Confounding-Robust Deferral Policy Learning

Ruijiang Gao[1], **Mingzhang Yin**[2]

[1]Naveen Jindal School of Management, University of Texas at Dallas, Richardson, TX 75082
[2]Warrington College of Business, University of Florida, Gainesville, FL 32611
ruijiang.gao@utdallas.edu, mingzhang.yin@warrington.ufl.edu

Abstract

Human-AI collaboration has the potential to transform various domains by leveraging the complementary strengths of human experts and Artificial Intelligence (AI) systems. However, unobserved confounding can undermine the effectiveness of this collaboration, leading to biased and unreliable outcomes. In this paper, we propose a novel solution to address unobserved confounding in human-AI collaboration by employing sensitivity analysis from causal inference. Our approach combines domain expertise with AI-driven statistical modeling to account for potentially hidden confounders. We present a deferral collaboration framework for incorporating the sensitivity model into offline policy learning, enabling the system to control for the influence of unobserved confounding factors. In addition, we propose a personalized deferral collaboration system to leverage the diverse expertise of different human decision-makers. By adjusting for potential biases, our proposed solution enhances the robustness and reliability of collaborative outcomes. The empirical and theoretical analyses demonstrate the efficacy of our approach in mitigating unobserved confounding and improving the overall performance of human-AI collaborations.

1 Introduction

In recent years, policy learning has emerged as a powerful tool for learning and optimizing decision-making policies across a diverse range of applications, including healthcare, finance, and marketing (Imbens 2024). One of the most promising avenues for leveraging machine learning is policy learning on observational data (Athey and Wager 2021), which aims to infer optimal decision rules from historical data without the need for costly randomized experiments. Observational data, generated from real-world systems, is abundant and easily accessible, making it an attractive source for training models that can guide policy decisions.

Many algorithms have been proposed for efficient policy learning from observational data (Joachims, Swaminathan, and de Rijke 2018; Gao et al. 2021b; Kallus 2021), usually under the *unconfoundness* assumption. It assumes no hidden confounders that simultaneously influence both the treatment assignment and individual outcomes (Rubin 1974). This assumption is defensible in certain domains such as automated recommendation or pricing systems (Biggs, Gao, and Sun 2021) where we have full control of the historical algorithm, but may rarely hold true for domains where the observational data is generated by human decision-makers.

Consider a healthcare scenario, where observational data is generated by human experts as the electronic health records (EHRs). These records contain a wealth of information about patients' medical histories, treatments, and outcomes to inform policy learning for personalized medical interventions. However, human experts, such as physicians, may seek additional information when making decisions about patient care, such as the patient's lifestyle, mental well-being, or other contextual factors like the bedside information that might influence their decision-making process as well as the patient's health outcomes. This additional information, though crucial for decision-making, may not be recorded in the EHRs, leading to potential confounding issues in the observed data. For example, a physician may prescribe specific medication to patients with a specific lifestyle, so the observed treatment might be confounded by the potentially unrecorded lifestyle factor. In this case, the unobserved confounding can result in suboptimal actions and reduce the reliability of learned policies. In the causal inference literature, the marginal sensitivity model has been proposed under the unmeasured confounding to bound the possible value of the true propensity score (Tan 2006). This idea was recently applied to policy learning without humans in the loop (Kallus and Zhou 2021).

In this paper, we propose a human-AI collaboration system that learns a policy robust to unmeasured confounding. The system uses a deferral component to decide task allocations to human experts or algorithms. The learned policy improves over the algorithm-alone and the human-alone approaches. Supposing the historical data in our motivating example are all generated by human decision-makers, an AI-only algorithm is likely to be inferior to humans in cases where external information, such as patients' lifestyles, is necessary for optimal decision-making. The benefit of human involvement stands out in the confounding setting as human decision-makers are adept at making choices based on (unobserved) confounding factors (Holstein et al. 2023). In contrast, a human-only system often incurs a high operation cost. An essential problem is how to jointly learn a rule to choose decision-makers and a rule to assign treatment once the AI system is chosen, especially when the observed data have

missing confounders. We refer to this problem as *deferral collaboration under unobserved confounding* (Gao et al. 2023).

By adopting a human-AI collaborative approach, we can alleviate the impact of these unobserved confounders in the traditional deferral collaboration. In addition, the external information of human experts can be leveraged by the AI system beyond the observed data to obtain a more accurate estimate of the optimal policy. This collaborative framework ensures that the learned policies better account for the missing confounders and yield more reliable decision-making.

We make the following contributions in the paper: We are the first to propose leveraging the learning-to-defer framework to tackle the policy learning under unobserved confounding problem. We propose a novel algorithm for the problem of *deferral collaboration under unobserved confounding*, where our algorithm works under an uncertainty set over the nominal propensity scores. The proposed algorithm leverages human decision-makers who have the capacity to acquire additional unrecorded information to aid their decision-making and a trained algorithmic policy. Theoretically, we prove it is guaranteed to offer policy improvements over a baseline policy based only on the available features or only on the incumbent human policy. In addition, we generalize our algorithm to personalized settings where each instance can be routed to a specific human decision-maker by exploiting the diverse expertise of humans. We theoretically and empirically validate the efficacy of the proposed method.

2 Related Work

Policy Learning with Unconfoundness. Deducing an optimal personalized policy from offline data has been extensively explored in various domains, including e-commerce, contextual pricing, and medicine (Dudík et al. 2014; Athey and Wager 2017; Kallus 2018, 2019; Gao et al. 2021a; Sondhi, Arbour, and Dimmery 2020; Swaminathan and Joachims 2015a). These studies usually assume the historical data were generated by a previous decision-maker, focusing on estimating treatment effects or optimizing an algorithmic policy without human involvement. It is yet underdeveloped for scenarios that could benefit from a combined human-AI team to enhance decision performance.

Sensitivity Analysis. Sensitivity analysis is widely used in causal inference that evaluates unconfoundedness assumption (Cornfield et al. 1959). A popular framework models the confounding effect on the treatment assignment nonparametrically. Among them, the marginal sensitivity model (MSM), generalizing the Rosenbaum sensitivity model (Rosenbaumn 2002), assumes a bound on the odds ratio of the propensity score conditional on the observed variables and a true propensity score conditional on all the confounding variables (Tan 2006). The MSM has been applied in estimating heterogeneous treatment effects (Yin et al. 2021; Jin, Ren, and Candès 2021), robust optimization (Namkoong, Ma, and Glynn 2022; Guo et al. 2022), and policy learning without human in the loop (Kallus and Zhou 2021). We adopt the MSM to quantify the deviation of unconfoundedness in the context of human-AI collaboration.

