

FEDERATED EXPLAINABLE MULTI-AGENT GOVERNANCE FRAMEWORK FOR ADAPTIVE PERSONALIZED LEARNING EQUITY IN HIGHER EDUCATION

Kamal Singh Kunwar

Tribhuvan University ,Nepal

kamalsinghkunwar560@gmail.com

ORCID: (<https://orcid.org/0009-0007-5674-6900>)

RESEARCH ARTICLE | DOI: <https://doi.org/10.37792/rxtajs29>

Received: May 27, 2026 | Revised: September 15, 2026 | Accepted: September 16, 2026 | Published: September 2026
License: Creative Commons Attribution 4.0 International (CC BY 4.0)

ABSTRACT

The rapid deployment of artificial intelligence (AI) in higher education raises questions of equity, transparency, accountability, privacy, and governance in adaptive personalized learning. While AI-enabled personalization can increase educational responsiveness, top-down personalized learning architectures may reinforce inequality through decentralized institutional data representation, esoteric decision-making, and insufficient stakeholder agency. Prior work has advanced federated learning, multi-agent systems, explainable AI (XAI), and algorithmic fairness, yet such components have not been fully integrated into a governance architecture for equitable higher education. This paper proposes an adaptive personalized learning federated explainable multi-agent governance architecture for equitable higher education. A theory-building research methodology leveraging critical realism and abductive inference combines concepts in distributed machine learning, multi-agent coordination, XAI, educational equity, and AI governance. The architecture consists of four interacting layers: federated learning for distributed institutional data; multi-agent governance comprising student, instructor, policy, and optimizer agents; explainability for decision traceability and interpretation; and equity evaluation to inform ongoing bias mitigation. Theoretical triangulation, construct-to-function mapping, cross-domain consistency testing, and boundary-condition analysis serve as the analytical validation process. Privacy, distributed governance, XAI, and equity are systemically positioned as mutually constraining properties. The architecture can inform design principles and empirical evaluation for agency-aware, privacy-preserving, equitable institutional AI governance.

Keywords: *artificial intelligence in education, federated learning, multi-agent systems, explainable AI, educational equity, AI governance*

1. INTRODUCTION

1.1 Global Context

Higher education is gradually integrating artificial intelligence in learning analytics, recommendation systems, predictive analytics, intelligent tutoring, automated assessment and adaptive learning environments. A review of 146 research articles on AI in higher education revealed that research has increased considerably, but also highlighted a significant prevalence of computer-science-related research and less research focusing on educational issues and teachers (Zawacki-Richter, 2019). This still remains a relevant issue, because AI systems for education are not just computing tools, but they are also involved in decisions related to student guidance, academic pathways, assessment and progression.

Thus, while the proliferation of personalized learning has both benefited from and further accelerated a data-intensive paradigm of personalization, it has also exposed learners to new risks concerning data collection, representational content, algorithmic bias and opacity. The policy briefs produced by UNESCO on AI use in education stress that AI should be designed for inclusion, equity, transparency and human agency rather than efficiency in education (Miao, 2021). The crux is therefore not 'will AI personalize learning?' but 'what institutional or technical conditions enable personalized learning to be beneficial and accountable'.

A model-based architecture raises a governance issue: if data from different institutions are pooled in a central repository to build a single model, those with more digital infrastructure or more students could dominate the model-creation process. Educational data are variable; different institutions have different curricula, student populations, assessment practices, technological resources and digital culture, so a model trained on pooled data may not work in the same way for various institutions or learners.

This is one solution to this challenge - federated learning, where many clients collaborate on training one model while the training data remains decentralized (Kairouz, 2021). Educational use cases provide an example of what is possible. Studies on federated learning for educational analytics have shown the performance of decentralized training to be similar to that of centralized approaches, while not requiring sensitive educational data to be centralized in a single place.

But federated learning itself does not tell us how educational decisions should be governed, explained and checked for equity.

Likewise, explainable AI has risen in speed as a response to black box machine learning. Explainability in the eye of the XAI research community is in part focused on using explanation to increase transparency, facilitate debugging and auditing, and allows users to comprehend model behavior (Arrieta, 2020). But transparency doesn't necessarily mean fairness. A model may be transparent but still produce systematic disparities or unequal results. The problem in theory therefore becomes an integration issue. All of these tools are intricately related and different facets of the same socio-technical problem. Disjoining them in silos is inadequate in explaining how they interact and influence one another.

1.2 Problem Statement

Three issues raise the importance of this research effort. Firstly, a centralized AI architecture for education can lead to information asymmetry. Different educational institutions may vary in the amount of educational data they generate, access, label, and utilize. Then, a centralized architecture may make the personalization system privileged in the distribution of educational information.

