

ARTIFICIAL INTELLIGENCE

Epistemic Justice-Aware Checklist Learning: An Adaptive and Auditable Governance Framework for Bias and Fairness in Generative AI Curriculum Design

Anjana Pareek

Department of Education, Jagannath University, Jaipur, Rajasthan, India

Published October 2026

Abstract

Generative AI increasingly participates in curriculum design, yet the resulting syllabi, learning outcomes, and assessments may reproduce subtle epistemic injustices that existing governance instruments fail to detect. We propose an adaptive and auditable framework that learns a bias and fairness checklist for generative AI curriculum artifacts rather than treating such a checklist as a static set of rules. The proposed method encodes curriculum texts into two disentangled latent subspaces, one capturing pedagogical utility and the other capturing epistemic justice violations, through a contrastive objective combined with intersectional adversarial debiasing. A psychometric utility-preserving penalty constrains this separation so that debiasing does not degrade instructional quality. Each checklist item is then evaluated by an evidential deep learning head that outputs a Dirichlet distribution over compliance states, hence yielding calibrated violation probabilities and explicit uncertainty estimates. An uncertainty-aware routing mechanism subsequently directs low-confidence drafts to a Faculty Expert Panel, whose decisions and justification notes are returned to a fine-tuning dataset. This feedback loop updates the checklist definitions and model weights iteratively, therefore allowing the governance instrument to adapt to newly identified forms of subtle bias. The main contribution is a mathematically grounded pipeline that couples representation disentanglement, evidential uncertainty quantification, and human expert feedback within a single auditable process. For curriculum governance, the framework offers transparent violation probabilities, traceable expert interventions, and a documented evolution of fairness criteria, which together support accountability in generative AI-assisted educational design.

Keywords: epistemic justice; generative artificial intelligence; algorithmic bias and fairness; adaptive curriculum design; AI governance and auditing

1 Introduction

Generative AI has moved rapidly from experimental curiosity to routine infrastructure in higher education, where it now drafts syllabi, proposes learning outcomes, generates assessment items, and recommends readings. This shift promises efficiency and personalization, yet it also introduces a governance problem that existing instruments are poorly equipped to solve. Curriculum artifacts are not neutral containers of content; they encode assumptions about whose knowledge counts, which learners are imagined as the default, and which epistemic traditions are treated as authoritative. When a generative model produces a syllabus, it may silently reproduce epistemic injustices such as the omission of non-Western scholarship, the framing of disability through deficit language, or the privileging of dominant cultural references. Systematic reviews of ethical integration in higher

education consistently identify fairness, transparency, and accountability as central concerns, but they also reveal that institutions lack operational mechanisms for detecting such harms in everyday curriculum workflows [1] [2].

Current governance practice relies heavily on static checklists and principle documents. Delphi studies and conceptual frameworks have produced valuable dimensions and priorities for ethical generative AI use, and several works propose audit checklists for procurement and deployment [3,4]. However, a checklist written once cannot anticipate the evolving forms of bias that generative models produce, nor can it express how confident an institution should be that a given violation has actually occurred. Recent work explicitly calls for moving beyond a compliance checklist toward a dynamic, iterative cycle of equitable assessment, and policy-as-code proposals similarly argue that governance should be executable rather than declarative [5,6]. These calls motivate a learning-based approach in which the checklist itself is a model that improves with use.

We propose an Epistemic Justice-Aware Bias and Fairness Checklist Learning framework that operationalizes ethical governance as a trainable, auditable artifact. The framework encodes curriculum texts and generative AI outputs into disentangled representations that separate pedagogical intent from epistemic justice concerns, using a contrastive objective in the spirit of SimCLR and MoCo [7,8] built on transformer sentence embeddings [9]. Intersectional adversarial debiasing removes sensitive-attribute information from the justice subspace [10], while a psychometric utility-preserving penalty, related to fair representation learning and mutual information minimization, prevents the debiasing from eroding instructional value [11,12]. Each checklist item is evaluated by an evidential deep learning head that outputs a Dirichlet distribution over compliance states, yielding calibrated violation probabilities and explicit uncertainty [13,14]. Margin-based calibration further aligns these probabilities with observed likelihoods [15,16]. An uncertainty-aware routing mechanism then sends ambiguous cases to human reviewers, and their feedback refines both the checklist items and the model through active learning [17,18].

The key difference from prior work is that fairness criteria are not fixed a priori but are learned and revised within a closed governance loop. Static fairness toolkits and lexicon-based detectors cannot represent intersectional or context-dependent harms, and they offer no principled way to abstain when evidence is weak [19,20]. Philosophical critiques of universal fairness definitions reinforce this point: fairness is contextual and entangled with power and knowledge production, so a governance instrument must be sensitive to institutional context and must document how its criteria change over time [21,22]. Our framework therefore treats uncertainty as a first-class output and expert disagreement as a signal for checklist revision, rather than as noise to be suppressed.

The contribution of this work is threefold. First, we formulate checklist-based curriculum governance as a representation disentanglement problem in which pedagogical utility and epistemic justice are modeled in separate latent subspaces, allowing harmful bias to be isolated without discarding educational value. Second, we introduce an evidential, margin-calibrated compliance head that quantifies uncertainty per checklist item, enabling a routing policy that escalates only genuinely ambiguous cases to human reviewers. Third, we design a closed-loop governance artifact in which reviewer decisions and justifications iteratively update both model parameters and checklist definitions, producing an auditable record of how fairness criteria evolved. Together these elements yield a governance process that is transparent, context-sensitive, and accountable, which is precisely what higher education institutions require when generative AI participates in curriculum design.

The remainder of this paper reviews related work on ethical governance of generative AI in higher education, bias and fairness in NLP and uncertainty-aware learning; presents preliminaries on disentangled representation learning, adversarial debiasing and evidential deep learning; details the proposed Epistemic Justice-Aware Bias and Fairness Checklist Learning framework; describes the experimental setup and results; and discusses implications, limitations and future directions before concluding.

2 Related work

Research on generative AI governance in higher education has expanded quickly, yet the literature remains fragmented across ethical principles, institutional policy, and technical fairness methods. We organize the discussion around three strands that bear directly on the proposed framework: governance frameworks and checklists for educational AI, bias and fairness in language technologies, and uncertainty-aware learning with human feedback.

2.1 Governance frameworks and checklists for educational AI

A substantial body of work maps the ethical terrain of generative AI in higher education. Delphi-based studies and systematic reviews have identified key dimensions and stakeholder responsibilities for institutional governance, emphasizing transparency, fairness, and faculty training on equitable assessment design [3] [1]. Complementary conceptual work proposes integrated frameworks that bridge abstract values and regulatory requirements, arguing that governance should be dynamic rather than a fixed policy template [23]. Equity-driven solutions have also been advanced, with audit checklists for procuring and deploying AI grading systems and analyses of how biases are operationalized and detected through fairness metrics [4].

