measurable against what follows. Because hollow adoption is detectable, sincere adoption is credible, and credibility is what gives the record its defensive weight. That discipline runs on scrutiny, not on the route of adoption, whether voluntary or required by statute or executive order; and scrutiny likewise predicts where adoption will concentrate: where it meets uncertainty, in domains that demand more certainty

[26] William A. Niskanen, Bureaucracy and Representative Government (1971); George J. Stigler, The Theory of Economic Regulation, 2 Bell J. Econ. & Mgmt. Sci. 3 (1971). For adoption mechanisms in the adaptive-regulation literature that rely on designed statutory structure, see Karkkainen, supra note 9 (penalty defaults inducing negotiated adaptive arrangements); Lori S. Bennear & Jonathan B. Wiener, Periodic Review of Agency Regulation, Report to the Admin. Conf. of the U.S. (June 7, 2021) (finding periodic review more likely when specifically required by statute). The account here locates the adoption incentive in review scrutiny that already exists.

than the evidence can supply and where pretended precision no longer holds. The framework is hardest to fake precisely where it is most likely to be used.

Because the framework operates through documentation and analytic structure and not procedural mandates, adoption can occur incrementally—rule by rule, scaling to agency capacity. The preceding analysis suggests that this incremental path aligns with institutional incentives: transparency, once structured, becomes the path of least resistance for the actors who must implement it. The following section examines whether this is more than conjecture: whether elements of this architecture have already emerged independently in federal practice, adopted by agencies facing precisely the uncertainty-dominant conditions the framework is designed to address.

VIII. Precedent in Practice

The framework proposed in the preceding sections might appear to require novel analytic practices. This section examines whether elements of the architecture already exist in federal rulemaking. The investigation is preliminary: characterizing another agency's rule from the outside involves interpretive judgment, and the classifications offered here are intended as a starting point for further analysis and not a definitive assessment.[27] That said, the pattern that emerges is suggestive. Elements of all four pillars appear across existing practice, though no single rule implements them together as a coherent architecture.

Table 2: Precedent in Practice: Agency Elements by Pillar

Agency / Rule	P1: Adequacy	P2: Adaptive Statement	P3: Decision-Rule Table	P4: Review Spine	CC: Cross-cutting	Basis
EPA NAAQS	✓	(✓)	(✓)	✓	(✓)	S
FWS Polar Bear ITR	✓	✓	✓	✓	(✓)	S
FDA Benefit-Risk	✓	✓	✓	✓	✓	S/V*
SEC Reg SHO Pilot	✓	✓	(✓)	✓	✓	V
NHTSA CAFE Midterm	✓	✓	✓	✓	✓	S/V*
Interagency CRA Final Rule (2023)	(✓)	(✓)	✓	—	(✓)	S/V

*S = Statutory mandate; V = Voluntary design; S/V = Statutory authority, voluntary design; ✓ = element present (not necessarily full adoption); (✓) = partial or implicit evidence; — = explicitly absent. *FDA's framework reflects a negotiated statutory commitment (FDASIA § 905); NHTSA's basis is split (EPA's MTE was regulatory; NHTSA's participation was discretionary). No single rule implements all four pillars as an integrated architecture. The classifications are preliminary and interpretive.*

EPA's NAAQS process may be the most mature and familiar of these architectures: the Clean Air Act's five-year review cycle (Pillar 4), the Integrated Science Assessment's causality

[27] Prior empirical work has examined particular mechanisms in depth. See, e.g., Bennear & Wiener, Periodic Review, supra note 26 (assembling a database of federal periodic-review practice across agencies). The question here is different: whether the elements of the full architecture co-occur within individual rules. See also Joseph E. Aldy, Learning from Experience: An Assessment of the Retrospective Reviews of Agency Rules and the Evidence for Improving the Design and Implementation of Regulatory Policy, Report to the Admin. Conference of the United States (Nov. 17, 2014).

grading, and the Administrator's documented reasoning give it strong adequacy, learning, and review elements. Bennear and Wiener identify EPA's periodic NAAQS review as among the periodic-review programs "integrated most successfully into agency culture and process."[28] Its clearest gap may be Pillar 3—as Dudley and Peacock document,[29] the normative judgments remain intertwined with the scientific record rather than surfaced as a separable premise, the separability the Decision-Rule Table is meant to supply. Similar statutory architectures appear in other domains: FWS operates on five-year incidental take regulation (ITR) cycles under the Marine Mammal Protection Act;[30] FDA's benefit-risk framework operates within Prescription Drug User Fee Act (PDUFA) reauthorization cycles; [31] NHTSA voluntarily joined EPA's midterm evaluation of the coordinated Corporate Average Fuel Economy (CAFE)/greenhouse gas (GHG) standards.[32] Table 2 includes additional examples, such as the banking agencies' 2023 Community Reinvestment Act (CRA) final rule, which adopted benchmark-based performance ranges, fixed test weights, and standardized performance scores for CRA evaluations.[33] Together, these examples show that agencies can construct—and, in several

