

Learning Complementary Policies for Human-AI Teams

Ruijiang Gao ^{*} Maytal Saar-Tsechansky^{*} Maria De-Arteaga ^{*} Ligong Han [†]

Wei Sun [‡] Min Kyung Lee ^{*} Matthew Lease ^{*}

February 7, 2023

Abstract

Human-AI complementarity is important when neither the algorithm nor the human yields dominant performance across all instances in a given context. Recent work that explored human-AI collaboration has considered decisions that correspond to classification tasks. However, in many important contexts where humans can benefit from AI complementarity, humans undertake course of action. In this paper, we propose a framework for a novel human-AI collaboration for selecting advantageous course of action, which we refer to as Learning Complementary Policy for Human-AI teams (LCP-HAI). Our solution aims to exploit the human-AI complementarity to maximize decision rewards by learning both an algorithmic policy that aims to complement humans by a routing model that defers decisions to either a human or the AI to leverage the resulting complementarity. We then extend our approach to leverage opportunities and mitigate risks that arise in important contexts in practice: 1) when a team is composed of multiple humans with differential and potentially complementary abilities, 2) when the observational data includes consistent deterministic actions, and 3) when the covariate distribution of future decisions differ from that in the historical data. We demonstrate the effectiveness of our proposed methods using data on real human responses and semi-synthetic, and find that our methods offer reliable and advantageous performance across setting, and that it is superior to when either the algorithm or the AI make decisions on their own. We also find that the extensions we propose effectively improve the robustness of the human-AI collaboration performance in the presence of different challenging settings.

1 Introduction

Supervised learning and algorithmic decision-making have shown promising results when solving some tasks traditionally performed by humans (Chen et al. 2018, Swaminathan and Joachims 2015a, He et al. 2015). Yet, most AI research focuses on improving the algorithm performance rather than optimize the human-AI *team's* performance (Bansal et al. 2020). Human-AI complementarity is particularly promising when neither the state-of-the-art algorithm nor the human yields dominant performance across all instances in a given domain. While humans are known to be imperfect decision-makers, AI models are often of limited capacity, meaning that improved performance for some subset of instances can come at the cost of sacrificing performance in others (Menon and Williamson 2018). Together, these properties of humans and AI algorithms present opportunities to leverage the complementary abilities of humans and AI to yield better performance by selectively allocating instances to either a human or an AI, depending on which entity is most likely to provide a correct assessment. Such a design is motivated by contexts in which actions can be undertaken autonomously by either a human or an AI, such as moderation of social media comments (Lai et al. 2022), pricing offers for individual buyers (Elmachtoub et al. 2021), or microloan application assessments (Achab et al. 2018).

However, there are challenges to leverage potential complementarity. In particular, it is necessary to develop effective methods for human-AI collaboration that can: 1) reliably *identify* instances for which the human and AI

^{*}University of Texas at Austin

[†]Rutgers University

[‡]IBM Research

can complement one another to yield better rewards, and then 2) effectively *exploit* these human-AI complementarity opportunities to maximize the benefits (Wilder et al. 2020, Madras et al. 2018). Recent work on human-AI collaboration (Mozannar and Sontag 2020, Keswani et al. 2021, Madras et al. 2018, Wilder et al. 2020) has proposed to learn and exploit human-AI complementarity by optimally assigning instances to either a human or an AI. Learning AI models for such human-AI collaboration has been referred to in prior work as *learning to defer* (Madras et al. 2018) or *learning to complement humans* (Wilder et al. 2020). We will henceforth refer to it as *deferral collaboration*.

Research on deferral collaboration has thus far considered problems that correspond to traditional classification settings. Specifically, prior work considered predictions that correspond to inferring the correct ground truth label for a given instance (*e.g.*, whether a post contains hateful language), and assumed that labels are available for all instances in the training set. However, in many important contexts in which it is desirable to develop algorithmic decisions that complement a human decision-maker, the nature of the task and the corresponding historical data are meaningfully different. In particular, often in practice decisions do not correspond to classification labels, but to the selection of a course of action that yields the best reward in which the realized reward depends on the action that is taken¹. In this setting, the corresponding AI task to complement the human is to learn a *personalized decision policy* from historical data that infers choices of actions for future instances to yield the best expected reward. In such scenarios, the observational data only contains instances corresponding to actual past choices made by decision-makers, along with their corresponding, observed, rewards. Such data is often rich, abundant and preferred by practitioners because data from randomized controlled trials could be costly and scarce (Kallus and Zhou 2018). For example, decisions regarding microloan applications may pertain to a specific interest rate offered to each customer, and historical data may only include the profit / loss under the different assigned interest rates (Ban and Keskin 2021). Similarly, historical data on pricing offers might include past offers made, along with outcome customer buying decisions in response to these offers (Bertsimas and Kallus 2016, Gao et al. 2021). Importantly, the historical data would not include pricing choices and corresponding rewards for price offers that were not made, and it is unclear how to adapt deferral collaboration methods for classification to recommend actions here.

In this paper, we consider the problem of developing a human-AI collaboration method to improve the performance of policy learning by leveraging human-AI complementarity. While prior work has proposed a variety of methods to address the problem of *policy learning from observational data* (Dudík et al. 2014, Kallus 2018, Athey and Wager 2017, Shalit et al. 2017), to the best of our knowledge, we are the first to consider the problem of learning a policy to *complement* human. In particular, we consider decision tasks involving the selection of actions to maximize reward (as measured by a predefined reward function) in which the observational data reflect the human decision-makers’ past choices and their corresponding rewards. For this setting, our deferral collaboration method allows for future actions to be undertaken autonomously by either a human or an AI. Our goal is to discover and then exploit opportunities for human-AI complementarity in order to achieve higher rewards than would be possible for either a purely human-based or purely algorithmic-based approach. We refer to this problem as *Learning Complementary Policies for Human-AI teams* (LCP-HAI). To our knowledge, this is the first work to formulate and tackle this problem.

Figure 1 outlines the deferral collaboration framework we consider in which each instance can be routed to either the algorithmic policy or the human decision maker to recommend the best action. Learning complementary policies to optimally leverage human-AI complementarity requires several steps. First, the AI must learn a policy that specializes in instances for which the algorithm can achieve superior decisions compared to those that can be made by the human. Second, we need an algorithm to route decisions to an agent (human or algorithmic) expected to yield the best reward. In this work, we develop methods for both tasks by proposing a training algorithm that uses observational data to simultaneously learn both the human-complementing policy and the router.

The practical contexts in which LCP-HAI can be applied to improve decision performance also give rise to additional challenges. First, we discuss contexts in which the human-AI team include multiple human decision makers. Rather than consider human decision makers as a sole entity, we propose an algorithm that considers each individual’s strengths. Particularly, when different human decision makers exhibit varying decision-making skill over different subsets of instances, it is beneficial to route different instances to different people in order to optimize performance. We discuss these settings and how to adapt our approach to these decision-making

¹We use the terms “decision” and “action” interchangeably in this paper.

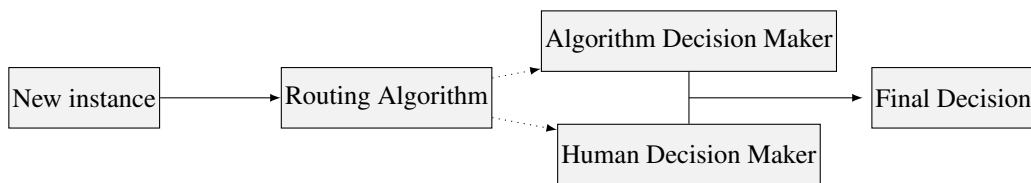


Figure 1: Decision-Making Process for Human-AI Collaboration

contexts.

In addition, we propose two variants of our proposed LCP-HAI in order to improve reliability of deferral collaboration. First, consider a context in which human decision-makers may consistently choose the same action for some instances, indicating deterministic behavior. In general, if historical human decision makers prescribe deterministic actions to a subset of instances, this hinders learning advantageous courses of action from the data (Langford et al. 2008, Swaminathan and Joachims 2015a, Kallus 2021) because the reward under a different action is never observed. Interestingly, however, prior work has discussed contexts in which human consistency reflects valuable domain expertise and thereby encodes advantageous choices (De-Arteaga et al. 2021). In this work, we present a novel way to leverage consistent human decisions in such contexts that leads to more robust and reliable human-AI teams (Langford et al. 2008).

Second, we consider settings of distribution shift in which the underlying data distribution used to train a policy is not the same as the distribution during deployment. For instance, this could arise if consumer behavior shifts due to changes in the environment after the model is deployed. In this case, it is difficult to accurately estimate the deferral collaboration system’s expected performance to reliably route decisions to the most suitable decision-maker. In particular, not only do AI algorithms often yield poor performance for out-of-distribution data, but this failure can be difficult for the routing algorithm to assess given that such instances are poorly represented in the historical data (He et al. 2015, Geirhos et al. 2018). In contrast, human reasoning capabilities often enable them to effectively adapt to new instances. For example, humans have been shown to outperform algorithms when pricing individualized quotes that involve unique characteristics (Karlinsky-Shichor and Netzer 2019). It is thus desirable that such instances will be routed to a human (Huber et al. 2021). To enhance the human-AI team’s reliability, we propose solutions to detect and route out-of-distribution instances to human decision makers.

Overall, our work makes the following contributions:

- We formulate the problem of *Learning Complementary Policies for Human-AI teams* (LCP-HAI) in which: 1) decisions can be made autonomously by either an algorithmic policy or by human decision makers, 2) decisions correspond to selecting advantageous course of actions, and 3) the goal is to leverage human-AI complementarity to maximize total reward. The problem is to learn such a system using observational data reflecting human historical decisions and their corresponding rewards.
- We propose an approach to solve the LCP-HAI problem. This includes a routing algorithm that assigns each instance to either a human decision maker or the algorithmic policy, and a new method that produces a decision-making policy based on the observational data and that can effectively complement the human. The router and the decision-making policy are jointly trained to best exploit human decision-making abilities, while complementing the human by optimizing the algorithmic policy to perform well on instances that are more challenging for human. We further develop variants of our approach to yield robust human-AI team performance in the presence of multiple decision makers, deterministic decisions, and distribution shift between the training and deployment covariate distributions.
- We empirically demonstrate the performance of the proposed solutions for the LCP-HAI problem relative to alternatives, using data with both synthetic and real human decisions. Our results demonstrate that our approaches yield significant improvement over baselines by effectively learning and exploiting human-AI decision-making complementarity. Furthermore, our approaches and the principles they are based on offer promising foundations on which future work on human-AI collaboration can build upon.

- We discuss managerial implications of our research and the potential of the proposed hybrid system to inform future algorithmic solutions to practical business problems.

2 Related Work

Methods for Human-AI Collaboration. Recent research on methods for human-AI collaboration have considered how to allow human-AI teams to achieve better classification performance, such as accuracy and fairness, by leveraging their respective complementarity. Bansal et al. (2020) study the problem of human-AI interaction and demonstrate that the best classifier is not always the one that leads to the best human decisions when the classifier is used as decision support. Recent work has studied how to develop algorithms in contexts in which the human is the sole and final decision-maker (Bansal et al. 2019, Ibrahim et al. 2021, Wolczynski et al. 2022). We consider a setting that does not involve human-AI interaction, and in which instead decisions are made by either a human or an algorithm. Prior work has considered the challenge of routing instances to an algorithm or a human (Madras et al. 2018, Wilder et al. 2020, Raghu et al. 2019, De et al. 2020, Wang and Saar-Tsechansky 2020). Madras et al. (2018), Wilder et al. (2020), Raghu et al. (2019) consider the problem of optimizing overall classification performance, while De et al. (2020) study human-AI collaboration for a regression task, and Wang and Saar-Tsechansky (2020) consider jointly augmenting human accuracy and fairness. All of these works rely on the estimated prediction uncertainty of the algorithm and the human, but the problem of quantifying this uncertainty when learning from bandit observation data remains an open problem. The core difference between these works and ours is that they consider contexts in which the AI’s learning task is a traditional supervised classification task. In this work, we focus on the problem of learning a decision *policy*, which assigns optimal actions for decision subjects to optimally complement a human’s decision policy, based on data with bandit feedback.

Policy Learning from Observational Data. The problem of inferring an optimal personalized policy from offline observational data has been studied extensively in many domains ranging from e-commerce, contextual pricing, and personalized medicine (Dudík et al. 2014, Athey and Wager 2017, Kallus 2018, Kallus and Zhou 2018, Kallus 2019, Gao et al. 2021, Sondhi et al. 2020, Swaminathan and Joachims 2015a). Most of these works assume that the historical data is generated by a previous decision maker and studies how to estimate the treatment effect or find an optimal algorithmic policy using proposed estimators and a specific policy class. Importantly, this line of work has not considered nor developed a learning algorithm for contexts that can benefit from a hybrid team of humans and AI to enhance decision performance. In the AI learning literature, policy learning from observational data is also known as Offline Policy Learning/Optimization (OPL/OPO), Counterfactual Risk Minimization (CRM), or Batch Learning from Bandit Feedback (BLBF) (Swaminathan and Joachims 2015a, Joachims et al. 2018, Wang et al. 2019b, Lawrence et al. 2017, Kato et al. 2020, Si et al. 2020). While OPL/OPO sometimes also refer to reinforcement learning problems with state transitions (Fujimoto et al. 2019), CRM and BLBF typically refer to the problem setting which can be framed as an offline contextual bandit problem. Inverse propensity weighting (IPW) (Rosenbaum 1987) is utilized to account for the bias in the actions reflected in the data. Swaminathan and Joachims (2015a) introduce the CRM principle with a variance regularization term derived from an empirical Bernstein bound for finite samples to constrain the learned policy to be much different from the historical policy. In order to reduce the variance of the IPW estimator, Swaminathan and Joachims (2015b) propose a self-normalized estimator, while Joachims et al. (2018) proposes an estimator that is easy to optimize using stochastic gradient descent for deep nets. In this paper, we extend the traditional problem of policy learning from observational data, also known as CRM and BLBF, and propose a method to incorporate human-algorithm decision complementarity in the optimization. We show that human-AI collaboration can yield further improvements over either the human or the algorithm’s performance, and propose a method that includes a router along with a policy optimized to complement a human. Furthermore, we exhibit how a joint optimization of both elements leads to a further improvement and extend this approach to demonstrate how decision rewards can be further improved when applied to leverage complementarity of a human-AI team with multiple human decision-makers, deterministic actions and covariate shift. Our setup is also related to ensemble bandits (Pacchiano et al. 2020), which aims to identify the optimal bandit algorithm in an online fashion. Our goal is distinct in that it aims to learn the best human-AI hybrid system and it aims to do so from observational data (*i.e.*, offline).