Secondly, personalization raises issues of decision opacity. Learners and teachers may receive recommendations without an understanding of the logic behind a learning recommendation, interventions, and predictions. Explainability then becomes relevant not only for the explainability of algorithms but also for educational accountability. Thirdly, pedagogical institutions are frequently detached from AI architecture. Educational policies on student privacy, ethics, fairness, supervision, and accountability may exist externally to the operational system while the technical optimization continues. For example, algorithm audits have shown the importance of embedding accountability at all levels of the AI lifecycle (Raji, 2020).

1.3 Research Gap

The literature uncovers four, partially fragmented, research streams. On federated learning: main focus on decentralized training, privacy, communication efficiency, heterogeneous data and optimization (Kairouz, 2021; Li, 2020); educational applications showing good promising privacy-preserving potential, but federation not by itself an institutional governance mode for fairness and pedagogical decision-making; on multi-agent systems: main focus on autonomous agents, coordination, negotiation, distributed problem solving, decentralized intelligence (Wooldridge, 2009); on XAI: main focus on transparency, interpretation of models (Arrieta, 2020), explaining different stakeholders needs (Liao, 2020); on educational equity and AI-governance: focus on principles of fairness, accountability, inclusion, human oversight (Miao, 2021); Together, these streams of research suggest that privacy, interpretability, equity and institutional governance are respectively elaborated concepts within each of the sub-areas, but not in combination. Theorization of complex governance institutions. The gap is therefore not the lack of research on any of the four components of this model, but rather the lack of an explanation of how the four components can cohesively constitute an institution for equitable and adaptive personalization.

1.4 Research Questions

The research will answer four questions:

1. How can federated learning promote institutional data sovereignty and equity in AI-supported personalized learning?
2. How can multi-agent architecture allocate decision-making power in education while preserving institutional accountability?
3. How can explainability be a governance tool instead of simply an interpretability feature?

4. How can equity metrics be incorporated into a self-reflective adaptive governance loop for keeping AI-supported learning decisions in check?

1.5 Objectives

The overarching goal of the project is to:

- 1) create an integrated federated explainable multi-agent governance approach;
- 2) identify the functional linkages between distributed learning, coordination among agents, explainability, and equity assessment;
- 3) implement an equity-driven set of indicators for adaptive personalized learning; and 4) set the theoretical validity criteria and constraints as conditions for further empirical testing.

1.6 Contributions

There are four contributions. First, it synthesizes four already-debated dimensions of educational AI across all the existing literature into a single comprehensive governance architecture. Second, it theorizes explainability as interconnected with accountability, instead of simply being a feature of user-interfaces. Third, it theorizes an interdisciplinary, multidimensional equity evaluation model, connecting each indicator of fairness to adaptive system feedback. Fourth, it specifies explicit theoretical validation and future empirical validation requirements instead of framing this framework as empirically validated.

2. THEORETICAL FOUNDATIONS

2.1 Federated Learning

Federated learning involves many clients collectively training a machine-learning model without sharing each client's training data locally (Kairouz et al., 2021). Because student records may include sensitive academic, behavioral, demographic, and interactions data, federated learning is a compelling education application. Theoretically, what federation preserves is institutional participation, not simply privacy. Institutions remain a data-governance unit, rather than become a raw-data source. This distinction matters because it allows participation for institutions with disparate data systems and structures, without requiring direct control of local records. But federation does not automatically democratize data.

Federated data can still remain non-independence and non-identically-distributed, and institutions may have large or systematically-different datasets. The model considers federation as necessary but not sufficient for equity. Relevant education studies on this front indicate that simultaneous decentralization of data governance and federated learning for predictive modeling is possible.

Using federated learning while maintaining data decentralization, the study can compare the accuracy of models on pre-collected data, with no accuracy loss for federated models in comparison with models trained on centralized data. The research question thus transitions from centralized or decentralized to the governance of heterogeneous decentralized participation.

2.2 Multi-Agent Systems

Multi-agent systems comprise autonomous computing entities capable of interacting and collaborating within an environment (Wooldridge, 2009). They are relevant to educational AI, where it is possible to delegate functions to multiple specialized agents instead of relying on a single model making all decisions.

The proposed framework defines four principal agents:

- **Student agent:** represents learner preferences, progress, engagement, and expressed learning needs.
- **Instructor agent:** incorporates pedagogical judgment and instructional objectives.
- **Policy agent:** represents institutional rules, privacy requirements, fairness constraints, and governance policies.
- **Optimization agent:** evaluates system-level performance subject to explicit governance constraints.

The agents do not have equal power, in fact, no. Instead, their collaborations follow a policy hierarchy. This is an important difference, as unconstrained autonomy would merely create yet another layer of algorithmic opacity. There is emerging empirical evidence that multi-agent learners can enable differentiated engagements and thereby narrow the performance gaps across students with different levels of prior knowledge. The framework we propose has built upon this literature, with the institutionalization of agent coordination in a governance layer.