Human-AI Collaboration. Recent studies on human-AI collaboration methods improved classification performance, such as accuracy and fairness, by capitalizing on the complementary strengths of humans and AI (Bansal et al. 2019; Ibrahim, Kim, and Tong 2021; Wolczynski, Saar-Tsechansky, and Wang 2022). We focus on the setting without human-AI interaction, where decisions are made by either a human or an algorithm. Previous research has also addressed the task of routing instances to either a human or an algorithm (Madras, Pitassi, and Zemel 2018; Wilder, Horvitz, and Kamar 2020; Raghu et al. 2019; De et al. 2020; Wang and Saar-Tsechansky 2020). The primary distinction between these studies and ours is that they explore contexts where the AI's learning task is a conventional supervised classification task while we focus on policy learning. Gao et al. (2021b, 2023) study how to design a deferral collaboration system similar to ours under the unconfoundness assumption, and does not consider the bias due to unmeasured confounding that is often leveraged by humans (Holstein et al. 2023).

3 Confounding-Robust Deferral Policy

3.1 Problem Setup

Assume we have access to the observed tuples $\{X_i, T_i, Y_i\}_{i=1}^N$, where the covariates $X_i \in \mathcal{X}$, the treatment arm $T_i \in \{0, \cdots, m-1\}$, and a scalar outcome $Y_i \in \mathbb{R}$. Using the potential outcome framework, we assume $Y_i = Y_i(T_i)$, *i.e.*, the SUTVA assumption (Rubin 1980). We consider Y as the risk and aim to minimize the risk aggregated over the population. In practice, humans often utilize additional information for decisions. For example, a customer service representative may use emotion information in the phone call to decide the compensation plan, but such information cannot be recorded in the past due to the legacy computer system. We assume such unobserved confounder is U_i and the unconfoundedness assumption would hold if we account for both U_i and X_i. The U_i can be postulated as the unmeasured covariate or as the unobserved potential outcome itself, *i.e.*, $U_i = Y_i(t)$ (Zhao, Small, and Bhattacharya 2019). The data is generated by the human decision maker with the *behavior policy* π_0 as $T_i \sim \text{Categorical}(\pi_0(T_i|X_i, Y_i))$.

Due to the unobserved confounding, the true propensity $\pi_0(t|x, y) := P(T = t|X = x, Y(t) = y)$ generally cannot be identified using the observational data alone. We can only estimate the nominal propensity, denoted as $\tilde{\pi}_0(t|x) := P(T = t|X = x)$. The nominal propensity can be estimated from the observational data using a machine learning classifier such as logistic regression. To quantify the difference between the nominal and true propensity scores incurred by confounding, we adopt the MSM (Tan 2006) to assume an uncertainty set.

Assumption 1 (Marginal Sensitivity Model)**.**

$$\Gamma^{-1} \leq \frac{(1 - \tilde{\pi}_0(T|X)\pi_0(T|X, Y)}{\tilde{\pi}_0(T|X)(1 - \pi_0(T|X, Y))} \leq \Gamma. \tag{1}$$

The MSM quantifies the deviation from the true propensity scores by the scalar parameter $\Gamma \geq 1$. When $\Gamma = 1$, it corresponds to the unconfoundness setup. Γ can be determined using domain knowledge or estimated using empirical data, which we will discuss in Section 4.

Deferral collaboration (Madras, Pitassi, and Zemel 2018; Gao et al. 2021b) considers how to evaluate and learn a routing algorithm $\phi : \mathcal{X} \to [0, 1]$ that assigns tasks to the human decision-makers or AI system, and an algorithmic policy $\pi : \mathcal{X} \to \Delta^m$ that decides the treatment distribution. The element in simplex Δ^m is the probability over the treatment arms and $\phi(X)$ denotes the probability of routing to humans.

The routing algorithm is designed to *complement* human decision-makers. A successful deferral collaboration routes different instances to the entity that is likely to yield the best reward by $\phi(X)$, and it leverages the policy $\pi(X)$ for the instances routed to the AI. The human decision-maker may incur a cost of $C(X)$ for producing a decision on an instance.

In this paper, we consider a general setting with multiple human decision-makers $H \in \{1, \cdots, K\}$. Accordingly, the data is generated by first assigning an instance with covariates X_i to different human decision-makers by the rule $d_0(H_i|X_i) : \mathcal{X} \to \Delta^K$. Each human decision-maker H_i chooses the treatment by the behavior policy $\hat{\pi}_0(T_i|X_i, H_i, Y_i)$. The observed data become $\{X_i, H_i, T_i, Y_i\}_{i=1}^N$. The routing algorithm ϕ is generalized to $\phi : \mathcal{X} \to \Delta^{K+1}$ where $\phi(A|X), \phi(H|X)$ means the probability of routing instances to the algorithm and a specific human expert H. The goal is to learn an optimal routing algorithm ϕ and policy π that minimizes the risk. The process is illustrated in Section 3.2.

3.2 Our Method: Deferral Collaboration with Unobserved Confounding

We first consider the situation of homogeneous human experts who have similar decision performance. The expected team performance can be calculated by the self-normalized Hájek estimator (Swaminathan and Joachims 2015b)

$$\theta(\pi, \phi) = \mathbb{E}\phi(X)(Y + C(X)) + \sum_{t=0}^{m-1} \mathbb{E}\frac{\mathbb{I}(T = t)}{\pi_0(T|X, Y)}\pi(T|X)Y(1 - \phi(X))/\mathbb{E}\frac{\mathbb{I}(T = t)}{\pi_0(T|X, Y)}. \quad (2)$$

Throughout the paper, without further specification, the expectation is with respect to the underlying data distribution. The first term of Eq. (2) is the cost of assigning to human by ϕ and the second term is the cost of assigning to the algorithm with policy π. Note that now the propensity score depends on both X and Y because of the unobserved confounding. The equality is because $\mathbb{E}\frac{\mathbb{1}(T=t)}{\pi_0(T|X,Y)} = 1$ for every t.

