Adoption-Aware Crop Recommendation: Expected Realized Benefit as a Direct-Method Offline Policy under Unidentifiable Compliance

Anonymous authors
Paper under double-blind review

Abstract

A recommendation only helps if it is taken. Most recommender systems, though, are trained on observational data to predict an outcome and then deployed to pick an action for each user as if that action will be followed. We study the resulting gap in smallholder agriculture, where the standard approach recommends the (crop, input) bundle with the highest predicted yield and says nothing about whether the farmer will adopt it. Our objective, **Expected Realized Benefit (ERB)**, instead recommends the action with the largest predicted yield gain *weighted by the probability the farmer adopts it*. We build an ERB-maximizing recommender from two calibrated models, one for yield and one for adoption, over a five-tier action space, and we treat it as a direct-method off-policy estimator: farmer-side propensities are not identifiable from survey data, so inverse-propensity and doubly-robust evaluation simply do not apply. On a real LSMS-ISA Ethiopia panel (6,770 households), the adoption-aware policy beats a conventional accuracy-only baseline by +50 kg/ha (cluster-robust 95% CI $[+33, +70]$, cluster permutation $p < 10^{-3}$, $n = 800$ enumeration-area-disjoint test plots). Every number here is a model-believed offline estimate, not a field-measured return. The gain holds up under four sensitivity sweeps, a Rosenbaum analysis ($\Gamma \geq 5$), a marginal-sensitivity-model pessimistic policy (out to $\Lambda = 5$), a partial-identification floor anchored to the Duflo–Kremer–Robinson Kenya fertilizer RCT (+26 kg/ha $[+13, +39]$), and four CATE-metalearner baselines that, without an adoption model, never recover it. Reported conservatively—propagating the yield split-conformal interval and a bootstrap uncertainty set on the adoption head jointly—the policy still guarantees a +18 kg/ha $[+9, +29]$ lift over status quo, and the headline is robust to $\pm 50\%$ perturbation of the yield-head literature priors (not only the adoption penalties). An education-shuffle placebo shows the per-subgroup gradient is descriptive rather than identified, so we treat the aggregate gain as the central claim. Validating the adoption model against the same RCT then exposes the paper's principal limitation: the model gets the ordering of the treatment arms right but is badly miscalibrated in absolute terms on out-of-distribution actions, off by roughly 50 percentage points. We frame that gap as a concrete benchmark for future work that fuses observational and experimental evidence, and we release the code, data pipelines, and RCT-validation harness.

1 Introduction

A deployed recommender only helps if its advice is taken. Yet the usual pipeline trains a model on observational data to predict some outcome, then deploys it to pick an action for each user as though the user will comply. Whether they comply, their *adoption*, is recorded only for the action actually taken, never for the alternatives the system might have proposed. That asymmetry drives a wedge between how a policy looks in-sample and how it does in the field: a recommender can ace its training objective and still under-deliver by advising actions people will not take. The problem bites hardest when adoption varies a lot across users, when the action space mixes cheap easy options with expensive high-friction ones, and when the logging pol-

icy's propensities cannot be identified. That last case is exactly when the inverse-propensity-scoring (IPS) and doubly-robust (DR) corrections that would otherwise clean up outcome-model bias are off the table.

We study this gap in smallholder crop recommendation, where adoption-blindness is still the default. Conventional systems output the *agronomic optimum*, the (crop, input) bundle that maximizes predicted yield at a location, and say nothing about whether the target farmer will take it up (Lobell et al., 2009). In practice uptake swings widely with a household head's education, gender, landholding, market access, and extension contact; a meta-analysis of 367 adoption studies across 41 countries in Sub-Saharan Africa and Asia finds education and extension contact to be the most reliable positive predictors (Ruzzante et al., 2021; Feder et al., 1985). The problem sits naturally in the offline contextual-bandit frame, with each plot a context, each input bundle an action, and observed yield the reward, and the actions carry real structure because input tiers escalate in cost and friction. The setting also carries an unusual advantage: agricultural field experiments (Duflo et al., 2011) supply externally identified counterfactual adoption rates, so we can check an offline-trained policy against an RCT ground truth in a way most offline-policy benchmarks cannot.

Expected Realized Benefit. We propose **Expected Realized Benefit (ERB)** as an offline policy objective that internalizes per-action adoption likelihood:

$$\begin{aligned}\mathrm{ERB}(a \mid x) = \mathbb{P}(\text{adopt}\, a \mid x) \cdot \mathbb{E}[Y \mid a, x] \\ + \big(1 - \mathbb{P}(\text{adopt}\, a \mid x)\big) \cdot \mathbb{E}[Y \mid a_0, x],\end{aligned} \tag{1}$$

where a is a candidate action, a_0 the user's status-quo action, and Y the outcome (yield). Equivalently, the ERB *lift* over status quo is

$$\Delta\mathrm{ERB}(a \mid x) = \mathbb{P}(\text{adopt}\, a \mid x) \cdot \big(\mathbb{E}[Y \mid a, x] - \mathbb{E}[Y \mid a_0, x]\big). \tag{2}$$

ERB collapses to mean predicted yield when $\mathbb{P}(\text{adopt}) \equiv 1$, and otherwise penalizes recommendations the user is unlikely to follow; a recommendation generates positive lift only if it is *both* yield-improving and adoption-likely. Because the status-quo term $\mathbb{E}[Y \mid a_0, x]$ is constant across the candidate actions for a given plot, $\Delta\mathrm{ERB}(a \mid x) = \mathrm{ERB}(a \mid x) - \mathbb{E}[Y \mid a_0, x]$, and the two objectives induce the *same* argmax policy; we use them interchangeably and report results in ΔERB (lift) units throughout. The argmax-ERB policy automatically backs off from agronomically aggressive recommendations when adoption probability is low.

Contributions. This paper makes four contributions.

1. **An adoption-aware recommender, formalized as a direct-method offline-policy estimator under unidentifiable compliance** (§3). Both heads, the yield head $\mathbb{E}[Y \mid a, x]$ and the adoption head $\mathbb{P}(\text{adopt}\, a \mid x)$, must answer counterfactual queries on actions never observed for a given context. We address this with a hybrid architecture (an ML-learned baseline plus literature-anchored response components) and position the estimator explicitly as direct-method (DM) rather than IPS/DR (§3.5), because farmer-side logging propensities are not identifiable from observational survey data. Heads are calibrated with Mondrian-isotonic recalibration (adoption) and split-conformal intervals (yield).

2. **Offline-policy results on real LSMS-ISA Ethiopia data** (§4): a statistically significant aggregate ERB lift of +50 kg/ha over an accuracy-only baseline (plot-level bootstrap 95% CI $[+24, +74]$; EA-cluster-robust 95% CI $[+33, +70]$, cluster permutation $p < 10^{-3}$; $n = 800$ EA-disjoint test plots), robust to four orthogonal sensitivity sweeps (§C).

3. **A transparent separation of identified from descriptive claims** (§4.4). Rosenbaum signed-rank bounds ($\Gamma \geq 5$, applied to the model-based matched-pair contrast rather than a realized outcome), a sign-exchangeable permutation test, and a SAFI-anchored partial-identification floor (+26 kg/ha $[+13, +39]$) bracket the aggregate effect; an education-shuffle placebo shows the per-subgroup gradient is *not* distinguishable from random label assignment, so we report it as descriptive concentration, not equity targeting. We argue this placebo should be default practice in offline-policy applications with sociotechnically loaded subgroups.

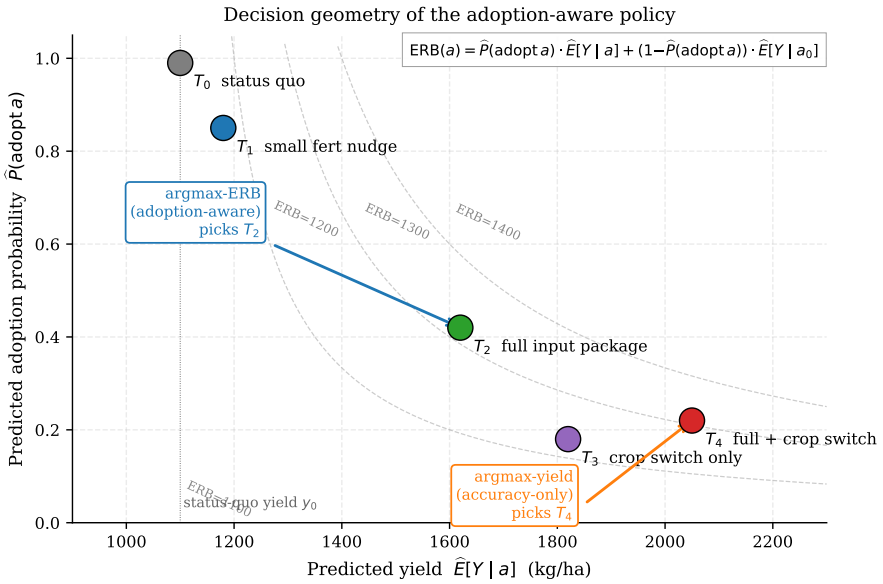

Figure 1: **Decision geometry of the offline policy.** For one plot, the five candidate actions (T_0 status quo through T_4 full-package crop-switch) occupy distinct points on the (predicted yield, predicted adoption probability) plane. Dashed curves are iso-ERB contours anchored at status-quo yield y_0. The accuracy-only policy picks the rightmost point (T_4, highest predicted yield); the ERB policy picks the point on the highest iso-ERB contour (T_2, the best yield-vs-adoption trade-off).

4. **Direct external validation against an identified-propensity RCT** (§4.5). Evaluated against the Duflo–Kremer–Robinson (Duflo et al., 2011) Kenya SAFI fertilizer experiment, the offline-trained adoption head reproduces the experimentally measured directional ordering of treatment effects but with ~ 50 percentage points of absolute miscalibration on out-of-distribution actions, a concrete, reproducible characterization of an observational-to-experimental calibration gap that any future method targeting this regime can be benchmarked against.

Our contribution is in framing, evaluation methodology, and empirical findings rather than new ML primitives; the components (gradient boosting, Mondrian-isotonic recalibration, split-conformal prediction, Rosenbaum bounds) are established. We make this scope explicit throughout.

2 Related Work

Off-policy evaluation and learning. ERB is a direct-method (DM) off-policy estimator (Dudík et al., 2011; Jiang & Li, 2016; Thomas & Brunskill, 2016): rather than re-weighting observed rewards by inverse propensity (Horvitz & Thompson, 1952; Swaminathan & Joachims, 2015), we model $\mathbb{E}[Y \mid a, x]$ and $\mathbb{P}(\text{adopt}\, a \mid x)$ separately and combine them. We do not apply IPS/SNIPS/DR because farmer-side logging propensities $\mathbb{P}(a^{\text{obs}} \mid x)$ are confounded with unobserved plot-level potential yield and are not identifiable from the survey instruments (§3.5). The CATE / uplift / metalearner literature (Künzel et al., 2019; Nie & Wager, 2021; Athey & Imbens, 2016; Wager & Athey, 2018; Kennedy, 2023; Rzepakowski & Jaroszewicz, 2012) is conceptually adjacent but does not model adoption-conditional uptake; T-/X-/R-learner and DR-learner baselines on the same panel are a natural next comparison (§G).

Robust and pessimistic OPL under unobserved confounding. This is the literature most directly addressing our central limitation. Partial-identification (Manski, 2003) and marginal-sensitivity-model (MSM) approaches (Tan, 2006; Kallus & Zhou, 2018; Yadlowsky et al., 2022; Namkoong et al., 2020; Oprescu et al., 2023) construct sets of policy values consistent with observational data under bounded hidden confounding

and recommend the policy maximizing the worst-case lower bound; pessimistic offline RL (Jin et al., 2021; Rashidinejad et al., 2021; Buckman et al., 2021) formalizes the same intuition. Proxy / bridge-function methods (Tchetgen Tchetgen et al., 2020; Cui et al., 2024), including proximal off-policy evaluation (Bennett & Kallus, 2024), relax the no-unobserved-confounders assumption via negative-control structure. Our DM ERB is an observational-only baseline along these axes; an MSM-pessimistic ERB variant (Λ-MSM with $\Lambda \in [1.5, 5]$ matching our Rosenbaum critical-Γ regime) is named in §G.

Conformal and risk-aware decision-making. Conformal off-policy prediction (Taufiq et al., 2022) combined with Mondrian-conformal heads (Vovk et al., 2005; 2003) would yield a properly distribution-free joint ERB lower bound, replacing the partial joint uncertainty quantification we currently have (§5). Capacity- and risk-aware recommendation (de Nijs et al., 2018; Cheng et al., 2024) treats action friction as a first-class component; we operationalize cost as adoption probability and show (§C) that ERB beats a yield-minus-λ·complexity heuristic at every λ.

