Agentic AI for Research Evaluation Systems: A Decision-Sciences Framework with Ownership, Responsibility, and Organisational Strength

Lanz CWJ Chan

Finamatrix Quant-Lab, Singapore

Corresponding author. Email: lanz@finamatrix.net

Shin Hung Pan

Department of Information Management, Chaoyang University of Technology, Taiwan

Email: vincentpan@cyut.edu.tw

Published on October 2, 2026

Abstract

Purpose — Universities continue to reward journal prestige and publication counts, although responsible-assessment movements have shown that these proxies mismeasure individual contribution and render translation, infrastructure, and societal engagement structurally invisible. This paper reconceptualises research evaluation as a decision-sciences design problem and asks whether agentic artificial intelligence (AI) can make broader, evidence-centred evaluation operationally feasible without weakening accountability. The central claim is conditional: earlier adoption strengthens the organisation only when governance quality and ownership clarity exceed identifiable thresholds; below them, earlier adoption scales confusion and bias.

Design/methodology/approach — The paper develops a conceptual framework with a formalised propositional model. Responsible-assessment principles (DORA, the Leiden Manifesto, CoARA, and the Singapore Statement on Research Integrity) are synthesised with a five-agent system architecture (narrative parsing, multi-label classification, knowledge-graph construction, reviewer recommendation, and explanation over a shared evidence registry). Twelve equations define organisational evaluation strength as a bounded [0,1] composite of decision quality, fairness, strategic alignment, and organisational learning. An explicit adoption-stock bridge connects adoption timing and intensity to model capability and evidence richness, so the adoption effect is derived as a total derivative rather than asserted. Proposition 1 and Hypotheses H1–H5 state the testable implications, with full operationalisation, identification, validity, and ethics provisions.

Findings — Organisational evaluation strength is a joint function of AI capability, evidence richness, governance quality, ownership clarity, and explainability. The sign of the adoption effect is determined by a race between a capability-and-learning channel (positive) and a governance-erosion channel (negative when adoption dilutes accountability or amplifies bias). Thresholds G* and O* are therefore not arbitrary constants but sign-change conditions solvable from the model. Under stated assumptions,

advancing the adoption date (holding the investment path fixed) weakly increases learning and strength at every horizon — a corollary that gives the familiar slogan 'the sooner, the stronger' a precise, falsifiable meaning.

Research limitations/implications — The model is formal and propositional; coefficients and thresholds require empirical calibration. The paper specifies a four-stage validation programme linked hypothesis-by-hypothesis to the model (extraction benchmarking, graph verification, committee human-factors experiments, and adversarial bias audits), together with difference-in-differences and instrumental-variable strategies for the self-selection problem that threatens any naive 'early adopters perform better' comparison. Field heterogeneity and cross-institutional portability beyond the Singapore deployment context remain open.

Practical implications — Institutions can redesign evaluation workflows around inspectable evidence packets, assign five accountable human roles (system owner, decision owner, model steward, audit lead, domain expert reviewer), and benchmark decision-system quality with the ownership-clarity and traceability indices defined here. Singapore's RIE2030 (S$37 billion, 2026–2030) provides a policy-aligned deployment context whose four strategic domains supply the mission taxonomy for the relevance construct.

Social implications — Appropriately governed AI-augmented evaluation can reduce prestige-proxy dependence, surface overlooked contributions (software, data, standards, mentoring, policy influence), and realign academic incentives toward genuine knowledge creation and translational impact, while preserving due process, contestability, and appeal rights for those evaluated.

Originality/value — The paper offers (i) a decision-sciences formalisation of agentic-AI adoption in academic evaluation with bounded, internally consistent dynamics (the prior linear specification is replaced by logistic-bounded composites and a single stock–flow learning law with closed form); (ii) an explicit adoption bridge that repairs the missing link between adoption timing and organisational strength; (iii) a governance architecture that operationalises the Singapore Statement as auditable checkpoints aligned to NIST AI RMF and ISO/IEC 42001; and (iv) a hypothesis-linked empirical roadmap. It reframes evaluation reform as governance engineering, not merely a normative aspiration.

Keywords: agentic AI; research evaluation; decision sciences; narrative CV; knowledge graph; explainable AI; responsible research assessment; organisational learning

JEL Classification: I23; O33; M15; D83

1. Introduction

Decision systems determine what organisations reward, scale, and reproduce. In universities and research-intensive institutions, evaluation systems are especially consequential because they govern hiring, promotion, tenure, resource allocation, and legitimacy (Hicks et al., 2015; Moher et al., 2018).

Yet many such systems continue to rely on journal impact factors, publication counts, and prestige heuristics despite longstanding evidence that these proxies are conceptually inappropriate for evaluating individual researchers and narrow the kinds of contributions they can detect (Garfield, 1955; Seglen, 1997). The cost is not merely measurement error. When evaluation rewards venue prestige, researchers rationally optimise for venue prestige; infrastructure, replication, data stewardship, mentoring, and translation — the activities through which knowledge changes the world — are systematically under-supplied because the decision system cannot see them.

The journal impact factor was created by Eugene Garfield as part of the Science Citation Index to help librarians identify which journals to prioritise for acquisition — not to judge the merit of individual scholars or articles (Garfield, 1955, 2006). Nevertheless, it became embedded in appointment, promotion, and funding cultures across higher education globally, often as a shorthand for quality (Grzybowski & Patryn, 2017; Larivière & Sugimoto, 2019; Seglen, 1997). The San Francisco Declaration on Research Assessment (DORA), developed in December 2012 and published in May 2013, explicitly rejects this usage and advocates assessing research on its own merits while recognising a broader range of outputs including software and data (San Francisco Declaration on Research Assessment, 2013). The Leiden Manifesto (Hicks et al., 2015) and, more recently, the Coalition for Advancing Research Assessment (CoARA, 2022) extend the same normative programme into institutional commitments; as of the mid-2020s, hundreds of organisations have signed the CoARA Agreement. The normative case, in short, is settled. What remains unsettled — and what this paper addresses — is the operational and organisational question: broader evaluation is information-intensive, and committees working under severe time constraints will revert to prestige heuristics unless the evidence-processing burden is fundamentally reduced and the accountability structure for any machine assistance is made explicit.

This article treats research evaluation as a decision-sciences problem. The central question is whether organisations can design AI-assisted decision systems that produce better, fairer, and more strategically aligned judgments than legacy heuristics while preserving accountability and due process (Hicks et al., 2015; Moher et al., 2018). Three propositions organise the framework. First, responsible research assessment has established that broader evidence should be considered in evaluation (CoARA, 2022; San Francisco Declaration on Research Assessment, 2013). Second, AI and agentic tools make such broader evaluation operationally feasible (Devlin et al., 2019; Vaswani et al., 2017; Wang et al., 2024). Third — and most importantly — the timing benefit of adoption is conditional: the sooner these tools are implemented appropriately, with explicit ownership and responsibility, the stronger the organisation becomes; implemented without those conditions, earlier adoption compounds disorder rather than capability. Making the third proposition precise, bounded, and falsifiable is the paper's main formal contribution (Sections 6 and 8).

The paper proceeds as follows. Section 2 reviews four literature streams and states the search approach. Section 3 formulates the research problem as six design requirements (R1–R6). Section 4 presents the conceptual model (Dual Dashboard and Impact Matrix). Section 5 specifies the five-agent AI

architecture. Section 6 develops the formal model with explicit assumptions, bounded equations, operationalisation, and Hypotheses H1–H5. Section 7 details the ownership and responsibility architecture. Section 8 states and proves (in sketch) Proposition 1, the conditional adoption–strength result, and its early-adoption corollary. Section 9 situates deployment in Singapore's RIE2030 context. Section 10 discusses contributions, a hypothesis-linked validation roadmap, identification and validity, ethics, and limitations. Section 11 concludes.

2. Literature Review

2.1 Search approach

This is a conceptual paper; the review is purposive rather than exhaustive, and its logic is stated here so that construct provenance can be audited. Four streams were synthesised: (i) bibliometric critique and responsible assessment (Science Citation Index / impact-factor historiography, DORA, Leiden Manifesto, CoARA); (ii) narrative-CV and broader-evidence evaluation research; (iii) AI for scholarly knowledge structuring (transformer language models, scientific information extraction, knowledge graphs, explainable AI); and (iv) organisational research on AI adoption, accountability, and technology acceptance. Sources were drawn from bibliometric and science-policy journals (e.g., Nature, Research Evaluation, Minerva, JASIST), computer-science venues (NAACL-HLT, EMNLP, NeurIPS, CHI), management outlets (Harvard Business Review, California Management Review, MIS Quarterly), and authoritative governance documents (Singapore Statement, NIST AI RMF, ISO/IEC 42001). Table 1 maps each model construct to its primary literature source, so that the formal model in Section 6 can be read as a synthesis rather than an invention ex nihilo. Where grey literature is cited (e.g., a United Nations University prototype report), it is flagged as such and never used as sole support for a theoretical claim.