Practically, we are often interested in human-AI systems that can outperform either the human, or a candidate algorithmic policy. Suppose in addition there is a baseline policy $\pi_c(T|X)$, such as the never-treat policy $\pi_c(0|x) = 1$ or a candidate algorithmic policy learned from data, that the proposed human-AI system aims to improve upon. The objective can be written as the improvement over $\pi_c(T|X)$,

$$R(\pi, \phi, \pi_c) = \mathbb{E}\phi(X)(Y + C(X)) + \sum_{t=0}^{m-1} \frac{\mathbb{E}\frac{\mathbb{I}(T=t)}{\pi_0(T|X,Y)}Y[(1 - \phi(X))\pi(T|X) - \pi_c(T|X)]}{\mathbb{E}\frac{\mathbb{I}(T=t)}{\pi_0(T|X,Y)}}. \quad (3)$$

Let $\tilde{W}_i := \frac{1}{\hat{\pi}_0(T_i|X_i)}$ and $W_i := \frac{1}{\pi_0(T_i|X_i,Y_i)}$. By the MSM, our key observation is that the true weights W_i are bounded in the uncertainty set $\mathcal{W}_n^\Gamma = \{W : 1 + \Gamma^{-1}(\tilde{W}_i - 1) \le W_i \le 1 + \Gamma(\tilde{W}_i - 1), \forall i = 1, \cdots, n\}$. Hence, the worst-case empirical estimator $\hat{R}_n(\pi, \phi, \pi_c, \mathcal{W}_n^\Gamma) = \sum_{t=0}^{m-1}$

$$\frac{\frac{1}{n}\sum_i \mathbb{I}(T_i = t)[(1 - \phi(X_i))\pi(T_i|X_i) - \pi_c(T_i|X_i)]W_iY_i}{\frac{1}{n}\sum_i \mathbb{I}(T_i = t)W_i} + \max_W \frac{1}{n}\sum_{i=1}^n \phi(X_i)(Y_i + C(X_i)) \quad (4)$$
$$s.t. \;\; 1 + \Gamma^{-1}(\tilde{W}_i - 1) \le W_i \le 1 + \Gamma(\tilde{W}_i - 1),$$

The algorithm chooses the policy and router that minimize the robust regret bound $\bar{\pi}(\Pi, \Phi, \pi_c, \mathcal{W}_n^\Gamma), \bar{\phi}(\Pi, \Phi, \pi_c, \mathcal{W}_n^\Gamma) =$

$$\arg\min_{\pi\in\Pi, \phi\in\Phi} \hat{R}_n(\pi, \phi, \pi_c, \mathcal{W}_n^\Gamma). \quad (5)$$

In this case, the algorithm will choose to select the robust routing and decision policy considering that humans may use unobserved information for their decision making. The resulting system is confounding-robust in the sense that it considers the worst risk when humans' behavior cannot be point-identified because of the unobserved confounding.

Similarly, if we are interested in the policy improvement over the human's policy, we can optimize the future decision and routing policy by minimizing $\hat{R}_n^H(\pi, \phi, \mathcal{W}_n^\Gamma)$ as

$$\max_W \sum_{t=0}^{m-1} \frac{\frac{1}{n}\sum_i \mathbb{I}(T_i = t)[(1 - \phi(X_i))\pi(T_i|X_i)]W_iY_i}{\frac{1}{n}\sum_i \mathbb{I}(T_i = t)W_i} + \frac{1}{n}\sum_{i=1}^n (\phi(X_i) - 1)(Y_i + C(X_i)). \quad (6)$$

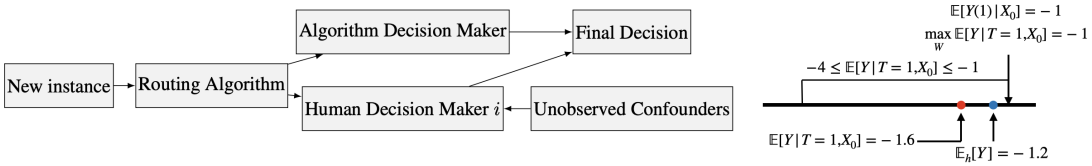
by removing the baseline policy π_c and contrasting the future system's performance with the performance of the human's decision policy with $\phi^H(X) \equiv 1$.

In practice, we would want the resulting human-AI system to outperform both the incumbent human policy and a candidate algorithmic policy. Then after our system is optimized, we can check whether Eq. (4) and Eq. (6) are both smaller than 0, which indicates the resulting human-AI system has a better performance compared to human working alone or the candidate algorithm working alone. We offer a theoretical improvement guarantee in Section 4.

3.3 An Illustrative Example

We use a toy example to illustrate how our method works. Assume a single context $X = X_0$ and we observe repeated observations for it. $P(U = 1|X_0) = P(U = 0|X_0) = 0.5$. With some abuse of notations, $Y(1) = -2, Y(0) = 0$ when $U = 1$, and $Y(1) = 0, Y(0) = -1$ when $U = 0$. Humans follow $P(T = 1|X_0, U = 1) = 0.5+\gamma$ and $P(T = 1|X_0, U = 0) = 0.5 - \gamma$ ($\gamma = 0$ means no unobserved confounding). $C(x) \equiv 0$. Here the two potential algorithms' performance are $\mathbb{E}[Y(1)|X_0] = -1$, $\mathbb{E}[Y(0)|X_0] = -0.5$.

With some simple algebra, the human performance is $\mathbb{E}_h[Y] = -0.75 - 1.5\gamma$. The nominal propensity score is



(a) Workflow between Human-AI Teams (b) Toy Example

Figure 1: Human-AI Collaboration with Unobserved Confounders

$P(T=1|X_0)=0.5$. If we want to evaluate the AI policies using observational data (no U), by the inverse propensity score weighting (or from Bayes theorem), $\mathbb{E}[Y|T=1,X_0] =$

$$\mathbb{E}\frac{\mathbb{I}(T=1,Y=-2)}{P(T=1|X_0)}(-2) = \frac{-2P(T=1,U=1|X_0)}{P(T=1|X_0)}$$

Plugging in the nominal propensity, $\mathbb{E}[Y|T=1,X_0] = -1-2\gamma$. Similarly, $\mathbb{E}[Y|T=0,X_0] = -0.5+\gamma$. When $\gamma=0.3$, $\mathbb{E}[Y|T=1,X_0] = -1.6 < -1.2 = \mathbb{E}_h[Y]$, so an algorithm assuming unconfoundness will incorrectly think the AI policy $T=1$ is the optimal policy, but human is actually better ($\mathbb{E}_h[Y] = -1.2 < -1 = \mathbb{E}[Y(1)|X_0]$).