[28] Lori S. Bennear & Jonathan B. Wiener, Pursuing Periodic Review of Agency Regulation, The Regulatory Review (Nov. 9, 2021), https://www.theregreview.org/2021/11/09/bennear-wiener-periodic-review/ (summarizing Periodic Review, supra note 26).

[29] Dudley & Peacock, supra note 14. Dudley and Peacock document EPA's NAAQS process as a mature regulatory learning system developed through iterative Clean Air Act implementation and D.C. Circuit review. See, e.g., *Lead Indus. Ass'n v. EPA*, 647 F.2d 1130 (D.C. Cir. 1980) (upholding EPA's standard-setting methodology while establishing expectations for structured scientific assessment); *Am. Trucking Ass'ns v. EPA*, 175 F.3d 1027 (D.C. Cir. 1999), rev'd on other grounds sub nom. *Whitman v. Am. Trucking Ass'ns*, 531 U.S. 457 (2001).

[30] U.S. Fish & Wildlife Serv., Marine Mammals; Incidental Take During Specified Activities; North Slope, Alaska, 86 Fed. Reg. 42,982 (Aug. 5, 2021) (final rule); Letter of Authorization (LOA) regulations at 50 C.F.R. Part 18. FWS established an adaptive management structure for polar bear incidental take, accepting directional estimates rather than precise predictions, with monitoring requirements including 48-hour incident reports, den surveys via aerial infrared, and annual LOA evaluations.

[31] FDA, Structured Approach to Benefit-Risk Assessment in Drug Regulatory Decision-Making, PDUFA V Plan (draft, Feb. 2013). Developed under a negotiated PDUFA commitment codified in FDA Safety and Innovation Act (FDASIA) § 905, the framework grades evidence quality through structured qualitative assessment rather than requiring specific quantitative thresholds. It also includes adaptive elements (post-approval updating, Change Control Board governance) and a staged implementation timetable with scheduled evaluation, operating within PDUFA reauthorization cycles as a meta-review spine.

[32] Light-Duty Vehicle Greenhouse Gas Emission Standards and Corporate Average Fuel Economy Standards for Model Years 2017–2025, Draft Technical Assessment Report, EPA-420-D-16-900 (July 2016). EPA voluntarily exercised its discretion to codify a midterm evaluation in the 2012 rule; once codified, 40 C.F.R. § 86.1818-12(h), promulgated at 77 Fed. Reg. 62,624 (Oct. 15, 2012), required EPA to perform it. The formal sequence comprised a Draft Technical Assessment Report issued jointly by EPA, NHTSA, and the California Air Resources Board; an EPA Proposed Determination with public comment; and an EPA Final Determination. NHTSA voluntarily participated and used the shared record in its separate, statutorily required de novo CAFE rulemaking.

[33] Community Reinvestment Act, 89 Fed. Reg. 6,574, 6,792, 6,958, 7,026–27 (Feb. 1, 2024) (joint OCC, Federal Reserve, and FDIC final rule). The rule adopted a metrics-based CRA evaluation framework: performance ranges incorporating market and community benchmarks to set thresholds for conclusion categories; fixed performance-test weights of 40 percent for the Retail Lending Test, 10 percent for the Retail Services and Products Test, 40 percent for the Community Development Financing Test, and 10 percent for the Community Development Services Test; and standardized performance scores aggregated to state, multistate-MSA, and institution levels. The agencies deferred certain thresholds—notably for the Community Development Financing Test—citing data limitations. The rule was enjoined before its effective date and never applied to any bank; the agencies proposed its rescission in 2025, and the pre-2023 regulations remain in effect. See Community Reinvestment Act Regulations, 90 Fed. Reg.

settings, operate—these components. But most required decades of domain-specific litigation and legislative refinement, unique statutory mandates, or both. The question for the framework is whether these elements can be adopted voluntarily, within existing notice-and-comment authority, as a matter of analytic design and not statutory compulsion.

