AI Co-Pilots in Low-Resource Team Decisions: Fairness, Skill Drift, and Evidence-Ready Governance

Akshaya Jayasankar

Abstract

AI assistants are starting to support complex, collaborative work, but we still know little about when they genuinely improve human expertise versus when they quietly erode it. This study designs an AI "co-pilot" for a moderation-like team task in low-resource settings (voice / IVR / WhatsApp, intermittent connectivity, low-end Android devices) and sketches a preregistered, multi-session randomized experiment to evaluate it.

The co-pilot shows calibrated recommendations, short natural-language rationales, and targeted "second look" prompts that nudge workers to re-check high-stakes, high-confidence suggestions. We plan to measure effects on decision quality, group fairness (e.g., equalized-odds style gaps in TPR/FPR), trust and reliance, and whether human judgment skills drift or improve over time.

To meet governance expectations, the design includes a Model Card, Datasheet, and a NIST AI RMF–aligned checklist, plus a public artifact package with synthetic demo data and code. We also propose an access-aware disparity index (ADI) that tracks how performance and fairness change across connectivity, device class, and language strata in real deployments, a longitudinal Skill-Drift Index (SDI) for unaided human performance, and a community-governed evaluation loop that ties these metrics to local advisory structures and post-study handoff.

1 Introduction

Organizations are rapidly experimenting with AI co-pilots to support day-to-day judgment, but evidence is mixed on when assistance lifts human skill versus quietly eroding it through over-reliance. We study the socio-technical effects of AI co-pilots in team workflows, focusing on two tensions: (i) fairness vs. performance under assistance, and (ii) trust vs. automation bias that can produce skill drift over time.

Research questions. *When do co-pilot design choices (calibration, rationales, counter-arguments) improve equity and decision quality without eroding human skill?* We propose a preregistered team experiment that varies co-pilot assistance styles and estimates heterogeneous effects using modern ML-based causal methods [1–3].

Contributions. This brief makes three contributions. (1) It proposes a fairness-aware assistance architecture for low-resource, voice-first contexts that explicitly tracks access-aware parity and calibration over time. (2) It links AI governance and sociotechnical measurement by embedding AI management concepts (RMF/ISO) into the system design, including explicit controls for calibration, override logging, and access-aware fairness monitoring. (3) It provides a small, publicly available GitHub repository with synthetic data and analysis scripts, illustrating what an "evidence-ready" deployment study could look like even before a full-scale RCT is run; an OSF preregistration and archive snapshots are treated as *planned* extensions rather than current deliverables (see App. A, B).

1.1 From design-for to design-with: participatory AI governance

Much work on AI assistance in organizations still treats communities as passive recipients: systems are "designed for" rather than "designed with" the people who bear the risks. In contrast, ICTD and participatory HCI emphasize long-horizon partnerships, co-specification of workflows, and community-led governance over data and models. Our proposed co-pilot follows this lineage by foregrounding ASHAs and local NGOs as *co-governors* of the system rather than only "end users."

Practically, this means three things. First, formative and ongoing co-design shape not only UI details but also what is logged, which metrics are monitored, and how assistance is framed in local languages. Second, fairness and calibration are not treated as purely technical objectives: the Access-aware Disparity Index (ADI) and Skill-Drift Index (SDI) are intended to be read and debated in community advisory panels, not just in internal dashboards. Third, our governance mapping (NIST AI RMF; ISO/IEC 42001) is framed as a tool for *local* accountability: we explicitly plan handoff paths where community partners retain control of artifacts, thresholds, and redlines after the research phase.

Finally, we design explicitly for *globally equitable AI.* The co-pilot is tailored to offline-first, multilingual voice/SMS/IVR on low-end Android phones, motivated by constraints faced by community health workers (e.g., ASHA-like cadres in rural Karnataka) and prior work on voice systems for underserved communities [4, 5]. The same design choices (voice/IVR, WhatsApp/SMS, offline-first, access-aware fairness) target similar constraints observed across parts of Sub-Saharan Africa and Latin America, where usage rather than nominal coverage drives the mobile internet gap [6].

2 Related Work

2.1 ICTD voice systems, CHWs, and WhatsApp ecosystems

Voice- and phone-first systems have long supported low-resource communities in India. Sangeet Swara [4] demonstrated community-moderated voice forums running over basic phones; more broadly, Vashistha's survey distills design lessons for *voice interfaces for underserved communities* (offline-first, low-literacy UI, locally adapted vocabularies) [5]. For health workflows, in-situ participatory design with CHWs highlights the value of co-designed feedback and voice/web hybrids in longitudinal deployments [7].

Our proposed co-pilot builds on these directions with fairness-aware, calibrated assistance that offers *voice/IVR and WhatsApp/SMS* interaction, and explicitly situates assistance within WhatsApp social norms and misinformation dynamics [8]. Beyond the motivating India context, prior work on voice/phone-first civic systems (e.g., Ila Dhageyso in Somaliland) suggests that IVR-style designs can transfer to other regions, informing our focus on access-aware design across Sub-Saharan Africa and parts of Latin America [5].

2.2 Algorithmic fairness and reporting

We focus on parity gaps consistent with *equal opportunity* and *equalized odds* [9]. For transparency, we specify a *Model Card* [10] and a *Datasheet* structure for the evaluation data [11], aligning the planned artifact package with emerging reporting standards in algorithmic fairness.

2.3 Human–AI teaming, trust, and automation bias

Automation bias and algorithm aversion show that people may over-trust or under-use automation, especially after observing errors [12–15]. Motivated by this literature, our design exposes calibrated probabilities and adds counter-argument prompts that encourage a second look when model confidence is high, aiming to support appropriate reliance rather than blind trust or blanket rejection.

2.4 Causal estimation and heterogeneous effects

Our experimental design targets intent-to-treat (ITT) effects and heterogeneous treatment effects (HTEs) learned with generalized/causal forests [1, 2], using tools such as EconML and DoWhy for policy-relevant analyses [3, 16]. This connects the study to modern causal-inference practice that emphasizes treatment heterogeneity and model-based policy evaluation.

2.5 Participatory AI and community governance in ICTD

Participatory methods in ICTD and HCI have long argued that technical systems should be co-designed and co-governed with communities, especially where power and resource asymmetries are stark. In low-resource health settings, longitudinal collaborations with CHWs and local NGOs have shown that shared ownership over protocols, content, and data flows is critical for sustained use and trust [7]. Work on documentation and reporting (e.g., model cards and datasheets) similarly emphasizes explicitly surfacing context, stakeholders, and limitations rather than presenting models as context-free artifacts [10, 11].

Our design adopts this participatory AI lens in three ways: (i) formative work and workshops with ASHAs and supervisors shape prompts, challenge scripts, and fallback behavior; (ii) governance artifacts (Model Card, Datasheet, NIST/ISO mappings) are intended as shared objects in community advisory meetings, not just internal compliance documents; and (iii) fairness metrics such as ADI and SDI are parameterized to reflect local notions of inclusion and harm, and can be re-weighted or re-stratified as partners identify new axes of disparity.

3 System: Co-pilot for Team Decisions

Overview. We design a lightweight web app that surfaces (a) ranked suggestions with calibrated probabilities, (b) concise rationales, and (c) a *counter* panel offering counter-arguments or alternative labels to reduce over-reliance. All user actions and model outputs are event-logged (timestamps, features, model/version hashes) to support reproducibility and governance.

3.1 Fairness–aware calibrated recommendations

Our design trains a transformer encoder (Sec. 3.2) with a joint objective that trades off accuracy, group parity, and probability calibration:

$$\mathcal{L}(\theta,\lambda) = \underbrace{\mathcal{L}_{\text{acc}}(\theta)}_{\text{cross-entropy}} + \lambda_1 \underbrace{\mathcal{L}_{\text{fair}}(\theta)}_{\text{equalized-odds gap}} + \lambda_2 \underbrace{\mathcal{L}_{\text{cal}}(\theta)}_{\text{Dirichlet calibration}} . \tag{1}$$

Algorithm 1: FairCal: Fairness-aware calibrated training with online weight updates

```
Input: data $(x_i, y_i, A_i)$; calibration head $\mathrm{DirCal}(\cdot)$; weights $\lambda_1, \lambda_2$
for epoch do
    sample minibatch $\mathcal{B}$; logits $z_\theta(x)$; probs $\hat{p} = \mathrm{DirCal}(z)$;
    $\mathcal{L}_{\text{acc}} \leftarrow -\sum_{(x,y)\in\mathcal{B}} \log \hat{p}_y$;
    compute EO gaps $\Delta_{\text{TPR}}, \Delta_{\text{FPR}}$ on $\mathcal{B}$;
    $\mathcal{L}_{\text{fair}} \leftarrow |\Delta_{\text{TPR}}| + |\Delta_{\text{FPR}}|$;
    $\mathcal{L}_{\text{cal}} \leftarrow \mathrm{NLL}(\hat{p}, y)$;
    $\mathcal{L} \leftarrow \mathcal{L}_{\text{acc}} + \lambda_1 \mathcal{L}_{\text{fair}} + \lambda_2 \mathcal{L}_{\text{cal}}$;
    update $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}$;
Online after session $s$: if parity worsens, $\lambda_1 \leftarrow \lambda_1 + \eta_\lambda$; if ECE worsens, $\lambda_2 \leftarrow \lambda_2 + \eta_\lambda$.
```

Differentiable parity surrogate. Directly constraining equalized-odds is non-differentiable, so we use a smooth proxy that penalizes between-group differences in mean scores conditional on y:

$$\widehat{\mathrm{EO}}(\theta) = \sum_{y\in\{0,1\}} \left|\mathbb{E}[\hat{p}_\theta(x) \mid A{=}1, y] - \mathbb{E}[\hat{p}_\theta(x) \mid A{=}0, y]\right|.$$

Our training loss becomes

$$\mathcal{L} = \mathcal{L}_{\text{acc}} + \lambda_1 \, \widehat{\mathrm{EO}} + \lambda_2 \, \mathcal{L}_{\text{cal}},$$

which preserves gradients while we track TPR/FPR gaps ex post in evaluation.