Here, the MSM assumption corresponds to $\frac{1}{1+\Gamma} \le P(T=1|X_0,U) \le \frac{\Gamma}{1+\Gamma}$, so $\gamma=0.3$ means $\Gamma=4$. Plugging in the confidence interval, we have $-4 \le \mathbb{E}[Y|T=1,X_0] \le -\frac{0.5+0.3}{0.8} = -1$. Since our method adopts the pessimistic principle, the worst risk of the AI algorithm is $-1 > \mathbb{E}_h[Y]$, thus our algorithm can choose the best decision maker robustly in the presence of the uncertainty. This is illustrated in Fig. 1b.

3.4 Personalization

In the collaborative objective Eq. (2), we assume the experts have similar performance. However, this may not be the case in real-world scenarios. Experts often possess different areas of expertise, and may get different levels of confounding information. Therefore, implementing a personalized routing model may enhance the performance of the human-AI team.

Rather than indiscriminately assigning an expert to evaluate a given instance, the routing algorithm can make a decision to either delegate the instance to an algorithm or to a human, and, more importantly, determine the most suitable human decision-maker for the task at hand accounting for varied degrees of confounding for each human. We assume the odds ratio of each human decision-maker $H \in \{0,\cdots,K-1\}$'s propensity scores are associated with the confounding bound Γ_H. Similarly, the policy improvement with personalization has confounding-robust objective $R_n^P(\pi,\phi,\pi_c,\mathcal{W}_n^{\Gamma_H})$ is

$$\mathbb{E}\frac{\phi(H|X)}{d_0(H|X)}(Y+C(X)) + \sum_{t=0}^{m-1}\frac{\mathbb{E}\frac{\mathbb{I}(T=t)}{\pi_0(T|X,Y,H)}[\phi(a|X)\pi(T|X)-\pi_c(T|X)]Y}{\mathbb{E}\frac{\mathbb{I}(T=t)}{\pi_0(T|X,Y,H)}} \quad (7)$$

Let $\tilde{W}_i = \frac{1}{\tilde{\pi}_i(T_i|X_i,H_i)}$, $W_i = \frac{1}{\pi_i(T_i|X_i,Y_i,H_i)}$, then the worst-case estimator $\hat{R}_n^P(\pi,\phi,\pi_c,\mathcal{W}_n^{\Gamma_H})$ is

$$\max_W \sum_{t=0}^{m-1}\frac{\sum_i \mathbb{I}(T_i=t)W_i[\phi(a|X_i)\pi(T_i|X_i)-\pi_c(T_i|X_i)]Y_i}{\sum_i \mathbb{I}(T_i=t)W_i} + \frac{1}{n}\sum_{i=1}^{n}\frac{\phi(H_i|X_i)}{d_0(H_i|X_i)}(Y_i+C(X_i)) \quad (8)$$
$$s.t.\ \ 1+\Gamma_{H_i}^{-1}(\tilde{W}_i-1) \le W_i \le 1+\Gamma_{H_i}(\tilde{W}_i-1).$$

The policy and router can be similarly found by optimizing the following objective, $\bar{\pi}(\Pi,\Phi,\pi_c,\mathcal{W}_n^{\Gamma_H}), \bar{\phi}(\Pi,\Phi,\pi_c,\mathcal{W}_n^{\Gamma_H}) =$

$$\arg\min_{\pi\in\Pi,\phi\in\Phi} \hat{R}_n^P(\pi,\phi,\pi_c,\mathcal{W}_n^{\Gamma_H}). \quad (9)$$

Compared to Eq. (5), Eq. (9) further considered how to leverage individual human expertise to minimize the human-AI team's risk. When the historical and future human assignment is fully randomized and each human decision maker has the same Γ, Eq. (9) recovers Eq. (5).

3.5 Implementations

Optimizing the Objectives. To optimize the deferral collaboration system in Eq. (5) and Eq. (9), first, we need to solve the inner maximization in Eq. (4) and Eq. (8).

To simplify notations, we consider the following problem

$$\hat{Q}_t(r,\mathcal{W}) = \max_{W\in\mathcal{W}}\frac{\sum_{i=1}^n r_i W(T_i,X_i,Y_i)}{\sum_{i=1}^n W(T_i,X_i,Y_i)} \quad s.t.\ a_i^{\Gamma_i} \le W(T_i,X_i,Y_i) \le b_i^{\Gamma_i} \quad (10)$$

When $r_i = \mathbb{I}(T_i=t)[(1-\phi(X_i))\pi(T_i|X_i) - \pi_c(T_i|X_i)]Y_i$, $a_i^{\Gamma_i} = 1+\Gamma^{-1}(\tilde{W}_i-1)$, $b_i^{\Gamma_i} = 1+\Gamma(\tilde{W}_i-1)$, solving Eq. (10) is equivalent to optimizing W for the empirical $\hat{R}_n(\pi,\phi,\pi_c,\mathcal{W}_n^{\Gamma})$ in Eq. (4), and when $r_i = \mathbb{I}(T_i=t)[\phi(a|X_i)\pi(T_i|X_i) - \pi_c(T_i|X_i)]Y_i$, $a_i^{\Gamma_i} = 1+{\Gamma_{H_i}}^{-1}(\tilde{W}_i-1)$, $b_i^{\Gamma_i} = 1+\Gamma_{H_i}(\tilde{W}_i-1)$, solving Eq. (10) is equivalent to optimizing W for $\hat{R}_n^P(\pi,\phi,\pi_c,\mathcal{W}_n^{\Gamma_H})$ in Eq. (8).

The optimization problem in Eq. (10) is known as a *linear fractional program* (Chadha and Chadha 2007). Taking the derivative of the objective in Eq. (10) w.r.t. $W_i = W(T_i,X_i,Y_i)$, the objective is monotonically increasing (decreasing) with W_i if $r_i\sum_{j\neq i}W_j - \sum_{j\neq i} r_j W_j$ is greater (less) than zero. Hence the optima is achieved when all the W_i are taking the value at the boundary. Furthermore, the objective can be viewed as a weighted combination of r_i with the weights adding up to one. So the objective is maximized when the weights $W_i/\sum_i W_i$ are high for the large r_i and

Algorithm 1: Confounding-Robust Deferral Collaboration (ConfHAI/ConfHAIPerson)

Input: number of iterations N, π, ϕ, Γ (Γ_H), $\{X_i, T_i, Y_i\}_{i=1}^N$, π_c
Output: π_θ, ϕ_ρ
for $i \leftarrow 1$ to N **do**
$W \leftarrow \arg\max_{W \in \mathcal{W}}$ Eq. (4) (Eq. (8) for ConfHAIPerson).
$\theta, \rho \leftarrow \nabla \hat{R}_n(\pi, \phi, \pi_c, \mathcal{W}_n^\Gamma)$ ($\nabla \hat{R}_n^P(\pi, \phi, \pi_c, \mathcal{W}_n^{\Gamma_H})$ for ConfHAIPerson).
end for

are low for the small r_i. Based on these insights, the optimal weights $\{W_i\}$ of the linear fractional program can be characterized by the following theorem.