2.3 Explainable Artificial Intelligence

The goal of XAI is to provide AI actions and logic in a way that makes sense to stakeholders. According to Arrieta et al. (2020) there are varied types of interpretabilities and explainability, and explanation needs are dependent on properties of systems and stakeholder needs.

The proposed framework uses explanation at three levels:

1. **Model level:** which variables contribute to a prediction or recommendation?
2. **Decision level:** why was a particular educational action selected?
3. **Governance level:** was the decision consistent with institutional policy and equity constraints?

SHAP and LIME can act as candidates for the explanation method, but the method is not said to be universally enough. Instead, a validity of the explanation is to be tested for coherence, stability, intelligibility, and on the basis of the decision. This difference speaks to a conceptual flaw in the first paper, as explainability is not the same thing as fair. If an explanation admits to an unfair decision, it does not eliminate it. An explanation has to contribute to governance and equity assessment.

2.4 Educational Equity

Equity in education is not the same as equality since equitable systems acknowledge that students and school organizations may not start at the same place. There are competing measures of algorithmic fairness such as demographic parity, equal opportunity, equalized odds, and individual fairness (Mehrabi, 2021). The framework does not argue in favor of a single criterion of fairness that can define educational equity. But instead views in equity as a multi-dimensional concept with the following:

- outcome fairness;
- opportunity fairness;
- representation fairness;
- recommendation fairness;
- institutional participation fairness; and
- procedural accountability.

This distinction prevents a technically convenient metric from being treated as a complete measure of educational justice.

2.5 AI Governance

AI governance refers to the mechanism, responsibilities, procedures, and accountability structures that are embedded to influence the development and deployment of AI. The work on algorithmic audits advocates for lifecycle accountability instead of static compliance (Raji et al., 2020). UNESCO has similarly advocated policy measures to ensure inclusion, equity, transparency and human agency in AI-facilitated learning (Miao, 2021). This framework therefore considers governance as an embedded system function. Governance guidelines impose restrictions on model-fine tuning, agent actions, explanations, and fairness assessment, as opposed to purely external regulations.

3. RESEARCH GAP AND THEORY-BUILDING PROPOSITION

The main theoretical lacuna is an architectural lacuna. While current solutions only address single problems, they lack a comprehensive understanding of how they may work together. The proposed framework comes to the following theoretical hypothesis: Reasonable adaptive personalization can best be achieved if data decentralization, distributed decision-making,

explainability, and the ongoing assessment of equity inhibit rather than promote each other in a unified architectural system of governance.

This proposition generates four subsidiary propositions.

P1: Distributed Data Governance Proposition.

Federated learning can enhance institutional data sovereignty and offset reliance on central data collection, but equity demands explicit attention to participation and representation of federated nodes.

P2: Distributed Decision Governance Proposition.

Multi-agent coordination can distribute educational decision authority, but equity demands explicit policy constraints for agents' priorities, conflicts and escalation.

P3: Explainability-Accountability Proposition.

Explainability can enhance accountability through links between explanations, decision traceability and governance review, but not if explanations are only post-hoc model interpretability.

P4: Recursive Equity Proposition.

Equity can be an adaptive property when fairness assessments produce feedback that can influence model aggregation, agent decisions or recommendation policy.

These propositions evolve the "artificial intelligence as an ensemble of technologies" framework into a "theory of governed adaptive intelligence."

4. PROPOSED FEDERATED EXPLAINABLE MULTI-AGENT GOVERNANCE FRAMEWORK

4.1 Architecture

The framework includes four vertically integrated layers:

Federated Learning Educational data are left in their local environments.

Each node is used to train a local model, but only allowed model parameters or updates, not raw student records, are transmitted. Contributions are combined via a federated aggregation procedure. Two additional institutional governance constraints are added to ordinary federated learning. Equity-sensitive aggregation. This ensures that an institution is not treated as just one more computation node. It is evaluated according to its representation, data quality, and influence upon the model.

Multi-Agent Governance.

The second layer represents the four agents described earlier, with their results fed through the institutional governance constraints. For example, the learning optimization agent

produces a learning recommendation that maximizes predicted learning gain, but the policy agent acts to limit that recommendation if it exceeds rules or fairness thresholds, as does the instructor agent through pedagogical override and the student agent via learner-specific information. This architecture switches the optimization goal from maximize predicted learning performance to maximize learning utility conditioned on privacy, explainability, fairness, and institutional governance.

Explainability and Traceability

The third layer records and interprets decisions. Each consequential recommendation should produce an explanation record containing:

- decision identifier;
- model version;
- relevant input features;
- principal feature contributions;
- responsible agent;
- governing policy;
- explanation method;
- confidence or uncertainty;
- timestamp; and
- available human-review pathway.

This transforms explainability into an auditable governance mechanism.

Equity Evaluation

The fourth layer evaluates system behavior continuously. It does not merely calculate whether the model is accurate.

The equity layer evaluates:

- group-level outcome disparities;
- opportunity disparities;
- recommendation disparities;
- institutional participation;
- explanation consistency; and
- audit completeness.