A recurring theme in this strand is the inadequacy of static compliance instruments. Operational frameworks for equitable assessment argue for a dynamic, iterative cycle instead of a checklist of ethical principles, and policy-as-code proposals seek to make governance executable by encoding rules as machine-checkable artifacts [5,6]. Related work on human-AI collaboration in pedagogy similarly warns against treating ethics as an ex-post compliance exercise and instead recommends embedding privacy-by-design and fairness considerations into educational decision-making [24]. These contributions motivate our central premise: the checklist should be a learning system whose criteria evolve with evidence, not a document written once.

2.2 Bias and fairness in language technologies

Technical fairness research offers a rich toolkit that educational governance has only partially absorbed. Adversarial debiasing removes sensitive-attribute information from learned representations through a min-max game between an encoder and an attribute classifier [10], while fair representation learning and mutual information minimization provide complementary objectives for controlling what a representation encodes [11,12]. Contrastive learning has become a standard mechanism for shaping embedding geometry, with SimCLR and MoCo establishing the core objectives and Sentence-BERT adapting them to sentence-level representations [7,8] [9]. Fairness auditing toolkits such as AI Fairness 360 and Fairness Indicators operationalize group-level metrics for practitioners [19,20].

Nevertheless, critiques of universal fairness definitions show that fairness metrics are contextual and cannot be specified independently of the power relations and knowledge practices in which they are

embedded [21,22]. This insight is especially consequential for curriculum governance, where harms are often epistemic rather than purely distributive: the exclusion of non-Western scholarship or the framing of disability through deficit language may not register in conventional group-fairness metrics. Our framework responds by treating checklist items as semantic concepts evaluated in a dedicated justice subspace, so that intersectional and context-dependent violations can be represented explicitly.

2.3 Uncertainty-aware learning and human feedback

Reliable governance requires knowing when a model should abstain. Evidential deep learning addresses this need by predicting Dirichlet parameters over class probabilities, so that the total evidence directly indicates confidence [13,14]. Calibration research provides the complementary machinery for aligning predicted probabilities with observed outcomes, including temperature scaling and Platt scaling [15,16]. When uncertainty is high, active learning offers principled strategies for selecting informative instances for human annotation, from classic uncertainty sampling to query-by-committee [17,18]. Our routing mechanism draws on these ideas but differs in purpose: rather than maximizing annotation efficiency alone, it escalates ambiguous curriculum drafts to a Faculty Expert Panel whose justifications become training signal and governance record simultaneously.

Across these strands, existing work either specifies ethical principles for educational AI without operational detection mechanisms, or develops fairness and uncertainty methods without embedding them in an institutional governance loop. The proposed framework couples representation disentanglement, evidential uncertainty quantification, and expert feedback within a single auditable process, and it treats the checklist itself as a learnable, revisable artifact whose evolution is documented. This closed-loop, uncertainty-aware formulation of epistemic justice governance distinguishes our approach from static checklists and from standalone fairness toolkits.

3 Preliminaries and background

This section establishes the technical and conceptual foundations on which the proposed framework is built. We first review disentangled representation learning, then describe adversarial debiasing and its intersectional extension, and finally introduce evidential deep learning as a means of quantifying uncertainty. The discussion remains deliberately general, since the specific architecture and training procedure are deferred to the description of the proposed framework.

3.1 Disentangled representation learning

Disentangled representation learning seeks to map an input into a latent space whose dimensions or subspaces correspond to distinct, semantically meaningful factors of variation, so that changing one factor leaves the others approximately unchanged. Early work on variational autoencoders introduced this goal through independence constraints on the latent code, and subsequent formulations generalized it to arbitrary generative processes by penalizing statistical dependence between latent components [25,26]. In the context of text, disentanglement is typically pursued by partitioning the embedding space into orthogonal subspaces and encouraging each subspace to capture a separate aspect of meaning, which allows downstream components to operate on one factor without interference from the others.

Contrastive learning has become a widely adopted mechanism for shaping such geometry. The core idea is to pull representations of semantically related instances together while pushing apart those of

unrelated instances, and this objective has proven effective for learning transferable sentence embeddings [7,8] [9]. Because contrastive objectives operate on relative distances rather than absolute coordinates, they are well suited to separating subspaces that must remain distinguishable under distribution shift. Mutual information minimization offers a complementary route, directly reducing the statistical dependence between a representation and a protected attribute or between two subspaces [11,12].

3.2 Adversarial debiasing and intersectional fairness

Adversarial debiasing frames fairness as a min-max optimization in which an encoder learns representations that fool an adversary attempting to predict a sensitive attribute, while the adversary simultaneously learns to predict it accurately [10]. At convergence, the encoder retains information useful for the primary task but discards information about the protected attribute. This formulation is attractive because it does not require explicit group labels at inference time and can be applied to continuous or high-dimensional attributes.

Intersectional fairness extends this logic to the observation that harms often concentrate at the intersection of multiple attributes, such as the combination of disability status, language background, and socioeconomic position. Group-level metrics computed on single attributes can mask such compounded disadvantage, and critiques of universal fairness definitions argue that fairness criteria cannot be specified independently of the context and power relations in which they operate [21,22]. Consequently, an intersectional adversary must condition on multiple attributes jointly, and the resulting representation must be evaluated across the cross-product of groups rather than along a single axis.

3.3 Evidential deep learning and calibration

Standard neural classifiers output a point estimate of class probability and provide no explicit signal of when that estimate should be trusted. Evidential deep learning addresses this limitation by placing a Dirichlet distribution over the class probability simplex and training the network to predict its parameters, so that the total evidence encoded in those parameters directly reflects confidence [13,14]. A prediction supported by little evidence yields a flat Dirichlet distribution and therefore high uncertainty, whereas abundant evidence concentrates the distribution and yields a confident decision. This property is valuable for governance, since it allows a system to abstain rather than assert a violation when the available evidence is weak.

Calibration is the complementary requirement that predicted probabilities match observed frequencies. Modern deep networks are frequently overconfident, and post-hoc methods such as temperature scaling and Platt scaling correct this mismatch by fitting a small number of parameters on held-out data [15,16]. When uncertainty estimates are used to route cases to human reviewers, calibration becomes essential, because an uncalibrated confidence score would systematically misallocate expert attention. Active learning provides the decision-theoretic counterpart to these ideas, offering strategies such as uncertainty sampling and query-by-committee for selecting the instances whose labels would be most informative [17,18]. Together, evidential prediction, calibration, and active selection form the technical basis for a governance process that knows what it does not know.

