



The Adaptive Opportunity Floor: Distributionally Robust Safeguards for Inclusive AI Tutoring

Kwan Hong TAN

Singapore University of Social Sciences, Singapore

DOI: <https://doi.org/10.68050/JAMS.2026.768>

Received: 02 October 2026; Accepted: 12 October 2026; Published: 20 October 2026

ABSTRACT

Purpose: Generative artificial intelligence (AI) tutors can personalize learning, but improved average performance cannot establish equitable benefit for learners facing accessibility, language, connectivity, or time constraints. **Methodology:** This theoretical design-science study develops the Adaptive Opportunity Floor (AOF), combining causal potential-outcome comparisons, subgroup distributional robustness, accessible instructional alternatives, and educator-controlled fallback. We derive a conditional lower bound on subgroup benefit and distinguish direct empirical Wasserstein uncertainty from a separately calibrated population-shift model. **Results:** A formal bound and a hypothetical worked example demonstrate how average-gain optimization can reject a subgroup-safe policy. A finite-sample extension makes sampling uncertainty, candidate-policy selection, and probabilistic assumptions explicit. A comparative synthesis distinguishes AOF from predictive parity, Universal Design for Learning, conventional causal policy learning, and generic robust optimization. **Implications:** Institutions can predeclare accessible comparators, absolute attainment minima, subgroup margins, and evidence thresholds before deploying AI tutoring. The proposed safeguards are sufficient only under identified causal effects, defensible error and shift bounds, and valid outcome measurement. No learner data were collected and empirical effectiveness has not been demonstrated.

Keywords: inclusive education; AI tutoring; adaptive learning; distributionally robust optimization; algorithmic fairness; learning outcomes

1. INTRODUCTION

Artificial intelligence is creating new possibilities for differentiated teaching, formative feedback, and personalized learning support. The educational rationale is older than contemporary generative AI: individually responsive instruction has long been regarded as a potentially powerful alternative to one-size-fits-all teaching (Bloom, 1984). Recent large language models extend the set of possible explanations, conversational supports, and feedback pathways, but they also introduce errors, dependence risks, interface barriers, and uneven access (Kasneci et al., 2023; Miao & Holmes, 2023). Accordingly, higher average course performance cannot by itself establish that an AI tutor is educationally inclusive.

The distributional problem is practical. A learner who can interact with a conversational system in a preferred language over reliable broadband may receive more usable instructional support than a learner relying on a shared device, assistive technology, low bandwidth, or short intermittent study sessions. Even when a model's answers are nominally identical in quality, the effective opportunity to learn differs across settings. Algorithmic bias can arise at several stages of educational system development and deployment, including the measurement of learning, the selection of training data, and the consequences of automated decisions (Baker & Hawn, 2022). CAST's Universal Design for Learning (UDL) Guidelines 3.0 emphasize the removal of exclusionary learning barriers and the design of meaningful alternatives (CAST, 2024). These are design objectives, not automatic consequences of optimization.

Prior work by Tan (2025) argues for reframing business education around human-AI capability development. Tan (2026a) highlights the educational value of structured reflective dialogue, while Tan (2026b) places human interpretive and contestatory agency at the centre of algorithmic governance. Tan (2026c) shows the



relevance of embedding human oversight and governance into AI-enabled workflows rather than treating control as an afterthought. Together these contributions motivate a more precise question for adaptive instruction: what should a university require before replacing an accessible existing pathway with an AI-mediated one? They do not themselves answer the subgroup safety question addressed here.

This article proposes the **Adaptive Opportunity Floor (AOF)**: for each prespecified educational access stratum, an AI policy should meet a conservative lower bound on its learning benefit relative to a feasible baseline pathway, allowing only a publicly justified non-inferiority margin. The floor is a policy-selection safeguard, not an assertion that each learner will improve. The paper contributes: (i) a counterfactual target tied to delayed, independent learning, (ii) a conditional distributional safety certificate, (iii) an explicit governance design with accessible alternatives, and (iv) a research protocol capable of challenging and potentially falsifying the proposed approach. Unlike approaches centred on verifying AI outputs, the AOF concerns equitable intervention choice and the distribution of educational benefit.

2. LITERATURE REVIEW AND CONCEPTUAL GAP

2.1. From personalization to educational opportunity

A review of AI applications in higher education found substantial attention to automated assessment, profiling, and adaptive systems, alongside limited inclusion of educators' perspectives in the reviewed literature (Zawacki-Richter et al., 2019). Cukurova (2025) argues that the value of AI in education is not restricted to outsourcing cognition, but also concerns hybrid intelligence and the development of human capabilities. The educational outcome of interest should therefore be an independently demonstrable capability, not merely a fluent AI-assisted response. In applied adult education, the structured reflective dialogue intervention reported by Tan (2026a) illustrates the importance of explanation, challenge, and reflection as pedagogical processes, although that study does not test the framework introduced here.