Combining observational and experimental evidence. Methods that combine observational and experimental data and transport trial findings to target populations (Colnet et al., 2024; Dahabreh et al., 2019; Bareinboim & Pearl, 2016) are the natural extension class for the calibration gap we characterize empirically (§4.5).

Adoption/compliance-aware decisions across domains. The gap between a recommendation and its uptake recurs far beyond agriculture: medication non-adherence is widely predicted from observational health records (Gu et al., 2021), and recommender-systems research treats user acceptance as a first-class design concern (Ricci et al., 2012), alongside cost- and capacity-aware recommendation (de Nijs et al., 2018; Cheng et al., 2024). ERB is, in this light, an adoption-aware decision objective whose form is domain-agnostic: any setting with a learned outcome model and heterogeneous, unidentified compliance fits the same template.

Contrast: compliance-aware uplift, IV-based recommendation, and joint-uncertainty decisions. ERB sits between three adjacent lines of work, and it helps to distinguish it from each. *(a) Compliance-aware uplift.* Uplift modeling (Rzepakowski & Jaroszewicz, 2012) and CATE metalearners (Künzel et al., 2019) estimate the effect of an action on those who take it; ERB differs by *re-weighting* a modeled effect by a separately estimated per-action adoption probability rather than estimating an effect on compliers, which is what lets it score counterfactual actions (tiers) that were never assigned in the data. Our metalearner baselines (§E.1) make the distinction operational: sharpening the compliance-conditional contrast on the one identified action does not recover ERB's gain. *(b) IV / LATE-based recommendation.* The econometrics of imperfect compliance (Angrist et al., 1996) identifies a local effect on compliers via an instrument; we have no instrument in the LSMS panel, so an IV/LATE estimand is unavailable and ERB is explicitly a direct-method surrogate—the gap that our RCT validation (§4.5) and the RCT-augmented extension (§G) target. *(c) Joint-uncertainty decision-making.* Robust/pessimistic OPL (Tan, 2006; Kallus & Zhou, 2018) and conformal decision-making (Taufiq et al., 2022) bound decision value under one source of uncertainty; ERB's decision consumes *two* learned quantities, so its natural robust form is a *joint* lower bound over yield and adoption uncertainty together—which we construct and foreground (§4.4, App. E.4) rather than bounding only the yield side.

3 Methods

Data. All experiments use real Living Standards Measurement Study panel micro-data harmonized into a canonical 80-column schema. The primary panel is LSMS-ISA Ethiopia ESS Wave 4 (Central Statistical Agency of Ethiopia & World Bank LSMS-ISA Team, 2019) (2018–19; 19,339 plot-level rows from 6,770 households across $\sim$ 430 enumeration areas; 8,744 cultivated plots after filtering). Cross-country replication uses LSMS-ISA Tanzania NPS Wave 4 (Tanzania National Bureau of Statistics & World Bank LSMS-ISA Team, 2015). Both panels are joined to CHIRPS rainfall (Funk et al., 2015), WorldClim temperature (Fick & Hijmans, 2017), NDVI, FAO soil quality, plus farmer-profile features (education, age, sex, household size, market distance, extension contact) and plot-level features. Documented LSMS-handling fixes (education-

code-to-years conversion, per-hectare ratio winsorization) and EA-centroid GPS privacy offsets are described in §A.

Notation. A farm-plot observation i is $(x_i, a_i^{\text{obs}}, y_i)$, where x_i is the feature vector (farmer + plot + agro-ecological), a_i^{obs} is the input bundle the farmer actually chose, $y_i \in \mathbb{R}_{\geq 0}$ is realized yield (kg/ha), and plot i nests in enumeration area $g(i)$. The deployment problem: given a new (x_*, a_*^{obs}), output $\hat{a}_* \in \mathcal{A}(x_*)$. Conventional accuracy-only systems set $\hat{a}_* = \arg\max_a \mathbb{E}[Y \mid a, x_*]$; we argue this is the wrong objective when $\mathbb{P}(\text{adopt } \hat{a}_*) < 1$.

3.1 Structured five-tier action space

For each plot we enumerate $\mathcal{A}(x_*)$ as five candidates ordered by adoption complexity: T_0 status quo; T_1 small fertilizer nudge (+30 kg/ha to the current crop); T_2 full input package on the current crop (EA-mean fertilizer rate among adopters + improved seed); T_3 crop switch only (to the EA-best-yielding observed crop); T_4 full package + crop switch ($T_2 + T_3$). Tier ordering encodes complexity-from-status-quo and is the input to the adoption head's action-complexity feature subvector. Discrete tiers (rather than continuous rates) produce interpretable policies and admit a clean head-to-head comparison without requiring counterfactual yield curves at every fertilizer level.

EA-level aggregates and deployment. Three quantities are anchored on the enumeration area: the peer-adoption rate (an adoption-head feature), the EA-mean fertilizer rate among adopters (T_2), and the EA-best-yielding observed crop (T_3). Under our EA-disjoint protocol these are computed for each *held-out* EA from that EA's own observed plots with the focal plot left out, so they introduce no train-to-test leakage and never borrow from the training EAs. At deployment in a genuinely new EA, the same three scalars are estimated from a small local survey sample; where local data are sparse they back off to agro-ecological-zone or regional aggregates, or to a normative anchor (e.g. GAEZ-derived agronomic potential) that avoids the selection bias the EA-observed "best crop" inherits (§A). The framework requires only these per-EA scalars, not membership in the training set, and they need a small local sample rather than a census: App. E.8 shows the policy near-invariant down to $k = 5$–10 sampled plots per EA (the ERB-argmax shifts for $< 1.5\%$ of plots and the aggregate lift stays $+50 \pm 2$ kg/ha), with the EA-best-crop anchor backing off to an agro-ecological-zone default where local plots are sparse.

3.2 Yield head

A **hybrid** architecture is required because a pure histogram gradient-boosting regressor in the EA-disjoint regime exhibited near-zero predicted-yield deltas on counterfactual action perturbations (75% of plots show identical prediction at +30 kg/ha vs. status quo), which collapses the policy-comparison signal. The hybrid restores action-sensitivity by adding a literature-anchored linear response:

$$\begin{aligned}\hat{f}_{\text{hybrid}}(x, a) = \hat{f}_{\text{base}}(x) \cdot \big(1 + \alpha_{\text{seed}} \mathbb{1}[\text{seed}] + \alpha_{\text{irr}}(x) \mathbb{1}[\text{irr}]\big) \\ + \beta_{\text{crop}(a)} \cdot \Delta_{\text{fert}}(a),\end{aligned} \tag{3}$$

with β_{crop} from agronomic meta-analysis (maize 7, wheat 6, teff 4, legumes 2 kg yield/kg fertilizer) (Vanlauwe et al., 2014; Mueller et al., 2012), $\alpha_{\text{seed}} = 0.15$ (an improved-variety yield premium broadly consistent with Sub-Saharan adoption-impact studies (Khonje et al., 2015)), and rainfall-modulated α_{irr}. The base $\hat{f}_{\text{base}}$ is a histogram gradient-boosting regressor with Gamma loss (Ke et al., 2017); squared error in log-space introduces a Jensen bias we observed empirically. We wrap predictions in Mondrian-by-crop split-conformal regression intervals (§3.5). Dose-response caveats and tier-specific perturbation rules are in §A.

3.3 Adoption head

The adoption head $\hat{g} : \mathcal{X} \times \mathcal{C} \to [0, 1]$ predicts $\mathbb{P}(\text{adopt } a \mid x, c(a))$, where $c(a)$ is a 7-dimensional complexity vector (tier, Δfert, $\log(1 + \Delta\text{fert})$, crop-switch indicator, new-seed indicator, count of inputs introduced, EA peer-adoption rate). A pure gradient-boosted classifier with these action features collapses to near-binary outputs by tier, because the action-complexity features deterministically encode the realized label in

observational data; we therefore decompose into a baseline propensity (logistic, farmer/plot features only) plus a literature-anchored action-complexity penalty:

$$\operatorname{logit} \hat{\mathbb{P}}(\text{adopt } a \mid x) = \operatorname{logit} \hat{p}_{\text{base}}(x) + \pi(a, x_{\text{peer}}, x_{\text{edu}}), \tag{4}$$

with tier-specific log-odds penalties $\{T_0{:}0,\ T_1{:}-0.4,\ T_2{:}-1.5,\ T_3{:}-3.2,\ T_4{:}-2.8\}$ calibrated to the Ruzzante et al. (2021) meta-analysis, modulated additively by the EA leave-one-out peer-adoption rate (coefficient 1.0 (Foster & Rosenzweig, 1995)) and by education ($-0.25 \log(1 + \text{edu})$). Raw outputs are Mondrian-isotonic recalibrated by education quintile, reducing expected calibration error (ECE) on the fertilizer head from $0.166 \rightarrow 0.138$ and on the improved-seed head from $0.131 \rightarrow 0.061$. Counterfactual queries on actions a outside a farmer's observed training distribution are prone to extrapolation error: this is the central technical risk we validate against the Kenya RCT in §4.5.

3.4 Policy head

The policy head $\pi : \mathcal{X} \rightarrow \mathcal{A}$ enumerates the action space, scores ERB for each candidate using the calibrated yield and adoption heads, and returns the argmax:

$$\pi^{\text{aware}}(x) = \operatorname*{arg\,max}_{a \in \mathcal{A}(x)} \widehat{\text{ERB}}(a \mid x).$$

This is a deterministic decision rule (an offline contextual-bandit policy), not a learned policy network; there is no reward-driven optimization loop, and we do not claim one. We additionally evaluate two variants:

- **Accuracy-only:** $\pi^{\text{acc}}(x) = \arg\max_a \widehat{\mathbb{E}}[Y \mid a, x]$, the conventional system, which ignores adoption probability entirely.
- **Conformal-aware (risk-averse):** $\pi^{\text{conf}}(x) = \arg\max_a \widehat{\text{ERB}}_{\text{lo}}(a \mid x)$, where $\widehat{\text{ERB}}_{\text{lo}}$ uses the lower bound of the yield-head conformal interval. Appropriate for subsistence contexts where downside-risk avoidance dominates expected-gain maximization.

We report all three policies side by side in §4.

3.5 Calibration, splits, and position relative to OPE

Calibration. The yield head wears Mondrian-by-crop split-conformal regression intervals (88.7% empirical coverage at a 90% nominal target). The adoption head receives Mondrian-isotonic recalibration by education quintile but no conformal interval, because split-conformal binary classification returns prediction *sets* on the label, not intervals on the probability; that proper extension is named as future work (§G).

Fairness note on calibrating by education. Stratifying the adoption recalibration by education quintile is a deliberate choice with a two-sided fairness profile, and we state the trade-off explicitly. On the benefit side, a single pooled isotonic map would inherit the majority (no-education) stratum's error and systematically misserve the smaller educated strata; per-quintile calibration equalizes calibration error across education groups (ECE improves most on the rarer improved-seed target, Table 1), which is a group-calibration fairness gain. On the risk side, conditioning any part of the pipeline on a sensitive-correlated attribute can encode disparity if the attribute proxies for structural disadvantage rather than capability. Two safeguards bound this risk here: the education-shuffle placebo (§4.4) shows the *policy's* per-subgroup gradient is not identified by the education feature (only $\sim 9\%$ of recommendations change when education is counterfactually manipulated), so the calibration stratification does not translate into education-driven recommendation disparity; and calibration acts only on probability *scale*, not on the tier ordering the policy consumes. We nonetheless flag that group-calibration and individual-fairness objectives can conflict, and that a deployment auditing recommendation rates by education (not just calibration error) is the right guard (§H).

Splits. All experiments use EA-disjoint train/calibration/test splits: every enumeration area contributes plots to exactly one of train, calibration, or test. Random within-village splits leak peer-norm and micro-climate effects; we report within-EA performance as a robustness check, where AUROC and R^2 rise substantially (§C), indicating that spatial transfer rather than intrinsic predictability is the binding constraint.

	n	base	AUROC	ECE (uncal → cal)	Brier (uncal → cal)
Adoption: fertilizer	8,744	0.41	0.70	0.166 → **0.138**	0.251 → 0.246
Adoption: improved seed	8,744	0.12	0.73	0.131 → **0.061**	0.114 → **0.088**

Table 1: Adoption-head test metrics under Mondrian-isotonic recalibration. The proper conformal binary-classification extension (prediction sets with marginal label coverage) is future work.