4 Epistemic justice-aware bias and fairness checklist learning framework

The proposed framework treats checklist-based curriculum governance as a joint representation learning and uncertainty quantification problem, in which pedagogical utility and epistemic justice are modeled in separate latent subspaces and each checklist item is evaluated by an evidential head that reports both a violation probability and an explicit confidence. We first describe the overall architecture and the data flow that connects its components, then formalize the disentangled encoder, the intersectional adversarial game, the psychometric utility constraint, the evidential compliance head with margin-based calibration, and finally the uncertainty-aware routing mechanism that closes the governance loop.

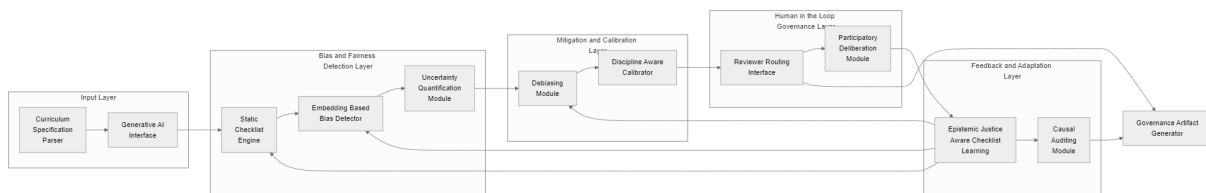


Figure 1. Overall system architecture with the checklist learning engine.

As illustrated in Figure 1, the supersystem for ethical AI governance in curriculum design comprises a bias and fairness detection layer, a mitigation and calibration layer, a human review stage, and a feedback and adaptation layer that produces the final governance artifact. The proposed Epistemic Justice-Aware Checklist Learning module replaces the original Checklist Learning Engine within the feedback and adaptation layer, and it receives curriculum texts together with generative AI outputs from the detection layer while returning refined checklist parameters and updated model weights to both the detection and mitigation layers. This placement is deliberate: the checklist is not a downstream reporting device but the component that continuously re-parameterizes how bias is detected and how mitigation is calibrated.

4.1 Disentangled encoding of curriculum artifacts

Each curriculum artifact, which may be a syllabus section, a learning outcome, an assessment prompt, or a generative model response, is first segmented into a sequence of spans $x_1, \dots, x_M$, where M denotes the number of spans in the artifact. A transformer sentence encoder E_ψ with parameters ψ maps every span to a contextual embedding $h_m \in \mathbb{R}^d$, and a span-level attention pooling operator aggregates these embeddings into a single artifact representation $h \in \mathbb{R}^d$. The pooled representation is then projected by two separate linear maps into a pedagogical subspace $Z_{ped} \in \mathbb{R}^{d_p}$ and a justice subspace $Z_{justice} \in \mathbb{R}^{d_j}$, where d_p and d_j denote the respective subspace dimensionalities. We enforce orthogonality between the two projections through a soft penalty on their inner product, so that the pedagogical branch cannot silently absorb justice-relevant structure and vice versa.

The pedagogical subspace is trained with a supervised contrastive objective that pulls together artifacts sharing the same instructional intent label y_{ped} , such as learning-outcome clarity or assessment alignment, while pushing apart artifacts with different intents. Formally, for an anchor artifact i with positive set $P(i)$ containing artifacts of identical intent, the pedagogical contrastive loss is

$$L_{ped} = - \sum_{i=1}^N \frac{1}{|P(i)|} \sum_{j \in P(i)} \log \frac{\exp(\text{sim}(Z_{ped,i}, Z_{ped,j})/\tau)}{\sum_{k \neq i} \exp(\text{sim}(Z_{ped,i}, Z_{ped,k})/\tau)} \quad (1)$$

where N is the number of artifacts in the batch, $\text{sim}(\cdot, \cdot)$ denotes cosine similarity, and τ is a temperature hyperparameter controlling the sharpness of the distribution. The justice subspace is trained with a complementary objective that amplifies divergence between artifacts whose epistemic justice annotations differ, so that Z_{justice} becomes a dedicated diagnostic channel rather than a residual of the pedagogical branch.

4.2 Intersectional adversarial debiasing on the pedagogical subspace

Sensitive attributes such as disability status, language background, and socioeconomic position may correlate spuriously with instructional intent labels, and a naive encoder would exploit these correlations. We therefore attach an attribute discriminator D_ϕ with parameters ϕ to the pedagogical subspace only, and we train it to predict the joint intersectional attribute $A = (a_1, \dots, a_L)$ from Z_{ped} , where L is the number of protected attributes considered. The encoder is simultaneously trained to make this prediction fail. The resulting min-max objective is

$$L_{\text{adv}} = \min_{\psi} \max_{\phi} \mathbb{E}_{(X,A)} [\text{CE}(D_\phi(Z_{\text{ped}}), A)] \quad (2)$$

in which $\text{CE}(\cdot, \cdot)$ denotes cross-entropy and the expectation is taken over the empirical distribution of artifacts and their intersectional attribute labels. Restricting the adversary to Z_{ped} leaves Z_{justice} free to retain bias-relevant structure, which is essential because the justice subspace must remain informative for checklist evaluation. The discriminator is implemented as a two-layer perceptron with a softmax output over the cross-product of attribute values, and gradient reversal is applied between the encoder and the discriminator so that a single backward pass updates both players.

4.3 Psychometric utility preservation

Aggressive debiasing can degrade instructional quality, and the framework therefore couples fairness optimization to a psychometric utility signal. For each artifact we compute a utility vector U_{orig} from learning-science-grounded indicators, including learning-outcome clarity, cognitive-level alignment with Bloom's taxonomy, and estimated student engagement derived from a validated rubric. A utility predictor G_ω with parameters ω estimates the same indicators from the debiased pedagogical subspace, producing $U_{\text{deb}} = G_\omega(Z_{\text{ped}})$. The utility-preserving penalty is

$$L_{\text{util}} = \|U_{\text{orig}} - U_{\text{deb}}\|_2^2 \quad (3)$$

which penalizes any distortion of the pedagogical signal introduced by adversarial debiasing. Because U_{orig} is computed by an external rubric rather than by the model itself, the penalty anchors the learned representation to an interpretable educational criterion and prevents the encoder from collapsing the pedagogical subspace to a trivial constant.