2.2. Why conventional fairness metrics are insufficient

Predictive fairness can ask whether an algorithm's error rates are similar between groups. Yet an adaptive tutor is not simply a prediction model: its recommendation changes instructional exposure and thus potentially changes outcomes. A common prediction metric can remain stable while learners experience different accessibility burdens or educational gains. Baker and Hawn (2022) document the variety of bias mechanisms in educational algorithms. Selbst et al. (2019) warn that abstract fairness criteria can obscure institutional relationships and the social setting of technical decisions. Accordingly, equality in recommendation scores or interaction rates does not imply equality in the opportunity to learn.

The more actionable question is whether the proposed adaptive pathway would yield an adequate learning outcome compared with the alternative that would actually be available to a learner. AOF makes this counterfactual baseline explicit. Its objective is not identical outcomes for all learners, which would overlook legitimate differences in preparation and support needs. Rather, its primary guardrail is avoidance of unjustified educational loss for prespecified strata; disparity analysis supplies a complementary diagnostic.

2.3. Inclusive design and the limits of technical certification

UDL emphasizes multiple means of engagement, representation, and action or expression (CAST, 2024). This suggests that an accessible non-AI pathway, a tutor-guided alternative, or a low-bandwidth format may be a necessary component of a responsible intervention. UNESCO's guidance recommends human-centred pedagogical design, safeguards, and governance in educational generative AI (Miao & Holmes, 2023). Tan's (2013) competency mapping report also illustrates why differentiated instructional needs should be identified before an intervention is selected, although that case is not evidence for the proposed AOF safety bound.

Distributionally robust optimization (DRO) offers a technical route for acknowledging distribution shift. Wasserstein ambiguity sets and their tractable formulations are established mathematics (Mohajerin Esfahani & Kuhn, 2018). Causal policy learning similarly evaluates intervention allocation under substantive constraints



(Athey & Wager, 2021). Neither body of work, by itself, specifies the educationally accessible comparator, subgroup delayed-learning non-inferiority requirement, policy-version governance, or humane fallback required here. AOF is an application-level synthesis with a distinct educational decision object, not a claim to invent DRO or causal inference. No mathematical certificate alone can repair inaccessible interfaces, unreliable assessments, or faulty identification.

Table 1. Comparison of established approaches and the distinct educational decision addressed by AOF

Approach	Principal object	Important limitation	AOF contribution
Predictive fairness audits (Baker & Hawn, 2022)	Parity of model errors, classifications, or scores	Do not alone identify causal learning effects or accessible counterfactual pathways	Subgroup lower bound on learning gain relative to feasible instruction
Universal Design for Learning (CAST, 2024)	Accessible choices in engagement, representation, and expression	Does not itself supply a statistical guarantee of incremental learning effects	Preserves accessibility options and educator fallback alongside quantitative checks
Causal policy learning (Athey & Wager, 2021)	Selection of interventions using identifiable welfare or treatment effects	Does not itself prescribe an absolute or subgroup-specific educational opportunity threshold	Makes a minimum subgroup increment and comparator governance explicit
Wasserstein DRO (Mohajerin Esfahani & Kuhn, 2018)	Conservative optimization under probabilistic distribution shift	A mathematical uncertainty set is not itself an educational outcome or fair baseline	Connects subgroup robustness to delayed learning, non-inferiority, and instructional choice
Adaptive Opportunity Floor (this article)	Conditional robust minimum subgroup gain and defensible educational alternative	Requires valid identification, credible uncertainty bounds, and independent validation	Operational safeguard, not an unconditional proof of inclusion or learner-level safety

3. METHODOLOGY AND DESIGN APPROACH

3.1. Research design and epistemic status

This article uses design science and analytical model development. It defines the educational decision, derives a conditional result, tests internal coherence using a transparent hypothetical worked example, and proposes a future empirical validation protocol. It does not report a recruitment exercise, observed trial outcomes, simulations fitted to real learners, or institutional ethics approval. The mathematical result is an implication of its assumptions, not a demonstrated field effect. This separation protects the distinction between conceptual feasibility and empirical evidence.

3.2. Learning pathway, baseline, and outcomes

Consider learners with measured baseline covariates $X \in \mathcal{X}$ and prespecified, responsibly collected educational access strata $G \in \{1, \dots, K\}$. Covariates might include prior knowledge, device access, preferred response modality, and session constraints. Group definitions are reviewed with educational stakeholders and are used for auditing; they are not proxies for ability or a mandate to restrict opportunities. Let π denote a policy assigning an instructional pathway, and let π_0 denote a feasible, accessible reference pathway that remains genuinely available, such as instructor-mediated support with offline materials.

The primary outcome $Y(\pi)$ is a prespecified, normalized score on a delayed assessment completed without assistance from the evaluated tutor, adjusted only through a declared scoring protocol that preserves comparability across accessible modalities. It must measure a curriculum construct rather than typing speed or



access to a particular device. Outcomes are bounded between zero and one; high-stakes certification decisions require separate validation. A policy's expected group outcome is

$$V_g(\pi) = \mathbb{E}[Y(\pi) | G = g], \quad g \in \{1, \dots, K\}. \quad (1)$$

The conditional incremental benefit over the baseline is

$$h_{\pi,g}(x) = \mathbb{E}[Y(\pi) - Y(\pi_0) | X = x, G = g]. \quad (2)$$

For a target distribution Q of learner characteristics in stratum g , the causal group-average benefit is

$$\Delta_g(\pi; Q) = \int_{\mathcal{X}} h_{\pi,g}(x) dQ(x). \quad (3)$$

These potential outcomes concern an educational intervention defined consistently enough for treatment comparison. In particular, changing the tutor model, prompt template, teacher facilitation, or allowed accessibility mode changes the intervention and must be documented. The group average is not a guarantee about any individual learner.

