

Article

The Architecture Premium: A Real Options Framework for Multi-Tier Research Data Infrastructure Systems

Sung-Seok Ko ¹, Bangwon Ko ² and Yong Hee Han ^{3,*}

¹ Department of Industrial Engineering, Konkuk University, 120 Neungdong-ro, Gwangjin-gu, Seoul 05029, Republic of Korea; ssko@konkuk.ac.kr

² Department of Statistics and Actuarial Science, Soongsil University, 369 Sangdo-ro, Dongjak-gu, Seoul 06978, Republic of Korea; bko@ssu.ac.kr

³ Department of Entrepreneurship and Small Business, Soongsil University, 369 Sangdo-ro, Dongjak-gu, Seoul 06978, Republic of Korea

* Correspondence: amade@ssu.ac.kr

Highlights

Please indicate how your work links to systems science via your contributions to systems practice, theory, and/or methodology.

- Research data infrastructure is modeled as a three-tier system whose architecture endogenously lowers exercise friction and thereby raises real option value, offering a systems perspective on digital infrastructures.
- The framework couples option pricing with explicit friction and environmental penalty functions, offering a systems methodology for evaluating national research data infrastructures.

What are the main findings and/or the implications of the main findings?

- Advancing data through the three-tier architecture lowers exercise friction, which creates an Architecture Premium and, once friction falls deep below the intrinsic asset value, sharply compresses the sensitivity of realizable scientific value to volatility; both are standard option properties applied to an architecture-dependent strike.
- Tier-3 AI sandboxes require both sufficient cross-disciplinary reuse and satisfaction of a total-cost boundary in which annualized upgrade cost reduces the admissible environmental penalty one-for-one.

Abstract

This paper proposes the Architecture-Aware Real Options Framework (AA-ROF), a systems-based model for valuing research data in multi-tier socio-technical infrastructures (Institutional Archive, Federated Mesh, AI Sandbox). AA-ROF treats architecture as an endogenous generator of exercise friction and defines the “Architecture Premium” as the gain in realizable option value; two of its three named effects are standard option properties applied to this endogenous strike, and the novelty lies in coupling that strike to the environmental penalty of compute-intensive architecture. The analysis yields two conditional implications: low friction can sharply compress value sensitivity to scientific uncertainty, and Tier-3 infrastructure can offset redundant-data-creation losses once cross-disciplinary reuse exceeds an illustrative threshold. Its central result is a symbolic total-cost boundary: the admissible Tier-3 environmental penalty equals the Tier-2 penalty plus the option-value gain minus annualized capital cost. Thus, the reported value of 60 is only a zero-capital-cost illustration, not a general adoption threshold; tier parameters follow literature ratios, not inverted from the model. A two-period extension reports opposing conditional feedback



Academic Editors: Raquel Ibar-Alonso, Maria-Carmen García-Centeno and Raquel Quiroga-García

Received: 10 August 2026

Revised: 16 September 2026

Accepted: 22 September 2026

Published: 24 September 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

effects rather than an unconditional tier preference. AA-ROF thus offers a systems-level baseline for governing environmental burden and realizable option value in research data infrastructure.

Keywords: research data valuation; real options framework; multi-tier architecture; socio-technical systems; systems methodology; global R&D governance; AI foundation models

1. Introduction

In the era of Data-Driven Science and Artificial Intelligence for Science (AI4S), the exponential generation of research data has exposed a structural inefficiency in the systems that store and mobilize it. While data is universally acknowledged as a paramount intangible asset, a significant portion of global scientific data devolves into “dark data”, stored in isolated silos and rarely reused. This linear lifecycle of data incurs substantial environmental and financial costs through redundant experiments and unoptimized data storage, imposing a system-level deadweight loss on national R&D ecosystems. To foster a true *Circular Economy of Research Data*, global infrastructure initiatives such as the National Artificial Intelligence Research Resource (NAIRR) in the United States, the European Open Science Cloud (EOSC) and Gaia-X in Europe, and sovereign high-performance computing (HPC) projects publicly describe multi-tier digital infrastructures that facilitate continuous data reusability.

However, the economic justification for investing in such advanced, multi-tier infrastructures remains analytically constrained. The existing Information Systems (IS) literature evaluates digital infrastructure primarily through the lens of operational efficiency and cost reduction. Traditional valuation methods, such as Net Present Value (NPV), measure infrastructure investments by how much they reduce the deterministic costs of data storage or network transfer. This cost-centric paradigm often misinterprets the economic nature of data. Advanced distributed architectures do not merely reduce the cost of conducting *existing* research; they reduce the friction that keeps disparate datasets from being merged with AI foundation models, rendering previously impractical combinations exercisable and thereby enabling researchers to answer novel scientific questions.

Simultaneously, the global push toward highly integrated AI infrastructures sharpens the exchange between such a system and its physical environment. Modern AI4S workloads and foundation model training demand immense power, massive water consumption, and extensive cooling resources, resulting in escalating carbon footprints that increasingly challenge national carbon-neutrality targets. As highlighted by international energy projections, global data center electricity consumption is set to more than double by 2030, with renewables expected to meet only about half of this additional demand while many electricity grids are already under strain [1]. Therefore, science policymakers face a complex and urgent dilemma: How can public R&D infrastructure policies balance the economic and scientific necessity of reducing the friction that keeps existing research data from being exercised against the energy and carbon burden that the same architectural choices impose?

To address this, we propose the Architecture-Aware Real Options Framework (AA-ROF). We move beyond the paradigm of merely applying existing financial formulas to empirical data. Instead, we establish a formal deductive analytical framework where the infrastructure architecture shapes economic friction. We demonstrate that transitioning data through a proposed three-tier distributed architecture (Archive, Federated, AI Sandbox) can act as a systematic driver of real option value rather than merely a cost-saving mea-

sure. In doing so, we adopt an explicitly systems-theoretic perspective: the research data ecosystem is modeled as a multi-tier system whose economic behavior emerges from the interaction of accessibility, governance readiness, and computing proximity. That ecosystem is socio-technical in the strict sense, since governance readiness is an institutional and organizational property whereas accessibility and computing proximity are technical ones, and the realizable value of the whole is set by how these heterogeneous components jointly enter the single priced friction quantity of Equation (2) rather than by any component in isolation. The framework is therefore offered as a reusable systems methodology for infrastructure policy analysis, linking real options economics to the systems-science tradition of studying structure-driven behavior.

The primary contributions of this paper are three structural properties that follow from treating research data infrastructure as a multi-tier system rather than as a static asset store. First, we provide an analytical proposition that architectural progression generates an “Architecture Premium,” a friction-reduction premium whose economic significance is that previously prohibitive combinations of datasets and AI models become exercisable, rather than a distinct combinatorial-value mechanism. Second, we formalize a conditional “Volatility Compression” effect, indicating that low-friction infrastructure sharply reduces, without eliminating, the sensitivity of realizable value to scientific uncertainty. Third, we introduce the “Sustainability Premium,” showing analytically under stated assumptions that advanced tier infrastructures can mitigate the severe deadweight loss associated with redundant data creation when reuse conditions are met. What is new is not the option-pricing machinery itself but the specific functional coupling in Equations (2) and (3) that maps architectural state jointly onto exercise friction and environmental penalty. We emphasize that Principles 1 and 2 below are, mathematically, disciplined applications of standard option-pricing properties (strike monotonicity and deep-in-the-money Vega compression) composed with that coupling; the theoretical contribution lies in the systematic mapping of architectural states into that calculus, and in the policy vocabulary this mapping yields, rather than in new option-pricing theorems.

Given that sovereign-scale Tier-3 AI Sandboxes are currently in early developmental phases globally, empirical ex-post evaluation remains structurally infeasible. Consequently, this study is explicitly positioned as an *ex-ante analytical framework* rather than a retrospective empirical validation. By establishing the theoretical boundary conditions and structural mechanisms of data infrastructure value, this paper provides a deductive baseline for future empirical calibration and infrastructure design. The remainder of this paper is structured as follows. Section 2 positions this contribution against prior work in data valuation, global R&D frameworks, and digital infrastructure. Section 3 develops the formal analytical framework, detailing the shift from cost reduction to option-based valuation and embedding governance readiness. Section 4 conducts a numerical boundary-condition analysis and policy scenario evaluation. Section 5 discusses theoretical contributions, public sector evaluation implications, and empirical calibration strategies, and Section 6 concludes.

2. Literature Review and Theoretical Positioning

To elucidate the theoretical necessity of the AA-ROF framework, we synthesize and critique recent advancements across distinct streams of peer-reviewed literature. Crucially, we focus on how the infrastructure architecture and research data management (RDM) status, rather than abstract policy declarations, dictate the realizable value of data within our proposed three-tier distributed model.

From a systems-science perspective, the relevant unit of analysis is not an isolated dataset or computing component but a bounded configuration of interdependent elements whose properties depend on relations among parts and on the level of abstraction at which

the system is described [2,3]. We use this perspective narrowly: it motivates evaluating data, infrastructure, and RDM arrangements jointly, but it does not by itself validate the particular three-tier taxonomy or the functional forms used in AA-ROF.

2.1. Data Valuation and Infonomics in the AI Era

The burgeoning field of Infonomics [4] established the premise that data must be treated as a formal economic asset. Recently, this has evolved into the precise quantification of data quality metrics, particularly through the lens of the FAIR (Findable, Accessible, Interoperable, and Reusable) principles [5]. Contemporary studies attempt to assign financial value based on these intrinsic data readiness metrics [6].

However, this “intrinsic determinism” assumes data possesses realizable value regardless of its infrastructural environment. In the AI4S era, a perfectly curated dataset is economically inert if trapped in a disconnected cold archive. The systematic bibliometric review by Qin et al. [7] shows that value is progressively co-created as data moves through assetization, commoditization, and capitalization stages. Crucially, in our proposed three-tier model, this realizable value depends fundamentally on the underlying infrastructure and the actual status of research data management rather than merely on policy initiatives or intrinsic data quality. Recent peer-reviewed work supports this view: Majithia et al. [8] and Greenberg et al. [9] show that FAIR compliance is necessary but not sufficient, as readiness is heavily mediated by an infrastructural metadata ecosystem that must be machine-actionable.

2.2. Real Options Theory in Environmental and IT Investments

Real options theory adapts financial option pricing to evaluate strategic investments under high uncertainty [10]. While prior IS literature primarily treats the IT infrastructure itself as the “underlying asset” [11], our three-tier framework inverts this paradigm: the *research data* is the underlying asset, and the *infrastructure architecture* is the endogenous mechanism that alters the “exercise cost” of that option.

Furthermore, the conceptual linkage between environmental costs, irreversibility, and option value traces its roots to classical environmental economics literature [12,13]. While these foundational works addressed the irreversible loss of natural resources, this paper extends the irreversibility concept to the modern context of carbon-intensive computing infrastructures and the deadweight loss of redundant data creation. Most directly relevant to data valuation is Coyle and Gamberi [14], who apply the Black–Scholes framework to value commercial data assets based on empirical exercise costs. Where the existing literature asks what the exercise cost was for a specific use case, AA-ROF endogenizes the exercise cost as a continuous function of the proposed three-tier architectural movement, extending classical environmental option value into a systems-level evaluation methodology for digital infrastructures.

2.3. Digital Infrastructure, Data Gravity, and Research Data Management

Information infrastructure is not merely a technical artifact but a relational property that dictates how scientific work is accomplished [15]. Classical socio-technical systems research likewise showed that work-system outcomes depend on the relation between social organization and technological content rather than on technology alone [16]. In the present setting, this insight motivates treating RDM readiness and infrastructure architecture as coupled conditions; it does not imply that evidence from industrial work systems directly establishes effects in research-data infrastructures. This is acutely relevant in the context of “Data Gravity” [17], which posits that moving petabyte-scale data to external compute nodes incurs prohibitive latency.

Recent studies emphasize the urgent need to co-locate computation with data storage, that is, to place compute and storage resources within the same physical data center or HPC facility, to overcome these bottlenecks. Belcastro et al. [18] show that multi-layered edge-cloud architectures are essential for distributed workflows, yet adoption remains hindered by infrastructural fragmentation. Frameworks integrating HPC, cloud, and data-analytics resources [19], alongside AI-assisted metadata management [20] and FAIR-aligned research data infrastructures for high-throughput experimentation [21], illustrate how distributed edge-cloud-HPC continuums support diverse scientific workflows in practice. By formalizing computing proximity and readiness, AA-ROF translates these specific RDM and infrastructural realities into the quantifiable structural parameters of the three-tier model.

2.4. Positioning of AA-ROF Relative to Existing Literature

The reviewed literature reveals a well-established machinery for pricing uncertain assets and a growing consensus that architecture shapes data usability. It is important to acknowledge that advanced Net Present Value (NPV) extensions and existing real options models already address managerial flexibility and the combinatorial value of assets. The primary theoretical departure of AA-ROF is not the assertion that real options apply to data, but rather the explicit mathematical coupling of *architectural tier movement* with the friction and environmental penalty functions.

AA-ROF should also be distinguished from established systems methodologies. Checkland's shift from optimizing to learning cautions that quantitative models for intervention in human affairs do not themselves settle problem framing or stakeholder purposes [22]. AA-ROF therefore assumes that the evaluation purpose and system boundary have already been specified; it does not claim to replace a learning-oriented systems inquiry. System dynamics, in turn, explains policy resistance and behavior over time through endogenous system responses and feedback [23]. The readiness-reuse loop and the minimal two-period extension are therefore best viewed as a limited feedback representation, not as a calibrated system dynamics model.

Instead of relying on abstract policy frameworks, AA-ROF embeds the concrete RDM realities of the specific three-tier distributed model directly into the exercise cost term. To delineate this theoretical refinement without oversimplifying prior literature, Table 1 summarizes the specific conceptual pivot from generalized intrinsic valuation to the architecture-driven valuation of our three-tier model.

Table 1. Positioning of AA-ROF Relative to Existing Paradigms.

Analytical Dimension	Generalized IT/Valuation Models	The AA-ROF Approach (Three-Tier Model)
Focus of Value	Intrinsic Data Value or General IT Flexibility	Architecture-Driven Value (Tier progression)
Infrastructure Role	Operational constraint or underlying asset	Endogenous setter of exercise friction that raises realizable option value across the three-tier RDM progression
Evaluation Focus	Policy/Financial return on investment	Infrastructural friction & RDM status reduction
Environmental View	Exogenous operational constraint	Endogenous Sustainability Premium/Penalty (Irreversibility)

3. The Architecture-Aware Real Options Framework (AA-ROF)

To establish AA-ROF as a practical baseline for policy evaluation rather than a mere application of existing financial models, we note that the strike-monotonicity property underlying Principle 1 does not depend on a specific stochastic process for scientific

uncertainty, whereas the Vega-compression property of Principle 2 is established here only for the standard Black–Scholes geometric Brownian motion benchmark that the numerical illustrations of Section 4 also adopt, and whose limitations are discussed in Section 5.4. We construct a generalized analytical framework based on explicit assumptions, analytical properties, and logical derivations.

3.1. The Three-Tier Infrastructure Architecture

Before formalizing the friction and value functions, we first define the three-tier architectural gradient that anchors the analysis throughout this paper. We characterize each tier along the three dimensions of the infrastructure vector $\vec{I} = (A, R, P)$ (Accessibility, Readiness, and Computing Proximity), introduced formally in Section 3.3 and visually summarize this dynamic progression in Figure 1.

- **Tier 1 (Institutional Archive):** Data resides in isolated, institution-specific repositories with minimal metadata standardization and no direct co-location with compute resources. Accessibility and readiness are low, and computing proximity is minimal ($P \approx P_{min}$, formalized in Section 3.4), so exercise friction $X(\vec{I})$ is near its structural maximum. This tier corresponds to the conventional “cold storage” model that underlies dark-data accumulation.
- **Tier 2 (Federated Mesh):** Data is exposed through interoperable, FAIR-aligned interfaces across institutions (e.g., federated catalogs, shared identifier and metadata schemas), substantially raising accessibility (A) and readiness (R). Computing proximity remains moderate, as data is discoverable and interoperable but not necessarily co-located with high-performance compute. This tier reflects federated governance models, of which EOSC and Gaia-X are illustrative, not formally classified, examples. Federated architectures of this kind preserve institutional data sovereignty and organizational autonomy, since each institution retains custody of its own holdings while exposing them through shared interfaces; the corresponding cost is a governance overhead of sustaining cross-institutional consensus on schemas and identifiers, which can manifest as schema fragmentation and lags in data freshness relative to a fully centralized architecture.
- **Tier 3 (AI Sandbox):** Data is co-located with sovereign or institutional GPU clusters within a governed, compute-adjacent environment, maximizing computing proximity (P) and minimizing latency for AI model training. This tier minimizes exercise friction but, as formalized in Section 3.4, incurs the steepest environmental penalty $E(\vec{I})$.