Personalizing for Diverse Worker Skills. Many supervised learning tasks require human labeling, which is often imperfect (Huang et al. 2017, Yan et al. 2011, Gao and Saar-Tsechansky 2020). Crowdsourcing research has studied the problem of learning from multiple noisy labelers across various settings and goals (Yan et al. 2011). For example, Yan et al. (2011) proposed a probabilistic framework to infer the label accuracy of workers and choose the most accurate worker for annotation. Huang et al. (2017) allocated workers with different costs and labeling accuracies to cost-effectively acquire labels for classifier induction. The importance of sociocultural diversity in the machine learning pipeline has also received increasing attention (Fazelpour and De-Arteaga 2022). Recent work has empirically demonstrated the implications of *annotator identity sensitivities* (Sachdeva et al. 2022) and developed multi-annotator methods as a way to better learn from diverse perspectives (Davani et al. 2022). All of these works, however, consider the context of supervised classification tasks in which the human input is limited to labeling historical data for model training, rather than performing decision-making at test time. Also related to our work is decision-theoretic active learning that considers human-AI hybrid configurations for classification (Nguyen et al. 2015). Our work focuses on training models from observational data, considering that humans can be asked to make decisions after the model is trained. This enables the human-AI team to achieve the highest average reward, making models integral to the human-AI decision-making team.

Decision-Making with Deterministic Actions In the offline bandit literature, deterministic actions often bring difficulty in estimating counterfactuals. This is known as a violation of the positivity assumption, and it has been shown that it is impossible to infer rewards for unseen actions in the observational data when actions are deterministic (Langford et al. 2008, Lawrence et al. 2017). To deal with this issue, Sachdeva et al. (2020) proposes to constrain the future policy to be similar to historical policy. This idea is similar to behavior cloning (Torabi et al. 2018), which is widely used in reinforcement learning. However, this is infeasible in settings in which the algorithmic designers do not have control over the instances that may be encountered in the future. In this paper, we explore the possibility of utilizing the fact that in many contexts, deterministic actions mean that humans are confident and knowledgeable in their actions (De-Arteaga et al. 2021) to design efficient algorithms in the presence of missing counterfactuals.

Decision-Making with Domain Shifting Machine learning algorithms often assume that training data and testing data follow the same distribution, but this assumption may be violated in practice. To tackle this challenge, Faury et al. (2020), Si et al. (2020) consider distributionally robust off-policy contextual algorithms that minimize the maximum risk of all distributions of a \mathcal{F} -divergence neighborhood of the empirical distribution. Kato et al. (2020) addresses this by estimating the density ratio between training and evaluation data, which requires access to the data density after covariate and concept shifting. Different from these methods, we propose to allocate some unseen instances to humans, who might generalize better compared to machine learning algorithms. Note that this does not require us to make any specific assumptions about the generation process of shifting, which is more general and practical in reality. Conveniently, these methods (Faury et al. 2020, Si et al. 2020, Kato et al. 2020, Swaminathan and Joachims 2015a) can also be integrated into our objective by adding the proposed regularization as a distributionally robust alternative to the baseline inverse propensity score weighting. In order to determine whether an instance is unlikely to be generated from the same distribution as the training data, an out-of-distribution detection (OOD) method is needed. Many such methods have been proposed. A common post-hoc approach is to design an uncertainty or confidence score from a supervised model to measure the likelihood of a data instance being an outlier (Oberdiek et al. 2018, Lee et al. 2018, Bahat and Shakhnarovich 2018). There are also works that only utilize in-distribution data for outlier detection, such as one-class classification (Li et al. 2003, Swersky et al. 2016). We refer readers to a more thorough discussion of OOD algorithms by Pang et al. (2021). In this paper, we use one-class SVM (Schölkopf et al. 1999) and deep one-class classification algorithms (Ruff et al. 2018).

3 Problem Statement and Main Algorithms

We consider settings in which either a human or an AI can make decisions autonomously, and in which each instance (and corresponding feature set) for which decisions must be taken, $x = x_1, \dots, x_N \in \mathcal{X}$ is drawn independently from $\mathbb{P}(x)$, in which \mathcal{X} represents an abstract space. We consider decisions that correspond to selecting a discrete action $a \in \mathcal{A}$, which gives rise to a reward observed following the action, $r = r(x, a)$ in

which $r \in [0, 1]$. Henceforth, $r(x, a)$ denotes the potential reward for action a and instance x , and r denotes the observed (realized) reward for instance x . We assume the observational data was generated by human decision makers $h \in \mathcal{H}$, who selected actions for all instances in the historical data available. The decision policy historically used by human decision makers is denoted as $\pi_0(a|x)$, which corresponds to what is often known as *propensity score* or *behavior policy* in policy learning from observational data and *logging policy* in CRM/BLBF. Consequently, the observational data includes instances with features x , an action a taken for each instance x , and the corresponding observed reward, r . We also consider that a human’s decision-making may incur on a predefined cost $C(x)$, reflecting the human’s effort and the corresponding cost of expertise.

Given the observational data, we aim to learn a deferral collaboration system for future instances. The system is composed of a routing algorithm, $d_\phi(h|x)$, that allocates each instance to either a human decision maker h or a learned algorithmic policy, $\pi_\theta(a|x)$. The task is to learn the routing algorithm, $d_\phi(h|x)$, and the algorithmic policy, $\pi_\theta(a|x)$, in order to maximize expected reward, and with the goal of achieving higher expected rewards than could be otherwise achieved by either a human or an algorithm alone. In this paper, we focus on differentiable policy and routing model classes in which π_θ , d_ϕ are parametrized by θ and ϕ , respectively.

We now discuss our proposed solutions to the deferral collaboration problem. We begin with preliminaries on the task of learning a decision policy from observation data. This is followed by our approach to learn an algorithmic decision policy that offers advantageous complementarity with respect to the human decision-maker, and a means to effectively exploit this complementarity. We first propose a solution to sequentially learn π_θ and d_ϕ , followed by a solution in which we optimize them jointly.

3.1 Preliminaries: Learning a Decision Policy from Observational Data

Learning a decision policy $\pi_\theta(a|x)$ from observational data corresponds to the goal of learning a policy that maps instances to actions to maximize the expected reward given by:

$$\mathbb{E}_{x \sim \mathbb{P}(x), a \sim \pi_\theta(a|x)} r(x, a) \quad (1)$$

Since the observational data is generated by $x \sim \mathbb{P}(x)$, $a \sim \pi_0(a|x)$, $r(x, a) \sim \mathbb{P}(r|x, a)$, learning the optimal policy from observational data has a distribution-mismatch problem (Shalit et al. 2017, Gao et al. 2021). Under assumptions that we outline below, inverse propensity weighting (Rosenbaum and Rubin 1983) can be used to optimize for the policy’s parameters θ over the observational data produced by the human decision makers’ policy, $\pi_0(a|x)$ (Swaminathan and Joachims 2015a, Rosenbaum 1987), as follows:

$$\mathbb{E}_{x \sim \mathbb{P}(x), a \sim \pi_\theta(a|x), r(x, a) \sim \mathbb{P}(r|x, a)} r(x, a) = \mathbb{E}_{x \sim \mathbb{P}(x), a \sim \pi_0(a|x), r(x, a) \sim \mathbb{P}(r|x, a)} \frac{\pi_\theta(a|x)}{\pi_0(a|x)} r(x, a) \quad (2)$$

Thus, using the subscript i to indicate the i -th example in the empirical observational data containing N instances, we can learn a policy $\pi_\theta(a|x)$ by optimizing the objective above for θ using the estimator:

$$\max_{\theta} \frac{1}{N} \sum_{i=1}^N r_i \frac{\pi_\theta(a_i|x_i)}{\pi_0(a_i|x_i)} \quad (3)$$

Then the optimization problem can be solved using gradient descent in which the gradient is

$$\sum_{i=1}^N \frac{\nabla \log \pi_\theta(a_i|x_i)}{N} \frac{\pi_\theta(a_i|x_i)}{\pi_0(a_i|x_i)} r_i. \quad (4)$$

Here, $\pi_\theta(a|x)$ can be any differentiable model, including logistic regression, neural network (*e.g.*, a multi-layer perceptron (Hastie et al. 2009) or a convolutional neural network (He et al. 2015)) in which the input is the feature x and the output is the action assignment probability $\pi_\theta(a|x)$ normalized by a softmax layer.

When $\pi_0(a|x)$ is not given, it is usually estimated by an additional classifier, such as logistic regression or random forest, which yield $\hat{\pi}_0(a|x)$ trained on observational data (Kallus 2021, Athey and Wager 2017, Shalit et al. 2017). With observational data, the conditional average treatment effect (CATE) is not always identifiable. We

can thus build on Rubin’s potential outcome framework and assume consistency and strong ignorability, which are commonly considered in prior work and that constitute sufficient conditions for identifying CATE (Pearl 2017). For completeness, we formally present the assumptions below (Rubin 2005):

Assumption 1. (*Ignorability*) $r(x, a) \perp\!\!\!\perp a|x, \forall a \in \mathcal{A}$.

Assumption 2. (*Overlap / Positivity*) $\mathbb{P}(A = a|x) > 0, \quad \forall a \in \mathcal{A}$

Assumption 3. (*Consistency*) For an instance with features x and $\forall a \in \mathcal{A}, \mathbb{E}(r(x, a)|x, a) = \mathbb{E}(r|x, a)$ in which r is the realized reward.

The ignorability assumption is a standard assumption made in the causal inference literature. It states that the historical decision was chosen as a function of the observed covariates X and that there are no unmeasured confounding variables which affect both the decision and the reward. Previous work has argued that this assumption is particularly defensible in some contexts of prescriptive analytics (Bertsimas and Kallus 2020). In particular, this assumption holds when past actions were based on information observed by the decision maker. This information is recorded and used as features for the machine learning model. When it is violated, an alternative ought to be considered, such as the availability of an instrumental variable (Angrist et al. 1996), or a worst-case risk defined on an uncertainty set (Kallus and Zhou 2018). The overlap assumption (also known as the positivity assumption), requires that all actions must have a non-zero chance of being prescribed to all instances observed historically. This assumption allows unbiased evaluation of the policy reward (Langford et al. 2008, Kallus 2021). In Section 4, we propose an extension in which we relax this assumption. Finally, the consistency assumption states that the potential outcome under action a matches the actual value, which bridges the potential outcome and the observed reward, permitting us to use empirical data for policy learning. The above assumptions have been established as sufficient conditions for the identifiability of conditional average treatment effect or individual treatment effect (Angrist et al. 1996, Hirano et al. 2003, Pearl 2010, Swaminathan and Joachims 2015a).

3.2 Learning and Exploiting Complementary Policy for Deferral Collaboration

In this section, we consider how to train an algorithmic policy that can *complement* human decision makers. Simultaneously, we will address how to productively exploit such potential complementarity in our context of deferral collaboration between humans and an algorithmic decision-maker. A key observation about our problem is that the algorithmic decision policy is not given. Rather, different algorithmic decision policies can be produced to offer different abilities and subsequent complementarity with respect to the human decision-maker. Such differential abilities can be achieved by developing an algorithmic policy that aims to produce better performance on different subset of instances at the possible cost of worse performance over others. Thus, our goal is to produce an algorithmic policy, $\pi_\theta(a|x)$, that can best complement the human. An important element towards our goal, is that leveraging the algorithmic policy by effective routing of different instances to the entity that is likely to yield the best reward by $d_\phi(h|x)$. Lastly, recall that the human decision-maker may incur a cost, $C(x)$, for producing a decision for an instance with features x . This allows us to capture the cost of the human’s time and expertise in producing the decision. The human’s decision cost can be set to zero in contexts in which the best decision is desired, irrespective of the human’s cost.

We propose two alternative methods for learning an algorithmic decision policy that best complements the human and for effectively exploiting the complementarity it gives rise to.

3.2.1 Two-Stage Collaboration.

The first approach we propose considers learning a complementary algorithmic policy first, followed by learning a routing policy that best exploits the potential human-AI complementarity. Specifically, we first produce an algorithmic policy $\pi_\theta(a|x)$ from the observational data using Equation (3), as in traditional policy learning methods (Swaminathan and Joachims 2015a, Joachims et al. 2018). An additional routing algorithm $d_\phi(h|x)$ can be subsequently learned to decide whether an instance x would be best assessed by a human decision maker h or by the algorithm $\pi_\theta(a|x)$, as shown in Figure 1. We then learn $d_\phi(h|x)$ to estimate the likelihood that the human’s decision (or, alternatively, the algorithmic policy) will yield a superior reward. Consequently, given an existing

algorithmic policy $\pi_\theta(a|x)$, maximizing the LCP-HAI system’s reward entails learning $d_\phi(h|x)$ that maximizes the following objective:

$$\max_{\phi} \sum_{i=1}^N d_\phi(h|x_i)(r_i - C(x_i)) + \frac{(1 - d_\phi(h|x))\pi_\theta(a_i|x_i)}{\hat{\pi}_0(a_i|x_i)} r_i \quad (5)$$

Thus, for each instance x_i , we have a probability $d_\phi(h|x_i)$ to rely on a human decision maker or $1 - d_\phi(h|x_i)$ to use the algorithm’s decision. The LCP-HAI’s objective thus is a weighted average of human and algorithm reward. During deployment of the deferral system, an instance is routed to a human decision maker when $d_\phi(h|x_i) > 0.5$ and to the algorithmic policy otherwise. A description of the algorithm is detailed in Algorithm 1. We refer to this method and corresponding optimization objective in Equation 5 as Two-Stage Collaboration (TS), since the algorithmic decision policy and the routing model $d_\phi(h|x)$ are trained *sequentially*.

3.2.2 Joint Learning of a Decision and Routing policies.

Thus far, we considered learning a complementary decision policy and routing policy sequentially. We now consider whether a joint optimization of both the decision and routing policies can further improve the deferral collaboration performance by optimizing the same objective outlined in Equation (5), while this objective is optimized jointly with respect to both the algorithmic decision-making model parameters θ and those of the routing policy, ϕ ,

$$\max_{\phi, \theta} \sum_{i=1}^N d_\phi(h|x_i)(r_i - C(x_i)) + \frac{(1 - d_\phi(h|x))\pi_\theta(a_i|x_i)}{\hat{\pi}_0(a_i|x_i)} r_i \quad (6)$$

To motivate how a joint learning procedure can outperform the two-stage approach, we outline a simple example that is illustrated in Figure 2. A similar rationale for classification decision tasks was proposed in Mozannar and Sontag (2020). Here, we outline its connection to our context in which we aim to learn a complementary decision policy. Consider a setting in which the action yields a non-negative reward for all instances. For instance, a promotion may result in purchase decisions for some customers while not affecting other customer decisions. Therefore, some instances can be viewed as treatment responders and others are non-responders (Angrist et al. 1996, Kallus 2019). Thus, an action will only yield a positive reward if the instance is a treatment responder. Consequently, given the treatment monotonicity, and given the action and outcome are both binary, the policy learning problem reduces to binary classification. This holds because it is sufficient to identify which instances will comply with the action to identify the instances for which actions can yield positive reward (Kallus 2019).

Let us assume: 1) the feature space is two-dimensional (for visualization purposes), 2) the model class used to learn a decision policy is chosen to be hyperplanes, 3) responders and non-responders to the actions are distributed as shown in Figure 2, and 4) the human decision-making expert yields a non-linear decision boundary to classify instances (namely, whether to prescribe the action). For simplicity of the exposition, let us also assume that the human can classify all instances perfectly. In Figure 2 the dashed lines represent the solution produced by the joint learning approach, and Green, Red, Black represent Algorithm, Human expert, and Routing algorithm respectively. The solid lines represent the solutions returned by the two stage learning method and Blue, Orange represent Algorithm and Routing algorithm respectively. Here, human decision makers can perfectly identify treatment responders but will incur a cost. Therefore, it is desired human experts solve fewer instances. Given that the model class used to learn the decision policy has limited capacity, using a two stage procedure might first estimate an algorithm similar to the blue solid line (since there is no perfect hyperplane that can identify all instances correctly). Consequently, one possible solution for the routing model is a hyperplane (orange) to the left of all points, assigning all tasks to the (more costly) human, and yielding suboptimal performance. With joint optimization, the routing algorithm can separate the left part and right part of the instances by the black dashed line, therefore reducing the human effort and increase the average team reward. Overall, given that both the algorithmic decision policy and router are learned and interdependent on the other policy’s performance, joint learning can offer tangible benefits over those that can be produced when this interdependence is not accounted for. It is also possible that two stage training has a comparable performance with joint training when the algorithmic policy trained from the first stage coincides with the joint training solution.