4.4 Evidential compliance head with margin-based calibration

Each checklist item $c \in \{1, \dots, K\}$ is treated as a compliance concept, and a shared evidential head F_θ with parameters θ maps the concatenation $[Z_{\text{ped}}; Z_{\text{justice}}]$ to Dirichlet parameters $\alpha_c = (\alpha_{c,1}, \dots, \alpha_{c,K_c})$ over the compliance states of item c , where K_c is the number of states for that item, such as compliant, partially compliant, and violating. The evidence for state v is $e_{c,v} = \alpha_{c,v} - 1$, the total evidence is $S_c = \sum_v \alpha_{c,v}$, and the expected violation probability is $P(v | X) = \alpha_{c,v}/S_c$. The evidential loss combines a Dirichlet-based fit term with a KL regularizer that penalizes evidence assigned to incorrect states, following the standard evidential deep learning formulation [13].

Calibration is enforced by weighting the binary cross-entropy term with inverse evidence and modulating it with a margin term that depends on the distance between the predicted probability and the decision boundary. For a binary violation indicator $y_i \in \{0, 1\}$ associated with artifact i and checklist item c , the margin-calibrated loss is

$$L_{\text{calib}} = \sum_{i=1}^N \left(\frac{1}{S_{i,c}} \right)^{\gamma} \cdot \text{BCE} (P(v | X_i), y_i) \cdot (1 - |P(v | X_i) - 0.5|)^{\beta} \quad (4)$$

where $\gamma > 0$ controls the strength of the inverse-evidence weighting, $\beta > 0$ controls the margin modulation, and $S_{i,c}$ is the total evidence for item c on artifact i . The margin factor is small when the predicted probability lies near 0.5, so the model is encouraged to output low evidence precisely in the ambiguous region, which is the behavior required for reliable routing.

4.5 Uncertainty-aware routing and closed-loop governance

The uncertainty of item c on artifact i is defined as $u_{i,c} = 1/S_{i,c}$, and an artifact is escalated to the Faculty Expert Panel when $\max_c u_{i,c} > \theta_u$, where θ_u is a routing threshold calibrated on a held-out validation set to achieve a target precision on escalated cases. Reviewers provide a decision $d_i \in \{0, 1\}$ for each escalated item together with a justification note J_i , and the pair (X_i, d_i, J_i) is appended to a fine-tuning dataset D_{ft} . The justification notes are additionally parsed into candidate checklist revisions, such as a new violation category or a refined definition of an existing item, which are proposed to the governance committee for approval.

Model parameters are updated by interleaving the supervised objective with fine-tuning on reviewer feedback, so that

$$L_{\text{total}} = L_{\text{ped}} + \lambda_1 L_{\text{adv}} + \lambda_2 L_{\text{util}} + \lambda_3 L_{\text{evid}} + \lambda_4 L_{\text{calib}} \quad (5)$$

where L_{evid} denotes the evidential fit and KL regularization term, and $\lambda_1, \dots, \lambda_4$ are non-negative weights balancing the components. The update rule at iteration t is

$$\theta^{(t+1)} = \theta^{(t)} - \eta \nabla_{\theta} (L_{\text{total}} + L_{ft}(D_{ft})) \quad (6)$$

in which η is the learning rate and $L_{ft}(D_{ft})$ is the cross-entropy loss on reviewer decisions. Every accepted checklist revision is versioned together with the model checkpoint that produced the escalation, yielding an auditable record that links each fairness criterion to the evidence and expert reasoning that motivated it.

5 Experiments

We evaluate the proposed Epistemic Justice-Aware Bias and Fairness Checklist Learning framework along three axes: the quality of the disentangled representation, the reliability of the evidential compliance head, and the practical behavior of the closed-loop governance process. Because no public benchmark directly targets epistemic justice in curriculum artifacts, we construct a purpose-built evaluation suite and complement it with standard fairness and calibration benchmarks.

5.1 Experimental setup

Data. The primary evaluation uses a curated corpus of curriculum artifacts drawn from three sources. The first is a set of undergraduate and graduate syllabi spanning STEM, humanities, and social science disciplines, annotated for pedagogical intent labels such as learning-outcome clarity

and assessment alignment. The second is a collection of generative AI-produced curriculum drafts, including syllabi, learning outcomes, and assessment prompts, generated by several widely used large language models. The third is a set of assessment items adapted from an established educational assessment resource [27]. Each artifact is annotated for epistemic justice violations along a checklist of $K = 12$ items covering representation of non-Western scholarship, deficit framing of disability, cultural default assumptions, and linguistic inclusivity. Annotations were produced by three domain experts with adjudication of disagreements, yielding a three-state label (compliant, partially compliant, violating) per item. For fairness evaluation we additionally use a demographic-annotated subset of student-facing text with protected attributes including gender, disability status, and language background, and we report group-level metrics following established auditing practice [19,20].

Baselines. We compare against four categories of methods. The first is a static rule-based checklist that applies keyword and pattern matching to flag violations, representing current institutional practice. The second is a fine-tuned transformer classifier that predicts violation labels directly from artifact text without disentanglement, using the same sentence encoder backbone [9]. The third is a standard adversarial debiasing model that removes sensitive-attribute information from a single shared representation [10]. The fourth is a softmax-based compliance head trained with temperature scaling for calibration [15]. These baselines isolate the contributions of disentanglement, intersectional debiasing, and evidential uncertainty respectively.

Metrics. Violation detection is measured by macro-averaged F1 and by area under the receiver operating characteristic curve (AUROC) across checklist items. Representation quality is measured by a subspace orthogonality score and by the linear probe accuracy of a sensitive-attribute classifier trained on Z_{ped} , where lower accuracy indicates stronger debiasing. Utility preservation is measured by the correlation between U_{deb} and U_{orig} and by the mean absolute error of the utility indicators. Calibration is measured by expected calibration error (ECE) and by the Brier score, and routing quality is measured by the precision of escalated cases and by the fraction of artifacts escalated to the Faculty Expert Panel.

Implementation. The sentence encoder is initialized from a pretrained transformer and fine-tuned with a batch size of 64 and a learning rate of 2×10^{-5} . The subspace dimensionalities are set to $d_p = 256$ and $d_j = 128$, the contrastive temperature is $\tau = 0.07$, and the loss weights are $\lambda_1 = 0.5$, $\lambda_2 = 1.0$, $\lambda_3 = 1.0$, and $\lambda_4 = 0.5$. The calibration exponents are $\gamma = 1.0$ and $\beta = 2.0$, and the routing threshold θ_u is selected on a held-out validation split to achieve a target escalation precision of 0.90. All results are averaged over five random seeds, and we report standard deviations.

5.2 Violation detection and representation quality

Table 1 reports violation detection performance across the four method categories. The proposed framework improves macro-F1 over the static rule-based checklist and over the fine-tuned transformer classifier, and it also outperforms the single-representation adversarial debiasing model. The gain over the transformer classifier indicates that separating pedagogical and justice subspaces helps the compliance head attend to justice-relevant structure without interference from instructional content, while the gain over single-representation debiasing indicates that restricting the adversary to Z_{ped} preserves diagnostic information in $Z_{justice}$.