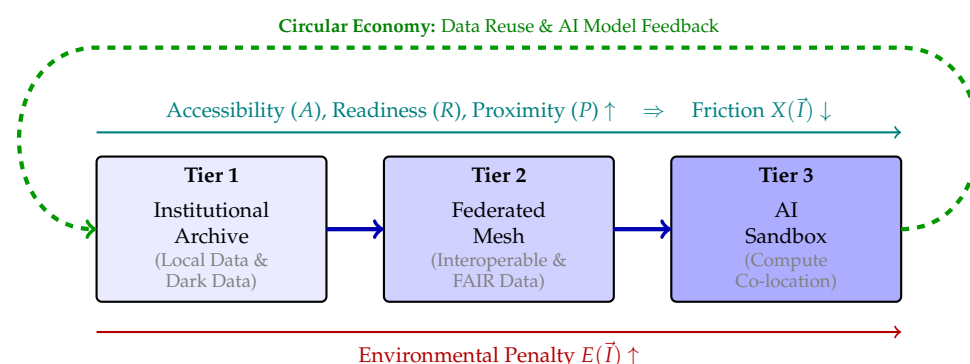


Figure 1. The three-tier infrastructural gradient integrating the Circular Data Economy. As data advances from Tier 1 to Tier 3, accessibility, readiness, and proximity increase, reducing exercise friction $X(\vec{I})$ (top) while environmental penalty $E(\vec{I})$ rises (bottom). The dashed feedback loop illustrates the Sustainability Premium achieved through continuous data and model reuse.

The dashed return path in Figure 1 represents the architecture as a closed loop rather than a one-directional pipeline: the outputs of the Sandbox tier, in the form of reused datasets and refined models, re-enter the archival base as newly curated holdings, raising the average readiness of the corpus that subsequent curation cycles draw on rather than altering the low-readiness definition of Tier 1 itself. This return path is a reinforcing loop in the qualitative system-dynamics sense, and we state plainly that the static formulation developed below does not endogenize it: neither $X(\vec{I})$ nor $E(\vec{I})$ makes the infrastructure state depend on prior exercise, so the loop is represented here as a structural property of the architecture. Section 4.6 takes a first step toward that loop in a minimal two-period form (reuse feeding back into each tier's own readiness and opportunity set, not yet into the archival base of another tier), and the remaining dynamic formalization is set out as a modeling extension in Section 5.4. As depicted in Figure 1, this tier structure is illustrative rather than a claim about any specific national system: real infrastructures may occupy intermediate or hybrid positions along the gradient. The choice of exactly three tiers is an analytical rather than an empirical judgment. Along the (A, R, P) space of Section 3.3, the tiers need only be numerous enough to trace the friction-versus-environmental-penalty trade-off of Section 3.4 through qualitatively distinct regimes: a high-friction, low-penalty regime in which data is largely inert (Tier 1), an intermediate regime with substantial interoperability but limited compute co-location (Tier 2), and a low-friction, high-penalty regime in which both option value and environmental cost are maximized (Tier 3). Three tiers are the minimum number that spans this trade-off end to end while retaining an interior point at which it is genuinely ambiguous; a finer partition would add intermediate points along the same gradient rather than a qualitatively new regime, at the cost of additional free parameters in Section 4.1. The tiers are accordingly representative archetypes chosen for analytical and expository sufficiency, not an exhaustive taxonomy of real infrastructures. Section 3.4 operationalizes this qualitative progression into the continuous friction function $X(\vec{I})$, and Section 4 discretizes it into the tier-indexed parameters (X_1, X_2, X_3) used in the numerical analysis.

3.2. Theoretical Core: Why Option Value over Cost Reduction?

In standard IS literature, an infrastructure upgrade is evaluated by the marginal decrease in processing costs (ΔC). Under this paradigm, the intrinsic value of the data asset remains static, and the infrastructure merely improves the profit margin of utilizing that data. This linear, operational perspective is adequate for predictable business environments but often falls short in scientific R&D, which is characterized by extreme uncertainty and the potential for discontinuous innovation.

In a dedicated three-tier distributed infrastructure, architecture governs which data-and-model combinations are exercisable rather than creating those combinations outright. When research data is elevated to a Tier 3 AI Sandbox, where computing proximity is maximized and latency is minimized, it gains the capacity to be instantly coupled with heterogeneous datasets and generative AI models. This state of high readiness does not just make existing research cheaper; it renders a *Real Option to Expand* economically exercisable. The infrastructure analytically transforms previously prohibitive combinatorial choices into viable ones for the researcher, allowing them to delay decisions until new scientific paradigms emerge and then execute complex analyses.

3.3. The Infrastructure Vector: Accessibility, Readiness, and Proximity

To operationalize the three-tier architecture, we explicitly define the dimensions of the infrastructure vector $\vec{I} = (A, R, P)$. A robust national R&D ecosystem demands not only data curation but also seamless institutional access and computational co-location.

The advancement of research data through the tiers is governed by the following three foundational pillars:

1. **Accessibility (A):** Overcoming institutional silos requires structural and legal interoperability. Accessibility is driven by the implementation of open APIs, standardized data-sharing agreements, and federated identity management (e.g., Single Sign-On across national R&D grids). High accessibility is fundamentally required to transition from a Tier-1 Archive to a Tier-2 Federated Mesh, enabling disparate research entities to seamlessly discover and query distributed datasets without administrative bottlenecks.
2. **Governance Readiness (R):** As the second pillar, readiness is not a binary state but a composite governance index. It is evaluated using three interdependent Data Maturity Indicators:
 - *FAIR Maturity Score (m_{fair})*: Quantified by the presence of persistent identifiers and machine-actionable metadata.
 - *Provenance Granularity (m_{prov})*: The availability of automated lineage logs, which strictly reduces the effort required to validate experimental reproducibility.
 - *Compute-Ready Encoding (m_{ai})*: The presence of standardized, GPU-accessible data structures (e.g., HDF5, Zarr).

We model readiness R as a Cobb–Douglas-style composite index:

$$R = \prod_{i=1}^3 m_i^{\omega_i} \quad (1)$$

where $(m_1, m_2, m_3) = (m_{fair}, m_{prov}, m_{ai})$, each $m_i \in [0, 1]$, and the ω_i are normalized weights satisfying $\omega_i > 0$ and $\sum_{i=1}^3 \omega_i = 1$, to be derived via multi-criteria decision-making methods based on surveys of national data governance experts. The geometric form is deliberately non-compensatory at the index level: a maturity component of zero drives R to zero, regardless of the other two, so no single governance dimension can be fully substituted by strength elsewhere. This is a statement about the readiness index and not about total friction, which stays bounded even at $R = 0$: as Section 3.4 shows, the readiness term of $X(\vec{I})$ then contributes at most $X_{max}\beta$ rather than diverging. Empirical implementations may impose a small floor, $m_i \geq \epsilon > 0$, so that R remains strictly positive, and the geometric mean is numerically well behaved, which relaxes the index-level statement without affecting the friction bound.

3. **Computing Proximity (P):** Driven by the principle of “Data Gravity,” moving petabyte-scale scientific data across standard networks is both economically and temporally infeasible. Computing proximity quantifies the physical and network closeness between data storage and HPC nodes. Maximizing P is strictly required for Tier-3 AI Sandboxes, ensuring that high-value datasets are co-located with sovereign GPU clusters to eliminate I/O latency during large-scale foundation model training.

3.4. Formalizing Infrastructural Friction and Environmental Penalties

To translate the conceptual infrastructure vector $\vec{I}$ into an analytical model, we formally specify the endogenous friction function $X(\vec{I})$ and the environmental penalty function $E(\vec{I})$.

The friction function, defined in Equation (2), dictates that exercise cost decreases as accessibility, readiness, and proximity improve:

$$X(\vec{I}) = X_{max} \left[\alpha(1 - A) + \beta e^{-\lambda R} + \gamma \left(\frac{1}{P} - 1 \right) \right] \quad (2)$$

where α, β, γ are structural weight parameters. The variables and parameters are restricted to the domains $A \in [0, 1]$, $R \in [0, 1]$, and $P \in [P_{min}, 1]$ with $P_{min} > 0$, together with $\alpha, \beta, \gamma > 0$,

$\lambda > 0$, and $X_{max} > 0$. Under these restrictions, $X(\vec{I})$ is finite and strictly positive, attaining its structural maximum $X_{max}[\alpha + \beta + \gamma(1/P_{min} - 1)]$ at $(A, R, P) = (0, 0, P_{min})$ and its structural minimum $X_{min} \equiv X(1, 1, 1) = X_{max} \beta e^{-\lambda} > 0$ at the fully developed corner $(1, 1, 1)$. Two naming clarifications follow. The parameter X_{max} is a friction scale rather than a literal upper bound of the function, as the expression for the attained maximum makes clear. The fully developed corner is not a frictionless one: a strictly positive floor, X_{min} , persists there because standardized metadata reduces, but never eliminates, the interpretive cost of using someone else's data. While this functional form is deductively formulated within this framework, it structurally reflects established principles of information economics. Specifically, the linear term for accessibility captures basic network access dynamics; the exponential decay of readiness (λ) captures the diminishing marginal returns of metadata curation; and the proximity term ($\frac{1}{P}$) mathematically models "Data Gravity," where transmission friction grows steeply as compute nodes are separated from storage. The unrestricted form would diverge as $P \rightarrow 0^+$, which is precisely why the domain restriction $P \geq P_{min}$ is imposed: Tier-1 friction is thereby kept large but finite, consistent with the tier description in Section 3.1.

Furthermore, a critical mathematical feature of Equation (2) is its behavior at the boundary condition where data maturity is entirely absent ($R \rightarrow 0$). In scenarios where governance is non-existent, the readiness friction term does not diverge to infinity but resolves to a finite ceiling, contributing at most $X_{max}\beta$ to total friction. From an operational policy perspective, this bounded form is consistent with operational experience: an absolute lack of metadata necessitates the maximum possible manual labor (e.g., reconstructing or reverse-engineering the dataset from scratch), which remains logically bounded by the finite structural maximum of Equation (2). This ensures that the framework remains robust when evaluating the integration of legacy "dark data."

Conversely, the environmental penalty $E(\vec{I})$ is defined in Equation (3) as strictly convex with respect to computing proximity P :

$$E(\vec{I}) = E_{base} + \eta P^\kappa \quad (3)$$

where E_{base} is the baseline archive energy cost, $\eta > 0$ is a hardware-specific energy scaling coefficient, and $\kappa > 1$ ensures strict convexity in P over the domain $P \in [P_{min}, 1]$. The convex power-law term models the non-linear energy overheads typical of high-density GPU cooling and power usage. We stress that P enters Equation (3) as a reduced-form index: greater compute co-location proxies for the far larger absolute IT workload that co-located clusters host, so $E(\vec{I})$ measures the total energy burden rather than facility efficiency alone. The vector argument is retained in $E(\vec{I})$ for symmetry with $X(\vec{I})$, but the present specification is deliberately univariate in P : accessibility and readiness are assumed to govern discovery and reuse friction rather than energy draw. Letting the environmental penalty depend explicitly on workload and utilization alongside proximity is a specification extension taken up in Section 5.4. This formalizes the central macroeconomic trade-off of national AI policies: advancing to Tier 3 minimizes infrastructural friction $X(\vec{I})$ but sharply increases the environmental penalty $E(\vec{I})$ under the assumed power law, due to the immense power demands of sovereign compute clusters.

Equation (3) is best read as a stylized reduced-form penalty rather than a first-principles emissions model: it deliberately compresses the environmental burden of a tier into the single scalar $E(\vec{I})$ rather than itemizing the distinct physical quantities that contribute to it. Six such quantities can be distinguished: energy efficiency (facility-level Power Usage Effectiveness, on which Tier-3 facilities typically score better), absolute computational demand (the far larger IT workload that higher-tier, GPU-dense facilities host, which drives the assumed $E_1 < E_2 < E_3$ ordering), transmission energy (the network cost of moving data to and between compute nodes), facility-level energy consumption (the

metered draw of the hosting data center), embodied emissions (the manufacturing and hardware life-cycle footprint of compute and storage equipment), and total burden (the aggregate quantity that $E(\vec{I})$ is calibrated to approximate). Equation (3) does not itemize transmission energy or embodied emissions separately: as specified, $E(\vec{I})$ approximates only the operational energy burden of the co-located compute and storage, and transmission energy and embodied emissions are not included in the current functional form; disentangling all six components empirically and extending $E(\vec{I})$ to incorporate them is left to the calibration roadmap of Section 5.4.

The transformation of data maturity is represented by a monotonic decline of the friction function $X(\vec{I})$ across the infrastructure state space, mapping the transition from high-friction institutional archives (Tier 1) to the low-friction, AI-ready environment of the Sandbox (Tier 3); Section 4.1 reports the tier-level values of $X(\vec{I})$ actually attained under the illustrative calibration and, in its worked structural-consistency example, the (A, R, P) coordinates from which Equation (2) reproduces them.

3.5. Conceptual Formalization of Value Expansion, Sustainability, and Environment

To quantify the dynamic value of research data, we define the real options value function $C(S, X, \Theta)$ using the standard Black–Scholes call option pricing model [24,25]. Let S be the intrinsic epistemic value, let $X(\vec{I})$ be the endogenous infrastructural friction acting as the strike price, and let σ be the scientific volatility. Throughout, $\Theta = (\sigma, r, T)$ collects the option-pricing parameters, namely the volatility, the risk-free rate, and the time horizon. The option value is given by:

$$C(S, X, \Theta) = S\Phi(d_1) - Xe^{-rT}\Phi(d_2) \quad (4)$$

where $d_1 = \frac{\ln(S/X) + (r + \sigma^2/2)T}{\sigma\sqrt{T}}$, $d_2 = d_1 - \sigma\sqrt{T}$, and $\Phi(\cdot)$ is the cumulative standard normal distribution.

Because research data is a non-market asset, applying standard financial option pricing requires the Marketed Asset Disclaimer (MAD) assumption [26]. Under MAD, we assume that the present value of the public investment required to curate the data and provision the infrastructure acts as the perfectly correlated replicating portfolio, so that this present value can stand in for the price of a traded twin security. We emphasize that this substitution equates S with an *expenditure-based proxy* for value, not with value itself: the underlying MAD assumption is that public R&D investment is efficiently allocated in expectation, so that the present value of resources committed to curating and provisioning a dataset approximates the market's best estimate of its capitalized epistemic value. This is a substantive identifying assumption rather than a definitional equivalence: if public R&D spending is systematically mispriced relative to the scientific value it generates (e.g., due to political misallocation or Baumol-type cost disease in research), S measured as expenditure will over- or understate true epistemic value, and the resulting option values inherit that bias. We adopt MAD because it is the standard resolution to the non-tradability problem in the real-options literature [26], but we flag the expenditure-value gap explicitly as an identification assumption to be tested, not assumed away, in the empirical calibration program of Section 5.4. Two boundaries follow for measurement. First, the proxy inherits a directional caveat: because higher expenditure mechanically raises S , a naive application would assign a higher option value to an inefficiently expensive project, which is precisely why efficient allocation in expectation is stated here as a testable assumption rather than a definition. Second, S prices the capitalized epistemic value of the dataset itself, whereas $X(\vec{I})$ and $E(\vec{I})$ price the friction and the energy of the infrastructure that makes that value exercisable; each expenditure item must therefore be assigned to exactly

one of these accounts, and provisioning cost already charged to X , E , or $\Delta Capex$ must not re-enter S .

To keep the first of these caveats visible in the formalism rather than only in prose, we write the proxy with an explicit allocative-efficiency coefficient,

$$S = \psi \cdot PV(\text{curation and provisioning expenditure}), \quad \psi \in (0, 1], \quad (5)$$

where ψ measures the fraction of committed public resources that is converted into capitalized epistemic value; the MAD assumption of efficient allocation in expectation is the special case $\psi = 1$, which the illustrative calibration of Section 4 adopts, so no reported quantity depends on introducing ψ . Its role is to make the identification burden explicit and directionally signed: since $\partial C / \partial S > 0$, evaluating a program whose true $\psi < 1$ at $\psi = 1$ overstates S ; hence, C and every indicator that depends on C (namely NSV and AP , though not the reuse-based accounts NEB and REA), and the proportional overstatement of C exceeds that of S , most strongly at the least advanced tier, so a common $\psi < 1$ compresses the measured Architecture Premium as well as the levels. Appendix A derives this elasticity result from Equation (4) and states the conditions under which cross-program and within-program comparisons remain valid when ψ is unknown.

