A Proofs

A.1 Additional Lemmas

Lemma 1 (Monotonicity)**.** *If a utility function u satisfies Eq. 1, then u is monotone with respect to the probability that $Y = 1$,* i.e.*, for any $P, P' \in \mathcal{P}(\{0, 1\})$ such that $P(Y = 1) \leq P'(Y = 1)$, it holds that $\mathbb{E}_{Y\sim P}[u(1, Y)] \leq \mathbb{E}_{Y\sim P'}[u(1, Y)]$.*

Proof. We readily have that

$$\begin{aligned}\mathbb{E}_{Y\sim P}[u(1,Y)] &= P(Y=1)\cdot u(1,1) + (1-P(Y=1))\cdot u(1,0)\\ &\leq P'(Y=1)\cdot u(1,1) + (1-P'(Y=1))\cdot u(1,0)\\ &= \mathbb{E}_{Y\sim P'}[u(1,Y)],\end{aligned}$$

where, in the above inequality, we use that $u(1, 1) > u(1, 0)$ and $P(Y = 1) \leq P'(Y = 1)$. □

Lemma 2 (Trivial policies are not always optimal)**.** *If a utility function u satisfies Eq. 1, then there exist $P, P' \in \mathcal{P}(\{0, 1\})$ such that the trivial policies π that either always decide $T = 1$ or always decide $T = 0$ are suboptimal. In particular, for any $P, P' \in \mathcal{P}(\{0, 1\})$ such that $P(Y = 1) < c$ and $P'(Y = 1) > c$, where*

$$c = \frac{u(0,0) - u(1,0)}{u(1,1) - u(1,0) + u(0,0) - u(0,1)} \in (0, 1), \tag{8}$$

it holds that

$$\mathbb{E}_{Y\sim P}[u(1,Y)] < \mathbb{E}_{Y\sim P}[u(0,Y)] \quad \textit{and} \quad \mathbb{E}_{Y\sim P'}[u(1,Y)] > \mathbb{E}_{Y\sim P'}[u(0,Y)]. \tag{9}$$

Proof. Let P be any distribution such that

$$P(Y=1) < c = \frac{u(0,0) - u(1,0)}{u(1,1) - u(1,0) + u(0,0) - u(0,1)},$$

where $c \in (0, 1)$ because, by assumption, u satisfies Eq. 1. Now, by rearranging the above inequality, we have that

$$P(Y=1)\cdot u(1,1) + (1-P(Y=1))\cdot u(1,0) < P(Y=1)\cdot u(0,1) + (1-P(Y=1))\cdot u(0,0),$$

and, using the definition of the expectation, it immediately follows that

$$\mathbb{E}_{Y\sim P}[u(1,Y)] < \mathbb{E}_{Y\sim P}[u(0,Y)].$$

The same argument can be used to show that, for any distribution P' such that $P'(Y = 1) > c$, it holds that $\mathbb{E}_{Y\sim P'}[u(1, Y)] > \mathbb{E}_{Y\sim P'}[u(0, Y)]$. Finally, note that, since $c \in (0, 1)$, we know that such distributions P and P' exist. □

A.2 Proof of Theorem 3

Before proving Theorem 3, we rewrite the expected utility with respect to the probability distribution $P^{\mathcal{M}}$ in terms of confidence H and B by using the law of total expectation,

$$\mathbb{E}_\pi[u(T,Y)] = \mathbb{E}_{H,B\sim P^{\mathcal{M}}(H,B)}\left[\mathbb{E}_\pi[u(T,Y)|H,B]\right].$$

Here, to simplify notation, we will write

$$\mathbb{E}_{H,B}\left[\mathbb{E}_\pi[u(T,Y) \mid H,B]\right],$$

where note that, using the law of total expectation, we can write the inner expectation in the above expression in terms of the utilities of the trivial policies, *i.e.*,

$$\begin{aligned}\mathbb{E}_\pi[u(T,Y) \mid H,B] = \mathbb{E}[u(1,Y) \mid H,B]\cdot P_\pi(T=1 \mid H,B)\\ + \mathbb{E}[u(0,Y) \mid H,B]\cdot P_\pi(T=0 \mid H,B),\end{aligned} \tag{10}$$

and we will use P to refer to probabilities induced by SCM $\mathcal{M}$, *e.g.*, $P(H, B)$ to denote $P^{\mathcal{M}}(H, B)$. Now, we restate and prove Theorem 3.

Theorem 3. There exist (infinitely many) AI-assisted decision making processes $\mathcal{M}$ satisfying Eqs. 2 and 3, with utility functions $u(T, Y)$ satisfying Eq. 1, such that f_B is perfectly calibrated and f_H is monotone but any AI-assisted decision policy $\pi \in \Pi(H, B)$ that satisfies monotonicity is suboptimal, *i.e.*, $\mathbb{E}_\pi[u(T, Y)] < \mathbb{E}_{\pi^*}[u(T, Y)]$.

Proof. To prove the above claim, we construct a monotone confidence function f_H, perfectly calibrated confidence function f_B and distribution $P^{\mathcal{M}}$ for which any monotone AI-assisted decision policy $\pi \in \Pi(H, B)$ achieves strictly lower utility than a carefully constructed non monotone AI-assisted decision policy $\tilde{\pi} \in \Pi(H, B)$.

We will present the proof in three parts. First, we will introduce the main building block and idea behind the proof by a small construction of f_H, f_B and $P^{\mathcal{M}}$ with $|\mathcal{H}| = |\mathcal{B}| = 3$, where $\mathcal{B} \subseteq [0, 1]$ denotes the (discrete) output space of the classifier's confidence function. We then construct examples of f_H, f_B and $P^{\mathcal{M}}$ for arbitrary $|\mathcal{H}| = k$ and $|\mathcal{B}| = m$ with $m, k \in \mathbb{N}$, $m > k \geq 2$. Lastly, we construct examples where $\mathcal{B}$ is non-discrete and $|\mathcal{H}| = k$ with $k > 2$.

Main building block and small example.