The equity layer then produces feedback signals.

Table 1: Integrated Governance, Literature Comparison, and Equity Evaluation Framework

Layer / Dimension	Core Components	Governance Function	Proposed Indicator	Prior Limitation	Framework Advancement
Federated learning	Institutional nodes, local training, aggregation	Data sovereignty and distributed participation	Institutional Representation Ratio	FL primarily addresses privacy and optimization	Adds equity-sensitive participation monitoring
Multi-agent governance	Student, instructor, policy, optimization agents	Distributed decision authority	Agent Governance Compliance Score	MAS research rarely specifies educational accountability	Embeds agent decisions within policy constraints
Explainability	SHAP/LIME, decision traces, audit logs	Transparency and traceability	Explanation Stability Score	XAI often treated as post-hoc interpretation	Connects explanations directly to governance
Equity evaluation	Fairness metrics, bias detection	Continuous fairness monitoring	Adaptive Equity Score	Fairness often evaluated after model development	Makes equity part of the adaptive feedback loop
Institutional governance	Policies, committees, human review	Accountability and oversight	Auditability Index	Governance frequently remains external to technical systems	Embeds governance into system architecture
Adaptive personalization	Recommendations, learner pathways	Individualized learning	Recommendation Equity Score	Personalization can amplify existing disparities	Makes personalization conditional on equity constraints

Note. The indicators are conceptual constructs proposed for subsequent empirical operationalization; they are not presented as validated scales.

4.2 Equity Evaluation Model

The equity model is intentionally multidimensional.

Let:

- (F_o) = outcome fairness,
- (F_p) = opportunity fairness,
- (F_r) = recommendation fairness,
- (F_i) = institutional representation fairness,
- (F_x) = explanation consistency, and
- (A) = accountability/auditability.

A conceptual composite equity score can be represented as:

$$[EAI = w_oF_o + w_pF_p + w_rF_r + w_iF_i + w_xF_x + w_aA]$$

where $(w_o + w_p + w_r + w_i + w_x + w_a = 1)$.

The equation is not presented as an empirically validated universal fairness formula. Instead, it specifies the proposed construct structure and establishes a basis for future empirical calibration.

4.2 Equity Evaluation Dimensions

This equity evaluation layer captures fairness as a property of governance rather than as a statistical measure. In adaptive personalized learning, equity may be a matter of disparities in outcomes, access to opportunities, suggestions, institutional representation, explanation, or accountability. As each learning decision carries its own kinds and degrees of risk, there is no one-size-fits-all fairness measure, so the proposed paradigm considers six equity dimensions.

4.2.1 Outcome Fairness

Outcome fairness is concerned with whether learning recommendations, assessments or interventions delivered by an algorithm produce systematically different outcomes for different relevant groups of learners. A relevant group may be defined with respect to what is legally and institutionally appropriate, for example socioeconomic background, gender identity, disability status, and/or place of residence, among other contextual considerations. For outcome fairness, the two most common educational measurements used for measuring relevant group differences are completion, achievement, retention, learning gains and access to effective intervention(s). It's worth noting that relevant group outcome differences may not be directly indicative of Algorithmic Discrimination since educational outcomes are driven by both structural and contextual factors.

4.2.2 Opportunity Fairness

Opportunity fairness measures if learners with similar observed performance, learning needs or participation criteria are provided similar opportunities for more advanced learning, academic support, intervention or progression. This dimension is especially relevant as adaptive systems might play a role in determining access to enrichment and other opportunities, or to remedial provision. Opportunity fairness will, therefore, examine if similar learner profiles have been treated equally across groups, while taking into account contextually valid pedagogical differences.

4.2.3 Recommendation Fairness

Equity Is recommendation fairness achieved if a system suggests some learners follow paths, learn materials, interventions, or other trajectories that are not equally suggested to others? Recommendations might seem personalized but reaffirm historical trends present in institutional data. This criterion, as a result, considers distribution of recommendations across relevant groups and institutional environments: disparities should be analyzed by differences in learner characteristics, ability level, prior achievement, preferences, and pedagogical goals. In cases where significant disparities are unexplained, the equity layer can flag recommendations for further model audit, explanation, or governance review.

4.2.4 Institutional Representation Fairness

Institutional representation fairness considers whether the development of a federated model is skewed towards or by a limited number of participating institutions. While federated learning keeps the data at its source, protecting a set of data from revealing others' does not necessarily mean the development is representative of all contributors. Institutions' student body, technical ability and organizational capacity, curriculum structures, participation rates, etc. may vary significantly. In federated learning, a few larger and/or more comprehensive sets of data have potentially the greatest influence over the final model in the aggregation process. The broader framework therefore considers institutional participation, contribution balance, data diversity, and influence over aggregation as governance issues.