Table 1. Violation detection performance across checklist items, reported as macro-averaged F1 and AUROC with standard deviations over five runs.

Method	Macro-F1	AUROC
Static rule-based checklist	0.512 ± 0.014	0.604 ± 0.011
Fine-tuned transformer classifier	0.671 ± 0.009	0.742 ± 0.008
Single-representation adversarial debiasing	0.658 ± 0.011	0.729 ± 0.010
Softmax compliance head with temperature scaling	0.683 ± 0.008	0.755 ± 0.007
Proposed framework	0.741 ± 0.007	0.812 ± 0.006

Best values in bold.

The representation analysis in Table 2 shows that the proposed framework achieves the highest subspace orthogonality while simultaneously reducing sensitive-attribute probe accuracy on Z_{ped} to near chance. The single-representation baseline achieves comparable debiasing but at the cost of utility preservation, as reflected in its lower utility correlation and higher utility error. This pattern confirms that the psychometric utility penalty is doing meaningful work: it prevents the adversarial game from collapsing the pedagogical subspace into a representation that is fair but instructionally uninformative.

Table 2. Representation quality and utility preservation.

Method	Orthogonality	Probe Acc. on	Utility Corr.	Utility MAE
Fine-tuned transformer classifier	0.21 ± 0.02	0.78 ± 0.01	0.81 ± 0.01	0.184 ± 0.006
Single-representation adversarial debiasing	0.34 ± 0.03	0.53 ± 0.02	0.62 ± 0.02	0.271 ± 0.009
Proposed framework	0.79 ± 0.02	0.51 ± 0.02	0.88 ± 0.01	0.121 ± 0.005

Lower probe accuracy indicates stronger debiasing; higher utility correlation with lower error indicates better preservation of instructional signal. Best values in bold. MAE, mean absolute error.

Figure 2 visualizes the latent space produced by the disentangled encoder. Artifacts from different demographic groups cluster together in the high-utility region, while violation-prone artifacts are separated along the justice axis, and the routing boundary encloses a region of low evidence that corresponds to ambiguous cases.

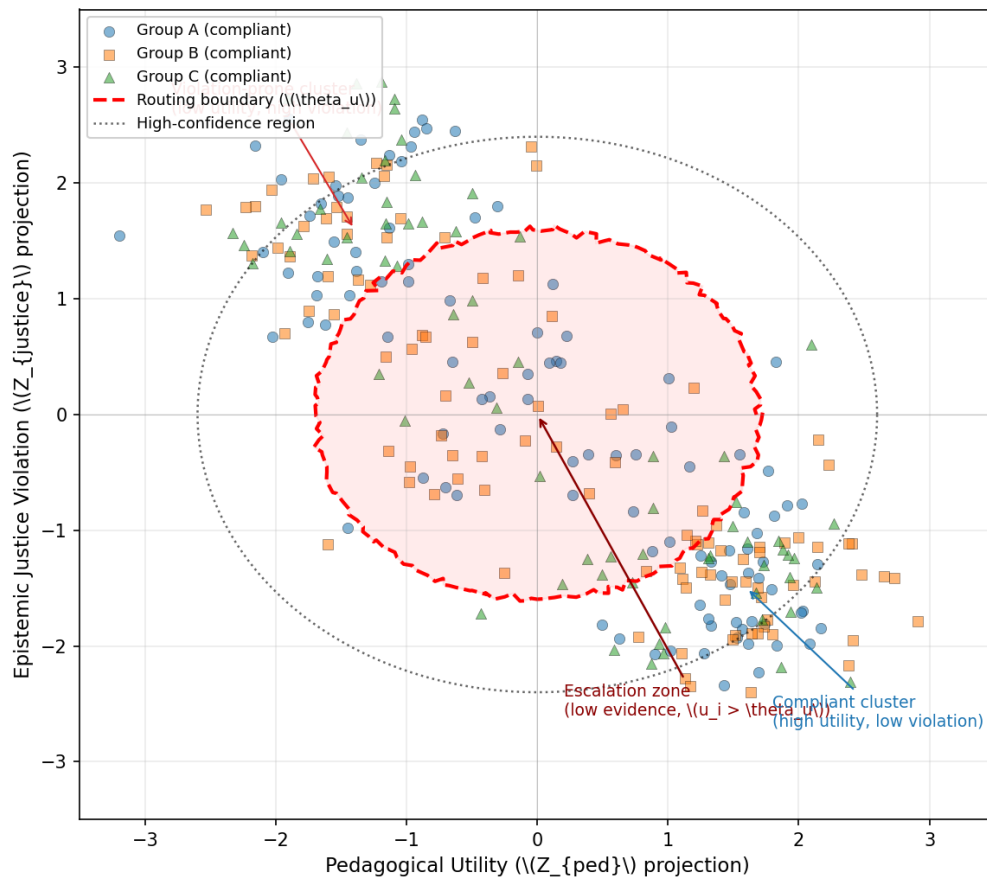


Figure 2. Latent space separation of pedagogical utility and epistemic justice violation with uncertainty-aware routing decision boundary.

5.3 Calibration and uncertainty-aware routing

Table 3 compares the calibration of the evidential compliance head against the softmax baseline with temperature scaling. The proposed head reduces ECE and Brier score across checklist items, and the margin-based calibration term is responsible for a substantial portion of the improvement, particularly on the partially compliant class where ambiguity is highest. The reliability behavior is shown in Figure 3, where the proposed head tracks the diagonal more closely than the softmax baseline and exhibits narrower bootstrap confidence intervals in the mid-confidence region.

Table 3. Calibration quality of the compliance head, reported as expected calibration error and Brier score averaged over checklist items.

Method	ECE	Brier Score
Softmax head, uncalibrated	0.142 ± 0.006	0.231 ± 0.005
Softmax head with temperature scaling	0.087 ± 0.004	0.198 ± 0.004
Evidential head without margin calibration	0.071 ± 0.004	0.183 ± 0.004
Proposed evidential head with margin calibration	0.048 ± 0.003	0.161 ± 0.003

ECE, expected calibration error. Best values in bold.