We start by presenting the main idea of the proof using an example with a small set of confidence values $\mathcal{H}$ and $\mathcal{B}$. Let the values of the decision maker's confidence H be in $\mathcal{H} = \{h_1, h_2, h_3\}$ and the values of the classifier's confidence B be in $\mathcal{B} = \{b_1, b_2, b_3\}$, with order $h_i < (h_i + 1)$ and $b_i < (b_i + 1)$ respectively.

Our main building block, consists of two distributions $P^-, P^+ \in \mathcal{P}(\{0, 1\})$ with $P^-(Y = 1) < c$ and $P^+(Y = 1) > c$, where c depends on utility u as described by Eq. 8 in Lemma 2. We use these distributions for our constructions of f_H, f_B and $P^{\mathcal{M}}$, so that for some realizations of H, B distribution $P(Y = 1 \mid H, B)$ is either P^- or P^+. Using Lemma 2 and from Eq. 10, we have that:

(I) For any h_i, b_i such that $P(Y \mid H = h_i, B = b_i) = P^-$, it holds that

$$\mathbb{E}[u(1, Y) \mid H = h_i, B = b_i] < \mathbb{E}[u(0, Y) \mid H = h_i, B = b_i].$$

Hence, *decreasing* $P_\pi(T = 1 \mid H, B)$ *increases* $\mathbb{E}[u(T, Y) \mid H = h_i, B = b_i]$.

(II) For any h_i, b_i such that $P(Y \mid H = h_i, B = b_i) = P^+$, it holds that

$$\mathbb{E}[u(1, Y) \mid H = h_i, B = b_i] > \mathbb{E}[u(0, Y) \mid H = h_i, B = b_i].$$

Hence, *increasing* $P_\pi(T = 1 \mid H, B)$ *increases* $\mathbb{E}[u(T, Y) \mid H = h_i, B = b_i]$.

Intuitively, suppose we now have that, for confidence values h_2, b_2, $Y \sim P^+$ and, for confidence values h_3, b_2, $Y \sim P^-$, *i.e.*, $P(Y \mid H = h_2, B = b_2) = P^+$ and $P(Y \mid H = h_3, B = b_2) = P^-$. Then, any non-monotone AI-assisted decision policy $\tilde{\pi}$ with $P_{\tilde{\pi}}(T = 1 \mid H = h_2, B = b_2) > P_{\tilde{\pi}}(T = 1 \mid H = h_3, B = b_2)$ will have higher expected utility than any monotone AI-assisted decision policy given confidence values h_2, b_2 and h_3, b_2. Finally, under an appropriate choice of distribution $P(H, B)$, such non-monotone AI-assisted decision policies $\tilde{\pi}$ will offer higher overall utility in expectation.

We formalize this intuition with the following lemma:

Lemma 3. *Let $\mathcal{M}$ be any AI-assisted decision making process satisfying Eqs. 2 and 3, with utility function $u(T, Y)$ satisfying Eq. 1. If f_H, f_B and $P^{\mathcal{M}}$ are such that there exists confidence values $b \in \mathcal{B}$, $h_i, h_j \in \mathcal{H}$, with $h_i < h_j$, which satisfy*

$$\begin{aligned} P(H = h_i, B = b) > 0, \quad P(H = h_j, B = b) > 0, \\ P(Y \mid H = h_i, B = b) = P^+ \quad \textit{and} \quad P(Y \mid H = h_j, B = b) = P^-, \end{aligned} \tag{11}$$

for some distributions P^-, P^+ with $P^-(Y = 1) < c$ and $P^+(Y = 1) > c$, where

$$c = \frac{u(0, 0) - u(1, 0)}{u(1, 1) - u(1, 0) + u(0, 0) - u(0, 1)}. \tag{12}$$

Then, for any monotone AI-assisted decision policy $\pi \in \Pi(H, B)$, there exists an AI-assisted decision policy $\tilde{\pi} \in \Pi(H, B)$ which is not monotone and achieves a stricly greater utility than π, i.e.*,* $\mathbb{E}_\pi[u(T, Y)] < \mathbb{E}_{\tilde{\pi}}[u(T, Y)]$.

Proof. Let π be a monotone AI-assisted decision policy, then it must hold that $P_\pi(T = 1 \mid H = h_i, B = b) \leq P_\pi(T = 1 \mid H = h_j, B = b)$ (see Eq. 4). Let $\tilde{\pi}$ be an identical AI-assisted decision

policy to π up to the decision for confidence values h_i, b and h_j, b. We distinguish between three cases.

— **Case 1**: $P_\pi(T = 1 \mid H = h_i, B = b) < P_\pi(T = 1 \mid H = h_j, B = b)$.

Let the probability of $T = 1$ under $\tilde{\pi}$ for confidence values h_i, b and h_j, b be switched compared to π, *i.e.*,

$$P_{\tilde{\pi}}(T = 1 \mid H = h_i, B = b) = P_\pi(T = 1 \mid H = h_j, B = b),$$
$$P_{\tilde{\pi}}(T = 1 \mid H = h_j, B = b) = P_\pi(T = 1 \mid H = h_i, B = b).$$

Then, $\tilde{\pi}$ is not monotone, as Eq. 4 is not satisfied, and it holds that

$$P_{\tilde{\pi}}(T = 1 \mid H = h_i, B = b) > P_\pi(T = 1 \mid H = h_i, B = b),$$
$$P_{\tilde{\pi}}(T = 1 \mid H = h_j, B = b) < P_\pi(T = 1 \mid H = h_j, B = b).$$

As we decreased $P(T = 1 \mid H = h_j, B = b)$ and increased $P(T = 1 \mid H = h_i, B = b)$, by properties (I) and (II), it must hold that the expected utility of $\tilde{\pi}$ given confidence values h_i, b and h_j, b is higher than the one of π, *i.e.*,

$$\mathbb{E}_{\tilde{\pi}}[u(T, Y) \mid H = h_i, B = b] > \mathbb{E}_\pi[u(T, Y) \mid H = h_i, B = b] \quad \text{and} \tag{13}$$
$$\mathbb{E}_{\tilde{\pi}}[u(T, Y) \mid H = h_j, B = b] > \mathbb{E}_\pi[u(T, Y) \mid H = h_j, B = b]. \tag{14}$$

— **Case 2**: $0 < P_\pi(T = 1 \mid H = h_i, B = b) = P_\pi(T = 1 \mid H = h_j, B = b) \leq 1$.