The expenditure-based proxy is theoretically defensible precisely because S enters $C(S, X, \Theta)$ as the value of the underlying asset in a real-options sense: what is being priced is the option to exploit whatever epistemic value that expenditure has created, so S needs only to track the correct order of magnitude and direction of that value, not to equal it exactly, for the comparative-static results of Principles 1–3 to hold. This is the same identifying role that replacement cost or accumulated investment plays as an asset-value proxy elsewhere in the real-options literature under MAD [26]. Because ψ renders the resulting bias explicit and directionally signed, an empirical operationalization can partially discipline it by cross-checking expenditure against independent value proxies: cumulative citation counts, downstream reuse counts, technology-transfer activity (licensing or spin-off events attributable to the dataset), and other downstream research outputs (e.g., derivative datasets or publications) are candidate proxies for the empirical calibration roadmap of Section 5.4, none of which is implemented in the present illustrative analysis. The baseline value $S = 100$ used throughout Section 4 is accordingly an illustrative normalization of the index unit, chosen so that other quantities can be read as ratios to it, and not an estimate of any particular program's expenditure.

Table 2 fixes the model-level definitions and unit conventions used throughout the analysis. In particular, all value quantities (S , X , E_i , K_{econ} , $p_c K_{carbon}$) are expressed as flows per evaluation horizon T in a common normalized index unit, so that additive comparisons among them are dimensionally meaningful, and all option values are read as normalized option-value indices under the geometric Brownian motion and MAD assumptions rather than as monetary valuations.

Based on this formalization, the structural propositions below rest on one standard property of option pricing and one analytical property derived from our framework.

Standard Property (Strike Monotonicity): Within the Black–Scholes model, $C(S, X, \Theta)$ is continuously differentiable in X with $\frac{\partial C}{\partial X} = -e^{-rT} \Phi(d_2) < 0$ for any fixed $S > 0$ and any $\Theta = (\sigma, r, T)$ with $X > 0$, $\sigma > 0$, and $T > 0$, meaning the option value is strictly decreasing as exercise friction increases. The strict inequality requires these interior conditions: $\Phi(d_2)$ is strictly positive at every finite d_2 and degenerates only in the limits $\sigma \rightarrow 0$ or $T \rightarrow 0$, where it tends to zero or to one according to whether the option is out of or in the money on a discounted basis. This is a textbook property of call options, derived rather than assumed, and it is not specific to this paper.

Analytical Property (Monotonic Friction Reduction): The friction function $X(\vec{I})$ is continuously differentiable in the infrastructure vector $\vec{I} = (A, R, P)$, and $\nabla X(\vec{I}) < 0$ componentwise. This is a direct algebraic corollary derived from the functional form in Equation (2): improving accessibility, readiness, or computing proximity inherently reduces exercise friction.

The strike monotonicity is standard; the monotonic friction reduction is a modeling choice specific to AA-ROF, and its plausibility depends on the functional form of $X(\vec{I})$ remaining monotonic outside the illustrative range explored in Section 4, a point we return to in Section 5.

Principle 1 (Friction-Reduction Premium). This principle is a direct composition of the standard strike-monotonicity property stated above with the model-specific friction function of Equation (2), rather than a new option-pricing result. For an arbitrage-free real options valuation model, and holding the intrinsic asset value S and the remaining option parameters Θ fixed, advancing the infrastructure vector $\vec{I}$ strictly increases the gross option value of the research data, securing the option to capitalize on future scientific volatility. The proposition is stated before environmental and transition costs. The net comparison, which subtracts the environmental penalty $E(\vec{I})$ and the capital cost of the transition, is carried out in Sections 4 and 5.3; because $E(\vec{I})$ rises along the same architectural gradient, the net ordering is not implied by the gross one. It happens to survive at the illustrative calibration of Table 3, where NSV still increases across tiers at every volatility level examined, but Scenario C identifies the environmental penalty at which it fails, and the robustness analysis identifies a friction perturbation that reverses it. Whether the gross ordering carries over to the net comparison is therefore an empirical question about E and the transition cost, not a corollary of this proposition, and exposing that gap is precisely what the framework is built to do.

Table 2. Model-level definitions and unit conventions.

Object	Definition and Unit Convention
S	Intrinsic epistemic value. Canonical proxy: present value of public curation and provisioning expenditure under the MAD assumption; citation- or reuse-based measures serve only as alternative cross-checks. Normalized index units.
$X(\vec{I})$	Exercise friction (strike price), an endogenous function of $\vec{I}$. Normalized index units per evaluation horizon.
σ	Annualized volatility of the value process of S under geometric Brownian motion; empirical proxies (citation variance, grant-funding fluctuation) belong to the calibration roadmap of Section 5.4.
r, T	Assumed scenario constants ($r = 0.05$, $T = 1.0$); not empirical estimates.
$E_i, K_{econ}, p_c K_{carbon}$	Flows per evaluation horizon T in the same normalized index unit as S and X , stated as present values at the valuation date. The discount factor e^{-rT} therefore appears explicitly only inside C , where it applies to the strike because that amount is paid at exercise; every other cost and benefit term is already expressed in present-value units and enters NSV , NEB , and REA without further discounting.
$C(S, X, \Theta)$	Normalized option-value index under the geometric Brownian motion and MAD assumptions; not a monetary valuation.

Analytical Implication. Let $C(S, X, \Theta)$ be a continuously differentiable call option value function. By the strike monotonicity property ($\frac{\partial C}{\partial X} < 0$) and the monotonic friction reduction property ($\nabla X(\vec{I}) < \vec{0}$), the total derivative of the option value with respect to any forward

architectural movement ($d\vec{I} \geq \vec{0}$ with at least one strict component, taken within the interior of the domain of Section 3.4) is strictly positive:

$$dC = \left(\nabla X(\vec{I}) \cdot d\vec{I} \right) \frac{\partial C}{\partial X} > 0$$

Economically, this compact derivation implies that infrastructure upgrades directly translate into the expansion of gross option value, bypassing the need for intrinsic asset modification. This result holds the research opportunity set fixed; Section 4.6 partially endogenizes that set through $N'_i = 1 + \nu\rho_i$ (Equation (9)), while a fuller combinatorial expansion of the opportunity set itself remains part of the theoretical roadmap of Section 5.4.

Table 3. Illustrative normalized parameters for the structural sensitivity analysis (literature-informed).

Category	Parameter	Assumed Normalized Value (Rationale)
Asset & Environment	Intrinsic Epistemic Value (S)	100 (Normalized index; expenditure-based proxy under MAD)
	Risk-free Rate (r) & Horizon (T)	$r = 0.05$, $T = 1.0$ (Assumed scenario values)
Infrastructure Friction	Tier 1 (Archive) Friction (X_1)	150 (Proxy: 80% wrangling overhead)
	Tier 2 (Federated) Friction (X_2)	70 (Proxy: Reduced by FAIR cataloging)
	Tier 3 (AI Sandbox) Friction (X_3)	20 (Proxy: Compute-adjacent automation)
Environmental Costs	Redundant Creation Cost (K_{econ})	50 (Proxy: Grant-funded reproduction)
	Carbon Penalty Conversion ($p_c K_{carbon}$)	10 (Proxy: Social Cost of Carbon)
	Tier Energy Penalties (E_1, E_2, E_3)	5, 15, 50 (Assumed total energy burden: IT load $\times$ PUE)

Principle 2 (Conditional Volatility Compression). This principle is the standard deep-in-the-money Vega-compression property of call options, applied here to the endogenous friction X_i rather than to an exogenously observed strike. Let $X_i \equiv X(\vec{I}_i)$ denote the residual friction actually attained at tier i , the same symbol used in the numerical calibration of Section 4; it is distinct from the structural floor X_{min} of Section 3.4, which no attainable tier reaches, and in general strictly exceeds it. Conditional upon the infrastructure advancing to a tier where this residual friction is strictly dominated by the intrinsic asset value, that is $X_i \ll S$, which places the option deep in the money and in particular implies $d_1 > 0$, the architecture sharply compresses, without eliminating, the option's sensitivity to volatility (Vega, $\mathcal{V}$); this is the conditional volatility-compression effect anticipated in Section 1. The label is a shorthand and should be read literally: what the architecture compresses is $\partial C / \partial \sigma$, the valuation's responsiveness to scientific uncertainty, and not σ itself, which is a property of the science and is unaffected by infrastructure.

Analytical Implication. In option pricing theory, Vega ($\mathcal{V}$) is defined as $\mathcal{V} = \frac{\partial C}{\partial \sigma} = S\sqrt{T}\phi(d_1)$, where $\phi(d_1) = \frac{1}{\sqrt{2\pi}}e^{-d_1^2/2}$ and

$$d_1 = \frac{\ln(S/X_i) + (r + \sigma^2/2)T}{\sigma\sqrt{T}}.$$

Two statements must be distinguished. The exact limit holds only as friction vanishes: for fixed S, σ, r, T , letting $X \rightarrow 0^+$ drives $d_1 \rightarrow \infty$ and hence $\mathcal{V} \rightarrow 0$. At any finite Tier-3 friction ($X_3 > 0$), Vega is compressed but strictly positive: under the illustrative calibration of Section 4 ($S = 100$, $T = 1$), Vega falls from approximately 31.5 (Tier 1, $X_1 = 150$) to 0.003 (Tier 3, $X_3 = 20$) at $\sigma = 0.4$, and from approximately 39.9 to 1.9 at $\sigma = 0.8$. This is a large but finite compression, and it is the standard deep-in-the-money behavior of call options composed with the friction function, rather than a separate hedging mechanism.

The compression is monotone in friction only within that regime: Vega attains its maximum at the friction level $X^* = Se^{(r+\sigma^2/2)T}$ at which $d_1 = 0$, approximately 113.9 at $\sigma = 0.4$ under the same calibration, so reducing friction from a level above X^* first raises Vega toward that peak before compressing it. Because $r > 0$ and $\sigma^2/2 > 0$, $X^* = Se^{(r+\sigma^2/2)T}$ always exceeds the intrinsic asset value S ; this is a structural property of the Black–Scholes formula itself, not an artifact of the calibration in Table 3. X^* is therefore a Vega-maximizing turning point rather than a moneyness boundary: the option is not yet in the money there, and the deep-in-the-money regime of Principle 2 begins only at friction levels well below it. At the baseline calibration, the tier progression reported here crosses X^* between Tier 1 and Tier 2 ($X_1 = 150 > X^* \approx 113.9 > X_2 = 70$), and Principle 2 therefore claims compression only along the deep-in-the-money segment beyond that crossing; unlike the general inequality $X^* > S$, this specific crossing location is a property of the calibrated X_1 and X_2 values and is not guaranteed to hold outside them. In particular, it is not robust across the full X_1 range used in the joint robustness check of Section 4.5 ($X_1 \in [112.5, 150]$): at the lower end of that range, $X_1 = 112.5$ falls below $X^* \approx 113.9$, so the crossing would occur at or before Tier 1 rather than strictly between Tier 1 and Tier 2. The one-at-a-time sensitivity table of Section 4.5 varies each input independently and does not itself test this crossing margin, nor does it capture interaction effects among inputs that move simultaneously; the joint Monte Carlo test partially addresses the latter by drawing all inputs together, and Section 4.5 reports how often the $X_1 > X^*$ placement underlying this crossing survives under the calibrated draw range (approximately 96% of draws at $\sigma = 0.4$ with S held at 100), together with the scope limits of that figure. In the same deep-in-the-money regime, the option value converges to its deterministic exercise value, $C \rightarrow S - Xe^{-rT}$; the framework therefore nests the static NPV benchmark criticized in Section 1 as its low-friction limiting case, and the option lens is most informative in intermediate-friction, high-uncertainty regimes (e.g., Tier 1 to Tier 2 transitions), where time value remains material.

Principle 3 (Net Environmental Benefit and Data Circularity). Let K_{econ} be the financial cost, and let K_{carbon} be the environmental emission cost of redundantly reproducing an existing dataset due to accessibility failure. K_{econ} and K_{carbon} price only this reuse-failure redundancy, that is, the re-creation that occurs because an existing dataset could not be found or accessed; they do not price redundancy that is deliberately engineered for fault tolerance or availability, such as planned multi-site replication of critical holdings, which carries its own resilience benefits and lies outside the scope of this model. Under the condition that Tier-1 friction leaves the data’s realizable value below the recreation cost ($C_1 < K_{econ}$), and noting that, whenever the residual friction itself exceeds the recreation cost ($X_1 > K_{econ}$), a rational actor prefers re-creation over reuse, the data devolves into “dark data,” forcing redundant investments. The two conditions are distinct and neither implies the other. The first is a value-side barrier: the option on the existing dataset is worth less than rebuilding it. The second is a choice-side barrier: paying the access friction costs more than rebuilding, irrespective of what the option is worth. Either condition alone is sufficient to drive a rational actor toward re-creation, and the illustrative Tier-1 calibration of Section 4 satisfies both ($C_1 \approx 5 < 50$ and $X_1 = 150 > 50$ at $\sigma = 0.4$). Tier 3 infrastructure is assumed to enable a Circular Data Economy and thereby to generate a macroeconomic *Net Environmental Benefit (NEB)*; we refer to this architectural surplus as the “Sustainability Premium” introduced in Section 1.

Analytical Implication. To internalize circularity and maintain dimensional consistency within the model, we introduce p_c as a conversion factor (e.g., the Social Cost of Carbon) that translates the physical carbon emissions (K_{carbon}) into monetary equivalents. Let ρ_{reuse} be the *incremental* cross-disciplinary reuse rate, defined as the expected number of reuse events per dataset within the evaluation horizon T that the advanced tier enables

and that would not have occurred under Tier-1 friction; by construction $\rho_{reuse} = 0$ at Tier 1, so the reuse benefit credited below is attributable in full to the tier transition and is not double-counted against a pre-existing baseline. The scenarios of Section 4 evaluate $\rho_{reuse} \geq 1$, with $\rho_{reuse} = 1$ as the conservative case of a single incremental reuse event. The reuse benefits and the energy penalties below are flows over the same horizon, so the aggregation is symmetric. At Tier 3, as $X_3 \rightarrow X_{min}$, the realizable option value exceeds the recreation cost at the illustrative parameterization ($C_3 \approx 81 > K_{econ} = 50$), removing the value-side barrier to reuse; whether reuse actually occurs additionally depends on demand and researcher choice, which the static framework treats as exogenous through ρ_{reuse} ; Section 4.6 relaxes that treatment by tying the reuse rate to moneyiness. The total Net Environmental Benefit (*NEB*) is mathematically quantified as the sum of avoided economic and environmental deadweight losses over ρ_{reuse} cycles, minus the operational penalty of the infrastructure (E_3):

$$NEB = [\rho_{reuse} \cdot (K_{econ} + p_c K_{carbon})] - (E_3 - E_1)$$

Two accounting conventions in this formulation are made explicit. First, each reuse event is assumed to avert exactly one full redundant re-creation, a one-to-one counterfactual. Three related quantities are worth distinguishing here: the potential reuse value is the option value C_3 that low Tier-3 friction makes exercisable; the avoided redundant activity is the $K_{econ} + p_c K_{carbon}$ credited per reuse event in the *NEB* expression above; and the actual environmental saving is the carbon component ($p_c K_{carbon}$) of the fraction of that avoided activity that materializes, governed by the avoidance conversion rate θ introduced below; the K_{econ} component of the same credit is an economic, not an environmental, saving. The one-to-one counterfactual identifies the third quantity with the second at $\theta = 1$; whenever $\theta < 1$, the avoided-activity credit overstates the actual environmental saving, so the displayed *NEB* is an upper bound on the environmental account; the option value C_3 and the avoided-activity credit are kept as separate accounts throughout and are not equated with each other. Second, *NEB* is kept as a separate environmental-economic account from the single-exercise option value C , and per-event access frictions are not netted out of it. Both simplifications can overstate *NEB* when a reuse event would not otherwise have triggered re-creation or when per-event access costs are material; we return to this in Section 5.4. Let $\theta \in (0, 1]$ denote the avoidance conversion rate, namely the proportion of reuse events that actually substitute for redundant re-creation. The generalized account is $\rho_{reuse} \theta (K_{econ} + p_c K_{carbon}) - (E_3 - E_1)$, and the displayed *NEB* expression is the special case $\theta = 1$, so every numerical result reported in this paper is unaffected. The positive-benefit condition correspondingly generalizes to $\rho_{reuse} > 0.75/\theta$ at the illustrative baseline: at $\theta = 0.5$, the requirement rises to 1.5 incremental reuse events, which the baseline $\rho_{reuse} = 1$ does not meet. From an empirical measurement perspective, this formulation is highly operationalizable: K_{econ} can be quantified by the labor and grant funding required for data reproduction, $p_c K_{carbon}$ can be monetized via life cycle assessment of the redundant experimental hardware, and E_i can be directly tracked using data center Power Usage Effectiveness (PUE) metrics and GPU workload logs. This explicit measurability ensures that the Circular Economy concept within AA-ROF is not merely abstract, but an analytically quantifiable objective function. The reuse condition follows directly from the formulation: $NEB > 0$ holds if and only if $\rho_{reuse} > (E_3 - E_1)/(K_{econ} + p_c K_{carbon})$. At the illustrative baseline of Section 4, this ratio equals $45/60 = 0.75$, so every reuse level evaluated there ($\rho_{reuse} \geq 1$) satisfies the condition, whereas a transition that generated less than three quarters of one incremental reuse event per dataset would not; the reuse requirement binds, however, as the energy differential grows or as the recreation costs shrink, which is exactly the regime that the screening indicator of Section 5.3 is designed to flag.