Theorem 1. *Let (i) be the ordering such that $r_{(1)} \leq r_{(2)} \leq \cdots \leq r_{(n)}$. $\hat{Q}_t(r, \mathcal{W}) = \lambda(k^*)$, where $k^* = \inf\{k = 1, \cdots, n+1 : \lambda(k) < \lambda(k-1)\}$ and*

$$\lambda(k) = \frac{\sum_{i<k} a_{(i)}^\Gamma r_{(i)} + \sum_{i \geq k} b_{(i)}^\Gamma r_{(i)}}{\sum_{i<k} a_{(i)}^\Gamma + \sum_{i \geq k} b_{(i)}^\Gamma} \tag{11}$$

See Appendix for the proof. Theorem 1 provides an efficient way to solve Eq. (10) by line search: first sort r_i in ascending order and initialize all $W_i = a_i^\Gamma$, then change W_k to b_k^Γ for $k = n, n-1, \cdots, 1$ until the first time when $\lambda(k)$ decreases. After solving the inner maximization problem, we can optimize the minimization problem in Eq. (5) and Eq. (9). In this paper, we consider differentiable policies $\Pi = \{\pi_\theta : \theta \in \Theta\}$ and router class $\Phi = \{\phi_\rho, \rho \in \mathrm{P}\}$, such as logistic policies with $\pi_{\{\alpha,\beta\}}(x) = \sigma(\alpha + \beta^T x)$ or neural networks, so the following optimization problem can be efficiently solved by gradient descent. For every iteration, our algorithm starts by finding the weights W given the current model parameters through line search, then uses gradient descent to update policy and router jointly. We call our main algorithm assuming all decision makers can only be queried randomly as ConfHAI and its variant considering the diverse expertise of individual human decision makers as ConfHAIPerson.

In practice, the value of Γ can also be estimated in a data-driven way, which we discuss in detail in Appendix.

4 Theoretical Analysis

Improvement Guarantees. We first show the worst-case empirical regret is an asymptotic upper bound for the population regret. We assume the outcome and true propensity score is bounded for analysis, *i.e.*, $|Y| \leq B, \pi_0(t|x,y) \geq v, \forall t \in \{0, \cdots, m-1\}, x \in \mathcal{X}, y \in \mathcal{Y}$. The following theorem guarantees the improvement over the population regret by solving the minimax optimization for the empirical regret.

Theorem 2. *Suppose the true inverse propensities $1/\pi_0(T_i|X_i, Y_i) \in \mathcal{W}_n^\Gamma, i = 1, \cdots, n$, $|Y| \leq B, C(X) \leq \bar{c}, \pi_0(t|x,y) \geq v, \forall t \in \{0, \cdots, m-1\}, x \in \mathcal{X}, y \in \mathcal{Y}$ and denote policy and router's class Π and Φ's Rademacher Complexity as $\mathfrak{R}_n(\Pi)$ and $\mathfrak{R}_n(\Phi)$, then for $\delta > 0$, with probability $1-\delta$, we have*

$$R(\pi, \phi, \pi_c) \leq \hat{R}_n(\pi, \phi, \pi_c, \mathcal{W}_n^\Gamma) + 2(B + \bar{c})\mathfrak{R}_n(\Phi).$$
$$+ 2\frac{B}{\nu}\mathfrak{R}_n(\Pi) + (3B + 3\bar{c} + \frac{5B+1}{\nu^2})\sqrt{\frac{2\log\frac{8m}{\delta}}{n}}$$

The proof is included in Appendix and can be extended for the improvement guarantee over the personalized version of the algorithm. Note that the global optima of the empirical objective is never positive when $\pi_c \in \Pi$ since we can take $\pi = \pi_c$ and $\phi(X) = 0$. If we only consider Π, Φ with vanishing Rademacher Complexity (*i.e.*, $O(n^{-1/2})$), then Theorem 2 implies that given enough samples, if the empirical objective is negative, we can get an improvement over π_c under well-specification. We also provide the improvement guarantee over the incumbent human policy in Corollary 2.1.

Corollary 2.1. *Under the condition of Theorem 2 and suppose $\mathfrak{R}_n(\Pi) = O(\frac{1}{\sqrt{n}})$, $\mathfrak{R}_n(\Phi) = O(\frac{1}{\sqrt{n}})$, then for $\delta > 0$, with probability $1-\delta$, we have*

$$R_H(\pi, \phi, \pi_c) \leq \hat{R}_n^H(\pi, \phi, \pi_c, \mathcal{W}_n^\Gamma) + O(\sqrt{\frac{\log\frac{m}{\delta}}{n}}). \tag{12}$$

What Kind of Instances are routed to Humans. One interesting question under the proposed deferral collaboration framework is what kind of instances should be solved by humans and what should be solved by algorithms. Here we provide a theoretical analysis of the routing decision under the optimal AI policy. Assume we have access to the true human behavior policy and for an AI policy $\pi(T|X)$ given only X, we can compare the expected risk of routing the instance to human and the expected risk of routing the instance to the AI and choose the one with lower expected risk. This produces a closed-form solution for the routing decision.

Theorem 3 (Instances routed to the humans)**.** *Assume we have access to the true weight W^* and an AI policy $\pi(T|X)$ given only X, to minimize Eq.* (3)*, the routing system should send the decision task to humans when*

$$\mathbb{E}_{U \sim P(U|X), T \sim \pi_0(T|X,U)}[Y + C(X)|X] < \mathbb{E}_{T \sim \pi(T|X)}[Y|X]$$

The proof is in Appendix. The left term is the expected risk of routing the instance to human when humans have access to the unobserved confounder U and the right term is the expected risk of routing the instance to the AI when it only has access to X. The theorem has an interesting implication that the routing system should always send the instance to human when humans can utilize U to improve their decision making performance and surpass the best possible decision performance when only X is available. Compared to Gao et al. (2021b), where the main source of complementarity comes from model misspecification, here the main source of complementarity comes from the unobserved confounder and humans may be irreplaceable even if we have access to the optimal AI policy with confounded data.