3.3. Distributional uncertainty and model error

Let $\hat{P}_g$ be a held-out reference distribution for the covariates of stratum g , equipped with an educationally justified metric d and finite first moments. The distance should be defined on transformed, commensurable covariates; arbitrary scaling can make a Wasserstein radius meaningless. Define the uncertainty set

$$\mathcal{U}_g(\rho_g) = \{Q: W_1(Q, \hat{P}_g) \leq \rho_g\}, \quad \rho_g \geq 0. \quad (4)$$

Here W_1 denotes the first Wasserstein distance:

$$W_1(P, Q) = \inf_{\gamma \in \Gamma(P, Q)} \int_{\mathcal{X} \times \mathcal{X}} d(x, x') d\gamma(x, x'). \quad (5)$$

Suppose $\hat{h}_{\pi,g}$ estimates the conditional policy benefit, is $L_{\pi,g}$ -Lipschitz with respect to d , and has a group-specific uniform modeling error bound $\eta_g \geq 0$ over the admissible support. More precisely, the condition is

$$\sup_{x \in \mathcal{X}} |h_{\pi,g}(x) - \hat{h}_{\pi,g}(x)| \leq \eta_g \quad (6)$$

Such a uniform error bound is demanding and must be estimated or conservatively justified; it is not obtained merely because a prediction model fits a dataset. Likewise, a radius ρ_g requires a shift model, independent validation data, and sensitivity analysis. The framework does not certify abrupt changes to the conditional causal response itself, including those arising from a major language-model update.

3.4. The Adaptive Opportunity Floor

Define the observable-model reference mean $\hat{m}_g(\pi)$ and its conservative opportunity floor $F_g(\pi)$ by

$$\hat{m}_g(\pi) = \int_{\mathcal{X}} \hat{h}_{\pi,g}(x) d\hat{P}_g(x). \quad (7)$$

When $\hat{P}_g$ is the empirical distribution of an independent reference sample of size n_g , the corresponding plug-in estimator is

$$\hat{m}_g(\pi) = \frac{1}{n_g} \sum_{i=1}^{n_g} \hat{h}_{\pi,g}(X_i). \quad (8)$$

$$F_g(\pi) = \hat{m}_g(\pi) - L_{\pi,g}\rho_g - \eta_g. \quad (9)$$

For a predeclared maximum tolerable decrement $\delta_g \geq 0$, the policy is *AOF-certified under the stated assumptions* only if

$$F_g(\pi) \geq -\delta_g \quad \text{for every audited stratum } g. \quad (10)$$

For a low-stakes pilot, a zero-decrement margin ($\delta_g = 0$) is a reasonable default when supported by precision and substantive considerations. Any nonzero margin requires a prespecified educational rationale and governance review, rather than retroactive selection to rescue a desired policy. A comparator must be a real, accessible pathway with comparable instructional objectives, time, support, and scoring accommodations. The primary certificate is relative non-harm, not absolute educational adequacy. A baseline can be accessible yet pedagogically ineffective; therefore a separate minimum attainment requirement is desirable when the baseline falls below a defensible curricular threshold. Neither group-average statement protects every individual learner.

Proposition 1 (conditional subgroup lower bound). Assume (i) the target group's covariate distribution Q belongs to $\mathcal{U}_g(\rho_g)$, (ii) the estimated conditional benefit $\hat{h}_{\pi,g}$ is $L_{\pi,g}$ -Lipschitz, and (iii) the stated uniform modeling error bound holds. Then

$$\Delta_g(\pi; Q) \geq \hat{m}_g(\pi) - L_{\pi,g}\rho_g - \eta_g = F_g(\pi). \quad (11)$$

Proof. The uniform error assumption yields $\mathbb{E}_Q h_{\pi,g} \geq \mathbb{E}_Q \hat{h}_{\pi,g} - \eta_g$. By Kantorovich-Rubinstein duality, a Lipschitz function has expectation difference bounded by its Lipschitz constant times W_1 . Thus $\mathbb{E}_Q \hat{h}_{\pi,g} \geq \mathbb{E}_{\hat{P}_g} \hat{h}_{\pi,g} - L_{\pi,g}W_1(Q, \hat{P}_g) \geq \hat{m}_g - L_{\pi,g}\rho_g$. Combining the inequalities gives the claimed bound. $\square$

Proposition 1 gives a sufficient deterministic statement conditional on its assumptions, not a necessary test. In Equation (4), the direct radius is measured from the empirical reference distribution and must cover the entire difference to the deployment distribution, including sampling discrepancy. A direct bound of this kind already accounts for that discrepancy and should not be combined with a second sampling correction for the same source of uncertainty. A failed certificate means insufficient evidence under the chosen assumptions, not proof of educational harm. Without statistical coverage for the radius and model error, the quantity F_g is a scenario-based lower bound rather than a confidence bound. Section 3.7 provides an alternative population-referenced construction with an explicit finite-sample correction.