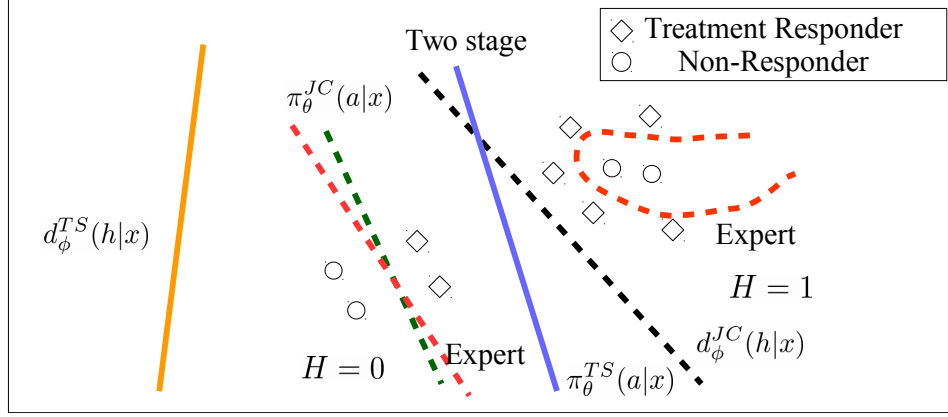


Figure 2: Joint Learning versus Two Stage Procedure: Dashed line represents the solution from joint learning (Green: Algorithm, Red: Expert, Black: Router) and the blue solid line represents the fixed algorithm solution from the two stage procedure, orange solid line represents the routing model for two stage training.

Algorithm 1 LCP-HAI: Learning Complementary Policy for Deferral Human-AI Collaboration

Require: Observational data $\mathcal{D} = \{x_i, a_i, r_i\}_{i=1}^N$, policy, router, propensity score model class: $\pi_\theta, d_\phi, \hat{\pi}_0$.

Learn $\hat{\pi}_0$ from observational data.

if Two Stage Training: **then**

Via gradient ascent, solve: $\hat{\theta} = \arg \max_{\theta} \sum_{i=1}^N \frac{\pi_{\theta}(a_i|x_i)}{\hat{\pi}_0(a_i|x_i)} r_i$.

With fixed $\hat{\theta}$, via gradient ascent, solve:

$\hat{\phi} = \arg \max_{\phi} \sum_{i=1}^N d_{\phi}(h|x_i)(r_i - C(x_i)) + \frac{(1-d_{\phi}(h|x))\pi_{\theta}(a_i|x_i)}{\hat{\pi}_0(a_i|x_i)} r_i$

end if

if Joint Training: **then**

Via gradient ascent, solve:

$\hat{\theta}, \hat{\phi} = \arg \max_{\theta, \phi} \sum_{i=1}^N d_{\phi}(h|x_i)(r_i - C(x_i)) + \frac{(1-d_{\phi}(h|x))\pi_{\theta}(a_i|x_i)}{\hat{\pi}_0(a_i|x_i)} r_i$

end if

Testing: for new instance x , query the human decision if $d_{\phi}(h|x) > 0.5$, otherwise output $\arg \max_a \pi_{\theta}(a|x)$.

4 Extensions

In this section, we develop three extensions to our main LCP-HAI algorithms to address specific challenges arising in important practical contexts. Specifically, we consider: 1) contexts with multiple and diverse human decision makers (Section 4.1), 2) contexts in which there exists deterministic human actions in observational data (Section 4.2), and 3) contexts in which instances encountered during deployment can be out-of-distribution as a result of covariate shift (Section 4.3).

4.1 Personalization: Learning Complementary Policy for Multiple, Diverse Humans

Thus far, we implicitly considered contexts in which the human-AI team consists of a single human, or equivalently, of multiple homogeneous human decision makers. In some practical contexts, a human-AI team may include multiple human decision-makers having different expertise. For instance, one pricing specialist may produce superior rewards for pricing commercial realty in a given locale, while another specialist may have greater success in pricing residential realty across neighborhoods. In such settings, it is desirable to learn a complementary algorithmic policy that can effectively identify opportunities to complement all of the experts. This allows us to effectively learn to exploit different complementarity both between the algorithmic decision policy and the human, as well as amongst different human experts. We propose a personalized approach that accounts for the varying expertise of different humans so as to further improve the human-AI team’s performance. For the settings we consider here, our learning of the algorithmic policy need not be adapted. This is because one of our goals remains to learn an algorithmic decision policy to complement the decision-maker(s) that produced the observational data. However, the routing algorithm must now decide whether to defer an instance to the algorithm or to a human, and, in the latter case, which human decision-maker should be selected to make the decision.

Recall that one challenge in the overall problem we address is that typically the human’s underlying decision-making policy, $\pi_0(a_i|x_i, h_i)$, is not given. Thus, in the case of as a single human (or, equivalently, homogeneous human decision makers) we model the human’s policy with $\pi_0(a|x)$. In the context of multiple diverse human decision-makers, we model each individual human decision maker’s policy by training a classifier on the observational data, conditioned on their identity, h_i . That is, we include the decision-maker’s identity as an additional covariate so that we estimate $\hat{\pi}_0(a|x, h_i)$. Alternatively, one can also train a separate supervised learning model for each human decision maker. In general, a single model, learned from the observational data conditioned on their identity, h_i , is likely to be advantageous when the human decision makers exhibit similar decision behaviors (Künzel et al. 2019). In this case, instances undertaken by one human decision-maker are informative of the choices others may take for the same instances. In contrast, separate models for each human’s policy may perform better when different individuals’ decision patterns vary significantly. In that case, decisions instances handled by one decision-maker are not informative of the choices others may take. In the experimental results we report here, we model decision-makers’ policies using a single model, induced from the observational data and conditioned on the human’s identity. We can thus train the LCP-HAI algorithm by:

$$\max_{\theta, \phi} \sum_{i=1}^N (r_i - C_j(x_i)) \frac{d_\phi(h_i|x_i)}{\hat{d}_0(h_i|x_i)} + r_i \frac{d_\phi(\perp|x) \hat{\pi}_\theta(a_i|x_i)}{\hat{d}_0(h_i|x_i) \hat{\pi}_0(a_i|x_i, h_i)} \quad (7)$$

where \perp denotes that the algorithmic policy is used to make a decision. In addition, $\hat{\pi}_0(a|x_i, h_i)$ denotes the estimated likelihood that action a_i is selected by the human decision-maker h_i who undertook the decision for instance x_i . Similarly, $\hat{d}_0(h_i|x_i)$ denotes the estimated likelihood that, historically, h_i would have been assigned to undertake the decision for instance x_i . Here we consider that the assignment of human decision makers to instances depends on the observed features x , and there are no unobserved confounders. This assumption is frequently met in practice; for instance, when the assignment of customer service representative to complaint cases is non-random, variables that inform the assignments, such as type of insurance or nature of the complaint, can often be included in the data. When assignments are randomized, then $d_0 = 1/K$, where K is the total number of human decision makers.

At testing time for a new arriving x , we can then choose $h = \arg \max_{h_i} d(h_i|x)$ as the decision maker for a given instance x .

4.2 Violation of the Overlap Assumption / Deficient Support

In some contexts, human experts may assign actions deterministically to some instances. For example, an airline customer service specialist may never offer an expensive compensation, such as twice the airfare, to a customer who is not severely impacted by a delay. In such contexts, there is a violation of a core assumption of overlap, defined in Assumption 2, necessary for using inverse propensity score based methods for policy learning on observational data (Langford et al. 2008, Sachdeva et al. 2020). This is because when there exists an x such that $\pi_0(a|x) = 1$ for a given a , we never have the chance to observe the potential rewards for different actions $a' \in \mathcal{A}, a' \neq a$, which results in a biased reward estimation. Specifically, Proposition 1 formally characterizes this bias in the inverse propensity score estimator.

Proposition 1 (Bias in Policy Evaluation with Deterministic Action (Sachdeva et al. 2020)). *If there exists a non-empty set of instances $x \in \mathcal{X}$ for which there are actions with zero support, $\mathcal{U}(x, \pi_0) = \{a \in \mathcal{A} : \pi_0(a|x) = 0\}$, then the reward estimation of the inverse propensity score weighting estimator has a bias of*

$$\mathbb{E}_{x \in \mathcal{X}} \left[- \sum_{a \in \mathcal{U}(x, \pi_0)} \pi(a|x)r(x, a) \right]. \quad (8)$$

It can be observed that, as the number of deterministic actions or actions with zero support increases, the (asymptotic) bias tends to be larger.

However, in some important contexts, expert deterministic choice of an action is rather informative, given that it may be driven by domain knowledge that supports the selection as the optimal action (De-Arteaga et al. 2021). For instance, in content moderation, certain posts may always be removed by human moderators because they are confident about its negative impact on the platform, *e.g.*, based on ample prior evidence or strict legal directives that justify this decision. Here we propose an extension of our approach that allows us to leverage this simple but useful observation of a common phenomenon in many important contexts.

Specifically, we propose a data augmentation algorithm that utilizes consistent human expert decisions to reduce the estimation bias from deterministic actions in the setting of binary rewards and actions. While not all encompassing, this setting corresponds to important business and policy decision-making domains (Kallus 2019) in which predictive models are already deployed or considered for deployment, ranging from content moderation to e-commerce. For instance, decisions in personalized marketing could correspond to whether or not to send a coupon to the customer, with the reward being the customer’s purchase decision.

In such contexts, we propose to impute the unseen counterfactual of the deterministic actions in the observational data with the *sub-optimal* reward. That is, we assume that if an action is never chosen by humans for a given instance, that action would yield a suboptimal reward. Subsequently, the unbiased inverse propensity weighting objective with all known counterfactuals can be formally written as,

$$\frac{1}{N} \sum_{i=1}^N \left(\frac{\pi_\theta(a_i|x_i)}{\hat{\pi}_0(a_i|x_i)} r_i + \sum_{a \in \mathcal{U}(x_i, \pi_0)} \pi_\theta(a|x_i)r(x_i, a) \right) \quad (9)$$

Assuming that humans make deterministic actions on instances from the subset $\mathcal{S} \subset \mathcal{X}$, and given the suboptimal and optimal binary rewards: $r^s < r^o$, we impute \mathcal{S} from observational data to include the unseen instances with the suboptimal actions a_i^c and corresponding suboptimal reward r^s . Consequently, the evaluation objective of the reward with expert consistency assumption (denoted as EC-IPS) can be written as:

$$\frac{1}{N} \left(\sum_{x_i \in \mathcal{S}} \left(\pi_\theta(a_i|x_i)r_i + \pi_\theta(a_i^c|x_i)r^s \right) + \sum_{x_i \in \mathcal{X} \setminus \mathcal{S}} \frac{\pi_\theta(a_i|x_i)}{\hat{\pi}_0(a_i|x_i)} r_i \right) \quad (10)$$

In settings where domain knowledge justifies the assumption stating that consistent experts often produce optimal deterministic actions, we can thus reduce the inductive bias. Similarly, we can write the human-AI collaboration objective with deterministic actions as,

$$\begin{aligned}
& \sum_{i=1}^N d_\phi(h|x_i)(r_i - C(x_i)) + \\
& \sum_{x_i \in \mathcal{S}} \left(\pi_\theta(a_i|x_i)r_i + \pi_\theta(a_i^c|x_i)r_s \right) (1 - d_\phi(h|x_i)) + \sum_{x_i \in \mathcal{X} \setminus \mathcal{S}} \frac{\pi_\theta(a_i|x_i)}{\hat{\pi}_0(a_i|x_i)} r_i (1 - d_\phi(h|x_i))
\end{aligned} \tag{11}$$

Implications of Biases in consistent Human Decisions It is useful to understand how our proposed extension for a setting with deterministic, consistent human decisions might be impacted when the assumption that such behaviors are optimal is violated. Hence, we aim to characterize how bias of human decision makers translates into the algorithmic decision policy.

Theorem 1. *Assume that for a subset of the instances in which humans exhibit consistent actions, $x \in \mathcal{B} \subset \mathcal{S}$, human decision makers make suboptimal actions on \mathcal{B} , yielding an average regret $\delta = (r^s - r^o)\mathbb{E}_{x \in \mathcal{X}} \mathbb{1}(x \in \mathcal{B})$. With an exact propensity score $\pi_0(a|x)$, the imputation of our objective in Equation 10 has asymptotic bias of*

$$\delta' = (r^s - r^o)\mathbb{E}_{x \in \mathcal{S}, a \sim h(a|x)} \pi_\theta(a^c|x) \mathbb{1}(x \in \mathcal{B}) \leq \delta. \tag{12}$$

Proof. Note that Equation (10) is unbiased when $\delta = 0$, and the second term in Equation (10) with inverse propensity weighting is unbiased when propensity score is correct. Thus, the asymptotic bias of imputed estimator Equation (10) in estimating policy π_θ 's reward is,

$$\mathbb{E}_{x \in \mathcal{X}, a \sim h(a|x)} \left(\pi_\theta(a|x)r(x, a) + \pi_\theta(a^c|x)r^s - \pi_\theta(a|x)r(x, a) - \pi_\theta(a^c|x)r(x, a^c) \right) \mathbb{1}(x \in \mathcal{S}) \tag{13}$$

$$= \mathbb{E}_{x \in \mathcal{X}, a \sim h(a|x)} \pi_\theta(a^c|x) (r^s - r(x, a^c)) \mathbb{1}(x \in \mathcal{S}) \tag{14}$$

$$= (r^s - r^o) \mathbb{E}_{x \in \mathcal{S}, a \sim h(a|x)} \pi_\theta(a^c|x) \mathbb{1}(x \in \mathcal{U}) = \delta' \tag{15}$$

Since $0 \leq \pi_\theta \leq 1$, then $\delta' \leq \delta$. □

Our analysis shows that the only source of estimation bias in our imputation objective comes from the human bias itself. When human decision makers make no mistakes, the imputation objective will be unbiased. Importantly, when they do make mistakes, the bias of the imputation objective will be upper-bounded by the human bias. Since there are no principled established methods for learning with deterministic actions from observational data (Langford et al. 2008, Swaminathan and Joachims 2015a), our results constitute a potential pathway to address this challenge in settings where deterministic actions are driven by domain expertise. Since the bias of the policy evaluation for future deferral collaboration policy leveraging expert consistency is bounded asymptotically by the human decision maker's bias in making consistent decisions, this assumption needs to be carefully examined in practice if practitioners want to apply it.