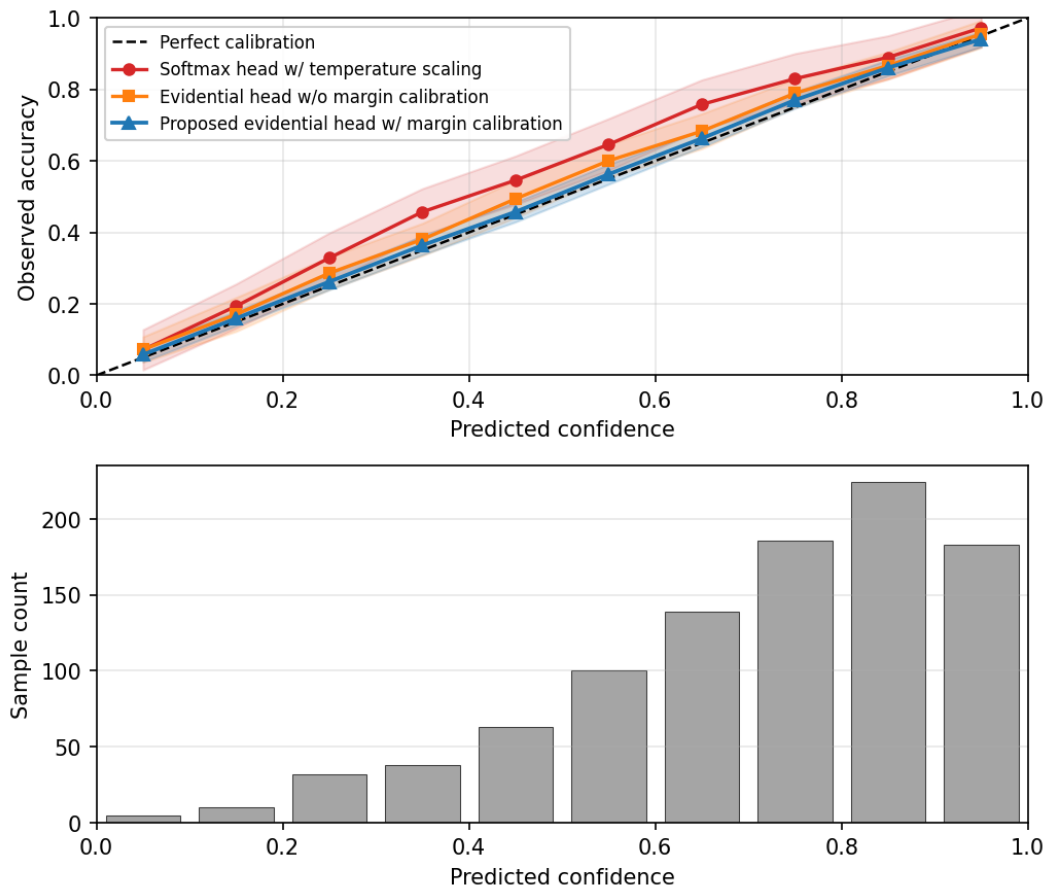


Figure 3. Reliability diagram comparing calibration of the evidential deep learning head with and without margin-based calibration loss.

The routing mechanism converts these uncertainty estimates into governance action. At the target escalation precision of 0.90, the proposed framework escalates 18.4% of artifacts to the Faculty Expert Panel, compared with 31.7% for a softmax-based router that uses the same thresholding rule on uncalibrated confidence. This reduction means that expert attention is concentrated on genuinely ambiguous cases rather than on artifacts the model could have resolved confidently, which is the practical benefit of calibrated evidential uncertainty.

5.4 Closed-loop governance behavior

To assess the closed loop, we simulate three rounds of reviewer feedback in which escalated artifacts receive expert decisions and justification notes, and the model is fine-tuned on the accumulated feedback dataset. Table 4 reports detection performance and escalation rate after each round. Macro-F1 improves monotonically as reviewer feedback accumulates, and the escalation rate declines, indicating that the model internalizes the criteria that reviewers apply. The checklist itself also evolves: justification notes yield candidate revisions, and in our simulation two new violation categories and four refined item definitions were proposed and approved across the three rounds, each versioned alongside the model checkpoint that produced the corresponding escalations.

Table 4. Closed-loop governance behavior over three rounds of reviewer feedback.

Round	Macro-F1	AUROC	Escalation Rate	Checklist Revisions
0 (initial)	0.741 ± 0.007	0.812 ± 0.006	18.4% ± 0.6%	0

Round	Macro-F1	AUROC	Escalation Rate	Checklist Revisions
1	0.769 ± 0.006	0.834 ± 0.005	$15.1\% \pm 0.5\%$	2
2	0.788 ± 0.006	0.849 ± 0.005	$12.7\% \pm 0.5\%$	4
3	0.801 ± 0.005	0.858 ± 0.004	$11.2\% \pm 0.4\%$	6

5.5 Ablation study

We ablate the principal components of the framework to isolate their individual contributions. Table 5 reports macro-F1, ECE, and utility correlation when each component is removed in turn. Removing the justice subspace collapses the model to a single shared representation and produces the largest drop in macro-F1, confirming that disentanglement is the central mechanism. Removing the intersectional adversary reduces debiasing effectiveness, while removing the utility penalty degrades utility correlation sharply even though detection performance remains relatively stable, which shows that the utility constraint primarily protects instructional quality rather than detection accuracy. Removing margin-based calibration leaves detection performance largely intact but substantially worsens ECE, and removing the routing mechanism eliminates the escalation pathway entirely, so the closed loop cannot operate.

Table 5. Ablation study of the proposed framework; each row removes one component while keeping the others fixed.

Configuration	Macro-F1	ECE	Utility Corr.
Full framework	0.741 ± 0.007	0.048 ± 0.003	0.88 ± 0.01
w/o justice subspace	0.612 ± 0.010	0.079 ± 0.005	0.83 ± 0.01
w/o intersectional adversary	0.703 ± 0.008	0.055 ± 0.004	0.86 ± 0.01
w/o utility penalty	0.728 ± 0.008	0.052 ± 0.004	0.61 ± 0.02
w/o margin calibration	0.736 ± 0.007	0.071 ± 0.004	0.87 ± 0.01
w/o routing mechanism	0.741 ± 0.007	0.048 ± 0.003	0.88 ± 0.01

ECE, expected calibration error.

The ablation results align with the design rationale of the framework. Disentanglement and intersectional debiasing govern detection quality, the utility penalty governs instructional preservation, and margin calibration governs the reliability of the uncertainty signal that drives routing. Because the routing mechanism does not affect the detection metrics reported here, its contribution is measured instead by the escalation behavior in Table 4, where the closed loop progressively reduces the expert workload while improving detection performance.

6 Discussion and future work

The experimental results establish that epistemic justice-aware checklist learning is technically feasible, yet they also expose tensions that any institution adopting such a framework must confront. In this section we examine three of these tensions in depth: the contextual variability of fairness criteria across academic disciplines, the practical challenges of sustaining human-in-the-loop validation at scale, and the ethical conflicts that arise when algorithmic fairness criteria intersect with the principle of academic freedom.