Position relative to OPE. ERB is a *direct-method* (DM) off-policy estimator (Dudík et al., 2011). We do not apply IPS/SNIPS/DR because the logging-policy propensities $\mathbb{P}(a^{\text{obs}} \mid x)$ are not identifiable from LSMS observational data: farmer choices depend on unobserved heterogeneity (private plot-quality knowledge, idiosyncratic risk preferences, weather expectations) not captured in any survey instrument. Moreover, even with surrogate propensities, the many near-zero adoption probabilities on counterfactual actions would make IPS and doubly-robust estimators extremely high-variance; the direct method is thus also the more *stable* choice here, not only the more identifiable one. Where partial proxy structure is available, sensitivity-aware doubly-robust or proximal off-policy evaluation (Bennett & Kallus, 2024) are the natural refinements; our proxy-control check (App. E.5) is a step in that direction, and a full proximal estimator awaits validated negative controls. We compensate with (i) Mondrian-conformal uncertainty on the yield head, which provides distribution-free guarantees under misspecification; (ii) external validation against an identified-propensity RCT (§4.5); and (iii) Rosenbaum sensitivity and a partial-identification bound (§4.4).

Inference. Headline experiments report bootstrap 95% CIs ($n_{\text{boot}} = 2{,}000$) on the matched-pair difference $d_i = \Delta\text{ERB}_i^{\text{aware}} - \Delta\text{ERB}_i^{\text{acc}}$. Because plots nest in enumeration areas, we additionally report an EA-cluster bootstrap (resampling whole EAs) and an EA-cluster sign-exchangeable permutation test (§4.4). Rosenbaum signed-rank sensitivity reports the critical Γ at which the one-sided p-value crosses 0.05.

4 Results

4.1 Calibration of the two heads

Adoption head. Test-set metrics (EA-disjoint) are in Table 1. Mondrian-isotonic recalibration substantially reduces ECE, with the largest gains on the rarer improved-seed target. We report calibration metrics but do *not* claim a conformal coverage guarantee on the calibrated probabilities themselves (§3.5).

Yield head. Pooled $R^2 = +0.022$ (MAE 1,051 kg/ha) on the EA-disjoint test set ($n = 1{,}457$), with statistically significant ranking quality (Spearman $\rho = +0.180$, $p \approx 5 \cdot 10^{-12}$; Kendall $\tau = +0.121$). Within-major-staple ranking correlations are stronger ($\rho \in [+0.28, +0.43]$ for teff, maize, sorghum, wheat, millet). Within-EA R^2 rises to $+0.238$ (§C), localizing most of the gap to spatial transfer rather than intrinsic predictability. **The relevant fidelity question for an ERB-maximizing policy is whether the head ranks candidate actions correctly on a given plot, not whether absolute predictions are accurate**; the rank-correlation result establishes that. Because the policy's choice depends on the within-plot *ordering* of the five tiers — mostly same-crop tiers, where even the conformal interval cancels in the contrast (App. E.4) — the modest pooled rank correlation is less consequential than it appears: ΔERB decisions are stable wherever the head orders the tiers correctly, and the B3 ablation (§C.3) shows the literature-anchored response, not the noisy ML level, supplies that ordering. Conformal coverage matches the nominal target (88.7% empirical at 90% nominal).

4.2 Head-to-head: adoption-aware vs. accuracy-only

This is the central experiment. We score both policies on the same 800 held-out EA-disjoint plots over the same five-tier action space, compute the matched-pair difference $d_i = \Delta\text{ERB}_i^{\text{aware}} - \Delta\text{ERB}_i^{\text{acc}}$ per plot, and report bootstrap 95% CIs. Figure 2 visualizes the result; Table 2 gives the numbers.

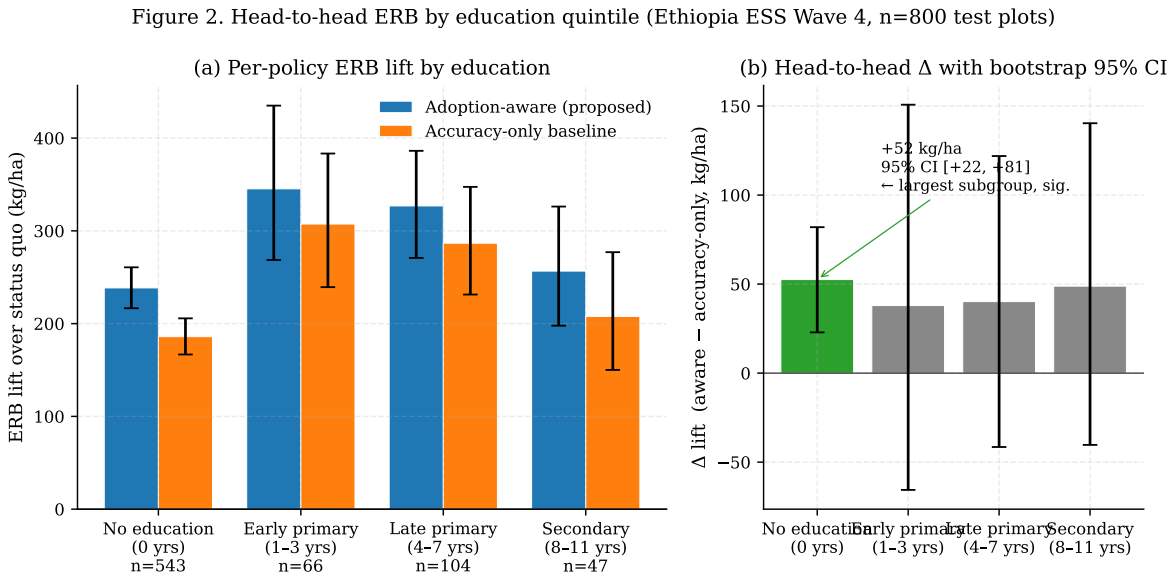

Figure 2: Head-to-head ERB lift by education quintile (Ethiopia ESS Wave 4, $n = 800$ test plots). **(a)** Per-policy lift over status quo with bootstrap 95% CIs. **(b)** The head-to-head Δ with bootstrap 95% CIs. The aggregate and the no-education subgroup are the cells whose plot-level CIs exclude zero (see Table 2 and §4.4 for cluster-robust inference).

Subgroup	n	Aware lift (kg/ha)	Accuracy-only lift	Δ aware − acc
All farmers	800	$+263$ $[+244, +282]$	$+213$ $[+196, +230]$	$\mathbf{+50}$ $\mathbf{[+24, +74]}^*$
No education	543	$+239$ $[+217, +261]$	$+186$ $[+167, +206]$	$\mathbf{+53}$ $\mathbf{[+23, +82]}^*$
Early primary	66	$+345$ $[+269, +435]$	$+307$ $[+239, +383]$	$+38$ $[-66, +151]$
Late primary	104	$+327$ $[+271, +386]$	$+287$ $[+231, +347]$	$+40$ $[-42, +122]$
Secondary	47	$+257$ $[+198, +326]$	$+208$ $[+150, +277]$	$+49$ $[-40, +140]$

Table 2: Head-to-head ERB lift over status quo by household-head education quintile (hybrid heads). Plot-level bootstrap 95% CIs in brackets ($n_{\text{boot}} = 2{,}000$). * CI excludes zero. Both the aggregate and the no-education subgroup are significant. Smaller subgroups ($n \in \{47, 66, 104\}$) show consistent positive point estimates ($+38$ to $+49$) but do not individually clear plot-level significance, a power constraint rather than an effect-direction failure (§4.4 reports cluster-robust subgroup CIs). The four quintiles shown cover 760 of the 800 test plots; the remaining 40 (39 records with missing household-head education, plus a singleton post-secondary stratum, $n = 1$, $\Delta = 0$) enter the aggregate row only.

Headline claim. The adoption-aware policy delivers a statistically significant aggregate ERB lift over the accuracy-only baseline of $+50$ kg/ha (plot-level bootstrap 95% CI $[+24, +74]$), and within the no-education subgroup specifically ($+53$ kg/ha; 95% CI $[+23, +82]$, $n = 543$, the largest stratum and the most policy-relevant target population for smallholder advisory systems). The result is reproducible across random seeds (seed 42: $+50$ $[+24, +74]$; seed 43: $+42$ $[+5, +83]$). *All ERB figures are model-believed offline estimates reported in kg/ha*: they depend on the calibrated yield and adoption heads and are not field-measured returns (§5); the economic translations in App. F are illustrative orders of magnitude only. A conformal-aware variant (selecting the argmax of the ERB conformal lower bound) trades roughly 30–50% of the headline gain for distribution-free worst-case coverage (§C). **As a default conservative report** (App. E.4), a risk-averse policy scored under a joint lower bound that propagates *both* the yield split-conformal interval and a bootstrap uncertainty set on the adoption head still guarantees $+18$ kg/ha $[+9, +29]$ over status quo; we recommend this joint lower bound as the number to act on, with the $+50$ point estimate as the expected-value counterpart.

4.3 Mechanism: when policies disagree

Across 800 test plots the two policies agree 54% of the time; on the 46% ($n = 368$) where they diverge, the dominant transition (93%, $n = 341$) is accuracy-only's "full package + crop switch" (T_4) re-routed by adoption-aware to "full package, current crop" (T_2), with a +114 kg/ha mean lift gain and +0.17 adoption-probability gain (App. E.2). Because the adoption-aware action is by construction the ERB-argmax under the same model that scores both policies, the near-deterministic win on disagreement plots is largely definitional; what is informative is which transitions drive it, and the pattern matches the technology-adoption literature—when the argmax-yield action requires a crop switch, the complexity penalty discounts its uptake and the policy trades marginal switching yield for the higher adoption at T_2. The headline evidence remains the aggregate matched-pair lift of §4.2 and the checks below, all model-believed offline estimates (§5) until a field trial observes realized yields.

4.4 Robustness: identified vs. descriptive claims

Cluster-robust inference. Because plots nest in enumeration areas, plot-level bootstrap CIs can understate uncertainty. Resampling whole EAs (44 clusters) gives an aggregate cluster-bootstrap 95% CI of $[+33, +70]$ kg/ha (99% CI $[+27, +77]$), and an EA-cluster sign-exchangeable permutation test rejects the no-effect null at $p < 10^{-3}$. Under the cluster bootstrap, all four education subgroups' lifts are individually positive with CIs excluding zero (no education +53 $[+33, +74]$; early primary +38 $[+9, +90]$; late primary +40 $[+23, +60]$; secondary +49 $[+25, +71]$); see §D. The aggregate effect is thus robust to the most consequential dependence structure in the data.

Rosenbaum sensitivity to unmeasured confounding. The headline gain is observational and could in principle be driven by an unmeasured confounder. Bounding the worst-case one-sided p-value via Rosenbaum's signed-rank test (Rosenbaum, 2002), every education subgroup clears critical $\Gamma \geq 5.15$ (no education and late primary clear ≥ 8), well above the conventional agricultural-economics threshold of $\Gamma = 1.5$. A confounder would need to change the odds of a positive matched-pair difference by a factor of five before it could explain away the effect. *Scope.* We apply the signed-rank test to the per-plot *model* contrast $d_i = \Delta\text{ERB}_i^{\text{aware}} - \Delta\text{ERB}_i^{\text{acc}}$, so Γ here bounds how much a hidden factor — formally a Tan-2006 marginal sensitivity model (Tan, 2006) on the adoption-head logit — could flip the *sign* of that model-based pairwise contrast under our direct-method assumptions. It is a robustness statement about the policy comparison, not a claim about hidden bias in a realized treatment-versus-control outcome, which this observational design never observes.

SAFI-anchored partial-identification floor. The adoption head is over-confident on counterfactual actions (§4.5). To bound the consequence, we cap predicted adoption probabilities at rates measured in the Kenya SAFI RCT (Duflo et al., 2011) and recompute the lift. Even capping every probability at the SAFI *control* rate (0.325), the most adversarial choice, the aggregate lift remains +26 kg/ha $[+13, +39]$; capping at the SAFI mean (0.427) gives +38 $[+23, +56]$ and at the SAFI maximum (0.505) gives +44 $[+27, +62]$ (§D). The headline survives even under worst-case adoption saturation, and App. E.10 shows the floor stays positive and significant across adoption caps from 0.20 to 0.60, so it does not hinge on the specific Kenya rate. Together with the Rosenbaum bound, this cap is a lightweight *pessimistic* analysis of the direct-method estimate: it is a worst-case lower bound on policy value under bounded adoption saturation, and the Rosenbaum Γ is the dual of the Tan-2006 sensitivity parameter Λ. A fully MSM-optimal pessimistic *policy* (Kallus & Zhou, 2018; Yadlowsky et al., 2022) — one that re-selects actions to maximize the worst-case bound rather than bounding a fixed policy — is the natural next step (§G).