Let the probability of $T = 1$ under $\tilde{\pi}$ for confidence values h_j, b be strictly lower compared to π and be the same as π for h_i, b. Then, $\tilde{\pi}$ is not monotone, since by case assumption

$$P_{\tilde{\pi}}(T = 1 \mid H = h_i, B = b) = P_\pi(T = 1 \mid H = h_j, B = b) > P_{\tilde{\pi}}(T = 1 \mid H = h_j, B = b)$$

and the inequality in Eq. 14 holds by property (I).

— **Case 3**: $P_\pi(T = 1 \mid H = h_i, B = b) = P_\pi(T = 1 \mid H = h_j, B = b) = 0$.

Let the probability of $T = 1$ under $\tilde{\pi}$ for confidence values h_i, b be strictly higher compared to π and be the same as π for h_j, b. Then, $\tilde{\pi}$ is not monotone, since by case assumption

$$P_{\tilde{\pi}}(T = 1 \mid H = h_j, B = b) = P_\pi(T = 1 \mid H = h_i, B = b) < P_{\tilde{\pi}}(T = 1 \mid H = h_i, B = b)$$

and the inequality in Eq. 13 holds by property (II).

As in all three cases at least one of the strict inequalities in Eqs. 13 or 14 holds and $\tilde{\pi}$ is equivalent to π (*i.e.*, it has the same expected conditional utility) given any other pair of confidence values $h' \in \mathcal{H}$, $b' \in \mathcal{B}$, we have that

$$\mathbb{E}_{\tilde{\pi}}[u(T, Y)] = \mathbb{E}[\mathbb{E}_{\tilde{\pi}}[u(T, Y)]|H, B] > \mathbb{E}[\mathbb{E}_\pi[u(T, Y)|H, B] = \mathbb{E}_\pi[u(T, Y)].$$

□

Before proceeding further, we would like to note that we may also state Lemma 3 using $h \in \mathcal{H}$, $b_i, b_j \in \mathcal{B}$, with $b_i < b_j$, the proof would follow analogously.

Now, we construct an AI-decision making process $\mathcal{M}$, with $\mathcal{H} = \{h_1, h_2, h_3\}$ and $\mathcal{B} = \{b_1, b_2, b_3\}$, such the decision maker's confidence f_H is monotone, the classifier's confidence f_B is perfectly calibrated, and the conditions of Lemma 3 are satisfied. First, let f_H, f_B and $P^{\mathcal{M}}$ be such that

$$P(f_B(Z) = b_j) = \begin{cases} 3/6 & \text{if } j = 1 \\ 2/6 & \text{if } j = 2 \\ 1/6 & \text{if } j = 3 \\ 0 & \text{otherwise} \end{cases} \quad \text{and}$$

$$P(H = h_i \mid B = b_j) := P_{X,V}(H = h_i \mid f_B(Z) = b_j) = \begin{cases} \frac{1}{4-j} & \text{if } i \geq j \\ 0 & \text{otherwise.} \end{cases}$$

Then, it readily follows that $P(H = h_i, B = b_j) = 1/6$ for $i \geq j$ and $P(H = h_i, B = b_j) = 0$ otherwise. Moreover, for each pair of confidence values (h_i, b_j) with positive probability $P(H = h_i, B = b_j)$, we set

$$P(Y = 1 \mid H = h_i, B = b_j) = \begin{cases} P^+ & \text{if } i = j = 2 \text{ or } (i = 3 \text{ and } j \in \{1, 3\}) \\ P^- & \text{if } (j = 2 \text{ and } i = 3) \text{ or } (j = 1 \text{ and } i \in \{1, 2\}), \end{cases}$$

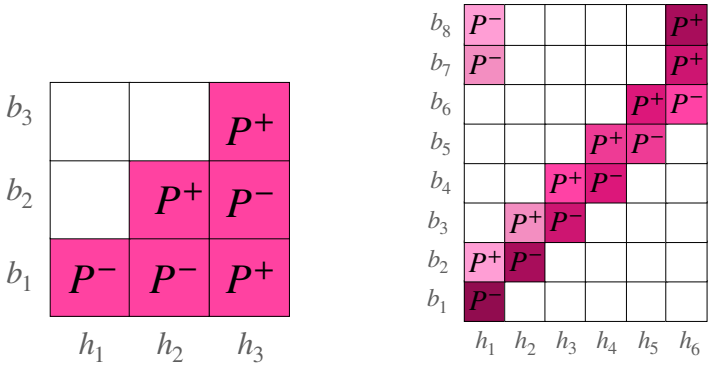


Figure 4: Nonzero values of $P(Y = 1|H = h_i, B = b_j)$ and $P(H = h_i, B = b_j)$ for every $h_i \in \mathcal{H}$ and $b_j \in \mathcal{B}$ used in the first (left) and second (right) part of the proof of Theorem 3. In each cell (h_i, b_j) in both panels, P^+ or P^- is the value of $P(Y = 1|H = h_i, B = b_j)$ and lighter color means lower value of $P(H = h_i, B = b_j)$, where white means $P(Y = 1|h = h_i, B = b_j) = 0$ and $P(H, B) = 0$. In both panels, the assignment of values is very stylized to facilitate the proof—the classifier's confidence function f_B partitions the feature space in a way such that a rational decision maker is unable to take decisions that maximize utility for almost all confidence values. However, less stylized examples also satisfy the conditions of Lemma 3. For example, as long as there is one triplet of confidence values b_2, h_2, h_3 (or h_3, b_1, b_2 in the left example) for which a rational decision maker is unable to take decisions that maximize utility, Lemma 3 can be applied.

as shown in Figure 4 (left). Then, it readily follows that f_H is monotone with respect to the probability that $Y = 1$, *i.e.*, $P(Y = 1 \mid H = h_i) \leq P(Y = 1 \mid H = h_{i+1})$), and we have that the classifier's confidence values

$$
\begin{aligned}
b_j &:= \sum_{i:i\geq j} P(H = h_i \mid B = b_j) \cdot P(Y = 1 \mid H = h_i, B = b_j) \\
&= \begin{cases} 2/3 \cdot P^- + 1/3 \cdot P^+ & \text{if } j = 1 \\ 1/2 \cdot P^- + 1/2 \cdot P^+ & \text{if } j = 2 \\ P^+ & \text{if } j = 3 \\ 0 & \text{otherwise} \end{cases}
\end{aligned}
$$

are perfectly calibrated and satisfy that $b_j < b_{j+1}$.