4.2.5 Explanation Consistency

Consistency of explanation asks whether similar decisions are given "close enough" stable and intelligible explanations in similar situations. Explainability is not just viewed as a means of communication but as a way of being responsible. For example, if a significantly similar learner case receives a different explanation without a reason related to the context, the model might be unstable, sensitive to the features, or the agent might be reasoning in an inconsistent way. As a result, the framework suggests observing the different explanations by learner group, institution and decision type. Consistent explanation should be considered in conjunction with explainability fidelity, since an explanation can be consistent but not necessarily precise.

4.2.6 Accountability

Accountability can be determined by mapping material decisions back to the responsible model, agent, institutional policy, and decision procedure. When a significant intervention occurs, an acceptable system should produce an accessible record that can identify the model version, agent, policy decision space and constraints, explanation result, and escalation or human review status. This record can then be used by an institution to audit decisions, address equitable or harmful errors, and identify accountability. Accountability, in this context, is a link between technical functioning and institutional governance.

4.3 Feedback Mechanisms

The architecture functions by three feedback loops between evaluation of the equitable outcome, explainability, multi-agent governance, and federated learning mechanisms. They provide the system with this property as opposed to a series of packages where the data are processed once and then passed onto the learners.

Equity Federated Learning. Bias signals identified when aggregating data will be used to perform model re-weighting, alternative aggregation strategies, extended local validation, or disaggregated focus on underrepresented institutional population subsets. Such signals are not intended to suggest individual bias correction, but serve as a mechanism by which bias signals inform the overall federated learning governance framework. Equity signals are thus used as monitoring and intervention triggers. Explanation. Multi-Agent Governance. Explanation output provides evidence for student, instructor, policy, and optimization agent relating to how a specific recommendation or intervention was generated.

Agents are enabled to rationalize features that might be problematic, contradictory decisions or inconsistent between optimization goals and institutional equity criteria. If consequential concerns are detected, the decision can be brought to human validation rather than automated resolution. Multi-agent Governance Federated Learning Governance agents ingest institutional policies and equity goals and translate them into model development constraints (which could be in the form of local updates acceptance, rejection, reweighting, escalation, and re-evaluation). Governance infrastructure can monitor or restrict model optimization based on institutional governance rather than only predictive performance.

So, the architecture is both recursive not linear: the model provides explanations, explanations inform governance, governance creates corrective constraints, and equity assessment tells us whether future system outputs continue to be consistent with institutional fairness standards. The recursive nature allows for constant monitoring and intervening without losing data sovereignty or assuming an institutional understanding of fairness is a fixed property defined in model development, but an ongoing governance process.

5. METHODOLOGY

5.1 Research Design

This study is a theory-building conceptual research design. It is not claiming to be empirical validated nor does it invent any student, institutional, LMS data. The difference is highlighted here because the conceptual framework is designed to develop testable propositions and an operational architecture for further empirical study.

The methodological process consists of four stages:

1. domain convergence mapping;
2. construct extraction and integration;
3. architecture derivation; and
4. theoretical validation.

5.2 Philosophical Position

What is the philosophical basis? Critical realism is the philosophical basis of the methodology. It is suitable because the system I am proposing has tangible, observable effects because of the interaction of agents within the system, and the technical and institutional mechanisms that support the implementation. You can observe that the recommendation has an effect, but the generative mechanisms that produce it are – model parameters, agent preferences, institutional policy, distribution of data, governance mechanisms.

5.3 Abductive Reasoning

Abduction is relevant as the research will involve "seeking the best available explanation for a whole set of research observations over a number of various fields". This is the case because the research does not try to understand a single hypothesis, or theory from one domain, but switches between many.

The abductive process asks:

1. What problem does each literature address?
2. What assumptions does each literature make?
3. What does each literature leave unexplained?
4. Which constructs can be integrated without theoretical contradiction?
5. What new relationships emerge from the integration?

5.4 Conceptual Synthesis

The synthesis follows a construct-to-function procedure.

Analytical Step	Question	Output
Domain mapping	What are the central constructs?	Domain construct inventory

Analytical Step	Question	Output
Functional analysis	What does each construct do?	Function map
Gap identification	What relationships are missing?	Integration gaps
Construct integration	Which constructs can interact?	Relational architecture
Governance alignment	What constraints are required?	Governance rules
Equity operationalization	What should be measured?	Indicator framework
Validation	Are relationships theoretically consistent?	Validated conceptual model

5.5 Theoretical Validation

Validation of the study is therefore of an analytical nature rather than statistical since the developed governance framework is not tested as an implemented system with empirical data. The governance framework is validated through four complementary types of theoretical validation to test its conceptual coherence, functional fit, internal logical consistency and conditions of applicability.

Theoretical triangulation

Each key element of the conceptual framework is reviewed with the literature that supports its origins and implementation. For example: for federated learning that resides within the literature of distributed and privacy-preserving machine learning; for multi-agent governance that has roots in distributed coordination and agent-based systems; for explainability that is supported by XAI and algorithmic accountability literature; and for the equity aspect that draws from educational fairness and algorithmic bias literature. Triangulation of these literatures minimizes reliance on one theoretical trajectory and enhances reasons for the inclusion of all three elements.

