



Sovereign: A Socio-Technical Framework for Human Capability and Responsibility Coverage in AI-Assisted Work

Emily J. Barnes 

Capitol Technology University, Laurel, Maryland, United States

Abstract: Generative artificial intelligence (AI) can improve immediate task performance without establishing whether human capability develops, remains stable, or deteriorates. Sovereign proposes responsibility-capability alignment as a theoretical account of whether available capability covers consequential responsibilities under AI disruption. A provisional, time-indexed measurement and allocation bridge connects capability trajectories with coverage of specified responsibilities at their required dates; component profiles take priority over optional averages. Responsibility-capability alignment specifies proposed pathways through which vulnerability can increase despite stable individual competence, and substantial assistance benefits can coexist with adequate collective control. A purposive conceptual synthesis integrates research on technology-supported cognition, expertise, human-automation interaction, productivity, and AI governance. Four complementary modules support that central mechanism: the Capability Dynamics Model represents conditional capability trajectories; the Assisted Performance Gap Index describes assisted-unaided performance differences; the Cognitive Sovereignty Model represents exercised control capacity; and the Dependency Exposure Model organizes capability shortfalls and structural exposures against explicit workflow requirements. AI-supported learning and capability erosion are competing possibilities; a performance gap alone establishes neither deterioration nor unsafe dependency. Bounded equations, comparisons with established theories, contrasting analytical cases, and falsifiable propositions specify the contribution. No original empirical testing, parameter estimation, or validation of the proposed models is reported. Subsequent studies are intended to assess measurement quality, compare alternative specifications, estimate relationships, and determine predictive value against independently observed outcomes. The contribution is an analytically explicit foundation for research on capability governance, rather than an established risk assessment instrument.

Keywords: generative artificial intelligence, conceptual framework, human capability, cognitive offloading, AI dependency, human oversight

INTRODUCTION

Generative AI changes the relationship between task completion and the knowledge required to complete a task independently. Studies of professional writing, customer support, and software development document productivity benefits in specific settings [1, 2, 3]. Improved outputs, however, do not alone reveal what users learn or retain. An organization can observe rising productivity while independent competence improves, stagnates, or declines. Distinguishing those possibilities matters whenever people remain responsible for evaluating outputs, resolving exceptions, or maintaining operations during disruption. The relevant governance question concerns the capability required for those responsibilities, rather than the preservation of every task formerly performed manually.

The distinction between supported performance and durable learning has an established intellectual history. Salomon [4] distinguished cognitive effects during computer use from changes that persist beyond the interaction. Cognitive offloading research examines how external resources alter the demands of a task [5]. Human-automation research identifies difficulties associated with reliance, monitoring, and intervention [6, 7]. Sovereign builds on those traditions. The proposed contribution is a time-indexed measurement and allocation specification connecting retained capability with coverage of consequential responsibilities at their required dates. Available expertise, task requirements, and exercised control determine the proposed alignment; assisted-unaided performance differences provide descriptive context.

The central argument is conditional. Assistance that replaces reasoning, verification, and engagement with unfamiliar problems may weaken opportunities for capability development. Assistance that supplies feedback, exposes errors, or enables more demanding practice may strengthen those opportunities. Evidence of learning in customer support and evidence of impaired unassisted performance under some educational uses illustrate why neither trajectory should be assumed universally [2, 8]. Task characteristics, baseline competence, instructional design, and organizational incentives are plausible boundary conditions. The proposed framework therefore treats capability change as an empirical question.

Sovereign comprises four complementary modules: the Capability Dynamics Model (CDM), Assisted Performance Gap Index (APGI), Cognitive Sovereignty Model (CSM), and Dependency Exposure Model (DEM). CDM is an explanatory model of change in individual capability within a defined domain. APGI is a descriptive index of a contemporaneous difference between assisted and unaided performance. CSM is a component profile of the responsible team's exercised capacity to reason, verify, intervene, recover, and reconstruct decisions. DEM is a provisional exposure profile anchored to workflow requirements, available role coverage, and structural exposures; control effects remain separately testable. Their combination provides a research architecture, not an estimated causal system.

The article is theoretical and reports no original empirical tests of the four modules. Its contribution comprises construct distinctions, provisional mathematical specifications, conditional propositions, and a program for subsequent empirical evaluation. The equations make assumptions inspectable; their mathematical consistency does not establish empirical validity. Later studies are intended to determine whether the constructs can be measured reliably, whether the proposed relationships hold, and whether the integrated framework adds value beyond simpler explanations. Findings may support, revise, or reject individual models without requiring acceptance or rejection of the entire framework.

THEORETICAL FOUNDATIONS AND EVIDENTIAL BOUNDARIES

Productivity, Learning, and Output Compression

Productivity and learning require different observations. A faster or higher-quality assisted response demonstrates performance under an assistance condition. Evidence of retained capability requires assessments over time, including performance after the focal assistance is unavailable. Professional writing and programming experiments provide reasons to

investigate assistance benefits but cannot, without appropriate follow-up assessments, establish lasting capability change [1, 3]. Workplace evidence also cautions against assuming that assistance merely conceals deficits: Brynjolfsson et al. [2] report patterns consistent with learning alongside productivity gains. Educational evidence identifies a contrasting pathway when unrestricted assistance substitutes for productive engagement [8].

Output compression denotes lower between-person dispersion in assisted performance than in comparable unaided performance. Larger gains among initially weaker performers may produce compression, but average gains do not establish reduced variance. Ceiling effects, restricted task difficulty, and scoring rubrics can also compress observed scores. Even genuine compression does not establish convergence or divergence in underlying capability. The construct concerns a distribution of outputs; APGI concerns a within-person performance contrast. Keeping those levels separate prevents an aggregate pattern from becoming evidence of individual deterioration.

Capability Formation and Automation Reliance

Deliberate practice research emphasizes effortful activities directed toward improvement, including feedback and correction [9]. Transfer research further distinguishes success on familiar tasks from the ability to use knowledge under changed demands [10]. Those perspectives motivate attention to practice, verification, exception handling, and transfer in Sovereign. They do not establish that the four activities are independent, exhaustive, or adequately represented through an additive equation. The proposed functional form remains a modeling choice open to comparison with alternatives.

Automation can redistribute work while leaving people responsible for difficult interventions. Bainbridge [6] identified tensions between automated routine operation and demands placed on human operators. Parasuraman and Riley [7] distinguished appropriate use from misuse, disuse, and abuse of automation. These antecedents locate the proposed capability problem within established human-automation research. Sovereign extends that problem formulation toward repeated generative-AI use and capability development. A reduced need for routine execution can be beneficial; concern arises when responsibilities remain while the capability needed to discharge them is inadequately maintained.

Learning displacement denotes a hypothesized substitution pathway: repeated access to expert-like outputs accompanies displacement of the effort and feedback needed to develop relevant human capability. The mechanism concerns displaced learning opportunities rather than ordinary knowledge sharing. Evidence for the pathway would require observations of displaced learning activity and subsequent capability consequences. Strong assisted performance alone is insufficient. An opposing pathway, capability scaffolding, occurs when assistance supports explanation, correction, and progressively more demanding independent work. Both pathways can coexist across tasks or stages of expertise.

Dependency and Substantive Control

AI governance concerns socio-technical arrangements as well as model behavior. The National Institute of Standards and Technology (NIST) [11, p. 1] states that “AI systems are

inherently socio-technical in nature.” The quotation supports assessment of human capabilities and organizational authority alongside technical performance. A capability-centered framework complements that orientation through attention to the people and teams expected to supervise deployment. Formal assignment of human oversight does not establish effective oversight, a concern developed in Green’s [12] analysis of government algorithm policies. Sovereign consequently defines control through demonstrated actions under specified conditions, rather than policy documentation alone.

Technical vulnerabilities supply plausible pathways from dependence to operational harm. Pearce et al. [13] documented insecure generated code under selected security-relevant scenarios. OWASP Foundation [14] identifies application risks including prompt injection, improper output handling, and excessive agency. Such sources motivate exposure and control constructs, but do not demonstrate that capability erosion causes those vulnerabilities. Technical debt likewise concerns deferred liabilities and their management, rather than AI provenance alone [15]. Human capability, system architecture, permissions, and available recovery arrangements must therefore remain analytically distinguishable.

Table 1: Selected Evidence and Its Role in Framework Development

Source and Design	Relevant Contribution	Boundary on Inference
Noy & Zhang [1]; writing experiment	Improved assisted productivity and output quality in professional writing tasks.	Does not establish long-term capability erosion.
Peng et al. [3]; programming experiment, preprint	Faster completion of a bounded JavaScript task with Copilot.	A task-level productivity result does not demonstrate failure to learn.
Brynjolfsson et al. [2]; customer-support field study	Productivity gains, heterogeneous benefits, and evidence consistent with learning.	Counters a universal erosion account; does not establish identical effects across occupations.
Bastani et al. [8]; mathematics field experiment	Unguided assistance can impair later unassisted performance; learning-oriented safeguards alter outcomes.	Educational setting and assistance design limit transfer to professional work.
Pearce et al. [13]; code-generation security evaluation	Generated code can contain vulnerabilities in selected security scenarios.	Does not estimate contemporary population-wide vulnerability rates or human atrophy.
NIST [11]; governance framework	Organizes socio-technical AI risk management.	Normative framework, not empirical validation of Sovereign.
OWASP Foundation [14]; technical risk guidance	Identifies application exposure pathways and mitigation categories.	Technical guidance, not estimated effects of human capability on failures.