To bridge these theoretical propositions with operational policy evaluation, we formally define the exact measure of this value expansion by introducing the *Architecture Premium*.

Definition 1 (The Architecture Premium). *The Architecture Premium (AP) is defined as the discrete expansion of the option value resulting from the architectural movement from a base state, $\vec{I}_{base}$, to a target state, $\vec{I}_{target}$:*

$$AP = C(S, X(\vec{I}_{target}), \Theta) - C(S, X(\vec{I}_{base}), \Theta)$$

This discrete finite-difference formulation deliberately avoids the theoretical abstraction of continuous path integrals. It directly maps to the step-wise tier transitions (e.g., Tier 1 $\rightarrow$ Tier 3) typically utilized in public R&D infrastructure planning and serves as the exact objective function for our subsequent numerical simulations. Formally, *AP* is a friction-reduction premium: it follows from strike monotonicity composed with the friction function, and it captures no effect beyond the reduction in the exercise friction $X(\vec{I})$. The combinatorial interpretation, namely that lower friction is what permits heterogeneous datasets to be coupled with AI models at scale under high Computing Proximity (*P*) and Governance Readiness (*R*), is an economic reading of that reduction, not a separate mathematical object; extending the framework so that the size or arrival rate of the research opportunity set itself depends on $\vec{I}$ is taken up in the minimal dynamic extension of Section 4.6, where realized reuse scales the number of period-two opportunities, and its fuller treatment remains part of the theoretical roadmap in Section 5.4.

In practical public R&D policy evaluation and Pre-Feasibility Studies (PFS), *AP* is measured by operationalizing its core components: (1) Intrinsic value *S* is proxied canonically by curation and provisioning expenditure under the MAD assumption (Table 2), with historical citation impact and reuse valuation serving as alternative cross-checks; (2) Exercise friction $X(\vec{I})$ is estimated through average data-wrangling labor hours and network latency logs; and (3) Volatility σ is benchmarked using citation variance or grant-funding fluctuations within specific scientific domains.

4. Illustrative Structural Analysis of the Three-Tier Architecture

This section conducts a structural sensitivity analysis to operationalize the proposed three-tier architecture. Rather than aiming to empirically discover deterministic financial returns or evaluate specific national policy budgets, we utilize this numerical section as a comparative-static structural heuristic. By applying illustrative normalized values informed by IT infrastructure and RDM literature, we identify symbolic boundary conditions for comparing the three tiers. Unless stated otherwise, the numerical threshold tables report the zero-capital-cost intercepts of those boundaries; they do not establish a preferred tier without a program-level annualized transition cost.

To keep these illustrative results from being read as more than they are, three levels of claim should be kept distinct throughout this section. The first is the set of structural, deductive propositions of Section 3 (Principles 1–3 and the strike-monotonicity and friction-reduction properties), which hold under their stated conditions for any admissible parameterization of Equations (2) and (3) (Principle 2 within the geometric Brownian benchmark and its deep-in-the-money regime, Principle 3 under the reuse-failure conditions stated there) and are not specific to the numbers below. The second is the illustrative calibrated parameter set of Table 3, which fixes one admissible, literature-informed point in that space; the numerical results attached to it, including the value of 60 discussed in Section 4.4, are illustrative point values computed from a calibration that is not identified by the model, whose role is to demonstrate direction and order of magnitude rather than to estimate any specific infrastructure. The third is the non-empirical, illustrative narrative of Section 5 that maps AA-ROF’s vocabulary onto named programs (NAIRR, EOSC, Gaia-X);

those mappings carry no claim about the actual friction, cost, or carbon profile of any real system. Unless stated otherwise, the results below should be read at the second level.

4.1. Illustrative Normalized Calibration for the Three-Tier Model

To ensure the transparency and reproducibility of our sensitivity analysis, we parameterize the structural variables, namely the environmental penalty (E_i), the infrastructural friction (X_i), and the data reuse rate (ρ_{reuse}), using illustrative normalized values informed by relative metrics from empirical studies on data center efficiency and RDM status. Table 3 outlines these assumed values.

- Environmental Penalties (E_i) via PUE Proxies:** Rather than arbitrary index values, energy penalties are informed by empirical facility-efficiency and demand data. According to the Uptime Institute's Global Data Center Survey [27], the industry average PUE stood at 1.56 in 2024 and has remained essentially flat for five consecutive years, an average kept high by the large stock of aging legacy facilities, whereas recent optimized builds routinely achieve $PUE \approx 1.3$ and leading hyperscale fleets report fleet-wide averages near 1.1 [28]. PUE alone, however, moves in the opposite direction of the assumed E_i ordering: it measures facility overhead per unit of IT load, and Tier-3 facilities are the more efficient ones on that metric. The ordering $E_1 < E_2 < E_3$ is instead driven by the far larger absolute IT workload that higher tiers host: electricity consumption in accelerated servers, driven mainly by AI adoption, is projected to grow by approximately 30% annually and to account for almost half of the net increase in global data center electricity consumption, and a typical AI-focused data center already consumes as much electricity as 100,000 households [1]. The E_i values therefore represent assumed total energy burdens (IT load $\times$ PUE) per evaluation horizon, in which the growth of absolute compute demand dominates the partially offsetting PUE improvements; the assumed 1:3:10 progression of the E_i values is an illustrative encoding of this workload-driven divergence, not a measured ratio.
- Infrastructural Friction (X_i) via Data Management Labor Costs:** Infrastructural friction is calibrated using industry-standard RDM metrics. Practitioner accounts of data science workflows widely report that researchers spend approximately 60% to 80% of their time on data discovery, wrangling, and cleaning due to fragmented silos; the upper end of this range traces to early work on data cleaning, which places the combined exploratory-analysis and cleaning effort at about 80% [29]. We treat the range as an indicative order of magnitude rather than a measured constant. We map this maximum labor overhead to Tier 1 friction (X_1). Transitioning to a Tier 3 AI Sandbox, where metadata is standardized and co-located with compute, structurally compresses this manual labor constraint, reducing the friction parameter (X_3) to a baseline computational latency cost. The macroeconomic scale of this friction channel is independently documented: a European Commission study estimates the measurable cost of *not* having FAIR research data in Europe at a minimum of EUR 10.2 billion per year, driven by time spent, storage and license costs, research retraction, double funding, and forgone interdisciplinarity [30].
- Cross-Disciplinary Reuse Rate (ρ_{reuse}) via Open Data Premium:** The reuse rate is not assumed statically but motivated by bibliometric evidence regarding the "Open Data Citation Advantage." Reported estimates suggest a twofold to fivefold increase in cross-disciplinary reuse for datasets achieving high FAIR compliance and computational interoperability, compared to isolated data. We flag this range explicitly as an assumption rather than an established effect size. The published evidence on open data documents a citation advantage of roughly 9% to 25% [31,32] rather than a reuse multiple of this magnitude, and the closest directly measured reuse quantity is the

accumulation of more than 150 third-party reuse publications from 100 deposited datasets over five years [31]. Because NSV and the Tier-2/Tier-3 total-cost boundary are independent of ρ_{reuse} , as noted in Scenario B, this assumption affects only the NEB and REA accounts. Thus, we evaluate scenarios ranging from $\rho_{reuse} = 1.0$ (conservative baseline scenario) to $\rho_{reuse} \geq 3.0$ (high-reuse Tier 3 environments), where ρ_{reuse} counts incremental reuse events as defined in Section 3. The translation is itself an assumption and is stated as such: the cited evidence reports a relative multiple of reuse, whereas ρ_{reuse} is an absolute count of incremental events per dataset, so the scenario values are read as illustrative counts consistent with the reported multiples rather than as quantities the bibliometric studies measure directly.

A note on calibration status: The proxy values presented in Table 3 are derived from relative ratios found in infrastructure literature, designed to satisfy the qualitative dynamics ($X_1 > X_2 > X_3$ and $E_1 < E_2 < E_3$) inherent in the three-tier architecture. They function as structural anchors for analyzing the threshold behaviors of the model, rather than absolute point estimates of any specific institutional expenditure. The mapping from the cited ranges onto the index scale is fixed as follows: $X_1 = 150$ anchors the upper bound of the reported wrangling-overhead range, so that the 60–80% range corresponds linearly to $X_1 \in [112.5, 150]$ (with $112.5 = 150 \times 60/80$); $X_3 = 20$ represents residual computational latency; and the E_i values encode the assumed total-energy ordering described above. All quantities are expressed in the common normalized index unit of Table 2, in which only relative magnitudes are meaningful, and all of them are per-dataset allocations on a single scale: E_i denotes the marginal energy burden attributable to one dataset's storage and workload at tier i , not the absolute draw of a national facility, so that E_i is dimensionally comparable with S , K_{econ} , and X_i . The facility-level evidence cited above therefore motivates the ratio 1:3:10 between tiers, not the absolute level of any E_i . We note explicitly that these tier-level values are calibrated directly from the cited literature ratios rather than by numerically inverting the continuous friction function $X(\vec{I})$ of Equation (2) or the environmental penalty function $E(\vec{I})$ of Equation (3) from underlying (A, R, P) values; establishing that micro-to-macro linkage empirically (i.e., fitting $\alpha, \beta, \gamma, \lambda, \eta, \kappa$ to observed (A, R, P) trajectories of real infrastructures) is itself part of the empirical roadmap in Section 5.4. Because the tier-level values in Table 3 are themselves proxies rather than measured quantities, we treat them as uncertain inputs and formally examine how far the boundary-condition results of the following scenarios can move under plausible variation in these inputs; this robustness check is reported in Section 4.5.

A worked structural-consistency example: Although the tier-level values are not obtained by inverting the continuous functions, they are attainable states of those functions rather than free-standing assignments. As one admissible parameterization, let the tier coordinates be $\vec{I}_1 = (0.2, 0.2, 0.2)$, $\vec{I}_2 = (0.6, 0.6, 0.6)$, and $\vec{I}_3 = (0.9, 0.9, 0.9)$ with $P_{min} = 0.1$. Setting $X_{max} = 200$, $\lambda = 2$, and $(\alpha, \beta, \gamma) = (0.7385, 0.1481, 0.0150)$ in Equation (2) reproduces $X(\vec{I}_1) = 150.0$, $X(\vec{I}_2) = 70.0$, and $X(\vec{I}_3) = 20.0$ to one decimal, with all weights strictly positive as required in Section 3.4 and a positive friction floor $X_{min} = X_{max}\beta e^{-\lambda} \approx 4.0$; the small proximity weight reflects the steepness of the $1/P$ term at low proximity rather than the unimportance of Data Gravity. Likewise, setting $(E_{base}, \eta, \kappa) = (4.82, 66.55, 3.675)$ in Equation (3) reproduces $E = (5.0, 15.0, 50.0)$ at the same coordinates, with the convexity condition $\kappa > 1$ satisfied. These coordinates and parameters are chosen for demonstration, not measured; the example establishes only that the discrete calibration of Table 3 is structurally consistent with, and recoverable from, the functional forms of Section 3.4, while fitting the same parameters to observed (A, R, P) trajectories remains the calibration task of Section 5.4. One identification caveat applies to any such inversion: X_{max} and (α, β, γ) are identified only up to a common multiplicative

scale, since replacing $(X_{max}, \alpha, \beta, \gamma)$ by $(cX_{max}, \alpha/c, \beta/c, \gamma/c)$ for any $c > 0$ leaves $X(\vec{I})$ pointwise unchanged. The values reported above are accordingly one admissible normalization rather than a unique solution, and an empirical fit must impose a normalization, for instance $\alpha + \beta + \gamma = 1$, before the individual weights admit a structural interpretation. The same tier-level values are also reproduced under a second, non-uniform coordinate assignment, $\vec{I}_1 = (0.20, 0.35, 0.25)$, $\vec{I}_2 = (0.45, 0.55, 0.60)$, and $\vec{I}_3 = (0.85, 1.00, 0.90)$, but with the materially different weight vector $(\alpha, \beta, \gamma) = (0.3206, 0.2876, 0.1169)$, environmental parameters $(E_{base}, \eta, \kappa) = (4.56, 66.59, 3.63)$, and an implied friction floor of $X_{min} \approx 7.8$, rather than 4.0, which stays strictly positive in both cases. Structural consistency is therefore a necessary condition rather than an identification result, and the individual weights admit a structural reading only once the (A, R, P) coordinates have themselves been measured. The relationship between the continuous functions of Section 3.4 and the discrete tier-level values of Table 3 is accordingly an assumed consistency, not a demonstrated identification: what this example demonstrates is only that at least one, and in fact more than one, (A, R, P) parameterization reproduces the assigned tier values, not that the tier values were derived from a uniquely recovered (A, R, P) trajectory.

Interpreting the normalized scale: Because only relative magnitudes are meaningful on this index, the calibration is best audited by reading off the ratios it implies for observable quantities. Three are directly interpretable. First, $K_{econ}/S = 0.5$ states that re-creating a dataset costs half of what was spent curating and provisioning it, which is consistent with re-creation reusing established protocols and instrumentation while repeating the collection itself. Second, $X_1/S = 1.5$ states that the labor and access burden of exercising a Tier-1 holding exceeds its capitalized epistemic value, which is what renders Tier-1 data dark in the sense of Principle 3 rather than merely expensive. Third, and most demanding, $E_3/K_{econ} = 1$ states that the energy burden attributable to one dataset in a Tier-3 sandbox over the evaluation horizon equals the full cost of re-creating that dataset. We flag this last ratio as the most aggressive assumption in Table 3 and state the direction of the resulting bias plainly: it charges Tier 3 the largest environmental penalty the framework can plausibly carry, and therefore biases the analysis toward the caution about Tier-3 centralization that Section 5 goes on to recommend. The Tier-3 warnings reported here should accordingly be read as an upper bound on the environmental case against centralization rather than as a calibrated estimate of it. The index unit can also be given an order of magnitude. A synthesis of economic studies of three UK research data centers reports annual operating budgets alongside annual acquisition counts for the same centers and years [33]: GBP 698,000 against 444 datasets acquired at one center, and GBP 3.3 million against 750 collections acquired at another. Because those budgets fund delivery and preservation as well as acquisition, the implied per-dataset annual expenditure, of order one to several thousand pounds (approximately GBP 1600 and GBP 4400 for the two centers), is an upper bound rather than an acquisition cost, and the two centers are reported there as separate studies rather than as a comparison. Read against $S = 100$, this places one index unit in the range of tens of pounds, so that the assumed recreation cost $K_{econ} = 50$ and the Tier-3 penalty $E_3 = 50$ each correspond to a per-dataset annual burden of order one thousand pounds. We report this as a magnitude check rather than a calibration, and none of the results above is recomputed from it. At the aggregate level, the European Commission estimate cited above bounds the same friction channel at a minimum of EUR 10.2 billion per year across European research, but converting that aggregate into a monetary value for one index unit additionally requires the number of affected dataset-years, which that study does not report. Fixing the index unit against measured per-dataset expenditure, energy, and re-creation costs is accordingly the first task of the empirical roadmap in Section 5.4.

4.2. Scenario A: The Impact of the Cross-Disciplinary Reuse Rate (ρ_{reuse})

A common critique of massive computational infrastructure is that its environmental cost ($E_3 = 50$) outstrips immediate economic gains. However, when applying our framework, the net benefit depends heavily on data reuse. If the advanced tier enables only a single incremental reuse event ($\rho_{reuse} = 1$), the avoided redundant cost is $(50 + 10) = 60$. The net premium over Tier 1 is marginal. If the Tier 3 environment enables broad cross-disciplinary reuse ($\rho_{reuse} = 3$), the avoided loss scales to 180, more than offsetting the 45 net energy penalty ($E_3 - E_1$). This structurally implies that, although the baseline reuse level already satisfies the $NEB > 0$ condition (the threshold ratio of 0.75 derived in Section 3), the environmental margin at $\rho_{reuse} = 1$ is thin; the case for public AI sandboxes therefore strengthens materially only when policies promote high data reuse rates.