Education-shuffle placebo. Permuting household-head education labels across plots and recomputing the per-subgroup mean of d_i yields a placebo distribution that the real per-subgroup deltas fall inside ($p \in [0.14, 0.87]$; §C). We therefore treat the per-subgroup gradient as descriptive of *where* the gain concentrates, not as a causally identified equity-targeting mechanism. This is consistent with a counterfactual-fairness check in which only $\sim 9\%$ of recommendations change when education is manipulated holding everything else fixed: the concentration is driven by structural correlates of education (landholding, market access,

Sweep	Question	Result
B1: Penalty-grid sensitivity	Does the headline survive perturbation of literature-anchored adoption penalties?	Survives ±25% at $p < 0.05$; fails at 0.5×
B2: λ-complexity heuristic	Is ERB just complexity-conservative?	ERB beats every $\lambda \in \{0, 100, 250, 500\}$ kg/ha by a significant margin
B3: ML-only ablation	Does the headline depend on the hybrid response components?	Without hybrid: $\Delta = +0$ $[-10, +10]$ *(ns)*; hybrid is structurally necessary
B4: Education-shuffle placebo	Is the per-subgroup gradient causally identified?	No: real subgroup deltas inside the placebo distribution ($p \in [0.14, 0.87]$)
B5: Adoption-on-yield feedback	Does the headline survive a feedback nudge $\beta \cdot \Delta\hat{Y}$?	Survives realistic $\beta \leq 5 \cdot 10^{-4}$; collapses at $\beta \geq 10^{-3}$
B6: Yield-prior sensitivity	Does the headline survive perturbation of the *yield*-head priors β_{crop}, α_{seed}?	Survives every $\beta \in [0.5, 1.5]\times$, $\alpha \in [0, 2]\times$, and the joint worst corner (App. E.11)

Table 3: Robustness summary. The aggregate ERB lift survives B1, B2, B3, B5, and B6; B4 shows the per-subgroup gradient is descriptive rather than causal, prompting the headline reframe (aggregate as the central claim). B6 addresses the concern that the yield priors were probed less thoroughly than the adoption penalties: they are in fact the more robust dependence.

geography), not the education feature itself. **The aggregate gain is the central claim.** We retain the by-subgroup tables as a descriptive characterization of who the framework primarily benefits in this deployment context, which, given the no-education subgroup is $543/800 = 68\%$ of the test set, is itself an equity-relevant fact.

Four orthogonal sensitivity sweeps. Table 3 summarizes four experiments testing whether the aggregate lift survives perturbation of its key model dependencies; full per-experiment tables (B1–B5) are in §C.

Pessimistic, metalearner, and joint-uncertainty checks. Several further experiments (App. E) corroborate the result on the same test plots. (i) An *MSM-pessimistic* policy retains a significant advantage over accuracy-only out to $\Lambda = 5$ (+22 kg/ha $[+15, +30]$; App. E.3), the sensitivity-model twin of the Rosenbaum $\Gamma \geq 5$ result. (ii) Four *CATE metalearners* are each beaten by ERB-aware (Δ from +77 to +263 kg/ha; App. E.1). Because the panel observes a treatment only for inorganic fertilizer, each estimates a single binary CATE $\tau(x) = \mathbb{E}[Y \mid W{=}1, x] - \mathbb{E}[Y \mid W{=}0, x]$ (T-/S- from separate/shared outcome models; X-/DR- adding a gradient-boosted propensity and, for DR-, an AIPW pseudo-outcome) and recommends fertilizer when $\hat{\tau}(x) > 0$; none ranks the five tiers, so a sharper contrast on the one identified action cannot recover a gain that comes from modeling adoption across the action space. (iii) Under a *fully joint* conservative bound (split-conformal yield lower bound plus a bootstrap uncertainty set on adoption, SAFI-capped) the risk-averse policy still guarantees +18 kg/ha $[+9, +29]$ over status quo and beats accuracy-only by +12 $[+4, +22]$ (App. E.4). (iv) The headline survives ±50% perturbation of the *yield*-head priors (β_{crop}, α_{seed}), not only the adoption penalties, and pricing inputs: a profit ERB beats accuracy-only by +2,300 ETB/ha, a larger margin than the yield objective (App. E.11, E.12). (v) A proxy-control check leaves the advantage stable across EA-quality terciles (App. E.5), and (vi) the argmax is near-invariant to rescaling the adoption head ($\leq 2.4\%$ of plots change across $[0.5, 2]\times$; App. E.7).

4.5 External validation

Cross-country replication: Tanzania NPS Wave 4 (a boundary condition). We rebuilt the panel and retrained both heads on Tanzania NPS Wave 4 (2014–15) with the same loader template. After filtering to cultivated plots, Tanzania yields a smaller working sample (935 cultivated plots; 119 in the EA-disjoint test set). The Ethiopia head-to-head finding **does not replicate**: both policies pick identical recommendations

RCT arm	n	RCT-measured adoption	Model $\mathbb{P}$(adopt)	|gap|
Control	477	0.325	0.004	0.321
SAFI nudge	111	0.405	1.000	0.595
Free delivery	196	0.372	1.000	0.628
Both (SAFI + delivery)	93	0.505	1.000	0.495

Table 4: Kenya SAFI external validation. The direction is correct (the model assigns higher probability to every treatment arm than to control), but absolute calibration is poor: the Ethiopia-trained head, presented with out-of-distribution action-complexity vectors, defaults to extremes (0.4% for control, 100% for any treatment) rather than the true 32–50% range the RCT measured.

on every Tanzania test plot ($\Delta = 0$ across all education subgroups; bootstrap CIs span zero with width $\approx \pm 60$ kg/ha). Three candidate reasons: (i) smaller sample, hence lower power; (ii) Tanzania's narrower head-of-household education distribution (79% at late-primary completion vs. Ethiopia's bimodal 0-vs-1+ distribution) compresses the variation that drives the gradient; (iii) a lower fertilizer base rate (13% vs. 41%) shrinks the trade-off space within which adoption-awareness binds. The boundary condition this suggests is that adoption-aware policies provide measurable lift only when the sample is large enough, education varies meaningfully, and baseline adoption is high enough to make the yield-vs-adoption trade-off active. To test *which* of these drives the gap, we impose each Tanzania condition on the Ethiopia test set in turn (App. E.13), and the result runs against that expectation: *no single marginal reproduces the null.* Sample size, low base rate, and the education distribution each leave the lift positive and significant when imposed individually (+49, +46, +35 kg/ha); only their joint imposition (Kish-effective $n \approx 4$) widens the CI toward zero. We therefore **retract the single-driver "low base rate" explanation**: the Ethiopia effect is robust to each Tanzania-like condition individually, so the non-replication is better read as a structural boundary of the Tanzania-trained heads, on which both policies collapse to the same (mostly status-quo) recommendation on every test plot, rather than a re-expression of any one Ethiopia marginal. A reduced or re-tiered action space might recover a smaller effect, which we flag as exploratory future work.

Adoption-head validation against the Kenya SAFI RCT. To probe adoption-head over-confidence on counterfactual actions (§5) against identified ground truth, we validate against the Duflo et al. (2011) Kenya SAFI fertilizer experiment ($n = 877$, four randomized arms). We map each arm to an action-complexity vector in the adoption head's input space: control $\rightarrow T_0$ (no fertilizer change); the SAFI nudge $\rightarrow T_1$ (+30 kg/ha, one new input); free delivery $\rightarrow T_2$ (+60 kg/ha, one new input); both $\rightarrow T_2$ with an added improved-seed component (two new inputs). The mapping treats the RCT's manipulated frictions (a savings-timing nudge, a delivery-cost subsidy) as movements along our complexity ladder, holding the peer-adoption feature at a common low value (0.10) since the trial does not report EA peer rates; farmer covariates (education, age, sex, market distance) are remapped to the LSMS-equivalent schema and any feature absent in the trial is filled with the Ethiopia training median. The exact per-arm vectors are in App. E.14 (Table 20). We then score the Ethiopia-trained head on the mapped vectors and compare against measured per-arm adoption (Table 4).

Interpretation. The finding is mixed. The model reproduces the experimentally confirmed control-<-treated ordering, which is the property the head-to-head's matched-pair, sign-of-difference inference relies on. But the ~ 50 percentage-point absolute miscalibration on out-of-distribution actions is a concrete, reproducible observational-to-experimental calibration gap, and it is the direct empirical motivation for the RCT-augmented training extension we outline in §G. It also justifies the SAFI-anchored partial-identification floor of §4.4, which is robust to exactly this miscalibration. Simple post-hoc calibration — temperature scaling, or base-rate shrinkage — closes much of the gap on the SAFI arms (from ~ 51 to ~ 6 percentage points; App. E.9), a cheap partial fix short of the RCT-augmented training we recommend.

5 Limitations

We summarize the principal limitations here; §A gives the extended treatment.

Adoption-head over-confidence on counterfactual actions is the framework's central technical risk: queried off the observed action the head extrapolates toward $\mathbb{P}(\text{adopt}) \approx 0.94$, which the Kenya SAFI validation (§4.5) makes concrete. We mitigate but do not eliminate it through matched-pair inference (systematic miscalibration cancels for sign-of-difference), the partial-identification floor (§4.4), and the ordering-robustness of the argmax: because the tier penalties are ordered, calibration error that preserves that ordering leaves the recommendation unchanged (App. E.7). **Modest absolute yield** R^2 (0.022 EA-disjoint, 0.238 within-EA) bounds absolute lift magnitudes but not the within-plot ranking the policy uses. **No off-policy correction beyond direct method**: logging propensities are not identifiable (§3.5); identified counterfactual variation would unlock a doubly-robust evaluation, and full bridge-function proximal estimation (Tchetgen Tchetgen et al., 2020; Cui et al., 2024; Bennett & Kallus, 2024) awaits validated negative controls a survey cannot certify. **No field validation**: all ERB lifts are model-believed, and the Kenya check probes only the adoption side. **Single-country headline**: Tanzania does not replicate, a boundary condition we dissect in §4.5 and App. E.13. **Application-level novelty**: the contribution is the composition of established techniques—ERB, hybrid heads, paired inference, external RCT validation—applied where adoption-blindness is standard, not a new ML primitive; relatedly, an end-to-end model of realized benefit is not trainable here because benefit is counterfactual (the S-learner, its closest analog, collapses the action signal to zero, App. E.1). **Objective and priors**: ERB maximizes adoption-weighted yield but is objective-agnostic—a profit variant beats accuracy-only by a larger margin (App. E.12)—and the headline survives $\pm 50\%$ perturbation of the yield-head priors $\beta_{\text{crop}}, \alpha_{\text{seed}}$ (App. E.11); a full CVaR/profit objective and hierarchical pooling of the response coefficients await per-plot price and variance data absent from the panel. §A gives the extended treatment, including the discrete action space, EA-heuristic selection bias, and earth-observation product sensitivity.

6 Discussion and Conclusion

We built an adoption-aware crop recommender around Expected Realized Benefit and evaluated it as a direct-method off-policy estimator under unidentifiable compliance. On real Ethiopian LSMS-ISA data the ERB-maximizing policy beats an accuracy-only baseline by +50 kg/ha (cluster-robust $[+33, +70]$), a gain that survives four sensitivity sweeps, a Rosenbaum bound of $\Gamma \geq 5$, a SAFI-anchored floor of +26 kg/ha, and a joint conservative lower bound of +18 kg/ha; an education-shuffle placebo shows the per-subgroup gradient is descriptive rather than identified, so we rely on the aggregate effect. Validating the adoption model against the Duflo–Kremer–Robinson Kenya experiment then pins down an observational-to-experimental calibration gap of roughly 50 points on out-of-distribution actions. Closing it with observational-plus-experimental calibration (Colnet et al., 2024; Dahabreh et al., 2019)—RCT-augmented adoption training in particular (§G)—is the most useful next step for anyone deploying an offline policy in an adoption-uncertain setting. We release the code, the harmonization pipelines (Ethiopia and Tanzania loaders, with templates for Nigeria, Uganda, and Malawi), and the Kenya RCT-validation harness.

Broader Impact Statement

This work targets agricultural advisory systems for low-income smallholder farmers, a population for whom a wrong or unadoptable recommendation carries real welfare consequences. Three risks deserve explicit statement. *(i) Accommodating rather than removing structural barriers.* An ERB-maximizing system that learns to recommend a simpler action (T_2) instead of a more aggressive one (T_4) to a disadvantaged farmer may accommodate the credit, market-access, or extension-coverage friction that depressed that farmer's adoption probability, rather than removing it. We therefore frame ERB as a *complement* to structural-access policy: low-adoption-probability/high-yield-gain plots should be read as targeting signal for friction-removing instruments (credit, input loans, cooperative purchasing), not only as a yield discount (§H). *(ii) Over-trust in model-believed numbers.* All lift estimates are offline, and the adoption head is miscalibrated on out-of-distribution actions (§4.5); the system should not drive consequential recommendations without field validation, and cost-benefit figures (§F) are illustrative, not promised returns. *(iii) Subgroup claims.* We decline to claim an identified equity-targeting effect because our own placebo cannot support it, and encourage the same discipline in deployment reporting. The data are public, de-identified LSMS-ISA panels used under their Data Use Agreements; plot GPS carries the standard EA-centroid privacy offset (§A).