5 Experiments

We report empirical findings to examine the advantages of Human-AI complementary and being robust to unobserved

confounding. Our first experiment demonstrates the benefit of human-AI collaboration within a controlled environment. Our subsequent experiments consider two real-world examples in financial lending and healthcare industry. Code and appendix are available at https://github.com/ruijiang81/ConfoundL2D.

In our experiment, we examine the following decision-making configurations. For all baselines without using personalization, human experts are selected at random. Human Only (Human) solely queries human decision-makers randomly to output final decisions. Algorithm Only (AO) uses the inverse propensity score weighting method (Swaminathan and Joachims 2015a) to train a policy. Confounding-Robust Algorithm Only (ConfAO) trains a confounding-robust policy with no human involved to determine the final decisions (Kallus and Zhou 2018). Human-AI team (HAI) uses the deferral collaboration method assuming unconfoundedness (Gao et al. 2021b). Our method and its personalized variant are denoted as *ConfHAI* and *ConfHAIPerson* respectively. See Appendix for a more detailed discussion about the baselines. We use the logistic policies for the policy and router model classes. The baseline policy is set as the never-treat policy $\pi_c(0|x) = 1$ (Kallus and Zhou 2018).

5.1 Synthetic Experiment

We demonstrate the benefit of confounding-robust human-AI collaboration using the following data-generating process,

$$\xi \sim \text{Bern}(0.5), X \sim \mathcal{N}((2\xi - 1)\mu_x, I_5), U = \mathbb{I}[Y_i(1) < Y_i(-1)]$$
$$Y(t) = \beta_0^T x + \mathbb{I}[t = 1]\beta_{\text{treat}}^T x + 0.5\alpha\xi\mathbb{I}[t = 1] + \eta + w\xi + \epsilon$$

where $\beta_{\text{treat}} = [1.5, 1, 1.5, 1, 0.5]$, $\mu_x = [1, .5, 1, 0, 1]$, $\eta = 2.5, \alpha = -2, w = 1.5$ and $\epsilon \sim \mathcal{N}(0, 1)$ (Kallus and Zhou 2021). The nominal propensity $\pi_0(T = 1|X) = \sigma(\beta^T X)$, $\beta = [0, .75, .5, 0, 1, 0]$, T_i is generated by the true propensities by $\pi_0(T = 1|X, U) = \frac{(\Gamma U + 1 - U)\pi_0(T=1|X)}{[1+2(\Gamma-1)\pi_0(T=1|X)-\Gamma]U+\Gamma+(1-\Gamma)\pi_0(T=1|X)}$, where Γ is the specified level of confounding. In this setting, the human decision-maker acquires unobserved information to improve decisions. We set $\log(\Gamma) = 2.5$, $C(x) = 0$ and vary the log-confounding parameter in $\{0.01, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4\}$. To also test the personalized variant, we simulate three human decision makers with the same Γ.

The results are shown in Fig. 2a. When Γ is small (weak unmeasured confounding), the baselines not considering unobserved confounding have similar performance with our methods and ConfAO. When Γ is approaching the underlying confounding factor, we observe a significant policy improvement over the baseline policy (regret is smaller than 0). The personalized method has performance similar to ConfHAI since all human decision makers have the same performance here. In this example, interestingly, the ConfAO policy is actually worse than humans' performance, and almost never exceeds humans' performance with varying Γ, while human-AI complementarity can outperform humans' performance for a range of Γ close to its true value, which emphasizes the benefit of our confounding-robust deferral system.

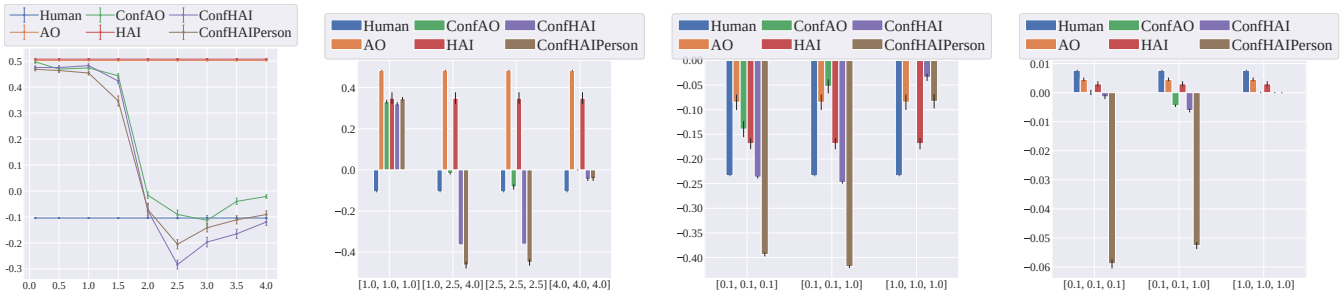
Next, we simulate three human workers with $\log(\Gamma) = 1, 2.5, 4$, respectively, which corresponds to the setting where different humans may acquire different unobserved information to aid their decision making and some experts may perform better than their peers. We examine four $\log(\Gamma)$ specifications with heterogeneous workers: $[1, 1, 1], [2.5, 2.5, 2.5], [1, 2.5, 4], [4, 4, 4]$ and show the results in Fig. 2b. With small Γ, we observe all methods perform suboptimally with no policy improvement. The personalized variant has an additional improvement over ConfHAI by leveraging the diverse expertise of human decision makers. With correctly specified and relatively large Γ, we observe that ConfHAI and ConfHAIPerson significantly outperform other baselines and demonstrate human-AI complementarity, where they outperform both human-only and algorithm-only teams.

5.2 Real-World Examples

We provide two real-world examples in this section with human cost set as $C(x) = 0.1$. See Appendix for more details about the datasets and training in the experiments.