4.3. Scenario B: Volatility Sensitivity Across Tiers

Table 4 illustrates how the Net Sustainability Value ($NSV = C - E$) behaves under varying levels of scientific volatility (σ). Note that NSV does not depend on ρ_{reuse} : the reuse channel enters the separate NEB account of Principle 3 and not the tier-level objective function, so the comparison below holds across the full range of reuse assumptions considered in Scenario A. To avoid spurious precision, output values are rounded to the nearest whole index unit. The three volatility levels examined ($\sigma = 0.2, 0.4, 0.8$) are an illustrative grid spanning low, medium, and high scientific uncertainty, not a calibrated standard deviation estimated from any specific research domain; Table 2 defines σ formally, and identifying empirical proxies for it is part of the calibration roadmap of Section 5.4.

Table 4. Boundary condition analysis: net sustainability value ($NSV = C - E$) under varying volatility. Within each tier block, the first column reports the option value C_i , and the second reports the resulting $NSV_i = C_i - E_i$, with E_i held at the value stated in the block heading.

Volatility (σ)	Tier 1 ($X_1 = 150, E_1 = 5$)		Tier 2 ($X_2 = 70, E_2 = 15$)		Tier 3 ($X_3 = 20, E_3 = 50$)	
	C_1	NSV_1	C_2	NSV_2	C_3	NSV_3
$\sigma = 0.2$ (Low)	0	−5	34	+19	81	+31
$\sigma = 0.4$ (Med)	5	0	36	+21	81	+31
$\sigma = 0.8$ (High)	20	+15	46	+31	81	+31

The structural evaluation reveals a nuanced policy insight. For highly stable data ($\sigma = 0.2$), the residual friction in Tier 1 renders the option value insufficient to cover the baseline energy penalty, resulting in a negative NSV_1 index of −5; a negative NSV indicates that the tier's option value does not cover its energy penalty, which we interpret as a waste-prone configuration rather than as an observed reuse outcome. At $\sigma = 0.4$, Tier 1 sits at the rounded break-even point ($NSV_1 \approx 0$). Transitioning to Tier 3 is contextually optimal in the operating-only account, before $\Delta Capex$ (Section 4.4), to prevent the data from devolving into a deadweight loss.

However, for highly volatile, disruptive data ($\sigma = 0.8$), the gross speculative value rises even under high friction ($C_1 \approx 20$), yielding an NSV_1 of +15; whether that value is realized through reuse rather than through redundant re-creation still depends on the friction-versus-recreation comparison of Principle 3, since X_1 remains above K_{econ} . Notably, in this high-volatility regime, the sustainability gap between Tier 2 (+31) and Tier 3 (+31) closes: the two tiers differ by less than one index unit before rounding, so the tabulated equality reflects genuine convergence rather than an artifact of the rounding policy. This indicates that, while Tier 3 remains scientifically vital for rapid processing, policymakers must carefully weigh whether the carbon expenditure is justified when Tier 2 already captures the sustainable option value.

4.4. Scenario C: The Total-Cost Tier-2/Tier-3 Indifference Boundary

To evaluate boundary conditions, we compare Tier 3 with Tier 2 under medium scientific volatility ($\sigma = 0.4$) and make the transition capital cost explicit. Let $\Delta\text{Capex}_{2 \rightarrow 3}$ denote the annualized incremental cost of upgrading from Tier 2 to Tier 3 on the same normalized per-dataset scale. Tier 3 is preferred on a total-cost basis only when its option-value gain covers both the environmental differential and this annualized upgrade cost:

$$\begin{aligned} C(S, X_3, \Theta) - E_3 - \Delta\text{Capex}_{2 \rightarrow 3} &\geq C(S, X_2, \Theta) - E_2 \\ \iff E_3 &\leq E_{3,\text{crit}}(\Delta\text{Capex}_{2 \rightarrow 3}) \\ &\equiv E_2 + C(S, X_3, \Theta) - C(S, X_2, \Theta) - \Delta\text{Capex}_{2 \rightarrow 3}. \end{aligned} \quad (6)$$

The sign is decisive: $\partial E_{3,\text{crit}} / \partial \Delta\text{Capex}_{2 \rightarrow 3} = -1$, so each additional unit of annualized upgrade cost lowers the admissible Tier-3 environmental penalty by one unit. Under the illustrative calibration, $E_2 = 15$ and the unrounded option-value gain is $C_3 - C_2 = 44.9608$, giving $E_{3,\text{crit}} = 59.9608 - \Delta\text{Capex}_{2 \rightarrow 3}$. The familiar value of 60 is therefore only the rounded, zero-capital-cost intercept. At $\Delta\text{Capex}_{2 \rightarrow 3} = 10$, the boundary falls to 49.9608, slightly below the assumed $E_3 = 50$, and the unrounded Tier-3 advantage reverses. No total-cost tier preference follows from the numerical illustration until a program-level $\Delta\text{Capex}_{2 \rightarrow 3}$ is supplied.

Figure 2 visualizes only the $\Delta\text{Capex}_{2 \rightarrow 3} = 0$ slice of Equation (6). In that slice, Tier 3 ceases to dominate Tier 2 once E_3 exceeds approximately 60. Positive transition cost moves the intersection left one-for-one. This relative boundary remains distinct from the absolute point at which Tier 3's operating NSV_3 turns negative, $E_3 = C_3 \approx 81$.

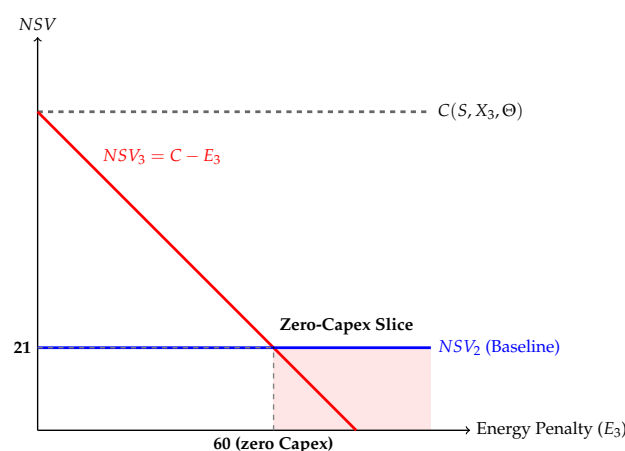


Figure 2. Zero-capital-cost slice of the Tier-2/Tier-3 total-cost boundary. At $\Delta\text{Capex}_{2 \rightarrow 3} = 0$, the intersection occurs at $E_3 \approx 60$ (i.e., $E_{3,\text{crit}}^{(0)} = E_2 + C_3 - C_2$, Equation (6), rounded from 59.9608) against the Tier-2 baseline $\text{NSV}_2 = 21$ (Table 4, $\sigma = 0.4$); a positive annualized upgrade cost shifts it left one-for-one according to Equation (6). Tier 3's stand-alone operating value remains positive up to $E_3 \approx 81$.

4.5. Robustness of the Calibration Thresholds

The numerical quantity examined in this subsection is the zero-capital-cost intercept $E_{3,\text{crit}}^{(0)} = E_2 + C_3 - C_2$, which equals approximately 60 under the baseline proxies of Table 3. It is not the complete adoption boundary: for any fixed annualized upgrade cost $k = \Delta\text{Capex}_{2 \rightarrow 3}$, Equation (6) gives $E_{3,\text{crit}}^{(k)} = E_{3,\text{crit}}^{(0)} - k$. Because the proxy inputs are literature-informed approximations rather than measured quantities, even the intercept risks conveying spurious precision. We therefore stress-test $E_{3,\text{crit}}^{(0)}$ against plausible variation in the four inputs that determine it (S, X_2, X_3, E_2), holding $\sigma = 0.4$ as in Scenario C. This exercise quantifies uncertainty in the operating-value intercept; it neither estimates nor validates $\Delta\text{Capex}_{2 \rightarrow 3}$.

We first conduct a one-at-a-time sensitivity analysis: Table 5 perturbs each input individually by $\pm 20\%$, a range consistent with the width of the empirical ranges cited in Section 4.1 (e.g., the 60–80% data-wrangling overhead underlying X_1 , and the PUE differential underlying E_i).

Table 5. One-at-a-Time Sensitivity of the Zero-Capital-Cost Intercept $E_{3,\text{crit}}^{(0)}$ to Calibration Inputs ($\sigma = 0.4$). The Baseline column reports the common, unperturbed intercept $E_{3,\text{crit}}^{(0)} = 59.9608$ (Section 4.4), rounded to 60.0 here at the one-decimal precision used for the perturbed entries; it is the same reference value in all four rows, not a separate estimate per input.

Perturbed Input	−20%	Baseline	+20%
Tier 2 Friction ($X_2 = 70$)	48.5	60.0	69.5
Tier 3 Friction ($X_3 = 20$)	63.8		56.2
Tier 2 Energy Penalty ($E_2 = 15$)	57.0		63.0
Epistemic Value ($S = 100$)	56.5		61.5

Across the four one-at-a-time perturbations, $E_{3,\text{crit}}^{(0)}$ ranges from 48.5 to 69.5, a deviation of up to 20% from its baseline value of 60. Under a fixed positive k , every value in the table falls by exactly k ; for example, $k = 10$ shifts the range to 38.5–59.5. In seven of the eight zero-capital-cost perturbations, the intercept remains above the baseline energy penalty $E_3 = 50$; the exception is the downward perturbation of Tier-2 friction ($X_2 = 56$), which pulls it down to 48.5. The intercept is comparatively insensitive to the epistemic-value proxy itself ($S \pm 20\%$ moves it by only -5.8% and $+2.5\%$), because C_3 and C_2 co-move with S and largely cancel in the difference. These comparisons describe operating-value sensitivity only. A total-cost tier preference still requires the program-level k that shifts the entire boundary.

We then apply a joint Monte Carlo stress test that perturbs all six calibration inputs ($X_1, X_2, X_3, E_1, E_2, E_3$) simultaneously: each input is drawn independently and uniformly within $\pm 20\%$ of its baseline, except X_1 , which is drawn uniformly on $[112.5, 150]$, the interval obtained by mapping the cited 60–80% wrangling-overhead range onto the index scale (Section 4.1). Under this range, the share of draws in which X_1 exceeds the Vega-maximizing point $X^* = Se^{(r+\sigma^2/2)T} \approx 113.88$ (Section 3) is $(150 - 113.88)/37.5 \approx 96.3\%$, so the placement of Tier 1 on the high-friction side of X^* (that is, $X_1 > X^*$, beyond the Vega peak) assumed in Section 3 holds for approximately 96% of draws. This share is an analytical proportion of the uniform draw interval rather than a sampled frequency (the seed-7 sample gives 96.5%), and it is specific to $\sigma = 0.4$ with S held at its baseline of 100: because X^* rises with σ , the same interval gives essentially 100% at $\sigma = 0.2$ ($X^* \approx 107.3$) but only about 14% at $\sigma = 0.8$ ($X^* \approx 144.8$), and when S is also perturbed by $\pm 20\%$ (the seven-input variant below) the share falls to approximately 83%. The Tier-1 placement beyond the Vega peak is therefore a property of the moderate-volatility calibration, not a general feature of the model. The model's required ordering ($X_1 > X_2 > X_3$ and $E_1 < E_2 < E_3$) is preserved automatically under these ranges, because the resulting draw intervals do not overlap; the accept-reject safeguard implemented in the simulation code therefore never binds in this design and is retained only to protect the ordering under future recalibration. Drawing the friction and energy inputs independently is a deliberately conservative choice rather than an inconsistency with Section 3.4, where both are functions of the same vector $\vec{I}$ and are therefore negatively dependent. Imposing that common dependence would tie a downward perturbation of X_2 to an upward perturbation of E_2 , and since $E_{3,\text{crit}}^{(0)}$ is increasing in both inputs, as Table 5 confirms, those two moves push the intercept in opposite directions and would partially cancel. The independent design is therefore expected to report a wider uncertainty envelope for the intercept than a design imposing that common dependence would, which is the direction that favors caution in reading the intercept; we state this as the expected direction of the difference rather than as a proven bound. We

note that $E_{3,\text{crit}}^{(0)}$ itself is a function of (S, X_2, X_3, E_2) ; among the six jointly perturbed calibration inputs only (X_2, X_3, E_2) enter it, while S is held at its baseline of 100 in the joint test, a choice supported by the mild S -sensitivity documented in Table 5. As a direct check on this choice, a complementary variant that additionally draws S uniformly within $\pm 20\%$ of its baseline (seven inputs perturbed jointly) leaves the intercept distribution essentially unchanged: the mean shifts from 59.6 to 59.2, the median from 59.9 to 59.3, the 90% interval from $[48.5, 70.0]$ to $[48.2, 70.4]$, and the share of draws with the intercept above $E_3 = 50$ from 91.5% to approximately 90%. The remaining inputs are perturbed jointly because they enter the accompanying tier-level NSV comparisons reported below. Over 20,000 draws, the zero-capital-cost intercept has a mean of 59.6, a median of 59.9, and a 90% interval of $[48.5, 70.0]$ (full range $[42.3, 75.6]$); these summary statistics are stable across independent random seeds, varying by less than 0.5 index units. The baseline value of 60 sits almost exactly at the center of this zero-capital-cost distribution rather than at an extreme. In 91.5% of draws, the intercept exceeds the fixed baseline penalty $E_3 = 50$, and in approximately 85% it exceeds that draw's perturbed E_3 ; both shares are operating-only, zero-capital-cost comparisons, not total-cost preference probabilities. For any fixed $k = \Delta Capex_{2 \rightarrow 3}$, the entire intercept distribution shifts left by k , and the corresponding preference shares must be recomputed rather than carried over.

We repeated the same joint perturbation for the low-volatility regime of Scenario B ($\sigma = 0.2$): the operating-only finding that Tier 1 exhibits a negative NSV_1 survives in 96.7% of draws, and Tier 3 has the highest operating NSV in 88.2% of draws. Taken together, these checks support treating the numerical quantities reported in Sections 4.2–4.4 as zero-capital-cost intercepts within a moderately wide envelope, not as complete operational cutoffs or total-cost tier rankings. Any funding decision requires empirically measured inputs and a program-level $\Delta Capex$, as discussed in Section 5.4. This Monte Carlo exercise is a deterministic sensitivity analysis over independently drawn calibration inputs rather than a controlled experiment; its summary statistics are reported above and the full distribution in Figure 3, and the draw-generating procedure, including fixed random seeds, is archived in the Zenodo replication package referenced in the Data Availability Statement.

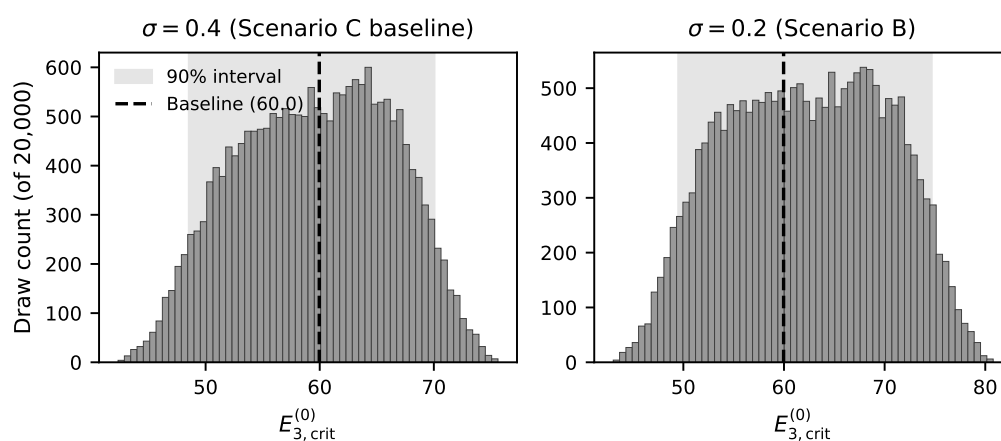


Figure 3. Distribution of the zero-capital-cost intercept $E_{3,\text{crit}}^{(0)} = E_2 + C_3 - C_2$ over the joint Monte Carlo stress test described above ($N = 20,000$ draws, seed fixed at 7). The **(left panel)** is the $\sigma = 0.4$ series underlying the summary statistics reported in the text; the **(right panel)** repeats the same exercise under the low-volatility regime of Scenario B ($\sigma = 0.2$) for comparison. The dashed vertical line in both panels marks the unrounded $\sigma = 0.4$ baseline intercept (59.9608), drawn as a common reference; the $\sigma = 0.2$ panel's own zero-capital-cost intercept is 62.44. The shaded band marks each panel's own 90% interval. This is a deterministic sensitivity analysis over calibration inputs, not a controlled experiment: the histograms display how the intercept varies across independently drawn combinations of the calibration proxies, not measurement variability in an observed quantity. The dark grey columns are the histogram bars themselves, each giving the number of draws that fall in that intercept bin.