Construct-to-function consistency

Every construct should serve a well-defined purpose in the architecture. For instance, if the scope of the architecture is federated model training (model-specific architecture; helping with decentralized training), data sovereignty (institutions' specific data requirements) and multi-agent governance (role-specific decision-making) would be the function, explainability (interpretability and traceability) and equity assessment (raise signals for differences) the corresponding constructs.

Cross-layer consistency

The framework is analyzed for compatibility across its four layers: privacy requirements should not interfere with providing meaningful explanations; agent autonomy should not interfere with institutional governance; and personalization should not interfere with having equity objectives. Any conflicts are handled explicitly as governance design issues rather than assumed to dissolve in the technological design.

Boundary-condition analysis

Lastly, the framework lists circumstances where its validity may be limited; for example, when there is a high degree of resource inequality across the institution, when there is inadequate or overwhelmingly heterogeneous local data, when institutional governance is incompatible, when the technical infrastructure is inadequately developed, or when the fairness conditions cannot all be met simultaneously. These limitations define the scope of the framework, and guard against the framework's conceptual claims from being taken as evidence. The four validation procedures serve as a clear, disciplined basis for evaluating theory adequacy, apart from treating theoretical rationale as evidence.

6. FRAMEWORK RESULTS AND ANALYTICAL VALIDATION

6.1 Governance Integration

The first analytical outcome of this template is that federated learning, multi-agent systems, XAI, and equity evaluation can all be composed, provided each of their roles are carefully distinguished within the same governance architecture.

The layered model of federated learning, multi-agent governance, explainability, and equity assessment can be mapped to the following chain of functions: the federated learning model decides where data should be processed, enabling participant institutions to retain sensitive learner data within their own borders, while collaborating with other institutions to develop a shared model; the multi-agent architecture assigns responsibilities to each of the decision-making sub-systems and determines their appropriate collaboration; explainability communicates the reasons why a decision has been made, and attributes to the system the rationale behind decision elements; the integrity of the overall decision, even in the face of the explainability and multi-agent architecture, is then evaluated in terms of whether it would produce fair distributional impacts in light of the relevant goals and constraints of the system. The normative structure established by the governance process acts as the tie binding these technical aspects together, ensuring each decision can be understood as a consequence of institutional commitments. The overarching process can then be outlined as distributed data processing, coordinated decision-making, explainable outputs, equity assessment, and system adjustment.

6.2 Equity as a System Property

The second contribution made by this analytical model is the reconceptualization of equity from a post hoc measurement tool into an ongoing systemic characteristic. Traditional fairness measurements are performed after the fact of model development or implementation. This limits the usefulness of the measures in change enabling contexts, wherein new operational behavior and future model development are the responsibility of the equation maker, rather than feature engineers and other human developers. This architecture accounts for that by placing equity in the feedback loop of prediction and recommendation.

Disparities in our models' predictions and recommendations are identified, those disparity findings relayed to the equity governance agents for consideration, and subsequently acted upon in future model development and recommendation generation. These three components (prediction, evaluation, and remedial governance) form an iterative feedback loop of prediction, evaluation, correction, and prediction again. This model does not suggest that all disparities are unfair, or that all disparities should lead to adjusted model development or predictions. It merely places detected disparities into the context of an explanation and evaluation step prior to their possible use in model development and recommendation generation.

This is the conceptual as opposed to measurement contribution of this model.

6.3 Explainability as Governance Infrastructure

The third result is about where XAI fits into the architecture. Our definition of explainability is: as a form of governance infrastructure, and is not a technical or GUI feature alone. Any explanation has greater governance potential when it can be traced back to a particular decision, model version, responsible agent, relevant institutional policy, and post-review path. As a result, generating a feature-importance plot or an explanation alone is insufficient for accountability. Explanation requires decision traceability and information to enable relevant actors to judge and challenge important decisions or recommendations. Here XAI acts as a 'middle ground' between autonomous systems and institutional accountability.

6.4 Distributed Decision Authority

The fourth data-mining result relates to distributed decision authority in the multi-agent tier. By spreading decision authority across a greater number of agents, the model can draw attention away from exclusive reliance on a single algorithmic decision-maker, and make room for design decisions and constraints specific to certain roles. This in turn gives rise to conflicts among pedagogical, institutional, learner-centered, equity, and optimization interests, which will not necessarily align.

A proposed hierarchy of policy precedence resolves this in the framework: institutional policy constraints trump all, followed by equity constraints, pedagogical constraints, learner interests, and finally optimization. This hierarchy of policy can serve as a conceptual decision-making guide in supporting boundaries to agent autonomy, though this is not offered as the optimal solution but as a testable governance principle for further empirical assessment in simulation and in institution.