4.3 Covariate Shift

Machine learning algorithms are known to not generalize well to out-of-distribution (OOD) test data. Such data may be frequently encountered during deployment, especially if there is a distribution shift that leads to covariates having a different domain than that observed in the observational data. If the algorithmic policy never observes certain types of instances during training, its performance for such instances cannot be reliably anticipated or guaranteed. Similarly, since the router's decision of whether the human is likely to select an action of higher reward is also estimated by a machine learning algorithm, these decisions are also less reliable for OOD instances. Thus, for both the algorithmic policy and the router algorithm, one depends on the extrapolation power of the routing model and the algorithmic policy for such instances, which requires strong functional form assumptions and is thus unreliable.

Meanwhile, prior work has demonstrated that humans can exhibit relatively superior performance in such contexts. For example, for the task of image recognition, recent work has documented how four-year-old children can outperform the state-of-the-art AI algorithm on OOD samples (Huber et al. 2021). This gives rise to opportunities to improve the reliability of a deferral collaboration system by detecting and then deferring OOD

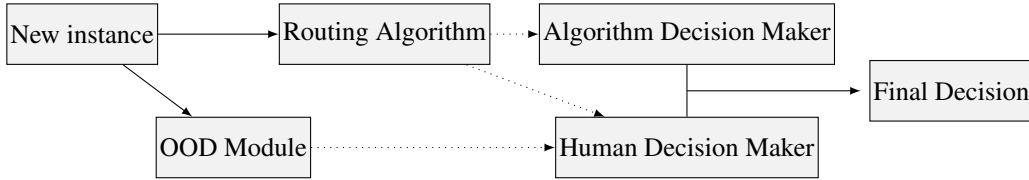


Figure 3: Decision Making Process for Human-AI Collaboration with Covariate Shift. The OOD Module represents any out-of-distribution detection algorithm.

instances to humans. In general, trustworthiness of any deferral collaboration system can benefit from reducing the risk that algorithms autonomously undertake decisions for OOD instances.

Motivated by the goals above, we propose a variant of our LCP-HAI system shown in Figure 3, which aims to detect OOD instances and defer them to the human. When a new instance arrives, an OOD detection module estimates whether an instance is an OOD data point by comparing it to the training data distribution. If the OOD detection algorithm decides this data instance is out-of-distribution, the instance will be deferred to the human decision-maker. The training process of the routing algorithm and algorithmic decision policy remain the same as before. Thus, in cases where the OOD module assesses that all instances are in-distribution samples, then the system is equivalent to LCP-HAI introduced before.

Post-deployment OOD parameter tuning As post-deployment data becomes available, it is possible to tune the OOD parameters to account for the fact that the initial deferral model may have some extrapolation power. Assuming we have access to a small set of data after the deployment of a deferral collaboration system, it is possible to do a quick tuning of the OOD module’s hyperparameter(s). Such tuning allows us to account for the fact that the algorithmic policy in the deferral collaboration system may still outperform its human counterpart in some OOD cases, especially if instances are not too distant from the original distribution.

Assume the out-of-distribution detection algorithm has a hyperparameter space \mathcal{P} . The shifted observational data can then be used to select the best hyperparameter that yields the highest overall reward. More specifically, after θ, ϕ are learned for our LCP-HAI approach, it is possible to learn the hyperparameter p by optimizing as follows:

$$\max_{p \in \mathcal{P}} \sum_{i=1}^N \mathbb{I}(\mathcal{O}_i(p) = 0) \left(d_\phi(h|x_i)(r_i - C(x_i)) + \frac{(1 - d_\phi(h|x_i))\pi_\theta(a_i|x_i)}{\hat{\pi}_0(a_i|x_i)} r_i \right) + \mathbb{I}(\mathcal{O}_i(p) = 1)(r_i - C(x_i)) \quad (16)$$

where $\mathcal{O}_i(p) = 0$ means that the OOD algorithm with hyperparameter p predicts the instance x_i to be an in-distribution instance, and $\mathcal{O}_i(p) = 1$ means that it predicts the instance to be an out-of-distribution sample. The hyperparameter in most out-of-distribution detection algorithms controls the boundary of the OOD algorithm \mathcal{O} . For instance, in one class SVM, we can use the upper bound on the fraction of training errors and a lower bound on the fraction of support vectors to control what points we consider to be outliers.

5 Experiments

We begin with experimental results of our main LCP-HAI algorithm, followed by evaluations of the different variants we propose. We later report experimental results of ablation studies and discuss the effect on human-AI complementarity of model class choice used to produce the algorithmic policy. The data statistics used in all of our experiments are available in Appendix B. We include significance tests for the experiments in Appendix C.

Methods We consider the following main baselines for comparison. We will introduce the variants we propose for personalization, deterministic action and covariate shift in Section 5.1, Section 5.3 and Section 5.4, respectively. Human experts are queried randomly with equal probabilities when no personalization is applied.

- **Human:** This baseline exclusively queries humans for decisions in the test set.
- **Algorithm Only (AO):** This baseline trains the algorithmic policy using Equation (3), and exclusively uses the algorithmic policy for decisions in the test set.
- **Two Stage Collaboration (TS):** The objective of the two stage baseline is shown in Equation 5. Note that the algorithmic policy for two stage is the same as AO; during test time, a routing model decides whether to query the human or the algorithm for a decision.
- **Joint Collaboration (JC):** The objective jointly optimizes both the algorithmic policy and the routing algorithm. Here, the algorithmic policy is trained together with the routing model; during test time, the routing model decides whether to query the human or the algorithm for a decision.

Creating observational data for empirical evaluations Given that observational data in practice never includes rewards for actions not taken, and given that such outcomes are necessary to evaluate methods involving policy learning from observational data, data must be produced for evaluation purposes. In this work, we adopt a common approach from prior work that evaluates policy learning from observational data (Swaminathan and Joachims 2015a, Joachims et al. 2018, Kato et al. 2020). Specifically, the approach involves a conversion of any multi-label (or multi-class) classification dataset into a semi-synthetic decision-making data set, with complete counterfactuals for all possible actions. The conversion builds on the following relationship between the multi-label classification task and the policy task: a policy’s choice of action out of all potential actions can be also formulated as determining which of a predefined set of classes an instance corresponds to. More specifically, given each instance x in a classification task with multi-label $y \in \{0, 1\}^l$, where l is the number of possible labels, and each instance can be associated with one or more labels, then each possible label can be considered as a possible *action* that a policy can select. Subsequently, for each instance x , if the policy selects one of the correct labels for an instance, the corresponding observed reward is 1. The reward is 0, otherwise. Thus, for a policy’s choice of action a , the reward y_a is revealed, producing the test instance (x, y_a) . We follow prior work to create an observational data that represent historical human decisions. We subsequently evaluate different policy choices of actions on the dataset following the above scheme and report the corresponding rewards.

Human Feedback In the experiments, we use both synthetic and real human responses to evaluate the proposed LCP-HAI algorithms. When using real human responses, we apply the approach described above to make use of classification datasets with real human annotations. Datasets that record worker identities allow us to evaluate the proposed personalization variant. We complement evaluations on real human responses with simulations, which allow us to evaluate the proposed methodologies on multiple benchmark datasets, even if these do not include human assessments, which ensures reproducibility via multiple publicly available datasets. Additionally, this enables us to evaluate the proposed work under different settings of variation in worker accuracy. We use two different human behavior models (HBM) in our simulations. First, we fit black box models on different subsets of 30% of the data with full (not partial) ground truth labels and use it to generate synthetic human decisions at both training and testing time. Note that this HBM can exhibit different conditional decision accuracy for different instances. We use random forest (default implementation in scikit-learn package) as the black box model to capture the potential non-linearity in the predictive relationship. The second HBM is motivated by labeling noise in the crowdsourcing literature (Zheng et al. 2017). We assume each expert has a uniform decision noise across all instances, such that for any given instance, the expert will make the correct decision with probability ρ , yielding a reward of 1, and an incorrect decision with a probability $1 - \rho$, resulting in zero reward. The human’s decision is then drawn at random from this distribution. The implementation details of both HBMs are included in the Appendix A.

5.1 Policy Learning for Deferral Collaboration

In this section, we evaluate our proposed deferral collaboration algorithms on both semi-synthetic and real data. We compare the proposed approaches against the human baseline and the algorithm-only baseline.

For all the experiments, we produce the algorithmic policy by learning a three-layer neural network, unless specified otherwise. We also use a three-layer neural network to learn the router model. We later explore the effect of the number of hidden neurons for the algorithmic policy. We use the Adam optimization algorithm to train the networks (Kingma and Ba 2014), with a learning rate of 0.001 for optimization, and train each method

Table 1: Reward on Focus dataset. For all settings, JC has competitive performance against all baselines, demonstrating the possibility of human-AI complementarity in LCP-HAI.

Data	Human	AO	TS	JC
Focus	235.3±4.2	231.2±2.0	231.3±2.0	237.5±1.8

using sufficient epochs until convergence. The estimation of the human’s decision policy, $\hat{\pi}_0$, is trained on the observational data via a random forest model, so as to capture the potential nonlinearity in human decisions. During deployment (after the deferral system is trained), both the algorithmic policy and router models are deterministic – *i.e.*, the router assigns a decision to the entity (the human or the algorithm) that it estimates is most likely to yield the highest decision reward. Similarly, the algorithm selects the action that it estimates will yield the highest reward.

5.1.1 Results for Data with Real Human Responses.

We first evaluate our approach with data containing real human responses. Towards this end, we use a text analysis dataset (Focus) annotated by 5 crowd workers (Rzhetsky et al. 2009). Specifically, Rzhetsky et al. (2009) offers multi-dimensional human assessment on text corpus extracted from biomedical journals. Each paragraph is annotated along six dimensions: Focus, Evidence, Polarity, Trend, and number of fragments. We use the focus dimension for our experiments. For each sentence, human labelers label each segment as generic (*e.g.*, financial support statement), methodology (*e.g.*, describing the approach used in the paper), and science (*e.g.*, whether our experiments support our claim). We use the science dimension annotation since we observe more variation in the worker decision performance (later for examining personalization). If the human labeler correctly annotates the science dimension, the labeler will receive a positive reward. In this task, different segments of a paragraph receive different annotations, and each annotator may partition the paragraph into a different number of sentences. We use the the first fragment in “science” annotation, in which the first fragment is the part of text all labelers agree to be the first segment (so that each labeler is annotating the same text). We set the test set ratio as 30% and randomly sample five training-testing splits to validate our proposed algorithms.

Table 1 shows results of the proposed approaches and its comparison to the baselines. In this experiment, the cost of querying the human set to $C(x) = 0.05$ for all x . The specific cost is chosen since it gives rise to complementarity; later we also conduct ablation study on the effect of cost on the human-AI team complementarity. In Table 1 we find that the human-AI team reward produced by LCP-HAI outperforms the alternatives in which either an algorithm or human decision makers work alone when optimizing the algorithmic policy and the router jointly, emphasizing both the benefits of human-AI collaboration and the importance of joint optimization when training the deferral collaboration system.

5.1.2 Varying Complementarity via Cost.

We also explore how LCP-HAI adapts to varying complementarity between the human and the AI. Specifically, to vary the complementarity, we vary the human decision cost between 0 to 0.5, thereby reducing the ultimate reward that the human can produce and making the human less advantageous, overall. The average rewards are shown in Table 2. For costs greater than 0.3, we notice that human experts have much lower reward than the algorithm in this case, and the human-AI team learns to only output algorithm decisions. In this specific domain, the human outperforms the algorithm when no cost is accounted for, so when the cost is set to 0, the deferral collaboration model performs as well as the human alone. At intermediate cost values, such as 0.05, the deferral collaboration model yields a higher reward than either baseline. In practice, the cost would be determined by external factors, including cost of labor and scarcity of human experts. For example, in the case of content moderation, human moderators are a limited resource, and thus choosing to query the human for a decision would incur a cost. The results shown illustrate the benefit of our proposed approach in a dataset containing real human annotations.

Table 2: Reward on Focus dataset with different expert costs. We examine the effect of human cost and set it from 0 to 0.5. For all settings, JC has competitive performance against all baselines. When the human costs are too high, the human-AI complementarity decreases and the deferral system chooses to only use algorithm decisions.

Data (cost)	Human	AO	TS	JC
Focus (0)	250.3 ±4.2	231.2±2.0	231.5±2.2	250.3±1.3
Focus (0.05)	235.3±4.2	231.2±2.0	231.3±2.0	237.5±1.8
Focus (0.1)	220.3±4.2	231.2±2.0	231.0±2.0	239.7±1.4
Focus (0.3)	160.3±4.2	231.2±2.0	231.2±2.0	230.33±1.8
Focus (0.5)	100.3±4.2	231.2±2.0	231.2±2.0	231.2±1.9

5.1.3 Additional Experimental Results.

We provide experimental results on three additional datasets. We first validate our findings using another dataset with real human responses used in Li et al. (2018) for Multi-Label Learning from Crowds (MLC). Because not all workers annotated the same instances in MLC, we cannot query every worker for each instance; hence, for MLC, we view all 18 workers as a group of workers. Specifically, when querying from a human, a worker is sampled at random from the workers who labeled this instance to offer a decision. The action (decision) made by this worker is sampled from the worker’s chosen label(s).² For MLC, we set the test set ratio as 15% and randomly sampled five training-testing splits to validate our proposed algorithms.

We also conduct experiments for synthetic human decision models on two multi-label datasets, Scene and TMC from LIBSVM repository (Elisseeff and Weston 2002, Boutell et al. 2004), which are used for semantic scene and text classification. This design of evaluation allows us to examine our framework on more empirical data distributions, even if these do not have real human responses. In each experiment, we assume there are three experts present and that the observational data is generated by a random assignment of these experts. Random assignment is frequent in domains in which instances are routed to the first available human, such as customer service representatives assigned to complaint cases (Ensign et al. 2018). To create the observational data, given an expert’s probability of making each decision (determined by either the black-box HBM or the decision noise HBM), we sample the decision according to that probability. When using decision noise HBM, we set ρ to 0.6, 0.7, 0.8, for each expert respectively. Thus, we assume experts exhibit varying decision-making accuracy and all have above-random accuracy. For each decision, we observe a reward of 1 or 0 according to the ground-truth label in the original dataset, and we compare the total reward that each system achieves on the test set. Results are shown for human decision cost at $C(x) = C = 0.3$, which makes the algorithm’s predictions cost-effective and gives us opportunities to exploit the complementarity. Each experiment is run over ten repetitions, and we report the average reward and standard error.

The results shown in Table 3 show that LCP-HAI often yields superior performance and is otherwise comparable to the human or algorithmic only alternatives. An important advantage of the proposed method is that when either the human or the algorithm are best for all instances, it will correctly identify this and route all instances to the relevant entity. The results in Table 3 also show that when the models are jointly optimized, the human-AI team performs significantly better than the algorithm or the human alone. The two stage method does not always offer a significant improvement. Our results on the benefit of the joint optimization further support our motivation of using it in our approach. Namely, given the router and complementary policy are interdependent, LCP-HAI can benefit from a joint optimization of the complementary policy and router.

5.2 Personalization

In this section, we examine the benefit of leveraging diverse expertise of different human decision makers. Therefore, we also evaluate the personalization method we propose in Section 4.1, as described below.

- Joint Collaboration with Personalization (JCP): This algorithm corresponds to the objective in Equation 7. When jointly learning the algorithmic policy and routing model, heterogeneous expertise are considered.

²We note that 5 instances in MLC had no annotation, and we remove them from the dataset.