In our design, λ_1 and λ_2 are updated online based on human overrides (Sec. 4.5): increasing λ_1 when group gaps widen, and λ_2 when reliability diagrams drift.

Fairness penalty. For binary $y \in \{0, 1\}$ and sensitive group A, the equalized-odds penalty matches TPR/FPR across groups:

$$\mathcal{L}_{\text{fair}} \;=\; \left|\mathrm{TPR}_{A=0} - \mathrm{TPR}_{A=1}\right| \;+\; \left|\mathrm{FPR}_{A=0} - \mathrm{FPR}_{A=1}\right|.$$

Calibration term. We add a learnable multiclass calibration head (Dirichlet calibration) so probabilities are calibrated, not just logits temperature-scaled:

$$\mathcal{L}_{\text{cal}} = \mathrm{NLL}\big(\mathrm{DirCal}(p_\theta(y \,|\, x)),\, y\big).$$

3.2 Model and uncertainty

We encode text with a transformer (e.g., BERT/RoBERTa) and a shallow head outputs class logits. To obtain calibrated probabilities, we attach a Dirichlet calibration layer to the logits (Sec. 4.4). We expose predictive uncertainty to trigger challenge prompts via: (i) Monte Carlo dropout at inference (approximate Bayesian), and (ii) deep ensembles for epistemic uncertainty. High uncertainty raises a "second look" panel in the UI.

3.3 Online adaptation and bandit selection

We log overrides and outcomes to adapt the system in real time. A contextual bandit allocates assistance style per task (None / Rationale / Calibrated / Counter-arguments), balancing exploration and utility; fairness constraints penalize styles that increase group gaps. We also update the loss weights λ_1, λ_2 online: if session-level parity (TPR/FPR gaps) worsens, we increase λ_1; if reliability drifts (ECE increases), we increase λ_2. This yields a closed loop: data → model → UI → feedback → update.

Algorithm 2: Training + online adaptation with fairness/calibration tradeoffs

```
Input: Data $\mathcal{D} = \{(x_i, y_i, a_i)\}$, group $a_i \in \{0,1\}$; base model $f_\theta$ (Transformer); calibrator
       $g_\phi$ (Dirichlet); uncertainty head $u_\psi$ (MC-dropout / ensemble); initial weights
       $\lambda_1, \lambda_2 \geq 0$; step sizes $\eta_\theta, \eta_\phi, \eta_\psi, \eta_\lambda$; fairness target $\tau_{\text{fair}}$, calibration target $\tau_{\text{cal}}$.
Output: Calibrated predictor $\hat{p}(x) = g_\phi(f_\theta(x))$, uncertainty $\sigma(x) = u_\psi(x)$.
▷ --- Offline supervised pretraining
  ------------------------------------------------
for epoch $= 1 \dots E$ do
    Sample minibatch $\mathcal{B} \subset \mathcal{D}$
    // Forward
    $z \leftarrow f_\theta(x);\quad \hat{p} \leftarrow g_\phi(z);\quad \sigma \leftarrow u_\psi(x)$
    // Losses use the paper's symbols
    $\mathcal{L}_{\text{acc}} \leftarrow -\frac{1}{|\mathcal{B}|} \sum_{(x,y)\in\mathcal{B}} \log \hat{p}_y$
    ▷ Equalized-odds style parity across groups
    $\mathcal{L}_{\text{EO}} \leftarrow \text{EOgap}(\hat{p}, a, y)$
    ▷ Reliability penalty (e.g., differentiable ECE/Brier/ACE surrogate)
    $\mathcal{L}_{\text{cal}} \leftarrow \text{RelErr}(\hat{p}, y)$
    // Total objective (matches your write-up)
    $\mathcal{L} \leftarrow \mathcal{L}_{\text{acc}} + \lambda_1 \mathcal{L}_{\text{EO}} + \lambda_2 \mathcal{L}_{\text{cal}}$
    // Update
    $(\theta, \phi, \psi) \leftarrow (\theta, \phi, \psi) - \eta \nabla \mathcal{L}$
▷ --- Online adaptation during deployment (Sec. 4.5) -------------------
for session $s = 1 \dots S$ do
    for task $t$ in session do
        $z \leftarrow f_\theta(x_t);\quad \hat{p} \leftarrow g_\phi(z);\quad \sigma \leftarrow u_\psi(x_t)$
        // Bandit selects assistance style/threshold given context; UI logs
           events
        $a^\star \leftarrow \textsc{BanditPolicy}(x_t, \sigma)$
        Show suggestion/rationale or trigger counter-arguments if $\sigma$ high
        Log outcome, override, latency → event store
    // Compute session-level monitors used elsewhere in the paper
    $\text{TPRgap}_s, \text{FPRgap}_s \leftarrow \textsc{ComputeParityGaps}(\text{logs})$
    $\text{ECE}_s \leftarrow \textsc{ReliabilityFromLogs}(\text{logs})$
    ▷ Closed-loop updates of trade-off weights $\lambda_1, \lambda_2$
    $\lambda_1 \leftarrow \max\{0,\ \lambda_1 + \eta_\lambda[(\text{TPRgap}_s + \text{FPRgap}_s) - \tau_{\text{fair}}]\}$
    $\lambda_2 \leftarrow \max\{0,\ \lambda_2 + \eta_\lambda[\text{ECE}_s - \tau_{\text{cal}}]\}$
```

3.4 Low-resource and accessible deployment

We target low-connectivity, low-end devices in the design of the co-pilot. The system is intended to support offline-first use: models are quantized and cached on-device; logs and updates sync opportunistically over 2G/3G. Interfaces include text UIs and voice/IVR and WhatsApp/SMS flows with multilingual prompts, plus screen-reader-friendly markup. We aim to co-design with local practitioners, with governance and data stewardship aligned to community norms.

Governance alignment (NIST AI-RMF). Our loop is intended to instantiate the AI-RMF functions: *Govern* (policies/roles; fairness constraints and override logging), *Map* (task/context

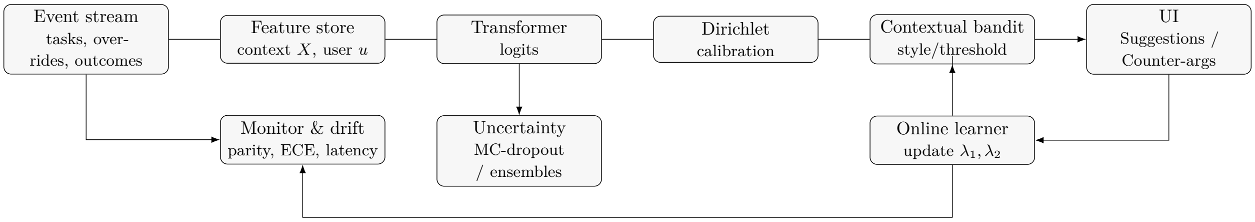


Figure 1: Proposed end-to-end co-pilot architecture: transformer + uncertainty (MC-dropout, deep ensembles), Dirichlet calibration, fairness-aware bandit selection, and online updates of loss weights.

features and stakeholder impacts in the feature store), *Measure* (parity gaps, ECE/reliability, drift monitors), and *Manage* (bandit allocation, online updates of λ_1, λ_2, incident handling). This planned mapping follows NIST AI RMF 1.0 and Playbook guidance on embedding risk and trustworthiness checks into the development/runtime pipeline [17, 18].

3.5 Interaction design to reduce bias

A one-click *challenge* action reveals counter-evidence when model confidence is high; editable recommendations allow small user tweaks, which reduces algorithm aversion in prior work [15].

3.6 User-adaptive co-pilot via meta-learning

We propose modeling per-reviewer behavior with MAML: initialize θ_0 on pooled tasks, then adapt to user u with one–two gradient steps on their override history, yielding θ_u. Confidence thresholds can be personalized with contextual bandits over features (x, u).

3.7 Governance hooks

We specify a NIST AI RMF checklist (GOVERN, MAP, MEASURE, MANAGE) and note alignment points to ISO/IEC 42001 (AI management systems) [17, 19, 20].

3.8 Community-led fairness auditing and local advisory structures

To move beyond purely technical monitoring, we design the co-pilot to support community-led fairness audits. In addition to internal dashboards, we envision a "governance console" view that aggregates ADI, SDI, EO gaps, and override logs in simple summaries that can be discussed with ASHAs, supervisors, and NGO staff during periodic advisory meetings.

Practically, this implies (i) curated views that highlight disparities across connectivity, device class, and language, with drill-down to sample cases; (ii) controls for local partners to flag problematic prompts, thresholds, or workflows, which then enter an incident and takedown queue; and (iii) annotations on metrics that record community feedback (e.g., whether a particular gap is seen as acceptable, urgent, or misaligned with local priorities). These structures are meant to turn fairness monitoring from a one-way reporting exercise into an ongoing, dialogic process in which community members can shape the criteria by which the co-pilot is judged and adapted.