7. DISCUSSION

The framework advances four central theoretical arguments. First, federated learning may help achieve educational equity by shifting the structure of data participation.

The federated-learning literature already provides evidence that decentralization and the reduction of data centralization can help (Kairouz, 2021), and educational applications demonstrate comparable performance (Chiu, 2020). Federation is not a silver bullet that will mean data locality does not matter. When an institution's data is underrepresented on the federation, it is underrepresented, whether the data are in the federation or not. The pedagogical rationale of representation monitoring is therefore supplemented with a rationale of representation management.

Second, multi-agent systems can allocate decision rights, but multi-agent interaction assumes coordination principles. Kamar (2016)'s work on multi-agent learning environments by which teachers interact with differentiated student pairs provides pedagogical support for the multiple-agent structure; this article, however, adds the distribution of governance responsibilities across multiple agents.

Third, explainability should serve as an accountability tool. Arrieta et al. (2020)'s survey of explainability techniques reveals the multiplicity of explanation methods and stakeholder needs; LIME and SHAP should not be seen as all-encompassing. Explanation should be linked to traceable decision-making and institutional validation.

Fourth, equity must be continually monitored. Algorithmic fairness is multidimensional and different fairness metrics may be incompatible (Mehrabi et al., 2021). This paper does not suggest that individual fairness metrics constitute educational equity. Instead, the model advances a portfolio of indicators on which relative importance should be empirically and institutionally determined.

By doing so, the article puts educational technology theories of adaptation in conversation with theories of distributed AI governance. While the literature on adaptive learning emphasizes personalization and learning effectiveness, the model proposed in this article discusses governance as an additional internal property of AI architecture.

8. THEORETICAL CONTRIBUTIONS

The four main contributions are: Integrated governance architecture of federated learning, multi-agent decision-making, explainable AI and educational equity. While the theoretical and technical innovations behind each of these components are well established, integrating them around an explicit governance problem is in its infancy. Conceptualization of adaptive educational equity system.

Equity is proposed not as an end-result but as a system constraint that influences model outcomes. Four-level governance taxonomy. Data governance, decision governance, explanation governance and equity governance, as well as the relations between them, are explicitly differentiated. Institutional accountability model. Policy rules, audit trails, human

review and agent responsibilities form integral components of the technical architecture. These are discussed as meaningful hypotheses that need empirical support rather than validated results.

9. POLICY IMPLICATIONS

The framework implies four institutional policy priorities First, where feasible, introducing privacy-preserving distributed AI infrastructures rather than default centralization architecture across the university. Second, requiring explainability and auditability of consequential AI-supported decisions in education - as the evidence basis of algorithmic auditing research advocates for lifecycle-based internal auditing, rather than post-deployment only review (Raji, 2020). Third, instituting multidisciplinary AI governance committees in universities bringing together educational, technical, legal, and student representatives.

Fourth, enabling cross-institution federated collaboration with agreements covering data sovereignty, ownership of models, security provisioning, fairness standards and governance conflict resolution procedures. The policies echo the international policy discourse on the inclusion, fairness and transparency on responsible use of AI in education (Miao, 2021).

10. LIMITATIONS

The main limitation is that the framework has yet to be empirically validated. The proposed equity indicators, agent hierarchy and feedback mechanisms therefore constitute hypothetical constructs not validated tools. Second, Federated learning presents technical challenges including non-IID data, communication overhead, local model heterogeneity, security issues and aggregation techniques (Kairouz, 2021).

Third, fairness criteria may be incompatible with each other; improving one criterion may lead to the decrease of another, which renders the construction of the proposed composite equity model context specific.

Fourth, the institutional contexts vary across countries in terms of governance, privacy legislations, technical capacity and cultural perceptions of educational equity. Therefore, the framework should be adapted accordingly. Fifth, Explainability methods such as LIME and SHAP have methodological limitations that should prevent us from equating these to causal models. Hence, the framework requires empirical testing of explanation techniques prior to implementation.

11. FUTURE RESEARCH

Next steps empirical work should be carried out in four stages. First, the framework should be empirically validated on LMS data, including heterogenous institutional data that is ethically and legally permissible. The Open University Learning Analytics Dataset and any other standard datasets should serve as a sandbox for simulation and benchmarking studies. Second, agent-based simulation should be used to stress-test conflicts between student, instructor, policy, and optimization agents. Agent-based simulation should also investigate the

extent to which different governance hierarchies lead to different fairness and learning outcomes.

Third, cross-country comparative work should investigate differences in institutional and regulatory contexts that lead to different governance arrangements.

Fourth, studies should examine the effectiveness of alternative privacy, security, and verification mechanisms.

Blockchain-based audit trails may serve as a comparative platform for recording governance events, but should not be regarded as automatically solving all governance challenges. The 2026 advent of federated and explainable educational analytics illustrates that empirical work on governances is already underway at a fast pace; the challenge now is to progress from component-level technical investigations to contextually valid governance design experiments.

