

**Computational Coherence in AI-Supported Educational Decision-Making under Policy
Disruption: Operationalizing the Coupled Framework and a Monte Carlo
Demonstration**

Ayman Abu Hajaj and Nadine Sheikh

Dr. Ayman Abu Hajaj

Affiliation: Ono Academic College, Kiryat Ono, Israel

Email: ayman2424@gmail.com

ORCID: 0009-0002-1944-8473

Role: First Author and Corresponding Author

Nadine Sheikh

Affiliation: Girne American University, Kyrenia, Cyprus

Email: nadinedia@gmail.com

ORCID: 0009-0006-1527-7514

Role: Co-author

Corresponding Author:

Dr. Ayman Abu Hajaj

Ono Academic College, Kiryat Ono, Israel

Email: ayman2424@gmail.com

Computational Coherence in AI-Supported Educational Decision-Making under Policy Disruption: Operationalizing the Coupled Framework and a Monte Carlo Demonstration

Abstract

The use of artificial intelligence (AI) in education raises a question that predictive accuracy alone cannot answer: Do inference, coordination, simulation, and human governance remain valid and actionable when procedures, authority, and evaluative criteria change? This study adapts computational coherence to a coupled educational decision framework and operationalizes it through Monte Carlo simulation using synthetic data. Procedural, role-based, and evaluative instability were modeled at three levels. Four architectures - disconnected, one-way pipeline, coupled, and coupled with threshold-based human governance - were each run across 5,000 episodes per level. Under low instability, decision quality was 86.59 in the coupled-governed system and 88.63 in the disconnected system; under high instability, the relationship reversed (82.14 versus 72.35). In the latter condition, adaptability rose from 0.278 to 0.463 and unsafe decisions fell from 12.06% to 0.14%, at the cost of 1.18 response hours and human review in 31.94% of cases. Targeted tests showed that coordination improved feasibility more than additional inference under role-based disruption, while semantic updating improved quality far more than increased accuracy based on outdated meanings. This proof of concept tests mechanism consistency and boundary conditions; it does not provide empirical validation.

Keywords: computational coherence; artificial intelligence in education; educational decision-making; policy disruption; Monte Carlo simulation.

1. Introduction

Schools face complex decisions in the face of quickly evolving data, process, resources, and authorities. Artificial intelligence (AI) systems are employed or considered to identify at-risk students, for assessment, resource allocation, timetabling, admissions, policy writing and leadership support (Rababah & Rababah, 2024). While reviews show a growing range of applications, there are enduring challenges related to teacher participation, evaluation of implementation, transparency, and teacher protection from harm (Bond et al., 2024; Sposato, 2025; Wang et al., 2024; Zawacki-Richter et al., 2019, 2024). The Organisation for Economic Co-operation and Development (OECD) also states that value does not necessarily stem from cutting edge technology but the purpose of education, professional capacity, the institutional framework and responsible governance (OECD 2026).

The quality of a system can be evaluated by model performance such as accuracy, sensitivity, specificity, calibration, or predictive ability. These are necessary, but not sufficient, indicators. A dropout-risk score does not constitute a student; a student's timetable recommendation does not result in the creation of personnel; and a therapeutic need does not mean anyone can act, a therapeutic service exists or how a student's needs can be challenged. Professional interpretation,

professional delivery, translation to action, and evaluation of output in terms of rights, values, and accountability are prerequisites for computational output to be a valid educational decision. It is then the socio-technical system involving people, software, data, procedures, roles and resources that determine the outcomes, not the predictive model itself.

The prediction-decision gap expands when there is a change in the curriculum, changing eligibility, emergency directives, changing instructional methods, budget cuts, staffing changes, or changes in authority. These conditions not only change data distributions, but also institutional meaning. A past risk factor can become a new one; an official can no longer have access to authorizing an intervention; or an assumed service can be closed. Research on the implementation of policy reveals that directives are not imposed on schools, but rather are negotiated, negotiated and negotiated again as they are understood locally, as routines, within the practices of professional communities and in relation to the authority dynamics in each school (Coburn, 2001; Spillane et al., 2002). Research on crisis-leadership also focuses on the elements of interpretation, coordination, flexibility, and trust beyond just a quick execution of a procedure (Harris & Jones, 2020).

The theoretical framework for this article is constituted by four layers of products that emerge with it: inference, coordination, simulation, and human governance, all of these in a state of procedural, role-based and evaluative instability. It suggests that, as instability rises, the interactions between the layers gain in importance relative to the roles of the individual layers. The two questions that follow are: How can alignment between different types of layers be translated into measurable; and can the theoretical relationships be expressed in a practical, transparent, reproducible computational model that can be sensitive to boundary conditions?

To answer these questions, this article offers the notion of computational coherence. The coherence does not imply all the components give the same answers. Instead, it implies that their depictions of the situation and rules of action can be aligned, that they are up-to-date in terms of current policies, and that they can guide action that is aligned with authority, capacity, and human oversight. Coherence is thus an operational and diagnostics property of the decision system, not a replacement for decision quality, fairness or legitimacy.

This article aims to align the concept of computational coherence with the coupled theoretical perspective and to show with a Monte Carlo simulation, how coupling and human governance relate to decision quality, adaptability, response time and risk under various levels of institutional instability. The research question is: How can computational coherence across the four layers of the framework be represented, and how do the roles of computational coupling change when it is procedural, role-based or evaluative?

The article has three contributions. First, it translates the framework into a computational model where four layers of partially known instabilities are updated with three dimensions of instabilities. Second, it frames coherence as a complementary diagnostic and a part of a larger framework of decision quality, adaptability, and response time. Third, it directly addresses propositions which stem from the framework: the contingent advantage of coupling; the relevance

of coordination under role-based disruption; diminishing returns to accuracy when meaning is out of date; and the relevance of human governance for errors.

To provide clarity on meanings, operational definitions are provided for key terms to clarify the concepts:

In this study, the term artificial intelligence in education is defined as computational systems that aid in analysis, prediction, recommendation, simulation and/or coordination in education settings. Educational decision-making is the process of interpreting evidence and choosing or authorizing an action that impacts or is possible with respect to learners, staff, resources, or institutional policy. Policy disruption refers to the change in procedures, authority or evaluative criteria, which renders some of the previous assumptions as partially obsolete while institutional instability refers to the strength of such change in procedures, roles and evaluation. Computational coherence (CC) is a measure of how well the representational and action pathways of inference, coordination, simulation and human governance are mutually reconcilable with the existing institutional state.