Finally, using Lemma 3 with $b = b_2$, $h_i = h_2$, $h_j = h_3$, we have that any monotone AI-assisted decision policy is suboptimal for any $\mathcal{M}$ with f_H, f_B and $P^{\mathcal{M}}$ as defined above.

Construction with arbitrary $|\mathcal{H}| = k$ and $|\mathcal{B}| = m$, $m > k \geq 2$.

In this second part of the proof, we construct an AI-assisted decision making processes $\mathcal{M}$, with $|\mathcal{H}| = k$ and $|\mathcal{B}| = m$ such that $m > k \geq 2$, such that the decision maker's confidence f_H is monotone, the classifier's confidence f_B is perfectly calibrated and the conditions of Lemma 3 are satisfied.

First, let the space of confidence values be $\mathcal{H} = \{h_i\}_{i\in[k]}$ and $\mathcal{B} = \{b_j\}_{j\in[m]}$, with order $h_i < h_{i+1}$ and $b_i < b_{i+1}$, respectively, and f_H, f_B and $P^{\mathcal{M}}$ be such that $P(f_B(Z) = b_j) = 1/m$ and

$$
P(H = h_i \mid B = b_j) := P_{X,V}(H = h_i \mid f_B(Z) = b_j) = \begin{cases} \frac{m-j+1}{m} & \text{if } j = i \\ \frac{m-j+1}{m} & \text{if } i = 1, j > k \\ \frac{j-1}{m} & \text{if } j = i+1, j \leq k \\ \frac{j-1}{m} & \text{if } i = k, j > k \\ 0 & \text{otherwise.} \end{cases} \tag{15}
$$

Moreover, for each pair of confidence values (h_i, b_j) with positive probability $P(H = h_i, B = b_j)$, we set

$$P(Y = 1 \mid H = h_i, B = b_j) = \begin{cases} P^- & \text{if } j = i \\ P^- & \text{if } i = 1, j > k \\ P^+ & \text{if } j = i+1, j \leq k \\ P^+ & \text{if } i = k, j > k, \end{cases} \tag{16}$$

as shown in Figure 4 (right). Further, we set the classifier's confidence values b_j to

$$b_j := \frac{m - j + 1}{m} \cdot P^- + \frac{j - 1}{m} \cdot P^+ .$$

Then, it holds that $b_j < b_{j+1}$ and f_B is perfectly calibrated as

$$P(Y = 1 \mid B = b_j) = \begin{cases} P(H = h_j \mid B = b_j) \cdot P^- + P(H = h_{j-1} \mid B = b_j) \cdot P^+ & \text{if } j \leq k \\ P(H = h_1 \mid B = b_j) \cdot P^- + P(H = h_k \mid B = b_j) \cdot P^+ & \text{if } j > k \end{cases}$$

and thus, using the definitions of $P(H \mid B)$ and $P(Y \mid H, B)$, we have that $P(Y \mid B = b_j) = b_j$.

To show that f_H is monotone with respect to the probability that $Y = 1$, first note that $P(H = h_i, B = b_i)$ decreases as i increases and $P(H = h_i, B = b_{i+1})$ increases as i increases. Moreover, further note that $P(Y = 1 \mid H = h_i, B = b_i) = P^- < P(Y = 1 \mid H = h_i, B = b_{i+1}) = P^+$. Hence, for any $i \in \{2, \ldots, k-1\}$, it readily follows that

$$\begin{aligned} P(Y = 1 \mid H = h_i) &= P^+ \cdot P(B = b_{i+1} | H = h_i) + P^- \cdot P(B = b_i | H = h_i) \\ &\leq P(Y = 1 \mid H = h_{i+1}), \end{aligned}$$

and, for $i = 1$, it is evident that $P(Y = 1 \mid H = h_1) < P(Y = 1 \mid H = h_2)$.

Finally, using Lemma 3 with any choice of confidence values $b = b_j$, $h_i = h_{j-1}$ and $h_j = h_j$ with $j \in \{2, \ldots, k\}$, we have that any monotone AI-assisted decision policy π is suboptimal for any $\mathcal{M}$ with $|\mathcal{H}| = k$ and $|\mathcal{B}| = m$, $m > k \geq 2$, and f_H, f_B and $P^{\mathcal{M}}$ as defined above. Here, note that, as we do not fix the exact distributions P^- and P^+, the above Lemma applies to infinitely many AI-assisted decision making processes $\mathcal{M}$.

Construction with $\mathcal{B} \subseteq [0, 1]$ and $|\mathcal{H}| = k$.

In this last part of the proof, we construct an AI-assisted decision making process $\mathcal{M}$, with $|\mathcal{H}| = k \geq 2$ and $\mathcal{B} \subseteq [0, 1]$, such that the decision maker's confidence function f_H is monotone, the classifier's confidence function f_B is perfectly calibrated and the conditions of Lemma 3 are satisfied.