3.5. Constrained educational policy selection

Let $C(\pi)$ be a prospectively specified educational or operational cost, and let $\omega_g \geq 0$ be transparently chosen group weights summing to one. A simple illustrative decision rule is

$$\pi^* \in \arg \max_{\pi \in \Pi} \left\{ \sum_{g=1}^K \omega_g \hat{m}_g(\pi) - \lambda C(\pi) \right\}. \quad (12)$$

This maximization is subject to a subgroup AOF requirement, a predeclared institutional cost ceiling, minimum accessibility requirements, human appeal, and a real educational alternative. Weights and cost terms must not relax an excluded group's threshold. The original comparator π_0 is feasible for the relative floor (its incremental effect is zero) if it also meets the other design constraints. An additional minimum-attainment rule can render the feasible set empty, which should trigger remediation or nondeployment, not undisclosed threshold relaxation. When a policy is selected using the same observations that estimate its safety, Equations (9)–(11) do not by themselves control selection-related estimation error. The protocol must therefore reserve independent certification data or support simultaneous inference over the prespecified candidate class, as detailed in Section 3.7.

As a complementary diagnostic, report the maximum modeled pairwise disparity

$$D(\pi) = \max_{g,g'} |\hat{m}_g(\pi) - \hat{m}_{g'}(\pi)|. \quad (13)$$

A small $D(\pi)$ does not ensure substantive inclusion; all groups could receive inadequate teaching. Conversely, the relative floor permits different subgroup gains when learners have different support needs. Disparity diagnostics and the floor should therefore be interpreted alongside absolute learning and access outcomes.

3.6. Architecture and accountability

Figure 1 depicts the AOF decision loop. Instructors first define learning outcomes and a real comparator. Designers then specify accessible modalities and identify relevant barriers through participatory consultation. A candidate AI policy is evaluated for independent learning, subgroup model validity, plausible distribution shift, implementation burden, and availability of human-led alternatives. The system can be piloted only where the evidence meets prespecified rules; otherwise the relevant lesson or pathway returns to instructor-controlled delivery. Logging records *which policy and version* each learner encountered without collecting unnecessary sensitive information.

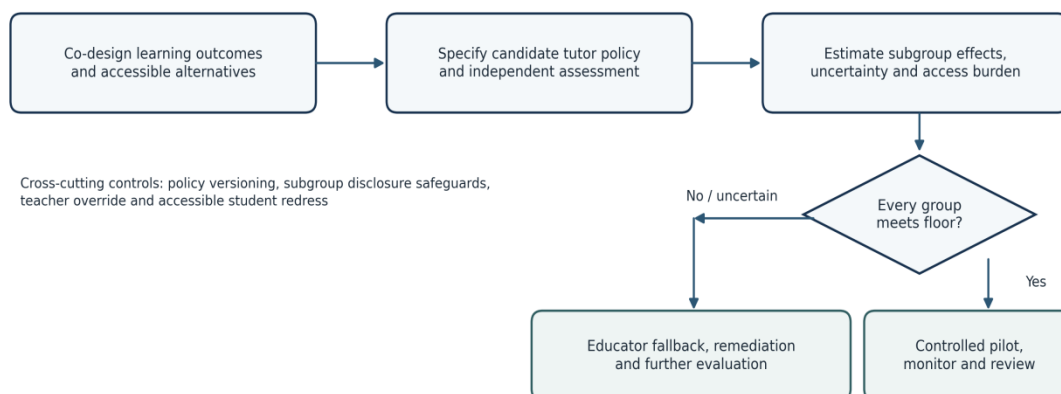


Figure 1. Adaptive Opportunity Floor architecture, from co-design through independent evaluation, robust subgroup certification, and educator-controlled fallback.

3.7. Finite-sample certification and adaptive policy selection

There are two different, valid calibration routes. Under the direct route used in Proposition 1, an empirical-centred radius ρ_g must satisfy $W_1(Q_g, \hat{P}_g) \leq \rho_g$ for the deployment law Q_g ; that radius already absorbs reference-sample variation. Alternatively, define a population reference law P_g^{ref} , and justify a shift-only radius ρ_g^{shift} for $W_1(Q_g, P_g^{\text{ref}}) \leq \rho_g^{\text{shift}}$. In the latter route, sampling uncertainty in the held-out estimate of the P_g^{ref} expectation must be treated separately. Do not simply add both versions of the same uncertainty adjustment.