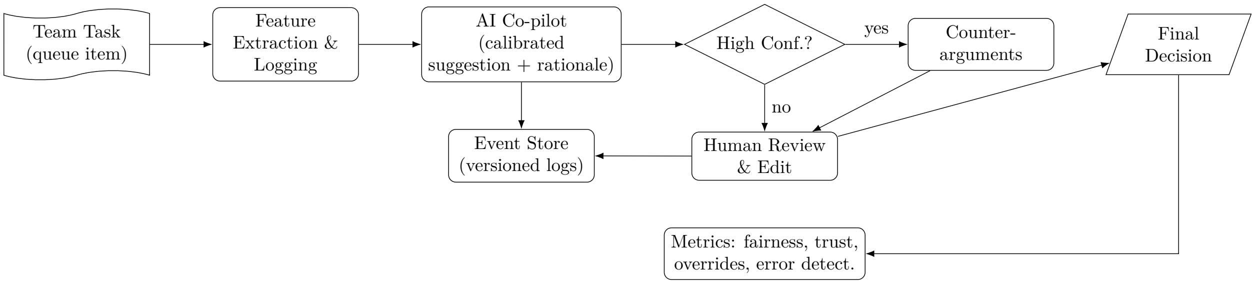
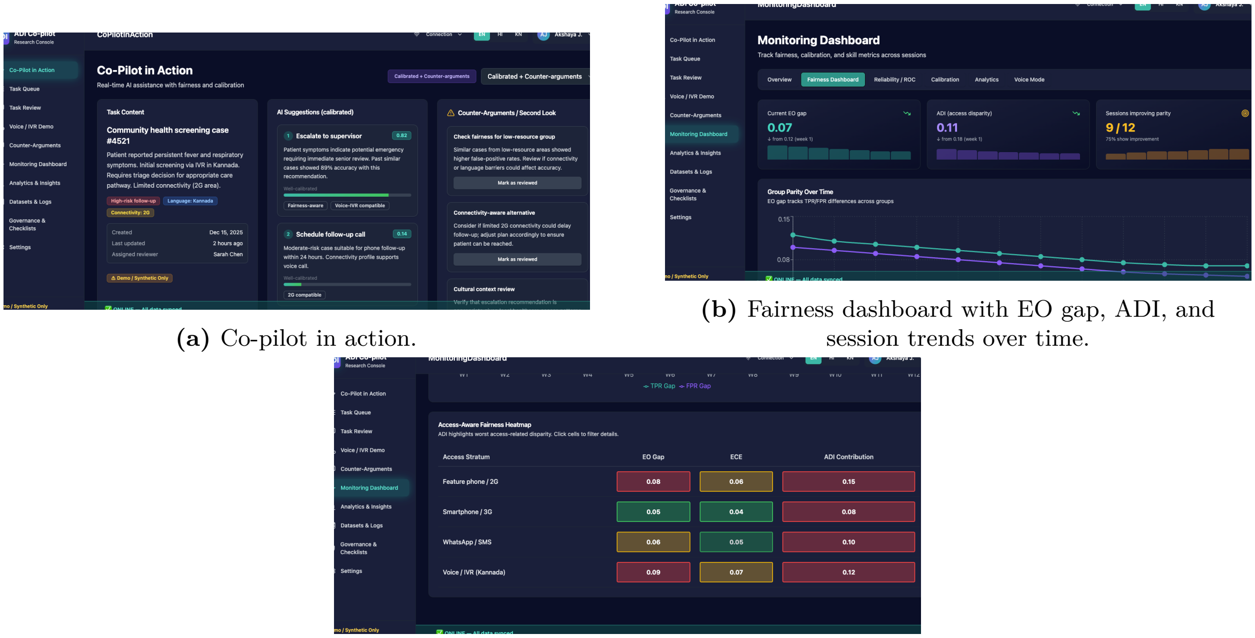


Figure 2: System flow for the co-pilot: calibrated suggestions, rationales, counter-arguments; comprehensive logging for reproducibility and governance.



Co-Pilot in Action

Real-time AI assistance with fairness and calibration

Task Content

Community health screening case #4521

Patient reported persistent fever and respiratory symptoms. Initial screening via IVR in Kannada. Requires triage decision for appropriate care pathway. Limited connectivity (2G area).

High-risk follow-up | Language: Kannada | Connectivity: 2G

Created	Dec 15, 2025
Last updated	2 hours ago
Assigned reviewer	Sarah Chen

⚠ Demo / Synthetic Only

AI Suggestions (calibrated)

1 Escalate to supervisor 0.82

Patient symptoms indicate potential emergency requiring immediate senior review. Past similar cases showed 89% accuracy with this recommendation.

Well-calibrated

Fairness-aware | Voice-IVR compatible

2 Schedule follow-up call 0.14

Moderate-risk case suitable for phone follow-up within 24 hours. Connectivity profile supports voice call.

Well-calibrated

2G compatible

⚠ Counter-Arguments / Second Look

Check fairness for low-resource group

Similar cases from low-resource areas showed higher false-positive rates. Review if connectivity or language barriers could affect accuracy.

Connectivity-aware alternative

Consider if limited 2G connectivity could delay follow-up; adjust plan accordingly to ensure patient can be reached.

Cultural context review

Verify that escalation recommendation is

(a) Co-pilot in action.

Monitoring Dashboard

Track fairness, calibration, and skill metrics across sessions

Current EO gap 0.07 ↓ from 0.12 (week 1)

ADI (access disparity) 0.11 ↓ from 0.18 (week 1)

Sessions improving parity 9 / 12 75% show improvement

Group Parity Over Time

EO gap tracks TPR/FPR differences across groups

(b) Fairness dashboard with EO gap, ADI, and session trends over time.

Access-Aware Fairness Heatmap

ADI highlights worst access-related disparity. Click cells to filter details.

Access Stratum	EO Gap	ECE	ADI Contribution
Feature phone / 2G	0.08	0.06	0.15
Smartphone / 3G	0.05	0.04	0.08
WhatsApp / SMS	0.06	0.05	0.10
Voice / IVR (Kannada)	0.09	0.07	0.12

(c) Access-aware fairness heatmap by device/connectivity and modality.

Figure 3: Interface snapshots.

4 Technical Background

4.1 Notation and problem setup

Let x denote task features, $y \in \{0, 1\}$ the ground truth label, $A \in \{0, 1\}$ a protected attribute (e.g., group), a the AI suggestion, and h the human decision. Sessions are indexed by $s = 1, \ldots, S$, with each session s containing items $i = 1, \ldots, N_s$. We denote unaided human decisions by h^{unaided} and AI-aided decisions by h^{aided}.

4.2 Skill drift and Net Skill Differential (SDI)

We study whether unaided human skill improves or decays over time under intermittent AI assistance. Let U_s denote the expected loss of unaided human decisions in session s for a proper loss ℓ (e.g., 0–1 loss or cross-entropy on human labels). Lower values of U_s indicate better unaided performance. We define the session-averaged skill-drift index:

$$\mathrm{SDI} \;=\; \frac{1}{S-1}\sum_{s=2}^{S}\big(U_s - U_{s-1}\big).$$

Positive SDI indicates net skill degradation (loss increasing over sessions), while negative SDI indicates net skill improvement.

Estimation. We estimate U_s in two complementary ways: (i) *AI-off practice windows* embedded in each session; and (ii) *doubly-robust (DML) counterfactual estimation* using learned outcome and propensity models with orthogonalization (Sec. 6). Under mild conditions for sequential data, this yields consistent estimates of how unaided skill evolves over time.

4.3 Dynamic fairness for interactive use

Static parity metrics are insufficient when decisions unfold over sessions. We therefore track the *fairness degradation rate*:

$$\mathrm{FDR} \;=\; \frac{1}{S-1}\sum_{s=2}^{S}\Big[\big|\mathrm{TPR}_\Delta\big|_s + \big|\mathrm{FPR}_\Delta\big|_s - \big(\big|\mathrm{TPR}_\Delta\big|_{s-1} + \big|\mathrm{FPR}_\Delta\big|_{s-1}\big)\Big]\,,$$

where TPR_Δ and FPR_Δ denote between-group gaps at session s. Positive FDR indicates that between-group parity is worsening over sessions (gaps increasing), while negative FDR indicates improving parity. We combine FDR with our fairness-aware objective (Sec. 4.4) to adapt penalties online.

Access-aware fairness. Beyond a protected group A, we stratify by access modalities M (e.g., connectivity tier, device class, language). Let $\Delta^{(M)}_{\mathrm{EO}}$ denote the equalized-odds gap within stratum M; we report the *access disparity index* $\mathrm{ADI} = \max_M |\Delta^{(M)}_{\mathrm{EO}}|$ and session-level change $\Delta\mathrm{ADI}_s = \mathrm{ADI}_s - \mathrm{ADI}_{s-1}$ to monitor inclusion over time.

4.4 Fairness-aware learning objective

We train the assistance model with a composite loss that trades off accuracy, parity, and calibration:

$$\mathcal{L}(\theta,\lambda) = \mathcal{L}_{\mathrm{acc}}(\theta) + \lambda_1\,\mathcal{L}_{\mathrm{fair}}(\theta) + \lambda_2\,\mathcal{L}_{\mathrm{cal}}(\theta), \tag{2}$$

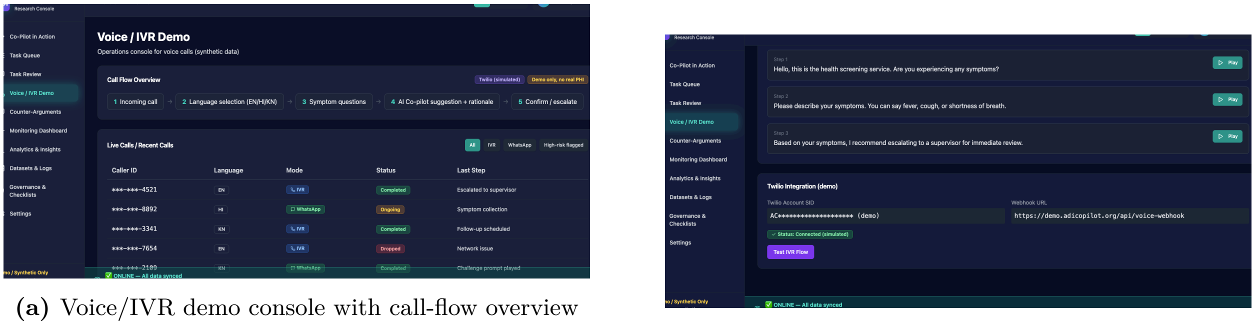
where θ are model parameters and $\lambda_1, \lambda_2 \geq 0$ are weights that can be adapted online (Sec. 4.5).

Differentiable parity surrogate. Directly optimizing equalized odds is non-differentiable, so we use a smooth proxy that penalizes between-group mean-score differences conditional on the true label y:

$$\widehat{\mathrm{EO}}(\theta) = \sum_{y\in\{0,1\}} \left|\mathbb{E}[\hat{p}_\theta(x)\mid A{=}1, y] - \mathbb{E}[\hat{p}_\theta(x)\mid A{=}0, y]\right|.$$

We set $\mathcal{L}_{\mathrm{fair}} := \widehat{\mathrm{EO}}$ during training and report TPR/FPR gaps ex post for evaluation.