Monte Carlo simulation refers to repeated runs of a probabilistic simulation model under specified conditions, for the purpose of exploring mechanism consistency and boundary conditions. Adaptability is defined as the percentage of gap filled by the final decision step in the disruption. Human governance is the legitimate management of the information, time and power to review, change, suspend, record and allow appeal of a consequential recommendation. These are operational definitions that apply specifically to this study and not as proposed measures of validation.

2. Algorithmic Accuracy to System-level Coherence

The educational decisions that shape and influence the curriculum are distributed across various stakeholders.

Cognitive and institutional responsibility of decision making are distributed in the education system. Distributed cognition views thoughts as distributed among people, tools, representations and routines and not residing in a single administrator or database (Hutchins, 1995; Hollan et al., 2000). The teacher is familiar with how the students function in class, the counselor with emotional situations, information systems track attendance and achievement, the leaders are aware of staffing limitations, and authorities determine eligibility and budgets. Only valid decisions can be made with respect to access restrictions and decentralised authority by coordinating these fragments.

While pattern recognition, information synthesis and comparison of alternatives can be extended through the use of AI, bounded rationality cannot be eliminated. Human and computer decision makers make decisions in constrained environments, with partial representations (Simon, 1997). Additional information doesn't mean that variables are the proper ones to solve the problem. Bias can creep in during the objective definition, data collection, selection of labels, model development, implementation or use (Baker & Hawn, 2022; Suresh & Guttag, 2021). Therefore,

calculation, recommendation, professional judgment and institutional decision must continue to be separate.

A Decision Support System (DSS) is designed to augment judgment in partially structured decision making tasks, rather than to take its place (Arnott & Pervan, 2008). This distinction has significance in education because the choices made have implications for the rights, opportunities and public resources of learners. The human-centered learning analytics, thus, focuses on the needs of the stakeholders, their capacity for action and their interpretive contexts (Alfredo et al., 2024; Rababah et al., 2023). Human-AI combinations do not automatically provide an advantage, rather, the advantages rely on the type of task, division of labor, the detection of error, and communication (Vaccaro et al., 2024). The question is not about which is "better", as much as it is about whether the architecture allows each actor to contribute where it knows something that is relevant or has something to offer in terms of authority. The material in this section draws on a wider group of sources. The information included here has been sourced from a broader range of sources. The model-based concept drift literature presents approaches to detecting shifts in the input-output relationship and implications of the degradation of the model performance (Bayram et al., 2022; Lu et al., 2019). However, a policy disruption can affect more than just a statistical distribution but can also change the meaning of a variable, its authority to act and the standard to be used when evaluating a decision. In this way, a model can be valid for historical labels but not be valid for institutional and/or educational labels. As in the overall framework, there are three aspects of institutional instability. Procedural instability (P) involves changes in workflows, eligibility, documentation, deadlines, and approvals; role-based instability (R) involves changes in authority, responsibility, availability, and reporting; and evaluative instability (E) involves changes in criteria for success, risk, achievement, priority, and/or fairness. It can be quantified as $v = w_p\delta_p + w_r\delta_r + w_e\delta_e$, with the values of the weights w and the magnitude of the changes in the different dimensions δ depending on the context. This is an analytical representation, not an empirical measure. Semantic change is not a fourth dimension, but a process that can occur mostly in the procedural or evaluative dimension: the name of the variable stays constant, its definition or allowed use of it changes. Capacity is also not a standalone dimension of instability but is a bound within the decision-making environment which might shift after disruption of procedure and / or roles. These differences will prevent conceptual duplication and help to determine which part of the system needs to be modified.

2.3 Defining Computational Coherence

Here, computational coherence is judged by how coherent the state representations, rules for state updates and action pathways of the four decision levels are, and how they are consistent with the current institutional context. The construct consists of three parts: These are the distance between layer representations, the current procedural, role-based, and evaluative state, the extent of contradiction between the layer representations themselves and the distance between the coordination activity and the authority of human approval and oversight. Coherence is a quality that should not be confused with uniformity. Any system with all parts repeating the same old assumption is uniform but not coherent as it is misaligned with the environment. At the opposite end of the spectrum, however, there can be a

temporary disagreement that exposes a misguided understanding or a valid normative disagreement. The measure thus punishes any deviation from the institutional state and any unexplainable contradiction but does not assume that all layers are to have the same information. Coherence does not necessarily mean decision quality. Validity, feasibility, timeliness, educational impact, legal and ethical considerations, explainability and correctability are all components of decision quality. Coherence is an investigation of mechanisms that make those properties possible. It may be used to assist in validation, but not a substitute for external validation, fairness testing, or impact assessment. As part of the criticism of social simulation, the present demonstration investigates the consistency of the theoretical argument with the code, but it does not assume that its parameters refer to a real institution (Anzola, 2023; Belfrage et al., 2024). Distinguish between computational coherence and related concepts. The concept is similar to calibration, interoperability, data consistency and organizational alignment but it cannot be reduced to these. A calibrated model can send a recommendation to an unprivileged user to perform an action; systems may interpret the data differently; and records may be internally consistent even if there is no service to carry out the recommendation. So the computational coherence then includes authority, constraints and consequences. Transparency and explainability are separate states. An explanation can help make an infeasible recommendation understandable, but a coherent episode does not necessarily have a sufficient explanation. The thinking, that is intelligible, contestable and stakeholder appropriate (Khosravi et al., 2022), is an important part of high-stakes educational decisions, as is institutional alignment. Coherence does not imply automatic compliance. The system should identify and report a contradiction if a procedure is old, outdated, or in conflict with protected rights and should not automate that procedure. Responsible coherence can thus be expressed in the form of abstaining from recommendation.