2.2 The misuse of journal-based metrics

Eugene Garfield introduced the impact-factor concept in a 1955 paper in Science as a bibliographic tool for evaluating journal coverage of the Science Citation Index (Garfield, 1955). Garfield himself later stated that the measure was not designed for evaluating individual researchers (Garfield, 2006). Seglen (1997) showed that a large fraction of citations derives from a small fraction of papers, making journal-level averages misleading as article-level proxies. Research systems that over-indexed on these metrics showed growing evidence of quality decline, integrity problems, and strategic misalignment between institutional incentives and genuine research contribution (Curry et al., 2022; Shu et al., 2022). DORA (2013) provides the most widely adopted formal response, with 18 recommendations covering funders, institutions, publishers, and researchers; central among these is the recommendation to avoid journal-based metrics in hiring, promotion, and funding decisions, and to assess research on its own merits across diverse output types (San Francisco Declaration on Research Assessment, 2013). CoARA (2022) has since converted this recommendation into signed institutional commitments with implementation timelines, which matters for the present paper because commitment devices change the adoption

calculus modelled in Section 8: evaluation reform is no longer only an aspiration but, for signatories, a scheduled organisational change whose timing can be chosen — or squandered.

2.3 Narrative CVs and responsible assessment

Narrative curriculum vitae (CVs) have emerged as a practical mechanism for surfacing contributions that publication lists obscure, including team science, infrastructure building, societal engagement, and mentoring (Bordignon et al., 2023; Larivière & Sugimoto, 2019). Bordignon et al. (2023), reviewing opinion pieces and experiments in Research Evaluation, found that narrative CVs address concerns about metrics-driven evaluation by contextualising achievements, broadening counted contributions, and supporting diversity — while warning that unstructured narratives can reintroduce prestige and rhetoric biases if committees lack processing support. Subsequent funder pilots confirm rapid diffusion: Aubert Bonn et al. (2025) document narrative-CV adoption in biomedical funding, Albert et al. (2025) scope what funders actually request in narrative formats, and Lee et al. (2026) validate AI-assisted evaluation of open-science practices against human experts. The key operational challenge is that narrative materials increase qualitative richness but impose interpretive load on committees (Albert et al., 2025), motivating the AI-assisted approach developed in Section 5: the richer the evidence base, the stronger the case for machine pre-processing — and the stronger the requirement that machine inferences remain inspectable and contestable (Requirement R3).

2.4 AI for knowledge structuring and evaluation

Natural language processing and large language models (LLMs) have matured sufficiently to extract structured information from unstructured scholarly text (Devlin et al., 2019; Vaswani et al., 2017). Scientific knowledge-graph construction demonstrates that research entities, relations, and claims can be converted into queryable graph structures (Luan et al., 2018; Martínez-Rodríguez et al., 2020). The agentic turn — LLM-based autonomous agents that plan, invoke tools, and coordinate sub-tasks — is now surveyed systematically (Wang et al., 2024; Xi et al., 2025), providing the architectural vocabulary for the five-agent design in Section 5. AI-assisted evaluation prototypes in CV screening and document review consistently show both operational promise (throughput, consistency, coverage of heterogeneous evidence) and governance risks (opacity, inherited prestige bias, accountability gaps) (Chan et al., 2024; Lee et al., 2026; United Nations University, 2026). Explainable-AI research supplies the inspection layer (Amershi et al., 2019; Arrieta et al., 2020; Rudin, 2019), and model-reporting standards supply the documentation layer (Mitchell et al., 2019). The literature thus supports feasibility; it does not, by itself, support any claim that adoption improves organisational outcomes. That claim requires the organisational theory and formal conditions developed next.

2.5 Organisational strength, adoption timing, and accountability

Management research consistently finds that AI adoption generates positive performance outcomes under appropriate governance, human oversight, user competence, and well-defined accountability (Fountaine et al., 2019; Shrestha et al., 2019). Technology-acceptance theory predicts that sustained,

effective use — the flow that feeds organisational learning in Section 6 — depends on perceived usefulness and social/organisational facilitating conditions (Davis, 1989; Venkatesh et al., 2003). Global AI-ethics and governance instruments converge on the same design variables the model formalises: transparency, human oversight, bias monitoring, and accountable ownership (International Organization for Standardization, 2023; Jobin et al., 2019; National Institute of Standards and Technology, 2023). The Singapore Statement on Research Integrity (Resnik & Shamoo, 2011; World Conferences on Research Integrity, 2010), produced at the Second World Conference on Research Integrity in Singapore in July 2010 (340 participants, 51 countries), articulates four principles — honesty, accountability, professional courtesy and fairness, and good stewardship — and fourteen researcher responsibilities. Section 7 converts these principles from normative decoration into auditable governance checkpoints. What the adoption literature has not provided is a formal account of when earlier adoption helps rather than harms. Section 6 supplies that account by distinguishing the adoption stock, the governance-erosion channel, and the learning-compounding channel.

Table 1. Construct provenance: each formal construct and its literature source.

Model construct	Primary literature source(s)	Section
Evidence richness (E); prestige-proxy critique	Garfield (1955, 2006); Seglen (1997); Hicks et al. (2015); DORA (2013); CoARA (2022)	§2.2, §6
Narrative evidence; committee load	Bordignon et al. (2023); Aubert Bonn et al. (2025); Albert et al. (2025)	§2.3, §5
AI capability (M); agentic architecture	Vaswani et al. (2017); Devlin et al. (2019); Wang et al. (2024); Xi et al. (2025)	§2.4, §5
Explainability (X); contestability	Arrieta et al. (2020); Rudin (2019); Amershi et al. (2019)	§5, §6
Bias intensity (B); fairness (F)	Buolamwini and Gebru (2018); Mitchell et al. (2019); Jobin et al. (2019)	§6
Governance (G); ownership (O); traceability (Tr)	Singapore Statement (2010); NIST (2023); ISO (2023); Shrestha et al. (2019)	§6, §7
Adoption & learning (K, U, L)	Davis (1989); Venkatesh et al. (2003); Fountaine et al. (2019)	§6, §8

3. Research Problem and Design Requirements

The research-assessment problem is a mismatch between three layers: institutional goals, available evidence, and committee decision processes. Universities claim to value originality, openness, societal relevance, mentoring, and problem-solving, but evaluation forms often reduce candidate review to publication-venue heuristics (Hicks et al., 2015; San Francisco Declaration on Research Assessment, 2013). Available evidence is richer than metrics suggest — spanning code repositories, datasets, technical standards, prototypes, and policy influence — but is fragmented and difficult to interpret consistently at scale (Larivière & Sugimoto, 2019). This creates six design requirements, each traced to the architecture (Section 5) and the formal constructs (Section 6) so that satisfaction can be audited rather than asserted:

R1: Heterogeneous evidence ingestion — the system must accept multiple output types, not article metadata alone (DORA, 2013). [Architecture: Evidence Registry; Construct: E]

R2: Multi-dimensional representation — the system must distinguish theory, tools, practice, policy, markets, and education as separable contribution forms (Bornmann, 2013). [Architecture: Classification agent; Construct: Impact Matrix IM1–IM6]

R3: Human interpretability — committees must inspect and contest the basis of AI inferences (Arrieta et al., 2020; Rudin, 2019). [Architecture: Explanation agent; Construct: X]

R4: Bias awareness — the system must detect and minimise prestige reinforcement (Buolamwini & Gebru, 2018; Mitchell et al., 2019). [Architecture: Audit lead + adversarial testing; Construct: B]

R5: Workflow scalability — broader assessment must be operable under realistic committee time constraints (Bordignon et al., 2023). [Architecture: five-stage workflow; Construct: U]