The SEC's Regulation SHO pilot[34] suggests they can. In 2004, the Commission faced a question it could not answer with existing data: whether short sale price test restrictions, in place since 1938, remained necessary for market integrity. Rather than modeling the effects of removal or deferring indefinitely, the Commission chose a structured learning approach. It suspended the tick test for approximately one-third of Russell 3000 stocks in a stratified pilot (stocks sorted by listing market and average daily dollar volume; every third stock designated as pilot security), pre-specified the variables to study (market volatility, price efficiency, and liquidity), designated the Office of Economic Analysis to gather and analyze data, arranged for pilot data to be made publicly available, and invited external researchers to provide independent analyses. The pilot was temporary by design and expired in August 2007. None of this was required by statute. The Securities Exchange Act gave the Commission broad rulemaking authority, but no provision mandated a pilot, specified variables, or required public data. The Commission adopted these elements as a matter of regulatory design.

Mapped to the framework's vocabulary, the pilot reflects elements of three pillars. The Commission explicitly acknowledged that existing data was insufficient to support a permanent decision—"[w]ithout empirical data relating to short selling in the absence of a price test in today's market, we believe that only broad conclusions could be derived"—and chose to act on adequate rather than precise information (Pillar 1). It identified specific variables, designated an analytical unit, and created a public evidence pathway (Pillar 2). And it built in a fixed expiration that created a concrete endpoint for reconsideration, as opposed to allowing the status quo to persist indefinitely (Pillar 4). The pilot delivered: OEA and four outside studies analyzed the data, and in 2007, after considering the pilot results, the Commission eliminated the federal tick

34,086, 34,088, 34,090 (July 18, 2025) (proposed rescission); 91 Fed. Reg. 509 (Jan. 7, 2026). The design pattern—establishing quantitative decision-rule criteria voluntarily—remains illustrative. See also Community Reinvestment Act, 87 Fed. Reg. 33,884 (June 3, 2022) (joint notice of proposed rulemaking); Community Reinvestment Act Regulations, 85 Fed. Reg. 34,734 (June 5, 2020) (earlier OCC-only final rule, rescinded December 2021).

[34] Short Sales, Exchange Act Release No. 50,103, 69 Fed. Reg. 48,008, 48,024 (Aug. 6, 2004); Order Suspending the Operation of Short Sale Price Provisions for Designated Securities and Time Periods, Exchange Act Release No. 50,104, 69 Fed. Reg. 48,032 (Aug. 6, 2004) (designating the pilot securities); Order Extending Term of Short Sale Pilot, Exchange Act Release No. 53,684, 71 Fed. Reg. 24,765 (Apr. 26, 2006) (extending the pilot—originally one year—through August 6, 2007, the expiration date of Rule 202T, the temporary Regulation SHO provision under which the Commission suspended price tests to study their effectiveness; SEC staff arranged for monthly public release of pilot data). The pilot suspended Rule 10a-1's tick-test requirements for the designated securities, identifying market microstructure variables (volatility, liquidity, price efficiency) to study before determining permanent rule changes. It named what to measure but did not pre-specify decision thresholds mapping observed values to regulatory outcomes.

test and barred exchanges and other market self-regulators from having short-sale price tests of their own.[35]

What the Commission did not do is equally instructive. The rule listed three possible post-pilot outcomes—eliminate the price test, adopt a uniform bid test, or retain current tests—but specified no criteria for choosing among them. What level of volatility increase would be tolerable? What liquidity improvement would justify permanent removal? The pilot named what to study but not how findings would map to regulatory action. This is the Pillar 3 gap: a structured learning exercise without a corresponding decision-rule table connecting evidence to outcomes. The framework proposed here would fill that gap—not by constraining the Commission's judgment, but by making the normative premises behind each alternative explicit at promulgation.

The Regulation SHO pilot demonstrates that an agency can voluntarily adopt elements of this architecture—acknowledging uncertainty, specifying learning variables, creating a public evidence pathway, and building in a review endpoint—within notice-and-comment rulemaking as it exists today. It also illustrates where the framework adds value: the Commission built strong elements of three pillars but, without a decision-rule table, the connection between evidence and action remained implicit. As Table 2 above suggests, similar patterns appear across the other rules surveyed. The characterizations are necessarily interpretive, and a deeper treatment would require rule-by-rule engagement with the administrative record.

IX. Conclusion — From Static Optimization to Structured Learning

Benefit–cost analysis was designed to discipline regulatory judgment under uncertainty—to ensure that agencies weigh evidence, consider alternatives, and articulate reasoned explanations for their choices. Anchored in Executive Order 12866 and the Administrative Procedure Act, this analytic architecture has provided a durable foundation for regulatory decision-making. Agencies model expected costs and benefits with increasing sophistication, and the analytic expectations of the rulemaking process have grown accordingly.