Table 1. Components of computational coherence in the demonstration

Components of Computational Coherence and Their Educational Significance

Component	Analytical question	Computational expression in the demonstration	Educational significance
Environmental alignment	Do layer representations correspond to the current procedural, role-based, and evaluative state?	Mean distance between layer representations and the institutional-state vector	Detecting reliance on outdated definitions of eligibility or attendance
Interlayer consistency	Can the representations of inference, coordination, simulation, and governance be reconciled?	Mean pairwise distance among the four layers	Aligning a recommendation with capacity, authority, and the workload scenario

Component	Analytical question	Computational expression in the demonstration	Educational significance
Governance alignment	Does the planned action accord with authority, oversight, and normative criteria?	Coordination-governance gap on the role-based and evaluative dimensions	Authorized approval, documentation, and the capacity to halt or appeal
Overall coherence (CC)	Is the system current, consistent, and mechanistically justifiable?	Weighted combination of environmental alignment, consistency, and governance alignment	A complementary diagnostic that does not replace decision quality or fairness

3. The Coupled Framework

3.1 Four Functional Layers

The inference layer estimates a state, risk or potential outcome and is a representation of uncertainty. It can be a statistical model, a rule-based system, a language model or a hybrid of these. It is not just about being correct, but about being faithful to the existing definition of variables, data provenance documentation and limitations disclosure. For complex systems with high dimensional impacts, a preference for interpretable models or at least interpretability and documentation that does not mask the process of decision-making (Rudin, 2019; Khosravi et al., 2022). The coordination layer translates recommendations into actionable steps, through the representation of roles, authority, capacity, dependencies, deadlines and escalation routes. It hinders recommendations which no authorised or competent player can take action on. This is a formal language for the problem in multi-agent systems, where agents with varying information, goals and roles interact in an interdependent manner (Wooldridge, 2009). However, in education, an "agent" can represent a role, a unit or a service, and the real authority is human. The simulation layer looks at how alternatives might play out over time and with constraints. It is not used to forecast "the one true path" for the future, rather it is used to compare scenarios, highlight bottlenecks, and uncover unintended consequences (Rababah, 2023). When a decision has to impact more than one group or when it has feedback loops, simulation can be very helpful. It encourages the development of theories by forcing a specification of mechanism and a study of their results under different parameters (Davis et al., 2007). However, transparency of the model is key: agents, environments, interactions, and time should be reported—not hidden behind the graphical output — (Dilaver & Gilbert, 2023; Grimm et al., 2020).

The human-governance layer defines objectives, boundaries, and authority; reviews consequential recommendations; enables modification, suspension, and appeal; and records responsibility. Meaningful governance is not a human who merely clicks approval: it requires information, time, genuine authority to depart from recommendations, and accountability (Amershi et al., 2019; Raji et al., 2020). International frameworks emphasize lifecycle risk

management, documentation, data protection, human oversight, and impact assessment, especially where rights or opportunities are affected (Autio et al., 2024; European Parliament & Council of the European Union, 2024; Tabassi, 2023; United Nations Educational, Scientific and Cultural Organization [UNESCO], 2023).

The concept of coupling, feedback and boundaries is discussed.

A change in the state or output of one layer can cause a change in the state or output of another layer by means of explicit rules, as in a coupling. When risk is inferred, the coordination process will be a capacity check; when a simulation is used, the question will be asked about how a referral threshold will impact workload; and when a governance process is used it will be determined whether the criterion and action are allowable and justifiable. If detected constraints return to inference or inform an alternative. This is a two-way process as opposed to a one-way pipe which would advance the output without challenging previous assumptions.

The coupling does not have to be fully technological, it can also happen in the software, team meetings, work protocols, shared records or a combination of these. The important thing is that information gets to a point of action and has the potential to fix a previous layer. Studies have shown that hybrid intelligence, the complementary capabilities of humans and computational systems, has value in task design and division of labor, and in the learning from each other, rather than in co-location alone (Dellermann et al., 2019; Hooshyar et al., 2025; Jarrahi, 2018; Molenaar, 2022).

Overcoupling is risky. Moderate changes can cause the system to slow, overload professionals and introduce error. Strong linkages can also cause accountability issues. For this reason, the framework calls for the introduction of controlled coupling: role-based permissions, expertise-sensitive trust weight, review thresholds, separation of duties, source documentation and safe disconnection. The simulation tests are performed for improved coherence and noise and workload added to the simulation.

Figure 1. The educational framework for using Artificial Intelligence in decision making.

Note. Policy disruption occurs as procedural, role-based and evaluation instability, and updates the inference, coordination and simulation layers of the policy. Human governance sets limits, validates or adapts recommendations and triggers feedback. There are aspects of coupling that feed into aspects of decision quality: adaptability, response time; and an operational measure of the quality of the relationships between the layers is computational coherence.

3.3 Theoretical Propositions

Increased institutional instability will lead to greater discrepancies in decision quality between a coupled system and one in which inference, coordination, and simulation are performed separately. If conditions are stable, then current alignments can be used; if

conditions are disrupted, then different layers lose their alignment at different rates, and coupling can be used to update other layers when a change is detected in one layer.

Proposition 2: If the disruption is mostly related to authority and responsibility changes, the quality of the coordination will be more influential on the feasibility of the decisions than an extra predictive accuracy improvement. Even the best recommendations are worthless if they fall into the wrong hands, if there is not enough capacity, or if there is no escalation path.

The convergence of improved predictive accuracy will lead to diminishing returns if the recommendation is based on a different meaning from the past, but not on a feasible action, or is not evaluated against future consequences. The effect of the interactions between the inference performance, semantic alignment, coordination and simulation should thus be studied.

Proposition 4: Meaningful human governance will reduce the likelihood of error propagation from the coupling, will enhance accountability, legitimacy, and the ability to appeal. Its effect is not from the presence of a person but from the approval thresholds, documentation, information, time, and authority to intervene.

Coupling will affect decision quality partly due to increased adaptability and RT changes. The relationship could be non-linear, as there can be other dependencies and reviews which may enhance correction, but can also extend it.

The Framework is designed to identify 3.4 Failure Modes.

The first failure mode is false coherence: The layers agree, because they all are based on the same outdated source. The second is a mismatch between inference (an identification of a need) and coordination (lack of service, capacity or authority to meet the need). In both ways the problem is not captured in the measurement of accuracy or agreement only, so that the measure of alignment with the institutional state is given significant weight.

The other failure mode is propagation of error (authority) in the one way pipeline and overload of review in excess coupling. If all discrepancies are reviewed, alert and delay fatigue can set in, or if they converge too early, local knowledge and professional disagreement can be lost. Such risks make it crucial that there is controlled coupling, information provenance is preserved and there is safe disconnection.

4. Methodology

4.1 Research Design and Status of the Demonstration

The study is a blend of conceptual development and a generative computational experiment. In the conceptual part, relationships among policy disruption, four decision layers and their computational coherence are defined. The experiment is to determine whether or not these relationships can be mathematically programmed and whether the system acts as per the

framework under certain conditions. Monte Carlo simulation takes a probabilistic model and repeatedly runs it - it is appropriate in cases where the rules for the mechanics are fixed and the instability and communication architecture are systematically varied.