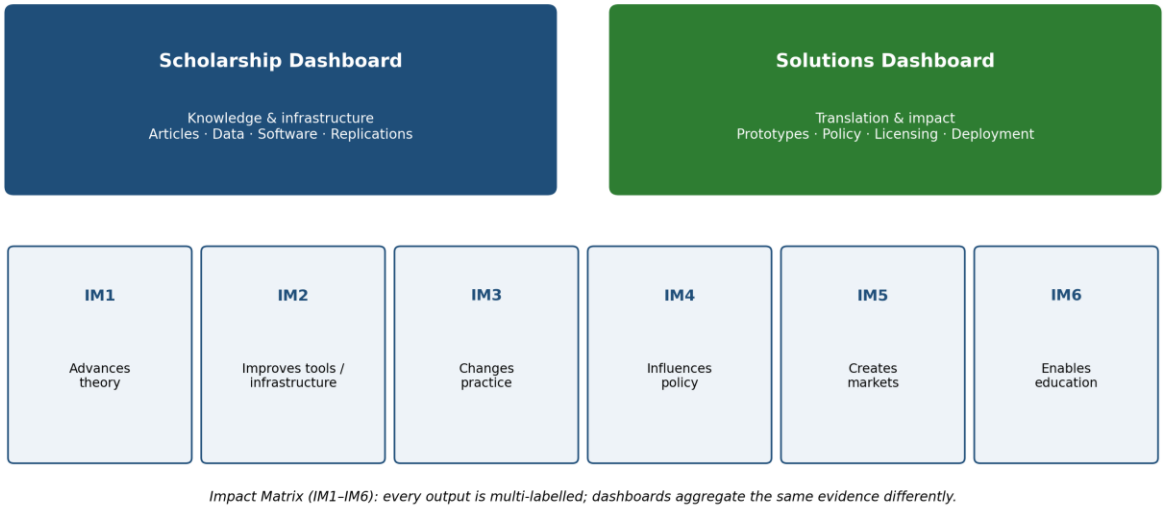
R6: Governance compatibility — the system must support research integrity and due process principles (World Conferences on Research Integrity, 2010; NIST, 2023). [Architecture: five accountable roles; Constructs: G, O, Tr]

4. Conceptual Model of Academic Contribution

4.1 Dual Dashboard

The framework evaluates candidates through two complementary dashboards. The Scholarship Dashboard captures advancement of knowledge and scientific infrastructure (articles, preprints, methods, datasets, software, replications); the Solutions Dashboard captures translation into practice, policy, and commercialisation (prototypes, pilots, standards, policy citations, licensing, deployed systems). The dashboards are views over the same evidence registry, not separate evaluations: a dataset may appear as scholarship (infrastructure) and, once adopted by a hospital network, as a solution (changed practice). Figure 1 visualises the dual view and the six-category Impact Matrix defined next. Author's framework note: the Dual Dashboard / Impact Matrix taxonomy was first articulated in an unpublished policy white paper (Chan, 2026); in this paper the taxonomy is grounded in the peer-reviewed societal-impact literature (Bornmann, 2013; Waltman & van Eck, 2019) and is treated as a design proposal to be validated, not as an established standard.

Figure 1. Dual dashboard and six-category impact matrix



Note: Every output is multi-labeled; the dashboards aggregate identical underlying evidence for different decision questions (see Table 2a).

Table 2a. Dual Dashboard: two views over one evidence registry.

Dashboard	Core purpose	Representative evidence
Scholarship Dashboard	Advancement of knowledge and scientific infrastructure	Articles, preprints, methods, datasets, software, replications
Solutions Dashboard	Translation into practice, policy, and commercialisation	Prototypes, pilots, standards, policy citations, licensing, and deployed systems

4.2 Impact Matrix

Each candidate's outputs are projected into an Impact Matrix across six categories (Table 2b). The sixfold structure follows the societal-impact literature's distinction between epistemic, instrumental, and capacity-building effects (Bornmann, 2013), extended with an explicit market-creation category for translational systems. Because field-specific impact pathways and timescales differ sharply (Waltman & van Eck, 2019), category assignments are probabilistic (Section 5.4) and field-normalised at the analytics layer; the matrix is a representation device, not a ranking formula. No candidate score is computed by summing categories — aggregation to a decision remains a committee judgement supported, not replaced, by the matrix (see Section 5.6 and the ethics provisions in Section 10.4).

Table 2b. Impact Matrix: six multi-label contribution categories with operational definitions.

Impact category	Operational definition	Sample evidence
IM1 Advances theory	Expands explanatory or predictive understanding	New models, benchmarks, generalisations
IM2 Improves tools/infrastructure	Produces reusable research-enabling assets	Software, datasets, protocols, platforms
IM3 Changes practice	Alters professional or clinical workflows	Adoption records, partner testimony, deployment metrics
IM4 Influences policy	Shapes public or regulatory decisions	Policy citations, committee roles, and standards references
IM5 Creates markets	Generates commercial activity or	Spinoffs, licensing, procurement,

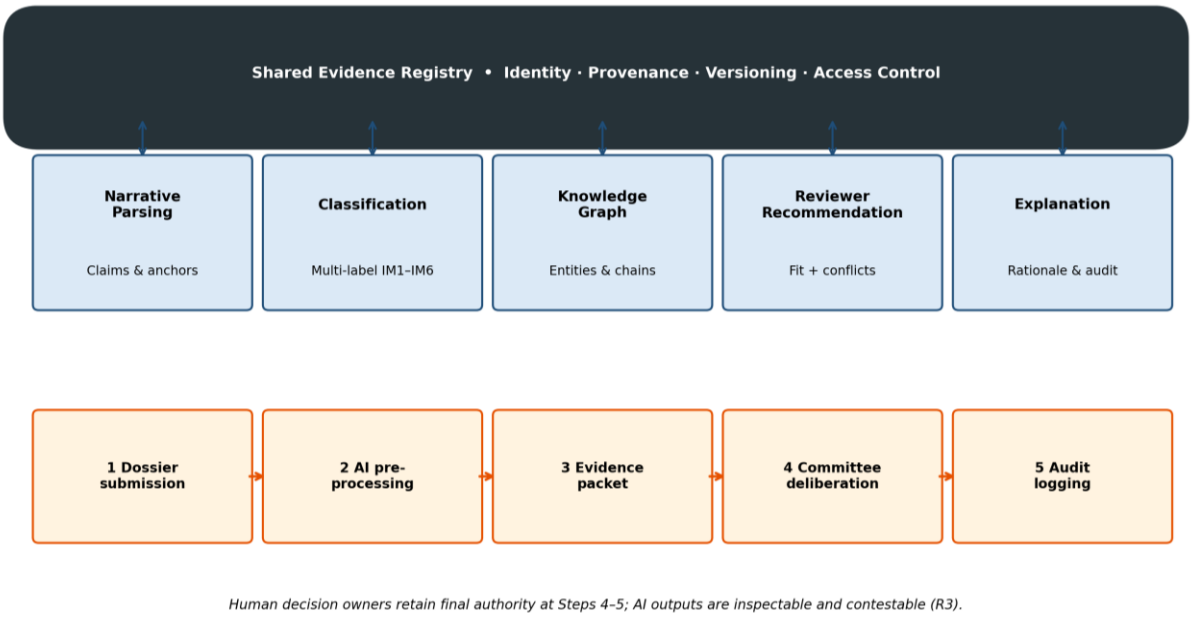
	enterprise value	revenue
IM6 Enables education	Develops capability through teaching and mentoring	Curriculum reuse, open resources, trainee outcomes

5. AI System Architecture

5.1 Overview

The proposed AI system contains five functional agentic modules operating over a shared evidence registry (Figure 2, Table 3). Following current agent-survey terminology (Wang et al., 2024; Xi et al., 2025), each module is an LLM-based agent with a bounded mandate, tool access, and an auditable output contract; no agent issues an evaluative decision. The registry enforces identity resolution, provenance, versioning, and access control — the technical precondition for traceability (Tr) in Section 6. Design follows human–AI interaction guidance that machine contributions be explainable, interruptible, and correctable (Amershi et al., 2019) and organisational-AI governance requirements for human oversight and accountability (NIST, 2023; ISO, 2023; Shrestha et al., 2019).

Figure 2. Agentic AI system architecture and the five-stage decision-support workflow.



Note: Human decision owners retain final authority; all AI inferences carry provenance and confidence for inspection (R3).

Table 3. Five agentic modules: mandates and auditable output contracts.