Calibration term. $\mathcal{L}_{\mathrm{cal}}$ is a calibration penalty (e.g., differentiable Brier or a Dirichlet-calibration objective) to better align predicted confidences with empirical accuracy; reliability is evaluated via ECE (Sec. 5.8).



Voice / IVR Demo

Operations console for voice calls (synthetic data)

Call Flow Overview

1 Incoming call → 2 Language selection (EN/HI/KN) → 3 Symptom questions → 4 AI Co-pilot suggestion + rationale → 5 Confirm / escalate

Live Calls / Recent Calls

Caller ID	Language	Mode	Status	Last Step
--4521	EN	IVR	Completed	Escalated to supervisor
--8892	HI	WhatsApp	Ongoing	Symptom collection
--3341	KN	IVR	Completed	Follow-up scheduled
--7654	EN	IVR	Dropped	Network issue
--2109	KN	WhatsApp	Completed	Challenge prompt played

Step 1
Hello, this is the health screening service. Are you experiencing any symptoms?

Step 2
Please describe your symptoms. You can say fever, cough, or shortness of breath.

Step 3
Based on your symptoms, I recommend escalating to a supervisor for immediate review.

Twilio Integration (demo)

Twilio Account SID: AC******************** (demo)

Webhook URL: https://demo.adicopilot.org/api/voice-webhook

✓ Status: Connected (simulated)

(a) Voice/IVR demo console with call-flow overview and recent calls across EN/HI/KN and IVR/WhatsApp modes.

(b) Simulated Twilio integration showing IVR script steps and demo webhook endpoint.

Figure 4: Voice/IVR operations console for the co-pilot, including simulated telephony integration.

4.5 Online adaptation (bandit control)

We propose updating (λ_1, λ_2) over sessions using a contextual bandit that observes per-session fairness degradation (FDR; Sec. 4.3), calibration (ECE), and latency, and selects assistance styles/thresholds that approximately minimize a proxy of Eq. (2) subject to operational constraints.

4.6 Reliability (calibration) plots

Expected Calibration Error (ECE). With B equal-width probability bins $\{I_b\}$, predicted confidences $\hat{p}_i$, and outcomes $y_i \in \{0, 1\}$,

$$\mathrm{ECE} = \sum_{b=1}^{B} \frac{|I_b|}{n} \left| \underbrace{\frac{1}{|I_b|} \sum_{i \in I_b} \nVdash[y_i = 1]}_{\text{emp. accuracy}} - \underbrace{\frac{1}{|I_b|} \sum_{i \in I_b} \hat{p}_i}_{\text{avg. confidence}} \right|.$$

We report ECE per access stratum M and overall.

4.7 Fairness under distribution and infrastructure drift

Fairness guarantees that hold on a static benchmark can degrade once deployment conditions shift. In our setting, both label distributions and infrastructure conditions (connectivity, device mix, language mix) may drift over time. We therefore treat fairness as a *dynamic* property: EO gaps, ADI, SDI, and calibration error are computed per session and summarized by the fairness degradation rate FDR (Sec. 4.3).

Infrastructure drift matters because access itself is a mediator of harm: when connectivity degrades or ASR quality drops for certain accents or languages, people in those strata may see reduced assistance quality or fewer opportunities to benefit from the co-pilot. Our monitoring loop is designed to capture such shifts by combining telemetry (latency, drop rates, device class) with stratified performance metrics, and by feeding these into the online adaptation and governance hooks described in Sec. 3.

5 Methods

5.1 Planned setting

We plan to partner with a public–health NGO working with Accredited Social Health Activists (ASHAs) in rural India. A typical deployment involves dozens of ASHAs across several villages (population on the order of 500–2,000 each), with intermittent 2G/3G connectivity and entry-level Android devices. Primary languages include Kannada and Hindi. The design is grounded in prior ICTD/HCI work on voice/phone systems and CHW deployments in India [4, 5, 7].

5.2 Participatory co-design with ASHAs (three-phase protocol)

We plan a three-phase participatory protocol with ASHAs and supervisors.

Phase 1: Workflow mapping and risk elicitation. Semi-structured interviews and shadowing sessions will document existing triage workflows, handoff patterns, and informal "back-channel" practices (e.g., senior nurses consulted via phone or WhatsApp). We will explicitly elicit where errors are most consequential, how workers currently double-check decisions, and what kinds of AI behavior would feel unacceptable or disrespectful.

Phase 2: Co-design of prompts, flows, and challenge scripts. Building on Phase 1, participatory workshops will be used to co-specify IVR prompts, WhatsApp micro-flows, and "second look" challenge scripts in Kannada and Hindi. Activities will include card-sorting of candidate messages, sketching call-flows on paper, and role-play of assisted vs. unassisted decisions. The goal is to align tone, prosody, and timing of prompts with local communication norms, and to avoid framings that could undermine CHW autonomy.

Phase 3: Governance and metric mapping. In later workshops, we will present simplified versions of the fairness and skill metrics (EO gaps, ADI, SDI, override selectivity) and discuss which strata and thresholds matter most to ASHAs and NGO partners. This phase is intended to seed the community advisory panel that later reviews metrics and incidents (Sec. 5.4, App. D), ensuring that governance criteria are co-defined rather than imposed unilaterally.

5.3 Practicalities in low-resource deployment

The design assumes intermittent connectivity, periodic power outages, and shared charging infrastructure. We therefore include offline caching with resumable uploads, SMS/WhatsApp fallbacks for short notices, and opportunistic sync over 2G/3G. Connectivity and latency logs are planned to support access-aware analyses (Sec. 4.3).

5.3.1 Language and cultural adaptation

We localize prompts in Kannada and Hindi with culturally adapted terms (e.g., "high blood pressure" → ಉಚ್ಚ ರಕ್ತದೊತ್ತಡ) and short audio explainers. Voice menus default to the participant's preferred language; the text UI includes screen-reader–friendly markup. Terminology, audio glossaries, and framing are intended to be co-designed with ASHAs in future workshops (Sec. 5.2), consistent with voice-first ICTD practice [21, 22].

Table 1: Language packs and culturally adapted terms (examples).

Concept	Kannada (audio text)	Hindi (audio text)
High blood pressure	ಉಚ್ಚ ರಕ್ತದೊತ್ತಡ	उच्च रक्तचाप
Second look	ಎರಡನೇ ಪರಿಶೀಲನೆ	दूसरी नज़र
Counter–argument	ವಿರುದ್ಧ ಕಾರಣ	प्रतिवाद

Why this fits low-resource settings. Affordability and usage gaps persist in contexts where nominal coverage exists but mobile internet is under-used. Voice/telco channels and offline-first designs reduce this burden while preserving auditability and fairness checks [6]. Prior ICTD/HCI deployments show that community-moderated voice systems and CHW-aligned participatory design can sustain engagement under similar constraints [7, 22].

Beyond lexical choices, we also treat prosody and timing as design variables for voice UIs. "Second look" prompts and counter-arguments will be tested with ASHAs to calibrate tone (e.g., neutral vs. apologetic), speaking rate, and pause length before and after critical suggestions. Prior work in HRI and conversational agents suggests that such cues shape whether prompts are experienced as supportive reminders or as interruptions; our co-design activities will therefore explicitly probe which prosodic patterns feel respectful, especially when the system is flagging a possible error by a senior worker.

5.4 Ethics and community governance

For any field deployment, we plan to seek IRB approval and obtain informed consent in local languages (e.g., Kannada/Hindi), with audio consent options for low-literacy participants. Phones would be tagged with non-identifying IDs, and logs would exclude free text unless explicitly consented for research. We envision a community advisory panel (ASHAs + NGO representatives) to review prompts and rationales periodically and approve policy changes. Data control and handoff plans would follow "globally equitable AI" guidance.

5.4.1 Participatory calibration: community input on confidence thresholds

Calibration thresholds and "high-confidence" cutoffs are not purely technical choices: they shape when the co-pilot is allowed to sound certain, when it must hedge, and when it should actively defer to humans. We therefore plan to run small-group sessions where ASHAs and supervisors review example cases with different probability displays and threshold settings, and indicate which combinations feel appropriate for their context.

Feedback from these sessions will be used to (i) set initial thresholds for "high confidence" warnings and "second look" triggers; (ii) define categories for community-facing dashboards (e.g., how to bucket confidence bands); and (iii) specify redlines, such as classes of cases where the co-pilot must always display uncertainty or encourage escalation, independent of model scores.

5.5 Global generalizability (brief)

Constraints similar to our motivating India context recur elsewhere: intermittent 2G/3G/4G, basic Android handsets, and high device-cost barriers. GSMA reports a persistent global *usage gap* (people living under coverage but not using mobile internet), and affordability efforts are underway with multi-stakeholder coalitions [6, 23]. We therefore parameterize language packs, IVR prompts, and fairness strata to port to regions such as East Africa or the Andes, with co-design with local partners as a first step.

5.5.1 Transfer protocol for other regions (East Africa, Latin America)

To support responsible transfer, we envision a lightweight protocol for adapting the co-pilot to new regions such as East Africa or the Andes. Step 1 is a scoping phase with local partners to inventory existing phone/IVR systems, languages, and community health structures. Step 2 is a minimal pilot with synthetic or low-stakes scenarios, focused on validating language packs, voice scripts, and error categories with local workers. Step 3 is a staged rollout in which governance artifacts (Model Card, Datasheet, risk register) are forked and updated to reflect the new context, rather than reusing India-specific assumptions.

Throughout, we treat ADI and SDI as *portable* but not context-free: strata and thresholds are redefined with local advisors, and transfer is viewed as an opportunity to surface new axes of disparity (e.g., rural–peri-urban, dialectal variation) that may not be salient in the original deployment.

5.6 Measures & outcomes

Primary outcomes.