Simulations allow testing with a known “ground truth” that can be produced on the computer, but is not always available from the field (Burton et al., 2006). The subject of the inner truth is not the empirical one. The model is not a specific school model, is not student data calibrated and estimates no real-world causal effect. It is a proof of concept and mechanism-consistency test: it demonstrates that the constructs can be encoded; it demonstrates that the measure reacts to the changes predicted; and it demonstrates that the model can generate results that test the theory. The purpose, the entities, the state variables, the updating rules, the experimental conditions and the outputs of the model are summarized below according to the guidance to be reported (Grimm et al., 2020).

In this unit, the student will identify the unit of analysis and describe the synthetic environment.

A unit of analysis is a decision episode, which refers to a situation where an educational institution needs to interpret information and make a decision to take action after a possible policy change. A new institutional-state vector $z = (p, r, e)$ is created in each episode.

In this sense, p stands for procedural aspects (definitions, processes, eligibility) r for roles, authorities, and availability and e for evaluative aspects (criteria, rights, priorities). They each express numbers between 0 and 1 and can be used as metaphoric coordinates to quantify distance—not in terms of procedures, authorities, or value.

A baseline state z_0 was sampled from a symmetric beta distribution. Then, a shock was applied on the three dimensions based on the probability and the intensity defined by $v = w_p\delta_p + w_r\delta_r + w_e\delta_e$. Low ($v = 0.15$), moderate ($v = 0.45$), and high ($v = 0.75$) instability were examined. In the primary experiment, the weights were balanced; for targeted tests, either the weight of the tasks was increased according to the task or the combined weight of the procedure and evaluation was increased.

4.3 Layer Representations and Partial Information

The four layers have four distinct representations, x_k , of the state vector. The representations start at the beginning of an episode roughly equal to the previous policy, z_0 . Following the shock, each layer is given an unobserved and noisy observation of the new state, o_k . As a result, a first gap arises between the changed institution and the earlier assumption-laden decision system. Observation accuracy is different for different functions. Procedural definitions and evaluative criteria are fairly accurate in terms of inference; role and process coordination is more accurate in terms of information; simulation has some combination of information and the lack of it; and the governance has more accurate information about authority and legal-normative criteria. Instability

results in an increase in observation noise. This is complementary skills, but not assuming any layer is seeing the institutional state as a whole. They will need to demonstrate with some of the information. If information is perfect, there is no need for coordination and if it is not, communication cannot fix a common erroneous source. The model explores the intermediate condition, where a consensus may be reached and expertise may be integrated—however, there may also be noise transfer and a false consensus.

4.4 Comparison Architectures

Four architectures have been compared. In the disconnected architecture, the representation of each layer is updated based on its local observation. In one way pipeline, the information flows from inference to coordination, simulation and governance without backward correction. In the coupled architecture, there is also a two-way communication, in which expertise is weighted. If in the coupled architecture with human governance, there is disagreement or high coordination-governance gap (over 0.10), then human review is triggered, but review observation is returned to all layers.

Table 2. Comparison architectures used in the simulation

Simulation Conditions and Differences among the Architectures

Architecture	Update rounds	Communication	Human-governance feedback	Organizational logic
Disconnected	1	None	None	Each layer operates on local information
One-way pipeline	3	Inference → Coordination → Simulation → Governance, with no backward correction	None	The recommendation proceeds through a sequential process
Coupled	4	Bidirectional and expertise weighted	None	Layers correct one another
Coupled + human governance	4	Bidirectional and weighted	Activated above a threshold of 0.10 and returns a correction	Selective, risk-based human review

In each round, every layer, inference, coordination, simulation, and human governance, updates its estimate using three sources: its estimate from the preceding round, the new information it observed locally, and estimates received from other layers, weighted by the degree of trust in their expertise.

The updating rule is $x_k(t+1) = \text{clip}[(1-\alpha-\lambda_k)x_k(t) + \alpha o_k + \lambda_k \sum_j w_{kj} x_j(t)]$. In this equation, $x_k(t+1)$ is the updated value of layer k . The parameter $\alpha = 0.48$ determines the weight assigned to new local information, whereas λ determines how strongly the layer is influenced by other layers. The weight w_{kj} represents layer k 's trust in layer j , based on the latter's expertise in the dimension under consideration.

No information is transferred among layers in the disconnected system ($\lambda = 0$). The one-way pipeline permits limited influence ($\lambda = 0.24$), whereas reciprocal influence is stronger in the coupled systems ($\lambda = 0.30$). The clip function constrains all values to the interval from 0 to 1. Parameter values were selected solely for the simulation and were not estimated from educational institutions.

4.5 Outcome Measures

The computational coherence measure (CC) captures how much the four layers of the system are coherent with each other and coherent with the actual state of the institutional system (z). It is divided into three components: environmental error (ϵ) is the mean distance from layer estimates to the current state; interlayer disagreement (D) is the distance between the different estimates of the layers; governance gap (G) is the difference between the role-based and evaluative dimensions of human governance and coordination. The measure was computed by means of the following:

$$CC = 1 - [0.50 \cdot \min(\epsilon/0.35, 1) + 0.30 \cdot \min(D/0.30, 1) + 0.20 \cdot \min(G/0.30, 1)]$$

A score near 1 means that the layers are well aligned, and aligned with institutional reality. The most weight is given to environmental error so that an agreement that is reached in error is not considered coherent. Since the weights and thresholds were selected for the theoretical demonstration and not based on empirical data in an educational institution, a sensitivity analysis was performed to determine if altering the weights and thresholds would affect the simulation findings.

Decision quality (DQ) was the sum of the four sub-scores of inference validity, coordination feasibility, simulation fidelity and human-governance quality, scaled from 0-100 and divided by coherence. A geometric mean was a "weakest-link" design: if one layer fails badly, the decision may be affected even if the other layers are performing well. Coherence, coordination feasibility, and governance quality scores < 0.55 were considered unsafe decisions. These synthetic measures are not accepted measures for assessing educational decisions.

The ability of the system to adapt was calculated as a percentage of the gap that the system closed by the end of the process as a result of the disruptions. The time to correct errors, review by humans, and update interlayer was also included in the response time. The analysis also included the percentage of cases sent for human review and the percentage of cases in need of rework. These steps enable the evaluation of whether coupling increases adaptability and hastens the route to an approved viable recommendation as opposed to just speeding up the technical processing.

4.6 Simulation Procedure, Uncertainty and Sensitivity

For each instability level, 5,000 scenarios were run with the same initial conditions, shocks and observations, and four different architectures. The first experiment consisted then of 60,000 runs. The results were always reproducible exactly for the same code and conditions with a fixed random seed (20260801). The means and computational 95% confidence intervals were computed for each measure. The intervals provided here are just for simulation variability and cannot be used to make inferences about educational institutions at the population level.