Module	Function	Output contract
Narrative parsing agent	Extracts claims, outputs, beneficiaries, and evidence anchors from CVs and impact statements	Structured assertion tuples with source spans
Classification agent	Maps outputs to Impact Matrix categories using multi-label prediction	Category probabilities with calibrated confidence
Knowledge graph agent	Links people, projects, outputs, organisations, and downstream	Typed triples with provenance and verification status

	outcomes	
Reviewer recommendation agent	Suggests panellists based on semantic expertise overlap and conflict screening	Ranked shortlist with conflict flags (human confirms)
Explanation agent	Generates inspectable rationale and provenance trails for inferences	Evidence packet: rationale, confidence, unresolved ambiguities

5.2 Data model

A minimal data model includes five entity types: Researcher (name, identifiers, affiliation, career stage), Output (article, software, dataset, patent, standard, prototype, policy brief), Project (funded or unfunded initiative), Organisation (university, company, hospital, ministry), and Evidence Event (citation, deployment, adoption, award). These entities are linked through typed edges including authored, cites, deployed at, adopted by, and informs policy, consistent with knowledge-graph approaches for scholarly corpora (Luan et al., 2018; Martínez-Rodríguez et al., 2020). Every edge carries provenance (source document, extraction method, verification status), which is what allows the graph to support contestation (R3) and audit (R6) rather than merely retrieval.

5.3 Narrative claim extraction

An LLM-assisted extraction pipeline decomposes narrative CVs into structured assertion tuples: (actor, contribution type, beneficiary, evidence anchor, claimed outcome). A claim such as ‘the project informed national digital health policy’ becomes a policy-influence assertion requiring corroboration against policy texts or advisory appointments (Devlin et al., 2019; Vaswani et al., 2017). The pipeline includes sentence segmentation, named-entity recognition, relation extraction, and claim normalisation. Extraction quality is an empirical quantity (precision/recall against an annotated corpus), not an assumption; it enters the formal model through evidence richness (E) and is benchmarked in validation Stage 1 (Section 10.2). Unverifiable claims are not deleted; they are surfaced to the committee as unresolved ambiguities — a deliberate design choice that converts a known LLM failure mode (fluent fabrication) into an explicit human-judgement prompt.

5.4 Multi-label classification

Because research outputs can contribute across multiple Impact Matrix categories simultaneously, the architecture uses multi-label classification. Feature inputs include textual embeddings from abstracts and narratives, metadata such as repository links or patent status, and entity neighbourhoods from the knowledge graph (Vaswani et al., 2017). A thresholded ranking model surfaces the most likely category assignments with calibrated confidence estimates, preserving uncertainty for committee review (Arrieta et al., 2020). Calibration matters organisationally, not only statistically: a committee that learns the system’s stated 0.8 confidence is empirically right four times out of five can rationally allocate scrutiny; a miscalibrated system trains either blind trust or blanket dismissal, both of which degrade effective use (U) and hence learning (Section 6). Prestige features (journal names, institutional brands) are excluded

from the classifier's feature set by design, and their indirect leakage is tested adversarially (Section 10.2, Stage 4), operationalising R4.

5.5 Knowledge graph reasoning

The knowledge graph enables relational queries over impact chains that bibliometrics cannot answer: which external organisations adopted the candidate's outputs, which policy documents cite work connected to the candidate, and which codebases are reused across institutions. Graph neural networks or transparent rule-based inference can be applied depending on interpretability requirements (Luan et al., 2018; Martínez-Rodríguez et al., 2020). For high-stakes evaluation, the default is rule-transparent path queries whose full evidence chain can be displayed (Rudin, 2019); learned graph embeddings, if used for retrieval, may rank candidates for display but may not constitute evidence. This asymmetry — opaque methods may retrieve, only transparent methods may justify — is a core governance rule of the architecture.

5.6 Decision-support workflow

The committee workflow has five stages: (1) dossier submission by candidate; (2) AI-assisted pre-processing to extract claims, classify outputs, and assemble corroborating evidence; (3) evidence-packet delivery to reviewers with confidence scores and unresolved ambiguities; (4) committee deliberation with inspection, challenge, and qualitative judgement; (5) audit logging of accepted and rejected AI inferences with rationale for future system improvement. Stages 4–5 are exclusively human authorities (Section 7). Following Amershi et al. (2019), the packet interface must support efficient correction (overrides with reason codes feed construct D), disclose uncertainty by default, and make the cost of contesting an inference low — because contestation behaviour is the behavioural foundation of the learning dynamics formalised next.

6. Formal Model of Organisational Evaluation Strength

This section replaces the original linear specification with a bounded, internally consistent propositional model. Three repairs are load-bearing. First, every construct is declared on [0,1], and every composite is bounded by construction, so the weighted strength index is meaningful. Second, the two inconsistent learning equations of the original draft are merged into a single stock–flow law with a closed-form solution. Third — and most consequentially — adoption is brought inside the model: an adoption stock K and adoption date $\tau 0$ drive usable model capability and processable evidence, so the adoption effect can be derived. Figure 4 (Section 6.6) summarises the causal structure; Table 4 defines all notation. The former 'Conditional Adoption-Strength Theorem' is restated as Proposition 1 in Section 8 with assumptions A1–A3 and a proof sketch; nothing in this paper is claimed as a theorem without proof.

6.1 Assumptions and notation

Assumption A1 (Bounded regularity). All index constructs lie in [0,1]; the logistic function σ(z) = 1/(1 + e^(−z)) is continuously differentiable with bounded derivative, so all partial derivatives below exist and are finite. Assumption A2 (Productive adoption inputs). At the margin, adoption investment weakly increases usable capability and processable evidence ($\mu_1 \geq 0$, $\eta_1 \geq 0$ in Eq. 11), with strict inequality once implementation is non-trivial. Assumption A3 (Governance response). Governance quality and bias intensity respond to the adoption stock, dG/dK and dB/dK exist (possibly zero); their signs are institutional choices, not technological constants — this is precisely where ownership architecture (Section 7) enters the mathematics. Time t indexes evaluation cycles (e.g., annual promotion rounds); τ_0 denotes the first cycle with positive adoption investment.

Table 4. Notation: one symbol per construct. Counts use N; the original R/C collisions are eliminated.

Symbol	Construct	Range
S	Organisational evaluation strength (composite index)	[0,1]
Q / F / A-align / L	Decision quality / Fairness / Strategic alignment / Organisational learning stock	[0,1] each
E / M	Evidence richness / AI model capability (task-relative, benchmark-standardised)	[0,1]
G / X / H / B	Governance quality / Explainability / Human oversight quality / Bias intensity	[0,1] (B: 0 = none detected)
Rel / Cons	Mission relevance / Strategy consistency (formerly R, C — renamed to end symbol collision)	[0,1]
O / Tr	Ownership clarity / Traceability (formerly R-trace — renamed)	[0,1]
N-crit / N-amb / N-log	Critical decision steps / Accountability-ambiguous steps / Logged steps (counts)	Non-negative integers
U / D	Effective-use intensity / Documented-feedback intensity (log-defined, see Table 5)	[0,1]
K / I / τ_0	Adoption stock / Per-period adoption investment / Adoption date	[0,1] / [0,1] / cycle index
G* / O*	Governance / ownership sign-change thresholds (solved, not assumed)	(0,1)

6.2 Strength, quality, fairness, and alignment

Organisational evaluation strength is a normatively weighted composite of decision quality, fairness, strategic alignment, and the learning stock. Weights are set by an institutional normative procedure (e.g., analytic hierarchy process with published sensitivity analysis) and are not estimated jointly with the structural coefficients; this choice is stated explicitly to avoid the scale non-identification that arises when index weights and regression coefficients are fit simultaneously.

$$S_t = \alpha Q_t + \beta F_t + \gamma A_t^{align} + \delta L_t. \quad (1)$$

Decision quality increases with evidence richness, task-relative model capability, and governance quality, with two complementarities (evidence × capability; capability × governance) computed on

mean-centred inputs. The logistic bound replaces the original unbounded linear form, whose intercept could place quality outside [0,1] and whose marginal effects never diminished:

$$Q_t = \sigma\big(\theta_0 + \theta_1 E_t + \theta_2 M_t + \theta_3 G_t + \theta_4 \tilde{\mathrm{E}}_t \widetilde{M}_t + \theta_5 \widetilde{M}_t \tilde{G}_t\big). \quad (2)$$