Note: The sources motivate constructs and boundary conditions. None evaluates the complete Sovereign framework.

CONCEPTUAL APPROACH AND SCOPE

Theory Development

The article uses purposive conceptual synthesis and model development. Following Jaakkola’s [16] discussion of conceptual research, the contribution depends on explaining the selection and role of theoretical perspectives and the logic connecting proposed constructs. Technology-supported cognition supplies the performance-learning distinction; expertise research motivates capability-building activities; human-automation research

motivates reliance and intervention concerns; AI governance and security sources motivate workflow exposures and control arrangements. Empirical studies establish relevant phenomena and counterexamples. They are not pooled into an effect estimate or treated as observations validating Sovereign.

Sources are selected for direct relevance to those theoretical roles, methodological transparency, and the inclusion of both beneficial and adverse assistance outcomes. Table 1 makes the principal evidential contributions and limits explicit. The synthesis is purposive rather than exhaustive; no reproducible systematic-search protocol, comprehensive screening process, coded research dataset, or meta-analysis is claimed. Industry forecasts are not used as evidence of observed capability decline. The resulting framework reflects an interpretive theoretical construction, and alternative selections or integrations may yield different specifications.

The development logic proceeds from distinctions to relationships. First, supported performance, retained capability, dependency exposure, and exercised control are defined separately. Second, candidate mechanisms connect those constructs across time and levels of analysis. Third, equations express bounded representations of the proposed mechanisms and composites. Finally, propositions identify observable outcomes that could challenge the theory. Competing explanations are evaluated analytically through contrasting cases: an assistance advantage without capability deterioration, declining role coverage without individual deterioration, and equal control averages with different intervention permissions. A direct gap-to-exposure mapping fails the first contrast; a capability-only mapping fails the second; uniform control attenuation obscures the third. Those failures motivate requirement-referenced shortfalls and separate exposure-control relationships. The cases test internal explanatory discrimination, not empirical truth. Table 3 records the source-to-construct-to-proposition chain and distinguishes inherited claims from author-proposed equations. Definitions, assumptions, mathematical implications, literature findings, and empirical predictions retain different evidential status throughout the article.

Units, Boundaries, and Contribution

The individual unit is a person performing a defined class of cognitive tasks; the organizational unit is a specified workflow and its responsible team. Time refers to a declared observation interval appropriate to the domain. Organizations containing heterogeneous workflows should be represented through separate workflow profiles before any institutional summary is attempted. A software review workflow, for example, may require different retained capabilities and recovery arrangements from an administrative drafting workflow. No universal capability scale or cross-sector risk threshold is assumed.

The framework is most relevant where AI assists consequential knowledge work and human or institutional responsibility for verification, intervention, or continuity remains. Capability means competence relevant to those responsibilities. Unaided assessment removes the focal generative-AI assistance while preserving explicitly permitted reference materials, established tools, and necessary accessibility accommodations. Such assessment is not a general test of intelligence or personal worth. Tasks permanently and safely delegated to reliable infrastructure need not remain individually executable merely to reduce an index value.

The proposed theoretical contribution is responsibility-capability alignment: organizational vulnerability depends on whether available human and technical arrangements satisfy the requirements attached to consequential responsibilities under specified disturbances. Individual learning, task allocation, service requirements, and control arrangements can change at different rates. Consequently, stable competence can coexist with increasing vulnerability when backup coverage disappears or responsibilities expand. Conversely, large assistance benefits can coexist with sufficient independent coverage and exercised control. Those contrasting configurations cannot be distinguished from assisted-unaided gaps alone. The integration therefore introduces a requirement-referenced bridge between individual assessment and workflow exposure. The mechanism extends established perspectives through an explicit conditional specification; no claim of priority for distributed capability or organizational resilience is advanced. Table 2 separates inherited ideas from the proposed synthesis and identifies relationships that the proposed operational specification makes explicit.

Table 2: Theoretical Inheritance and Proposed Explanatory Extension

Perspective	Inherited Explanation	Sovereign Extension and Contrast
Automation reliance Bainbridge [6]; Parasuraman & Riley [7]	Reliance, retained skill, and difficult human interventions.	Tracks responsibility requirements and availability; vulnerability may grow without individual decline.
Distributed cognition Hutchins [17]	Performance emerges across people, representations, and artifacts.	Specifies task coverage under a focal-AI disturbance, rather than inferring collective capability from an individual mean.
Organizational resilience Duchek [18]	Knowledge, accessible resources, redundancy, social coordination, and power support anticipation, coping, and adaptation.	Operational refinement within a compatible account: maps calibrated trajectories and task assignments to dated responsibility thresholds; no claim that resilience theory excludes these cases.
Human-AI complementarity Bansal et al. [19]	Joint performance and appropriate reliance require examination beyond standalone accuracy.	Separates routine assistance advantage from capacity to meet minimum responsibilities during disruption.
Responsibility-capability alignment Proposed synthesis	Requirements, available coverage, structural exposures, and exercised control jointly constrain action.	An improved AI score alone leaves capability exposure unchanged; losing critical coverage can increase exposure with unchanged competence.

Note: Comparisons concern the selected explanations and their focal questions, not claims that entire research traditions cannot accommodate the proposed cases.

The gap-only, individual-atrophy, and control-average specifications are deliberately restricted analytical baselines, not faithful descriptions of named scholarly theories. Their limitations diagnose information loss within particular representations. They cannot establish superiority over distributed cognition or organizational resilience. The principal scholarly comparator is therefore Duchek's [18] capability-based account, which connects

anticipation, coping, and adaptation with knowledge, available resources, social relationships, and power and responsibility. That account already accommodates redundant staffing, access to expertise, and organizational learning. Sovereign claims a narrower contribution: an explicit, time-indexed measurement and allocation specification for AI-related responsibilities. The comparison concerns additional precision within a compatible explanation, rather than theoretical displacement.

Deriving the Integration from Organizational Resilience

Duchek's [18] account connects knowledge and resources with the capacity to prepare for, manage, and learn from disruption. Sovereign adopts that explanatory direction and specifies the unresolved operational mapping for a bounded AI-assisted workflow. A statement that relevant knowledge and resources support resilience does not alone identify which assessed capability covers which responsibility at a particular time. The proposed derivation begins with an externally justified task requirement and deadline. A measurement relationship translates evolving capability into expected unaided task performance, with uncertainty. A feasible assignment then identifies whether available personnel can meet the requirement during a specified disturbance. Exercised authority and recovery arrangements determine whether assessed capability can support intervention. The resulting prediction concerns which task loses coverage, through which pathway, and during which observation interval. A distinguishing analytical case holds total staff, mean capability, and assessed controls constant while reallocating specialist capability across tasks. Consider two sequential tasks, each requiring a score of 0.80, and two available employees. In configuration C, their task-score vectors are (0.90, 0.90) and (0.50, 0.50); in configuration D, both vectors are (0.90, 0.50). Both configurations have four-score means of 0.70. Configuration C covers both tasks because one employee can complete both sequentially within the specified deadlines. Configuration D leaves the second task uncovered despite the same mean capability and headcount. Duchek's account can explain the importance of accessible resources; Sovereign supplies a particular task-allocation rule that makes the contrasting coverage prediction explicit. A strong empirical comparator must therefore operationalize accessible knowledge, redundancy, coordination, and authority faithfully, rather than substitute a mean-score caricature.

Four named components distinguish four questions and their evidential requirements. CDM supplies a provisional temporal account of retained capability; DEM links responsibility requirements, feasible coverage, and technical exposure; CSM specifies the exercised controls relevant to each exposure pathway. APGI describes assistance advantage alongside those mechanisms and has no necessary causal role in the coverage pathway. The names identify modules within one research architecture, rather than four independent theories or four indispensable scalar scores. Removing CDM preserves a cross-sectional coverage analysis but removes the proposed trajectory account. Removing task allocation prevents a capability trajectory from locating organizational shortfalls, while removing CSM leaves the conversion of exposure into failure underspecified. Removing APGI leaves the central responsibility-capability prediction intact but removes the explicit comparison with assisted productivity. Empirical parsimony may therefore favor fewer modules for a particular research question; four-component completeness is not a condition for accepting the central mechanism.

Table 3: Source-to-Construct-to-Proposition Reasoning

Source / Construct	Analytical Choice	Proposition / Boundary
Salomon [4]; Risko & Gilbert [5] Supported versus retained performance	Separate paired performance contrasts from capability trajectories; APGI is a descriptive transformation.	P1-P2; a gap establishes neither loss nor independent predictive information.
Ericsson et al. [9]; Bransford & Schwartz [10] Practice and transfer	Represent learning exposure through P, V, E, X. Saturation and additive rates are author-proposed simplifications.	P2; $\alpha = 0$ is primary; compare functional forms and the conditional extension.
Hutchins [17]; Duchek [18] Distributed capability and continuity	Reference task scores to requirements and scenario availability. The maximum rule assumes substitutable single-person tasks.	P3; joint and concurrent tasks require team assessment or feasible allocation.
Pearce et al. [13]; Li et al. [15]; OWASP Foundation [14] Technical exposure	Retain separately assessed liabilities, shared dependencies, persuasive inputs, and permissions.	P4; source findings do not establish the proposed weights or cross-domain causal effects.
Green [12]; Bansal et al. [19]; NIST [11] Substantive oversight	Assess demonstrated controls and test exposure-specific outcome interactions.	P5; control effects are estimated rather than imposed uniformly.