The sensitivity analysis was then performed to determine the changes in the results as a function of coupling strength and the threshold value required for human review. A total of 30 different combinations tested with 2,500 runs per combination, 75,000 runs.

Selected tests explored coordination based on changing roles and authority, and on the use of outdated data meanings. The six configurations were thus executed 30,000 times. In total, 165,000 executions were performed: 60,000 in the primary experiment ($3 \times 5,000 \times 4$), 75,000 in the sensitivity analysis ($30 \times 2,500$), and 30,000 in tests of Propositions 2 and 3 ($6 \times 5,000$). This is the number of executions, NOT unique scenarios.

Incorporate implementation protocols and transparency and be able to reproduce them.

The code was implemented and tested in Python with the use of NumPy, pandas, and Matplotlib. While computational implementation is not the necessary and sufficient condition of scientific validity, a number of conditions are required: correspondence between the constructs and code, clarity of parameter choices, sensitivity testing, and peer review. The code thus records the work performed on scenario generation, observations, architectures, measures and figures separately.

The following materials include the code, compressed episode-level data, summary tables, proposition tests, sensitivity analysis, metadata, and hash values. The automated checks confirmed 60,000 rows during the primary experiment, no missing values, CC was between 0-1, and DQ was between 0-100. With this documentation, researchers can change noise, thresholds and weights and determine if the results are consistent (Dilaver & Gilbert, 2023; Grimm et al., 2020).

5. Results

5.1 Main Results

Table 3. Main results by instability level and architecture

Results of the Primary Experiment by Instability Level and Architecture

Instability	Architecture	CC	DQ	Adaptability	Unsafe (%)	Response time
Low	Disconnected	0.873	88.63	0.067	0.00	3.12
Low	One-way pipeline	0.895	89.99	0.097	0.00	3.67
Low	Coupled	0.846	86.32	0.064	0.00	3.97
Low	Coupled + human governance	0.850	86.59	0.068	0.00	4.03
Moderate	Disconnected	0.828	82.02	0.170	1.02	4.07
Moderate	One-way pipeline	0.854	83.42	0.202	0.68	4.62
Moderate	Coupled	0.810	83.06	0.225	0.06	4.91
Moderate	Coupled + human governance	0.826	84.10	0.246	0.04	5.10
High	Disconnected	0.771	72.35	0.278	12.06	5.04
High	One-way pipeline	0.798	73.60	0.306	10.82	5.59
High	Coupled	0.779	79.84	0.412	0.42	5.86
High	Coupled + human governance	0.811	82.14	0.463	0.14	6.22

Under low instability, the disconnected system achieved $CC = 0.873$ and decision quality of 88.63, while the coupled-governed system achieved 0.850 and 86.59. The one-way pipeline performed best (0.895 and 89.99), suggesting that local observations and simple sequencing may suffice when change is limited, whereas coupling adds noise and overhead. Under moderate instability, quality reversed in favor of the coupled-governed system (84.10 versus 82.02); adaptability rose from 0.170 to 0.246, unsafe decisions fell from 1.02% to 0.04%, and review occurred in 16.84% of episodes. Similar CC values (0.826 versus 0.828) confirmed that coherence and quality are distinct. Under high instability, the coupled-governed system achieved $CC = 0.811$ and quality of 82.14, versus 0.771 and 72.35 in the disconnected system. The quality gap reached 9.78 points, adaptability rose from 0.278 to 0.463, unsafe decisions fell from 12.06% to 0.14%, and review occurred in 31.94% of cases, supporting Proposition 1.

Figure 2. Computational coherence and decision quality by architecture and instability level.

Note. The plots present synthetic means. Computational confidence intervals are negligible because of the number of runs and are therefore omitted; they should not be interpreted as empirical uncertainty concerning educational institutions.

5.2 The One-Way Pipeline and the Distinction between Coherence and Quality

The one-way pipeline was efficient and achieved high CC under stable conditions, but its advantage eroded as disruption intensified. Under high instability, its CC was 0.798—above the

disconnected system—yet decision quality was only 73.60, versus 79.84 in the coupled system and 82.14 in the coupled-governed system. Unsafe decisions reached 10.82%, compared with 0.42% and 0.14%, respectively. A pipeline can reduce disagreement by passing an early representation forward, but cannot use information about authority, capacity, or consequences to correct inference. Sequential agreement therefore does not ensure quality, supporting the use of CC alongside, rather than instead of, DQ.

5.3 Adaptability and Response Time

Adaptability was also raised in the face of disruption as there was a greater margin for error; however, there were differences between architectures regarding the extent of repair. The coupled-governed system mitigated 46.34% of the gap, while the coupled system did so at 41.20%, the pipeline at 30.64%, and the disconnected system at 27.83%, thus enacting adaptability. This gain came at the expense of coupled-governed system synthetic response time, which is 6.22 hours, and disconnected system synthetic response time, which is 5.04 hours (1.18 hours more), with the low instability being 0.91 hour more. The goal is not to be the fastest, but to be faster than the competition but not so fast that meaningful review is not possible.

Figure 3. Response time and adaptability of architecture and instability level.

Note. Hours are arbitrary units of time and are meant to show an operational trade off, not foretell the time it will take an educational institution to process.

5.5 Response to the Review Threshold

The sensitivity analysis revealed the dependency of benefits to calibration. With high instability, 67%–98% of episodes reached the threshold for activation of review, which enhanced CC, though it was a challenging human workload. But review happened in only 0%–4% of episodes (CC = 0.22) and, in some configurations, CC dropped to 0.69. Review was 67.48% with a 85.89 decision quality and maximum CC of 0.864, achieved at a coupling strength of 0.50 and threshold of 0.06. Approximating the primary experiment at $\lambda = 0.30$, the threshold 0.10, CC was 0.811 and review 32.56%. Human capacity and time should be considered when maximizing coherence.

Figure 4. Sensitivity of computational coherence and review rate to coupling strength and the review threshold under high instability.

Note. Darker areas indicate greater CC or review rates. The coherence may be unusually intense and necessarily an unsustainable level of review.

This is a 5.5 Targeted Tests of Propositions 2 and 3.

Role-based disruption is the type of disruption that is primarily concerned with changes in authority or responsibility or division of labor. As an example, a specific employee may no longer be able to make a decision on a request, support for a student may have passed to a

different employee or it can be unclear who is best suited to making a recommendation. What can matter under these circumstances is not only the quality of the AI system's prediction, but also the ability to make it implementable and legitimate actions.