Fairness increases with explainability and human-oversight quality and decreases with bias intensity (Buolamwini & Gebru, 2018). Explainability and bias also enter governance (Eq. 8); the two paths are conceptually distinct — Eq. 3 is the committee-experienced legitimacy of a decision, Eq. 8 is the institution's control capacity — and their empirical separation is tested via the measurement model in Section 10.3 rather than assumed away:

$$F_t = \sigma(\varphi_0 + \varphi_1 X_t + \varphi_2 H_t - \varphi_3 B_t). \quad (3)$$

Strategic alignment increases with mission relevance (alignment of evaluated outputs to the mission taxonomy, in Singapore the four RIE2030 domains; Section 9) and with consistency between evaluation criteria and organisational strategy:

$$A_t^{align} = \sigma(\psi_0 + \psi_1 Rel_t + \psi_2 Cons_t). \quad (4)$$

6.3 Ownership, traceability, and governance

Ownership clarity is the share of critical decision steps with unambiguous accountability. It is the original model's cleanest definition and is retained verbatim apart from notation; the case of zero critical steps is excluded by design (a cycle with no critical steps is not an evaluation cycle):

$$O_t = 1 - \frac{N_t^{amb}}{N_t^{crit}}. \quad (5)$$

Traceability is the logged share of the same critical-step population, so Eqs. 5 and 6 share a denominator by construction:

$$Tr_t = \frac{\sum_{i=1}^{N_t^{crit}} I(\text{Step } i \text{ possesses a verified audit trail})}{N_t^{crit}} = \frac{N_t^{crit,logged}}{N_t^{crit}}. \quad (6)$$

Governance quality is a formative composite index (not a regression) of ownership clarity, traceability, and explainability, net of residual bias. As a formative index, its weights are normative and published; the logistic map keeps the index on [0,1] so the threshold analysis in Section 8 is well-scaled. Governance enters quality contemporaneously in Eq. 2; for estimation, Section 10.3 specifies either lagged governance or simultaneous-equations estimation to break the simultaneity:

$$G_t = \sigma(\omega_0 + \omega_1 O_t + \omega_2 Tr_t + \omega_3 X_t - \omega_4 B_t),$$
$$\omega_0 = -\frac{1}{2}(\omega_1+\omega_2+\omega_3). \quad (7)$$

6.4 Organisational learning: one law, closed form

Learning is a stock–flow process with depreciation. The per-period learning inflow Π combines effective use, ownership clarity, and documented feedback, mapped via a logistic function to ensure smooth (0,1) bounds. The original draft's separate 'confusion' term is removed: accountability confusion is the complement of ownership clarity as operationalised in Eq. 5, and retaining both would induce near-perfect collinearity (Hypothesis H3 tests this redundancy empirically should an independent confusion measure be proposed). The stock updates as a convex combination of retained knowledge and new inflow, so boundedness is preserved by induction:

$$L_{t+1} = \rho L_t + (1-\rho)\prod_t, where\ \prod_t = \sigma(\lambda_0 + \lambda_1 U_t + \lambda_2 O_t + \lambda_3 D_t). \quad (8)$$

Iterating Eq. 8 yields the closed form. Earlier adoption lengthens the summation — the compounding intuition of the original draft — but now with depreciation and saturation: the stock converges to the average inflow rather than diverging, and early low-quality inflows (an immature system) are correctly discounted rather than celebrated. The symbol ρ is therefore interpreted once, as the knowledge-retention rate (1 − ρ = depreciation/forgetting), reconciling the recursion and its solution:

$$L_t = \rho^t L_0 + (1-\rho)\sum_{\tau=0}^{t-1} \rho^{t-1-\tau}\sigma(\lambda_0 + \lambda_1 U_\tau + \lambda_2 O_\tau + \lambda_3 D_\tau). \quad (9)$$

6.5 The adoption bridge — repairing the missing variable

In the original draft, adoption timing/intensity appeared only in the derivative of the concluding 'theorem' and nowhere in Eqs. 1–9, so that the derivative was undefined. The bridge below makes strength an explicit function of adoption. The adoption stock accumulates investment with implementation depreciation (module obsolescence, staff turnover); usable capability and processable evidence respond to the stock, with a capability–governance interaction capturing implementation quality:

$$K_t = (1-\delta_K)K_{t-1} + I_t, \quad (10)$$

$$M_t = \mu_0 + \mu_1 K_t + \mu_2(K_t \times G_t) + \nu_t; \quad E_t = \eta_0 + \eta_1 K_t + \xi_t. \quad (11)$$

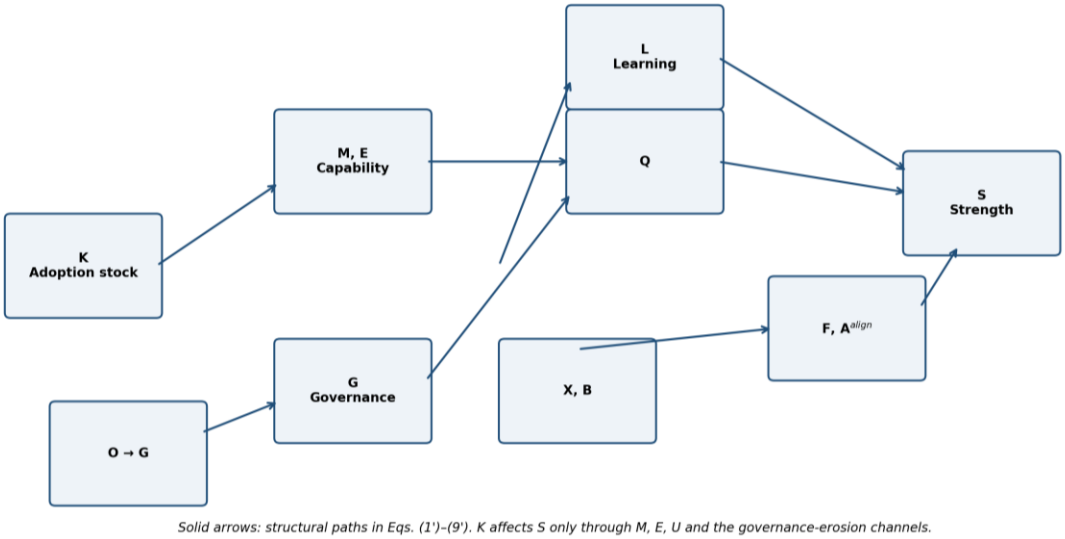
Effective use is likewise a function of the adoption stock and acceptance conditions (Davis, 1989; Venkatesh et al., 2003): U = u(K, facilitating conditions), u increasing in both arguments. With Eqs. 10–11 in place, every arrow in Figure 4 is an equation, and the adoption comparative statics in Section 8 follow the chain rule. Adoption timing τ0 is defined as the first cycle with I > 0; 'earlier adoption'

means a smaller τ0 holding the subsequent investment path fixed — the counterfactual, the corollary in Section 8 evaluates.

6.6 Hypotheses and operationalisation

Figure 3 displays the implied causal graph. Five hypotheses, each tied to an equation and a validation stage (Section 10.2), convert the framework into a falsifiable program:

Figure 3. Causal Structure of the formal model (DAG)



Note: Causal structure of Eqs. (1)–(11). The adoption stock K reaches strength S only through capability/evidence (M, E), effective use and learning (U, L), and the governance-erosion channels (G, B) — the two races whose net sign Proposition 1 characterises.

H1 (Complementarity). Evidence richness amplifies the quality effect of model capability ($\theta_4 > 0$ in Eq. 2). Rich evidence without capable models overwhelms committees; capable models without rich evidence merely automate prestige heuristics faster.

H2 (Governance moderation with threshold). The marginal quality effect of capability increases with governance ($\theta_5 > 0$), and there exists G* such that capability improvements fail to raise quality at or below G*. Mechanism: ungoverned capability inherits and scales prestige bias (B rises), offsetting accuracy gains.

H3 (Ownership two-path effect). Ownership clarity raises strength through governance (Eq. 7) and through learning inflow (Eq. 8); an independently measured accountability-confusion construct adds no explanatory power once O is controlled (discriminant validity test).

H4 (Compounding/timing). Holding total investment fixed, earlier τ0 predicts higher L and S at every finite horizon (tested as the corollary's comparative static, not as a raw early-vs-late mean difference, which is selection-contaminated; see Section 10.3).

H5 (Bias erosion and explainability buffering). Bias intensity lowers strength through both fairness (Eq. 3) and governance (Eq. 7); the total negative effect is stronger when explainability X is low (buffering interaction).

Table 5. Operationalisation: indicators, sources, and reliability/validity plans for all constructs.