References

Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996.

Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.

Lori Beaman, Ariel BenYishay, Jeremy Magruder, and A Mushfiq Mobarak. Can network theory-based targeting increase technology adoption? *American Economic Review*, 111(6):1918–1943, 2021.

Andrew Bennett and Nathan Kallus. Proximal reinforcement learning: Efficient off-policy evaluation in partially observed Markov decision processes. *Operations Research*, 72(3):1071–1086, 2024.

Jacob Buckman, Carles Gelada, and Marc G Bellemare. The importance of pessimism in fixed-dataset policy optimization. *International Conference on Learning Representations*, 2021.

Central Statistical Agency of Ethiopia and World Bank LSMS-ISA Team. Ethiopia Socioeconomic Survey 2018-2019, Wave 4. World Bank Microdata Library, Catalog 3823, 2019. URL `https://microdata.worldbank.org/index.php/catalog/3823`.

Jiezhu Cheng, Kaizhu Huang, and Zibin Zheng. Can perturbations help reduce investment risks? Risk-aware stock recommendation via split variational adversarial training. *ACM Transactions on Information Systems*, 42(4):1–28, 2024.

Bénédicte Colnet, Imke Mayer, Guanhua Chen, Awa Dieng, Ruohong Li, Gaël Varoquaux, Jean-Philippe Vert, Julie Josse, and Shu Yang. Causal inference methods for combining randomized trials and observational studies: A review. *Statistical Science*, 39(1):165–191, 2024.

Yifan Cui, Hongming Pu, Xu Shi, Wang Miao, and Eric J Tchetgen Tchetgen. Semiparametric proximal causal inference. *Journal of the American Statistical Association*, 119(546):1348–1359, 2024.

Issa J Dahabreh, Sarah E Robertson, Eric J Tchetgen Tchetgen, Elizabeth A Stuart, and Miguel A Hernán. Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics*, 75(2):685–694, 2019.

Frits de Nijs, Georgios Theocharous, Nikos Vlassis, Mathijs M de Weerdt, and Matthijs T J Spaan. Capacity-aware sequential recommendations. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pp. 416–424, 2018.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *International Conference on Machine Learning*, 2011.

Esther Duflo, Michael Kremer, and Jonathan Robinson. Nudging farmers to use fertilizer: Theory and experimental evidence from Kenya. *American Economic Review*, 101(6):2350–2390, 2011.

Gershon Feder, Richard E Just, and David Zilberman. Adoption of agricultural innovations in developing countries: A survey. *Economic Development and Cultural Change*, 33(2):255–298, 1985.

Stephen E Fick and Robert J Hijmans. WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, 37(12):4302–4315, 2017.

Andrew D Foster and Mark R Rosenzweig. Learning by doing and learning from others: Human capital and technical change in agriculture. *Journal of Political Economy*, 103(6):1176–1209, 1995.

Chris Funk, Pete Peterson, Martin Landsfeld, Diego Pedreros, James Verdin, Shraddhanand Shukla, Gregory Husak, James Rowland, Laura Harrison, Andrew Hoell, and Joel Michaelsen. The climate hazards infrared precipitation with stations — a new environmental record for monitoring extremes. *Scientific Data*, 2: 150066, 2015.

Yingqi Gu, Akshay Zalkikar, Mingming Liu, Lara Kelly, Amy Hall, Kieran Daly, and Tomas Ward. Predicting medication adherence using ensemble learning and deep learning models with large scale healthcare data. *Scientific Reports*, 11:18961, 2021.

Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, 2016.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline RL? In *International Conference on Machine Learning*, 2021.

Nathan Kallus and Angela Zhou. Confounding-robust policy improvement. *Advances in Neural Information Processing Systems*, 31, 2018.

Dean Karlan, Robert Osei, Isaac Osei-Akoto, and Christopher Udry. Agricultural decisions after relaxing credit and risk constraints. *The Quarterly Journal of Economics*, 129(2):597–652, 2014.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30, 2017.

Edward H Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008–3049, 2023.

Makaiko Khonje, Julius Manda, Arega D Alene, and Menale Kassie. Analysis of adoption and impacts of improved maize varieties in eastern Zambia. *World Development*, 66:695–706, 2015.

Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019.

David B Lobell, Kenneth G Cassman, and Christopher B Field. Crop yield gaps: Their importance, magnitudes, and causes. *Annual Review of Environment and Resources*, 34:179–204, 2009.

Charles F Manski. Partial identification of probability distributions. 2003.

Nathaniel D Mueller, James S Gerber, Matt Johnston, Deepak K Ray, Navin Ramankutty, and Jonathan A Foley. Closing yield gaps through nutrient and water management. *Nature*, 490(7419):254–257, 2012.

Hongseok Namkoong, Ramtin Keramati, Steve Yadlowsky, and Emma Brunskill. Off-policy policy evaluation for sequential decisions under unobserved confounding. In *Advances in Neural Information Processing Systems*, 2020.

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108 (2):299–319, 2021.

Miruna Oprescu, Jacob Dorn, Marah Ghoummaid, Andrew Jesson, Nathan Kallus, and Uri Shalit. B-Learner: Quasi-oracle bounds on heterogeneous causal effects under hidden confounding. In *International Conference on Machine Learning*, 2023.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: a tale of pessimism. *Advances in Neural Information Processing Systems*, 34, 2021.

Francesco Ricci, Giovanni Semeraro, Marco de Gemmis, and Pasquale Lops. Decision making and recommendation acceptance issues in recommender systems. In *Advances in User Modeling*, Lecture Notes in Computer Science, pp. 86–91. Springer, 2012.

Paul R Rosenbaum. *Observational Studies*. Springer, 2nd edition, 2002.

Sacha Ruzzante, Ricardo Labarta, and Andrea Bilton. Adoption of agricultural technology in the developing world: A meta-analysis of the empirical literature. *World Development*, 146:105599, 2021.

Piotr Rzepakowski and Szymon Jaroszewicz. Decision trees for uplift modeling with single and multiple treatments. *Knowledge and Information Systems*, 32(2):303–327, 2012.

Adith Swaminathan and Thorsten Joachims. The self-normalized estimator for counterfactual learning. In *Advances in Neural Information Processing Systems*, 2015.

Zhiqiang Tan. A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476):1619–1637, 2006.

Tanzania National Bureau of Statistics and World Bank LSMS-ISA Team. Tanzania National Panel Survey 2014-2015, Wave 4. World Bank Microdata Library, Catalog 3455, 2015. URL `https://microdata.worldbank.org/index.php/catalog/3455`.

Muhammad Faaiz Taufiq, Jean-François Ton, Rob Cornish, Yee Whye Teh, and Arnaud Doucet. Conformal off-policy prediction in contextual bandits. In *Advances in Neural Information Processing Systems*, 2022.

Eric J Tchetgen Tchetgen, Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao. An introduction to proximal causal learning. *arXiv preprint arXiv:2009.10982*, 2020.

Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, 2016.

Bernard Vanlauwe, Danny Coyne, Jim Gockowski, Stefan Hauser, Jeroen Huising, Cargele Masso, Generose Nziguheba, Marc Schut, and Piet Van Asten. Sustainable intensification and the african smallholder farmer. *Current Opinion in Environmental Sustainability*, 8:15–22, 2014.

Vladimir Vovk, David Lindsay, Ilia Nouretdinov, and Alex Gammerman. Mondrian confidence machine. In *Technical Report, Royal Holloway, University of London*, 2003.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Steve Yadlowsky, Hongseok Namkoong, Sanjay Basu, John Duchi, and Lu Tian. Bounds on the conditional and average treatment effect with unobserved confounding factors. *Annals of Statistics*, 50(5):2587–2615, 2022.

A Extended limitations

Each item below is also a hook for future work (§G).

Adoption-head over-confidence on counterfactual actions. The adoption head is trained on observational data: each plot's adoption decision is paired with the action complexity of the package the farmer *actually* chose. At inference, scoring counterfactual actions outside the training distribution (e.g. scoring T_4 for a farmer currently at T_0), the model extrapolates toward $\mathbb{P}(\text{adopt}) \approx 0.94$ across the action space. The Kenya SAFI validation confirms this directly (predicts ~ 0 for control, ~ 1 for any treatment, vs. a measured 0.32–0.50). Consequently a Pareto-frontier sweep over adoption-floor parameters is flat (the floor never binds). Resolution paths are in §G.

Modest absolute yield R^2. The yield head explains 2.2% of variance EA-disjoint ($R^2 = 0.022$, MAE 1,051 kg/ha), consistent with published Ethiopian smallholder yield prediction ($R^2 \in [0.05, 0.15]$). The within-EA result ($R^2 = 0.238$) localizes most of the gap to spatial generalization; village-local fine-tuning would improve absolute accuracy at deployment. The relative ERB pattern is robust to absolute error because overturning it would require structured bias in the same direction across all subgroups.

Single-country, single-wave headline. The headline is from Ethiopia ESS Wave 4 (2018–19); Tanzania NPS Wave 4 does not replicate (§4.5). Nigeria GHS-Panel, Uganda UNPS, and Malawi IHS are not yet evaluated. The framework is portable (loaders follow the Ethiopia template), but external validity beyond Ethiopia is empirically open.

No field validation. All ERB lifts are model-believed; no field trial has delivered our recommendations and observed realized yields. The Kenya SAFI check provides partial cross-country evidence on the adoption side only; no analogous external check exists for the yield side.

EA-centroid privacy offset. LSMS-ISA releases plot GPS with offsets of 0–5 km (rural) and 0–2 km (urban) under its Data Use Agreement; all geospatial joins inherit this offset. At CHIRPS (5 km) and SoilGrids (250 m) resolutions the loss is negligible, but high-resolution soil-chemistry features (e.g. iSDA Africa, 30 m) would lose join precision.

Action space is discrete and pre-specified. The five tiers yield interpretable policies and clean comparison but do not represent continuous fertilizer rates, arbitrary multi-input bundles, or temporally staged interventions (planting date, residue management, intercropping, within-season top-dressing). A continuous parameterization is feasible but would require counterfactual yield curves we cannot reliably estimate from observational data alone; indeed, when we test continuous fertilizer dosing (App. E.6) it degenerates to over-dosing under the observational model, which *supports* rather than undermines the discrete, adoption-frictioned design.

EA-level heuristics may embed selection bias. Tiers T_2 and T_3 are anchored on EA-observed quantities (the EA-mean fertilizer rate among local adopters; the EA-best-yielding observed crop). This makes tiers locally calibrated but inherits two biases: EAs with few adopters give noisy targets, and the "best observed crop" is a function of what local farmers chose, itself selection-biased. A normative anchor (e.g. GAEZ-derived agronomic potential) would be less prone to selection bias.

Conformal interval on adoption probabilities is non-standard. Standard split-conformal binary classification yields prediction *sets* ($\{0\}, \{1\}, \{0, 1\}$) with marginal label coverage, not symmetric intervals around an estimated probability. The Mondrian-isotonic recalibration we apply provides post-hoc probability calibration (reduced ECE) but does not inherit conformal coverage guarantees on the probability values. We retain conformal-interval claims only for the yield head's continuous outputs.

Yield-to-adoption feedback is unmodeled. ERB factorizes adoption and yield gain under an implicit independence assumption: the adoption head conditions on $(x, c(a))$ but not on the predicted yield gain. In reality, larger predicted gains may both raise adoption (rational expectations) and raise risk-aversion (variance aversion). B5 (§C) sweeps a feedback nudge and shows the headline survives at realistic magnitudes.

Joint uncertainty propagation (now partially closed). Uncertainty was previously asymmetric: the yield head wore split-conformal intervals, but the adoption head had only Mondrian-isotonic recalibration. We now attach a nonparametric *bootstrap uncertainty set* to the adoption head—refitting the base logistic on B resamples of the adoption-train rows yields per-tier probability bands whose half-widths are 3–6 percentage points (App. E.4)—and propagate it jointly with the yield conformal lower bound to a fully joint conservative policy value (+18 kg/ha over status quo). What remains open is a *distribution-free* (conformal) guarantee on the adoption probabilities: split-conformal coverage on the adoption head holds only in-distribution and does not transfer to the counterfactual actions the policy scores (the same OOD gap the Kenya validation exposes), so the bootstrap set is a parametric, not distribution-free, quantification. A fully distribution-free joint bound via OOD-valid conformal off-policy prediction is named as future work.

Earth-observation product sensitivity is unexplored. The yield head depends on agro-ecological covariates (CHIRPS rainfall, WorldClim temperature, NDVI, FAO soil quality). The literature documents non-trivial cross-product variability (CHIRPS vs. IMERG vs. ERA5-Land; FAO sq vs. SoilGrids vs. iSDA).