In the simulation, the feasibility score of coordination was raised from 0.848 to 0.879 with a 30% increase in information available to coordination. The system improved its ability to identify who to take action, to whom the recommendation should be directed, and what resources were needed, and unsafe decisions dropped to 0%. A similar gain in inference quality was not associated with an increase in feasibility, which remained at 0.846. This gain did not result in any changes to the unsafe-decision rate. The small error of 0.848 versus 0.846 is not significant, but rather simulation noise.

This is the result that demonstrates Proposition 2: When disruption is due to changes in roles and authority, enhanced coordination can add value beyond enhanced predictive accuracy. A correct recommendation is useless if it is mailed to someone who has no power to put it into effect or if there is no means of putting it into effect. This is a demonstration of the mechanism in a computational model, not evidence from actual schools.

The test of Proposition 3 focused on the other scenario: data were analyzed based on a different meaning than the new policy required. If the definition of "regular attendance" shifts once a transition to hybrid has taken place, for instance, a very high accuracy model using the previous definition may still yield an incorrect recommendation.

If data meaning was still missing in the simulation, the overall Decision Quality score (DQ) hardly changed, from 77.90 to 77.84 with a large improvement in the accuracy of the inference. Against this very small loss of accuracy, no harm would be implied, but a random fluctuation of no real substantive significance. In contrast, the data meaning was updated to align with the new policy, which led to an increase in decision quality to 83.16 and the change in the level of inference validity was from 0.722 to 0.904.

In conclusion, a computational accuracy is not going to solve the problem of using a construct that has become invalid. Validating the data and measures as representations of the educational phenomenon of interest is crucial as a prerequisite for improving the accuracy of the data.

6. Discussion

The theme of coherence is explored as a system level contingent property.

The findings confirm that AI-enabling the quality of educational decision making is not just about the accuracy of the model, but about alignment of procedures, roles, criteria, operational constraints and oversight. When instability was high, the coupling of complementary expertise provided information, and governance provided information when discrepancies transcended a threshold. It is also relevant that the boundary condition is a factor: The coupled-governed system performed worse than the disconnected system and the

simple pipeline performed best under low instability. Communication can carry with it noise, take up time, and be irrelevant in normal scenarios, as evidenced by non-synergistic human-AI pairs (Vaccaro et al., 2024). Coupling is then valuable contingently. In stable institutions, simple procedure and local expertise might be enough; if disrupted, coupling can reveal assumptions that are out of date and facilitate correction in both directions. Decision architecture should be tailored to type of disruption, risk, and institutional capacity, not be made too complicated.

6.2 Human Governance as an Information Mechanism

Policy and ethics scholarship look at the governance of humans from a perspective of accountability, appeal, fairness, and privacy (Holmes et al., 2022; Miao & Cukurova, 2024; Porayska-Pomsta, 2024). The model introduces an epistemic function: review can provide information that is missing from the earlier layers like temporary authority changes, unrepresentative student data, or local constraints, etc. and return it to the earlier layers. Governance enabled learning, it did not just enable approval. Observations are still noisy and thresholded and they still have value in providing independent knowledge, justifying departures, stopping action and placing responsibility. If there is no information, time or authority, a human reviewer will be a rubber stamp. Under low instability, review rose from 5.04% to 31.94% under high instability, while thresholds of very low levels referred almost every case. Thresholds are therefore required to be transparent, contestable and based on the severity of the consequences, reversibility and staff capacity (Tabassi, 2023; Autio et al., 2024).

The Conceptual Model to the Computational Model (6.3) describes the process of moving from the conceptual model to the computational model.

There are decisions to make when translating theory into code that prose leaves vague: what is disrupted, what is seen by each layer, how the disagreement is measured, and when the review will occur. Three instability dimensions and two mediators are maintained in order to maintain theoretical grounding. CC is not the objective manifestation of the theory; each weight and threshold is a priority and targeted tests are not a causal proof. Simulations bring to the fore the assumptions for critique, but do not (unless done numerically) resolve normative conflicts (Anzola, 2023). Limited computational verification is warranted: limited code executed, values kept in specified ranges, results repeatable, outputs were responsive to parameters. Themes of the preliminary structural correspondence are also used to provide theoretical relationships to the mechanisms. Field data, expert assessment, and research in real institutions are the criteria needed for external and predictive validity and effectiveness in education (Belfrage et al., 2024; Sargent, 2013).

6.4 Leadership implications, procurement, and system design implications

The framework moves the focus of attention from "model accuracy" to the decision chain for educational leaders. Prior to implementation, they should determine relevant data, interpretive authority, services that allow action, consequences and authority to stop or change the process. Policy versions and effective dates should be recorded on the mapping in order to get the coherence (which is time dependent). The requirements for procurement should include performance evidence, exportable assumptions, source and version information, uncertainty representation, records of human modification and safe disconnection. The buyer should decide whether a system is backward-feed or merely outputs through a pipeline which can lead to error propagation. The recommendation mechanism needs to be restored to capacity, and to the authority. Assessment reform being developed should include identification of new meanings of historical grades, rights and agreements for resource allocation that are not solely about optimizing and shared evidence in multidisciplinary coordination that is limited in disclosure. Remain domain specific with the measures and thresholds.

Professional deliberation should be the preserve of educational leadership. The implementation of AI in schools is not just a software project; it's a process of change that demands capacity, trust, and a shared professional vocabulary. AI implementation in schools is not about software; it's about change and transformation, which must involve capacity, trust, and a shared professional language. Coupled systems can sometimes turn knowledge to numbers and legitimate difference to a "data gap. Responsible use distinguishes between two types of factual differences and disagreements that averaging will not resolve: those that are correctable and those that are not normative. Students and families are rights holders and have contextual knowledge, not simply an observational source; therefore, an appeal must be integrated into the coherence architecture.

7. An Agenda for Empirical Research

Content validation should first question administrators, teachers, counselors, care professionals, students, parents, measurement specialists, and legal experts about the need for four state attributes of meaning, authoritativeness, capacity, and evaluation as well as the need for coherence measures to capture consequential contradictions. Vignettes depicting policy change could help to determine which layer updates are needed and what coherent responses are required. Second, the decision-log analysis should be retrospective, and follow a clear path from identification to action, with all the versions of the policy and the data, all handoffs, waiting times, exceptions, corrections, and appeals noted. These data could be used to fit error and delay distributions. Third, without impacting on decisions, comparisons of disconnected, pipeline, and coupled architectures on the same cases could be made through shadow deployment. Evaluation should cover: Action availability, workload, fairness of groups, quality of corrections, quality of explanation. Phased or longitudinal studies should

be conducted afterward, to differentiate the impacts of prediction, coordination, and governance, or gains from clearer roles/extra resources could be mistakenly attributed to AI.