PROPOSED MODULES AND INTEGRATION

Capability Dynamics Model: Conditional Capability Dynamics

CDM represents capability accumulation, stability, and decline. Let K at individual i and time t denote domain-specific retained capability on a calibrated interval from 0 to 1. The proposed dynamics separate capability-building exposure, L , from attrition pressure, F . Practice P , verification V , exception handling E , and transfer X are quality-adjusted exposures within the same observation interval. Assistance intensity A measures the extent of focal AI involvement; substitution orientation S measures displacement of relevant human cognitive work. Both range from 0 to 1 and must be measured separately.

$$L_{it} = \beta_1 P_{it} + \beta_2 V_{it} + \beta_3 E_{it} + \beta_4 X_{it} \quad (1)$$

$$F_{it} = \lambda \text{ (primary)} \quad (2)$$

$$F_{it} = \lambda + \alpha I_{it} \text{ (optional interference extension)} \quad (2a)$$

$$\frac{dK_{it}}{dt} = (1 - K_{it}) L_{it} - K_{it} F_{it} \quad (3)$$

All exposure variables, including I when measured, range from 0 to 1; β_1 through β_4 , λ , and α are non-negative rate parameters expressed in the chosen time unit. The factor $1 - K$ represents declining headroom for growth, while multiplication of F with K prevents erosion of capability below zero. At $K = 0$, the derivative equals L and cannot be negative; at $K = 1$, the derivative equals $-F$ and cannot be positive. Under non-negative, bounded rates, the continuous-time model therefore preserves the capability interval. Appendix A states further mathematical implications and an illustration.

The β parameters represent provisional contributions of capability-building exposures. Practice P denotes deliberate work without the focal AI; verification, exception handling, and transfer can occur with AI present when humans perform the relevant reasoning. The primary CDM represents capability dynamics conditional on those learning

exposures. Assistance intensity A and substitution orientation S influence the proposed trajectory through changes in P , V , E , and X ; their effects on those exposures remain separately hypothesized and require measurement. An engagement-preserving condition requires observable human explanation, verification, or reconstruction of solutions, whereas an engagement-displacing condition permits acceptance of outputs without those activities. The baseline rate λ permits ordinary attrition even without AI . The primary specification fixes $\alpha = 0$ because reduced learning exposure already provides a pathway from substitution to less favorable capability trajectories. The optional αI extension in Equation 2a represents a separate, testable interference process: repeated adoption of misleading procedures could weaken previously correct task representations despite comparable measured practice. Merely completing fewer independent exercises belongs in L and does not justify a second loss term. A positive α requires independently observed interference opportunities and outcome patterns that distinguish interference from omitted practice quality, fatigue, and task change. I denotes an independently coded, normalized rate of adopting demonstrably misleading procedures within a prespecified task-exposure window. Coding requires a correctness criterion and evidence of adoption, not merely encountering an incorrect output; the denominator and exposure window remain fixed across comparisons. Assistance intensity A and substitution S alone do not establish I . Without a defensible measure of I , Equation 2a is not estimated and the primary specification is retained.

Substitution must be operationalized through behavior, such as accepting solutions without reconstructing required reasoning, rather than inferred from low capability. Otherwise, the model becomes circular. Overlap among P , V , E , and X also requires careful coding; one activity may contribute to several learning processes without representing independent evidence for each coefficient. Controlled designs and repeated measurements are needed to separate mechanisms. Positive growth under measured interference exposure does not reject $\alpha > 0$, because capability-building gains may exceed attrition. The relevant comparison estimates the additional trajectory difference associated with I at comparable initial capability and learning exposure, against the nested $\alpha = 0$ specification. CDM should also be compared with simpler growth models and domain-specific learning curves.

K is a latent capability construct rather than an observed score. For a provisional bridge, let $K_{ir}(t)$ represent capability dimension r and let μ_{ikt} denote expected unaided performance on task k at assessment time t . The dimensions may include execution, transfer, diagnosis, and verification, provided measurement testing supports the proposed structure. Nonnegative task discrimination parameters ℓ_{kr} connect the dimensions to task performance, and b_k represents task difficulty; zero loadings exclude irrelevant dimensions. An illustrative bounded link is:

$$\text{logit}(\mu_{ikt}) = \sum_r \ell_{kr} \text{logit}(K_{ir}(t)) - b_k \quad (3a)$$

$$Q_{ikt}^U = \mu_{ikt} + \varepsilon_{ikt}, \quad E(\varepsilon_{ikt} | K, \text{task}) = 0 \quad (3b)$$

Here $\text{logit}(K)$ is defined for interior values $0 < K < 1$, with boundary behavior treated through limits or an explicitly specified estimation constraint. The inverse-logit link maps the linear predictor to an expected score between zero and one. For an assessment comprising conditionally independent binary-scored opportunities with a common success probability, a provisional observation model treats nQ^U as binomial with n trials and success probability μ . The resulting residual ε has conditional mean zero and variance $\mu(1 - \mu)/n$;

dependent items, rubric scores, or rater effects require a different observation model. The scale and origin of latent capability require anchor tasks and identification constraints, while difficulty and loading parameters require calibration before longitudinal comparison. Equal observed scores on different tasks need not represent equal capability. Repeated parallel forms permit estimation of latent change separately from task difficulty and measurement error. If a common capability dimension is unsupported, dimension-specific CDM trajectories replace a single K .

The measurement bridge connects assessment at time t_0 with coverage at a later responsibility date t_1 . CDM projects each relevant capability dimension across the assessment interval, and the calibrated task link produces expected task scores at t_1 . Equations 6b–6c use those estimated expected scores or joint uncertainty draws, and Equations 7–8 convert the selected representation into task and workflow shortfalls; Equation 6a remains the contemporaneous observed-score description. The raw-score maximum remains a transparent descriptive approximation when assessment and responsibility dates coincide and calibration is adequate. Selecting the highest noisy score can overstate coverage, so estimation should propagate calibration, trajectory, assessment, and selection uncertainty jointly. Studies should report an interval for expected-score coverage or, under a specified uncertainty model, the probability that expected-score coverage falls below the requirement. Such a probability concerns uncertainty about expected capability, not the chance of unsuccessful performance on a particular future attempt. Predicting realized task performance additionally requires the observation model to represent variation across future attempts; predicting operational failure requires the separately evaluated outcome relationship in Section 4.5. Neither an uncertain crossing nor a change in task difficulty establishes actual capability loss. The bridge yields testable responsibility-date forecasts; causal attribution to AI still requires an appropriate comparison design.

Assisted Performance Gap Index: An Assisted-Unaided Contrast

APGI is a descriptive index of the positive proportional difference between assisted and unaided performance. Let Q^A and Q^U denote comparable scores for individual i , dimension j , and time t . The four proposed dimensions are execution, transfer, diagnosis, and verification. Higher scores indicate better performance. Each score must use a prespecified, non-negative scale with a meaningful zero; rescaling an arbitrary ordinal rating to the unit interval does not establish that proportional differences are meaningful.

$$APGI_{it} = 100 \sum_{j=1}^4 w_j \max\left(0, \frac{Q_{ijt}^A - Q_{ijt}^U}{\max(Q_{ijt}^A, \kappa)}\right) \quad (4)$$

Weights w_j are non-negative and sum to 1. The denominator floor κ is fixed in advance, expressed on the same score scale, and greater than zero. With both scores in $[0, 1]$, every component lies in $[0, 1]$, giving APGI a range of 0 to 100. The result is an index in points, not a percentage of capability lost. The floor limits instability near zero but introduces a consequential analytic choice. Studies should report sensitivity to κ and to weights, together with every component score. When a defensible ratio interpretation is unavailable, standardized or raw paired differences should replace the proportional specification.

A positive APGI can arise from beneficial augmentation, stronger AI performance, weak independent competence, or several processes together. A zero APGI can describe equally excellent performance, equally poor performance, or assistance that makes performance worse. The positive-part transformation intentionally excludes the last possibility from the index. Consequently, reports must accompany APGI with absolute assisted and unaided scores, signed differences, uncertainty intervals, and longitudinal unaided trajectories. An increasing APGI does not demonstrate atrophy. For example, if unaided performance rises from 0.60 to 0.70 while assisted performance rises from 0.70 to 0.95, a single-dimension gap increases from 14.3 to 26.3 points despite independent learning.

Paired assessment requires matched task forms, balanced order, comparable incentives, and control of fatigue and practice effects. Assessment tasks should measure substantive verification and reasoning rather than reproducing identical answers remembered from the assisted condition. Baseline competence and model version must be recorded. APGI remains sensitive to the capability of the focal AI as well as the human participant. Comparisons across populations or time therefore require evidence of task and measurement comparability. APGI is a deterministic transformation of its underlying scores. Consequently, APGI supplies no additional information beyond the complete assisted and unaided inputs, scoring rules, and weights. Any predictive advantage over a restricted baseline would concern representation, aggregation, regularization, or finite-sample estimation, rather than a new information source. Comparisons should include flexible models using the complete inputs under matched training and evaluation conditions. APGI serves primarily as a compact description of assistance advantage; responsibility-relevant shortfalls are assessed separately.