First, let the space of confidence values be $\mathcal{H} = \{h_i\}_{i \in [k]}$, with order $h_i < h_{i+1}$, the feature space[10] $\mathcal{X} = [0, 1]$, and f^-, f^+ be two strictly monotone increasing functions with

$$f^- : [0, 1] \to [0, c) \quad \text{and} \quad f^+ : [0, 1] \to (c, 1], \tag{17}$$

where

$$c = \frac{u(0, 0) - u(1, 0)}{u(1, 1) - u(1, 0) + u(0, 0) - u(0, 1)}. \tag{18}$$

Further, let $Q_{k+1} = \{q_0, q_1, \ldots q_k, q_{k+1}\}$ be a set of quantiles such that $P(X \leq q_j) = j/(k+1)$ for all $j \in \{0, 1, \ldots, k+1\}$ and thus, we have that, for all $j \in [k+1]$,

$$\text{for} \quad I_j := (q_{j-1}, q_j], \quad \text{it holds that} \quad P(X \in I_j) = \frac{1}{k+1}.$$

Now, let f_H and $P^{\mathcal{M}}$ be such that

$$P_V(H = h_i \mid X, X \in I_j) = \begin{cases} 1/2 & \text{if } i \in \{j-1, j\} \\ 1 & \text{if } i = j = 1 \text{ or } (i = k \text{ and } j = k+1) \\ 0 & \text{otherwise,} \end{cases} \tag{19}$$

[10] For a more general feature space $\mathcal{X}$, we can use a mapping ϕ of $\mathcal{X}$ to $[0, 1]$. The proof works analogously by substituting X with $\phi(X)$.

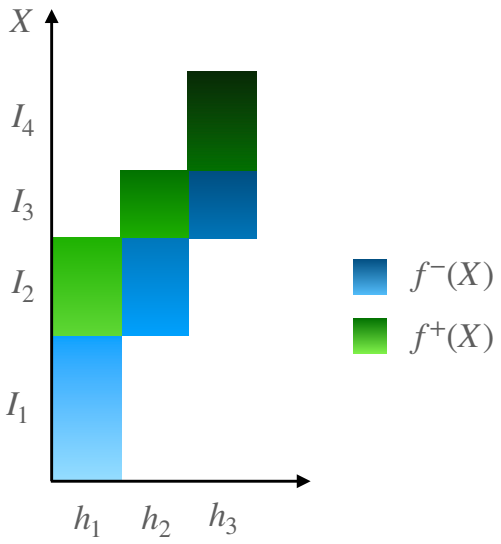


Figure 5: Nonzero values of $P(Y = 1|X, H = h_i, X \in I_j)$ for every $h_i \in \mathcal{H}$, with $|\mathcal{H}| = 3$, and $I_j = (q_{j-1}, q_j]$, with $q_j \in Q_4$ used in the last part of the proof of Theorem 3. Lighter color means lower value of f^- or f^+.

and let

$$P(Y = 1 \mid X, H = h_i, X \in I_j) = \begin{cases} f^-(X) & \text{if } j = i \text{ or } (i = j = 1\,) \\ f^+(X) & \text{if } j = i + 1 \text{ or } (i = k \text{ and } j = k + 1), \end{cases} \tag{20}$$

as shown in Figure 5. Next, we define

$$f_B(Z) = f_B(X) := P(Y = 1 \mid X) = \begin{cases} f^-(X) & \text{if } X \in I_1 \\ f^+(X) & \text{if } X \in I_{k+1} \\ (f^-(X) + f^+(X))/2 & \text{otherwise,} \end{cases}$$

which, by construction, is perfectly calibrated.

To show that the decision maker's confidence function f_H is monotone with respect to the probability that $Y = 1$, we first note that, using Eq. 19, we have that

$$P(X \in I_j \mid H = h_i) = \begin{cases} 1/2 & \text{if } 1 < i < k \text{ and } j \in \{i, i + 1\} \text{ and} \\ 1 & \text{if } i = j = 1 \\ 1 & \text{if } i = k \text{ and } j = k + 1 \\ 0 & \text{otherwise.} \end{cases} \tag{21}$$

Hence, using Eq. 21 and the law of total probability, for any $i \in \{2, \ldots, k - 2\}$, we have that

$$\begin{aligned} P(Y = 1 \mid H = h_i) &= \frac{1}{2}\left[P(Y = 1 \mid H = h_i, X \in I_i) + P(Y = 1 \mid H = h_i, X \in I_{i+1})\right] \\ &\leq \frac{1}{2}\left[f^-(q_i) + f^+(q_{i+1})\right] \\ &= \frac{1}{2}\left[f^-\left(\inf I_{i+1}\right) + f^+\left(\inf I_{i+2}\right)\right] \\ &\leq \frac{1}{2}\left[P(Y = 1 \mid H = h_{i+1}, X \in I_{i+1}) + P(Y = 1 \mid H = h_{i+1}, X \in I_{i+2})\right] \\ &= P(Y = 1 \mid H = h_{i+1}), \end{aligned}$$

where the inequalities follow from the fact that f^- and f^+ are strictly monotone increasing. Corner cases for $i = 1$ and $i = k - 1$ can be shown analogously by further using that $f^-(X) < c < f^+(X)$ for all X.

Finally, using Lemma 3 with any choice of confidence values $h_i = h_{j-1}\ h_j = h_j, j \in \{2, \cdots, k-1\}$ and $b = f_B(X)$ with $X \in I_j$, we have that any monotone AI-assisted decision policy π is suboptimal for any $\mathcal{M}$ with $|\mathcal{B}| \subseteq [0,1]$ and $|\mathcal{H}| = k$, $k \geq 2$ and f_H, f_B and $P^{\mathcal{M}}$ as defined above. □

A.3 Proof of Theorem 5

We prove the statement by contraposition. Let $\mathcal{M}$ be an AI-assisted decision making process satisfying Eqs. 2 and 3, with a utility function $u(T, Y)$ satisfying Eq. 1 and let $\mathcal{M}$ be such that f_B satisfies α-alignment with respect to f_H and f_B has output space $\mathcal{B} \subseteq [0,1]$. Assume there exists no (near-)optimal monotone AI-assisted decision policy for utility u. Thus, there must exist an optimal AI-assisted decision policy $\pi \in \Pi(H, B)$ which is not monotone and has strictly greater expected utility than any monotone policy. However, we show that we can modify π to a monotone AI-assisted decision policy $\hat{\pi} \in \Pi(H, B)$ with near-optimal expected utility.