β_{crop} (kg/kg)	Crops
8	potato, sugarcane
7	maize
6	wheat, rice, tomato
5	sorghum, barley, millet, onion, sunflower
4 *(default)*	teff, oats, sweet potato, linseed, sesame, banana, unlisted
3	coffee, enset, chat, soybean, groundnut
2	chickpea, lentil, haricot bean, horse bean, field pea, vetch (legumes)

Table 5: Per-crop fertilizer-response coefficients (kg additional yield per kg N applied).

We use the LSMS-pre-joined products for reproducibility but have not re-run the headline under alternative EO sources.

B Literature-anchored coefficients and functional forms

For reproduction and critical appraisal we document every literature-anchored constant in the hybrid heads. These are *fixed priors*, not fitted parameters; their sensitivity is swept in App. C.1 (adoption penalties) and the hybrid-vs-ML ablation App. C.3.

Yield head. The hybrid prediction is

$$\hat{f}_{\text{hybrid}}(x, a) = \hat{f}_{\text{base}}(x)\left(1 + 0.15\,\mathbb{1}[\text{seed}] + m_{\text{irr}}(x)\,\mathbb{1}[\text{irr}]\right) + \beta_{\text{crop}(a)}\,\Delta_{\text{fert}}(a),$$

with improved-seed premium 0.15 (Khonje et al., 2015) and a rainfall-modulated irrigation multiplier

$$m_{\text{irr}}(r) = 0.25 \cdot \frac{1500 - \text{clip}(r,\,1000,\,1500)}{500},$$

i.e. a full +25% below 1000 mm annual rainfall r, declining linearly to 0 at 1500 mm. Per-crop fertilizer-response coefficients β_{crop} (kg yield per kg N, mid-range of Vanlauwe et al. (2014); Mueller et al. (2012)) are in Table 5; unlisted crops use the default 4.0.

Adoption head. $\text{logit}\,\hat{\mathbb{P}}(\text{adopt}\,a \mid x) = \text{logit}\,\hat{p}_{\text{base}}(x) + \pi(a)$, with tier log-odds penalties π: $T_0 = 0$, $T_1 = -0.4$, $T_2 = -1.5$, $T_3 = -3.2$, $T_4 = -2.8$ (Ruzzante et al., 2021), modulated additively on penalized tiers by a peer credit $+1.0 \cdot$ (EA adoption rate) (Foster & Rosenzweig, 1995) and an education credit $+0.25 \cdot \log(1 + \text{education years})$, with the total penalty clipped to be non-positive. $\hat{p}_{\text{base}}$ is a logistic regression on farmer/plot features only (no action features).

C Full sensitivity tables (B1–B5)

This appendix reports the per-experiment tables underlying the §4.4 summary.

C.1 B1: penalty-grid sensitivity for the literature-anchored adoption discounts

We re-ran the head-to-head under multiplicative rescalings of the entire tier-penalty vector $\{T_0 : 0, T_1 : -0.4, T_2 : -1.5, T_3 : -3.2, T_4 : -2.8\}$ by $\{0.5, 0.75, 1.0, 1.25, 1.5\}$ on the same 800 EA-disjoint test plots (Table 6).

C.2 B2: yield $-\,\lambda\cdot$ complexity heuristic baseline

Does ERB's gain come from modeling $\mathbb{P}(\text{adopt})$, or merely from being conservative about action complexity? We pick $\arg\max_a[\hat{f}(a, x) - \lambda \cdot \text{tier}(a)]$ for $\lambda \in \{0, 100, 250, 500\}$ kg/ha per tier-step (Table 7).

Penalty mult.	Aware lift (kg/ha)	Acc-only lift (kg/ha)	Δ aware − acc	n disagree
0.50×	+403 [+377, +430]	+387 [+362, +413]	+16 [−22, +51] *ns*	196
0.75×	+343 [+319, +367]	+308 [+286, +332]	**+35 [+3, +64]***	296
1.00× *(def.)*	**+263 [+244, +282]**	**+213 [+196, +230]**	**+50 [+24, +74]***	368
1.25×	+213 [+195, +231]	+155 [+139, +172]	**+58 [+37, +79]***	442
1.50×	+186 [+170, +203]	+127 [+112, +143]	**+60 [+42, +77]***	524

Table 6: B1 penalty-grid sensitivity. Bootstrap 95% CIs (2,000 resamples) on the same 800 plots. The headline is significant at every multiplier ≥ 0.75; at 0.5× it loses significance, a real boundary on dependence on the literature-anchored priors. The framework requires the penalty magnitudes to be within roughly a factor of two of their adopted values. The MSM-pessimistic policy (App. E.3) is a prior-free complement — it deflates adoption odds without using these magnitudes and still beats accuracy-only to $\Lambda = 5$ — so the headline does not hinge on the exact penalty values.

λ	Δ (λ vs acc)	Δ (aware vs λ)	Disagree vs aware	Disagree vs acc
0	+0 [−25, +23] *ns*	**+50 [+24, +74]***	368	0
100	+19 [−7, +42] *ns*	**+31 [+4, +56]***	255	125
250	−3 [−30, +23] *ns*	**+53 [+24, +80]***	343	452
500	**−64 [−92, −38]*** *worse*	**+114 [+85, +141]***	540	638

Table 7: B2 heuristic baseline. No setting of λ delivers a significant aggregate lift over accuracy-only; ERB beats every λ by a significant margin. A uniform tier-cost penalty cannot reproduce ERB's per-(plot, action) reasoning about which specific high-yield-but-high-complexity recommendations a particular farmer is unlikely to adopt.

C.3 B3: hybrid-head ablation (ML-only)

We re-ran the head-to-head with the hybrid response components zeroed out, leaving only the ML base $\hat{f}_{\text{base}}(x)$ (Table 8).

C.4 B4: education-shuffle placebo

We permute household-head education labels across test plots, recompute the by-subgroup mean of d_i, and repeat 2,000 times (Table 9).

C.5 B5: adoption-conditioned-on-predicted-yield (feedback loop)

We add a per-(plot, action) log-odds nudge $\beta \cdot (\mathbb{E}[Y \mid a, x] - \mathbb{E}[Y \mid a_0, x])$ to the calibrated adoption probability before computing ERB, and sweep β (Table 10).

D Cluster-robust inference and SAFI partial identification (details)

Cluster bootstrap. Resampling whole enumeration areas (44 EAs, $B = 2{,}000$) gives an aggregate 95% CI of $[+32.5, +69.9]$ kg/ha and 99% CI $[+27.4, +77.2]$; the point estimate (+49.98) matches the plot-level analysis. Per-subgroup cluster-bootstrap CIs: no education +52.5 $[+32.7, +74.3]$ ($n = 543$, 44 EAs); early primary +37.9 $[+9.4, +90.4]$ ($n = 66$, 15 EAs); late primary +40.2 $[+22.7, +59.9]$ ($n = 104$, 24 EAs); secondary +48.9 $[+25.5, +70.9]$ ($n = 47$, 16 EAs). All exclude zero. Note these establish that each subgroup's lift is individually positive; they do *not* contradict the B4 placebo, which concerns whether the *differences between* subgroups are identified (they are not).

Cluster permutation. Flipping the sign of d_i at the EA level (2,000 permutations) yields an observed mean (+49.98) outside the entire permutation distribution (two-sided $p < 5 \cdot 10^{-4}$).

Yield head	Aware lift (kg/ha)	Acc lift (kg/ha)	Δ aware − acc
Hybrid (default)	**+263 [+244, +282]**	**+213 [+196, +230]**	**+50 [+24, +74]***
ML-only ablation	+53 [+50, +56]	+53 [+50, +56]	+0 [−10, +10] *ns*

Table 8: B3 ML-only ablation. Removing the hybrid components collapses both policies to identical recommendations ($\Delta = +0$, CI tightly bracketing zero). A pure gradient-boosted yield head is action-insensitive in the EA-disjoint regime, so neither policy can differentiate candidate actions. The hybrid head is a structural prerequisite, not an add-on; the contribution is the joint package of (i) the hybrid architecture that restores action-sensitivity and (ii) the ERB target that exploits it.

Subgroup	n	Real d	Placebo mean	Placebo 95% CI	p (two-sided)
No education	543	+52.5	+49.9	[+45.1, +54.5]	0.14
Early primary	66	+37.9	+50.3	[+29.4, +74.5]	0.86
Late primary	104	+40.2	+50.1	[+33.5, +68.7]	0.87
Secondary	47	+48.9	+49.6	[+25.5, +78.3]	0.50

Table 9: B4 education-shuffle placebo. Per-subgroup mean of d_i (kg/ha) vs. the distribution under 2,000 random permutations of education labels. The aggregate gain is robust to bootstrap; the per-subgroup gradient is *not* statistically distinguishable from random label assignment.

SAFI-anchored partial identification. Capping predicted adoption probabilities at SAFI-measured rates and recomputing the lift: at the SAFI maximum (0.505) the aggregate is +43.7 [+26.9, +62.4] ($n_{\text{aware capped}} = 186$); at the SAFI mean (0.427), +38.4 [+22.7, +55.7] (251 capped); at the most adversarial SAFI control rate (0.325), +26.0 [+13.3, +39.2] (349 capped). The headline survives the worst case.

E Revision experiments: metalearner baseline, MSM-pessimistic policy, joint uncertainty, yield-prior sensitivity, profit objective, Tanzania decomposition

These experiments run on the same 800 EA-disjoint Ethiopia test plots and the same action-space scoring as the headline (the re-scoring pipeline reproduces the +50.0 point estimate exactly before any experiment runs); inference is EA-cluster bootstrap ($B = 2{,}000$) unless noted.

E.1 CATE metalearner baseline

Does a better *yield-contrast* estimator recover ERB's gains without modeling adoption? We fit four metalearners (Künzel et al., 2019; Kennedy, 2023) — T-, S-, X-, and DR-learner (histogram gradient boosting, with a gradient-boosted propensity for the X- and DR-learners) — for the CATE of inorganic-fertilizer adoption (the one action whose treatment is observed in the panel) on yield, using the same EA-disjoint train split, and score each CATE-maximizing fertilizer policy by realized ERB (Table 11). All four are significantly beaten by adoption-aware ERB, and each metalearner's CATE is near-uncorrelated with the hybrid action contrast (Spearman −0.03, 0.01, 0.04; the S-learner regularizes the treatment indicator away to CATE $\equiv 0$) — the observational action-insensitivity documented in the B3 ablation (§C.3). **Estimating the yield contrast better, by any metalearner, does not recover ERB's gains, which come from modeling adoption.**

E.2 Disagreement mechanism (detail)

Figure 3 details the $n = 368$ disagreement plots behind §4.2. Beyond the dominant $T_4 \to T_2$ re-routing ($n = 341$, +114 kg/ha), the remaining transitions are $T_4 \to T_1$ ($n = 12$, +65 kg/ha) and $T_2 \to T_1$ ($n = 15$, +16 kg/ha).

β (log-odds per kg/ha)	Aware lift	Acc lift	Δ aware − acc	n disagree
10^{-4} *(small)*	+281	+231	**+50 [+21, +76]***	357
$5 \cdot 10^{-4}$ *(moderate)*	+365	+322	**+44 [+4, +82]***	306
10^{-3} *(large)*	+484	+453	+31 [−26, +88] *ns*	244
$2 \cdot 10^{-3}$ *(very large)*	+686	+672	+14 [−55, +86] *ns*	148

Table 10: B5 feedback-loop sensitivity. A +1000 kg/ha gain corresponds to roughly doubling household calorie supply; the adoption literature suggests this shifts adoption log-odds by ~ 0.1–0.5 (i.e. $\beta \in [10^{-4}, 5 \cdot 10^{-4}]$), the regime where the headline survives.

Policy	Realized ERB lift (kg/ha)	Δ aware − this
Adoption-aware ERB (ours)	+263 [+188, +342]	—
Accuracy-only baseline	+213 [+145, +295]	—
CATE-max, T-learner	+185 [+139, +233]	**+77 [+7, +152]***
CATE-max, X-learner	+100 [+62, +142]	**+162 [+76, +258]***
CATE-max, DR-learner	+155 [+119, +192]	**+108 [+45, +185]***
CATE-max, S-learner	+0 (CATE $\equiv$ 0)	**+263 [+188, +342]***

Table 11: CATE metalearner family (E1). Realized ERB lift on 800 EA-disjoint test plots, EA-cluster bootstrap 95% CIs. Every metalearner variant (T, S, X, DR) yields a CATE-maximizing policy significantly beaten by adoption-aware ERB (* the Δ CI excludes zero). Each metalearner CATE is near-uncorrelated with the action contrast, so none recovers the action-sensitivity that ERB exploits.