Table 3: Total reward for different Human Behavior Models. Model refers to the Black Box human behavior model and Noise refers to the uniform human behavior model. Results show averages over 10 runs and presented with standard error. LCP-HAI with joint collaboration is superior to all other alternatives.

	Scene (Model)	Scene (Noise)	TMC (Model)	TMC (Noise)	MLC
Human Only	341.3±7.9	294.8±12.0	4919.5±16.2	3435.1±28.3	53.8±3.6
Algorithm Only	376.3 ±9.3	358.1±7.8	5543.7±109.2	4438.1±131.5	66.5±1.1
Two Stage (TS)	379.2±9.2	364.7±7.7	5642.9±92.2	4361.5±113.6	77.1±0.7
Joint Collaboration (JC)	423.3±5.2	391.9±8.4	5736.1±87.5	4513.6±87.9	79.2±1.0

Table 4: Worker decision accuracy for the Focus dataset.

Worker ID	1	2	3	4	5
Decision Accuracy	0.82	0.93	0.89	0.90	0.64

At testing time, the routing model has the option to assign the new instance either to the algorithm or to one of the different human decision makers.

We conduct the same experiments as we did in Section 5.1, with the additional JCP method. In these results we exclude the MLC dataset since we do not have individual human responses for every instance in this dataset.

For semi-synthetic data, since we have control over all human behaviors, we can assess the benefits of personalization under different settings of human decisions. We consider the same settings of human behavior and modeling described in Section 5.1. For the Focus dataset, each instance is labeled by all workers, allowing us to evaluate our personalization objective using real human responses. In this dataset, the average decision accuracies of five workers are shown in Table 4, calculated as the percentage of accurate decisions for all instances. Most workers demonstrate high decision accuracy in the dataset, but Worker 5 has a relatively lower decision accuracy compared to others, which aligns with our motivation for designing the personalized routing objective with diverse worker expertise. The results for semi-synthetic data and real human responses are shown in Table 5 and Table 6, respectively.

Interestingly, we further observe that the human-AI team with personalized routing has significantly better performance over both baselines of the human or the algorithm alone, both on synthetic and real human responses. This confirms our intuition that better addressing “who is better at what” can increase the deferral system’s decision-making performance and the potential of human-AI collaboration with diverse human experts. We also find that in the Focus dataset, JCP allocates all instances only to Workers 1-4, and none to Worker 5, which shows that personalized routing can correctly identify when an entity (the human or the algorithm) underperforms all others.

Table 5: Reward on semi-synthetic dataset for different Human Behavior Models. Model means the Black Box human behavior model and Noise represents the uniform human behavior model. The decision reward of each method is reported with standard error. Personalization shows significantly better decision performance over JC, which is the best method over other baselines. Results are averaged over 10 runs.

Method	Scene (Model)	Scene (Noise)	TMC (Model)	TMC (Noise)
Joint Collaboration (JC)	423.3±5.2	391.9±8.4	5736.1±87.5	4513.6±87.9
JC with Personalization (JCP)	425.4±4.2	408.3±7.5	5787.8±92.0	4935.1±94.3

5.2.1 Varying complementarity with personalization.

To gain insights into the underlying allocation of decisions by JCP, we explore its allocation of instances to different human decision makers to achieve higher rewards. Towards this end, we simulate five labelers at each iteration with decision accuracy drawn uniformly at random from $[0.7, 1]$ and cost drawn uniformly at

Table 6: Reward on Focus datasets with different expert costs. We examine the effect of human cost and set it from 0 to 0.5. For all settings, JCP has competitive performance against JC, which is the best method over other baselines. When the human costs are too high, the human-AI complementarity decreases and the deferral system chooses to only use algorithm decisions.

Method	$C = 0$	$C = 0.05$	$C = 0.1$	$C = 0.3$	$C = 0.5$
Joint Collaboration (JC)	250.3 ± 1.3	237.5 ± 1.8	239.7 ± 1.4	230.33 ± 1.8	231.2 ± 1.9
JC with Personalization (JCP)	270.4 ± 2.0	257.3 ± 1.9	245.2 ± 1.5	229.8 ± 1.4	230.3 ± 1.5

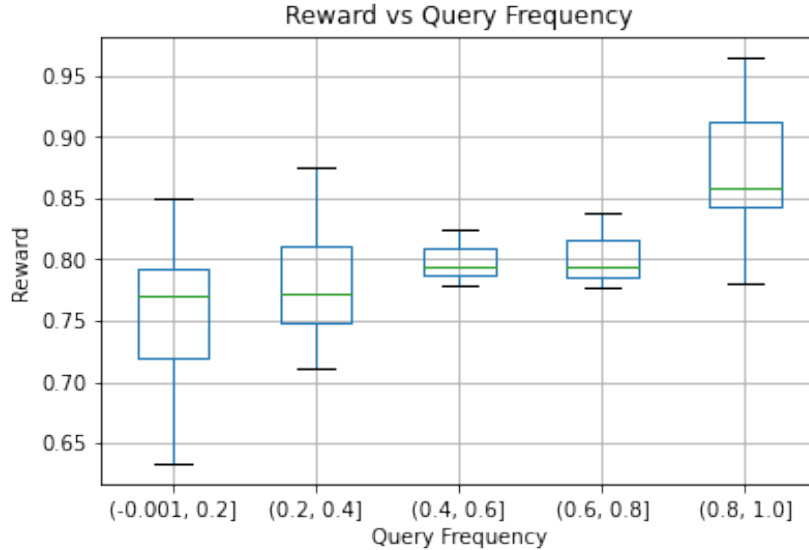


Figure 4: Personalization Effect. Labelers with higher decision reward is queried more often by Joint Collaboration with Personalization.

random from $[0, 0.4]$. After each iteration, we record the decisions allocated to each human expert and their corresponding reward (decision accuracy – individual cost). The results are shown in Figure 4 in which we show the box plot for average reward versus query frequency for 100 repetitions. This result shows that JCP effectively learns each human’s decision cost-effectiveness. When the human costs and decision performance yield higher expected reward, the proposed personalization variant, JCP, is more likely to query the corresponding human expert, thereby leveraging the expert’s cost-effectiveness. This result also shows that the cost can be used as a mechanism to prevent the algorithm from overburdening a single human with queries. By increasing the cost of querying the human with higher average decision reward, the algorithm can be incentivized to still make use of human with lower average decision reward, allocating them instances for which the loss incurred by querying them is lower.

5.3 Leveraging Expert Consistencies under Deterministic Actions

In this section we examine how the violation of the overlap assumption affects our proposed methods, and how our proposed extension of leveraging expert consistency can alleviate this problem. We evaluate the proposed method for considering expert consistency (EC), denoted as method-EC, which corresponds to the objective in Equation (11). We compare its performance against baselines, including human only, AO, TS and JC.

5.3.1 Synthetic Data

First, we construct a synthetic dataset to manually control the fraction of samples with deterministic actions in synthetic human decisions, and we assess how would it affect our system’s performance. This allows us to see the potential detrimental effect of deterministic actions on policy learning from observational data in a

Table 7: Rewards on Synthetic Data with Deterministic Actions. We vary the fraction of examples that humans will apply deterministic actions on. Due to the expert consistency assumption, human decision maker performance increases when the fraction of deterministic action increases, while it accompanies with a significant drop in algorithm’s and JC’s performance. Leveraging expert consistency can significantly help increase human-AI complementarity.

s	Human	AO	AO-EC	TS	TS-EC	JC	JC-EC
0.1	-1774.15±26.68	-664.60±30.74	-659.05±28.34	-2.20±305.55	425.75±341.88	0.80±309.39	423.50±342.28
0.2	-994.30±24.79	-699.10±50.14	-661.45±28.00	24.20±342.21	1724.90±128.06	131.45±401.70	1722.50±128.29
0.3	-243.25±29.10	-814.75±79.55	-656.50±30.41	-814.75±79.55	1765.55±178.78	-810.85±78.52	2116.85±236.28
0.4	495.95±41.03	-956.20±67.58	-643.60±32.07	-622.00±324.45	2159.30±200.83	-613.45±323.49	2752.25±253.44
0.5	1239.20±39.46	-1021.45±63.66	-672.10±30.90	-1021.45±63.66	2234.90±341.90	-1008.40±60.66	2416.85±398.56

controlled setting. The features are generated from a two-dimensional isotropic normal distribution $\mathcal{N}(0, I_2)$ in which I_2 is a two-dimensional identity matrix. We focus on the case of a binary action in which the human decision policy is decided as $P(a = 1|x) = \Phi(x_0)$, where $\Phi(\cdot)$ is the cumulative distribution function of a normal distribution. This corresponds to an imperfect decision maker which only uses partial information to infer decisions, where x_0 is the first dimension of the feature set. For samples in the top s -quantile of x_0 , we assume human decision makers apply deterministic actions, thus we can control how deterministic human decision makers are by controlling s (larger s leads to more deterministic human decisions). The potential rewards are defined as $r_0 = -0.5(\mathbb{1}(x_0x_1 > 0) \times 2 - 1)$ and $r_1 = 0.5(\mathbb{1}(x_0x_1 > 0) \times 2 - 1)$.

Following our assumption, we first examine the case in which the deterministic actions made by human experts are optimal. Specifically, we vary the fraction of examples in which experts make consistent decisions, ranging from 0.1 to 0.5. The results are shown in Table 7. We denote algorithms with expert consistency assumption as (algorithm)-EC. When the fraction of deterministic actions increases, the human performance also increases, since there are more optimal actions taken. As expected, the variants that do not leverage the information in the human consistent behavior, such as JC, fail to offer complementary performance in some cases, especially when the fraction of deterministic actions are high. By contrast, as the fraction of deterministic actions increases, the benefit of JC-EC increases.

However, in practice, consistent human experts may not be perfect, which results in some level of bias in the deterministic actions. To simulate such a scenario, we assume experts may make mistakes in deterministic actions with a uniform bias fraction α . We show the results for this setting in Table 8. With larger expert consistency bias, we observe a decline in algorithms with the EC assumption, while the AO baseline has a relatively stable performance. This is expected due to our theoretical analysis in Theorem 1. Surprisingly, we find that even with higher bias, JC-EC continues to offer complementary performance, while AO-EC no longer offers better performance compared to the AO baseline. In general, JC-EC is quite robust toward expert consistency bias in terms of complementarity but larger bias level will inevitably hurt the collaboration system’s performance.

Similarly, we vary the cost of human decision makers in Table 9 with overlap of 0.2 and training set size of 500 in which the complementary is relatively significant. With a higher human cost, the human team’s performance decreases rapidly and we observe the human-AI complementarity also decreases since there is less benefit of querying a human worker due to the increased cost.

In addition to our previous experiments with a stationary environment, we also vary the training set size and set the test set size to 10,000 to examine how the deferral system performs when the algorithm has access to a different number of samples. In some cases, violations of the overlap assumption may arise randomly, as a result of a small dataset. However, deterministic violations that result from humans never making certain decisions for certain types of instances may arise regardless of the dataset size. The results are shown in Table 10. With more data, we observe that the qualitative relationship between AO, TS, JC and their variants remains the same, while the EC variant consistently provides significant benefit over its counterparts. This is expected since the asymptotic bias introduced by the violation of the overlap assumption cannot be addressed away by adding more data in the training set.

5.3.2 Real Human Responses

In this section, we use two datasets with real human responses to examine whether humans exhibit deterministic responses and whether our algorithms leveraging expert consistency improve deferral collaboration performance

Table 8: Rewards on Synthetic Data with Deterministic Actions and Varying Expert Bias. We vary the bias level in deterministic actions from humans. With more consistent human bias, human decision performance decreases. While the deferral collaboration system’s performance also decreases (which is expected due to our theoretical analysis in Theorem 1), the human-AI complementarity is still possible, suggesting the deferral collaboration leveraging expert consistency may be robust to human bias in deterministic decisions.

Bias Level	Human	AO	AO-EC	TS	TS-EC	JC	JC-EC
0.0	-994.30±24.79	-699.10±50.14	-661.45±28.00	24.20±342.21	1724.90±128.06	131.45±401.70	1722.50±128.29
0.1	-1290.10±22.68	-698.50±50.30	-653.35±29.42	-201.25±295.64	1469.75±222.13	-200.50±295.87	1469.45±218.66
0.2	-1588.90±19.55	-682.60±45.43	-652.15±29.49	-164.80±319.19	1418.00±221.89	-162.25±320.23	1421.90±218.59
0.3	-1891.30±17.19	-673.60±46.74	-679.30±30.00	-184.15±300.12	1403.75±236.50	-184.75±301.91	1435.25±251.28
0.4	-2209.00±21.40	-675.70±37.39	-675.40±30.93	-187.00±298.85	881.90±339.14	-188.20±300.07	969.35±348.86
0.5	-2508.85±20.96	-675.85±37.32	-680.05±32.31	-185.95±299.46	605.60±345.76	-206.50±288.60	608.00±345.15

Table 9: Rewards on Synthetic Data with Deterministic Actions and Varying Expert Cost. We vary the cost of human decision makers. With larger human cost, the human decision maker performance decreases rapidly and the human-AI complementarity also decreases.

Cost	Human	AO	AO-EC	TS	TS-EC	JC	JC-EC
0.0	-994.30±24.79	-699.10±50.14	-661.45±28.00	24.20±342.21	1724.90±128.06	131.45±401.70	1722.50±128.29
0.05	-1494.30±24.79	-699.10±50.14	-661.45±28.00	-234.98±274.91	724.88±353.24	-234.24±275.07	726.04±352.91
0.1	-1994.30±78.38	-699.10±50.14	-661.45±28.00	-431.23±205.54	200.94±324.94	-213.15±270.22	330.73±367.94

Table 10: Rewards on Synthetic Data with Deterministic Actions. We vary the training set size from 300 to 2000. With deterministic actions, increased sample size has a limited effect on algorithm’s performance and the human-AI complementarity. By leveraging expert consistency, there is a significant improvement on human-AI complementarity.

Train	Human	AO	AO-EC	TS	TS-EC	JC	JC-EC
300	-994.30±24.79	-762.55±44.41	-695.95±42.68	-494.80±237.38	1575.65±240.39	-407.65±320.83	1573.40±239.39
500	-994.30±24.79	-699.10±50.14	-661.45±28.00	24.20±342.21	1724.90±128.06	131.45±401.70	1722.50±128.29
1000	-994.30±24.79	-766.45±91.76	-639.10±31.31	-515.80±233.36	1511.00±277.50	-521.80±228.82	1619.15±296.08
2000	-994.30±24.79	-644.50±38.67	-623.20±29.54	364.10±476.89	1902.50±205.20	369.95±479.31	2109.80±225.37

Table 11: Propensity Score and Average Reward on Hate Speech Recognition Dataset. The optimal reward is set as 0.5. Actions with more extreme propensity scores achieve significantly higher reward. The consistently deterministic responses achieve nearly optimal decision performance.

$\max\{\text{Propensity Score}, 1-\text{Propensity Score}\}$	Number of Samples	Average Reward
(0.99, 1]	463	0.4784
(0.95, 99]	1	0.5000
(0.9, 0.95]	55	0.4091
(0.8, 0.9]	350	0.3171
(0.7, 0.8]	215	0.1744
(0.6, 0.7]	200	0.0850
(0.5, 0.6]	132	0.0379

in such cases. Notably, this also means that we are evaluating all proposed algorithms in two additional datasets of real human responses.