As π is not monotone, there must exist confidence values $h_1, h_2 \in \mathcal{H}$, $h_1 \leq h_2$, and $b_1, b_2 \in \mathcal{B}$, $b_1 \leq b_2$, such that

$$\pi(h_1, b_1, w) > \pi(h_2, b_2, w) \quad \text{for some} \quad w \in \mathcal{W}, \tag{22}$$

where $\mathcal{W}$ denotes the space of noise values. In what follows, let $\tilde{\mathcal{W}}_{h_1,b_1}^{(\pi,h_2,b_2)} \subseteq \mathcal{W}$ denote the set containing any such w and let $\tilde{\mathcal{W}}^{(\pi,h_2,b_2)} = \bigcup_{h,b \in \mathcal{H}\times\mathcal{B}} \tilde{\mathcal{W}}_{h,b}^{(\pi,h_2,b_2)}$.

For any confidence value $h', b' \in \mathcal{H} \times [0,1]$, we modify policy π to a policy $\hat{\pi}$ as follows. Let $\{\tilde{\mathcal{S}}_h\}_{h\in\mathcal{H}}$ denote the sets satisfying the α-alignment condition for f_B with respect to f_H and, given confidence h', let $\hat{b}_{h'}$ denote the smallest confidence value of f_B, such that there exist $h \leq h'$ with $P(Y = 1 \mid B = \hat{b}_{h'}, Z \in \tilde{\mathcal{S}}_h) \geq c$, *i.e.*,

$$\hat{b}_{h'} := \min\{b \in \mathcal{B} \mid P(Y = 1 \mid B = b, Z \in \tilde{\mathcal{S}}_h) \geq c \text{ for } h \leq h'\}. \tag{23}$$

Now, we define a new AI-assisted policy $\hat{\pi}$ from π as follows,

$$\hat{\pi}(h', b', w) := \begin{cases} 1 & \text{if } b' \geq \hat{b}_h \text{ and } w \in \bigcup_{h\leq h', b\in\left[\hat{b}_{h'}, b'\right]} \tilde{\mathcal{W}}^{(\pi,h,b)} \\ 0 & \text{if } b' < \hat{b}_h \text{ and } w \in \bigcup_{h\geq h', b\in\left[b', \hat{b}_{h'}\right)} \tilde{\mathcal{W}}^{(\pi,h,b)} \\ \pi(h', b', w) & \text{otherwise.} \end{cases} \tag{24}$$

Next, we show that $\hat{\pi}$ is monotone and $\mathbb{E}_{\hat{\pi}}[u(T, Y)] \geq \mathbb{E}_{\pi}[u(T, Y)] + \alpha \cdot a$ for some constant a.

Proof $\hat{\pi}$ is a monotone assisted policy.

To prove that $\hat{\pi} \in \Pi(H, B)$ is a monotone AI-assisted decision policy, we show that, for all $h', h'' \in \mathcal{H}, b', b'' \in \mathcal{B}$, with $h' \leq h'', b' \leq b''$, it holds that $\tilde{\mathcal{W}}_{h',b'}^{(\hat{\pi},h'',b'')} = \emptyset$. We distinguish between three cases.

— **Case 1**: $b' \geq \hat{b}_{h'}$ and $b'' \geq \hat{b}_{h''}$.

Since $h' \leq h''$, $b' \leq b''$ and, by definition, $\hat{b}_{h''} \leq \hat{b}_{h'}$ since $h' \leq h''$, we have that

$$\bigcup_{h\leq h', b\in\left[\hat{b}_{h'}, b'\right]} \tilde{\mathcal{W}}^{(\pi,h,b)} \subseteq \bigcup_{h\leq h'', b\in\left[\hat{b}_{h''}, b''\right]} \tilde{\mathcal{W}}^{(\pi,h,b)}.$$

Hence, we can conclude that

$$\hat{\pi}(h', b', w) \leq 1 = \hat{\pi}(h'', b'', w) \text{ for all } w \in \bigcup_{h\leq h'', b\in\left[\hat{b}_{h''}, b''\right]} \tilde{\mathcal{W}}^{(\pi,h,b)}. \tag{25}$$

Further, for any other $w \in \mathcal{W} - \bigcup_{h\leq h'', b\in\left[\hat{b}_{h''}, b''\right]} \tilde{\mathcal{W}}^{(\pi,h,b)} \subseteq \mathcal{W} - \tilde{\mathcal{W}}_{h',b'}^{(\pi,h'',b'')}$, we have that $\hat{\pi}(h', b', w) = \pi(h', b', w)$ and $\hat{\pi}(h'', b'', w) = \pi(h'', b'', w)$ and, by definition of $\tilde{\mathcal{W}}_{h',b'}^{(\pi,h'',b'')}$, it follows that

$$\hat{\pi}(h', b', w) \leq \hat{\pi}(h'', b'', w) \text{ for all } w \in \mathcal{W} - \bigcup_{h\leq h'', b\in\left[\hat{b}_{h''}, b''\right]} \tilde{\mathcal{W}}^{(\pi,h,b)}. \tag{26}$$

From Eqs. 25 and 26, it follows that $\tilde{\mathcal{W}}_{h',b'}^{(\hat{\pi},h'',b'')} = \emptyset$.

— **Case 2**: $b' < \hat{b}_{h'}$ and $b'' \geq \hat{b}_{h''}$.

By definition of $\hat{\pi}$, we have that

$$\hat{\pi}(h', b', w) \leq 1 = \hat{\pi}(h'', b'', w) \text{ for all } w \in \bigcup_{h \leq h'', b \in [\hat{b}_{h''}, b'']} \tilde{\mathcal{W}}^{(\pi,h,b)} \tag{27}$$

and

$$\hat{\pi}(h', b', w) = 0 \leq \hat{\pi}(h'', b'', w) \text{ for all } w \in \bigcup_{h \geq h', b \in [b', \hat{b}_{h'})} \tilde{\mathcal{W}}^{(\pi,h,b)} \tag{28}$$

Analogously to case 1, since the values of w below are also in $\mathcal{W} - \tilde{\mathcal{W}}_{h',b'}^{(\pi,h'',b'')}$ and $\hat{\pi}$ is equivalent to π for these values, we have that

$$\hat{\pi}(h', b', w) \leq \hat{\pi}(h'', b'', w) \text{ for all } w \in \mathcal{W} - \bigcup_{h \leq h'', b \in [\hat{b}_{h''}, b'']} \tilde{\mathcal{W}}^{(\pi,h,b)} - \bigcup_{h \geq h', b \in [b', \hat{b}_{h'})} \tilde{\mathcal{W}}^{(\pi,h,b)} \tag{29}$$

From Eqs. 27, 28 and 29, it follows that $\tilde{\mathcal{W}}_{h',b'}^{(\hat{\pi},h'',b'')} = \emptyset$.