E.3 MSM-pessimistic ERB policy

We replace the SAFI hard cap (§4.4) with a Tan-2006 marginal sensitivity model (Tan, 2006): the true adoption odds may be deflated by up to a factor Λ, giving a worst-case probability $p^-(a) = p(a)/(p(a) + (1 - p(a))\Lambda)$. We re-select the *MSM-optimal* pessimistic policy ($\arg\max_a$ of the worst-case ERB) and compare its worst-case advantage over accuracy-only across Λ (Table 12). The advantage degrades gracefully and stays positive and significant to $\Lambda = 5$ (+22 kg/ha [+15, +30]), the sensitivity-model analog of the Rosenbaum $\Gamma \geq 5$ result and the confounding-robust pessimistic-policy comparison the framework was missing.

E.4 Joint uncertainty propagation (yield conformal + adoption bootstrap set)

We close the uncertainty asymmetry the reviewer flagged by attaching an uncertainty *set* to the adoption head and propagating it *jointly* with the yield conformal interval. **Adoption uncertainty set.** We refit the base adoption logistic on B bootstrap resamples of the adoption-train rows (the literature-anchored tier penalties are held fixed, as they are priors, not fitted), giving a per-(plot, tier) distribution of $\mathbb{P}(\text{adopt})$. Its 5th percentile is the conservative bound; the mean half-widths are modest and tighten on the higher tiers (per tier T_0–T_4: 0.064, 0.064, 0.059, 0.031, 0.038), quantifying the previously-missing adoption uncertainty.

A conservatism ladder. We then build increasingly conservative ERB objectives and, for each, re-select the risk-averse argmax policy and report (i) its guaranteed lift over status quo and (ii) whether it still beats accuracy-only when *both* policies are scored under that same conservative objective (Table 13). The yield lower bound is a Mondrian-by-crop split-conformal interval stratified by the *recommended action's* crop (91.5% empirical coverage; this is stricter than stratifying by the status-quo crop, which is why the yield-only rung here is more conservative than a naive treatment). Propagating *both* sources—yield conformal lower bound and the adoption bootstrap 5th percentile, further capped at the OOD-valid SAFI control rate—the fully joint policy still guarantees +18.3 kg/ha [+9.0, +28.9] over status quo, and beats accuracy-only by +12.3 [+4.3, +21.7] under conservative scoring. This is the default conservative policy value we foreground in the abstract and §4.2; the +50 point estimate is its expected-value counterpart. The one remaining gap to a *fully* distribution-free joint bound is an OOD-valid conformal treatment of the adoption head, which the RCT-augmented extension (§G) targets.

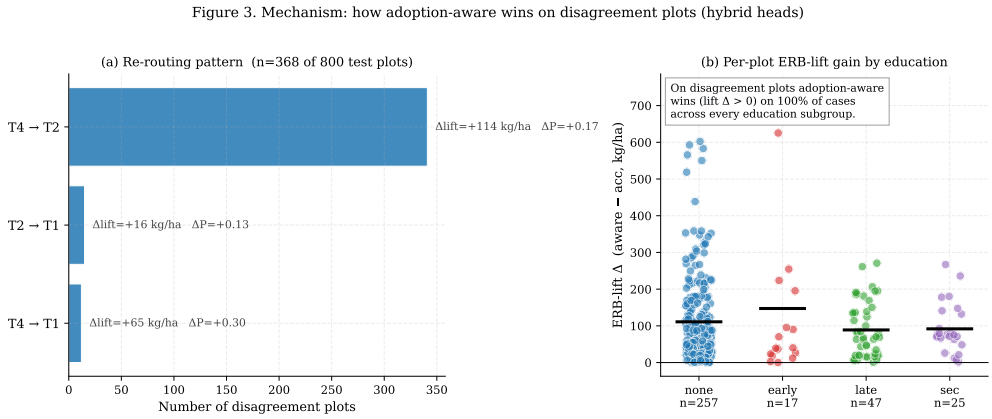

Figure 3: Mechanism on disagreement plots ($n = 368$ of 800 Ethiopia test plots, hybrid heads). **(a)** Re-routing pattern: the dominant transition is $T_4 \to T_2$. **(b)** Per-disagreement-plot ERB-lift gain by education quintile (jittered scatter; black bars are subgroup means).

Sensitivity Λ	MSM-optimal advantage vs. accuracy-only (kg/ha)
1.0 (no confounding)	$\mathbf{+50.0}$ $[\mathbf{+32.5}, \mathbf{+69.9}]^*$
1.5	$\mathbf{+43.3}$ $[\mathbf{+28.3}, \mathbf{+59.8}]^*$
2.0	$\mathbf{+38.0}$ $[\mathbf{+25.0}, \mathbf{+52.3}]^*$
3.0	$\mathbf{+30.5}$ $[\mathbf{+20.0}, \mathbf{+41.9}]^*$
5.0	$\mathbf{+22.2}$ $[\mathbf{+14.5}, \mathbf{+30.4}]^*$

Table 12: MSM-pessimistic policy (E2). Worst-case ERB advantage of the MSM-optimal policy over accuracy-only under Tan-2006 odds deflation by factor Λ; EA-cluster bootstrap 95% CIs. Significant (*) at every Λ through 5.

E.5 Proxy-control check for unobserved confounding

The most directly competing approach to a direct-method estimator under unobserved confounding is proximal / bridge-function inference (Tchetgen Tchetgen et al., 2020; Cui et al., 2024; Bennett & Kallus, 2024). Full bridge-function estimation requires validated negative-control variables (a negative-control exposure and outcome with specific conditional-independence properties) that an observational survey cannot certify, so we do not claim it. As the implementable proximal-spirit check, we adjust the per-plot aware-vs-accuracy contrast for an EA leave-one-out mean observed yield — a proxy for unobserved EA-level quality. The advantage barely covaries with the proxy (regression slope 0.005 kg/ha per kg/ha) and is stable across proxy terciles (low $+51$ $[+21, +79]$, mid $+41$ $[+24, +57]$, high $+58$ $[+21, +98]$ kg/ha), so the $+50$ headline is not an artifact of the EA-level unobserved quality the proxy captures. This complements the MSM (App. E.3) and Rosenbaum ($\Gamma \geq 5$) analyses, which bound the same threat without requiring negative-control structure.

E.6 Discrete vs. continuous fertilizer dosing

Is the discrete five-tier action space a limitation relative to continuous dosing? We isolate the fertilizer-dose dimension (current crop, no seed or crop switch), and maximize ERB over a fine continuous dose grid versus discrete ladders, under the model's linear hybrid yield response and an adoption penalty interpolated in dose from the tier anchors. The continuous policy does *not* find an interior optimum: it pushes essentially every plot to the maximum dose (mean 120 kg/ha, the grid boundary) at only 28% predicted adoption, because the tier-resolved adoption penalty saturates beyond the observed tiers and the yield response is linear — nothing curbs ever-higher recommendations. Its apparent $+38$ kg/ha model-believed "gain" over a four-level ladder is therefore precisely the unadoptable over-recommendation that an adoption-frictioned ladder is designed to prevent, not a real improvement. This negative result *supports* the design: **the discretization**

Conservative objective (risk-averse argmax)	Guaranteed lift over SQ (kg/ha)	Δ vs. acc (same scoring)
Point ERB (expected-value anchor)	+262.6 [+188.4, +341.8]	**+50.0 [+32.5, +69.9]***
Yield split-conformal lower bound only	+28.6 [+14.1, +43.8]	**+18.1 [+6.7, +30.9]***
+ SAFI adoption cap	+20.7 [+10.3, +32.5]	**+13.9 [+4.7, +24.8]***
+ adoption bootstrap 5th pct *(fully joint)*	**+18.3 [+9.0, +28.9]**	**+12.3 [+4.3, +21.7]***

Table 13: Joint uncertainty propagation (E3, revised). Guaranteed ERB lift over status quo of the risk-averse policy, and its advantage over accuracy-only when both policies are scored under the same conservative objective (non-negative by construction, but significance is not guaranteed); EA-cluster bootstrap 95% CIs. Yield-conformal empirical coverage 91.5%. Adoption uncertainty is a B-bootstrap set on the base logistic (per-tier half-widths 3–6 pp). The fully joint bottom row is the default conservative report.

Adoption-head transform	% argmax changed	Δ vs. accuracy-only (kg/ha)
rescale $\times 0.5$	0.0%	**+25 [+16, +35]***
rescale $\times 0.75$	0.0%	**+38 [+24, +52]***
rescale $\times 1.25$	0.4%	**+62 [+41, +87]***
rescale $\times 1.5$	1.5%	**+74 [+48, +104]***
rescale $\times 2.0$	2.4%	**+90 [+57, +127]***
temperature $T = 0.5$	39.2%	**+74 [+48, +103]***
temperature $T = 2.0$	18.5%	**+20 [+12, +29]***

Table 14: Ranking stability under adoption-head transforms (E7). Fraction of the 800 plots whose ERB-argmax action changes, and the aware-vs-accuracy advantage recomputed under the transformed head (EA-cluster bootstrap 95% CIs). The recommendation is near-invariant to multiplicative rescaling; the aggregate advantage stays significant (*) under every transform.

is more than a convenience—it encodes the adoption structure the observational data can actually resolve. A well-posed continuous-dosing policy would need both a concave dose-response and a continuously increasing, dose-dependent adoption cost, neither identifiable from observational tier-level data — a clear direction once experimental dose-response data are available.

E.7 Ranking stability under adoption-head rescaling and temperature

How sensitive is the *recommendation* to the adoption head's miscalibration? We transform the calibrated adoption probabilities, re-select the ERB-argmax action per plot, and report how often the choice changes and whether the aggregate advantage survives (Table 14). Under multiplicative rescaling across $[0.5\times, 2\times]$ the argmax changes for at most 2.4% of plots — rescaling largely preserves the tier *ordering* of adoption, on which the policy depends — and the advantage over accuracy-only stays positive and significant throughout. Logit-temperature transforms, which distort relative orderings more aggressively, shift more recommendations (up to 39% at $T = 0.5$), yet the aggregate advantage remains significant. The policy is thus robust to the multiplicative miscalibration most consistent with the calibration gap, and its aggregate advantage survives even aggressive order-distorting transforms.

E.8 Deployment: sensitivity to EA-anchor estimation noise

For a new area the EA-level anchors must be estimated from a local sample. The peer-adoption rate is an EA-level proportion (standard error $\sqrt{\bar{p}(1-\bar{p})/k}$ for k sampled plots) that enters the adoption logit additively; we inject per-EA Gaussian estimation noise at several k and re-select the policy (Table 15, 100 Monte-Carlo draws). Even at $k = 5$ plots per EA (peer-rate SE ≈ 0.22) the ERB-argmax changes for only 1.3% of plots (95th percentile 1.8%) and the aggregate lift stays $+50 \pm 2$ kg/ha; by $k = 10$–30 it is essentially exact. **A local sample of roughly five to ten plots per enumeration area therefore suffices** to estimate the anchors without materially perturbing the policy. (The EA-mean fertilizer and EA-best-crop anchors are a mean and an argmax over the same sample and stabilize at least as fast as this proportion.)

k plots/EA	peer-rate SE	% argmax changed (mean; p95)	aggregate lift (5–95%)
5	0.22	1.3%; 1.8%	+50 [+48, +51]
10	0.16	1.0%; 1.6%	+50 [+48, +51]
20	0.11	0.7%; 1.1%	+50 [+49, +51]
30	0.09	0.5%; 1.0%	+50 [+49, +51]

Table 15: Sensitivity to peer-anchor estimation noise (E8). The policy is near-invariant to the local sample size used to estimate the EA anchors.

Adoption cap	Aware − accuracy floor (kg/ha)
0.20	**+15 [+9, +20]***
0.25	**+21 [+13, +30]***
0.30	**+28 [+17, +39]***
0.325 *(SAFI control)*	**+31 [+19, +44]***
0.40	**+39 [+24, +55]***
0.50	**+45 [+29, +63]***
0.60	**+48 [+31, +68]***

Table 16: Transportability of the adoption cap (E10). EA-cluster bootstrap 95% CIs; * excludes zero at every cap.

E.9 Calibrating the adoption head beyond isotonic recalibration

Mondrian-isotonic recalibration is in-distribution; the out-of-distribution gap on the SAFI arms ($\sim$ 51 percentage points mean absolute) motivates alternatives. Applying simple post-hoc schemes to the SAFI-mapped predictions: logit-temperature scaling reduces the mean absolute gap to $\sim$ 27 pp (by damping the head's over-confident 0/1 extrapolations toward 0.5), and base-rate shrinkage toward a realistic prior ($\bar{p} = 0.41$) reduces it to $\sim$ 6 pp. The optimal shrinkage is $\lambda \to 1$: on these out-of-distribution actions the base rate *alone* predicts adoption better than the model's extrapolation — a sharp, quantified statement of the calibration gap, and evidence that shrinking toward realistic base rates is a cheap partial correction short of full RCT-augmented training. These figures are tuned post-hoc on the four SAFI arms, so they are illustrative of the achievable reduction rather than a validated deployable calibrator (which would fit the correction on held-out trials).