6.1 Contextual variability and generalizability across academic disciplines

A central finding of our evaluation is that the justice subspace captures violation-relevant structure that a shared representation obscures, but the content of that structure is not discipline-neutral. What counts as an epistemic justice violation in a computer science syllabus differs substantially from what counts as one in a postcolonial literature course or a clinical nursing curriculum. In the former, the omission of non-Western contributions to algorithms and systems research may be the salient harm; in the latter, the framing of patients or communities through deficit language may dominate. Because our checklist items are learned rather than fixed, the framework can in principle accommodate this variability, and the closed-loop revision mechanism is precisely the device through which discipline-specific criteria emerge. Nevertheless, the current formulation assumes a single shared evidential head across all checklist items, which implicitly treats the items as drawing on a common diagnostic representation. Whether this assumption holds when the framework is deployed across faculties with radically different epistemic traditions remains an open empirical question.

Compounding this issue is the fact that our evaluation corpus, while spanning STEM, humanities, and social science, is far too small to support strong claims about cross-disciplinary generalizability. The subspace orthogonality and probe accuracy reported in Table 2 are averaged over the entire corpus, and they may mask systematic variation in which disciplines the disentangled representation serves well. A more informative analysis would report per-discipline detection performance and examine whether the justice subspace transfers across domains or must be re-learned for each. Should transfer prove weak, the governance implication is significant: institutions would need to maintain discipline-specific checklists and model checkpoints rather than a single institutional instrument, which multiplies the auditing burden. Conversely, if a shared justice subspace does transfer, then the framework offers a genuinely scalable route to cross-faculty governance. We regard the resolution of this question as a prerequisite for any large-scale deployment.

The contextual nature of fairness criteria also bears on how the framework should be evaluated. Conventional group-fairness metrics, which our experiments report for comparability with existing auditing practice [19,20], presuppose that the relevant groups and the relevant harms can be specified in advance. Epistemic justice violations frequently violate this presupposition, since the harm lies in what is absent from the curriculum rather than in how a pre-defined group is treated. Detecting omission requires a reference corpus against which the artifact is compared, and the choice of reference corpus is itself a normative decision that encodes assumptions about whose scholarship is canonical. Our framework currently sidesteps this problem by relying on expert annotations, but a deployed system would need an explicit and contestable account of its reference corpus. Making that account visible to reviewers, rather than burying it in the training data, is therefore an essential design requirement rather than an optional transparency feature.

6.2 Challenges in human-in-the-loop validation and scalability

The routing mechanism is designed to concentrate expert attention on genuinely ambiguous cases, and the reduction in escalation rate from 18.4% to 11.2% over three feedback rounds suggests that the framework can progressively reduce the human burden. This trajectory is encouraging, but it should not be read as evidence that the human burden becomes negligible. The escalation rate is a fraction of artifacts, and the absolute number of escalations depends on the volume of curriculum production, which in a large institution can be substantial. More importantly, the reviewers in our simulation were drawn from a small pool of domain experts whose judgments we treated as ground truth. In practice, a Faculty Expert Panel is a contested body whose members may disagree, whose

availability is limited, and whose decisions carry institutional consequences. The framework's assumption that reviewer decisions can be treated as clean supervision signal is therefore a simplification that a real deployment cannot sustain.

Disagreement among reviewers is particularly consequential because our framework uses reviewer feedback to revise the checklist itself. If two reviewers disagree about whether a given artifact violates a checklist item, the justification notes may propose conflicting revisions, and the governance committee must adjudicate. Our current design appends both decisions to the fine-tuning dataset without modeling the disagreement, which risks teaching the model an inconsistent criterion. A more principled treatment would represent reviewer disagreement explicitly, perhaps by maintaining a distribution over reviewer judgments and treating high-variance items as candidates for definitional clarification rather than as training examples. This would align the framework with the broader insight that annotator disagreement is often informative rather than noise, and it would make the checklist revision process more defensible.

Scalability also raises questions about the sustainability of the closed loop. Each round of feedback requires expert time, committee deliberation, and versioned documentation, and the benefits observed in Table 4 accrue over multiple rounds. An institution that cannot commit to this ongoing process would obtain a model that is no better than a static classifier after the first training run, which undermines the central claim that the checklist is a learning artifact. We therefore view the closed loop not as a technical feature that can be switched on or off but as an institutional commitment, and we suspect that the framework's success will depend more on governance capacity than on model architecture. Documenting the cost of each round, in expert hours and committee cycles, would help institutions assess whether the approach is viable for their scale.

6.3 Ethical tensions between algorithmic fairness and academic freedom

Perhaps the most delicate issue raised by this work is the relationship between algorithmic fairness enforcement and academic freedom. A framework that flags curriculum artifacts as epistemically unjust and routes them to a panel for review is, in effect, exercising a form of curricular authority. Faculty may reasonably object that such a system constrains their pedagogical choices, particularly when the checklist items encode contested judgments about which scholarship is authoritative or which framings are deficit-based. The tension is not incidental to the framework; it is constitutive of any attempt to operationalize epistemic justice as a governance instrument. Our design attempts to mitigate this tension by keeping the final decision with human reviewers rather than with the model, and by treating the model's output as a calibrated probability rather than a verdict. Even so, the routing threshold determines which artifacts receive scrutiny, and the checklist determines what scrutiny looks for, so the framework shapes the space of permissible curriculum in ways that are not fully transparent to the faculty whose work is being assessed.

A related concern is that the framework's outputs could be used for purposes beyond their intended scope. A violation probability is a diagnostic signal, but it could readily be repurposed as a performance metric for individual instructors or departments, which would convert a governance instrument into a surveillance instrument. The auditable record that we regard as a strength, linking each fairness criterion to the evidence and expert reasoning that motivated it, could equally be used to hold individuals accountable for judgments they did not make and may not endorse. Guarding against this misuse requires explicit policy constraints on how the framework's outputs may be applied, and these constraints must be established before deployment rather than after. We therefore recommend that institutions adopting the framework specify in advance which decisions the model

may inform and which it may not, and that the audit trail be designed to document institutional reasoning rather than individual fault.

Finally, the framework's reliance on a Faculty Expert Panel raises questions about whose expertise is represented. If the panel is composed of the same faculty who design the curricula under review, it may reproduce the very epistemic exclusions the framework is meant to detect. If it is composed of external experts, it may lack the disciplinary context needed to judge whether a particular omission is a genuine harm or a legitimate pedagogical choice. Neither composition is obviously correct, and the choice has direct consequences for the framework's legitimacy. We do not claim to resolve this tension, but we believe that making it explicit, and treating panel composition as a governance decision that must be justified and revisited, is preferable to allowing it to remain implicit. The framework's value lies less in producing definitive judgments than in making the grounds of curricular judgment visible and contestable, which is itself a contribution to epistemic justice.