Financial Lending. In financial lending, loan officers can obtain additional information for decision-making by visiting the loan applicants. However, such information may not be recorded in the historical data. We use the Home Equity Line of Credit(HELOC) dataset which contains anonymized information about credit applications by real homeowners. Some of the used features are the average months since the account opened, maximum delinquency, and number of inquiries in the last 6 months. We assume there are three human decision makers with $\log(\Gamma) = [0.1, 0.1, 1]$, which means two of them rarely seek external information to improve their decision making and another decision maker is more likely to get external risk estimation when evaluating applications. We train a logistic regression on 10% of the data to simulate nominal policies, which can be a guideline policy of the insurance company, and the actual treatments taken are generated using the same procedure in Section 5.1 and the fitted nominal propensity is estimated using logistic regression on actual treatments. The outcome of the dataset is a binary outcome indicating whether the applicant was 90 days past due. We build a risk function where the loan company will receive a risk $Y \sim \mathcal{N}(0, 1)$ if not approving the loan, $Y \sim \mathcal{N}(-2, 1)$ if approving for an applicant with good credit and $Y \sim \mathcal{N}(2, 1)$ if approving for an applicant with bad credit.

Acute Stroke Treatment. In this healthcare example, the doctors need to treat patients with acute stroke. Experienced doctors may observe bedside information, and past patient behaviors to aid their decision-making, which are not recorded in the historical records. We use the data from the International Stroke Trial (Group 1997) and focus on two treatment arms: the treatment of both aspirin and heparin (medium and high doses), and the treatment of aspirin only. Since the trial only has the outcome under action taken, we create potential outcomes by fitting a separate random forest model for each treatment as in (Biggs, Gao, and Sun 2021; Elmachtoub, Gupta, and Zhao 2023). The outcome is a composite score including variables like death, recurrent stroke, pulmonary embolism, and recovery. Some of the features used by the algorithm include age, sex, deficient symptoms, stroke types, and cerebellar signs. Similarly, we assume there are three human physicians with $\log(\Gamma) = [0.1, 0.1, 1]$ prescribing treatments.



(a) Syn: homogeneous humans. (b) Syn: heterogeous humans. (c) HELOC. (d) IST.

Figure 2: Policy Regret. ConfHAI and ConfHAIPerson offer consistent and significantly better policy improvement for a range of Γ compared to baseline algorithms over synthetic and real datasets. X-axis: Specified $\log(\Gamma)$. Y-axis: Policy Regret.

The results are shown in Fig. 2c and Fig. 2d respectively. For each experiment, we try three $\log(\Gamma)$ specifications: $[0.1, 0.1, 0.1]$, $[0.1, 0.1, 1]$ and $[1, 1, 1]$, which correspond to under, correct and over specifications. In HELOC, the baselines not considering unobserved confounding can still achieve policy improvement, while it is not consistent across settings, *e.g.*, in IST. We observe that ConfHAI and ConfHAIPerson achieve the best performance with correctly-specified Γ and the personalized variant achieves significantly better performance than other methods. With over-specification, the performance of confounding-robust methods decreases but can reliably provide policy improvement. Similarly, ConfAO can provide policy improvement with different specifications of Γ, however, its performance is often much worse than the human-AI methods we propose.

5.3 Real Human Responses

In addition, we use the real human responses to validate our approach. We use the scientific annotation dataset FOCUS (Rzhetsky, Shatkay, and Wilbur 2009) with responses from five human annotators. The features are sentences extracted from a scientific corpus, and each labeler was asked to label the sentence as scientific or not. We transform the text into feature representations using Glove embeddings (Pennington, Socher, and Manning 2014). We assume if the human annotator considers the sentence scientific, they will apply action I (e.g., retweet the paper), if the sentence is indeed scientific, the risk is $\mathcal{N}(-1, 1)$, otherwise the risk is from $\mathcal{N}(1, 1)$. On the other hand, if the human annotator considers the sentence as non-scientific, they will apply action II (e.g., ignore the paper), if the sentence is indeed non-scientific, the risk is $\mathcal{N}(-1, 1)$ (otherwise the risk is $\mathcal{N}(1, 1)$). Since each sentence is annotated, we know whether the human annotator considers the sentence scientific and is able to derive the simulated human behavior policy. The confounding is created by removing samples with 20% top outcomes in the treated group and 20% bottom outcomes in the control group. We specify the same Γ for each human and vary it.

The results are shown in Fig. 3. This dataset is different from our simulations since humans' true propensities may not reflect the worst case indicated in the MSM optimization. However, we still observe our methods consistently offer the best performance with a wide range of Γ.

5.4 Ablation Studies

We examine the effect of human cost on the risk of each method in Fig. 3. We use the synthetic data setup and vary the human cost from 0 to 0.3. As the cost becomes higher, the Human baseline's performance gets worse. Human-AI systems' performance (HAI, ConfHAI, ConfHAIPerson) is also impacted when human cost is higher, but the proposed methods consistently outperform other baselines.

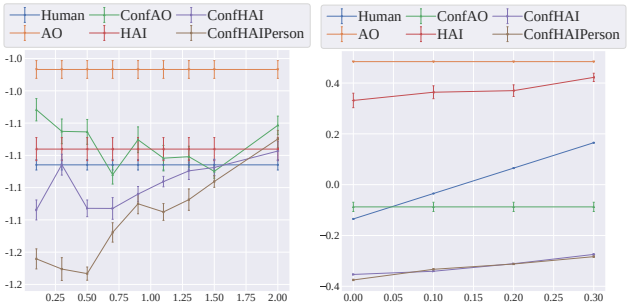


Figure 3: Left: Policy Regret vs. Specified $\log(\Gamma)$ (Real Data); Right: Policy Regret vs. Human Cost (Ablation Studies).

6 Conclusion and Future Work

In this paper, we study a new problem of unmeasured confounding in Human-AI collaboration. We propose ConfHAI as a novel confounding-robust deferral policy learning method to address this problem. ConfHAI optimizes policy decisions by selectively deferring decision instances to either humans or the AI, based on the context and capabilities of both and the strength of the unmeasured confounding. We demonstrate the policy improvements of ConfHAI in theory and through a variety of synthetic and real data simulations. Nevertheless, a potential limitation of the proposed method is the relatively strict constraints on the marginal sensitivity model. Future work can improve the sharpness of the MSM (Dorn and Guo 2023) or consider adopting more interpretable sensitivity models (Imbens 2003). Broadly speaking, our findings indicate the importance of explicitly accounting for the information discrepancy between human decision-makers and AI algorithms to improve human-AI complementarity.

References

Athey, S.; and Wager, S. 2017. Efficient policy learning. Technical report.

Athey, S.; and Wager, S. 2021. Policy learning with observational data. *Econometrica*, 89(1): 133–161.