1. **Fairness.** Between–group gaps in TPR/FPR and Equalized Odds (EO). For group $g \in \{0, 1\}$ at session s, $\text{TPRgap}_s = |\text{TPR}_{g=1,s} - \text{TPR}_{g=0,s}|$ and similarly for FPRgap_s.
2. **Trust / use.** Validated trust/affect scales; usage telemetry: *accept*, *override*, *second–look*.
3. **Skill drift / lift.** (i) *Override selectivity*: share of overrides concentrated on hard/ambiguous items; (ii) *Error detection (Hi–Conf.)*: fraction of high–confidence AI errors detected and corrected by humans; (iii) *SDI* (Sec. 4.2): net change in unaided performance over sessions.

Secondary outcomes. Task time; disagreement rates; post–hoc calibration metrics (Brier/ACE); and robustness of decisions after counter–arguments.

Trust repair (latency). Following an assistance failure at time t_0, we measure time-to-accept the next correct suggestion (median seconds/items). Shorter latency indicates faster trust repair.

Team coordination and conflict markers. We treat coordination quality as a secondary but important outcome. From voice logs and observer coding (where available), we will derive counts of interruptions, overlapping turns, and repeated explanations during aided decision points. Higher rates of overlapping speech or repeated justifications can indicate friction around the co-pilot's suggestions (e.g., disagreements between juniors and seniors, or between workers and the system). We will summarize these markers by assistance arm and session, and test whether designs that foreground challenge prompts and rationales reduce conflictual patterns while preserving appropriate dissent.

Dynamic fairness. We track session–level *fairness degradation rate* FDR across S sessions:

$$\text{FDR} = \frac{1}{S-1} \sum_{s=2}^{S} \Big(\text{TPRgap}_s + \text{FPRgap}_s - (\text{TPRgap}_{s-1} + \text{FPRgap}_{s-1}) \Big).$$

We use FDR alongside our fairness–aware objective (Sec. 4.5) to adapt assistance online.

Access–aware fairness (ADI). Let M index access modalities (connectivity tier, device class, language). For each M, let $\Delta_{\text{EO}}^{(M)}$ be the EO gap. Define $\text{ADI}_s = \max_M |\Delta_{\text{EO},s}^{(M)}|$ and $\Delta\text{ADI}_s = \text{ADI}_s - \text{ADI}_{s-1}$.

Calibration/reliability. We report Expected Calibration Error (ECE), Brier score, and ACE overall and per stratum M. Reliability diagrams are stratified by modality to visualize miscalibration across access conditions.

5.6.1 Voice-specific fairness: ASR quality across accents and regions

Because part of the co-pilot runs over voice and IVR, fairness also depends on automatic speech recognition (ASR) quality. We therefore plan to log simple ASR quality proxies (e.g., intent recognition success, need for repetition, fallback to DTMF) by language, dialect/region, and connectivity tier. Where feasible, a small, consented subset of calls will be manually coded for transcription errors and intent mismatches.

We will report ASR-sensitive fairness metrics such as EO gaps and ADI restricted to voice interactions, and compare them to text-only sessions. If certain dialects or regions exhibit systematically worse ASR quality and downstream assistance performance, this will trigger both technical mitigation (e.g., updated language models) and governance escalations via the advisory panel.

5.7 Analysis

Design. We plan a preregistered multi–arm RCT with teams performing N tasks per session over S sessions. Randomization at team×session assigns assistance *style* $\mathcal{T} \in \{\textsc{None}, \textsc{Rationale}, \textsc{Calib}, \textsc{Counter}\}$.

Intent–to–treat (ITT). For outcome Y_{is} (team i, session s), we estimate

$$Y_{is} = \alpha + \sum_{t \in \mathcal{T} \setminus \{\textsc{None}\}} \beta_t \mathbb{1}[T_{is} = t] \; + \; \gamma_i \; + \; \delta_s \; + \; \varepsilon_{is},$$

with team fixed effects γ_i, session fixed effects δ_s, and cluster–robust standard errors at the team level.

Heterogeneous effects (HTE). We learn CATEs $\tau(x)$ using generalized/causal forests with honesty and out–of–bag risk, and report strata by baseline accuracy, group, and access modality M.

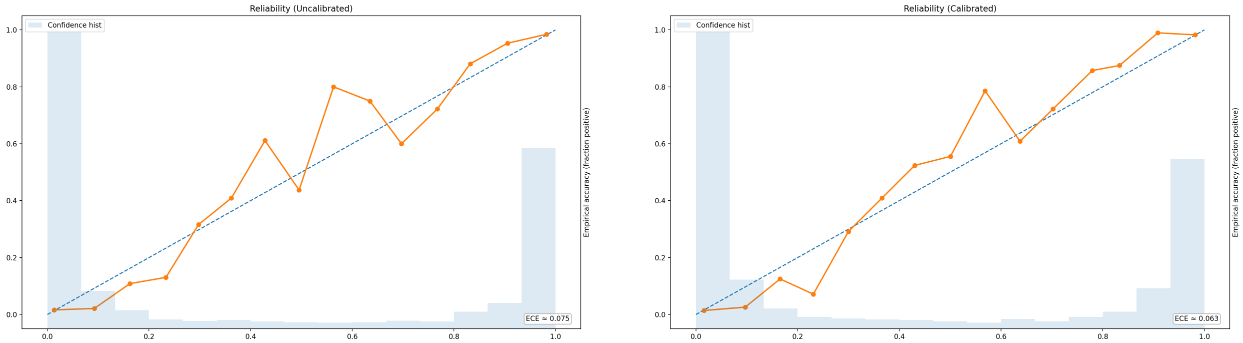
Doubly–robust estimation (DML). For selected outcomes, we fit outcome $m(x, t)$ and propensity $\hat{e}(x)$ models (neural nets), form orthogonalized scores, and cross–fit to obtain ATE/HTE estimates that are robust to high–dimensional nuisance functions [24]. Neural causal baselines include TARNet/DragonNet and CEVAE [25–27].

Mediation. To test whether *trust* or *effort* mediates effects, we specify $T \to Z \to D \to Y$ with Z a trust/effort proxy learned from interaction signals. We plan to report natural direct and indirect effects with sensitivity analysis; DAG assumptions are summarized in Fig. 5.

Multiple testing and sensitivity. We apply Benjamini–Hochberg FDR control across primary outcomes, inspect session-level learning curves, and consider per–team random effects to assess robustness.

Power sketch. Assuming a baseline TPR gap of 0.12, session SD 0.20, two-sided $\alpha = 0.05$, and intra-team ICC $\rho \approx 0.08$, the design effect is $DE = 1 + (m-1)\rho$ with m sessions per team. For m=10 and N_{teams}=20, the effective $n_{\text{eff}} \approx \frac{20\times10}{1+9\times0.08} \approx 67$ session-blocks per arm, yielding MDE ≈ 0.06 on TPR gap at 80% power. Exact counts and a resampling script will appear in the artifact (e.g., `/notebooks/power.ipynb`).

5.8 Reliability (Calibration) Plots



(a) Before (uncalibrated)

(b) After (calibrated)

Figure 6: Reliability diagrams (calibration curves). Calibration improves alignment to the diagonal; ECE is shown in-panel.

6 Causal Estimation

We plan to estimate effects of assistance style $T \in \{\textsc{None}, \textsc{Rationale}, \textsc{Calib}, \textsc{Counter}\}$ on outcomes (accuracy, fairness gaps, overrides) using modern causal estimators.

6.1 Doubly-robust / Double Machine Learning (DML)

We will compute ATEs and HTEs using orthogonalized (Neyman-orthogonal) scores with cross-fitting: flexible outcome models $\hat{m}(x,t)$ and propensity models $\hat{e}(x)$ (neural nets) are learned, then plugged into a doubly-robust score to obtain $\hat{\tau}$. This yields valid inference under mild conditions even when nuisance models are high-dimensional; see Chernozhukov et al. for the formal framework [28].

6.2 Neural causal baselines

We benchmark against deep representation and latent-variable approaches: TARNet/DragonNet (representation learning with targeted regularization) and CEVAE (latent-confounder VAE) for individual and average treatment effects [29–31].

6.3 Mediation and mechanisms

To test whether *trust* or *effort* mediates effects, we specify a small structural model $T \to Z \to D \to Y$ with Z (trust/effort proxy) learned from interaction signals. We plan to report natural direct and indirect effects with sensitivity analysis; DAG assumptions are shown in Fig. 5.

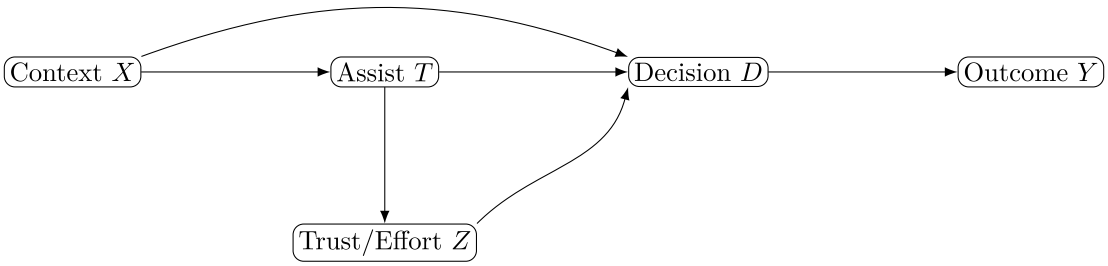


Figure 5: Conceptual structural model of assisted decisions used for mediation/identification.

Table 2: Planned ablations. Higher is better for **Accuracy**; lower is better for **EO gap** and **ECE**.

Variant	**Accuracy** ↑	**EO gap** ↓	**ECE** ↓
Full (Fair+DirCal+Bandit)	.XX	.XX	.XX
w/o fairness penalty	.XX	.XX↑	.XX
w/o Dirichlet calibration	.XX	.XX	.XX↑
Static assistance (no bandit)	.XX	.XX	.XX

Template shown with placeholder values; actual results will be filled in once models are trained on the study data.

6.4 Robustness

We will report sensitivity to unobserved confounding (e.g., Rosenbaum-style bounds), overlap diagnostics, and placebo tests using pre-treatment outcomes. For sequential sessions, effects are summarized per session and aggregated with cluster-robust standard errors; dynamic parity is tracked via FDR (Sec. 4.3).