Cognitive Sovereignty Model: Exercised Control Capacity

Cognitive sovereignty denotes retained human and institutional capacity to evaluate and govern an AI-mediated workflow. The responsible team is the unit of assessment. Five proposed components are independent reasoning (IR), verification capability (VC), override capability (OC), fallback capability (FC), and auditability capacity (AC). Each component is scored on [0, 1] through prespecified task or scenario criteria. Documentation, nominal permissions, and self-reported confidence alone do not demonstrate successful control.

$$CSM_{ot} = \delta_1 IR_{ot} + \delta_2 VC_{ot} + \delta_3 OC_{ot} + \delta_4 FC_{ot} + \delta_5 AC_{ot} \text{ (optional scalar)} \quad (5)$$

The non-negative δ weights sum to 1, so CSM lies in [0, 1]. Independent reasoning concerns whether the designated team can explain and assess consequential decisions without relying on the focal AI. Verification concerns detecting and resolving seeded or naturally occurring errors. Override requires both practical authority and successful intervention. Fallback concerns maintaining a defined minimum service or safe shutdown under disruption. Auditability concerns reconstructing relevant decisions from available records. Effective fallback may rely on another qualified team, established non-AI systems, or genuinely independent infrastructure; every worker need not reproduce the full workflow alone.

The weighted average assumes compensability: strength in one component can offset weakness in another. That assumption may be unsuitable for critical workflows. A high

average must therefore be reported alongside the component profile and domain-specific minimum requirements. If intervention authority is indispensable, an override component below a justified threshold prevents a claim of adequate control regardless of the average. Thresholds require external justification and subsequent validation; no universal cutoff is proposed. Alternative bottleneck or multiplicative specifications should be compared empirically.

CSM is specified independently of DEM, with the component requirements above determining adequacy for each scenario. Individual verification capability may contribute to team control, but team authority, coordination, and infrastructure also matter. Neither APGI nor the control average enters DEM as an automatic risk multiplier or offset. Exposure-specific outcome models provide the proposed tests of control effects.

Dependency Exposure Model and Responsibility Requirements

DEM represents specified workflow exposures without converting assistance advantage into a capability deficit. Let o identify the workflow, k a required task, s a disruption scenario, and t the observation period. The requirement R ranges within $(0, 1]$ on a calibrated task-performance scale, with a documented minimum acceptable standard and completion deadline. Requirements should derive from task analysis, service obligations, or externally justified standards before observed scores are inspected. A quality score is insufficient when the requirement also involves throughput, concurrent tasks, or response time; those dimensions require separate constraints. The focal AI is unavailable or unreliable in the scenario, while permitted reference materials and alternative tools remain explicitly identified. Requirement selection is a normative and operational decision whose sensitivity must be reported. Changes in required service levels must remain distinguishable from changes in human capability.

For a task that one eligible person can perform independently, Equation 6a defines contemporaneous descriptive coverage B^{obs} from observed task-specific unaided scores. Equation 6b defines point forecast coverage B^{pred} at responsibility date t_1 using expected task scores estimated from information available at assessment date t_0 . The binary indicator a equals 1 when the person is eligible for the task and available within the scenario deadline, and 0 otherwise. Eligibility concerns permissible task assignment rather than demonstrated sufficiency; successful override authority is assessed separately within CSM. Scores Q^U refer to the required task, not an undifferentiated average across unrelated activities. The maximum over an empty eligible set is defined as zero. In Equation 6b, μ is the estimated conditional expected task score and $a(t_1)$ specifies scenario availability at the responsibility date.

Equation 6c propagates uncertainty through joint draws m of expected scores and, where uncertain, availability; known availability remains fixed across draws. The point forecast in Equation 6b is a maximum of expected scores, not the expected maximum across draws. Equation 6c describes uncertainty about the highest expected capability within the available eligible set; the maximizing person may differ across draws. That available-capability envelope does not establish coverage achievable through an assignment selected before execution.

For assignment-specific forecasts, studies should select an eligible person using information available at t_0 and evaluate that same assignment across draws, treating unavailability as zero coverage unless a prespecified reassignment rule applies. An adaptive assignment procedure must specify which additional observations become available before selection or reassignment; the unknown capability values in a draw cannot themselves supply that information. Equation 7 applies to the declared coverage representation, with R evaluated at the matching task, scenario, and responsibility date; draw-specific shortfalls are calculated before summarizing uncertainty. Reports must distinguish available-capability envelopes from assignment-specific forecasts and identify the information and selection rule supporting the latter.

$$B_{kost}^{obs} = \max_i(a_{ikost} Q_{ikt}^U) \quad (6a)$$

$$B_{kos}^{pred}(t_1 | t_0) = \max_i[a_{ikos}(t_1) \hat{\mu}_{ik}(t_1 | t_0)] \quad (6b)$$

$$B_{kos}^{(m)}(t_1) = \max_i[a_{ikos}^{(m)}(t_1) \mu_{ik}^{(m)}(t_1)] \quad (6c)$$

$$d_{kost} = \max(0, \frac{R_{kost} - B_{kost}}{R_{kost}}) \quad (7)$$

The maximum operator assumes substitutability for a single task and cannot represent joint work or simultaneous demand. A team jointly performing a task should instead receive a directly observed team score under an explicit coordination protocol. Multiple concurrent tasks require a feasible personnel-to-task allocation that respects availability and workload; the same specialist cannot cover several simultaneous responsibilities through repeated maxima. Where feasible allocation is unresolved, separate task profiles are required and the scalar workflow summary should be withheld. Scenario analysis should include loss of an indispensable specialist, common infrastructure failure, and correlated unavailability where relevant. Those restrictions place staffing configuration inside the proposed mechanism rather than relegating coverage to an unmodeled caveat. An employee can retain identical competence while B declines because a changes. Conversely, a larger assisted score does not change d when the requirement and unaided coverage remain fixed.

A descriptive workflow shortfall D is the weighted mean of task shortfalls, with non-negative criticality weights ω summing to 1. The maximum shortfall must accompany the mean. A critical task with a binding minimum requirement remains a failed requirement even when the average shortfall is small. The summary therefore supports comparison but cannot certify acceptability.

$$D_{ost} = \sum_k \omega_{kost} d_{kost} \quad (8)$$

DEM primarily represents a five-component exposure profile $x = (D, SM, TD, PE, PP)$, with each component in $[0, 1]$ under a prespecified domain calibration. D describes human coverage shortfalls; SM describes shared failure propagation through common models, infrastructure, or data; TD describes independently assessed unresolved technical liabilities; PE describes consequential inputs exposed to misleading AI outputs; PP describes permissions whose scope or duration exceeds task requirements. The organizing distinction concerns the location of an initiating vulnerability: human response capacity, common dependency structure, latent technical defects, decision content, or action authority. The five channels provide an initial taxonomy rather than an exhaustive or mutually independent

decomposition. Security events can activate several channels, and domain analysis may require additional exposures. The primary analysis retains the component profile and an explicit pathway map. The scalar DEM in Equation 9 is optional and describes a fixed exposure mix rather than a probability or expected loss.

Dependency intensity DI is excluded from the shortfall calculation. An unmet responsibility remains an unmet responsibility even when routine reliance on AI is low. Multiplying shortfall with routine reliance would obscure that condition and lacks justification for the stated task-sufficiency estimand. Where research examines disruption burden across a portfolio, affected workload share may be measured separately as the proportion of service units touched under a defined scenario and horizon. Multiplying that share with a shortfall is defensible only for an explicitly additive workload-burden estimand with compatible denominators, not for task sufficiency. Disturbance occurrence, affected workload, loss severity, and conditional task shortfall answer different questions. A rare disturbance or small affected workload cannot certify that a binding responsibility remains covered. The worked examples therefore use D directly.

$$DEM_{ost} = 100 \sum_{l=1}^5 \gamma_l x_{lost} \quad (\text{optional scalar}) \quad (9)$$

For the optional scalar, the non-negative γ weights sum to 1, giving DEM a range of 0-100 descriptive exposure points. APGI is absent from the expression because assistance advantage alone does not establish unmet responsibility. An AI improvement that raises Q^A , with requirements, coverage, and other exposures held fixed, leaves DEM unchanged. If the organization expands promised service or reduces backup staffing following that improvement, DEM can increase through changed requirements or availability. The proposed mechanism therefore identifies the organizational change that creates exposure rather than assigning exposure directly to better AI performance. Severity, threat prevalence, and broader consequences require separate modeling before any estimate of actual risk. Structural exposures may remain substantial even when the capability-shortfall channel equals zero. Overlapping pathways require explicit coding: for example, an unpatched shared component can register as both TD and SM, but the same resulting outage is counted once as an outcome. A pathway diagram should record the shared cause and any interaction rather than assuming five independent contributions. When observations cannot distinguish two channels, researchers should combine or omit redundant indicators before considering a scalar. A linear average cannot identify joint failure propagation and must be compared with component-only and interaction specifications.

Exposure-Specific Control and Empirical Outcomes

Control effectiveness is a separate empirical relationship rather than a uniform deduction from every exposure. Verification may affect whether misleading content escapes review, whereas independent fallback arrangements may affect continuity during common infrastructure failure. Access limitation or successful revocation may constrain privilege-related consequences without improving verification performance. Auditability may accelerate reconstruction after failure without preventing the initiating event. Independent reasoning supports some interventions but cannot replace unavailable permissions or infrastructure. CSM remains a descriptive profile of five control capacities, not a complete inventory of every technical safeguard. Studies should add directly measured controls when

an exposure pathway requires them. A common interaction coefficient provides a parsimonious benchmark for prespecified pathways on the log-odds scale. The component-specific alternative permits different interactions, including no interaction, as specified below.