For the population-referenced route, suppose the candidate policy π and the fitted conditional-benefit function $\hat{h}_{\pi,g}$ are fixed before examining the independent certification sample $X_1, \dots, X_{n_g}$, drawn i.i.d. from P_g^{ref} within each audited stratum. Suppose $|\hat{h}_{\pi,g}(x)| \leq b$ for all admissible x , with b specified before certification. By Hoeffding's inequality and a union bound over K strata, a simultaneous two-sided reference-mean deviation bound at error budget α_m is:

$$t_g(\alpha_m) = b \sqrt{\frac{2 \log \left(\frac{2K}{\alpha_m} \right)}{n_g}}. \quad (14)$$

Define the population-referenced, finite-sample lower certificate by the following quantity, where η_g is a valid uniform error allowance for the learned conditional causal effect:

$$\underline{F}_g(\pi) = \hat{m}_g(\pi) - t_g(\alpha_m) - L_{\pi,g} \rho_g^{\text{shift}} - \eta_g \quad (15)$$

Proposition 2 (simultaneous probabilistic subgroup certification). Suppose (i) the independent reference sample and the fixed-policy conditions above hold, (ii) for all K strata simultaneously, the target covariate distribution lies within the population-referenced shift radius with probability at least $1 - \alpha_\rho$, (iii) a valid uniform conditional-benefit error bound holds simultaneously with probability at least $1 - \alpha_\eta$, and (iv) the conditional causal response is transportable between the reference and deployment settings. Then the union bound yields:

$$\Pr\{ \Delta_g(\pi; Q_g) \geq \underline{F}_g(\pi) \text{ for all } g \} \geq 1 - \alpha_m - \alpha_\rho - \alpha_\eta. \quad (16)$$

Proof. Conditional on the policy and fitted function fixed before certification, Hoeffding's inequality establishes the reference-mean event simultaneously for K groups. On that event, $\mathbb{E}_{P_g^{\text{ref}}}[\hat{h}_{\pi,g}]$ is at least $\hat{m}_g - t_g$. The assumed Lipschitz property and shift event then lower-bound the target expectation by another $L_{\pi,g} \rho_g^{\text{shift}}$, and the model-error event subtracts at most η_g . Intersection of the three simultaneous events has probability at least $1 - \alpha_m - \alpha_\rho - \alpha_\eta$. This conclusion depends on verified joint or union-adjusted error and radius guarantees. ■

If J policies were fixed before an independent common reference sample and evaluated on that sample, replace the logarithmic K term with KJ to control the reference-mean sampling component simultaneously over all audited groups and candidates:

$$t_{g,J}(\alpha_m) = b \sqrt{\frac{2 \log \left(\frac{2KJ}{\alpha_m} \right)}{n_g}}. \quad (17)$$

The multiplicity correction in Equation (17) does not automatically certify conditional-benefit model error or Wasserstein coverage for all candidates. Those components also require simultaneous validity. For a policy learned from data, the preferred design uses an exploration/training partition, an optional policy-selection partition, and a final untouched certification partition; after certification, model or prompt updates require reassessment. Where subgroups are too small, the confidence bound may be uninformative, correctly producing “not certified.” Independence, randomized identification or justified causal assumptions, overlap, and valid subgroup measurements remain indispensable.

3.8. Relative non-harm versus absolute educational adequacy

An accessible comparator can still deliver learning below a curricular minimum. To prevent a relative-safety certificate from legitimizing an inadequate status quo, specify an independently justified attainment target τ_g and fit $\hat{v}_{\pi,g}(x)$ an estimate of $\mathbb{E}[Y(\pi) | X = x, G = g]$. If the attainment prediction has verified error η_g^Y , Lipschitz constant $L_{\pi,g}^Y$ and independent-sample deviation bound t_g^Y , a distinct sufficient condition is:

$$\underline{A}_g(\pi) = \hat{\mu}_g^\pi - t_g^Y - L_{\pi,g}^Y \rho_g^{\text{shift}} - \eta_g^Y \geq \tau_g, \quad \forall g \quad (18)$$

Here $\hat{\mu}_g^\pi$ is the held-out mean of $\hat{v}_{\pi,g}(X_i)$. Equation (18) is a separate conditional or high-probability requirement only when its own modeling, shift and sampling assumptions have been established; it is not implied by Equation (10). An institution should distinguish three decision states: sufficient evidence for both criteria; sufficient evidence for relative non-harm but insufficient absolute attainment; and insufficient evidence for either. In the latter two states, fallback and instructional redesign must be available, with human oversight of any exceptional deployment.

4. ANALYTICAL RESULTS

4.1. Demonstration of average-gain versus subgroup safety conflict

Table 2 provides a fully hypothetical calculation, not observed or simulated research data. Values are invented solely to verify the arithmetic and explain the decision criterion. Two candidate tutoring policies are compared. The first has greater nominal average improvement but uncertain or unfavorable outcomes for the access-constrained stratum. The second has lower nominal improvement but a nonnegative lower bound for both strata. The margin is set to zero.

Table 2. Hypothetical subgroup illustration of the AOF rule (normalized learning-gain units)

Policy	Prespecified stratum	Mean predicted benefit $\hat{m}_g$	Shift allowance $L_{\pi,g}\rho_g$	Model-error allowance η_g	Floor F_g
A: maximum-mean tutor	Higher access barriers	0.035	0.04	0.015	-0.02
A: maximum-mean tutor	Fewer recorded barriers	0.105	0.035	0.015	0.055
B: accessibility-first tutor	Higher access barriers	0.055	0.025	0.015	0.015
B: accessibility-first tutor	Fewer recorded barriers	0.08	0.025	0.015	0.04

Assuming equal group weights only for this illustration, policy A has mean predicted benefit 0.070 and policy B has 0.0675. Yet A fails the zero-decrement subgroup rule because its lower bound is -0.020 in the higher-barrier stratum, whereas B satisfies both bounds. The comparison does not say that policy A harms actual students. It establishes that the decision rule can distinguish average predicted benefit from conservatively certified subgroup benefit.