6.5 Sensitivity analysis for community-level spillovers

Because ASHAs operate in overlapping social and geographic networks, treatment assignment at the team×session level may induce spillovers: assistance styles used by one team could affect expectations, norms, or informal advice channels for others. While our primary estimands treat interference as limited, we will probe sensitivity to community-level spillovers in two ways.

First, we will construct village- or cluster-level summaries of treatment exposure (e.g., share of sessions using Calib or Counter) and include these as additional covariates in outcome models, checking whether estimates of direct effects change materially. Second, we will explore simple partial-interference specifications that treat villages as clusters and compare results to the main DML and forest-based analyses. These checks will not fully identify spillover structures but will help bound how much our conclusions rely on strict no-interference assumptions.

7 Results

Planned analyses. Because this is a design and preregistration brief, we report *planned* rather than realized results. We will summarize treatment effects with 95% confidence intervals, reliability diagrams (calibration curves), usage trajectories over sessions, and HTE plots stratified by experience and resource access. Robustness tables will include placebo checks, overlap diagnostics, and per-session trends for fairness and calibration metrics (EO gaps, ADI, ECE).

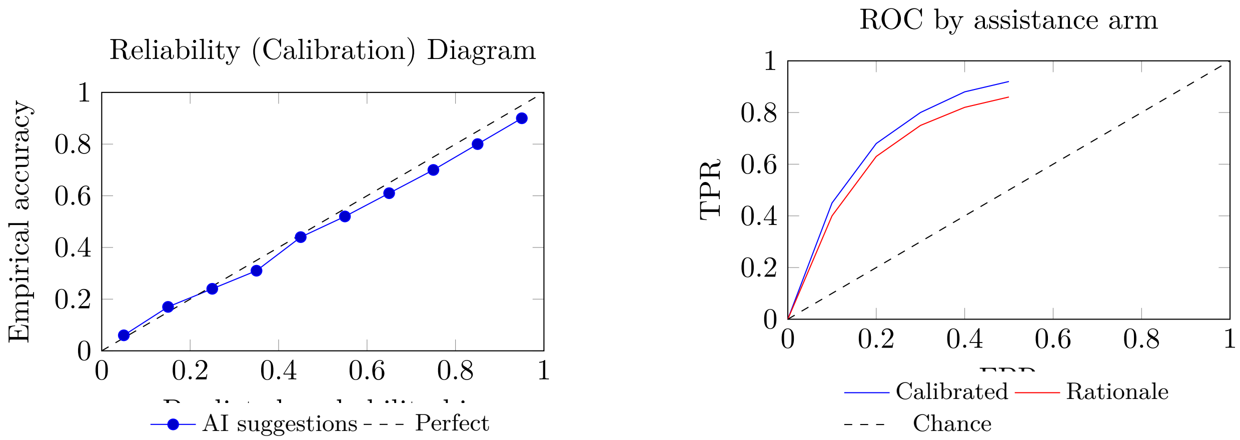


Figure 7: Planned reliability (left) and ROC (right) plots by assistance arm (illustrative placeholders with synthetic values).

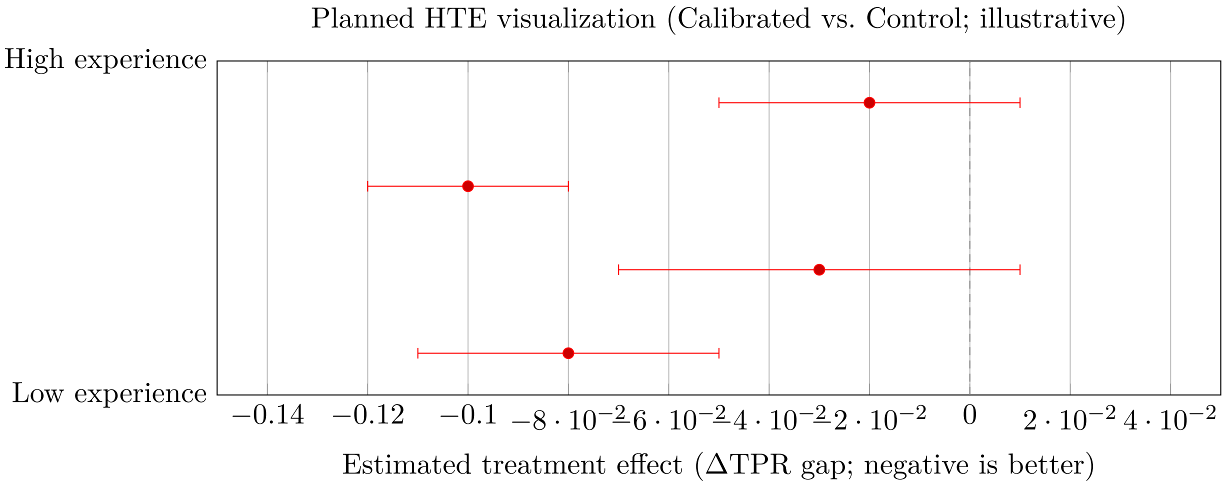


Figure 8: Illustrative HTE forest plot across experience and resource strata. Points show planned effect estimates; whiskers show 95% CIs; dashed line at 0.

8 Discussion

Voice-first pragmatics. Our design is motivated by ICTD findings that voice/IVR and WhatsApp often remain the most robust substrates for equitable access under intermittent connectivity and heterogeneous devices. Prior work shows that voice-first systems and WhatsApp-based workflows can help overcome literacy, device, and connectivity barriers in low-resource settings. In our planned study, we will compare assistance delivered via voice-first flows and text-centric flows, with the expectation that well-calibrated, co-designed voice/IVR support can improve both measured performance and perceived usability.

Design levers. The system surfaces calibration and challenge prompts as explicit design levers: uncertainty is made visible, and counter-evidence is presented through a dedicated "second look" panel rather than being buried in logs. If the planned experiment confirms that these features improve equity and error detection, this would suggest that teams benefit when uncertainty is explicit and counter-arguments are salient at the moment of decision.

Table 3: Planned access-strata outcomes for a 12-week deployment (template; values illustrative). Higher Acc is better; lower EO gap/ECE/ADI is better.

Stratum	Acc	EO gap	ECE	ADI
Feature phone / 2G	.XX	.XX	.XX	.XX
Smartphone / 3G	.XX	.XX	.XX	.XX
Kannada (voice/IVR)	.XX	.XX	.XX	.XX
Hindi (text UI)	.XX	.XX	.XX	.XX

Algorithmic dependency vs. skill reinforcement. The planned study is motivated by a tension between algorithmic dependency and skill reinforcement. If workers come to rely on high-confidence suggestions as authoritative, unaided skill may quietly erode, and junior staff may never build the tacit judgment that senior colleagues possess. Conversely, if assistance is framed as a structured practice aid—with AI-off windows, second-look prompts, and explicit rationales—it may help codify good habits and make expert heuristics more widely accessible.

By tracking SDI and related indicators over sessions, we aim to distinguish these regimes empirically. In low-resource settings, the stakes are high: if AI co-pilots induce dependency without building underlying capacity, they risk deepening vulnerability when connectivity fails, models drift, or funding ends. Our design therefore treats skill reinforcement, not just short-run accuracy, as a first-order goal.

Why ADI and verbal accountability matter. Static parity can mask exclusion in low-connectivity strata; our access disparity index (ADI) is intended to surface where access drives harm rather than simply model predictions. In voice-first flows, accountability must be audible rather than visual: rationale read-backs and challenge prompts are designed to make "second looks" explicit and logged, potentially improving error detection without adding visual UI overhead.

Governance. Our RMF/ISO alignment is intended to illustrate how an organizational AI management system (AIMS) could integrate logging, monitoring, and human-in-the-loop controls for a future production deployment [17, 20]. By tying fairness metrics (EO gaps, ADI), calibration (ECE), and override logging to NIST AI RMF functions (Govern–Map–Measure–Manage), the design highlights how responsible rollout and oversight can be embedded into the technical stack from the outset.

Socio-emotional design implications. Human–robot interaction and group dynamics research suggests that assistance systems are not only informational tools but also social actors: apology timing, gaze, and turn-taking cues can influence transient trust and conflict, which in turn shape downstream performance and adherence. Our co-pilot design adopts this lens by treating challenge prompts, rationales, and (in future work) embodied cues as socio-emotional signals, with the goal of supporting appropriate reliance and reducing unnecessary friction in team decision making.

Post-study handoff: building local capacity for fairness monitoring. A central design goal is that, after the formal study ends, local partners should be able to continue monitoring and adapting the co-pilot without relying on the original research team. The governance console and artifact package are therefore structured as capacity-building tools: dashboards expose interpretable metrics; model cards and datasheets document assumptions; and the NIST/ISO mappings point to concrete operational practices (e.g., incident response, prompt review).

During the deployment, we envision joint review sessions where NGO staff and ASHAs walk through metrics and logs together, practice interpreting trends, and decide on follow-up actions (e.g., revising challenge scripts, retuning thresholds, or pausing assistance in certain strata). The intent is that, by the time of handoff, community members have both the artifacts and the experience needed to exercise ongoing oversight, turning a one-off research deployment into a locally governed AI management system.

Ethics limitations. We plan to obtain ethics/IRB approval before any field data collection and to secure informed consent from all participants in their preferred language, with audio options for low-literacy users. Deployment logs would be de-identified (non-identifying device IDs; no free text without explicit consent) and stored under standard security controls. In line with NIST AI RMF guidance, we will document residual risks (e.g., automation bias, access-driven disparities, domain shift across regions or teams) and clearly state the limits of generalization for any findings (e.g., specific CHW programs, languages, and connectivity patterns).

WhatsApp misinformation dynamics and trust calibration. Because our setting sits inside dense WhatsApp ecosystems, trust in the co-pilot will be shaped by broader experiences with mobile information flows. Prior work documents how health advice, rumor, and misinformation circulate through family and community groups, often blurring boundaries between official and informal sources. Our design therefore treats trust calibration as a socio-technical problem: we frame the co-pilot explicitly as a tool endorsed by the NGO and health system, use consistent branding across IVR and WhatsApp channels, and avoid anthropomorphic cues that could make the system seem like an infallible "expert."