12. CONCLUSION

This work has proposed a federated explainable multi-agent governance architecture for adaptive personalized learning equity in higher education. Central to this is the central weakness of educational AI literature to date, which has largely regarded privacy, distributed intelligence, explainability, fairness and governance as discrete normative or technical issues. The resulting architecture unifies a layered architecture that combines federated learning for decentralized data and processing, multi-agent governance for distributed decision-making, explainability for decision-level transparency and traceability, and equity evaluation for continuous monitoring of the distributional and procedural downstream effects.

Accordingly, the contribution is an architectural one rather than a technological one. Federated learning, multi-agent systems and explainable AI are well established lines of research; however, the contribution demonstrates how this can be orchestrated through an equity-aware governance architecture for adaptive learning. It must be noted that the architecture itself cannot be regarded as empirically validated, but rather as providing the theoretical base for generating testable hypotheses, operationalized indicators, forms of governance and boundary conditions for subsequent empirical validation. This makes the contribution richer and more robust rather than weaker. It provides a canvas for future experimentation with LMS data, multi-agent simulations, institutional piloting and cross-country comparative research. If validated, it holds the potential to facilitate a move from largely centralized opaque personalization to adaptive educational AI that is distributed, explainable, accountable and equitable.

REFERENCES

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* (pp. 1–13). Association for Computing Machinery. <https://doi.org/10.1145/3290605.3300233>
- Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- 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>
- Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., Ramage, D., Segal, A., & Seth, K. (2017). Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* (pp. 1175–1191). Association for Computing Machinery. <https://doi.org/10.1145/3133956.3133982>
- Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (pp. 214–226). Association for Computing Machinery. <https://doi.org/10.1145/2090236.2090255>
- Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28, 689–707. <https://doi.org/10.1007/s11023-018-9482-5>
- Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86–92. <https://doi.org/10.1145/3458723>
- Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, 29. Curran Associates, Inc.
- Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudík, M., & Wallach, H. (2019). Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). Association for Computing Machinery. <https://doi.org/10.1145/3290605.3300830>
- Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1, 389–399. <https://doi.org/10.1038/s42256-019-0088-2>

- Jodayree, M., Ghafi, A. K., Atashafrouz, M., & Shafiabadi, M. H. (2026). Privacy-preserving federated learning for interpretable student at-risk prediction across schools. *Scientific Reports*, 16, Article 26448. <https://doi.org/10.1038/s41598-026-56609-z>
- Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, A. B., Eichner, H., El Rouayheb, S., Evans, D., Gardner, M. J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., ... Zhao, S. (2021). Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2), 1–210. <https://doi.org/10.1561/22000000083>
- Kizilcec, R. F., & Lee, H. (2022). Algorithmic fairness in education. In *The ethics of artificial intelligence in education*. [Use the final publisher/journal version required by the source used in the manuscript.]
- Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3), 50–60. <https://doi.org/10.1109/MSP.2020.2975749>
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, 30. Curran Associates, Inc.
- McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics* (pp. 1273–1282). PMLR.
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), Article 115, 1–35. <https://doi.org/10.1145/3457607>
- Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Gebru, T., & Mitchell, M. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220–229). Association for Computing Machinery. <https://doi.org/10.1145/3287560.3287596>
- Mittelstadt, B. D. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, 1, 501–507. <https://doi.org/10.1038/s42256-019-0114-4>
- Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2), 1–21. <https://doi.org/10.1177/2053951716679679>
- Ouyang, F., & Jiao, P. (2021). Artificial intelligence in education: The three paradigms. *Computers and Education: Artificial Intelligence*, 2, Article 100020. <https://doi.org/10.1016/j.caeai.2021.100020>

- 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>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?”: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). Association for Computing Machinery. <https://doi.org/10.1145/2939672.2939778>
- 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>
- Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 59–68). Association for Computing Machinery. <https://doi.org/10.1145/3287560.3287598>
- Suresh, H., & Guttag, 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* (pp. 1–9). Association for Computing Machinery. <https://doi.org/10.1145/3465416.3483305>
- UNESCO. (2021). *AI and education: Guidance for policy-makers*. UNESCO.
- van Haastrecht, M., Brinkhuis, M., & Spruit, M. (2024). Federated learning analytics: Investigating the privacy-performance trade-off in machine learning for educational analytics. In *Artificial intelligence in education* (pp. 62–74). Springer. https://doi.org/10.1007/978-3-031-64299-9_5
- Williamson, B., & Eynon, R. (2020). Historical threads, missing links, and future directions in AI in education. *Learning, Media and Technology*, 45(3), 223–235. <https://doi.org/10.1080/17439884.2020.1798995>
- Wooldridge, M. (2009). *An introduction to multiagent systems* (2nd ed.). John Wiley & Sons.
- Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology*, 10(2), Article 12. <https://doi.org/10.1145/3298981>