The arithmetic in Table 2 uses only the deterministic, direct-radius floor of Equation (9); it does not establish a finite-sample confidence certificate. To illustrate the importance of calibration without inventing empirical evidence, suppose the population-referenced variant assigns an additional sampling allowance of 0.020 to both groups while retaining the displayed 0.025 shift allowance and 0.015 model-error allowance for policy B. Its higher-barrier floor would then decrease from 0.015 to -0.005, so B would not be certified at a zero margin. The value 0.020 is hypothetical, not estimated; no inference should be made about an actual sample size, α level, or learner population.



4.2. Analytically derived implications

Result 1: A robust floor may reverse a mean-optimal choice. The Table 2 calculation is sufficient to show that maximizing group-weighted means without constraints can select a policy excluded by a non-inferiority floor. The magnitude and frequency of such reversals are empirical questions.

Result 2: Uncertainty matters independently of expected gain. From the definition of F_g , the floor decreases when ρ_g , $L_{\pi,g}$, or η_g increases, all else being equal. Institutions must therefore distinguish poor estimated performance from inadequate evidence. The latter can be addressed by better evaluation or a narrower deployment context rather than by concealing uncertainty.

Result 3: Fallback restores relative feasibility, not educational adequacy. Choosing π_0 makes the true incremental benefit exactly zero by construction, assuming a consistent intervention and comparator. This provides a relative reference, but does not establish that the original pathway is fair, effective, or sufficient for learners facing structural disadvantage.

Result 4: Group-average protection and individual protection differ. Proposition 1 concerns a group-averaged distribution of effects. It cannot imply that every learner receives a positive effect. AOF should be supplemented by individual complaint channels, teacher review, and harm monitoring, particularly where group means mask concentrated disadvantage.

4.3. Falsifiable propositions

P1: Relative to a mean-only adaptive policy, enforcing an AOF criterion will reduce the frequency of materially negative subgroup average effects, conditional on adequate identification and uncertainty calibration. P2: Where interface or access barriers are substantial, a multimodal and offline-capable adaptive pathway will produce a higher certified floor than a text-only pathway, though possibly at greater cost. P3: AOF use will produce more teacher fallback decisions when subgroup evidence is weak than when that evidence is precise. P4: Policy rankings based only on immediate AI-assisted task completion will diverge more often from rankings based on delayed independent assessment in learning environments with substantial tutor dependence. Each proposition could be false; none is presented as an observed finding.

5. PROSPECTIVE EMPIRICAL VALIDATION PROTOCOL

5.1. Population, sampling, and governance

A suitable first study would recruit adult learners across several university or continuing-education modules, with prospective consent, institutional ethics review where required, and accommodation planning before randomization. The unit of randomization should normally be teaching group or instructional session where spillovers between learners or instructor behavior make individual assignment problematic. Stratification would use baseline attainment and a limited set of predeclared access conditions. Any subgroup variable related to disability, language, or socioeconomic circumstances must have a specific legitimate purpose, restricted access, and a plan for small-cell disclosure prevention. Students who decline disclosure should still receive the best available accessible pathway.

5.2. Intervention and comparator

The comparator must be an instructional pathway actually accessible to the participants, not an artificial no-support condition. The treatment would consist of a documented model and version, approved instructional prompts, specified educator supervision, response-modality choices, accessibility options, and escalation conditions. Both arms would cover identical learning objectives and comparable instructional time. All necessary assistive accommodations would be permitted in both arms, with measurement accommodations documented to avoid confounding access with knowledge.



5.3. Primary outcome, secondary measures, and estimands

The primary endpoint should be a delayed, tutor-free assessment, for example four weeks after the instructional period, with blind scoring and pretested measurement equivalence across accessible response modalities. Immediate task completion, engagement, access friction, time burden, autonomy, learner complaints, and instructor workload would be secondary endpoints. A transparent estimand is the intention-to-treat mean difference within each predefined stratum. Missing assessments require explicit attrition reporting and sensitivity bounds; complete-case analyses alone are insufficient when connectivity or accommodation barriers predict dropout.

5.4. Estimation and uncertainty calibration

Investigators should preregister policy versions, candidate classes, comparator, delayed-learning endpoint, subgroup definitions, educational margins, minimum-attainment targets, an intended certification route, and a total error budget α split across mean sampling (α_m), shift coverage (α_p) and causal-model error (α_η). Data should be partitioned before outcome analysis: use appropriately identified randomized or observational data to fit and select candidates, then reserve a separate certification sample. In clustered educational settings, independently sampled classrooms rather than individual observations may determine effective concentration and uncertainty. The i.i.d. Hoeffding illustration in Equation (14) cannot be applied mechanically to correlated classrooms or weighted estimators; cluster-valid alternatives are then required. Observational clickstream data do not identify $h_{\pi,g}$ without credible identification assumptions, positivity, and a stable treatment definition (Athey & Wager, 2021).