In the planned study, we will interpret trust and reliance measures in light of these dynamics, looking for cases where over-trust may mirror existing patterns of deference to forwarded audio messages or videos. Over time, community advisory panels can help recalibrate how the system is introduced and described in local languages to avoid reinforcing harmful misinformation norms.

A Appendix - Artifact & Reproducibility Package

Current artifact (GitHub only). We provide a small, public GitHub repository with fully synthetic data and analysis scripts aligned to this brief. The goal is to show what an "evidence-ready" package could look like for a future field study—without exposing any real deployment data or personal information.[1]

Repository.

- **GitHub (code, synthetic data, notebook, environment files)**:

What is included (high level). The repository is structured as follows (directory names may be shortened in the PDF for readability):

- **data/**
  Contains only synthetic CSVs:

[1] All records in the repository are synthetic and constructed for demonstration only.

  - `data/raw/sample_decisions_v2.csv`: row-level synthetic decision logs with fields such as worker ID, session, group, access tier, assistance arm, AI confidence, and human accept/override indicators.
  - `data/processed/session_metrics_v2.csv`: session × group × access-tier summary metrics (accuracy, TPR/FPR, ECE, counts).

- **src/**
  Minimal Python scripts used to generate and analyze the synthetic data:

  - `generate_synthetic_data.py` — builds the synthetic dataset, saving `sample_decisions_v2.csv` and `session_metrics_v2.csv` under `data/`.
  - `02_compute_metrics_and_plots.py` — loads the processed metrics, computes fairness and calibration summaries, and writes a small results table plus illustrative reliability / HTE plots into `figures/`.
  - `03_causal_and_power_demo.py` — sketch of how one might plug the synthetic data into causal / power–analysis routines (e.g., DML-style estimation and simple variance calculations), consistent with Secs. 5.7, 6.

- **notebooks/**
  `01_explore_synthetic_data.ipynb`: a small Jupyter notebook for interactively inspecting the synthetic dataset, checking distributions, and reproducing the summary tables corresponding to the measures in Sec. 5.6.

- **figures/**
  Contains PNGs generated from the analysis scripts, e.g., `synthetic_calibration_curves.png` and `synthetic_hte_forest.png`, which mirror the style of the calibration and HTE plots in the main text.

- **requirements.txt** and **environment.yml**
  Lightweight environment specifications (Python 3.9+; standard scientific Python stack) so that the synthetic analysis can be re-run in a clean virtual environment.

- **LICENSE**
  An MIT-style license granting permission to inspect, reuse, and adapt the synthetic materials.

Tagged versions. The GitHub repository uses semantic tags to mirror the staged evolution of the artifact described in this brief:

- **v1.0.0** — Minimal demonstrator: small synthetic dataset and a basic summary script showing accuracy and simple group gaps.

- **v2.0.0** — Richer synthetic data (more workers, sessions, and items) and a first pass at fairness and calibration metrics (TPR/FPR gap, EO gap, ECE) aligned with Sec. 5.6.

- **v3.0.0** — Reorganized code into a clearer data → metrics → plots pipeline and added a small causal / power-analysis sketch consistent with Secs. 5.7, 6.

None of these versions contain real human data; they are intended purely as a template for how a future deployment package could be structured.

Planned extensions (not yet implemented). To avoid over-claiming, we treat the following as *planned* rather than current:

- An OSF preregistration record that would host the pre–analysis plan and a PDF export of the registration (App. B).
- A Zenodo (or similar) archival snapshot of a future, de-identified deployment log, with semantic versioning and checksums.

These are described as part of the envisioned research workflow but are not yet live at the time of this writing.

B Preregistration (Planned)

Registry and template (planned). If the study is implemented as a field RCT, we plan to preregister hypotheses, arms, outcomes, and analysis on OSF using an AsPredicted-style template for social/behavioral experiments. The preregistration would point to the GitHub artifact in Appendix A and freeze the main estimands and analysis plan prior to inspecting outcomes.

Scope (planned). Units: teams × sessions; blocked randomization by baseline accuracy. Arms: None, Rationale, Calib, Counter. Primary outcomes: EO gap ($\max\{\Delta\mathrm{TPR}, \Delta\mathrm{FPR}\}$), TPR gap, ECE, Override Selectivity, Error Detection (Hi-Conf.), SDI. Secondary outcomes: time, disagreement, ACE.

Estimands and models (pre-specified, planned). ITT with team and session fixed effects; cluster-robust SEs. HTE via generalized / causal forests with honesty; strata by baseline accuracy, group, and access modality M. DML for selected outcomes (orthogonalization; cross-fitting). Multiple testing: BH FDR. Sensitivity: per-session learning curves; per-team random effects. Power sketch: baseline $\mathrm{TPR}_{\mathrm{gap}} = 0.12$, $\mathrm{SD} = 0.20$, $\alpha = 0.05$, power $= 0.80$ ($\mathrm{MDE} \approx 0.06$).

The preregistration itself is not yet filed for this writing sample; instead, this appendix describes the planned structure and estimands to keep the proposal aligned with best practices.

C Model Card: ADI Co-pilot Assistance System

Model overview. Encoder–decoder foundation model with a small task adaptor for triage decisions. Outputs class probabilities and calibrated confidence scores for suggested actions (e.g., escalate, follow-up call, self–care advice). Deployed as an assistance system for community health workers (CHWs), not for diagnosis.

Intended users and use cases. Primary users: Accredited Social Health Activists (ASHAs) and supervisors in low–resource settings. Intended use: screening support and referral triage during maternal/child–health workflows under intermittent connectivity. The system is designed to *augment*, not replace, human judgment; final decisions remain with the CHW/supervisor.

Out–of–scope uses. Not approved for autonomous diagnosis, emergency–only triage, or direct-to-patient use without a trained intermediary. Not validated outside the deployment context described in Sec. 5.1.

Factors and evaluation strata. Key strata: protected group A (e.g., community/region); access modality M (feature–phone/2G, smartphone/3G, WhatsApp/SMS, voice/IVR); experience level (novice vs. senior CHWs). Metrics are reported by group and modality (EO gaps, ECE, ADI, SDI; see Sec. 5.6).

Metrics. Primary: accuracy, EO/TPR/FPR gaps, Expected Calibration Error (ECE), Access–aware Disparity Index (ADI), Skill–Drift Index (SDI). Secondary: task time, override selectivity, high–confidence error detection, disagreement rates (Sec. 5.6).

Data and training. In a real deployment, models would be trained on de–identified historical triage logs and synthetic augmentations; no raw audio/text would be stored without explicit consent (Sec. 5.4). Training/validation splits would respect site and session structure. For this brief, only synthetic demo data and scripts are provided in the artifact package (App. A).

Limitations. Performance may degrade under domain shift (new symptoms, new regions/languages), rare classes, and unobserved confounding in treatment uptake. Fairness metrics focus on observed groups and access strata; unmeasured axes of disparity may remain.

Risk controls and governance. Controls include calibrated probabilities, challenge prompts, override logging, offline cache, access–aware fairness monitoring (EO, ADI, FDR), and monthly community–panel review of prompts and rationales (Sec. 3, 5.4, 5.3). This appendix links to governance evidence tables in App. D.

D Planned Governance Mapping: NIST AI RMF & ISO/IEC 42001

Table D.1: NIST AI RMF 1.0 mapping (Functions → paper/artifact evidence).

AI RMF Function	Evidence (location in paper/artifacts)
Govern	Planned IRB/consent, community oversight, and handoff plan (5.4); envisioned community advisory panel; model/documentation ownership fields specified in the planned Model Card (Appendix C).
Map	Context, stakeholders, tasks, languages, access modalities and constraints (5.1, 5.3.1, 5.5); risk/impact framing for low-resource deployment.
Measure	Pre-specified outcomes and metrics (EO gap, ECE, ADI, SDI) (5.6); reliability plots; power/sensitivity and analysis plan (5.7); preregistration sketch (Appendix B).
Manage	Planned online adaptation (4.5); fallbacks (IVR/SMS, offline-first caching) (3, 3.4); monitoring of parity/ECE/ latency (3); deployment playbook sketched in the GitHub README (Appendix A).

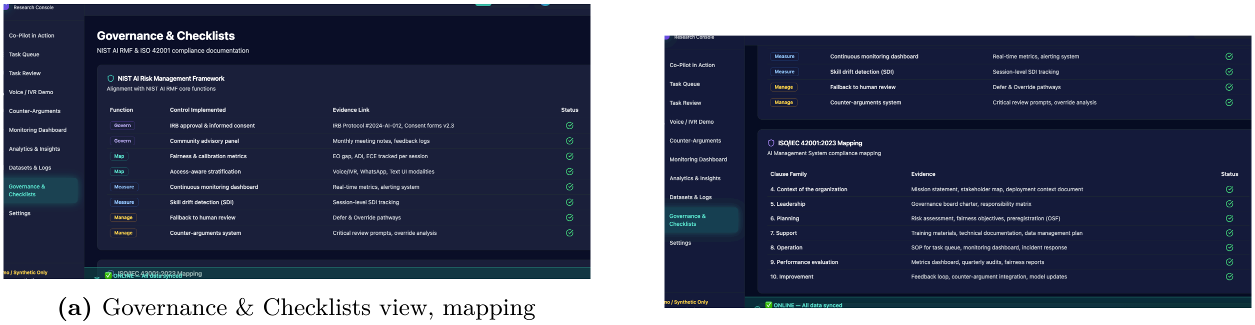
Shorthands: EO = Equalized Odds; ECE = Expected Calibration Error; ADI = Access-aware Disparity Index; SDI = Skill-Drift Index.

Scope. This appendix sketches how our proposed artifacts and methods could map to the *NIST AI Risk Management Framework (AI RMF 1.0)* functions (Govern, Map, Measure, Manage) and *ISO/IEC 42001:2023* clause families in a future field deployment.

Table D.2: ISO/IEC 42001:2023 mapping (AIMS clauses → paper/artifact evidence).