When regulatory impacts are reasonably estimable and key value premises broadly shared, this architecture works well—agencies gather evidence, model behavioral responses, and test sensitivity assumptions. But an increasing number of regulatory domains are characterized by conditions that exceed what ex ante analysis alone can organize: assumptions that are provisional rather than stable, effects that are context-sensitive, and feedback that arrives unevenly over

[35] Office of Economic Analysis, U.S. Securities and Exchange Commission, Economic Analysis of the Short Sale Price Restrictions Under the Regulation SHO Pilot (Feb. 6, 2007) (staff analysis of the pilot data; the Commission expressed no view regarding the staff analysis); Regulation SHO and Rule 10a-1, Exchange Act Release No. 55,970, 72 Fed. Reg. 36,348, 36,349, 36,359 (July 3, 2007) (noting four completed outside studies and that "[g]enerally, the Pilot Results supported removal of current price test restrictions"). The 2007 rule removed Rule 10a-1—the 1938 tick-test rule—and prohibited any self-regulatory organization, such as an exchange, from maintaining its own short sale price test.

implementation. In these uncertainty-dominant domains, the comparison of costs and benefits that BCA structures at promulgation can only be approximated at issuance. In addition, some evidence needed to improve those approximations may be too costly to gather at the time of rulemaking but feasible to collect eventually during implementation—yet the current analytic architecture provides no systematic mechanism for planning that learning at the point of promulgation or connecting the uncertainties identified ex ante to post-implementation measurement.

The Evidence-First framework provides that mechanism. It does not replace benefit–cost analysis; it puts that analysis to work on what can be estimated and organizes the rest. At promulgation, a rule does four things: it holds modeling to the point where added precision no longer changes the decision; it names the uncertainties whose resolution would most improve the next decision, for agencies, industry, or researchers to help fill as evidence allows; it states the value premises openly rather than burying them in technical parameters; and it marks in advance when revisiting the rule would be worthwhile.

What those things yield reaches beyond any single rulemaking. As implementation generates evidence, the review connects it back to the assumptions it tested, so the rule improves rather than freezes—closing the loop the regulatory process has long called for but rarely completes. Because the value premises were separated from the empirical record, a change of administration can revise values left to agency judgment without discarding what has been learned, and policy evolves without analytic destruction (Section III). Rules built this way convert policy oscillation from a random walk into a convergent process, each cycle more informed than the last (Section IV). Otherwise, the costs fall on regulated entities that invest in compliance under analytically thin rules, on agencies whose credibility erodes with each ungrounded reversal, and on the public when policy swings outpace measurement.

Emerging technologies increase the visibility of uncertainty by compressing feedback cycles and exposing assumptions more quickly. They do not create uncertainty, but they make its consequences more immediate. Uncertainty-dominant regulatory domains are multiplying—artificial intelligence, emerging contaminants, climate adaptation, algorithmic lending, autonomous systems—and they exhibit precisely the conditions under which agencies have historically developed framework-like elements: visible uncertainty that forecloses pretended precision, heightened litigation exposure, and iterative regulatory structures that reward investment in learning infrastructure.

The paper's contribution operates on three levels: the Three Decision States provide a diagnostic grammar for classifying regulatory components by evidentiary status (Section III); the Four Pillars provide an implementation architecture that translates that classification into adoptable procedural tools (Section IV); and the paper argues that this architecture is institutionally viable. On viability, the evidence is threefold: the framework operates within existing notice-and-

comment authority without expanding judicial review (Section VI); it aligns with the incentives of analysts, leadership, and reviewers (Section VII); and it finds preliminary support in the independent convergence of agencies already facing uncertainty-dominant conditions (Section VIII). The characterizations offered in Section VIII are interpretive and may benefit from deeper engagement with individual administrative records. But if the pattern is even approximately right, it suggests that agencies have been converging on similar structural responses to uncertainty—and the contribution of this paper is to name that pattern, make it portable, and lower the cost of adoption.

What emerges is not a replacement for benefit–cost analysis but a broader analytic architecture within which it can function more effectively. The framework improves promulgation-stage analysis itself. Adequacy grounds modeling in its marginal informational value; and making normative premises explicit rather than embedded clarifies what BCA is actually measuring and reduces the methodological proxy battles that stall action. And it extends that analysis over time, embedding the learning infrastructure that allows initial approximations to be tested and improved as implementation generates evidence. The architecture requires no new authority, meets agencies where they are, and can begin today—reorganizing the tools agencies already possess so they function more coherently for governance under deep uncertainty.