Last but not least, research needs to explore the impact of coherence on legitimacy and trust. If the stakeholders were not involved in setting the goals of a system and/or don't have the ability to appeal decisions in a system, then it can be fair to every participant, but unfair to the rest. Mixed methods research would enable the integration of some computational measures with interviews, observations and document analysis, thereby allowing for an assessment of whether the layers are in accord and whether that is due to an acceptable institutional process.

8. Limitations

The study is very limited. Its data are synthetic and make the institutional state a continuum of three dimensions, while real procedures, roles and criteria are categorical, hierarchical, and sometimes contradictory. Effect size is a parameter dependent object when prespecified noise and shock distributions are used, and sensitivity analysis is able to show this, but does not eliminate it. Governance is conceived as an information source and threshold, but in reality governance involves power, deliberation, relationships, and normative disagreement. Unsafe decisions are based on a theoretical threshold, and response time is in synthetic hours. The results represent examples of relationships and trade-offs, rather than forecasts of operations.

The model does not look at bias across groups nor does it learn across episodes. A system can have a high average with a negative impact on a specific group of people, and real institutions have memory, trust, and routines. Future research is needed to disaggregate the outcomes by group and to make thresholds, roles, and trust weights flexible as successes and failures occur, and as appeals are made.

Most importantly, a calculation does not constitute a solution that should be deployed. The simulation does not show that the joint system will be helpful in making decisions in schools, nor does it provide evidence for any specific threshold. It demonstrates the possibility of operationalizing the framework, and provides the opportunity for formulating the appropriate questions to be empirically examined. If a real system is to be used, it must be validated, assessed, protected, and involve stakeholders and have accountability mechanisms.

9. Conclusion

This study employed the computational coherence approach to the coupled approach for educational decision making when confronted with policy disruption. Coherence is not simply the accuracy of the predictor, nor is it simply an agreement between the four layers as a whole, but rather a dynamic alignment of the four layers' representation with the

procedural, role-based, and evaluative state, and of coordinated action with human governance.

The Monte Carlo demonstration illustrated the benefit of coupling is dependent on the depth of disruption. When the instability is low, the simple system had better quality, while the coupled system with human governance had better quality, better adaptability, and better safety, but required more time and review when the instability is high. Deliberate tests showed that the meaning of a message is critical for the situation in which it is disrupted and that more accurate messages cannot fix the old meaning.

The role of the simulation is not to establish effectiveness but to be used to illustrate theoretical propositions by helping to develop mechanisms that can be implemented, examined, and refined. This leads to the design implication controlled coupling: Bi-directional, but bounded, documented, risk-proportioned and safely disconnectable. The next step is to calibrate with field data and to undertake integrated assessment of outcomes, fairness, workload, legitimacy and appeal.

References

- Alfredo, R., Echeverria, V., Jin, Y., Yan, L., Swiecki, Z., Gašević, D., & Martinez-Maldonado, R. (2024). Human-centred learning analytics and AI in education: A systematic literature review. *Computers and Education: Artificial Intelligence*, 6, 100215. <https://doi.org/10.1016/j.caeai.2024.100215>
- Amershi, S., Weld, D., Vorvoreanu, M., Fournery, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Paper 3, pp. 1–13). Association for Computing Machinery. <https://doi.org/10.1145/3290605.3300233>
- Anzola, D. (2023). The theory-practice gap in the evaluation of agent-based social simulations. *Science in Context*, 34(3), 393–410. <https://doi.org/10.1017/S0269889722000242>
- Arnott, D., & Pervan, G. (2008). Eight key issues for the decision support systems discipline. *Decision Support Systems*, 44(3), 657–672. <https://doi.org/10.1016/j.dss.2007.09.003>
- Autio, C., Schwartz, R., Dunietz, J., Jain, S., Stanley, M., Tabassi, E., Hall, P., & Roberts, K. (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.AI.600-1>
- Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, 32, 1052–1092. <https://doi.org/10.1007/s40593-021-00285-9>

- Bayram, F., Ahmed, B. S., & Kassler, A. (2022). From concept drift to model degradation: An overview on performance-aware drift detectors. *Knowledge-Based Systems*, 245, 108632. <https://doi.org/10.1016/j.knosys.2022.108632>
- Belfrage, M., Johansson, E., Lorig, F., & Davidsson, P. (2024). [In]credible models—Verification, validation and accreditation of agent-based models to support policy-making. *Journal of Artificial Societies and Social Simulation*, 27(4), Article 4. <https://doi.org/10.18564/jasss.5505>
- Bond, M., Khosravi, H., De Laat, M., Bergdahl, N., Negrea, V., Oxley, E., Pham, P., Chong, S. W., & Siemens, G. (2024). A meta systematic review of artificial intelligence in higher education: A call for increased ethics, collaboration, and rigour. *International Journal of Educational Technology in Higher Education*, 21, Article 4. <https://doi.org/10.1186/s41239-023-00436-z>
- Burton, A., Altman, D. G., Royston, P., & Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine*, 25(24), 4279–4292. <https://doi.org/10.1002/sim.2673>
- Coburn, C. E. (2001). Collective sensemaking about reading: How teachers mediate reading policy in their professional communities. *Educational Evaluation and Policy Analysis*, 23(2), 145–170. <https://doi.org/10.3102/01623737023002145>
- Davis, J. P., Eisenhardt, K. M., & Bingham, C. B. (2007). Developing theory through simulation methods. *Academy of Management Review*, 32(2), 480–499. <https://doi.org/10.5465/amr.2007.24351453>
- Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid intelligence. *Business & Information Systems Engineering*, 61, 637–643. <https://doi.org/10.1007/s12599-019-00595-2>
- Dilaver, Ö., & Gilbert, N. (2023). Unpacking a black box: A conceptual anatomy framework for agent-based social simulation models. *Journal of Artificial Societies and Social Simulation*, 26(1), Article 4. <https://doi.org/10.18564/jasss.4998>
- European Parliament & Council of the European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*, L 2024/1689. <http://data.europa.eu/eli/reg/2024/1689/oj>
- Fullan, M., Azorín, C., Harris, A., & Jones, M. (2024). Artificial intelligence and school leadership: Challenges, opportunities and implications. *School Leadership & Management*, 44(4), 339–346. <https://doi.org/10.1080/13632434.2023.2246856>
- Grimm, V., Railsback, S. F., Vincenot, C. E., Berger, U., Gallagher, C. A., DeAngelis, D. L., Edmonds, B., Ge, J., Giske, J., Groeneveld, J., Johnston, A. S. A., Milles, A., Nabe-Nielsen, J., Polhill, J. G., Radchuk, V., Rohwäder, M.-S., Stillman, R. A., Thiele, J. C., & Ayllón, D. (2020). The ODD protocol for describing agent-based and other simulation models: A