Calibration must distinguish (a) a direct empirical radius with coverage for $W_1(Q_g, \hat{P}_g)$ and no additional correction for the same sample variation, from (b) a population shift radius with independent reference-mean uncertainty as in Equation (15). Define an educationally interpretable distance after prespecifying scales and rules for categorical access features. Estimate drift using a suitably representative independent deployment covariate sample or justify a scenario set; report the confidence level or clearly label a radius as scenario-based. A Lipschitz constant requires an analytical upper bound or defensible out-of-sample control, not merely the steepest slope observed in training. A uniform error bound η_g for conditional causal effects is substantially harder than a prediction residual or ordinary calibration score: if credible uniform coverage cannot be justified, report the floor as sensitivity analysis and do not assert statistical certification. Small-sample or intersectional groups may have vacuous bounds, requiring additional data or human-led alternatives.

Recommended decision workflow: first, lock learning outcomes and comparator with learner and disability-access consultation; second, preregister strata, δ_g , τ_g candidate policies, and acceptable failure rates; third, identify causal effects using an appropriate design and bound the model error; fourth, set the metric, plausible shift family, and calibrated uncertainty radius; fifth, certify the selected policy on independent held-out data with cluster-appropriate coverage and all subgroup adjustments; sixth, document accept, fallback, remediate, or stop decisions. A policy fails certification when uncertainty is too wide even if its point estimate is positive. Do not tune parameters after observing the certification result.

5.5. Decision rules, auditability, and stopping conditions

A pilot should specify a prospective rule that blocks or pauses expansion when subgroup learning falls below the predeclared margin, accessibility incidents accumulate, attrition is concentrated among constrained learners, or a model update invalidates the policy definition. Student complaints and educator override must be possible without penalizing the learner. Table 3 connects design obligations to observable audit evidence.

Table 3. Prospective AOF audit and implementation requirements

Audit dimension	Minimum evidence and operational decision
Learning validity	Delayed tutor-free, curriculum-aligned outcome with documented accessible scoring



Causal identification	Randomized or otherwise defensible comparator; report assignment, attrition, and spillovers
Group coverage	Predeclared access strata, sample counts, uncertainty, and small-cell privacy safeguards
Distributional robustness	Document $d, \rho_g, L_{\pi, g}$, and η_g with sensitivity analysis
Accessibility	Modality alternatives, assistive technology checks, low-bandwidth provision, and appeals
Human governance η_g	Named instructor owner, version control, explanation, override, and fallback logs
Stopping and redress	Prospective deterioration triggers, complaint pathway, and remediation timetable
Selection validity	Separate exploration and certification data, or simultaneous bounds for the full candidate policy class
Statistical calibration	Declared reference type, Wasserstein coverage or sensitivity, model-error evidence, cluster-adjusted sampling bounds
Baseline adequacy	Feasible accessible comparator plus independent absolute attainment standard and remedy if unmet

6. DISCUSSION

6.1. Educational contribution and theoretical boundary

The AOF contributes an operational distinction between inclusive intent and a defensible lower bound on subgroup educational benefit. It integrates accessible instructional choice, causal counterfactual learning targets, and precommitted governance with mathematical bounds. Its novelty is the coupled decision architecture and the proposed educational estimand, not a new optimal-transport theorem. Table 1 makes this delimitation explicit. Propositions 1 and 2 also distinguish a deterministic scenario certificate from a statistical statement dependent on independently justified sampling, shift, and model-error assumptions. The framework does not demonstrate that any particular AI tutor is inclusive.

The framework extends the logic of governance-integrated AI deployment advanced by Tan (2026c) from enterprise workflows to educational intervention decisions, while respecting a different objective and outcome structure. Its emphasis on learner alternatives and appeals is consistent with the agency-centred concerns articulated by Tan (2026b). It also supports UDL-style multiple pathways without presuming that a group descriptor determines any individual's learning preference (CAST, 2024). The direction is consonant with the view that AI should augment, rather than simply replace, human educational capability (Cukurova, 2025).

6.2. Practical use and institutional decision rights

Implementation can start without sophisticated causal software: instructors can first identify a defensible accessible comparator, agree on the independent learning outcome, specify the groups likely to be excluded, and map the action taken when evidence is insufficient. Subsequent quantitative work can estimate effects and uncertainty. The formal model then disciplines a governance decision already made explicit through instructional design. Institutional leaders should decide which access conditions matter, how scarce intervention resources are allocated, and who may override a system. These decisions cannot be delegated to a single fairness metric.

6.3. Limitations and possible failure modes

First, Wasserstein covariate robustness does not protect against changes in conditional causal effects, teacher behavior, interface functioning, language-model versions, or assessment validity. Second, choosing a distance metric and globally defensible Lipschitz constants is difficult for mixed categorical access features. Third, a uniform conditional-effect error bound is particularly demanding for small or intersectional groups, so the



framework may legitimately refuse certification. Fourth, classroom dependence, selection among many candidates, learner dropout, and reuse of certification data can invalidate convenient i.i.d. confidence formulas. Fifth, group averages may conceal important individual harms. Sixth, an accessible comparator may remain educationally inadequate even when the relative threshold is satisfied, requiring Equation (18) or a corresponding absolute learning target. Seventh, technical optimization cannot compensate for unavailable accommodations or institutional redress.