— **Case 3**: $b' < \hat{b}_{h'}$ and $b'' < \hat{b}_{h''}$.

Since $h' \leq h''$, $b' \leq b''$ and, by definition, $\hat{b}_{h''} \leq \hat{b}_{h'}$ since $h' \leq h''$, we have that

$$\bigcup_{h \geq h'', b \in [b'', \hat{b}_{h''})} \tilde{\mathcal{W}}^{(\pi,h,b)} \subseteq \bigcup_{h \geq h', b \in [b', \hat{b}_{h'})} \tilde{\mathcal{W}}^{(\pi,h,b)}.$$

Hence, we can conclude that

$$\hat{\pi}(h', b', w) = 0 \leq \hat{\pi}(h'', b'', w) \text{ for all } w \in \bigcup_{h \geq h', b \in [b', \hat{b}_{h'})} \tilde{\mathcal{W}}^{(\pi,h,b)} \tag{30}$$

Again analogously to case 1, since the values of w below are also in $\mathcal{W} - \tilde{\mathcal{W}}_{h',b'}^{(\pi,h'',b'')}$ and $\hat{\pi}$ is equivalent to π for these values, we have that

$$\hat{\pi}(h', b', w) \leq \hat{\pi}(h'', b'', w) \text{ for all } w \in \mathcal{W} - \bigcup_{h \geq h', b \in [b', \hat{b}_{h'})} \tilde{\mathcal{W}}^{(\pi,h,b)} \tag{31}$$

From Eqs. 30 and 31, it follows that $\tilde{\mathcal{W}}_{h',b'}^{(\hat{\pi},h'',b'')} = \emptyset$.

Note that, we cannot have a case where $b' \geq \hat{b}_{h'}$ and $b'' < \hat{b}_{h''}$, as this would imply $b'' < b'$. Since, in all three possible cases, we have shown that $\tilde{\mathcal{W}}_{h',b'}^{(\hat{\pi},h'',b'')} = \emptyset$, we can conclude that $\hat{\pi} \in \Pi(H, B)$ is monotone.

Proof $\hat{\pi}$ is near optimal.

First, we rewrite the inner expectation in Eq. 10 as

$$\mathbb{E}_{\pi}[u(T, Y) \mid H, B] = \mathbb{E}[u(0, Y) \mid H, B] + (\mathbb{E}[u(1, Y) \mid H, B] - \mathbb{E}[u(0, Y) \mid H, B]) \cdot P_{\pi}(T = 1 \mid H, B).$$

Further, recall that $|\tilde{\mathcal{S}}_h| \geq (1 - \alpha/2)|\mathcal{S}_h|$ for all $h \in \mathcal{H}$ and, for all $h', h'' \in \mathcal{H}$, $h' \leq h''$ and all $b', b'' \in [0, 1]$, $b' \leq b''$, we have that

$$P(Y = 1 \mid f_B(Z) = b', Z \in \tilde{\mathcal{S}}_{h'}) - P(Y = 1 \mid f_B(Z) = b'', Z \in \tilde{\mathcal{S}}_{h''}) \leq \alpha \tag{32}$$

Now, for any $h' \in \mathcal{H}, b' \in \mathcal{B}$, we show an upper bound on $\mathbb{E}_{\pi}[u(T, Y) \mid H = h', B = b'] - \mathbb{E}_{\hat{\pi}}[u(T, Y) \mid H = h', B = b']$. We distinguish between three cases.

— **Case 1**: $b' \geq \hat{b}_{h'}$ and $P(Y = 1 \mid H = h', B = b') \geq c$.

Using Lemma 2, we have that

$$(\mathbb{E}[u(1,Y) \mid H=h', B=b'] - \mathbb{E}[u(0,Y) \mid H=h', B=b']) \geq 0 \tag{33}$$

Moreover, as $b' \geq \hat{b}_{h'}$, the distribution of positive decisions in $\hat{\pi}$ may also increases for h', b' compared to π (see Eq. 24), *i.e.*,

$$P_\pi(T=1 \mid H=h', B=b') - P_{\hat{\pi}}(T=1 \mid H=h', B=b') \leq 0$$

Hence, it follows that

$$\begin{aligned}
&\mathbb{E}_\pi[u(T,Y) \mid H=h', B=b'] - \mathbb{E}_{\hat{\pi}}[u(T,Y) \mid H=h', B=b'] \\
&\quad = (\mathbb{E}[u(1,Y) \mid H=h', B=b'] - \mathbb{E}[u(0,Y) \mid H=h', B=b']) \\
&\quad \times (P_\pi(T=1 \mid H=h', B=b') - P_{\hat{\pi}}(T=1 \mid H=h', B=b')) \leq 0.
\end{aligned} \tag{34}$$

— **Case 2**: $b' \geq \hat{b}_{h'}$ and $P(Y=1 \mid H=h', B=b') < c$.