Clause family	Evidence (location in paper/artifacts)
Context of the organization	Stakeholders, scope, communities, constraints (5.1, 5.5).
Leadership & governance	Planned roles, ownership, advisory panel, and accountability mechanisms (5.4; App. C).
Planning (risks & objectives)	Risk register (connectivity, over-reliance), mitigation and KPIs (EO gap/ECE/ADI), documented in the artifact and governance evidence tables (App. A, App. D).
Support (competence, docs)	Training for local champions (planned); documentation sets (artifact README, Model Card) (App. A, App. C).
Operation (lifecycle)	Planned data pipelines, calibration, online learner, and fallbacks (3, 3.4).
Performance evaluation	Monitoring dashboards/telemetry, preregistered outcomes (planned), and planned ablations (5.7).
Improvement	Drift detection and iterative updates (3, 4.5); monthly prompt reviews (5.4).

AIMS = AI Management System. Rows reflect clause families commonly summarized for ISO/IEC 42001 evidence.



Governance & Checklists

NIST AI RMF & ISO 42001 compliance documentation

NIST AI Risk Management Framework

Alignment with NIST AI RMF core functions

Function	Control Implemented	Evidence Link	Status
Govern	IRB approval & informed consent	IRB Protocol #2024-AI-012, Consent forms v2.3	✓
Govern	Community advisory panel	Monthly meeting notes, feedback logs	✓
Map	Fairness & calibration metrics	EO gap, ADI, ECE tracked per session	✓
Map	Access-aware stratification	Voice/IVR, WhatsApp, Text UI modalities	✓
Measure	Continuous monitoring dashboard	Real-time metrics, alerting system	✓
Measure	Skill drift detection (SDI)	Session-level SDI tracking	✓
Manage	Fallback to human review	Defer & Override pathways	✓
Manage	Counter-arguments system	Critical review prompts, override analysis	✓

ISO/IEC 42001:2023 Mapping

(a) Governance & Checklists view, mapping implemented controls to NIST AI RMF functions (Govern, Map, Measure, Manage).

Measure	Continuous monitoring dashboard	Real-time metrics, alerting system	✓
Measure	Skill drift detection (SDI)	Session-level SDI tracking	✓
Manage	Fallback to human review	Defer & Override pathways	✓
Manage	Counter-arguments system	Critical review prompts, override analysis	✓

ISO/IEC 42001:2023 Mapping

AI Management System compliance mapping

Clause Family	Evidence	Status
4. Context of the organization	Mission statement, stakeholder map, deployment context document	✓
5. Leadership	Governance board charter, responsibility matrix	✓
6. Planning	Risk assessment, fairness objectives, preregistration (OSF)	✓
7. Support	Training materials, technical documentation, data management plan	✓
8. Operation	SOP for task queue, monitoring dashboard, incident response	✓
9. Performance evaluation	Metrics dashboard, quarterly audits, fairness reports	✓
10. Improvement	Feedback loop, counter-argument integration, model updates	✓

(b) ISO/IEC 42001:2023 mapping panel showing AI management system evidence by clause family.

Figure D.1: Governance console connecting the prototype to NIST AI RMF and ISO/IEC 42001 evidence.

E NIST AI RMF / ISO 42001 Checklist (Abbrev.)

Framing. We sketch how our *proposed* governance controls would map to the NIST AI Risk Management Framework 1.0 functions (*Govern, Map, Measure, Manage*) and indicate touchpoints with an AI management system per ISO/IEC 42001:2023.

Table E.1: Abbreviated checklist: NIST functions and matching ISO 42001 touchpoints.

NIST Function	Planned control in this design (ISO/IEC 42001 touchpoint)
Govern	Planned IRB/ethics approval; envisioned community advisory panel; role-based access controls; incident/takedown workflow (ISO 5.2 leadership, 6.2 objectives).
Map	Context definition (target users such as ASHAs, languages, connectivity tiers); intended use limits (non-diagnostic); data lineage for logs (ISO 8.1 operations).
Measure	KPIs: EO/TPR/FPR gaps, ECE/ACE, ADI, drift metrics; planned ITT/HTE/DML analyses; periodic re-runs of key checks (ISO 9.1 monitoring).
Manage	Mitigations in the proposed system: Dirichlet calibration; challenge prompts; second-look workflow; offline cache; planned periodic review loop over metrics and incidents (ISO 10.1 improvement).

Artifact evidence. In the current artifact (Appendix A), we provide a public GitHub repository with synthetic demo data, analysis scripts, and an environment file. As the work matures toward a field deployment, we plan to add explicit policy documents (e.g., a `/governance/` directory), a risk register (`/risk/`), enriched model card and datasheet templates, and CI logs to support external audit and reproducibility.

References

[1] Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *Annals of Statistics*, 47(2):1148–1178, 2019. URL https://projecteuclid.org/journals/annals-of-statistics/volume-47/issue-2/Generalized-random-forests/10.1214/18-AOS1709.

[2] Susan Athey and Stefan Wager. Estimating treatment effects with causal forests: An application. *arXiv:1902.07409*, 2019. URL https://arxiv.org/abs/1902.07409.

[3] Vasilis Syrgkanis et al. Causal inference and machine learning in practice with econml and causalml (tutorial). In *KDD '21*, 2021. doi: 10.1145/3447548.3470792. URL https://kdd.org/kdd2021/tutorials.html.

[4] Aditya Vashistha and et al. Sangeet swara: A community-moderated voice forum in rural india. In *CHI '15*. ACM, 2015. doi: 10.1145/2702123.2702341.

[5] Aditya Vashistha and Agha Ali Raza. Voice interfaces for underserved communities. In *An Introduction to Development Engineering*. Springer, 2021. URL https://www.adityavashistha.com/uploads/2/0/8/0/20800650/deveng-springer-2021.pdf.

[6] GSMA. The state of mobile internet connectivity 2024. https://www.gsmaintelligence.com/research/the-state-of-mobile-internet-connectivity-2024, 2024.

[7] Brian DeRenzi and et al. Designing for community health workers: In-situ participatory design in india. In *CHI '17*. ACM, 2017. doi: 10.1145/3025453.3025923.

[8] Niharika Varanasi and et al. Understanding whatsapp practices in india: Misinformation, trust, and social norms. In *CHI '22*. ACM, 2022. doi: 10.1145/3491102.3501890.

[9] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *NeurIPS*, 2016. URL https://papers.neurips.cc/paper/6374-equality-of-opportunity-in-supervised-learning.pdf.

[10] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *FAT**, 2019. doi: 10.1145/3287560.3287596. URL https://arxiv.org/abs/1810.03993.

[11] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *arXiv:1803.09010*, 2018. URL https://arxiv.org/abs/1803.09010.

[12] Linda J. Skitka, Kathleen L. Mosier, and Mark D. Burdick. Does automation bias decision-making? *International Journal of Human-Computer Studies*, 51(5):991–1006, 1999. doi: 10.1006/ijhc.1999.0252.

[13] Kathleen L. Mosier and Linda J. Skitka. Automation bias: decision making and performance in high technology environments. *International Journal of Human-Computer Studies*, 1997.

[14] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 2015. URL https://marketing.wharton.upenn.edu/wp-content/uploads/2016/10/Dietvorst-Simmons-Massey-2014.pdf.

[15] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. Overcoming algorithm aversion: People will use imperfect algorithms if they can modify them. *Management Science*, 2018. URL https://faculty.wharton.upenn.edu/wp-content/uploads/2016/08/Dietvorst-Simmons-Massey-2018.pdf.

[16] PyWhy / DoWhy. Tutorial: Using dowhy + econml for causal inference, 2024. URL https://www.pywhy.org/dowhy/v0.11/example_notebooks/tutorial-causalinference-machinelearning-using-dowhy-econml.html.

[17] National Institute of Standards and Technology. Artificial intelligence risk management framework (ai rmf 1.0), 2023. URL https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf.

[18] National Institute of Standards and Technology. Nist ai rmf playbook. https://www.nist.gov/itl/ai-risk-management-framework/nist-ai-rmf-playbook, 2023.

[19] National Institute of Standards and Technology. Artificial intelligence risk management framework: Generative ai profile (nist.ai.600-1), 2024. URL https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf.

[20] International Organization for Standardization. Iso/iec 42001:2023 — artificial intelligence management systems (aims), 2023. URL https://www.iso.org/standard/42001.

[21] Aditya Vashistha and Nithya Sambasivan. Designing voice-based social computing for low-resource communities. In Karthik Srinivasan et al., editors, *Voice-Based Social Media*. Springer, 2021. Overview of IVR/voice-first systems in low-resource settings.

[22] Aditya Vashistha, Edward Cutrell, Gaetano Borriello, and William Thies. Sangeet swara: A community-moderated voice forum in rural india. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*, pages 417–426, Seoul, Republic of Korea, 2015. ACM. doi: 10.1145/2702123.2702191.

[23] Reuters. Telecom industry coalition to boost access to smartphones in poor countries. https://www.reuters.com/business/media-telecom/telecom-industry-coalition-boost-access-smartphones-poor-countries-2024-07-10/, 2024.

[24] Victor Chernozhukov and et al. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 2018. doi: 10.1111/ectj.12097.

[25] Uri Shalit, Fredrik Johansson, and David Sontag. Estimating individual treatment effect: Generalization bounds and algorithms. In *ICML '17*, 2017.

[26] Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. In *NeurIPS '19 Workshop*, 2019.

[27] Christos Louizos and et al. Causal effect inference with deep latent-variable models. In *NeurIPS '17*, 2017.

[28] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018. doi: 10.1111/ectj.12097.

[29] Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: Generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pages 3076–3085, Sydney, Australia, 2017. PMLR. URL https://proceedings.mlr.press/v70/shalit17a.html.

[30] Claudia Shi, David M. Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. *arXiv preprint arXiv:1906.02120*, 2019. URL https://arxiv.org/abs/1906.02120.

[31] Christos Louizos, Uri Shalit, Joris M. Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017. URL https://papers.nips.cc/paper/2017/hash/076a0c97d09cf1a0ec3e19c7f2529f2b-Abstract.html.