These limitations constrain the interpretation of Proposition 1 and motivate participatory review, measurement-equivalence testing, subgroup sensitivity analyses, individual redress, and absolute outcome targets. The proposed mechanism should not be used for high-stakes admissions, grading, or disability determination without separate safeguards and empirical validation.

6.4. Future research directions

Future studies should test whether AOF-constrained tutoring changes delayed subgroup learning outcomes, examine whether policy selection remains valid under independent certification, and compare direct ambiguity-set versus population-shift calibration in realistically small educational strata. Methodological work is needed on cluster-aware finite-sample bounds, defensible causal effect calibration, transportability, and outcome invariance across accommodations. Empirical work should quantify the costs of false assurance and excessive nondeployment, as well as whether joint relative and absolute floors detect inadequacy that relative non-inferiority alone misses.

7. CONCLUSION

Inclusive AI tutoring requires more than average satisfaction, fluent outputs, and short-term task completion. AOF frames educational inclusion as a policy-selection problem with accessible comparators, subgroup non-inferiority floors, and educator-controlled fallback. Proposition 1 gives a conditional deterministic certificate. Proposition 2 shows how an alternative population-referenced implementation can account for independent-sample uncertainty without double-counting direct empirical Wasserstein coverage. A separate attainment requirement guards against normalizing an inadequate baseline. The purely hypothetical example illustrates logical policy reversal, not observed effectiveness. Prospective, ethically reviewed educational evaluations with identified causal effects, credible error bounds, and independent policy certification remain necessary before AOF can be claimed as an evidence-backed deployment safeguard.

DECLARATIONS

Funding: No external funding was received for the preparation of this theoretical manuscript.

Conflict of interest: The author declares no competing interests.

Ethics approval and consent: No human participants, personal data, or animal subjects were involved in this theoretical analysis. Ethics review would be sought as required before any prospective evaluation.

Data availability: No empirical dataset was generated or analysed. All worked-example inputs are hypothetical and disclosed in Table 2.

Code availability: Not applicable to this analytical manuscript.

Author contributions: The sole author is responsible for conceptualization, formal analysis, methodology, writing, and manuscript approval.

REFERENCES

1. Athey, S., & Wager, S. (2021). Policy learning with observational data. *Econometrica*, 89(1), 133–161. <https://doi.org/10.3982/ECTA15732>

2. Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, 32(4), 1052-1092. <https://doi.org/10.1007/s40593-021-00285-9>
3. Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, 13(6), 4-16. <https://doi.org/10.3102/0013189X013006004>
4. CAST. (2024). *Universal design for learning guidelines version 3.0*. <https://udlguidelines.cast.org/>
5. Cukurova, M. (2025). The interplay of learning, analytics and artificial intelligence in education: A vision for hybrid intelligence. *British Journal of Educational Technology*, 56(2), 469-488. <https://doi.org/10.1111/bjet.13514>
6. Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., ... Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, Article 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
7. Miao, F., & Holmes, W. (2023). *Guidance for generative AI in education and research*. UNESCO. <https://unesdoc.unesco.org/ark:/48223/pf0000386693>
8. Mohajerin Esfahani, P., & Kuhn, D. (2018). Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171, 115-166. <https://doi.org/10.1007/s10107-017-1172-1>
9. 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>
10. Tan, K. H. (2013). *Strategic competency mapping and needs analysis in a boutique consulting firm: A case study of the Strawberry Strategy Consulting Group* [Technical report]. ResearchGate. <https://doi.org/10.13140/RG.2.2.26149.20967>
11. Tan, K. H. (2025). *The AI-augmented MBA paradigm: Reconceptualizing business education relevance in the era of artificial intelligence* [ResearchGate thesis]. ResearchGate. <https://doi.org/10.13140/RG.2.2.24863.06566>
12. Tan, K. H. (2026a). Improving critical thinking skills among adult learners through structured reflective dialogue: An action research study in a university social sciences course in Singapore. *International Journal of Research Publication and Reviews*, 7(6), 5322-5329. <https://doi.org/10.55248/gengpi.07.0626.17a03>
13. Tan, K. H. (2026b). Human agency under algorithmic governance: A humanities framework for law, management and public life. *International Journal of Law Management & Humanities*, 9(4), 559-571. <https://doi.org/10.63108/IJLMH.12537>
14. Tan, K. H. (2026c). Governance-integrated agentic automation for enterprise innovation engineering: A productivity-governance control cycle and risk-adjusted evaluation model. *International Journal of Sciences and Innovation Engineering*, 3(7), 172-182. <https://doi.org/10.70849/ijsci03072617871>
15. Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education: Where are the educators? *International Journal of Educational Technology in Higher Education*, 16, Article 39. <https://doi.org/10.1186/s41239-019-0171-0>