The first dataset we use is a hate speech recognition dataset with 1,471 Twitter posts for the MTurk survey by Keswani et al. (2021). The text corpus is a subset of a previously collected public dataset (Davidson et al. 2017). Each post in the dataset was labeled by around 10 different annotators, with 170 workers in total. Thus, for each post we treat the worker population as a whole and randomly query one annotator’s response for a human decision whenever we need to obtain a human assessment. For further details of the MTurk experiment, see Keswani et al. (2021). Posts are labeled as positive if labelers believe they contain hate speech or offensive language. This setup refers to a content moderation setting in which the decision maker has to decide whether a post should be deleted before becoming public. We use the pre-trained GloVe embeddings (Pennington et al. 2014) to serve as text representation for our model, which results in a 100-dimensional vector. We use the labels from Davidson et al. (2017) as the gold labels for this dataset, following Keswani et al. (2021).

However, among all posts labeled as hate speech, only 26% of them are coded unanimously in Davidson et al. (2017), which suggests that the ground-truth labels are questionable. Therefore, we also conduct experiments on the dataset CIFAR-10H (Peterson et al. 2019) to examine the effectiveness of our proposed method in a dataset with more reliable labels. CIFAR-10H contains the 10,000 test images from the original CIFAR-10 dataset (Krizhevsky 2009) and 511,400 human labels collected via Amazon Mechanical Turk from 2,571 labelers. Every image has 51 annotations on average, ranging from 47 to 63. An attention check is conducted for every 20 trials. In the experiment, we filter workers if they fail attention checks more than half of the time. The ground-truth labels come from the original CIFAR-10 dataset, which is widely used in computer vision tasks and considered to include relatively high-quality ground-truth labels.

First, for the hate speech dataset, we examine whether humans demonstrate deterministic actions on some instances. We stratify human decisions by propensity scores, the results of which are shown in Table 11. Note that the optimal average reward is set as 0.5. Interestingly, more extreme propensities correspond to near optimal performance, which is consistent with the motivation of leveraging expert consistency in De-Arteaga et al. (2021) and justifies our assumption. When the propensity scores become less extreme, there is a strong decline in human decision maker performance. Such differences in human performance across samples may come from the inherent aleatoric uncertainty in the data, which cannot be explained away using more data.

For our experiments, we compare 50%, 60%, and 70% of the data to examine the potential difference of each method with respect to different data size, setting human cost as 0. The results are shown in Table 12. We report the total reward on the test set. Algorithms considering the deterministic actions of human decision makers significantly outperform baselines ignoring such behavior. Furthermore, collaboration between algorithm and human decision makers can still offer a significant performance improvement compared to either AO or Human Only baselines and conclusions are stable with increased data size.

Similarly, we examine deterministic human actions and their corresponding reward on the CIFAR-10H dataset, which we show in Table 13. The experimental results when applying the different proposed algorithms and variants to this dataset are shown in Table 14. We use ResNet-18 (He et al. 2016) as the underlying model for algorithmic policy and routing model, since the input features are more complex images. Even for such complex model, we observe complementary performance from TS and JC. Here we observe JC and TS have similar performance, which can happen when the algorithmic policy trained on observational data coincides with the

Table 12: Total Reward on Hate Speech Detection Data with Deterministic Actions. We vary the training set size. The human reward observed has a relatively large variance and with more data. The algorithms with expert consistency are significantly better than their counterparts without it.

Train Size	Human	AO	AO-EC	TS	TS-EC	JC	JC-EC
50%	36.40±43.56	92.40±10.92	152.00±12.84	92.40±10.92	152.00±12.84	85.20±14.51	146.00±12.66
60%	28.75±40.72	85.00±5.46	108.20±10.21	85.00±5.46	108.20±10.21	85.40±6.18	102.60±8.06
70%	17.90±43.02	76.40±7.64	104.00±6.15	76.40±7.64	104.00±6.15	76.40±8.75	102.80±5.48

Table 13: Propensity Score and Average Reward on CIFAR-10H Dataset. The optimal reward is set as 1. Actions with more extreme propensity scores have significant higher reward and the consistently deterministic responses have nearly optimal decision performance.

$\max\{\text{Propensity Score}, 1-\text{Propensity Score}\}$	Number of Samples	Average Reward
(0.99, 1]	4308	0.9995
(0.9, 99]	4413	0.9283
(0.8, 0.9]	589	0.7046
(0.7, 0.8]	258	0.4868
(0.6, 0.7]	149	0.2402

joint training solution. EC algorithms also offer a significant improvement over its counterparts, which shows the importance of the expert consistency assumption in real-world applications.

5.4 Non-Stationary Data under Covariate Shift

In this section, we examine how each method would be impacted by covariate shift. We evaluate the proposed method considering out-of-distribution data (OD) with joint collaboration, denoted as JC-OD, which corresponds to the objective in Equation (16). We evaluate its performance against baselines assuming a stationary distribution, comparing human only, AO, TS and JC.

5.4.1 Synthetic Data

To simulate non-stationary data with covariate shift, we follow a similar setup as in Section 5.3. The decision policy is defined as $P(a = 1|x) = \Phi(0.5 * x_0)$, where $\Phi(\cdot)$ is the cumulative distribution function of a normal distribution. The potential rewards are defined as $r_0 = x_0 + \epsilon_0$ and $r_1 = 2x_0 + x_1 + \epsilon_1$, where $\epsilon_0, \epsilon_1 \sim \mathcal{N}(0, 1)$, and the human cost is set to 0.1. Additionally, we assume training data is sampled from $x_0 \sim \mathcal{N}(0, 1), x_1 \sim \mathcal{N}(\mu, 1)$ and testing data is sampled from $x_0, x_1 \sim \mathcal{N}(0, 1)$. By controlling for μ , we are able to vary the degree of covariate shift in our testing data. We refer to the method using outlier detection module according to the process in Figure 3 as JC-OD.

The results with synthetic data are shown in Table 15. When the covariate shift becomes more severe (μ is larger), the algorithm’s performance on the test data becomes significantly worse, which shows the algorithm’s poor generalizability. The proposed JC-OD performs the best among all methods, which shows the benefit of explicitly considering out-of-distribution samples. This benefit becomes more significant when the covariate shift is larger, and thus relying on decisions made by the algorithm is less reliable.

Table 14: Rewards on CIFAR-10H Dataset with Deterministic Actions. We vary the training set size and report the total reward on test set. The algorithms considering expert consistency are significantly better than their counterparts without it.

Train Size	Human	AO	AO-EC	TS	TS-EC	JC	JC-EC
70%	2739.20±23.34	2779.60±4.57	2808.40±7.53	2779.20±3.98	2809.60±7.30	2779.20±3.98	2818.79±3.64
80%	1790.00±9.63	1828.00±4.17	1851.20±5.50	1830.80±3.86	1850.00±5.79	1830.80±3.86	1858.40±6.62
90%	904.80±13.54	916.40±4.40	924.40±3.26	916.80±4.62	925.20±3.42	916.80±4.62	933.60±4.36

Table 15: Rewards on Simulation Data with Covariate Shift. We vary the level of covariate shift by adjusting μ . Due to the assumption, human performance is stable while algorithmic performance is impacted with large μ . We find human-AI complementarity is possible for deferral collaboration under covariate shift and our out-of-distribution extension can further improve deferral system’s performance with severe shifting.

	Human	AO	TS	JC	JC-OD
$\mu = 1$	1820.83±229.41	5090.11±136.50	5166.47±104.17	5321.83±102.18	5253.00±111.96
$\mu = 3$	1820.83±229.41	3489.32±216.65	3489.32±216.65	3740.62±314.49	3774.19±214.43
$\mu = 5$	1820.83±229.41	2987.95±191.52	2987.95±191.52	3089.54±206.99	3184.60±135.12
$\mu = 7$	1820.83±229.41	2561.03±219.45	2561.03±219.45	2650.99±206.91	2949.96±99.35
$\mu = 9$	1820.83±229.41	2555.47±250.51	2555.47±250.51	2654.25±235.88	2978.24±89.94

5.4.2 Semi-Synthetic Data

To simulate covariate shift in the real world, we construct a semi-synthetic experiment using two image datasets, MNIST (LeCun et al. 1998) and SVHN (Netzer et al. 2011). These are widely used in domain adaptation tasks (Han et al. 2019, Zou et al. 2019) to simulate two domains with covariate shift. MNIST consists of 70000 handwritten digits (60,000 for training and 10,000 for testing), while SVHN consists of more than 90,000 digits from street view images(73,257 for training, 26,032 for testing). The sample images are shown in Figure 5 in the Appendix. Here, we use the same approach used in previous experiments to convert the classification task into a policy learning task. Given that the data sources are simple images, it is reasonable to assume human decision makers may have the same performance on both sources. Thus, we use our human decision noise behavior model (See Appendix A) with human decision makers whose decision accuracy is randomly drawn from $U[0.4, 0.8]$ in each run.

Both datasets are used for digit recognition. Since they come from different sources, the data distribution might be different and contain samples that are OOD for the other dataset. A policy induced from MNIST data might not generalize well to SVHN, regardless of the training set size, due to covariate shift. We experiment with two settings: training on SVHN and testing on MNIST ($S \rightarrow M$), and MNIST to SVHN ($M \rightarrow S$).

The results are shown in Table 16. Since SVHN has more complex digit images from street views than MNIST, the generalization performance of algorithm on $S \rightarrow M$ is better. Interestingly, we find that JC and JC-OD have similar performance improvement on $S \rightarrow M$, which is consistent with our previous finding that JC seems to have some level of robustness against covariate shift, while JC-OD can further improve decision performance over JC.

When learning on M and attempting to generalize to S ($M \rightarrow S$), it seems to be more challenging for algorithms to adapt knowledge extracted from clean, simple images to more sophisticated domains, resulting in suboptimal performance for the AO baseline. In this setup, JC is also unable to match the human performance, since the test environment is quite different from training. Meanwhile, JC-OD shows significantly better performance by matching the performance of the human only team through cross validation (JC-OD chooses to route all instances to humans in this case). This demonstrates the robustness of the proposed method.

5.4.3 Real Human Responses

We use the same text analysis dataset in Rzhetsky et al. (2009) (that we also used in Section 5.1) as our dataset with real human responses to examine the benefit of our proposed systems. To simulate covariate shift, we follow Kato et al. (2020) and split data into training set and testing set using probability $P(\text{train}|x) = 1/2(1 + \exp(-\gamma(x) + \epsilon))$, where $\gamma(x) = \sum_{i=1}^{200} x_i$ and $\epsilon \sim \mathcal{N}(0, 1)$. We set a validation set with size 150 to tune the OOD parameter.

The results are shown in Table 17. With covariate shift, we can no longer consistently observe the complementarity of TS and JC, while JC-OD still successfully detects that the human has a much better performance compared to the algorithm and allocates all instances to human experts by tuning the OOD parameter. Due to the covariate shift, JC performs suboptimally and cannot route samples effectively. The results are consistent with varying training set sizes.

Table 16: Rewards on Semi-Synthetic Data with Covariate Shift. We vary the human cost. $S \rightarrow M$ indicates learning algorithm on SVHN and test methods on MNIST, which is an easier task compared to $M \rightarrow S$ since images in SVHN is more complex. The algorithm’s performance is worse than the human’s for $S \rightarrow M$ when $C = 0.05$ and OD extension can further increase human-AI complementarity in the deferral collaboration system. For $M \rightarrow S$, due to the challenging task, the algorithm has a significantly worse performance while methods considering OD can still achieve robust performance.

$S \rightarrow M$	Human	AO	TS	JC	JC-OD
$C = 0.05$	34835.32±1855.82	30098.00±928.34	35070.06±739.53	35108.74±794.08	35120.22±785.03
$C = 0.1$	31985.26±1855.77	30098.00±928.34	31762.88±871.42	32165.10±269.29	32188.52±279.41
$C = 0.15$	30157.93±1915.97	30098.00±928.34	31668.61±1169.51	32998.87±894.16	33022.16±892.68
$M \rightarrow S$	Human	AO	TS	JC	JC-OD
$C = 0.05$	43066.44±986.51	6083.20±764.79	19806.89±8020.84	19912.49±7982.95	43066.44±986.51
$C = 0.1$	39553.28±986.48	6083.20±764.79	15852.96±6015.54	18279.41±7082.87	39399.20±994.83
$C = 0.15$	36040.84±986.50	6083.20±764.79	11671.13±4906.86	11632.13±4895.47	36040.84±986.50

Table 17: Rewards on Real Data with Covariate Shift. We vary the size of the training set size and report the total reward. Due to the covariate shift, algorithm solution performs worse than human significantly and JC-OD can still match humans’ decision performance.

	Human	AO	TS	JC	JC-OD
train ratio = 0.3	425.00±2.80	346.80±2.80	358.72±2.80	326.72±2.80	425.00±2.80
train ratio = 0.5	273.60±3.33	233.40±3.33	240.48±3.33	251.86±3.33	273.60±3.33
train ratio = 0.7	117.60±1.00	113.00±1.00	112.98±1.00	110.50±1.00	117.70±1.00

5.5 Hybrid System Improvement

In this section, we conduct ablation studies to better understand the limitations of the proposed approach and understand how model selection affects the human-AI team performance. We use our real dataset MLC from above to better understand when human-AI collaboration will achieve a greater improvement. Intuitively, if the algorithm alone can achieve perfect performance, a human-AI team will not improve performance with respect to the optimized objective. We fix the routing model’s architecture as before and set the policy model to a different number of neurons to represent different model capacities. By the universal approximation theorem, with a sufficiently large non-linear layer, we can approximate any continuous function f (Bengio et al. 2017). We run each experiment for 10 runs, and consider the number of neurons in a small range to avoid potential overfitting, since the dataset is not difficult to learn. The average rewards across algorithms are reported in Table 18. We also consider a `ReLU` activation after the linear layer as an alternative to increase model capacity. When the model has limited capacity (2 units), we observe a greater benefit of using human-AI team. When the model has a stronger capacity, we observe a stronger performance in the AO baseline, and the joint optimization and hybrid team’s improvement drops. This finding confirms our intuition that the human-AI collaboration helps to a lesser degree when the model has a strong capacity and no overfitting is observed.

Table 18: Rewards on different model capacity on MLC. Human-AI complementarity is more significant when the model class is simpler.

# Hidden Units	AO	TS	JC
2	66.5±1.1	77.1±0.7	79.2±1.0
2 (w/ ReLU)	66.9±1.2	75.9±1.6	77.7±1.3
8	77.7±1.0	86.3±0.7	86.4±0.9

6 Conclusions, Managerial Implications, and Future Work

Decision-making under uncertainty is a core function of management, and advances in machine learning have presented new opportunities to bring data-driven algorithms to bear to autonomously undertake decision and improve decision outcomes. However, in contexts in which humans and algorithms can undertake decisions autonomously, research has also established that humans and algorithmic-based decision-making can often exhibit complementarity such that neither the human nor an algorithm offers dominant performance across all decision instances. A key path to leverage such HAI complementarity is to effectively defer decisions to either the algorithm or the human, based on the expected outcomes. Recent research has shown how leveraging varying *human* abilities by deferring tasks to suitable individuals can meaningfully improve outcomes (Wang et al. 2019a). However, productive deferral collaboration between humans and AI, including specifically for course of action decisions, introduce challenges that we both highlight and aim to address in this work.