- second update to improve clarity, replication, and structural realism. *Journal of Artificial Societies and Social Simulation*, 23(2), Article 7. <https://doi.org/10.18564/jasss.4259>
- Harris, A., & Jones, M. (2020). COVID-19—School leadership in disruptive times. *School Leadership & Management*, 40(4), 243–247. <https://doi.org/10.1080/13632434.2020.1811479>
- Hollan, J., Hutchins, E., & Kirsh, D. (2000). Distributed cognition: Toward a new foundation for human-computer interaction research. *ACM Transactions on Computer-Human Interaction*, 7(2), 174–196. <https://doi.org/10.1145/353485.353487>
- Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Buckingham Shum, S., Santos, O. C., Rodrigo, M. T., Cukurova, M., Bittencourt, I. I., & Koedinger, K. R. (2022). Ethics of AI in education: Towards a community-wide framework. *International Journal of Artificial Intelligence in Education*, 32, 504–526. <https://doi.org/10.1007/s40593-021-00239-1>
- Hooshyar, D., Šír, G., Yang, Y., Kikas, E., Hämäläinen, R., Kärkkäinen, T., Gašević, D., & Azevedo, R. (2025). Towards responsible AI for education: Hybrid human-AI to confront the elephant in the room. *Computers and Education: Artificial Intelligence*, 9, 100524. <https://doi.org/10.1016/j.caeai.2025.100524>
- Hutchins, E. (1995). *Cognition in the wild*. MIT Press.
- Jarrahi, M. H. (2018). Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Business Horizons*, 61(4), 577–586. <https://doi.org/10.1016/j.bushor.2018.03.007>
- Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y.-S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S., & Gašević, D. (2022). Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence*, 3, 100074. <https://doi.org/10.1016/j.caeai.2022.100074>
- Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., & Zhang, G. (2019). Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering*, 31(12), 2346–2363. <https://doi.org/10.1109/TKDE.2018.2876857>
- Miao, F., & Cukurova, M. (2024). AI competency framework for teachers. UNESCO. <https://unesdoc.unesco.org/ark:/48223/pf0000391104>
- Molenaar, I. (2022). Towards hybrid human-AI learning technologies. *European Journal of Education*, 57(4), 632–645. <https://doi.org/10.1111/ejed.12527>
- OECD. (2026). OECD digital education outlook 2026: Exploring effective uses of generative AI in education. OECD Publishing. <https://doi.org/10.1787/062a7394-en>
- Porayska-Pomsta, K. (2024). A manifesto for a pro-actively responsible AI in education. *International Journal of Artificial Intelligence in Education*, 34, 73–83. <https://doi.org/10.1007/s40593-023-00346-1>

- Rababah, L. M. (2023). Stepping into language mastery: Virtual reality simulations as catalysts for EFL pronunciation enhancement. *Journal of Applied Learning and Teaching*, 6(Special Issue 1), 110–122.
- Rababah, L. M., Al-Shboul, O. K., & Ahmed Banikalef, A. A. (2023). Examination of the use of feedback in EFL writing instruction: A case study of Jordan. *Journal of Language Teaching and Research*, 14(1), 263–268.
- Rababah, L. M., & Rababah, M. A. I. (2024). Online social learning and instructional presence: Enhancing English education at a Jordanian university. *Forum for Linguistic Studies*, 6(6), 729–741.
- Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 33–44). Association for Computing Machinery. <https://doi.org/10.1145/3351095.3372873>
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1, 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
- Saltelli, A., Aleksankina, K., Becker, W., Fennell, P., Ferretti, F., Holst, N., Li, S., & Wu, Q. (2019). Why so many published sensitivity analyses are false: A systematic review of sensitivity analysis practices. *Environmental Modelling & Software*, 114, 29–39. <https://doi.org/10.1016/j.envsoft.2019.01.012>
- Sargent, R. G. (2013). Verification and validation of simulation models. *Journal of Simulation*, 7(1), 12–24. <https://doi.org/10.1057/jos.2012.20>
- Simon, H. A. (1997). *Administrative behavior* (4th ed.). Free Press.
- Spillane, J. P., Reiser, B. J., & Reimer, T. (2002). Policy implementation and cognition: Reframing and refocusing implementation research. *Review of Educational Research*, 72(3), 387–431. <https://doi.org/10.3102/00346543072003387>
- Sposato, M. (2025). Artificial intelligence in educational leadership: A comprehensive taxonomy and future directions. *International Journal of Educational Technology in Higher Education*, 22, Article 20. <https://doi.org/10.1186/s41239-025-00517-1>
- Suresh, H., & Gutttag, J. V. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (Article 17, pp. 1–9). Association for Computing Machinery. <https://doi.org/10.1145/3465416.3483305>
- Tabassi, E. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.AI.100-1>

- UNESCO. (2023). Guidance for generative AI in education and research.
<https://unesdoc.unesco.org/ark:/48223/pf0000386693>
- Vaccaro, M., Almaatouq, A., & Malone, T. W. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 8, 2293–2303. <https://doi.org/10.1038/s41562-024-02024-1>
- Wang, S., Wang, F., Zhu, Z., Wang, J., Tran, T., & Du, Z. (2024). Artificial intelligence in education: A systematic literature review. *Expert Systems with Applications*, 252, 124167. <https://doi.org/10.1016/j.eswa.2024.124167>
- Wooldridge, M. (2009). *An introduction to multiagent systems* (2nd ed.). Wiley.
- Zawacki-Richter, O., Bai, J. Y. H., Lee, K., Slagter van Tryon, P. J., & Prinsloo, P. (2024). New advances in artificial intelligence applications in higher education? *International Journal of Educational Technology in Higher Education*, 21, Article 32. <https://doi.org/10.1186/s41239-024-00464-3>
- Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education—Where are the educators? *International Journal of Educational Technology in Higher Education*, 16, Article 39. <https://doi.org/10.1186/s41239-019-0171-0>