Construct	Operational indicator	Source/reliability plan
E Evidence richness	Non-paper evidence share × completeness rate, standardised	Registry logs; convergent validity vs. expert ratings
M Model capability	Benchmark F1 (extraction/classification) on a fixed annotated corpus	Stage-1 benchmark; fixed corpus avoids time-trend collinearity
G Governance quality	Formative index of O, Tr, X, (1−B); published weights	Governance audit; weight sensitivity analysis
X Explainability	Committee sufficiency scale (3 items) + explanation coverage	Survey + logs; Cronbach α, inter-rater ICC
H Human oversight	Substantive review rate; reasoned override rate	Workflow logs; CFA vs. O for discriminant validity
B Bias intensity	Group outcome gaps + prestige-proxy leakage rate	Adversarial audit (Stage 4)
Rel / Cons	Mission-taxonomy agreement (classifier + human review, κ); criteria–strategy semantic alignment	Knowledge-graph tags; double-coding reliability
O / Tr	1 − N-amb/N-crit; N-log/N-crit	Step inventory (double-coded); system logs
U Effective use	Meetings substantively using evidence packets (log-defined; never post-hoc 'usefulness')	System logs; removes the original circular definition
D Documented feedback	Reason-coded acceptance/override share	Audit logs
K Adoption stock	Depreciated module-coverage accumulation; τ0 = first live cycle	Deployment records
S Strength	AHP-weighted composite (Eq. 1); external committee-quality validation	Sensitivity (weight-perturbation) analysis

7. Ownership and Responsibility Architecture

AI adoption generates organisational strength only when accountability is explicitly architected. A minimum governance structure assigns five distinct roles (Table 6). The design aligns with NIST AI RMF's governing function and ISO/IEC 42001 management-system requirements (NIST, 2023; ISO, 2023) and with organisational-AI research showing that decision rights, not model sophistication, predict adoption success (Fountaine et al., 2019; Shrestha et al., 2019). Role ambiguity and overlap are the primary sources of governance failure; the ownership-clarity metric (Eq. 6) provides the measurable index of design quality, computed by double-coded inventory of critical steps each cycle (N-crit, N-amb), with disagreements adjudicated and reported as the index's measurement error.

Table 6. Five accountable roles (RACI summary)

Role	Responsibility (accountable for)	Key artefacts
System owner	Platform configuration, data boundaries, model deployment	Configuration log; data-boundary register
Decision owner	Final committee decisions and documented rationale	Signed decision rationale (human authority; non-delegable)
Model steward	Calibration, monitoring, and retraining	Calibration reports; benchmark

	controls	results (M)
Audit lead	Bias testing, override review, appeals logging	Bias audits (B); appeals register; Tr computation
Domain expert reviewer	Field-specific interpretation of substantive evidence	Annotated evidence packets; ambiguity resolutions

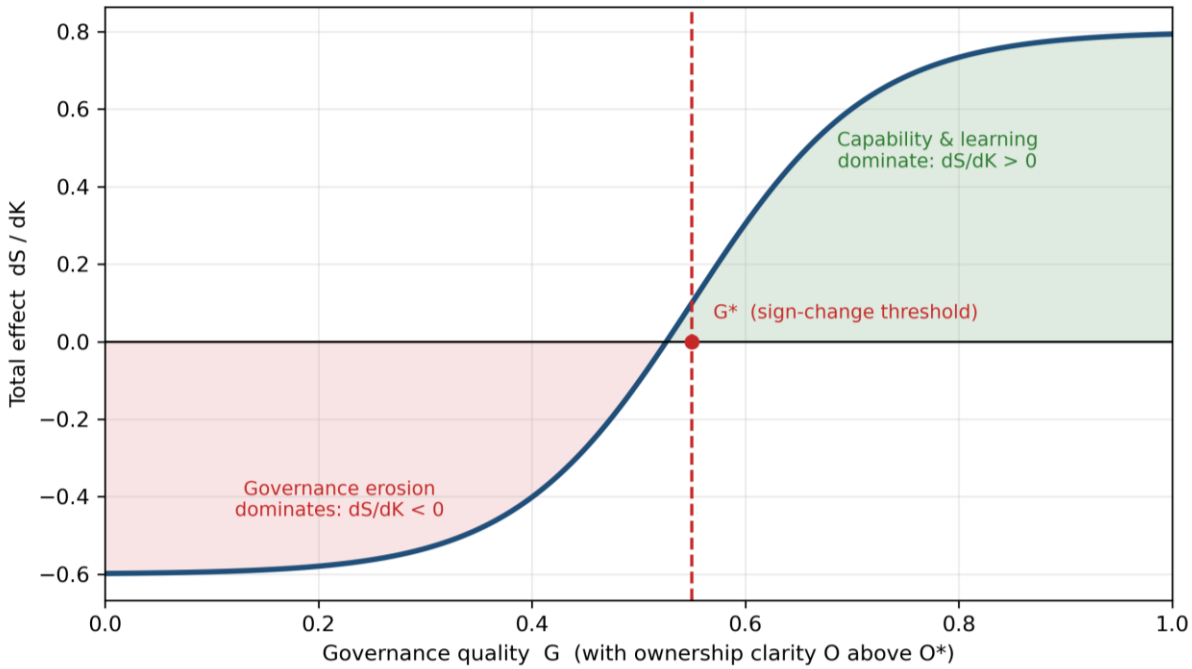
Note: Every critical decision step maps to exactly one accountable role; unmapped or multiply-mapped steps increment N-amb in Eq. (6).

The Singapore Statement's four principles map onto auditable checkpoints rather than slogans: honesty → provenance visibility (every packet claim traceable to a source span); accountability → non-delegable human decision ownership plus appeals logging; professional courtesy and fairness → scheduled bias audits with affected-group analysis; good stewardship → data minimisation, retention limits, and PDPA-compliant dossier handling (Jobin et al., 2019; Resnik & Shamoo, 2011; World Conferences on Research Integrity, 2010). Section 10.4 states the corresponding evaluatee rights. Governance quality G (Eq. 7) is thus not an attitude; it is the measured throughput of these checkpoints.

8. The Adoption–Strength Relationship: Proposition 1

With the adoption bridge in place, the adoption effect is a total derivative through two competing channels. Figure 4 visualises the sign change; Figure 5 (Section 10) shows how the thresholds are calibrated and tested.

Figure 4. Conditional adoption–strength: Sign of dS/dK Across Governance Quality



Note: the total effect dS/dK as a function of governance quality G (holding O above O*). Below G*, governance erosion dominates, and adoption reduces strength; above G*, capability and learning dominate. G* solves dS/dK = 0 in Eq. (12) and is calibrated empirically (Section 10.2).

Proposition 1 (Conditional adoption–strength). Under A1–A3 and Eqs. (1)–(11), the total effect of the adoption stock on organisational evaluation strength decomposes as:

$$\frac{dS_t}{dK_t} = \alpha \frac{\partial Q_t}{\partial M_t}(\mu_1 + \mu_2 G_t) + \alpha \frac{\partial Q_t}{\partial E_t}\eta_1 + \delta \frac{\partial L_t}{\partial K_t} + \left[\alpha\left(\frac{\partial Q_t}{\partial G_t} + \mu_2 K_t \frac{\partial Q_t}{\partial M_t}\right)\right]\frac{dG_t}{dK_t} + \beta \frac{\partial F_t}{\partial B_t}\frac{dB_t}{dK_t}, \quad (12)$$

where the first three terms (capability, evidence, and learning channels) are non-negative under A2, and the final two terms (governance-erosion channels) take the sign of dG/dK and −dB/dK respectively. Consequently: (i) if adoption is accompanied by governance investment such that dG/dK ≥ 0 and dB/dK ≤ 0 — operationally, sustained G > G* and O > O*, where G* solves Eq. (12) = 0 — then dS/dK > 0; (ii) if adoption dilutes accountability (dG/dK < 0) or amplifies bias (dB/dK > 0) sufficiently that the erosion terms exceed the capability terms in absolute value, then dS/dK < 0. Proof sketch: by A1, all components are differentiable; apply the chain rule through Eqs. (2), (3), (7), and (11) and collect terms; signs follow from σ′ > 0, A2, and the stated governance-response signs. Thresholds G* and O* are therefore implicit sign-change conditions of Eq. (12), estimable once coefficients are calibrated — not exogenous constants.