Our work aims to advance the state of the art by proposing and evaluating effective and reliable HAI deferral collaboration methods for course of action decisions in which historical data reflect only the consequences of past choices, and where the goal is to maximize the total reward from the actions selected. To our knowledge, ours is the first work that proposes and extensively evaluates methods to address this challenge. We then propose a personalization variant of our approach to further leverage differential human team members' decision-making abilities. Lastly, we propose two variants of our approach that aim to yield robust performance of the human-AI deferral collaboration in contexts in which human experts select advantageous actions deterministically, and in the presence of distribution shift between the training and deployment covariate distributions. These adaptations are important both because they improve the robustness of the system and because such robustness is key to the trustworthiness of the human-AI collaboration, which is crucial to its adoption and impact in practice. Overall, the methods we propose lay the groundwork for deferral human-AI collaboration for course of action decisions that future work can build on, and we propose a method that can be directly applied across business and societal contexts. Our work is timely and potentially impactful both because of advances in policy learning from observational data that offer opportunities to learn highly effective algorithmic policies, and because of the rich set of business and societal course of action decision contexts that can meaningfully benefit from these advances in the near term.

Our work also offers a comprehensive *empirical* evaluation framework that brings to bear practices from diverse applied machine learning research fields and that is instrumental for using real data to evaluate and compare methods for HAI deferral collaboration for choice of action decisions. Our evaluations use data with real human choices and principled approaches to simulate such choices, along with the means to vary key properties of our context to assess the robustness of proposed solutions in tackling them. This includes the means to vary the available human-AI complementarity that can be leveraged by competing methods, and by producing out-of-sample decisions instances during deployment to assess the trustworthiness of a alternative HAI collaboration systems. We hope that future work can adopt and build on our evaluations to allow for meaningful comparisons of alternative solutions. Finally, to our knowledge, our work is also the first to offer comprehensive empirical results that establish state-of-the-art performance on publicly available data sets of deferral human-AI collaboration method for course of action decisions. Our results establish the benchmark performance that future advances on HAI deferral collaborations systems can be compared against³.

Our results establish that the proposed LCP-HAI framework reliably produces effective human-AI deferral collaboration and that it can both learn effectively and leverage human-AI complementary abilities to yield superior decision rewards across different decision tasks. We find that our LCP-HAI offers advantageous performance, often yielding superior and, otherwise, comparable rewards than can be achieved by either the human or the algorithm on their own. Furthermore, LCP-HAI yields reliable performance across settings, including different data domains from which complementary policies and routing decisions must be learned, and across different human abilities and costs. Overall, our results demonstrate that our LCP-HAI approach and the principles it is based on offer foundations on which future human-AI deferral collaboration methods for course of action decisions can build upon.

Management research has proposed that in contexts in which tasks can be assigned to multiple entities with diverse abilities, deferring tasks to different entities creates an opportunity to exploit their inherent complementarity and thereby produce superior outcomes (Wang et al. 2019a). However, furthermore, our context of a human-AI deferral collaboration introduces additional opportunities and challenges. These include the oppor-

³The code for our proposed methods and evaluation procedures will be made available for replication and future work.

tunity to develop an algorithmic decision-maker that best complements human team members, and overcoming challenges of learning both complementary algorithmic policy and a router model from data that reflects limited outcomes of only past choices. The methods we propose are general-purpose, and can be directly applied to improve course of action decisions across managerial and societal contexts in which it is plausible for both humans and algorithms to undertake course of action decisions autonomously.

Similar to other research that develops methods to address novel challenges which have not been extensively studied, the problem we formulate and framework we develop offer abundant opportunities for future work to build on that are outside the scope of our study. In particular, our work and the proposed deferral human-AI collaboration system give rise to several interesting future research directions to address idiosyncratic challenges and opportunities arising in certain important contexts.

Specifically, we considered settings in which the historical data is generated by human decision makers. Future work can consider learning from data produced by deferral human-AI collaboration system in which instances were assigned to either the human or the AI the historical data is generated by a deferral system in which part of the instances are solved by the algorithm and the rest are taken by human decision makers. This brings another complexity since the data used for learning human behaviors and algorithmic policy follows different distributions, then the data distribution used for learning the deferral system is not aligned with the population distribution anymore and a more careful reweighting will be needed.

Another interesting future work is inspired by settings where decisions are undertaken by a group and are based on the group’s deliberation, such as panel discussions, where multiple decision-makers arrive at a consensus to produce the final decision. Hybrid human-AI decision policies may also become increasingly common. Learning from panel discussions is especially challenging due to the combinatorial number of possible human decision maker teams, which makes group decision rewards difficult to estimate.

Our work considers the task of learning to complement humans, where the decision makers who produced the observational data can also be queried again in the future. However, in some contexts, new decision makers may join a team while others may leave. It would be valuable to study how to adaptively revise the algorithmic policy and router to best complement the humans in such dynamic settings and solve the cold start problem of new human decision makers (Gong et al. 2019).

We focus on course of action decisions in the context of human-AI deferral collaboration, but our approach can also inform future work on contexts in which course of action decisions must always be undertaken by a final, human decision maker. Recent work has proposed methods for learning to advise a final human decision-maker (Wolczynski et al. 2022), with the aim of complementing the human decision maker by selectively recommending decisions to the human. It would be valuable to build on our work to learn to advise a human on course of action decisions, with the goal of achieving higher reward than can be achieved by the human without the AI’s advice.

Finally, our work considers a common objective in practice for the human-AI deferral collaboration, which is to achieve higher expected reward, such as higher profits. Future work can consider settings in which other objectives may be more desirable to optimize.

The potential value of human-AI collaboration to improve course of action decisions is tangible, and such advances are timely. We hope that our work will inspire more research to enhance the impact on practice of human-AI deferral collaboration and to benefit business and society at large.

References

- Achab M, Cl  men  on S, Garivier A (2018) Profitable bandits. *Asian Conference on Machine Learning*, 694–709 (PMLR).
- Angrist JD, Imbens GW, Rubin DB (1996) Identification of causal effects using instrumental variables. *JASA* 91(434):444–455.
- Athey S, Wager S (2017) Efficient policy learning. Technical report, Stanford University, Graduate School of Business.
- Bahat Y, Shakhnarovich G (2018) Confidence from invariance to image transformations. *arXiv preprint arXiv:1804.00657* .
- Ban GY, Keskin NB (2021) Personalized dynamic pricing with machine learning: High-dimensional features and heterogeneous elasticity. *Management Science* 67(9):5549–5568.
- Bansal G, Nushi B, Kamar E, Horvitz E, Weld DS (2020) Optimizing ai for teamwork. *arXiv:2004.13102* .
- Bansal G, Nushi B, Kamar E, Weld D, Lasecki W, Horvitz E (2019) A case for backward compatibility for human-ai teams. *arXiv preprint arXiv:1906.01148* .
- Bengio Y, Goodfellow I, Courville A (2017) *Deep learning*, volume 1 (MIT press Massachusetts, USA:).

- Bertsimas D, Kallus N (2016) The power and limits of predictive approaches to observational-data-driven optimization. *arXiv*
- Bertsimas D, Kallus N (2020) From predictive to prescriptive analytics. *Management Science* 66(3):1025–1044.
- Boutell MR, Luo J, Shen X, Brown CM (2004) Learning multi-label scene classification. *Pattern recognition* 37(9):1757–1771.
- Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T (2018) The rise of deep learning in drug discovery. *Drug discovery today* 23(6):1241–1250.
- Davani AM, Díaz M, Prabhakaran V (2022) Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics* 10:92–110.
- Davidson T, Warmusley D, Macy M, Weber I (2017) Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.
- De A, Koley P, Ganguly N, Gomez-Rodriguez M (2020) Regression under human assistance. *AAAI*, 2611–2620.
- De-Arteaga M, Dubrawski A, Chouldechova A (2021) Leveraging expert consistency to improve algorithmic decision support. *arXiv preprint arXiv:2101.09648*.
- Dudík M, Erhan D, Langford J, Li L, et al. (2014) Doubly robust policy evaluation and optimization. *Statistical Science* 29(4):485–511.
- Elisseeff A, Weston J (2002) A kernel method for multi-labelled classification. *NeurIPS*, 681–687.
- Elmachtoub AN, Gupta V, Hamilton ML (2021) The value of personalized pricing. *Management Science* 67(10):6055–6070.
- Ensign D, Friedler SA, Neville S, Scheidegger C, Venkatasubramanian S (2018) Decision making with limited feedback: Error bounds for predictive policing and recidivism prediction. *ALT*, volume 83.
- Fauray L, Tanielian U, Dohmatob E, Smirnova E, Vasile F (2020) Distributionally robust counterfactual risk minimization. *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 3850–3857.
- Fazelpour S, De-Arteaga M (2022) Diversity in sociotechnical machine learning systems. *Big Data & Society* 9(1):20539517221082027.
- Fujimoto S, Meger D, Precup D (2019) Off-policy deep reinforcement learning without exploration. *International conference on machine learning*, 2052–2062 (PMLR).
- Gao R, Biggs M, Sun W, Han L (2021) Enhancing counterfactual classification via self-training. *arXiv preprint arXiv:2112.04461*.
- Gao R, Saar-Tsechansky M (2020) Cost-accuracy aware adaptive labeling for active learning. *AAAI*, volume 34, 2569–2576.
- Geirhos R, Temme CRM, Rauber J, Schütt HH, Bethge M, Wichmann FA (2018) Generalisation in humans and deep neural networks. *arXiv preprint arXiv:1808.08750*.
- Gong W, Tschjatschek S, Nowozin S, Turner RE, Hernández-Lobato JM, Zhang C (2019) Icebreaker: Element-wise efficient information acquisition with a bayesian deep latent gaussian model. *Advances in neural information processing systems* 32.
- Han L, Zou Y, Gao R, Wang L, Metaxas D (2019) Unsupervised domain adaptation via calibrating uncertainties. *CVPR Workshops*, volume 9.
- Hastie T, Tibshirani R, Friedman JH, Friedman JH (2009) *The elements of statistical learning: data mining, inference, and prediction*, volume 2 (Springer).
- He K, Zhang X, Ren S, Sun J (2015) Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *Proceedings of the IEEE international conference on computer vision*, 1026–1034.
- He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. *CVPR*, 770–778.
- Hirano K, Imbens GW, Ridder G (2003) Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71(4):1161–1189.
- Huang SJ, Chen JL, Mu X, Zhou ZH (2017) Cost-effective active learning from diverse labelers. *IJCAI*.
- Huber LS, Geirhos R, Wichmann FA (2021) A four-year-old can outperform resnet-50: Out-of-distribution robustness may not require large-scale experience. *SVRHM 2021 Workshop@ NeurIPS*.
- Ibrahim R, Kim SH, Tong J (2021) Eliciting human judgment for prediction algorithms. *Management Science* 67(4):2314–2325.
- Joachims T, Swaminathan A, de Rijke M (2018) Deep learning with logged bandit feedback. *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. *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, Zhou A (2018) Confounding-robust policy improvement. *NeurIPS*, 9269–9279.
- Karlinsky-Shichor Y, Netzer O (2019) Automating the b2b salesperson pricing decisions: Can machines replace humans and when. Available at SSRN:3368402.
- Kato M, Uehara M, Yasui S (2020) Off-policy evaluation and learning for external validity under a covariate shift. *arXiv preprint arXiv:2002.11642* .
- Keswani V, Lease M, Kenthapadi K (2021) Towards unbiased and accurate deferral to multiple experts. *arXiv preprint arXiv:2102.13004* .
- Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. *arXiv* .
- Krizhevsky A (2009) Learning multiple layers of features from tiny images. Technical report, Citeseer.
- Künzel SR, Sekhon JS, Bickel PJ, Yu B (2019) Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences* 116(10):4156–4165.
- Lai V, Carton S, Bhatnagar R, Liao QV, Zhang Y, Tan C (2022) Human-ai collaboration via conditional delegation: A case study of content moderation. *CHI Conference on Human Factors in Computing Systems*, 1–18.
- Langford J, Strehl A, Wortman J (2008) Exploration scavenging. *Proceedings of the 25th international conference on Machine learning*, 528–535.
- Lawrence C, Sokolov A, Riezler S (2017) Counterfactual learning from bandit feedback under deterministic logging: A case study in statistical machine translation. *arXiv preprint arXiv:1707.09118* .
- LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86(11):2278–2324.
- Lee K, Lee K, Lee H, Shin J (2018) A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems* 31.
- Li KL, Huang HK, Tian SF, Xu W (2003) Improving one-class svm for anomaly detection. *Proceedings of the 2003 international conference on machine learning and cybernetics (IEEE Cat. No. 03EX693)*, volume 5, 3077–3081 (IEEE).
- Li SY, Jiang Y, Chawla NV, Zhou ZH (2018) Multi-label learning from crowds. *TKDE* 31(7):1369–1382.
- Madras D, Pitassi T, Zemel R (2018) Predict responsibly: improving fairness and accuracy by learning to defer. *NeurIPS* 31:6147–6157.
- Menon AK, Williamson RC (2018) The cost of fairness in binary classification. *Conference on Fairness, Accountability and Transparency*, 107–118 (PMLR).
- Mozannar H, Sontag D (2020) Consistent estimators for learning to defer to an expert. *arXiv:2006.01862* .
- Netzer Y, Wang T, Coates A, Bissacco A, Wu B, Ng AY (2011) Reading digits in natural images with unsupervised feature learning .
- Nguyen A, Wallace B, Lease M (2015) Combining crowd and expert labels using decision theoretic active learning. *HCOMP*.
- Oberdiek P, Rottmann M, Gottschalk H (2018) Classification uncertainty of deep neural networks based on gradient information. *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*, 113–125 (Springer).
- Pacchiano A, Phan M, Abbasi-Yadkori Y, Rao A, Zimmert J, Lattimore T, Szepesvari C (2020) Model selection in contextual stochastic bandit problems. *arXiv:2003.01704* .
- Pang G, Cao L, Aggarwal C (2021) Deep learning for anomaly detection: Challenges, methods, and opportunities. *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, 1127–1130.
- Pearl J (2010) Brief report: On the consistency rule in causal inference:” axiom, definition, assumption, or theorem?”. *Epidemiology* 872–875.
- Pearl J (2017) Detecting latent heterogeneity. *Sociological Methods & Research* 46(3):370–389.
- Pennington J, Socher R, Manning CD (2014) Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.
- Peterson JC, Battleday RM, Griffiths TL, Russakovsky O (2019) Human uncertainty makes classification more robust. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9617–9626.
- Raghu M, Blumer K, Corrado G, Kleinberg J, Obermeyer Z, Mullainathan S (2019) The algorithmic automation problem: Prediction, triage, and human effort. *arXiv:1903.12220* .
- Rosenbaum PR (1987) Model-based direct adjustment. *JASA* 82(398):387–394.
- Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1):41–55.