4.6. A Minimal Two-Period Extension: A First Step Toward the Reinforcing Loop

The static comparisons leave the dotted feedback path in Figure 1 outside the model. Although Section 5.4 identifies that path as a theoretical extension, a first step toward it can be taken here in a minimal form: reuse feeds back into the same tier's next-period state, while the cross-tier return into the archival base remains outside the model. This subsection introduces a two-period state transition in which reuse changes next-period governance readiness and the opportunity set, and examines how the zero-capital-cost intercept moves under that extension. Its purpose is to report conditional feedback comparative statics, not to establish a tier preference without transition-cost data.

The extension retains accessibility A and proximity P as infrastructure-investment choice variables and elevates only governance readiness R to an endogenous state variable. Because the period-one coordinates are identical to the worked example in Section 4.1, the period-one results reproduce the static results of Sections 4.2–4.5 to the precision reported there. The endogenous reuse rate is specified as

$$\rho_i = \bar{\rho} \Phi(d_2(X_i)). \quad (7)$$

This is an assumed, admissible reduced form rather than a derived result: it provides a monotonic connection in which greater moneyiness is associated with a greater reuse propensity, using the risk-neutral exercise probability $\Phi(d_2)$ as its proxy. The conservative baseline sets $\bar{\rho} = 1$. This ρ_i is distinct from ρ_{reuse} in Section 3, which denotes an incremental number of reuse events and includes scenarios with $\rho_{reuse} \geq 1$; by contrast, ρ_i is a probability-scale quantity satisfying $\Phi(d_2) \leq 1$. Learning and opportunity creation then operate through

$$R'_i = \min\{1, R_i + \delta \rho_i (1 - R_i)\}, \quad \delta \in [0, 1], \quad (8)$$

and

$$N'_i = 1 + \nu \rho_i, \quad \nu \geq 0. \quad (9)$$

Both transition equations are illustrative functional forms chosen for tractability rather than empirically established laws: the readiness update is a bounded linear learning rule and the opportunity update a linear scaling in the reuse propensity, and the conclusions drawn below are conditional on these choices (Section 5.4). The two-period objective is therefore

$$NSV_i^{2p} = [C(S, X_i, \Theta) - E_i] + e^{-rT} [N'_i C(S, X'_i, \Theta) - E_i], \quad (10)$$

where $X'_i = X(A_i, R'_i, P_i)$ and the energy penalty is incurred in each period. To preserve the total-cost boundary in the dynamic extension, let K_i denote the annualized capital burden of tier i on the same per-period scale and define $\Delta Capex_{2 \rightarrow 3} = K_3 - K_2$. The corresponding two-period total value is

$$TV_i^{2p} = NSV_i^{2p} - (1 + e^{-rT}) K_i. \quad (11)$$

Because energy and annualized capital burden enter with the same two-period factor, every Tier-2/Tier-3 intercept reported below falls one-for-one with $\Delta Capex_{2 \rightarrow 3}$. Black–Scholes prices a single-exercise European option, so the research opportunities in each period are treated as distinct European options on the same underlying asset and N'_i scales their number. This treatment does not resolve the tension between single exercise and repeated reuse identified in Section 5.4; it states that tension in a minimal form, while a complete treatment with optimal stopping remains a future task.

Under the illustrative calibration with $\bar{\rho} = 1$ and $\sigma = 0.4$, the endogenous rates are $\rho_1 = 0.14$, $\rho_2 = 0.79$, and $\rho_3 = 1.00$. These are probability-scale quantities bounded

above by 1, and they are not directly comparable in magnitude to the incremental-count threshold $\rho_{reuse} > 0.75$ derived in Section 3, since ρ_{reuse} counts incremental reuse events and is evaluated at $\rho_{reuse} \geq 1$ in the static scenarios. The comparison below therefore reads ρ_i only as an ordinal indicator of which tier's endogenous reuse propensity sits closest to saturation, not as a literal check against the 0.75 event-count threshold: on that ordinal reading, Tier 3's rate is closest to its own upper bound of 1, Tier 2's is substantially lower at 0.79, and Tier 1's is lowest at 0.14, consistent with the tier ordering of exercise friction. The exogenous baseline $\rho_{reuse} = 1$ adopted as a conservative reference in Section 3 remains a separate, incremental-count assumption and is not implied by, or interchangeable with, the probability-scale ρ_i values reported here.

The no-transition limit, $\delta = \nu = 0$, yields a zero-capital-cost dynamic intercept of 59.96, which agrees with the value of 60 reported in Section 4.4 under the rounding policy of Table 4. The value 60 follows from rounded $C_2 = 36$ and $C_3 = 81$, whereas 59.96 uses the unrounded option values. The extension thus nests the operating-only static model as a special case. Table 6 reports zero-capital-cost intercepts over the full grid of the two transition channels. With the learning channel alone, setting $\nu = 0$, the intercept falls from 59.96 to 58.79 as δ rises from 0 to 1, because Tier 2 has greater remaining improvement scope through $1 - R$ and consequently receives the larger learning gain. The small magnitude of this learning effect is a property of the calibration: the readiness weight $\beta = 0.1481$ in Equation (2) bounds the contribution of the readiness term to total friction, so readiness improvements can only reduce friction by a limited amount. Programs with a larger readiness weight would accordingly have larger dynamic gains. Once the opportunity-creation channel is activated, the zero-capital-cost intercept rises to approximately 65–66 for $\nu = 0.25$ and approximately 83–86 for $\nu = 1.0$. These are opposing feedback comparative statics, not evidence that 60 is generally conservative: readiness learning lowers the intercept, opportunity creation raises it, and any fixed $\Delta Capex_{2 \rightarrow 3}$ lowers every cell by that amount.

Table 6. Zero-capital-cost dynamic Tier-2/Tier-3 intercepts under the learning and opportunity-creation channels. Static zero-capital-cost benchmark = 60 (Section 4.4); subtract $\Delta Capex_{2 \rightarrow 3}$ from each cell for the total-cost boundary.

$\delta \backslash \nu$	0.00	0.25	0.50	1.00
0.00	59.96	66.35	72.74	85.52
0.25	59.59	65.91	72.23	84.87
0.50	59.27	65.54	71.80	84.32
0.75	59.01	65.23	71.44	83.87
1.00	58.79	64.97	71.15	83.50

The sequencing comparison starts from Tier 2 and considers remaining at Tier 2 in both periods, investing in Tier 3 from period one, and a staged transition that remains at Tier 2 in period one before moving to Tier 3 in period two. In the staged strategy, readiness carries forward from the Tier-2 learning state in period one. At the zero-capital-cost calibration, with $\delta = 0.5$ and $\nu = 0$, the corresponding two-period NSV functions of E_3 are 42.77 for remaining at Tier 2, $158.42 - 1.95 E_3$ for investing in Tier 3 from period one, and $96.59 - 0.95 E_3$ for the staged transition. Their pairwise intersections occur at $E_3 = 56.58$ for always Tier 2 and staged, $E_3 = 59.27$ for always Tier 2 and immediate Tier 3, and $E_3 = 61.83$ for staged and immediate Tier 3. The staged interval has no positive length on the upper envelope, so the staged transition is dominated in this operating-only comparison. The comparison can also be written in closed form. Let C_i denote the period-one option value at tier i , let C'_i denote the period-two option value when tier i is retained,

and let C'_{staged} denote the period-two option value at Tier-3 coordinates evaluated at the readiness accumulated under Tier 2. Staged then exceeds immediate Tier 3 only when $(E_3 - E_2) - (C_3 - C_2) + e^{-rT}(N'_2 C'_{staged} - N'_3 C'_3) > 0$, so the period-one energy saving must exceed the forgone period-one option value together with the present value of the period-two difference. Its left-hand side is -11.83 at the baseline. Staged exceeds the always-Tier-2 strategy when $e^{-rT}[E_2 + N'_2(C'_{staged} - C'_2) - E_3] > 0$, whose left-hand side is $+6.26$ at the baseline. A staged strategy is optimal only if an E_3 interval satisfies both conditions, and no such interval exists at the baseline. Across the 5×4 grid of 20 points, a strictly optimal staged interval appears at 10 points, all with $\nu > 0$; the $\nu = 0$ column remains dominated for every value of δ . For example, staged is optimal for $E_3 \in [70.59, 73.07]$ when $\delta = 0.5$ and $\nu = 0.5$, and for $E_3 \in [79.35, 89.56]$ when $\delta = 0.5$ and $\nu = 1.0$. These windows describe operating-only sequencing. Once capital burdens are introduced, each strategy's value changes according to its timing and tier occupancy, so a program-level cost path is required before any baseline sequencing recommendation follows.

The parameters δ , ν , and $\bar{p}$ are not measured values but free parameters introduced by this extension, so the results are reported over the full grid rather than as a single point. The two-period construction is the minimal form that renders the reinforcing loop a model object; infinite-horizon dynamics, a formal optimal-stopping problem for the transition time, and endogenous accumulation of accessibility and proximity remain tasks identified in Section 5.4. The extension supports only a conditional statement: the feedback loop can be represented as an explicit state transition, readiness learning and opportunity creation move the zero-capital-cost intercept in opposite directions, and the total-cost boundary additionally depends on the unmeasured $\Delta Capex_{2 \rightarrow 3}$. All numerical values in this subsection are deterministically regenerated by the accompanying replication script.

5. Discussion

5.1. Theoretical Contributions and Public R&D Governance

The primary contribution of this study is the formalization of the Architecture-Aware Real Options Framework (AA-ROF). By endogenizing both exercise friction and environmental penalties into the infrastructure state space, AA-ROF bridges the economic theory of IT investment with the systems-level governance of global research data initiatives such as EOSC and NAIRR. Methodologically, the contribution is a way of representing a socio-technical system in which institutional readiness and technical proximity are priced jointly: the friction function $X(\vec{I})$ is the interface at which organizational and technical states are converted into a single economic quantity. Because that interface is defined by the functional forms of Equations (2) and (3) rather than by anything specific to research data, the same construction transfers to other multi-tier infrastructures in which access friction falls and energy burden rises along a common architectural gradient, such as tiered HPC allocation hierarchies or scientific data transport networks; only the interpretation of (A, R, P) changes.

To state this contribution in the vocabulary of systems science rather than of option pricing alone: what is new is not that real options can price an uncertain asset, nor that Principles 1 and 2 are, formally, the standard strike-monotonicity and deep-in-the-money Vega-compression properties of call options (Section 3); what is new is the systems-level coupling that composes those standard properties with the architectural state through Equations (2) and (3). That coupling assembles five elements that a conventional real-options model, which takes its strike price and volatility as exogenously observed, cannot represent on its own: (i) interacting infrastructure layers, in which accessibility, governance readiness, and computing proximity enter a single priced friction quantity through Equation (2); although that function is additively separable in its arguments, their effects on realizable value interact through the nonlinearity

of the option value C in X , so no layer's contribution to realizable value can be assessed independently of the others (Section 3.3); (ii) an explicit reuse-readiness feedback channel, formalized in the two-period extension through the opportunity-set transition $N'_i = 1 + \nu\rho_i$ (Equation (9)), together with the readiness transition of Equation (8); (iii) endogenous exercise friction, in which the option's strike price is itself a function of the architecture rather than an externally observed market quantity; (iv) an environmental outcome, $E(\vec{I})$, priced jointly with the same architectural state that sets friction, so the two cannot be optimized independently of one another; and (v) an evolving research-opportunity set, in which the number of exercisable opportunities responds to each tier's reuse propensity rather than remaining fixed. The systems-science contribution is this joint architecture, not any one of its five components taken in isolation, each of which has analogues elsewhere in the real-options or systems literatures; it is what distinguishes AA-ROF from a standard real-options model in which infrastructure enters only through exogenously observed parameters plugged into an otherwise unchanged calculus.

From a policy perspective, AA-ROF provides a conceptual baseline for public R&D planners. It challenges the prevailing technological absolutism that all scientific data must eventually migrate to highly integrated AI Sandboxes. Instead, it offers an analytical justification for selective orchestration: allocating power-intensive GPU resources only when the marginal option value satisfies the total-cost boundary after both environmental and annualized transition costs are counted (Section 5.3). Furthermore, by linking infrastructure tiers to public institution performance metrics (e.g., data reuse rates, avoidance of redundant national R&D spending), the framework offers evaluation indicators for science policy administrators.

5.2. AA-ROF as a Lens for Global R&D Infrastructure Governance

Having formalized the tier structure and the friction and environmental functions in Section 3, we can now use AA-ROF's vocabulary as a comparative analytical lens for reading contemporary national infrastructure initiatives. We emphasize at the outset that the following mappings are conceptual illustrations of the framework's applicability, not empirical evaluations of the programs named; an assessment of that kind would require program-level data on friction, cost, and carbon intensity that lies outside the scope of this paper.

- **United States (NAIRR):** According to its own implementation plan, NAIRR is designed to provide researchers with expanded, direct access to compute, data, and AI models [34]. In AA-ROF terms, this stated aim is broadly consistent with a strategy of optimizing specific components of friction, particularly by raising computing proximity P , which lowers friction through the $\gamma(1/P - 1)$ term of Equation (2) (that is, $\partial X/\partial P < 0$); we offer this mapping as an illustrative interpretation of NAIRR's publicly stated goals, not as an independently verified assessment of its realized friction levels; as with the other programs mapped in this subsection, we do not claim that NAIRR, EOSC, Gaia-X, or any other named initiative formally endorses, adopts, or implements the AA-ROF framework.
- **Europe (EOSC and Gaia-X):** EOSC's Strategic Research and Innovation Agenda explicitly frames the initiative as building a federated "Web of FAIR data and services" rather than a single centralized repository, and Gaia-X similarly emphasizes federated, sovereignty-preserving data sharing [35]. This stated federated design is broadly consistent with a Tier-2-oriented strategy in AA-ROF terms (raising accessibility A through interoperability, which lowers friction through the $\alpha(1 - A)$ term of Equation (2) (that is, $\partial X/\partial A < 0$), while deferring the environmental cost (E) of full centralization), though, again, we do not have independent data on EOSC's or Gaia-X's actual friction or carbon

metrics, and this mapping should be read as illustrative rather than as an empirical finding about these programs.

- **Illustrative extension (HPC-intensive economies):** More speculatively, export-driven, computing-intensive economies that pair world-class HPC facilities with large-scale scientific datasets (e.g., Japan’s Fugaku system, or national public-led distributed scientific AI foundation model initiatives in South Korea) could in principle be analyzed as pursuing a Tier-3-oriented strategy. We flag this case explicitly as an unverified illustration: we are not aware of published, program-level friction or carbon-accounting data for these specific national systems, and the characterization here should not be read as an empirical claim about their actual operation.

These mappings are offered to illustrate how AA-ROF’s vocabulary could, in principle, be used to compare national R&D infrastructure strategies. We return to this comparative lens operationally in the next subsection, where it is translated into concrete performance indicators.

5.3. AA-ROF as a Policy Evaluation Tool: Performance Metrics

AA-ROF shifts the paradigm of public R&D infrastructure evaluation from simple “Capacity Utilization” to “Value-Generation Potential.” We propose the following guidelines for public institutions to integrate AA-ROF into their performance metrics and PFS frameworks:

1. **From Utilization Rate to Option-Creation Index (OCI):** Traditional metrics emphasize machine utilization (e.g., GPU hours). AA-ROF redefines success through the *OCI*, defined as the number of cross-disciplinary combinatorial analyses performed per dataset over the evaluation horizon, normalized by the counterfactual number of redundant re-creations that the same dataset would have induced under Tier-1 friction. The denominator is a fixed counterfactual benchmark rather than a realized count, so that averting more redundant experiments raises the index instead of depressing it, and the ratio is defined only where that benchmark is strictly positive. An infrastructure is deemed successful if it facilitates a measurable increase in cross-disciplinary data integration relative to that benchmark.
2. **Sustainability-Adjusted Return on Investment (S-ROI):** Public R&D investment should be evaluated by the Net Sustainability Value (*NSV*). Policymakers should calibrate their infrastructure tiering based on the *Redundant Experiment Avoidance (REA) Ratio*, directly linking our theoretical circularity parameters to policy evaluation:

$$REA = \frac{\rho_{reuse} \cdot (K_{econ} + p_c K_{carbon})}{E(\vec{I}) + X(\vec{I})}$$

$REA \geq 1$ indicates that avoided redundant activity alone covers the tier’s environmental and friction cost; how this signal enters the adoption decision is stated as a two-stage rule below.