Corollary (Early adoption, precisely stated). Fix an investment path {I} and governance responses satisfying Proposition 1(i). Advancing adoption by one cycle (τ0 → τ0 − 1, investment path shifted accordingly) weakly increases L and S at every subsequent horizon, strictly if inflows are positive. Proof sketch: by Eq. (9), the earlier path adds one non-negative inflow term, and re-weights retained terms by ρ; monotonicity of Eq. (1) in L completes the argument. The corollary — not the slogan — is the falsifiable content of 'the sooner, the stronger': it holds the investment path fixed, which naive early-vs-late comparisons do not (Section 10.3). Conversely, under Proposition 1(ii), earlier adoption compounds disorder: the same summation accumulates confusion when governance erodes. Timing amplifies the regime; it does not choose it. Ownership architecture (Section 7) chooses it.

9. Singapore-Aligned Deployment Context

Singapore's Research, Innovation and Enterprise 2030 (RIE2030) plan, announced on 5 December 2025 by Senior Minister Lee Hsien Loong, commits S$37 billion over five years — a 32% increase from RIE2025's S$28 billion — to advance research, innovation, and enterprise (National Research Foundation Singapore, 2025). RIE2030 prioritises four strategic domains: Human Health and Potential; Manufacturing, Trade and Connectivity; Urban Solutions and Sustainability; and Smart Nation and Digital Economy (National Research Foundation Singapore, 2025). These four domains function as an upper-level mission taxonomy. In the proposed system, each Impact Matrix category can be tagged by RIE domain, enabling both institutional and national-level analytics and giving the relevance construct (Rel) a concrete, auditable coding scheme. Singapore also authored the Singapore Statement on Research Integrity (2010), the first international framework from the Second World Conference on Research Integrity, representing 340 participants from 51 countries (Resnik & Shamoo, 2011; World Conferences on Research Integrity, 2010). The Statement's four principles map directly onto the governance architecture (Section 7): honesty governs provenance visibility, accountability governs audit

trails, fairness governs bias monitoring, and good stewardship governs data minimisation and long-term oversight. Two scope cautions apply. First, RIE2030 is a deployment context, not evidence: policy alignment supplies the Rel taxonomy and a plausible source of exogenous adoption-timing variation (funding-cycle instruments, Section 10.3), but no claim in Sections 6 or 8 is validated by Singapore's policy commitment. Second, portability beyond a high-governance-density city-state system is an empirical question (Section 10.5), not an assumption.

10. Discussion, Limitations, and Future Research

10.1 Contributions to decision sciences

This study contributes at three levels. At the normative level, it extends responsible research assessment by grounding broader evidence recognition in decision sciences theory (CoARA, 2022; Curry et al., 2022; Hicks et al., 2015; San Francisco Declaration on Research Assessment, 2013). At the systems level, it specifies how AI modules — LLMs, classifiers, knowledge graphs, and explainability pipelines — can reduce the operational burden that has impeded adoption of broader evaluation methods (Devlin et al., 2019; Vaswani et al., 2017; Wang et al., 2024), under the governance rule that opaque methods may retrieve but only transparent methods may justify. At the organisational level, it provides a bounded formal model in which governance quality, ownership clarity, and adoption timing jointly determine whether AI strengthens or weakens institutional decision capacity — with the timing claim demoted from slogan to corollary, and the sign conditions made estimable. The model also reframes a familiar implementation debate: 'pilot first vs. deploy early' is, in Eq. (12) terms, a bet on the sign of dG/dK and dB/dK under one's ownership architecture, which can be inspected before it is bet upon.

10.2 Empirical roadmap (hypothesis-linked)

Empirical validation proceeds in four stages, each mapped to the hypotheses it tests (Figure 5). Stage 1 benchmarks information-extraction and classification accuracy against an annotated narrative-CV corpus (estimating M; testing H1's inputs). Stage 2 validates knowledge-graph quality through sampled triple verification (estimating E; precondition for H1). Stage 3 runs human-factors studies comparing AI-assisted and manual committee workflows on speed, agreement, fairness perceptions, and override behaviour (Fountaine et al., 2019; Shrestha et al., 2019; Venkatesh et al., 2003) (estimating X, H, U, D; testing H3 and H5). Stage 4 conducts bias and robustness tests probing prestige-marker inheritance with adversarial dossiers (Buolamwini & Gebru, 2018; Mitchell et al., 2019) (estimating B; testing H2 and H5). Threshold calibration (G*, O*) uses Stage 3–4 estimates in Eq. (12) with bootstrap confidence sets; the early-adoption corollary (H4) is tested only under the identification strategies of Section 10.3, never by raw cohort comparison. A follow-on empirical study is planned using anonymised faculty dossier data from Singapore's Autonomous Universities, subject to the ethics protocol in Section 10.4.

Figure 5. Research design and empirical validation roadmap.

Note: Each stage estimates named constructs and tests named hypotheses; audit results feed back to threshold calibration (G*, O*) and model revision.

10.3 Identification, measurement, and validity

Identification Strategy. Naive early-versus-late comparisons conflate adoption timing with pre-existing governance quality. To achieve identification, we employ a two-pronged empirical strategy:

(1) *Staggered Difference-in-Differences (DiD):* We exploit the idiosyncratic rollout across autonomous universities and distinct faculties (e.g., Medicine vs. Humanities) using modern staggered DiD estimators robust to heterogeneous treatment effects (Callaway & Sant'Anna, 2021; Sun & Abraham, 2021). The parallel-trends assumption is formally tested against multi-period pre-adoption evaluation logs.

(2) *Micro-level Institutional IV:* Instead of macro-level research funding (which directly violates the exclusion restriction by financing laboratory infrastructure and researcher salaries), we instrument adoption timing τ_0 using pre-scheduled enterprise HR/IT legacy platform retirement and cloud-migration cycles. These contractual transition windows are locked in multi-year procurement schedules negotiated at the general university administration level years prior, making them plausibly orthogonal to individual academic departments' research potential and committee evaluation biases, yet highly predictive of the exact cycle when the AI-assisted dossier system can be technically integrated.

10.4 Ethics and evaluatee rights

AI-assisted evaluation of faculty dossiers engages the evaluated as rights-holders, not data sources. Four commitments, each tied to a Singapore Statement principle (Jobin et al., 2019; Resnik & Shamoo, 2011; World Conferences on Research Integrity, 2010), govern deployment: (1) Honesty/provenance — candidates may inspect every machine inference about them with its source evidence (the same packet the committee sees). (2) Accountability/contestability — final decisions are non-delegably human (Decision Owner, Table 6); candidates may contest inferences and appeal decisions, with appeals logged and reviewed by the Audit Lead; override and appeal rates feed D and Tr. (3) Fairness — scheduled bias audits report affected-group analyses; prestige-proxy leakage is a release-blocking defect. (4) Stewardship — dossier data are minimised, purpose-bound, retention-limited, and processed under applicable data-protection law (Singapore PDPA; GDPR where relevant); no dossier content is used for model training without separate consent and ethics approval. The empirical study (Section 10.2) requires institutional review board approval and data-sharing agreements before any dossier access.

10.5 Limitations

Three principal limitations apply. First, the formal model specifies relationships and threshold conditions but does not yet estimate coefficients; the equations provide a testable structure rather than calibrated predictions, and Proposition 1's thresholds are only as credible as the Stage 1–4 estimates that will populate them. Second, field heterogeneity is handled by probabilistic classification and field-normalised analytics (Waltman & van Eck, 2019) but not by discipline-specific model parameters; disciplines producing impact through different pathways and timescales may require hierarchical extensions. Third, the deployment context is Singapore's high-capacity governance environment; in systems with weaker audit infrastructure, the governance-erosion channel may dominate at adoption scales where Singapore institutions remain above G* — the model predicts this divergence (Proposition 1(ii)), but cross-system estimation is required to locate the thresholds comparatively.