For a prespecified exposure-control pair, a candidate test can model a subsequently observed binary failure Y using a logistic link. Let x_l denote the selected exposure, c_h the selected control component, and z a vector of prespecified contextual covariates. Suppressing workflow, scenario, and time indices for readability gives:

$$\text{logit } Pr(Y = 1) = b_0 + b_1 x_l + b_2 c_h + b_3 x_l c_h + z^T \xi \quad (10)$$

The interaction coefficient b_3 expresses moderation on the log-odds scale; probability-scale contrasts should also be reported because effect modification depends on scale. Neither b_3 nor the control main effect is constrained to be beneficial. With an adverse exposure association, a negative b_3 would support attenuation on the specified scale under the design's identifying assumptions; a precise null or positive value would challenge the hypothesized relationship. Multilevel extensions should account for repeated observations and organizational clustering. Candidate pairs must be selected before evaluation to limit opportunistic interaction searches. For comparisons across J prespecified pathways, a pooled extension applies Equation 10 to pathway-indexed binary outcomes Y_p , exposures x_p , and controls c_p , with $p = 1, \dots, J$. Each pathway has coefficients b_{0p} , b_{1p} , b_{2p} , b_{3p} and contextual coefficients ξ_p . The restricted model imposes $b_{3p} = b_3$ for every pathway p ; the heterogeneous model estimates each b_{3p} separately. Intercepts, main effects, and contextual coefficients remain pathway-specific in both models. Dependence among outcomes from the same workflow or organization requires joint modeling or cluster-appropriate inference. Comparisons require prespecified, substantively comparable exposure and control increments; equality concerns log-odds interactions, not equal probability-scale effects. Probability contrasts should therefore use declared increments and a common reference covariate distribution where supported. The equation is a proposed empirical test, not an estimated component of DEM.

Coverage and control measures can overlap unless their operational roles are specified carefully. Availability in B concerns eligible personnel within the task deadline; override concerns actual authority and successful execution of an intervention. If a fallback arrangement already increases assessed B, that same improvement should not receive a second automatic exposure reduction. A study can represent the arrangement through coverage, model a separately measured downstream control mechanism, or explicitly estimate a mediated pathway. The manuscript therefore reports DEM and CSM in parallel rather than combining their scores into a purported calibrated risk estimate. Figure 1 locates the control arrow on the exposure-outcome relationship to distinguish moderation from a direct effect. The direct effect permitted in Equation 10 remains separately identified in the figure. Feedback from failures to future task allocation is plausible but remains outside the current dynamic specification.

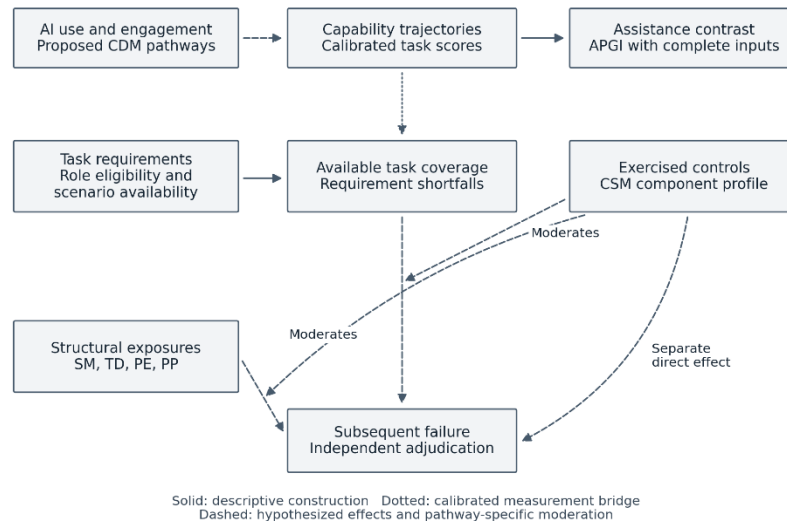


Figure 1: Proposed Capability, Coverage, and Control Relationships

Note: APGI also requires assisted assessment scores. Structural exposures originate separately from task coverage; each moderation arrow denotes a selected pathway-specific interaction, not uniform attenuation of every component. Controls moderate selected exposure-outcome pathways; the separate direct-effect path corresponds to the control main effect in Equation 10. The diagram does not imply a closed causal system.

Worked Example of Capability, Coverage, and Control

A hypothetical software-release workflow illustrates the proposed distinctions without supplying validation data. Two sequential critical tasks are required during a focal-AI outage: independent release review and recovery to a stable version. Their minimum task scores are 0.70 and 0.80, respectively, with fixed deadlines. Either task can be completed independently, so the single-task maximum is applicable. Task weights are equal, and the structural exposure profile is $SM = 0.40$, $TD = 0.20$, $PE = 0.30$, and $PP = 0.20$. Equal DEM weights yield 22 points when D equals zero. Assessment dimensions are separately calibrated; the reported APGI values summarize equivalent assisted-unaided contrasts across the four dimensions, while B uses task-specific scores. For all APGI illustrations in the article, $\kappa = 0.01$ on the 0-1 score scale; every illustrated assisted score exceeds that floor, so the denominators remain unchanged. All quantities are hypothetical and all comparisons concern the stated requirements and outage scenario.

Configuration A has primary and backup personnel scoring 0.80 on both required tasks, with availability maintained when the primary specialist is absent. Their assisted assessment score of 0.98 and unaided score of 0.80 across each gap dimension yield $APGI = 18.4$ points. Both task requirements remain covered, so D equals zero and DEM remains 22. The control profile (IR, VC, OC, FC, AC) is (0.80, 0.90, 0.90, 0.90, 0.80), giving an equal-weight CSM of 0.86. Successful seeded-error detection, exercised rollback permissions, and a demonstrated recovery procedure establish adequacy for the specified test criteria. The profile does not establish safety against every possible disturbance. If assisted performance rises to 1.00, $APGI$ increases to 20.0 while D and DEM remain unchanged. Thus, a greater assistance advantage need not imply a weaker capability foundation or greater exposure.

Configuration B has primary-task scores of 0.80 and assisted assessment scores of 0.82, yielding $APGI = 2.4$ points for the primary personnel. Review remains covered during

the outage, but the sole qualified recovery specialist is unavailable; the best available recovery performer scores 0.40. Recovery shortfall is $(0.80 - 0.40) / 0.80 = 0.50$, so D equals 0.25. The optional DEM increases to 27 points under the same structural exposures and weights. The assessment classifies the critical recovery requirement as unmet regardless of whether the five-point score difference appears small. The control profile is (0.80, 0.90, 0.00, 0.20, 0.80), giving $CSM = 0.54$; strong verification cannot compensate for absent rollback authority. Even if the specialist's absence were resolved, that authorization defect would remain until an effective intervention route was established. The low assistance gap consequently offers no assurance of continuity or control. A within-workflow comparison isolates coverage change from individual atrophy. Before the specialist's absence, Configuration B has $B = 0.80$ for recovery, $D = 0$, and $DEM = 22$. During the absence, every individual's assessed competence can remain unchanged while D rises to 0.25. Restoring an available, independently capable backup returns D to zero without changing the original staff's assessment scores. The causal hypothesis concerns whether that coverage restoration improves subsequent continuity outcomes, not whether the score decreases arithmetically. A team mean of personnel scores cannot represent the change without availability information. For a separate control contrast, profiles (0.90, 0.90, 0.40, 0.90, 0.90) and (0.80, 0.80, 0.80, 0.80, 0.80) both average 0.80. Under a hypothetical override minimum of 0.70, only the second profile meets that requirement, so a uniform reduction based on the average conceals a consequential difference. Table 4 summarizes the contrast and identifies where task profiles take precedence over aggregate scores.

Table 4: Hypothetical Coverage Configurations Under a Focal-AI Outage

Configuration	APGI	Review / Recovery B	D / Max D	DEM	Interpretation
A: Qualified backup available	18.4	0.80 / 0.80	0 / 0	22	Both requirements covered; exercised control meets the specified drill criteria.
A: Assisted score improves to 1.00	20.0	0.80 / 0.80	0 / 0	22	Gap increases without a changed requirement or capability shortfall.
B: Primary specialist available	2.4	0.80 / 0.80	0 / 0	22	Capability coverage present; absent override authority remains a separate control defect.
B: Recovery specialist unavailable	2.4	0.80 / 0.40	0.25 / 0.50	27	Critical recovery requirement fails; individual competence need not decline.
B: Qualified backup restored	2.4	0.80 / 0.80	0 / 0	22	Coverage restored; authorization still requires an effective intervention route.

Longitudinal Threshold Crossing with Unchanged Staffing

A complementary hypothetical example holds staffing, task requirements, and controls constant while capability changes. One eligible specialist covers a single sequential recovery task, with no available substitute and a requirement $R = 0.75$. For transparent arithmetic, the measurement link has one capability dimension, loading $\ell = 1$, and difficulty $b = 0$, so expected task performance μ equals K . Those calibration values are assumptions for the

example, not estimated properties. Let initial capability equal 0.90 and let constant monthly rates be $L = 0.02$ and $F = 0.06$. Set $\alpha = 0$ and $\lambda = 0.06$, demonstrating that the proposed threshold pathway does not require additional AI-specific attrition. The equilibrium equals 0.25, and Equation 12 gives $K(t) = 0.25 + 0.65 \exp(-0.08t)$. Expected capability crosses 0.75 at approximately 3.28 months; expected scores at months 0, 3, and 4 are 0.900, 0.761, and 0.722. The corresponding point-estimate shortfalls are 0, 0, and approximately 0.037, while availability and responsibility remain fixed.