The three sustainability quantities used in this paper form one coherent hierarchy rather than competing criteria. $NSV_i = C_i - E_i$ is the operating-only tier value used in Scenarios B–C; *NEB* isolates the incremental reuse channel of a Tier-1 to Tier-3 transition (Principle 3); and *REA* is a screening ratio that asks whether reuse benefits alone cover the full structural burden ($E + X$) without crediting any option value C . Its conservatism is exact only at the baseline $\rho_{reuse} = 1$, and this must be stated plainly: the numerator scales with ρ_{reuse} while the friction term in the denominator is charged once, so the ratio becomes progressively more permissive as reuse rises. Charging friction per reuse event instead, that is $\rho_{reuse}(K_{econ} + p_c K_{carbon}) / (E + \rho_{reuse}X)$, leaves the baseline value unchanged at $60/70 \approx 0.86$ but yields 1.64 rather than 2.57 at $\rho_{reuse} = 3$. *REA* should therefore be read as a screen at the baseline reuse level,

and the per-event variant is the stricter test where repeated access costs are material. Because *REA* ignores *C* while *NSV* is built on it, the two can disagree: at the baseline calibration with $\rho_{reuse} = 1$, Tier 3 has $NSV_3 \approx 31 > 0$ and $NEB = 15 > 0$, yet $REA = 60/70 \approx 0.86 < 1$. Such a configuration is value-positive but reuse-dependent. A failed *REA* screen flags that the sustainability case leans on option value rather than on demonstrated reuse, calling for the reuse-enforcing policies discussed in Scenario A. Note also that *X* appears in the *REA* denominator as a residual friction burden and, separately, inside *C* as the strike price; the two roles are not additive, because *REA* is a stand-alone screening ratio, not a component of *NSV*. The complete adoption rule is instead Equation (6), equivalently $AP_{i \rightarrow j} \geq \Delta Capex_{i \rightarrow j} + (E_j - E_i)$, where $\Delta Capex_{i \rightarrow j}$ denotes the annualized incremental capital burden of upgrading from tier *i* to tier *j*. The numerical stress tests set $\Delta Capex = 0$ and therefore characterize only the intercept of this rule. For Tier 2 to Tier 3, the exact intercept 59.9608 becomes 49.9608 when $\Delta Capex_{2 \rightarrow 3} = 10$; at the assumed $E_3 = 50$, the unrounded margin changes sign. Accordingly, none of the reported operating-only rankings establishes a total-cost tier preference without program-level transition-cost data.

For practitioners facing the conflicting signals illustrated above, we state the screening logic as two sequential stages rather than as a single composite score. Stage 1 (net-benefit screen) applies the complete adoption rule of Equation (6): a candidate transition $i \rightarrow j$ is admissible only if $AP_{i \rightarrow j} \geq \Delta Capex_{i \rightarrow j} + (E_j - E_i)$, and a transition failing this test is rejected regardless of *REA*, because *NSV* and *AP* already price the option value that *REA* deliberately excludes from its numerator. Stage 2 (reuse-dependence flag) applies only to transitions that clear Stage 1, and uses *REA* for the narrower question the baseline example above answers: whether the admissible case is self-sufficient on demonstrated reuse alone ($REA \geq 1$) or instead leans on projected option value that has not yet been realized through reuse ($REA < 1$), which is useful for ranking among several admissible candidates under a binding budget or for triggering the reuse-enforcing policies of Scenario A. *NEB* is reported alongside these two stages as the environmental account of the transition, but it does not enter either stage as a separate test, because the environmental penalty difference ($E_j - E_i$) that drives it already appears in the Stage-1 boundary. We do not adopt a lexicographic rule that screens on *REA* before *NSV* and *AP*: at the illustrative baseline, $REA \approx 0.86 < 1$ even though Tier 3 is net-benefit positive and clears Stage 1, so an *REA*-first rule would reject a transition that Equation (6) accepts. Appendix B restates this two-stage procedure as an explicit step-by-step algorithm.

Applied Scenario in PFS: Consider a proposed national initiative investing \$10 million to build a centralized Tier-3 AI Sandbox for materials science data. If the historical data reuse frequency is low (e.g., datasets are rarely queried across disciplines) and the estimated carbon penalty (E_3) exceeds the value of avoided redundant experiments, the calculated *REA* ratio will fall below 1, and the annualized share of the capital outlay additionally enters the adoption rule above. Under the AA-ROF logic, a pre-feasibility evaluation panel could use this indicator as one input for recommending against the Tier-3 centralization proposal, redirecting the budget toward a more carbon-efficient Tier-2 Federated Mesh until metadata maturity (*R*) and cross-disciplinary demand reach a justified threshold. We stress that this is a hypothetical, illustrative use case rather than a validated decision procedure.

3. **Governance Maturity as a Key Performance Indicator:** Maturity indicators (FAIR compliance, provenance tracking) are, under this framework, more than administrative overhead: AA-ROF suggests treating them as candidate *Key Performance Indicators* for infrastructure budget allocation. Within this analytical logic, institutions that do

not meet a threshold readiness component, R , would not be well served by Tier 3 high-performance computing allocations, since the high environmental penalty E_3 is harder to justify without high-readiness data. We frame this as a decision-support heuristic for funding agencies to weigh alongside other criteria, not as a categorical eligibility rule; translating it into an actual allocation policy would require the empirical calibration of R and E discussed in Section 5.4.

By institutionalizing these metrics, funding agencies can move toward a more strategic, performance-oriented governance of national R&D assets, ensuring that public investment in high-tier AI sandboxes is screened on net value at Stage 1 and held accountable to demonstrated data reuse and governance maturity through the Stage-2 flag and the reuse-enforcing policies it triggers.

5.4. Limitations and the Empirical Roadmap

First, a theoretical limitation of the AA-ROF model is the foundational assumption that scientific uncertainty follows a continuous geometric Brownian motion. In reality, scientific breakthroughs often exhibit discrete, discontinuous jumps, suggesting that jump-diffusion or compound option models might more accurately capture empirical epistemic volatility. However, the standard Black–Scholes framework was deliberately selected in this conceptual baseline to provide analytical tractability and isolate the directional mechanism of how infrastructural friction impacts value. Future theoretical extensions should adapt this framework to accommodate jump-diffusion processes for highly disruptive fields. This restriction also bounds Principle 2: under a jump-diffusion or stochastic-volatility process the deep-in-the-money compression of Vega need not persist, because jump risk or a stochastic variance factor keeps option value sensitive to uncertainty even where the diffusive component is nearly deterministic. The compression result should therefore be read as a property of the geometric Brownian baseline rather than as a general feature of low-friction architectures. Principle 1 stands on a different footing: it rests only on strike monotonicity, $\partial C / \partial X < 0$, a property that holds for essentially any arbitrage-free option-pricing model with a payoff that is monotonic in the strike, not only under geometric Brownian motion. We therefore expect the qualitative claim that reducing architectural friction raises option value to survive alternative volatility processes, even where the specific compression magnitude reported for Principle 2 does not.

While this framework establishes robust boundary conditions, its numerical application remains an analytical scenario, and Section 4.5 shows that the reported thresholds are stable within a moderate range but not immune to recalibration. To transition from a theoretical framework to an operational policy tool, future research must execute the following empirical roadmap:

1. **Calibration of Friction (X):** Future studies should quantify exercise friction using actual data-wrangling labor hours and network latency logs extracted from national data repositories and, going beyond the tier-level proxies used here, should fit the continuous parameters $(\alpha, \beta, \gamma, \lambda)$ of Equation (2) directly against measured (A, R, P) trajectories, so that the discrete X_1, X_2, X_3 values used in Section 4 become derived quantities rather than independently assigned proxies.
2. **Quantification of Environmental Costs (E):** The environmental penalty must be grounded using precise PUE metrics and GPU workload logs (e.g., SLURM scheduler data) from public supercomputing centers, similarly fitting (E_{base}, η, κ) in Equation (3) rather than assigning E_1, E_2, E_3 independently. The same measurements would support replacing the univariate proximity index of Equation (3) with an explicit dependence on served workload and utilization alongside facility efficiency and grid carbon intensity, a decomposition the present reduced form deliberately suppresses.

The present monotonic specification also omits data-transmission energy, which increases as computing proximity declines and would therefore raise E_2 relative to E_3 . Incorporating that component would raise rather than lower the zero-capital-cost intercept of Section 4.4, all else equal. A further boundary limitation is that E captures operational energy only: life-cycle assessments show that most emissions attributable to modern data center equipment arise from hardware manufacturing and infrastructure rather than from operation [36], so extending $E(\vec{I})$ to include embodied emissions, which would raise the environmental penalty of hardware-intensive higher tiers, is a priority extension for the framework's sustainability accounting.

3. **Domain-Specific Volatility (σ):** Defining scientific volatility requires the development of empirical proxies, such as variance in DOI citation networks, downstream grant-funding, or keyword burstiness specific to individual scientific disciplines.
4. **Validating the Expenditure-Value Proxy (S):** As flagged in Section 3, the MAD assumption substitutes public R&D expenditure for intrinsic epistemic value. This identifying assumption should itself be tested empirically, for instance, by checking whether expenditure-based S proxies correlate with independent, ex-post measures of realized scientific value such as citation impact or downstream technology transfer, before the framework's option values are used to justify specific funding allocations.
5. **Capital Costs of Tier Transitions ($\Delta Capex$):** Equation (6) includes annualized incremental transition cost symbolically, but the numerical illustration sets it to zero because no program-level estimate is available. Operational use therefore requires annualized capital and fixed-cost data on the same per-dataset, per-horizon scale as the modeled flow quantities; without those data, the reported values are intercepts rather than complete adoption thresholds.
6. **Endogenizing the Opportunity Set:** The present model channels all architectural effects through the strike price $X(\vec{I})$. A theoretical extension should let the size, composition, or arrival rate of the research opportunity set itself depend on $\vec{I}$, so that combinatorial option creation becomes a formal model object rather than an economic interpretation of friction reduction, and should reconcile the single-exercise option object with the non-rival, repeated-reuse accounting of *NEB* (including per-event access costs and the one-to-one avoidance counterfactual assumed in Principle 3). The same extension is what would formalize the return path of Figure 1: making the infrastructure state at one period a function of reuse realized in the previous one would convert the reinforcing loop described in Section 3.1 from a qualitative architectural property into a dynamic system with explicit state transitions. Section 4.6 takes a first step toward that conversion in a minimal two-period form with within-tier feedback only; what remains open is the cross-tier return path itself, the infinite-horizon version, a formal optimal-stopping treatment of the transition time, and endogenous accumulation of accessibility and proximity alongside readiness.
7. **Option Style and Early Exercise:** Research data can be reused at any point within the evaluation horizon, giving the reuse option an American character, whereas Black-Scholes prices a European option. For a call option with no interim payoffs prior to maturity, however, it is a standard result that the American value equals the European value, because early exercise of such a call is never optimal; option style by itself therefore introduces no bias into the reported C_i . A bias arises only when reuse generates an interim payoff before the horizon ends, analogous to a dividend paid by the underlying asset, in which case each tier carries its own non-negative early-exercise premium $a_i \geq 0$ and the American Architecture Premium becomes $AP_{i \rightarrow j}^{American} = AP^{European} + (a_j - a_i)$. The sign of this correction to AP then depends on the relative size of the intermediate reuse payoffs accruing to tier j versus tier i , and is conditional rather than determined

in advance by the framework. A conditional direction can nonetheless be stated: if the interim reuse payoff of a tier scales with its reuse propensity, so that the calibrated rates $\rho_1 = 0.14$, $\rho_2 = 0.79$, and $\rho_3 = 1.00$ of Section 4.6 proxy the ordering of a_i , then $a_3 \geq a_2 \geq a_1$, the correction $(a_j - a_i)$ is non-negative for every upward transition, and the European AP reported here understates the Tier-2 and Tier-3 gains. This direction rests on the proxy assumption that interim payoffs are ordered like reuse propensities, which the framework does not establish, and we therefore report it as a conditional statement rather than a result. The European form is retained here for analytical transparency; binomial-tree or least-squares Monte Carlo methods are the natural extension for pricing early exercise.

6. Conclusions

This study addresses a fundamental gap in the economics of digital science: the lack of a dynamic valuation framework that jointly prices the realizable option value and the environmental footprint of research data. By establishing the Architecture-Aware Real Options Framework (AA-ROF), we show analytically, under stated assumptions, that digital research infrastructure acts as an active economic catalyst rather than a passive cost-center. Advancing research data through a three-tier distributed architecture, from Archive through Federated Mesh to AI Sandbox, transcends operational cost reduction; it systematically raises the value of options to combine heterogeneous datasets with AI foundation models. This is the economic reading of the friction-reduction premium formalized in Section 3, not a claim that the research-opportunity set itself expands beyond the limited, explicit form given to that expansion in Section 4.6. Crucially, this architecture reframes infrastructure investments as proactive drivers of a circular data economy, closing a reinforcing loop in which repurposed data and refined AI models flow from the Sandbox back into the archival base and raise the readiness of the corpus that re-enters the system, sustaining the ecosystem's combinatorial potential. Section 4.6 takes a first step toward rendering that loop as an explicit state transition, with reuse feeding back into each tier's own readiness and opportunity set, while the cross-tier return into the archival base remains a qualitative architectural property awaiting the extension set out in Section 5.4.

Our boundary-condition analysis and structural sensitivity evaluation suggest two core macroeconomic takeaways. First, once friction has fallen past the Vega-maximizing level X^* at which $d_1 = 0$, moved the option deep into the money, and then approaches its structural lower bound ($X \rightarrow X_{min}$), the asset's volatility sensitivity (Vega, $\mathcal{V}$) is sharply compressed, though not eliminated, and the valuation converges toward its deterministic NPV benchmark. What the architecture stabilizes is the valuation's exposure to scientific uncertainty, not that uncertainty itself, which remains a property of the science rather than of the infrastructure. Second, the Tier-2/Tier-3 result is the total-cost boundary $E_{3,crit}(\Delta Capex_{2 \rightarrow 3}) = E_2 + C_3 - C_2 - \Delta Capex_{2 \rightarrow 3}$. The illustrative calibration gives $59.9608 - \Delta Capex_{2 \rightarrow 3}$: the familiar value of 60 and the Monte Carlo 90% interval [48.5, 70.0] are zero-capital-cost intercepts, not complete thresholds. A fixed positive transition cost shifts the entire distribution left one-for-one; at $\Delta Capex_{2 \rightarrow 3} = 10$, the exact boundary is 49.9608, just below the assumed $E_3 = 50$. The two-period extension supplies conditional feedback comparative statics around the same zero-capital-cost intercept: readiness learning lowers it, whereas opportunity creation raises it. These opposing movements do not support treating 60 as a general lower bound, and no total-cost tier preference follows until program-level transition cost is supplied.

For science policy administrators (e.g., NAIRR, EOSC, and national HPC initiatives), AA-ROF replaces technological absolutism with *selective orchestration*. Rather than mandating universal Tier-3 centralization, funding agencies should deploy carbon-intensive AI

sandboxes only when the resulting Architecture Premium covers both environmental and annualized transition costs. High-volatility datasets retain substantial option value even under Tier-1 friction, whereas stable datasets may gain more from Tier-3 integration on the value side; whether that gain warrants adoption remains conditional on the complete total-cost boundary.

We restate here the boundary already drawn in Section 1: both takeaways are standard option-pricing properties, strike monotonicity and deep-in-the-money Vega compression, evaluated along an architectural gradient. What AA-ROF contributes is the mapping that renders those properties measurable in infrastructure-policy terms, together with the environmental account that constrains them, and not a new option-pricing result. In conclusion, AA-ROF bridges economic real option theory and the systems science of research data infrastructures. While this study provides an analytical baseline, future empirical efforts validating friction logs, real-time PUE, and expert-derived governance weights will further translate these theoretical insights into national infrastructure decision systems.