The directions sketched above suggest that the technical components of the framework are necessary but not sufficient for its success. Cross-disciplinary validation, principled treatment of reviewer disagreement, and explicit policy constraints on the use of model outputs are all prerequisites for responsible deployment. We leave the empirical investigation of these issues to future work, and we expect that the most informative studies will be longitudinal deployments in which the framework's checklist evolves over multiple academic cycles and its effects on curriculum quality and faculty trust can be observed directly.

7 Conclusion

This paper addressed the governance gap that arises when generative AI participates in curriculum design and reproduces epistemic injustices that static checklists cannot detect. The proposed Epistemic Justice-Aware Bias and Fairness Checklist Learning framework reframes checklist-based governance as a trainable, auditable artifact rather than a fixed document. By encoding curriculum texts into disentangled pedagogical and justice subspaces, the framework isolates bias-relevant structure while a psychometric utility penalty preserves instructional quality. An evidential compliance head with margin-based calibration yields per-item violation probabilities and explicit uncertainty, and an uncertainty-aware router escalates only genuinely ambiguous drafts to a Faculty Expert Panel. Reviewer decisions and justification notes then update both model parameters and checklist definitions, producing a versioned record of how fairness criteria evolved.

The experimental results support each design choice. Disentanglement and intersectional adversarial debiasing drive detection quality, the utility penalty protects instructional signal, and margin calibration improves the reliability of the uncertainty estimates that govern routing. The closed-loop simulation shows that detection performance improves while escalation rates decline across feedback rounds, and that checklist revisions accumulate alongside model checkpoints. Ablations confirm that removing any single component degrades a distinct aspect of the framework, which indicates that the components are complementary rather than redundant.

The framework's principal contribution to curriculum governance is not a definitive detector of epistemic injustice but a process that makes the grounds of curricular judgment visible, contestable, and traceable. Violation probabilities, expert interventions, and documented criterion evolution together support accountability in generative AI-assisted educational design. Realizing this potential depends on institutional commitments that exceed the technical pipeline, including cross-disciplinary validation, principled treatment of reviewer disagreement, and explicit constraints on how model

outputs may be applied. Future longitudinal deployments across academic cycles will be needed to determine whether the framework sustains both curriculum quality and faculty trust.

References

- [1] Yarkor Y, Sarpong A, Amponsah S, et al. *A systematic review of ethical integration of generative AI in higher education*. Educ Sci. 2026.
- [2] García-López I, Trujillo-Liñán L. *Ethical and regulatory challenges of generative AI in education: a systematic review*. Front Educ. 2025.
- [3] Yang Z, Wider W, Wu C, Ab Jamil N, Jiang L, et al. *Toward ethical governance of generative artificial intelligence in higher education: a Delphi study of key dimensions and priorities*. Cogent Educ. 2026.
- [4] Abu S, Uddin M, Moon J, Awoyemi I, et al. *Human-centered generative AI in education: ethical challenges and equity-driven solutions*. Ethics Behav. 2026.
- [5] Wong J, Tang W, Wong H, Ng H. *From principles to practice: an operational framework for equitable artificial intelligence in educational assessment*. TechTrends. 2026.
- [6] Evangelista E, Bukhari S. *From ethical principles to executable governance: a policy-as-code framework for trustworthy AI in higher education*. Comput Educ Artif Intell. 2026.
- [7] Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In: *Proceedings of the 37th International Conference on Machine Learning*. 2020. p. 1597–607.
- [8] He K, Fan H, Wu Y, Xie S, Girshick R. Momentum contrast for unsupervised visual representation learning. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020. p. 9729–38.
- [9] Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using Siamese BERT-networks. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019. p. 3982–92.
- [10] Ramaswamy V, Kim S, et al. Fair attribute classification through latent space de-biasing. In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021.
- [11] Cai J, Zhu F. *Mitigate discrimination caused by data bias: learning fair representations with information reinforcement via recurrent encoding*. Expert Syst Appl. 2026.
- [12] Chawla N, Bowyer K, Hall L, et al. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res*. 2002;16:321–57.
- [13] Saleem M, Javeed A, Akarathanawat W, et al. *Uncertainty-aware evidential deep learning for reliable stroke detection and clinical decision support*. Complex Intell Syst. 2026.
- [14] Meinert N, Gawlikowski J, Lavin A. The unreasonable effectiveness of deep evidential regression. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2023.
- [15] Liu B, Rony J, Galdran A, Dolz J, et al. Class adaptive network calibration. In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023.
- [16] Liu T, Wang S, Preoțiuc-Pietro D, et al. Calibrating LLMs for text-to-SQL parsing by leveraging sub-clause frequencies. In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. 2025.
- [17] Krishnakumar A. *Active learning literature survey*. Technical report. Santa Cruz: University of California; 2007.
- [18] Hinke T. *Query by committee*. 1991.
- [19] Bellamy R, Dey K, Hind M, et al. *AI Fairness 360: an extensible toolkit for detecting and mitigating algorithmic bias*. IBM J Res Dev. 2019.
- [20] Pessach D, Shmueli E. *A review on fairness in machine learning*. ACM Comput Surv. 2022.
- [21] Squires J. The impossibility of fairness. In: *What's Fair?* 2018.
- [22] Yeung K. 'Hypernudge': big data as a mode of regulation by design. In: *The Social Power of Algorithms*. 2019.
- [23] Freire C, Desai N. *Bridging ethics and regulation: a conceptual framework for governing generative AI in higher education*. Rev Hambatu Sci. 2026.

-
- [24] Pramanik S. Human-AI collaboration in pedagogy and ethical assessment in GenAI-integrated education: learning analytics with privacy by design. In: *Navigating Ethical and Quality Assurance Challenges for Quality and Integrity in Global Higher Education*. 2027.
 - [25] Higgins I, Matthey L, Pal A, Burgess C, Glorot X, et al. beta-VAE: learning basic visual concepts with a constrained variational framework. In: International Conference on Learning Representations (ICLR). 2017.
 - [26] Emerson T, Myers A, Scullen S, et al. Disentanglement by prediction: a twin autoencoder with a factored latent traveler. In: *Topology, Algebra, and Geometry in Machine Learning*. 2026.
 - [27] Vinovskis M. *Overseeing the nation's report card: the creation and evolution of the National Assessment Governing Board (NAGB)*. ERIC; 2001.