Assessment uncertainty changes the strength of that conclusion. Suppose solely for illustration, the calibrated expected-score interval at month 4 spans 0.70-0.78. The associated expected-score shortfall interval spans 0-0.067, so a definite crossing cannot be inferred from the point estimate alone. The interval describes uncertainty in expected capability and excludes the additional variation of realized scores across future attempts. A lower-confidence-bound rule would represent a prespecified conservative decision policy, not a mathematical implication of CDM. Further independent assessment may distinguish a genuine crossing from score uncertainty. In the earlier availability example, expected capability remains constant while the eligible coverage set changes. The two cases therefore identify different intervention targets: maintaining capability in the first case and restoring qualified coverage in the second. A single average productivity or assistance-gap score cannot locate those targets. Independent subsequent recovery outcomes remain necessary to evaluate whether either coverage prediction explains actual failure.

Analytical Purpose of Profiles and Optional Summaries

Component profiles are the primary representations for CSM and DEM, and signed paired scores are primary for assistance effects. APGI optionally summarizes the positive proportional assistance advantage across a fixed set of commensurable dimensions; the summary can support comparisons of that particular benefit pattern but cannot represent net benefit when harmful effects occur.

Negative assistance effects must remain visible in signed differences and adverse-effect frequencies. CSM optionally summarizes change in a fixed control battery within the same workflow, provided binding control requirements are reported independently. DEM optionally summarizes a prespecified exposure mix for descriptive sensitivity analysis within a stable measurement design. Neither average determines whether a critical responsibility is covered, ranks unlike institutions, or supplies an intervention priority. A scalar is retained only when an explicit reporting or analytical purpose is served without obscuring the component conclusions.

Aggregation therefore introduces a testable restriction rather than additional information. A prospective study can compare a fixed-weight summary with a component-only model using identical inputs, outcome definitions, and evaluation samples. Any advantage must concern parsimony, stability, interpretability assessed with intended users, or held-out prediction under comparable tuning budgets. If the summary loses consequential distinctions without a compensating benefit, the scalar should be abandoned. The central responsibility-capability explanation survives that decision because task coverage and control pathways remain explicit.

PROPOSITIONS AND CONDITIONS FOR REJECTION

The propositions separate observed mechanisms from arithmetic properties of the proposed measures. Table 5 specifies candidate tests and potentially disconfirming findings. The effects concern later assessments or independently adjudicated operational outcomes rather than changes mechanically implied through a formula. Each test requires a defined population, task class, scenario, and smallest effect of substantive interest. Uncertainty intervals are necessary to distinguish evidence against a prediction from inadequate precision. Designs should also preserve the distinction between a mechanism, a descriptive indicator, and an intervention. The following propositions focus on the framework's consequential relationships.

P1: Compression is a conditional possibility rather than a consequence of larger average gains among weaker performers. Let U denote baseline unaided performance, G the signed assistance gain, and assisted performance equal $U + G$ on an equated scale. The variance identity gives:

$$\text{Var}(U + G) - \text{Var}(U) = 2 \text{Cov}(U, G) + \text{Var}(G) \quad (\text{P1})$$

Dispersion falls precisely when $2 \text{Cov}(U, G) + \text{Var}(G) < 0$. Under the illustrative gain model $G = a - cU + \eta$, with η uncorrelated with U , compression requires $0 < c < 2$ and $\text{Var}(\eta) < (2c - c^2)\text{Var}(U)$. Greater gains at lower baselines imply a negative slope but do not alone establish that inequality. The empirical hypothesis is that a prespecified assistance design produces sufficiently predictable compensatory gains to satisfy the variance condition on non-ceiling tasks. The condition must be forecast from theory or separate pilot data before confirmatory evaluation; checking the identity on the same outcomes is descriptive arithmetic. A precisely nonnegative variance contrast on held-out tasks challenges the design-specific compression hypothesis, while leaving the identity intact.

P2: Assistance that displaces capability-building activity is expected to produce less favorable unaided trajectories than comparable assistance that preserves active engagement. Reduced growth and absolute decline are distinct outcomes. The primary CDM fixes $\alpha = 0$. An optional interference extension compares $\alpha > 0$ with $\alpha = 0$ only where interference opportunities are measured independently, conditional on initial capability and measured learning exposures under a credible design. Positive growth alone cannot reject the additional loss mechanism. A precise absence of the predicted trajectory difference, or superior performance from a competing specification, would weaken the proposed term.

P3: During specified AI disruptions, unmet critical task requirements are expected to predict continuity or intervention failure after baseline competence, task difficulty, disturbance severity, and resource constraints are considered. Increased shortfall can follow reduced availability or expanded responsibility while individual competence remains stable. Restricted gap-only and individual-decline baselines omit that pathway; they do not represent the full scholarly alternatives. The requirement-referenced representation should be compared with flexible models using the same underlying scores, requirements, availability, and task-allocation information. Any advantage concerns useful structure or parsimony rather than information created through aggregation. Consistently successful critical-task performance despite substantial measured shortfalls would count against the preregistered requirement-to-outcome prediction; subsequent threshold changes require a new test.

P4: Conditional on prespecified disturbance type, disturbance intensity, assessed capability coverage, and measured controls, greater shared dependency exposure is expected to predict more correlated workflow failures; independently measured technical liabilities are expected to predict greater recovery burden. Before outcome observation, the protocol must identify the affected workflows, credible initiating disturbance, qualifying exposure levels, intervention window, and smallest meaningful outcome contrast. Controls expected to interrupt the selected pathway must be recorded and incorporated into the comparison in advance. A precise absence of the predicted contrast, including successful containment in the preregistered stratum lacking adequate pathway-specific controls, counts against the corresponding proposition. Unmeasured controls cannot be introduced retrospectively to preserve the original prediction. Measurement problems discovered afterward warrant a separately labeled revised test and do not erase the original unfavorable result. Imprecise estimates remain inconclusive rather than supportive. The proposition is pathway-specific and does not claim that every exposed organization fails.

P5: The association between exposure and subsequent failure is expected to vary with the specific control capable of interrupting that pathway. Verification is expected to attenuate misleading-output acceptance more directly than outage duration; exercised recovery arrangements are expected to affect outage consequences more directly than content accuracy. Different control profiles can therefore produce different outcomes despite identical CSM averages. Adequately estimated common effects across pathways would favor the simpler comparator; precise null or adverse interaction effects would challenge the specified mitigation hypothesis. Control improvements should be tested against external outcomes rather than a score that already deducts controls.

Table 5: Propositions and Discriminating Empirical Tests

Proposition	Primary Comparison	Potentially Disconfirming Result
P1: Conditional compression	Preregistered gain variability and baseline-gain covariance; held-out equated tasks.	Nonnegative held-out variance contrast under a preregistered compensatory-gain hypothesis; average gain differences alone are insufficient.
P2: Learning displacement	Primary displacement test ($\alpha = 0$): compare engagement conditions, allowing P, V, E, and X to vary. Separate interference test: estimate α_I at comparable initial capability and learning exposures, with I measured independently.	Displacement: precise absence of the predicted engagement-condition trajectory contrast. Interference: precise absence of additional attrition conditional on learning exposures. Positive growth alone does not reject interference.
P3: Unmet responsibility	Requirement and availability changes with stable competence; matched-information models.	Consistent success despite substantial preregistered shortfalls counts against the specified relationship; revisions require a new test.
P4: Structural exposure	Shared dependencies and liabilities before correlated failures or recovery burdens.	Precise absence of the registered pathway contrast under prespecified disturbances and control conditions; post hoc controls do not rescue the prediction.

P5: Specific control effects	Pooled pathway model: equal log-odds interaction coefficients b_{3p} versus prespecified pathway-specific coefficients; report probability contrasts separately.	Common effects fit adequately, or hypothesized pair-specific attenuation is precisely absent or adverse.
------------------------------	--	--

PROPOSED EMPIRICAL RESEARCH PROGRAM

The primary objective of the first proposed study is to test whether preregistered responsibility shortfalls predict independently assessed subsequent task failure in one bounded workflow. After measurement calibration, a prospective simulation could vary qualified backup availability while holding task requirements and assistance conditions constant. A longitudinal companion would assess whether capability trajectories predict later threshold crossings under unchanged staffing. Independent nonidentical recovery tasks would supply the outcomes, with adjudicators unaware of predicted coverage. A secondary objective is to compare predictive performance of the requirement-referenced specification with a flexible component model and a faithful resilience-informed model containing accessible expertise, redundancy, coordination, and authority. Preregistered outcomes, meaningful effect sizes, equivalence bounds, and uncertainty rules would define evidence against the predicted shortfall-failure relationship and, separately, the criteria for comparative predictive improvement. The study need not validate APGI, every structural channel, or every optional scalar before testing the central mechanism. The following stages describe the supporting measurement and extension program.