- Rubin DB (2005) Causal inference using potential outcomes: Design, modeling, decisions. *JASA* 100(469):322–331.
- Ruff L, Vandermeulen R, Goernitz N, Deecke L, Siddiqui SA, Binder A, Müller E, Kloft M (2018) Deep one-class classification. *International conference on machine learning*, 4393–4402 (PMLR).
- Rzhetsky A, Shatkay H, Wilbur WJ (2009) How to get the most out of your curation effort. *PLoS computational biology* 5(5):e1000391.
- Sachdeva N, Su Y, Joachims T (2020) Off-policy bandits with deficient support. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 965–975.
- Sachdeva PS, Barreto R, von Vacano C, Kennedy CJ (2022) Assessing annotator identity sensitivity via item response theory: A case study in a hate speech corpus. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 1585–1603.
- Schölkopf B, Williamson RC, Smola A, Shawe-Taylor J, Platt J (1999) Support vector method for novelty detection. *Advances in neural information processing systems* 12.
- Shalit U, Johansson FD, Sontag D (2017) Estimating individual treatment effect: generalization bounds and algorithms. *ICML*, 3076–3085 (PMLR).
- Si N, Zhang F, Zhou Z, Blanchet J (2020) Distributionally robust policy evaluation and learning in offline contextual bandits. *International Conference on Machine Learning*, 8884–8894 (PMLR).
- Sondhi A, Arbour D, Dimmery D (2020) Balanced off-policy evaluation in general action spaces. *International Conference on Artificial Intelligence and Statistics*, 2413–2423 (PMLR).
- Swaminathan A, Joachims T (2015a) Counterfactual risk minimization: Learning from logged bandit feedback. *ICML*, 814–823.
- Swaminathan A, Joachims T (2015b) The self-normalized estimator for counterfactual learning. *NeurIPS*.
- Swersky L, Marques HO, Sander J, Campello RJ, Zimek A (2016) On the evaluation of outlier detection and one-class classification methods. *2016 IEEE international conference on data science and advanced analytics (DSAA)*, 1–10 (IEEE).
- Tanno R, Saeedi A, Sankaranarayanan S, Alexander DC, Silberman N (2019) Learning from noisy labels by regularized estimation of annotator confusion. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 11244–11253.
- Torabi F, Warnell G, Stone P (2018) Behavioral cloning from observation. *arXiv preprint arXiv:1805.01954* .
- Wang G, Li J, Hopp WJ, Fazzalari FL, Bolling SF (2019a) Using patient-specific quality information to unlock hidden healthcare capabilities. *Manufacturing & Service Operations Management* 21(3):582–601.
- Wang L, Bai Y, Bhalla A, Joachims T (2019b) Batch learning from bandit feedback through bias corrected reward imputation.
- Wang T, Saar-Tsechansky M (2020) Augmented fairness: An interpretable model augmenting decision-makers’ fairness. *arXiv:2011.08398* .
- Wilder B, Horvitz E, Kamar E (2020) Learning to complement humans. *arXiv* .
- Wolczynski N, Saar-Tsechansky M, Wang T (2022) Learning to advise humans by leveraging algorithm discretion. *arXiv:2210.12849* .
- Yan Y, Rosales R, Fung G, Dy JG (2011) Active learning from crowds. *ICML*, volume 11, 1161–1168.
- Zheng Y, Li G, Li Y, Shan C, Cheng R (2017) Truth inference in crowdsourcing: Is the problem solved? *Proceedings of the VLDB Endowment* 10(5):541–552.
- Zou Y, Yu Z, Liu X, Kumar B, Wang J (2019) Confidence regularized self-training. *Proceedings of the IEEE International Conference on Computer Vision*, 5982–5991.

A Implementation Details of Human Behavior Models in Synthetic Dataset

Black-Box Human Behavior Model To create a black-box human behavior model, we use a random forest classifier on a random 30% subset of the data with full ground-truth labels. After the model is trained, at both training (data generating) and testing time, our synthetic expert will make a random action based on the output probability of the classifier. Since our task is a multi-label classification, the output probability is multiple predictive probabilities of each class which do not sum to 1. In order to make a choice, we use softmax with temperature to normalize the probabilities across classes, the temperatures are set to 10 and 20 for Scene and TMC respectively. When the number of actions increases, the probability output by softmax decreases and a larger temperature can ensure high confidence actions are still selected with relatively high probabilities. Here, high-confidence refers to the original output probability of the black-box model.

Decision Noise Human Behavior Model For decision noise human behavior models, motivated by learning from noisy labels under classification setup Tanno et al. (2019), we assume each expert will follow a uniform decision accuracy parameterized by ρ . Here we give an example in Table 19. Assume a customer specialist has $\rho = 0.6$, and needs to offer compensation plan to two customers, Customer 1 and Customer 2, with optimal compensation plan A and B respectively. Table 19 shows the decision probability for this decision maker under ρ . In such a setting, the decision maker is equally likely to make wrong decisions and have a decision accuracy of 0.6. Similarly, if there are multiple optimal actions, ρ is the summation of the decision accuracy of these actions and they all have the same decision accuracy.

In our experiments, we simulate three expert decision-makers and set ρ to be 0.6, 0.7, 0.8 respectively to see whether our personalization objective can help when experts have diverse skills Huang et al. (2017).

Optimal Action	A	B	C
Customer 1 - A	0.6	0.2	0.2
Customer 2 - B	0.2	0.6	0.2

Table 19: Decision Accuracy Example

B Dataset Statistics

	# Features	# Labels	Dataset Size
Scene	294	6	2407
TMC	30438	22	28596
MLC	1248	6	695
Focus	292	2	1000
Hate Speech	100	2	1471
CIFAR10-H	32x32	10	10000
MNIST	32x32	10	70000
SVHN	32x32	10	99289

Table 20: Dataset statistics.

C Significance Test

We report two sample t-tests for our experiments reported in the main paper. The significance level is set to 0.05 and the null hypothesis is that the difference between two methods is 0.

Table 21: Significance Tests on Focus dataset with Cost of 0.05. True means the averages are significantly different with confidence level of 0.05. For all settings, JC has competitive performance against all baselines, demonstrating the possibility of human-AI complementarity in LCP-HAI.

Data	Human/AO	Human/TS	Human/JC	AO/TS	AO/JC	TS/JC
Focus	False	False	False	False	True	True

Table 22: Significance Tests on Focus dataset with different expert costs. We examine the effect of human cost and set it from 0 to 0.5. True means the averages are significantly different with confidence level of 0.05. For all settings, JC has competitive performance against all baselines. When the human costs are too high, the human-AI complementarity decreases and the deferral system chooses to only use algorithm decisions.

Data (cost)	Human/AO	Human/TS	Human/JC	AO/TS	AO/JC	TS/JC
Focus (0)	True	True	False	False	True	True
Focus (0.05)	False	False	False	False	True	True
Focus (0.1)	True	True	True	False	True	True
Focus (0.3)	True	True	True	False	False	False
Focus (0.5)	True	True	True	False	False	False

C.1 Main Result and Additional Experiments

The significance test results for our experiments are reported in Table 1, Table 2 on the Focus dataset are shown in Table 21 and Table 22 respectively. When the human cost is low, the hybrid system demonstrates a significant improvement over the AO baseline. As the human cost increases, the benefit of hybrid team performance over the human-only baseline becomes increasingly significant, and the difference between the hybrid team and the AO baseline becomes insignificant.

The significance tests for the additional results reported in Table 3 are shown in Table 23. For joint collaboration, it always has a significant benefit over the human team and often has a significant benefit in reward over AO and TS.

C.2 Personalization

Similarly, the significant tests of the synthetic human responses and MLC in Table 3 and Table 5 are reported in Table 23 and Table 24 respectively.

C.3 Leveraging Expert Consistencies under Deterministic Actions

For simplicity, here we only conduct significance test for the EC variant and its corresponding algorithms here. The results for Table 7, Table 8, Table 9 and Table 10 are shown in Table 26, Table 27, Table 28 and Table 29 respectively.

Table 23: Significance Tests for different Human Behavior Models and MLC. True means the averages are significantly different with confidence level of 0.05. Model refers to the Black Box human behavior model and Noise refers to the uniform human behavior model. LCP-HAI with joint collaboration is superior to all other alternatives.

Data (HBM)	Human/AO	Human/TS	Human/JC	AO/TS	AO/JC	TS/JC
Scene (Model)	True	True	True	False	True	True
Scene (Noise)	True	True	True	False	True	True
TMC (Model) Two Stage	True	True	True	False	False	False
TMC (Noise)	True	True	True	False	False	False
MLC	True	True	True	True	True	False

Table 24: Significance Tests on different datasets for different Human Behavior Models. True means the averages are significantly different with confidence level of 0.05. Model means the Black Box human behavior model and Noise represents the uniform human behavior model. Personalization is never worse than JC and sometimes significantly outperforms JC.

Method	Scene (Model)	Scene (Noise)	TMC (Model)	TMC (Noise)
JC / JCP	False	False	False	True

Table 25: Significance Tests on Focus datasets with different expert costs. True means the averages are significantly different with confidence level of 0.05. We examine the effect of the human cost and set it from 0 to 0.5. For all settings, JCP has competitive performance against JC, which is the best method over other baselines. When the human costs are too high, the human-AI complementarity decreases and the deferral system chooses to only use algorithm decisions.

Method	$C = 0$	$C = 0.05$	$C = 0.1$	$C = 0.3$	$C = 0.5$
JC / JCP	True	True	True	False	False

The qualitative conclusions drawn from the main paper are further collaborated with the significance tests. We can observe the significance results do not change with the increasing bias level, a higher human cost leads to less significant human-AI complementarity, and our conclusions are stable with varying training set size.

C.4 Non-Stationary Data under Covariate Shift

The significance tests for our experiments in Table 15, Table 16 and Table 17 are reported in Table 30, Table 31 and Table 32 respectively. JC-OD has a significant benefit over other baselines with challenging covariate shift such as training on MNIST and testing on SVHN and the covariate shift happened on Focus dataset, which further validates our conclusions in the main paper.

D Samples from Semi-Synthetic Experiment with Covariate Shifting

Samples from SVHN and MNIST are shown in Figure 5. SVHN has digit images which is more realistic in life while MNIST has more regular digit images. It is reasonable to assume human decision maker can identify digits in both datasets with similar accuracy while machine learning trained on one dataset might be hard to generalize to the other (Zou et al. 2019, Han et al. 2019).

Table 26: Significance Tests on Synthetic Data with Deterministic Actions. True means the averages are significantly different with confidence level of 0.05. We vary the fraction of examples that humans will apply deterministic actions on. Due to the expert consistency assumption, human decision makers’ performance increases when the fraction of deterministic action increases, while it accompanies by a significant drop in algorithm’s and JC’s performance. Leveraging expert consistency can significantly help increase human-AI complementarity.

Ovp	AO/AO-EC	TS/TS-EC	JC/JC-EC
0.1	False	False	False
0.2	False	True	True
0.3	False	True	True
0.4	True	True	True
0.5	True	True	True

Table 27: Significance Tests on Synthetic Data with Deterministic Actions and Varying Expert Bias. True means the averages are significantly different with confidence level of 0.05. We vary the bias level in deterministic actions from humans. With more consistent human bias, human decision performance decreases. While the deferral collaboration system’s performance also decreases (which is expected due to our theoretical analysis in Theorem 1), the human-AI complementarity is still possible, suggesting the deferral collaboration leveraging expert consistency may be robust to the human bias in deterministic decisions.

Bias Level	AO/AO-EC	TS/TS-EC	JC/JC-EC
0.0	False	True	True
0.1	False	True	True
0.2	False	True	True
0.3	False	True	True
0.4	False	True	True
0.5	False	True	True

Table 28: Significance Tests on Synthetic Data with Deterministic Actions and Varying Expert Cost. True means the averages are significantly different with confidence level of 0.05. We vary the cost of human decision makers. With a larger human cost, the human decision makers’ performance decreases rapidly and the human-AI complementarity also decreases.

Cost	AO/AO-EC	TS/TS-EC	JC/JC-EC
0.0	False	True	True
0.05	False	True	True
0.1	False	False	False

Table 29: Significance Tests on Synthetic Data with Deterministic Actions. True means the averages are significantly different with confidence level of 0.05. We vary the number of training set size from 300 to 2000. With deterministic actions, increased sample size has a limited effect on algorithm’s performance, so as on the human-AI complementarity while complementarity is still possible. By leveraging expert consistency, there is a significant improvement on human-AI complementarity.

Train	AO/AO-EC	TS/TS-EC	JC/JC-EC
300	False	True	True
500	False	True	True
1000	False	True	True
2000	False	True	True

Table 30: Significance Tests on Simulation Data with Covariate Shift. True means the averages are significantly different with confidence level of 0.05. We vary the level of covariate shift by adjusting μ . Due to the assumption, humans’ performance is stable while algorithm’s performance is impacted with large μ . We find human-AI complementarity is possible for deferral collaboration under covariate shift and our out-of-distribution extension can further improve deferral system’s performance with severe shifting.

	H/AO	H/TS	H/JC	H/JC-OD	AO/TS	AO/JC	AO/JC-OD	TS/JC	TS/JC-OD	JC/JC-OD
$\mu = 1$	True	True	True	True	False	False	False	False	False	False
$\mu = 3$	True	True	True	True	False	False	False	False	False	False
$\mu = 5$	True	True	True	True	False	False	False	False	False	False
$\mu = 7$	True	True	True	True	False	False	False	False	False	False
$\mu = 9$	True	True	True	True	False	False	False	False	False	False

Table 31: Significance Tests on Semi-Synthetic Data with Covariate Shift. True means the averages are significantly different with confidence level of 0.05. We vary the human cost. $S \rightarrow M$ indicates learning algorithm on SVHN and test methods on MNIST, which is a easier task compared to $M \rightarrow S$ since images in SVHN is more complex. The algorithm’s performance is worse than humans’ for $S \rightarrow M$ when $C = 0.05$ and OD extension can further increase human-AI complementarity in the deferral collaboration system. For $M \rightarrow S$, due to the challenging task, the algorithm has a significantly worse performance while methods considering OD can still achieve robust performance.

$S \rightarrow M$	H/AO	H/TS	H/JC	H/JC-OD	AO/TS	AO/JC	AO/JC-OD	TS/JC	TS/JC-OD	JC/JC-OD
$C = 0.05$	True	False	False	False	True	True	True	False	False	False
$C = 0.1$	False	False	False	False	False	True	True	False	False	False
$C = 0.15$	False	False	False	False	False	True	True	False	False	False
$M \rightarrow S$	H/AO	H/TS	H/JC	H/JC-OD	AO/TS	AO/JC	AO/JC-OD	TS/JC	TS/JC-OD	JC/JC-OD
$C = 0.05$	True	True	True	False	False	False	True	False	True	True
$C = 0.1$	True	True	True	False	False	False	True	False	True	True
$C = 0.15$	True	True	True	False	False	False	True	False	True	True

Table 32: Significance Tests on Focus with Covariate Shift. True means the averages are significantly different with confidence level of 0.05. We vary the size of the training set size and report the total reward. Due to the covariate shift, algorithm solution performs worse than humans significantly and JC-OD can still match humans’ decision performance.

	H/AO	H/TS	H/JC	H/JC-OD	AO/TS	AO/JC	AO/JC-OD	TS/JC	TS/JC-OD	JC/JC-OD
train ratio = 0.3	True	True	True	False	True	True	True	True	True	True
train ratio = 0.5	True	True	True	False	False	True	True	True	True	True
train ratio = 0.7	True	True	True	False	False	False	True	False	True	True



(a) SVHN



(b) MNIST

Figure 5: Samples from SVHN and MNIST datasets. Each dataset has digit images with 10 classes.