Bansal, G.; Nushi, B.; Kamar, E.; Weld, D.; Lasecki, W.; and Horvitz, E. 2019. A case for backward compatibility for human-ai teams. *arXiv preprint arXiv:1906.01148*.

Biggs, M.; Gao, R.; and Sun, W. 2021. Loss functions for discrete contextual pricing with observational data. *arXiv preprint arXiv:2111.09933*.

Chadha, S.; and Chadha, V. 2007. Linear fractional programming and duality. *Central European Journal of Operations Research*, 15: 119–125.

Cornfield, J.; Haenszel, W.; Hammond, E. C.; Lilienfeld, A. M.; Shimkin, M. B.; and Wynder, E. L. 1959. Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer institute*, 22(1): 173–203.

De, A.; Koley, P.; Ganguly, N.; and Gomez-Rodriguez, M. 2020. Regression under Human Assistance. In *AAAI*, 2611–2620.

Dorn, J.; and Guo, K. 2023. Sharp sensitivity analysis for inverse propensity weighting via quantile balancing. *Journal of the American Statistical Association*, 118(544): 2645–2657.

Dudík, M.; Erhan, D.; Langford, J.; and Li, L. 2014. Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4): 485–511.

Elmachtoub, A.; Gupta, V.; and Zhao, Y. 2023. Balanced Off-Policy Evaluation for Personalized Pricing. In *International Conference on Artificial Intelligence and Statistics*, 10901–10917. PMLR.

Gao, R.; Biggs, M.; Sun, W.; and Han, L. 2021a. Enhancing Counterfactual Classification via Self-Training. *arXiv preprint arXiv:2112.04461*.

Gao, R.; Saar-Tsechansky, M.; De-Arteaga, M.; Han, L.; Lee, M. K.; and Lease, M. 2021b. Human-ai collaboration with bandit feedback. *arXiv preprint arXiv:2105.10614*.

Gao, R.; Saar-Tsechansky, M.; De-Arteaga, M.; Han, L.; Sun, W.; Lee, M. K.; and Lease, M. 2023. Learning complementary policies for human-ai teams. *arXiv preprint arXiv:2302.02944*.

Group, I. S. T. C. 1997. The International Stroke Trial (IST): a randomised trial of aspirin, subcutaneous heparin, both, or neither among 19 435 patients with acute ischaemic stroke. *The Lancet*, 349(9065): 1569–1581.

Guo, W.; Yin, M.; Wang, Y.; and Jordan, M. 2022. Partial identification with noisy covariates: A robust optimization approach. In *Conference on Causal Learning and Reasoning*, 318–335. PMLR.

Holstein, K.; De-Arteaga, M.; Tumati, L.; and Cheng, Y. 2023. Toward supporting perceptual complementarity in human-AI collaboration via reflection on unobservables. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1): 1–20.

Ibrahim, R.; Kim, S.-H.; and Tong, J. 2021. Eliciting human judgment for prediction algorithms. *Management Science*, 67(4): 2314–2325.

Imbens, G. W. 2003. Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review*, 93(2): 126–132.

Imbens, G. W. 2024. Causal inference in the social sciences. *Annual Review of Statistics and Its Application*, 11.

Jin, Y.; Ren, Z.; and Candès, E. J. 2021. Sensitivity Analysis of Individual Treatment Effects: A Robust Conformal Inference Approach.

Joachims, T.; Swaminathan, A.; and de Rijke, M. 2018. Deep learning with logged bandit feedback. In *ICLR*.

Kallus, N. 2018. Balanced policy evaluation and learning. *Advances in neural information processing systems*, 31.

Kallus, N. 2019. Classifying treatment responders under causal effect monotonicity. In *International Conference on Machine Learning*, 3201–3210. PMLR.

Kallus, N. 2021. More efficient policy learning via optimal retargeting. *Journal of the American Statistical Association*, 116(534): 646–658.

Kallus, N.; and Zhou, A. 2018. Confounding-robust policy improvement. In *NeurIPS*, 9269–9279.

Kallus, N.; and Zhou, A. 2021. Minimax-optimal policy learning under unobserved confounding. *Management Science*, 67(5): 2870–2890.

Madras, D.; Pitassi, T.; and Zemel, R. 2018. Predict responsibly: improving fairness and accuracy by learning to defer. *NeurIPS*, 31: 6147–6157.

Namkoong, H.; Ma, Y.; and Glynn, P. W. 2022. Minimax Optimal Estimation of Stability Under Distribution Shift.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Raghu, M.; Blumer, K.; Corrado, G.; Kleinberg, J.; Obermeyer, Z.; and Mullainathan, S. 2019. The algorithmic automation problem: Prediction, triage, and human effort. *arXiv:1903.12220*.

Rosenbaumn, P. R. 2002. *Observational Studies (2nd ed.)*. Springer, New York.

Rubin, D. B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5): 688.

Rubin, D. B. 1980. Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American statistical association*, 75(371): 591–593.

Rzhetsky, A.; Shatkay, H.; and Wilbur, W. J. 2009. How to get the most out of your curation effort. *PLoS computational biology*, 5(5): e1000391.

Sondhi, A.; Arbour, D.; and Dimmery, D. 2020. Balanced off-policy evaluation in general action spaces. In *International Conference on Artificial Intelligence and Statistics*, 2413–2423. PMLR.

Swaminathan, A.; and Joachims, T. 2015a. Counterfactual risk minimization: Learning from logged bandit feedback. In *ICML*, 814–823.

Swaminathan, A.; and Joachims, T. 2015b. The self-normalized estimator for counterfactual learning. *advances in neural information processing systems*, 28.

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

Wang, T.; and Saar-Tsechansky, M. 2020. Augmented Fairness: An Interpretable Model Augmenting Decision-Makers' Fairness. *arXiv:2011.08398*.

Wilder, B.; Horvitz, E.; and Kamar, E. 2020. Learning to Complement Humans. *arXiv*.

Wolczynski, N.; Saar-Tsechansky, M.; and Wang, T. 2022. Learning to Advise Humans By Leveraging Algorithm Discretion. *arXiv:2210.12849*.

Yin, M.; Shi, C.; Wang, Y.; and Blei, D. M. 2021. Conformal sensitivity analysis for individual treatment effects. *arXiv preprint arXiv:2112.03493*.

Zhao, Q.; Small, D.; and Bhattacharya, B. 2019. Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(4): 735–761.