Author Contributions: Conceptualization, S.-S.K.; methodology, S.-S.K.; software, S.-S.K.; validation, S.-S.K.; formal analysis, S.-S.K.; investigation, S.-S.K.; writing—original draft preparation, S.-S.K.; writing—review and editing, S.-S.K., B.K. and Y.H.H.; project administration, Y.H.H.; funding acquisition, Y.H.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (RS-2021-NR060140), and by the ANCHOR program through the Seoul ANCHOR Center, funded by the Ministry of Education (MOE) and the Seoul Metropolitan Government, Republic of Korea (2026-ANCHOR-01-020-04).

Institutional Review Board Statement: Not applicable. This study is a theoretical and conceptual analysis that did not involve humans or animals.

Informed Consent Statement: Not applicable. This study did not involve humans.

Data Availability Statement: No empirical datasets were created or analyzed in this study. The numerical analysis is fully specified by the parameters reported in Section 4. The Python (version 3.10 or later) scripts that reproduce the Monte Carlo robustness checks and the distribution figure of Section 4.5, the worked structural-consistency example of Section 4.1, and the two-period extension of Section 4.6, with the random seeds fixed in the code, are openly archived at Zenodo: <https://doi.org/10.5281/zenodo.22790343>.

Conflicts of Interest: The authors declare no conflicts of interest. The funding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to submit the results for publication.

Appendix A. Derivation of the Expenditure-Proxy Elasticity Result

Section 3 states, without derivation, that evaluating the allocative-efficiency coefficient ψ of Equation (5) at $\psi = 1$ when the true value is $\psi < 1$ overstates C by more than it overstates S : the elasticity of C with respect to S exceeds unity and is largest for the least advanced tier. This appendix derives that claim from the option-value function of Equation (4).

Differentiating $C(S, X, \Theta) = S\Phi(d_1) - Xe^{-rT}\Phi(d_2)$ with respect to S and using the standard identity $S\phi(d_1) = Xe^{-rT}\phi(d_2)$, which cancels the terms arising from the dependence of d_1 and d_2 on S , gives the exact result $\partial C/\partial S = \Phi(d_1)$. The elasticity of C with respect to S is therefore

$$\eta_{C,S} \equiv \frac{\partial C}{\partial S} \cdot \frac{S}{C} = \frac{S\Phi(d_1)}{C} = \frac{S\Phi(d_1)}{S\Phi(d_1) - Xe^{-rT}\Phi(d_2)}.$$

Because $X > 0$, $T > 0$, and $\Phi(d_2) > 0$ throughout the domain of Section 3.4, the subtracted term $Xe^{-rT}\Phi(d_2)$ is strictly positive, so $C < S\Phi(d_1)$ and hence, $\eta_{C,S} > 1$ at every tier and every calibration considered in this paper: a proportional overstatement of S always produces a strictly larger proportional overstatement of C , which is why Section 3 describes the resulting bias as more than proportional rather than as a simple pass-through. The elasticity is largest precisely where the subtracted term is largest relative to $S\Phi(d_1)$, that is, where the residual friction X is highest, and the option is furthest from the money. Under the tier ordering $X_1 > X_2 > X_3$ maintained throughout this paper, this places the largest elasticity, and, therefore, the largest sensitivity of C to a common misspecification of ψ , at Tier 1, the least advanced tier. This is a general property of call options rather than one specific to AA-ROF: it is the same mechanism by which options further from the money carry higher elasticity than those deep in the money, applied here to the architecture-dependent strike $X(\vec{I})$ of Equation (2) rather than to an externally observed one. As a numerical illustration of this compression, fix curation and provisioning expenditure at 100 and consider $\psi = 0.8$ (so $S = 80$ rather than the baseline $S = 100$): the resulting tier option values are $C = (1.39, 19.51, 60.98)$, and the Architecture Premium falls accordingly, with $AP_{2 \rightarrow 3}$ declining from 44.9608 to 41.4688 and $AP_{1 \rightarrow 3}$ declining from 76.1357 to 59.5828.

Two consequences follow for use. Comparisons across programs are valid only under a common ψ , and any operational deployment must either estimate ψ from independent value measures, such as citation- or reuse-based indicators, or restrict itself to within-program comparisons made under a common ψ , which preserve the ordering of the gross option values C_i but not the level of AP , nor necessarily the sign of the net accounts NSV and NEB once S is scaled down. The mild sensitivity of the zero-capital-cost intercept $E_{3,crit}^{(0)}$ to S documented in Section 4.5 limits how far this bias propagates into the tier-comparison boundary, but it does not repair the level of C itself.

Appendix B. Algorithmic Summary of the AA-ROF Valuation Procedure

For practitioners who need a step-by-step procedure rather than only the closed-form equations of Sections 3–5, Algorithm A1 restates those equations, and the two-stage screening rule of Section 5.3, as an explicit procedure; it introduces no new functional form and no numerical values beyond those already reported.

Algorithm A1 Applying AA-ROF to a candidate multi-tier infrastructure.

Input: an infrastructure vector $\vec{I}_i = (A_i, R_i, P_i)$ for each candidate tier i ; the friction-function parameters $(X_{max}, \alpha, \beta, \gamma, \lambda, P_{min})$ of Equation (2) and the environment-function parameters (E_{base}, η, κ) of Equation (3); the intrinsic epistemic value S and option parameters $\Theta = (\sigma, r, T)$; the redundant-recreation costs K_{econ} and $p_c K_{carbon}$; the incremental reuse rate ρ_{reuse} ; where available, the annualized transition cost $\Delta Capex_{i \rightarrow j}$; and, if Step 9 is exercised, the feedback parameters (δ, ν) and the other two-period inputs of Section 4.6.

1. For each candidate tier i , compute the exercise friction $X_i \leftarrow X(\vec{I}_i)$ from Equation (2).
 2. For each candidate tier i , compute the environmental penalty $E_i \leftarrow E(\vec{I}_i)$ from Equation (3).
 3. For each candidate tier i , compute the option value $C_i \leftarrow C(S, X_i, \Theta)$ from Equation (4).
 4. For a proposed transition $i \rightarrow j$, compute the Architecture Premium $AP_{i \rightarrow j} \leftarrow C_j - C_i$ and the operating-only Net Sustainability Value $NSV_i \leftarrow C_i - E_i$ for each tier under consideration.
-

Algorithm A1 *Cont.*

5. For the target tier j , compute the Net Environmental Benefit $NEB \leftarrow \rho_{reuse}(K_{econ} + p_c K_{carbon}) - (E_j - E_i)$, where ρ_{reuse} is measured as the incremental reuse rate relative to the origin tier i for the transition under consideration; this coincides with the Tier-1 baseline definition of Section 3 when $i = 1$, and for a transition from a higher origin tier the rate must be re-measured against reuse already realized at tier i . Then, compute the Redundant Experiment Avoidance ratio $REA \leftarrow \rho_{reuse}(K_{econ} + p_c K_{carbon}) / (E_j + X_j)$ (Section 5.3).
6. **Stage 1 (net-benefit screen):** accept the transition $i \rightarrow j$ only if $AP_{i \rightarrow j} \geq \Delta Capex_{i \rightarrow j} + (E_j - E_i)$ (Equation (6)); otherwise, reject it, regardless of REA . If $\Delta Capex_{i \rightarrow j}$ is not reported, Stage 1 instead evaluates only the zero-capital-cost intercept ($\Delta Capex_{i \rightarrow j} = 0$, as in Section 4.4) and the accept/reject decision remains undetermined until program-level transition-cost data are supplied.
7. **Stage 2 (reuse-dependence flag):** among transitions accepted at Stage 1, flag those with $REA < 1$ as value-positive but reuse-dependent, and use this flag to rank candidates under a binding budget or to trigger the reuse-enforcing policies of Scenario A.
8. (Optional robustness check, Section 4.5.) Redraw the tier-level inputs (X_i , E_i , S , σ , and the cost parameters) directly within their stated uncertainty ranges, as Section 4.5 does, entering at Step 3 with the redrawn X_i and E_i in place of the outputs of Steps 1–2, and repeat Steps 3–7 to report the resulting distribution of the Stage-1 boundary rather than a single point estimate.
9. (Optional dynamic extension, Section 4.6.) If reuse is expected to alter next-period readiness and the opportunity set, this extension does not replace the Stage-1 decision of Step 6; instead, report the two-period total values TV_i^{2p} and TV_j^{2p} of Equation (11) for the origin and target tiers, compared within that same two-period account (Table 6), as supplementary robustness information alongside the Stage-1 decision.

Output: an accept/reject decision for the candidate transition (or, where $\Delta Capex_{i \rightarrow j}$ is not supplied, the zero-capital-cost intercept with the decision left undetermined, as in Step 6), together with the reuse-dependence flag of Stage 2 and, where Step 8 or 9 is exercised, an uncertainty range or dynamic sensitivity around that decision.

References

1. International Energy Agency. *Energy and AI*; World Energy Outlook Special Report; IEA: Paris, France, 2025. Available online: <https://www.iea.org/reports/energy-and-ai> (accessed on 8 August 2026).
2. Boulding, K.E. General Systems Theory—The Skeleton of Science. *Manag. Sci.* **1956**, *2*, 197–208. [CrossRef]
3. Ackoff, R.L. Towards a System of Systems Concepts. *Manag. Sci.* **1971**, *17*, 661–671. [CrossRef]
4. Laney, D.B. *Infonomics: How to Monetize, Manage, and Measure Information as an Asset for Competitive Advantage*; Routledge: New York, NY, USA, 2017. [CrossRef]
5. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L.B.; Bourne, P.E.; et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016**, *3*, 160018. [CrossRef] [PubMed]
6. Tang, Y.; Liu, Y.; Liu, D.; Tsai, F.-S. Data asset valuation model based on generative artificial intelligence. *PLoS ONE* **2025**, *20*, e0328926. [CrossRef] [PubMed]
7. Qin, H.; Liu, Y.; Hu, Y. Commercial Value of Data Resource: An Overview of Current State of Research and Future Prospects. *SAGE Open* **2025**, *15*, 21582440251379511. [CrossRef]
8. Majithia, N.; Carey-Wilson, T.; Simperl, E.; Shadbolt, N. An actionable framework for AI-ready data. *AI Mag.* **2026**, *47*, e70054. [CrossRef]
9. Greenberg, J.; Pepper, J.; Zhao, X.; Marciano, R.; Breen, D.; An, Y. The metadata ecosystem and AI: Enabling FAIR and AI-ready data. *AI Mag.* **2026**, *47*, e70060. [CrossRef]
10. Trigeorgis, L. *Real Options: Managerial Flexibility and Strategy in Resource Allocation*; MIT Press: Cambridge, MA, USA, 1996.
11. Benaroch, M. Managing Information Technology Investment Risk: A Real Options Perspective. *J. Manag. Inf. Syst.* **2002**, *19*, 43–84. [CrossRef]
12. Arrow, K.J.; Fisher, A.C. Environmental preservation, uncertainty, and irreversibility. *Q. J. Econ.* **1974**, *88*, 312–319. [CrossRef]
13. Henry, C. Investment decisions under uncertainty: The irreversibility effect. *Am. Econ. Rev.* **1974**, *64*, 1006–1012.
14. Coyle, D.; Gamberi, L. A real options approach to data valuation. *Bus. Econ.* **2024**, *59*, 227–234. [CrossRef]

15. Star, S.L.; Ruhleder, K. Steps Toward an Ecology of Infrastructure: Design and Access for Large Information Spaces. *Inf. Syst. Res.* **1996**, *7*, 111–134. [\[CrossRef\]](#)
16. Trist, E.L.; Bamforth, K.W. Some Social and Psychological Consequences of the Longwall Method of Coal-Getting: An Examination of the Psychological Situation and Defences of a Work Group in Relation to the Social Structure and Technological Content of the Work System. *Hum. Relat.* **1951**, *4*, 3–38. [\[CrossRef\]](#)
17. McCrory, D. Data Gravity—in the Clouds. *Data Gravity Blog*, 2010. Available online: <https://datagravitas.com/2010/12/07/data-gravity-in-the-clouds/> (accessed on 7 August 2026).
18. Belcastro, L.; Marozzo, F.; Orsino, A.; Talia, D.; Trunfio, P. Navigating the Edge-Cloud Continuum: A State-of-Practice Survey. *IEEE Access* **2026**, *14*, 40622–40647. [\[CrossRef\]](#)
19. Ejarque, J.; Badia, R.M.; Albertin, L.; Aloisio, G.; Baglione, E.; Becerra, Y.; Boschert, S.; Berlin, J.R.; D’Anca, A.; Elia, D.; et al. Enabling dynamic and intelligent workflows for HPC, data analytics, and AI convergence. *Future Gener. Comput. Syst.* **2022**, *134*, 414–429. [\[CrossRef\]](#)
20. Yang, Y.; Fu, R.; Amin, M.B.; Kang, B. The Impact of Modern AI in Metadata Management. *Hum.-Centric Intell. Syst.* **2025**, *5*, 323–350. [\[CrossRef\]](#)
21. Gauthier, A.; Vancauwenberghe, L.; Cousty, J.-C.; Matthey-Doret, C.; Franken, R.; Maennel, S.; Riba Grognez, O. A FAIR research data infrastructure for high-throughput digital chemistry. *Digit. Discov.* **2025**, *4*, 3502–3514. [\[CrossRef\]](#)
22. Checkland, P. From Optimizing to Learning: A Development of Systems Thinking for the 1990s. *J. Oper. Res. Soc.* **1985**, *36*, 757–767. [\[CrossRef\]](#)
23. Sterman, J.D. System Dynamics Modeling: Tools for Learning in a Complex World. *Calif. Manag. Rev.* **2001**, *43*, 8–25. [\[CrossRef\]](#)
24. Black, F.; Scholes, M. The Pricing of Options and Corporate Liabilities. *J. Polit. Econ.* **1973**, *81*, 637–654. [\[CrossRef\]](#) [\[PubMed\]](#)
25. Merton, R.C. Theory of Rational Option Pricing. *Bell J. Econ. Manag. Sci.* **1973**, *4*, 141–183. [\[CrossRef\]](#)
26. Copeland, T.; Antikarov, V. *Real Options: A Practitioner’s Guide*; Texere: New York, NY, USA, 2001.
27. Uptime Institute. *Uptime Institute Global Data Center Survey Results 2024*; Uptime Institute: New York, NY, USA, 2024. Available online: <https://uptimeinstitute.com/resources/research-and-reports/uptime-institute-global-data-center-survey-results-2024> (accessed on 7 August 2026).
28. Google. *2025 Environmental Report*; Google LLC: Mountain View, CA, USA, 2025. Available online: <https://www.sustainability.google/reports/google-2025-environmental-report/> (accessed on 10 August 2026).
29. Dasu, T.; Johnson, T. *Exploratory Data Mining and Data Cleaning*; Wiley Series in Probability and Statistics; John Wiley & Sons: Hoboken, NJ, USA, 2003. [\[CrossRef\]](#)
30. European Commission, Directorate-General for Research and Innovation; PwC EU Services. *Cost-Benefit Analysis for FAIR Research Data: Cost of Not Having FAIR Research Data*; Publications Office of the European Union: Luxembourg, 2018. [\[CrossRef\]](#)
31. Piwowar, H.A.; Vision, T.J. Data reuse and the open data citation advantage. *PeerJ* **2013**, *1*, e175. [\[CrossRef\]](#) [\[PubMed\]](#)
32. Colavizza, G.; Hrynaszkiewicz, I.; Staden, I.; Whitaker, K.; McGillivray, B. The citation advantage of linking publications to research data. *PLoS ONE* **2020**, *15*, e0230416. [\[CrossRef\]](#) [\[PubMed\]](#)
33. Beagrie, N.; Houghton, J. *The Value and Impact of Data Sharing and Curation: A Synthesis of Three Recent Studies of UK Research Data Centres*; Charles Beagrie Ltd. and Jisc: Bristol, UK, 2014. Available online: <https://repository.jisc.ac.uk/5568/> (accessed on 9 August 2026).
34. National Artificial Intelligence Research Resource Task Force. *Strengthening and Democratizing the U.S. Artificial Intelligence Innovation Ecosystem: An Implementation Plan for a National Artificial Intelligence Research Resource*; National Science Foundation: Alexandria, VA, USA, 2023.
35. EOSC Partnership. *Strategic Research and Innovation Agenda (SRIA) of the European Open Science Cloud (EOSC)*, v1.3; Zenodo: Geneva, Switzerland, 2024. [\[CrossRef\]](#)
36. Gupta, U.; Kim, Y.G.; Lee, S.; Tse, J.; Lee, H.-H.S.; Wei, G.-Y.; Brooks, D.; Wu, C.-J. Chasing Carbon: The Elusive Environmental Footprint of Computing. In Proceedings of the 2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA), Seoul, Republic of Korea, 27 February–3 March 2021; pp. 854–867. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.