Construct and Measurement Development

The first stage should establish content validity and workable measurements within one clearly defined domain. Domain specialists, frontline practitioners, and assessment experts can evaluate whether the proposed dimensions cover consequential capabilities and distinguish learning activity from task completion. Cognitive interviews and pilot tasks can identify ambiguous items or inaccessible assessment conditions. Independent raters should apply explicit rubrics to explanations, verification actions, interventions, and recovery exercises. Reliability must be established for the component measures before composite scores receive substantive interpretation. Capability measures require evidence of temporal and group comparability, including task equating and checks for differential item functioning where appropriate. A common latent K should be retained only if the measurement structure supports that abstraction. CSM and DEM primarily organize component profiles; any optional averages are formative summaries rather than reflective latent traits. High internal consistency is therefore neither required nor sufficient for their validity. Studies should examine component coverage, redundancy, weighting sensitivity, missing-data handling, and associations with independent criteria. Failure to establish measurement quality should lead to revision before tests of the broader propositions.

Experimental and Longitudinal Tests

The second stage should compare assistance conditions designed to vary productive engagement. A possible design includes a conventional-workflow comparison, unrestricted

assistance, and assistance combined with explanation, verification, or graduated practice requirements. Assignment should be randomized where feasible; otherwise, a credible comparison strategy and explicit assumptions are needed. Baseline assessments, repeated task exposure, immediate assisted performance, delayed unaided performance, and novel transfer tasks would distinguish productivity from learning. The primary displacement test estimates the engagement-condition contrast without matching or adjusting away post-assignment P, V, E, and X, because those learning exposures carry the hypothesized pathway. The separate interference test estimates additional attrition associated with I at comparable initial capability and learning exposures; causal interpretation requires adequate control of confounding. The proposed design is a future protocol, not a study conducted for the present article.

Usage logs and process measures should assess A, S, P, V, E, and X independently of capability outcomes; testing Equation 2a additionally requires independent coding of I. Identical prompts or task sequences can create carryover, so assessment order and parallel forms require control. Longitudinal models should distinguish within-person change from differences between people and account for task difficulty, dropout, prior expertise, and model updates. Power analysis should target a prespecified practically meaningful effect using pilot variability. CDM should be compared with simpler growth or decay specifications, with $\alpha = 0$ as the primary specification and an independently motivated interference extension only where measurable. Mediation claims require temporal ordering and defensible assumptions concerning mediator-outcome confounding.

Workflow-Level Predictive and Intervention Studies

The third stage should examine workflows prospectively, retaining the nesting of observations within people, teams, and organizations. Task requirements, role availability, gap profiles, dependency exposures, and exercised control should be measured before independent outcome observation. Candidate outcomes include consequential errors escaping review, unsuccessful overrides, recovery time under controlled outages, and failures to maintain a specified minimum service. Coverage scores should derive from prior calibrated assessments, while outcomes concern subsequent nonidentical disturbances; reusing the same performance observation for both would make the test circular. Outcome adjudicators should be unaware of composite scores where feasible. Simulated disturbances may be appropriate when real failures are rare or unsafe to induce, but performance in drills should not be equated automatically with field reliability.

Testing should compare preregistered component models, simple baselines, and proposed representations on held-out observations, using matched information and comparable tuning budgets. Baselines should include a gap-only account, an individual-capability account, a flexible model of the complete assessment inputs, and a role-coverage model incorporating requirements and availability. Control tests should use the pooled pathway specification in Section 4.5, comparing equal b_{3p} coefficients with pathway-specific coefficients while retaining the same intercept and main-effect structure. The equality test concerns log-odds interactions; standardized probability contrasts should be reported separately. Weight selection and tuning must remain separate from final evaluation; choosing weights to maximize performance and reporting the same sample as validation would overstate evidence. Ranking performance, uncertainty, and decision usefulness

should be reported. Any conversion into failure probabilities requires a separately estimated calibration model and external validation. The current 0-100 scale alone supplies neither.

Replication and Revision

The fourth stage should examine transportability across tasks, expertise levels, institutions, and evolving AI systems. Replication should test whether construct meanings and predictor relationships remain comparable, rather than applying identical weights automatically. Sensitivity analyses should vary denominator floors, task requirements, task-criticality weights, component weights, and compensatory assumptions. Studies should report whether a small change in a modeling choice reverses substantive conclusions. Research should also assess burdens imposed through assessment and monitoring, including differential effects on workers using accessibility tools.

A cumulative program should retain useful components while revising unsuccessful ones. Failure of a compact score representation to outperform a flexible model of the same inputs would not negate evidence for capability change. Failure of the proposed CDM functional form would not establish that capability is static. Conversely, repeated evidence of beneficial independent learning would constrain claims about adverse substitution. The framework earns scientific value through comparisons that permit rejection, rather than through treating every outcome as compatible with the theory.

DISCUSSION AND GOVERNANCE IMPLICATIONS

Sovereign refines a resilience-compatible account of capability governance through a time-indexed mapping of assessed capability, feasible task allocation, and responsibility requirements. Responsibility-capability alignment specifies how stable competence can accompany greater vulnerability and how large assistance benefits can coexist with effective control. The distinction between output, capability, dependence, and control prevents several invalid inferences: productivity growth cannot establish retained competence; a performance gap cannot establish deterioration; and formal oversight cannot establish successful intervention. These distinctions also preserve beneficial cases. Assistance can improve both performance and learning, while specialized teams and reliable infrastructure can provide control without duplicating every automated function at the individual level.

The framework connects workforce development with system governance through responsibility for consequential tasks. A provisional practical implication is that organizations should identify the capabilities required to challenge outputs, handle exceptions, and maintain acceptable service. Periodic scenario-based assessments could then evaluate those responsibilities alongside ordinary productivity metrics. Where repeated assistance reduces opportunities to exercise relevant capabilities, structured practice or alternative team arrangements may warrant investigation. These are design implications of the theory, not interventions whose effectiveness has been established through Sovereign testing.

Governance evaluation should also distinguish the existence of controls from their performance. An override policy has limited practical meaning if the responsible team lacks

authority, information, time, or technical means to intervene. Audit records are useful only when they support reconstruction of consequential decisions. Alternative suppliers provide resilience only when their relevant failure modes are sufficiently independent. The proposed control profile encourages examination of those conditions. Such assessment can complement established AI governance activities rather than replace technical testing, security review, or broader evaluation of harms [11, 14].

Educational and organizational assessment should be aligned with purpose. Assisted tasks can evaluate effective collaboration with AI; carefully designed unaided tasks can evaluate retained knowledge and transfer. Neither assessment should substitute automatically for the other. Workforce measures should avoid interpreting a large assistance benefit as personal deficiency, particularly when the workflow intentionally relies on specialization or accommodations. Capability monitoring also creates questions of privacy, fairness, and employment consequences. Before operational use, affected groups should have meaningful input into the capabilities assessed, permissible tools, and interpretation of findings.

Theoretical precision limits policy claims. No mandatory threshold for acceptable APGI, CSM, or DEM follows from the current specifications. Procurement decisions, individual performance sanctions, or certification of organizational safety would require evidence of reliability, validity, incremental value, and consequences of use in the intended setting.

LIMITATIONS

The framework has several substantive limitations beyond the absence of original empirical testing. First, the selected studies span different populations, tasks, time horizons, and AI systems; their findings cannot be pooled informally into a general effect of AI on capability. Second, construct boundaries may overlap in practice. Verification can function as a learning activity, an assessed capability, and an organizational control, so operational definitions must distinguish process, state, and team-level action.

Third, the mathematical forms prioritize transparency over realism. CDM assumes bounded, smoothly changing capability and additive exposure contributions; actual learning may be discontinuous, multidimensional, or sensitive to sequencing. APGI depends on comparable score scales and is unstable near low assisted scores without a consequential floor choice. Optional CSM and DEM averages retain the aggregation restrictions specified in Section 4.8; DEM additionally assumes monotonic exposure contributions and depends on externally justified requirements, scenario definitions, and feasible task allocation. Control effects require separate estimation; the exposure score does not account comprehensively for threats and severity. Alternative forms may outperform every proposed specification.

Fourth, moving from individual assessment to workflow vulnerability risks aggregation error. The maximum-coverage rule applies only to substitutable single-person tasks; joint performance, workload, concurrent demand, and correlated unavailability require richer representations. Fifth, reverse causation and selection are plausible: weaker performers may choose more assistance, and organizations facing greater risk may strengthen controls. Sixth, rapidly changing tasks and models may alter both measurement and causal relationships. Subsequent studies must address those problems rather than

treating a statistically significant association as confirmation of the framework. Until then, Sovereign remains a provisional theoretical contribution.

CONCLUSION

Sovereign proposes a time-indexed measurement and allocation specification connecting retained human capability with coverage of consequential responsibilities under AI disruption. Within a framework compatible with organizational resilience, the specification links capability trajectories, task requirements, available personnel, and exercised controls; assisted-unaided performance contrasts provide complementary descriptive context. The framework permits both capability development and erosion, treats assistance gaps as descriptive contrasts, and locates capability-related exposure in unmet responsibilities under specified scenarios. Contrasting cases demonstrate how coverage and control can change vulnerability independently of individual learning. Subsequent experimental, longitudinal, and organizational studies are intended to evaluate the constructs, estimate relationships, compare competing specifications, and determine whether the integration provides explanatory or predictive value. The present contribution is a foundation for those tests, not a report of validated models.