11. Conclusion

Universities need a new operating system for academic brain power because the current system confuses journal prestige with intellectual contribution and leaves major forms of knowledge translation structurally invisible and unrewarded (Garfield, 1955; San Francisco Declaration on Research Assessment, 2013; Seglen, 1997). Responsible research assessment movements have established the normative case, but adoption has lagged because broader evaluation is administratively difficult at scale (Bordignon et al., 2023; CoARA, 2022). AI and agentic tools change the feasible frontier: evidence-centred evaluation becomes operationally viable when AI assists committees rather than replacing them (Devlin et al., 2019; Vaswani et al., 2017; Wang et al., 2024). Proposition 1 (Section 8) captures the governance condition precisely: organisations that implement AI with explicit ownership and responsibility accumulate capability, coordination, and learning advantages, and — by the corollary — gain more from each earlier cycle; those that implement without these conditions compound existing confusions at a greater scale. The practical implication is therefore conditional, and the condition is inspectable in advance: audit G and O, assign the five accountable roles, and only then does 'the sooner, the stronger' become a strategy rather than a slogan. For Singapore, this means aligning AI-assisted evaluation governance with the accountability principles of the Singapore Statement (World Conferences on Research Integrity, 2010) and the strategic priorities of RIE2030 (National Research Foundation Singapore, 2025). For universities globally, it means treating evaluation reform as a systems-design and governance-engineering challenge, not merely a normative aspiration.

Acknowledgments and disclosure: The Dual Dashboard / Impact Matrix taxonomy originated in an unpublished policy white paper by the first author (Chan, 2026). This self-citation is disclosed; the taxonomy is presented here as a design proposal grounded in Bornmann (2013) and Waltman and van Eck (2019) and subject to the validation programme in Section 10.2. Grey-literature sources (Chan, 2026; United Nations University, 2026) are flagged where cited and are never the sole support for a theoretical claim. No faculty dossier data were accessed for this conceptual paper.

References

Albert, M. A., Qayyum, A., MacKinnon, K., Ramos, J., de Paula Dídimo, G., Kalkmann, G. F., ... & Cobey, K. D. (2025). What goes in a funder's Narrative CV?: A scoping review. *F1000Research, 14*, Article 1345. https://doi.org/10.12688/f1000research.170507.1

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Article 3, pp. 1–13). Association for Computing Machinery. https://doi.org/10.1145/3290605.3300233

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Aubert Bonn, N., Morris, J. P., Sapcariu, S., & Stroobants, K. (2025). Are narrative CVs contributing towards shifting research culture? Workshop report from the 2023 Recognition and Rewards Festival. *F1000Research, 13*, Article 332. https://doi.org/10.12688/f1000research.146108.2

Bordignon, F., Chaignon, L., & Egret, D. (2023). Promoting narrative CVs to improve research evaluation? A review of opinion pieces and experiments. *Research Evaluation, 32*(2), 313–320. https://doi.org/10.1093/reseval/rvad013

Bornmann, L. (2013). What is societal impact of research and how can it be assessed? A literature survey. *Journal of the American Society for Information Science and Technology, 64*(2), 217–233. https://doi.org/10.1002/asi.22803

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research, 81*, 77–91. http://proceedings.mlr.press/v81/buolamwini18a.html

Callaway, B., & Sant'Anna, P. H. (2021). Difference-in-differences with multiple time periods. *Journal of econometrics*, *225*(2), 200-230. https://doi.org/10.1016/j.jeconom.2020.12.001

Chan, C., Zhao, Y., & Zhao, J. (2024). A case study on assessing AI assistant competence in narrative interviews. *F1000Research, 13*, Article 601. https://doi.org/10.12688/f1000research.151952.2

Chan, L. C. W. J. (2026). *A new operating system for academic brain power: A policy white paper and proposed Singapore Standard for Faculty Evaluation* (Finamatrix Policy Paper No. 1). Finamatrix Pte Ltd. [Grey literature — unpublished white paper]

Coalition for Advancing Research Assessment. (2022). *Agreement on reforming research assessment*. https://coara.eu/agreement/the-agreement-full-text/

Curry, S., Gadd, E., & Wilsdon, J. (2022). Harnessing the metric tide: Indicators, infrastructures, and priorities for responsible research assessment. *Research Evaluation, 31*(3), 375–386. https://doi.org/10.1093/reseval/rvac015

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly, 13*(3), 319–340. https://doi.org/10.2307/249008

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Fountaine, T., McCarthy, B., & Saleh, T. (2019). Building the AI-powered organization. *Harvard Business Review, 97*(4), 62–73.

Garfield, E. (1955). Citation indexes for science: A new dimension in documentation through association of ideas. *Science, 122*(3159), 108–111. https://doi.org/10.1126/science.122.3159.108

Garfield, E. (2006). The history and meaning of the journal impact factor. *JAMA, 295*(1), 90–93. https://doi.org/10.1001/jama.295.1.90

Grzybowski, A., & Patryn, R. (2017). Impact factor: Universalism and reliability of assessment. *Clinics in Dermatology, 35*(3), 331–334. https://doi.org/10.1016/j.clindermatol.2016.07.012

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature, 520*(7548), 429–431. https://doi.org/10.1038/520429a

International Organization for Standardization. (2023). *ISO/IEC 42001:2023 — Information technology — Artificial intelligence — Management system*. https://www.iso.org/standard/81230.html

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Larivière, V., & Sugimoto, C. R. (2019). The journal impact factor: A brief history, critique, and discussion of adverse effects. In W. Glänzel, H. F. Moed, U. Schmoch, & M. Thelwall (Eds.), *Springer handbook of science and technology indicators* (pp. 3–24). Springer. https://doi.org/10.1007/978-3-030-02511-3_1

Lee, D. Y. H., Parker, A. J., Norbury, C. F., & Shanks, D. R. (2026). Validating AI-assisted evaluation of open science practices in brain sciences: ChatGPT, Claude and human expert comparisons. *Royal Society Open Science, 13*(2), Article 250381. https://doi.org/10.1098/rsos.250381

Luan, Y., He, L., Ostendorf, M., & Hajishirzi, H. (2018). Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 3219–3232). Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1360

Martínez-Rodríguez, J. L., Hogan, A., & López-Arévalo, I. (2020). Information extraction meets the semantic web: A survey. *Semantic Web, 11*(2), 255–335. https://doi.org/10.3233/SW-180333

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220–229). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287596

Moher, D., Naudet, F., Cristea, I. A., Miedema, F., Ioannidis, J. P. A., & Goodman, S. N. (2018). Assessing scientists for hiring, promotion, and tenure. *PLOS Biology, 16*(3), Article e2004089. https://doi.org/10.1371/journal.pbio.2004089

National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (NIST AI 100-1). https://doi.org/10.6028/NIST.AI.100-1

National Research Foundation Singapore. (2025, December 5). *Singapore unveils S$37 billion RIE2030 Plan* [Press release]. https://www.nrf.gov.sg

Resnik, D. B., & Shamoo, A. E. (2011). The Singapore Statement on Research Integrity. *Accountability in Research, 18*(2), 71–75. https://doi.org/10.1080/08989621.2011.557296

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

San Francisco Declaration on Research Assessment. (2013). *San Francisco Declaration on Research Assessment*. American Society for Cell Biology. https://sfdora.org/read/

Seglen, P. O. (1997). Why the impact factor of journals should not be used for evaluating research. *BMJ, 314*(7079), 498–502. https://doi.org/10.1136/bmj.314.7079.497

Shrestha, Y. R., Ben-Menahem, S. M., & von Krogh, G. (2019). Organizational decision-making structures in the age of artificial intelligence. *California Management Review, 61*(4), 66–83. https://doi.org/10.1177/0008125619862257

Shu, F., Liu, S., & Larivière, V. (2022). China's research evaluation reform: What are the consequences for global science? *Minerva, 60*(3), 329–347. https://doi.org/10.1007/s11024-022-09468-7

Sun, L., & Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of econometrics, 225*(2), 175-199. https://doi.org/10.1016/j.jeconom.2020.09.006

United Nations University. (2026). *AI-assisted CV screening: A research prototype* [Blog post]. https://c3.unu.edu/blog/ai-assisted-cv-screening-a-research-prototype

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems, 30* (pp. 5998–6008).

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly, 27*(3), 425–478. https://doi.org/10.2307/30036540

Waltman, L., & van Eck, N. J. (2019). Field normalization and indicators. In W. Glänzel, H. F. Moed, U. Schmoch, & M. Thelwall (Eds.), *Springer handbook of science and technology indicators* (pp. 281–307). Springer.

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., & Wen, J.-R. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science, 18*(6), Article 186345. https://doi.org/10.1007/s11704-024-40231-1

World Conferences on Research Integrity. (2010). *Singapore Statement on Research Integrity*. Second World Conference on Research Integrity, Singapore, 21–24 July 2010. https://www.wcrif.org/guidance/singapore-statement

Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., ... & Gui, T. (2025). The rise and potential of large language model based agents: A survey. *Science China Information Sciences, 68*(2), Article 121101. https://doi.org/10.1007/s11432-024-4235-8