E.10 Transportability of the SAFI adoption cap

The partial-identification floor (§4.4) caps adoption at the Kenya SAFI control rate (0.325), assuming cross-country transportability. To test sensitivity to that anchor, we re-optimize the policy under a range of adoption caps and recompute the aware-vs-accuracy floor (Table 16). The floor stays positive and significant at every cap from 0.20 (below any SAFI arm) to 0.60. The qualitative conclusion therefore does not hinge on the specific Kenya rate; any plausible adoption ceiling leaves a positive, significant lift. (Re-optimizing the policy under each cap generalizes the fixed-policy floor of §4.4.)

E.11 Yield-head prior sensitivity (β_{crop}, α_{seed})

B1 (§C.1) sweeps the literature-anchored *adoption* penalties; a reviewer correctly noted the literature-anchored *yield*-head priors were not probed analogously. The hybrid yield prediction is $\hat{f}_{\text{base}}(x)\,(1 + \alpha_{\text{seed}}\mathbb{1}[\text{seed}] + m_{\text{irr}}\mathbb{1}[\text{irr}]) + \beta_{\text{crop}(a)}\Delta_{\text{fert}}(a)$; we rescale the whole per-crop response vector β_{crop} and the improved-seed premium α_{seed} (default 0.15) by multipliers and recompute the head-to-head on the same 800 plots (Table 17). The headline is significant at every single-parameter multiplier ($\beta \in [0.5, 1.5]\times$, $\alpha \in [0, 2]\times$) and at every cell of the 3×3 joint grid, including the most adversarial corner ($\beta\times0.5$, α=0: +27.8 [+17.8, +38.7]). The yield priors are therefore a *less* fragile dependence than the adoption penalties

Prior rescaling	Δ aware − acc (kg/ha)
$\beta_{\text{crop}} \times 0.5$	**+41.8 [+28.1, +56.5]***
$\beta_{\text{crop}} \times 0.75$	**+45.7 [+30.2, +63.1]***
$\beta_{\text{crop}} \times 1.0$ *(def.)*	**+50.0 [+32.5, +69.9]***
$\beta_{\text{crop}} \times 1.25$	**+54.6 [+35.2, +76.8]***
$\beta_{\text{crop}} \times 1.5$	**+59.4 [+38.0, +83.9]***
$\alpha_{\text{seed}} \times 0.0$ *(no seed premium)*	**+37.9 [+23.8, +54.5]***
$\alpha_{\text{seed}} \times 0.5$	**+43.2 [+27.7, +60.9]***
$\alpha_{\text{seed}} \times 1.0$ *(def.)*	**+50.0 [+32.5, +69.9]***
$\alpha_{\text{seed}} \times 1.5$	**+58.1 [+38.8, +79.5]***
$\alpha_{\text{seed}} \times 2.0$	**+66.6 [+45.0, +90.2]***
joint worst corner ($\beta\times0.5$, $\alpha\times0$)	**+27.8 [+17.8, +38.7]***

Table 17: Yield-prior sensitivity (E11). Aware-vs-accuracy head-to-head on the same 800 EA-disjoint plots under multiplicative rescaling of the yield-head literature priors; EA-cluster bootstrap 95% CIs. Significant (*) at every multiplier and at every joint-grid cell (worst corner shown).

Policy (scored on realized profit)	Profit lift over SQ (ETB/ha)	vs. accuracy-only
Profit-ERB (objective-consistent)	+3,518 [+2,473, +4,851]	**+2,326 [+1,605, +3,044]***
Yield-ERB (our headline, on profit)	+2,923 [+1,896, +4,196]	**+1,732 [+1,104, +2,354]***
Accuracy-only	+1,192 [+211, +2,595]	—

Table 18: Profit objective (E12). Realized profit lift over status quo of three policies, and the advantage over accuracy-only; EA-cluster bootstrap 95% CIs. Profit-ERB vs. yield-ERB differ by +594 ETB/ha [+233, +1,086] (policy agreement 84%). Robust across $p_{\text{fert}} \in [8, 24]$ ETB/kg.

(which lost significance at 0.5×, Table 6); intuitively, the argmax depends on the tier *ordering* of predicted yield, which uniform rescaling of the response vector preserves.

E.12 Profit- and cost-sensitive objective

ERB is objective-agnostic (§5); with the farm-gate prices used for the cost-benefit translation (App. F) we can substitute *profit* for yield directly. We define per-action profit $\rho(a, x) = \hat{f}(a, x) \cdot p_{\text{crop}(a)} - \Delta_{\text{fert}}(a) \cdot p_{\text{fert}}$ (farm-gate crop price p_{crop}; inorganic-fertilizer cost $p_{\text{fert}} = 16$ ETB/kg, the mid-point of the 2018–19 CSA/FAO 15–18 ETB/kg band), and maximize the profit ERB $\mathbb{P}(\text{adopt}) \cdot (\rho(a, x) - \rho(a_0, x))$. Table 18 scores three policies on realized profit lift. The profit-ERB policy beats accuracy-only by +2,326 ETB/ha [+1,605, +3,044] ($\approx$ \$83/ha); our *yield*-ERB headline policy, re-scored on profit, also significantly beats accuracy-only (+1,732 ETB/ha) and captures 84% of the profit-optimal policy's gain (the two policies agree on 84% of plots). The gap to profit-optimal (+594 ETB/ha) is what pricing inputs recovers. The advantage is stable across the fertilizer price (+2,393 $\rightarrow$ +2,260 ETB/ha as p_{fert} ranges 8 $\rightarrow$ 24). These figures inherit the adoption-head miscalibration (§4.5) and are model-believed, but they establish that adoption-awareness is not an artifact of ignoring input cost.

E.13 Decomposing the Ethiopia→Tanzania non-replication

Rather than only state that Tanzania fails to replicate (§4.5), we test which Tanzania condition drives it by imposing each on the Ethiopia test set (Table 19). Tanzania's test set is $n = 119$, $\sim 0\%$ no-education (71% late-primary), fertilizer base rate 0.11 vs. Ethiopia's 0.29. **No single marginal reproduces the null.** Subsampling to $n = 119$ leaves +49 [+35, +66] (Monte-Carlo over Ethiopia subsamples); restricting to Ethiopia's lowest-peer-adoption plots—the analog of Tanzania's low base rate—leaves +46 [+12, +80]; reweighting to Tanzania's education distribution is the largest single attenuator but still gives +35 [+18, +56]. Only the joint imposition (low-peer plots reweighted to Tanzania education; Kish-effective $n \approx 4$) widens the

Tanzania condition imposed on Ethiopia test set	Δ aware − acc (kg/ha)	n (eff.)
None (full Ethiopia baseline)	+50.0 [+32.5, +69.9]	800
(a) subsample to $n = 119$	+49.2 [+35.1, +65.5]	119
(b) reweight to Tanzania education distribution	+35.0 [+17.8, +55.6]	∼800
(c) low peer-adoption (< 0.15) plots only	+46.5 [+11.6, +80.1]	195
(d) low-peer + Tanzania-education reweight (joint)	+19.5 [+2.5, +52.9]	≈ 4 (Kish)

Table 19: Tanzania-gap decomposition (E13). Each row imposes one (or, in (d), two) Tanzania condition(s) on the Ethiopia test set. Subsample/(d) CIs are Monte-Carlo/weighted cluster bootstrap; others EA-cluster bootstrap. No single condition reproduces the Tanzania null.

SAFI arm	tier	Δfert (kg/ha)	crop-switch	new-seed	#new inputs
Control	0	0	0	0	0
SAFI nudge	1	30	0	0	1
Free delivery	2	60	0	0	1
Both	2	60	0	1	2

Table 20: Kenya SAFI arm → action-complexity vector (the values fed to the adoption head; peer-adoption feature fixed at 0.10 for all arms). Matches the `validate_kenya_rct` harness released with the code.

CI toward zero. We read the non-replication as a structural boundary of the Tanzania-trained heads—both policies collapse to status quo on every Tanzania test plot—rather than a re-expression of any single Ethiopia marginal, and we retract the earlier single-driver "low base rate" explanation accordingly.

E.14 Kenya SAFI arm-to-action-complexity mapping

Table 20 gives the exact action-complexity vectors used to score the Ethiopia-trained adoption head on the Kenya SAFI arms (§4.5). The mapping treats the trial's manipulated frictions as movements along our tier ladder; the peer-adoption feature is held at a common low value (0.10) because the trial does not report EA peer rates, and any adoption-head feature absent from the trial is filled with the Ethiopia training median. This mapping is a modeling assumption, not an identity; its purpose is to place the RCT arms in the head's input space so the OOD calibration gap can be measured, and the partial-identification floor (§4.4) is designed to be robust to exactly the miscalibration it exposes.

F Cost-benefit translation

For interpretability we translate the adoption-aware policy's realized lift over status quo (the per-subgroup aware-policy lifts in the "Aware lift" column of Table 2, *not* the head-to-head Δ) into economic and nutritional units. We use subgroup-specific crop mixes priced at 2018–19 Ethiopian farm-gate values (FAO / Central Statistical Agency of Ethiopia), USDA calorie densities, an exchange rate of 28 ETB/USD, and a per-capita requirement of 2,100 kcal/day for a five-person household. The no-education lift of 239 kg/ha corresponds to ≈ 3,600 ETB/ha (≈ \$128/ha), or ≈ 2.7 household-months of food; the early-primary (345 kg/ha), late-primary (327), and secondary (257) subgroups translate to ≈ 3.9, 3.6, and 2.8 household-months respectively. These are model-believed back-of-envelope translations, not field-measured returns, and they inherit the adoption-head miscalibration of §4.5; they should be read as orders of magnitude only. Price and calorie constants are released with the code.

G Future work, ranked by leverage

1. **RCT-augmented adoption head (observational-plus-experimental calibration).** The SAFI replication data contains explicit treatment-arm assignments and measured uptake. Augmenting adoption-head training with SAFI plus other public agricultural RCTs (Karlan et al., 2014; Beaman et al., 2021)

via trial-transport / generalizability methods (Colnet et al., 2024; Dahabreh et al., 2019) would directly address the over-confidence; we expect it to bring per-arm calibration gaps below 0.10. Single highest-leverage follow-up.

2. **MSM-optimal robust policy with formal guarantees.** We report a Λ-sweep of the MSM-optimal pessimistic policy (significant advantage through $\Lambda = 5$; App. E.3); a finite-sample worst-case guarantee across continuous Λ (Tan, 2006; Yadlowsky et al., 2022; Kallus & Zhou, 2018) remains future work.
3. **Fully distribution-free joint bound.** App. E.4 now reports a joint bound that combines the yield conformal interval with a *bootstrap* uncertainty set on the adoption head (capped at the SAFI RCT rate), because conformal coverage on the adoption head does not transfer out-of-distribution; the remaining step is to replace the parametric bootstrap set with a *fully* distribution-free, OOD-valid conformal treatment via conformal off-policy prediction (Taufiq et al., 2022).
4. **R-learner and semiparametric-efficient variants.** App. E.1 reports T-/S-/X-/DR-learner baselines (all beaten by ERB); extending to the R-learner (Nie & Wager, 2021) and to a multi-action (rather than single-fertilizer-arm) CATE formulation on the same panel would round out the comparison.
5. **Proxy / bridge-function relaxation.** Proximal causal learning (Tchetgen Tchetgen et al., 2020; Cui et al., 2024) using EA-aggregate proxies for unobserved plot quality would unlock semiparametric-efficient ERB estimation under interpretable confounding assumptions.
6. **Multi-country LSMS replication and a field pilot.** Nigeria GHS-Panel, Uganda UNPS, Malawi IHS-Panel would generalize or further bound the framework; a field pilot with an extension service is the only path to non-model-believed lift estimates.

H Sociotechnical fairness trade-off

A reviewer-relevant tension in any adoption-aware recommender is that *recommending simpler actions to disadvantaged subgroups can entrench the disparities that produced the disadvantage.* If a no-education farmer's adoption-probability landscape is depressed by credit constraints, market-access frictions, or extension-coverage gaps, rather than by any incapacity, then an ERB-maximizing system that recommends T_2 rather than T_4 to that farmer accommodates the structural barrier instead of removing it. We therefore frame ERB as a *complement* to structural-access policy: pair ERB recommendations with friction-removing instruments (credit subsidy, input loans, cooperative purchasing); report realized-benefit gap closure dynamically across deployment seasons rather than one-shot subgroup metrics; and treat low-$\mathbb{P}$(adopt)-on-high-yield-gain plots as targeting signal for structural intervention, not only as a yield-discount factor. The danger is treating ERB as a sufficient solution rather than one component of an integrated equity-aware advisory system.