DECLARATIONS

Ethics and Consent: Not applicable. The article presents conceptual theory development and involves no original human-participant research or personal-data analysis. Ethical approval and participant consent were not required.

Data and Code Availability: No original empirical datasets were generated or analyzed. All numerical illustrations use explicitly hypothetical values. Equations, assumptions, parameter values, and worked calculations are provided in the manuscript and Appendix A. No separate implementation code or trained model accompanies the conceptual framework.

Conflicts of Interest: The author declares no conflicts of interest related to the article.

Funding: No external funding was received for the research.

Author Contributions (CRediT): Emily J. Barnes: Conceptualization; Methodology; Formal analysis; Visualization; Writing - original draft; Writing - review and editing.

Use of Generative AI: OpenAI ChatGPT (Codex, GPT-6) assisted with journal template preparation and citation conversion. The author remains responsible for the accuracy and integrity of the final manuscript.

Preprint or Prior Version: No preprint or prior published or conference version exists.

REFERENCES

- [1] Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654), 187-192. <https://doi.org/10.1126/science.adh2586>
- [2] Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at work. *The Quarterly Journal of Economics*, 140(2), 889-942. <https://doi.org/10.1093/qje/qjae044>

- [3] Peng, S., Kalliamvakou, E., Cihon, P., & Demirer, M. (2023). The impact of AI on developer productivity: Evidence from GitHub Copilot [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2302.06590>
- [4] Salomon, G. (1990). Cognitive effects with and of computer technology. *Communication Research*, 17(1), 26-44. <https://doi.org/10.1177/009365090017001002>
- [5] Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences*, 20(9), 676-688. <https://doi.org/10.1016/j.tics.2016.07.002>
- [6] Bainbridge, L. (1983). Ironies of automation. *Automatica*, 19(6), 775-779. [https://doi.org/10.1016/0005-1098\(83\)90046-8](https://doi.org/10.1016/0005-1098(83)90046-8)
- [7] Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2), 230-253. <https://doi.org/10.1518/001872097778543886>
- [8] Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, 122(26), Article e2422633122. <https://doi.org/10.1073/pnas.2422633122>
- [9] Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, 100(3), 363-406. <https://doi.org/10.1037/0033-295X.100.3.363>
- [10] Bransford, J. D., & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. *Review of Research in Education*, 24(1), 61-100. <https://doi.org/10.3102/0091732X024001061>
- [11] 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>
- [12] Green, B. (2022). The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review*, 45, Article 105681. <https://doi.org/10.1016/j.clsr.2022.105681>
- [13] Pearce, H., Ahmad, B., Tan, B., Dolan-Gavitt, B., & Karri, R. (2022). Asleep at the keyboard? Assessing the security of GitHub Copilot's code contributions. In *2022 IEEE Symposium on Security and Privacy (SP)* (pp. 754-768). IEEE. <https://doi.org/10.1109/SP46214.2022.9833571>
- [14] OWASP Foundation. (2025). 2025 Top 10 risk & mitigations for LLMs and Gen AI apps. <https://genai.owasp.org/llm-top-10/>
- [15] Li, Z., Avgeriou, P., & Liang, P. (2015). A systematic mapping study on technical debt and its management. *Journal of Systems and Software*, 101, 193-220. <https://doi.org/10.1016/j.jss.2014.12.027>
- [16] Jaakkola, E. (2020). Designing conceptual articles: Four approaches. *AMS Review*, 10, 18-26. <https://doi.org/10.1007/s13162-020-00161-0>
- [17] Hutchins, E. (1995). How a cockpit remembers its speeds. *Cognitive Science*, 19(3), 265-288. https://doi.org/10.1207/s15516709cog1903_1
- [18] Ducheck, S. (2020). Organizational resilience: A capability-based conceptualization. *Business Research*, 13, 215-246. <https://doi.org/10.1007/s40685-019-0085-7>
- [19] Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. S. (2021). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery. <https://doi.org/10.1145/3411764.3445717>

APPENDIX A: MATHEMATICAL PROPERTIES AND HYPOTHETICAL ILLUSTRATION

For constant non-negative L and F , Equation 3 can be written as a linear differential equation in K . When $L + F$ is positive, the equilibrium and solution are:

$$K^* = \frac{L}{L + F} \quad (11)$$

$$K(t) = K^* + (K_0 - K^*) e^{-(L + F)t} \quad (12)$$

The solution remains between its initial value and equilibrium. Capability grows when K lies below the equilibrium and declines when K lies above the equilibrium. When $L = F = 0$, capability remains constant. These statements are consequences of the assumed equation and do not verify the proposed learning mechanism. Numerical implementations should preserve the bounds; a coarse unconstrained Euler update can overshoot even when the continuous-time model is bounded.

For illustration only, consider $K_0 = 0.60$ and rates measured per month. With $L = 0.08$ and $F = 0.04$, the equilibrium is approximately 0.667, and the initial derivative is 0.008 per month. Under the optional extension, the first case can include $\lambda = 0.01$ and $\alpha I = 0.03$, demonstrating positive growth despite an additional attrition term. With $L = 0.02$ and $F = 0.06$, the equilibrium is 0.25, and the initial derivative is -0.028 per month. All values are hypothetical, have not been estimated from observations, and provide no forecast of actual learning. Their purpose is to show that the same formal structure permits accumulation and decline.

For APGI, a positive numerator cannot exceed Q^A because Q^U is non-negative; the denominator is at least Q^A . Each component is therefore at most 1 and at least 0. Non-negative weights summing to 1 preserve the resulting 0-100 bound. CSM is a convex combination of bounded control components. For $R > 0$ and $B \geq 0$, every proportional shortfall lies in $[0, 1]$; non-negative task weights summing to 1 preserve that bound for D . DEM is a convex combination of bounded exposure channels multiplied with 100, giving the stated 0-100 range. Bounds establish numerical consistency, not interval-scale validity, predictive calibration, or legitimate thresholds.

The worked example provides two additional checks for the optional exposure average. With equal weights, the four structural components contribute $100 \times (0.40 + 0.20 + 0.30 + 0.20) / 5 = 22$ points. A requirement shortfall $D = 0.25$ adds $100 \times 0.25 / 5 = 5$ points, producing 27 points. Changing assisted scores alone changes neither term. For the longitudinal example, solving $0.75 = 0.25 + 0.65 \exp(-0.08t)$ gives $t = \ln(0.65/0.50)/0.08$, approximately 3.28 months. At month 4, K is approximately 0.722 and d is approximately 0.037. Those results establish consistency under stated assumptions; threshold legitimacy and future failure effects require separate evidence.

APPENDIX B: CANDIDATE OPERATIONAL DEFINITIONS

The following operationalizations are proposals for domain-specific instrument development. The common 0-1 range requires anchored rubrics or calibrated measures, not arbitrary normalization across unlike constructs. Missing observations should remain identified as missing; assigning zero can falsely indicate absent capability or absent exposure. Equal weights may be used as a transparent exploratory baseline, but do not establish equal substantive importance.

Table B1: Variables and Proposed Operational Definitions

Variable / Unit	Candidate Operationalization
K; person, domain, time	Calibrated repeated unaided capability assessments; assess dimensionality and comparability.
A, S; person, time	Focal AI involvement and displacement of relevant human effort, independently observed.
P, V, E, X; person, time	Quality-adjusted practice, verification, exception handling, and transfer exposures.
L, F; person, time	Learning and attrition rates. Primary $F = \lambda$; optional $F = \lambda + \alpha I$. I is independently coded adoption of misleading procedures per fixed task-exposure window, normalized to $[0, 1]$. Rate parameters use a declared time unit.
Q^A , Q^U , APGI; person, dimension, time	Comparable assisted and unaided scores; positive proportional contrast, accompanied with complete inputs and signed differences; $\kappa > 0$ is the prespecified denominator floor.
R; workflow task, scenario, time	Minimum acceptable task score and deadline, justified before observing assessment results.
a; person, task, scenario, time	Binary task eligibility and availability within the deadline; assess staffing, correlated absence, and assignment constraints.
B; task, scenario, time	B^{obs} uses contemporaneous observed scores; B^{pred} uses expected scores at the responsibility date; $B^{(m)}$ describes uncertainty in the available-capability envelope. Assignment-specific forecasts evaluate a declared selection or reassignment rule across draws. Expected-score uncertainty excludes future-attempt variation. Joint or concurrent tasks require team scores or feasible allocation.
d, D, ω ; task / workflow, scenario, time	Requirement shortfall, task-weighted mean shortfall, and prespecified criticality weights. Report every critical shortfall and the maximum.
μ , ℓ , b, ϵ ; task assessment	Expected task score, capability loadings, task difficulty, and observation error. Calibrate anchors and propagate uncertainty.
SM, TD, PE, PP; workflow, scenario, time	Shared failure exposure, unresolved technical liabilities, persuasive-output exposure, and excessive permission scope or duration.
IR, VC, OC, FC, AC; responsible team	Demonstrated independent reasoning, verification, override, fallback, and auditability; retain components and binding minimum criteria.
CSM, DEM; workflow, scenario, time	Primary component profiles; optional descriptive averages require a specified purpose, fixed weights, and comparison with component-only representations.
x_i , c_h , z; prospective test	Selected exposure, selected control, and prespecified confounders or contextual covariates for a pathway-specific model.
Y; workflow, subsequent outcome	Independently adjudicated critical failure. Define the horizon, scenario, and outcome before collecting predictor scores.