Since $b' \geq \hat{b}_{h'}$, there exists $h, b \in \mathcal{H} \times \mathcal{B}$, with $h \leq h'$, $b \leq b'$, such that $P(Y=1 \mid B=b, Z \in \tilde{\mathcal{S}}_h) \geq c$. Moreover, using the definition of α-alignment, we have that

$$P(Y=1 \mid B=b, Z \in \tilde{\mathcal{S}}_h) \leq P(Y=1 \mid B=b', Z \in \tilde{\mathcal{S}}_{h'}) + \alpha \tag{35}$$

Then, we can use this to lower bound the expected utility of $T=1$ given $B=b'$ and $Z \in \tilde{\mathcal{S}}_{h'}$ as follows:

$$\begin{aligned}
&\mathbb{E}[u(1,Y) \mid B=b, Z \in \tilde{\mathcal{S}}_h] - \mathbb{E}[u(1,Y) \mid B=b', Z \in \tilde{\mathcal{S}}_{h'}] \\
&\quad = u(1,1) \cdot (P(Y=1 \mid B=b, Z \in \tilde{\mathcal{S}}_h) - P(Y=1 \mid B=b', Z \in \tilde{\mathcal{S}}_{h'}) \\
&\quad + u(1,0) \cdot (P(Y=1 \mid B=b', Z \in \tilde{\mathcal{S}}_{h'}) - P(Y=1 \mid B=b, Z \in \tilde{\mathcal{S}}_h)) \\
&\quad \leq (u(1,1) - u(1,0)) \cdot \alpha,
\end{aligned} \tag{36}$$

where the last inequality due to Eq. 35 and the assumption that $u(1,1) - u(1,0) > 0$. Analogously, we can also upper bound the expected utility of $T=0$ given $H=h', B=b'$ and $Z \in \tilde{\mathcal{S}}_{h'}$ as follows:

$$\begin{aligned}
&\mathbb{E}[u(0,Y) \mid B=b, Z \in \tilde{\mathcal{S}}_h] - \mathbb{E}[u(0,Y) \mid B=b', Z \in \tilde{\mathcal{S}}_{h'}] \\
&\quad = u(0,1) \cdot (P(Y=1 \mid B=b, Z \in \tilde{\mathcal{S}}_h) - P(Y=1 \mid B=b', Z \in \tilde{\mathcal{S}}_{h'}) \\
&\quad + u(0,0) \cdot (P(Y=1 \mid B=b', Z \in \tilde{\mathcal{S}}_{h'}) - P(Y=1 \mid B=b, Z \in \tilde{\mathcal{S}}_h)) \\
&\quad \geq (u(0,1) - u(0,0)) \cdot \alpha,
\end{aligned} \tag{37}$$

where the last inequality holds due to Eq. 35 and the assumption that $u(0,1) - u(0,0) < 0$.

Now, as $P(Y=1 \mid B=b, Z \in \tilde{\mathcal{S}}_h) \geq c$, by Lemma 2, we have that

$$\mathbb{E}[u(1,Y) \mid B=b, Z \in \tilde{\mathcal{S}}_h] \geq \mathbb{E}[u(0,Y) \mid B=b, Z \in \tilde{\mathcal{S}}_h] \tag{38}$$

Combining Eqs. 36, 37 and 38, we obtain

$$\begin{aligned}
&\mathbb{E}[u(1,Y) \mid B=b', Z \in \tilde{\mathcal{S}}_{h'}] + \alpha(u(1,1) - u(1,0)) \\
&\quad \geq \mathbb{E}[u(0,Y) \mid B=b', Z \in \tilde{\mathcal{S}}_{h'}] + \alpha(u(0,1) - u(0,0))
\end{aligned} \tag{39}$$

In addition, note that we have following trivial bound for the expectation when $H=h'$ but $Z \notin \tilde{\mathcal{S}}_{h'}$

$$u(1,0) \leq \mathbb{E}[u(1,Y) \mid H=h', B=b'] \leq u(1,1), \tag{40}$$

$$u(0,1) \leq \mathbb{E}[u(0,Y) \mid H=h', B=b'] \leq u(0,0) \tag{41}$$

Moreover, since $b' \geq \hat{b}_{h'}$, the distribution of positive decisions in $\hat{\pi}$ may also increase for h', b' compared to π, *i.e.*,

$$P_\pi(T=1 \mid H=h', B=b') - P_{\hat{\pi}}(T=1 \mid H=h', B=b') \leq 0$$

Hence, we have that

$$\begin{aligned}
&\mathbb{E}_\pi[u(T,Y) \mid H=h', B=b'] - \mathbb{E}_{\hat{\pi}}[u(T,Y) \mid H=h', B=b'] \\
&\quad \leq (-1) \cdot (\mathbb{E}[u(1,Y) \mid H=h', B=b'] - \mathbb{E}[u(0,Y) \mid H=h', B=b']),
\end{aligned} \tag{42}$$

where the inequality follows since $\mathbb{E}[u(1,Y) \mid H=h', B=b'] - \mathbb{E}[u(0,Y) \mid H=h', B=b'] \leq 0$ by Lemma 2 as $P(Y=1 \mid H=h', B=b') < c$.

Finally, combining Eqs. 39, 40, 41 and 42 and using the law of total expectation, we obtain

$$\begin{aligned}
&\mathbb{E}_\pi[u(T,Y) \mid H=h', B=b'] - \mathbb{E}_{\hat{\pi}}[u(T,Y) \mid H=h', B=b'] \\
&\qquad \leq (1-\beta_{(h',b')})(\mathbb{E}[u(0,Y) \mid B=b', Z \in \tilde{\mathcal{S}}_{h'}] - \mathbb{E}[u(1,Y) \mid B=b', Z \in \tilde{\mathcal{S}}_{h'}]) \\
&\qquad + \beta_{(h',b')}(\mathbb{E}[u(0,Y) \mid H=h', B=b'] - \mathbb{E}[u(1,Y) \mid H=h', B=b']) \\
&\qquad \leq (1-\beta_{(h',b')})\alpha(u(1,1)-u(1,0)+u(0,0)-u(0,1)) + \beta_{(h',b')}(u(0,0)-u(1,0)),
\end{aligned} \tag{43}$$

where $\beta_{(h',b')}$ denotes the probability of $Z \notin \tilde{\mathcal{S}}_{h'}$ given $H=h', B=b'$, *i.e.*, $\beta_{(h',b')} = P(Z \notin \tilde{\mathcal{S}}_{h'} | H=h', B=b')$.

— **Case 3**: $b' < \hat{b}_{h'}$.

For all h, b, with $h \leq h'$, $b \leq b'$, we have that $P(Y=1 \mid B=b, Z \in \tilde{\mathcal{S}}_h) < c$. In particular, $P(Y=1 \mid B=b', Z \in \tilde{\mathcal{S}}_{h'}) < c$. Thus, by Lemma 2,

$$\mathbb{E}[u(1,Y) \mid B=b', Z \in \tilde{\mathcal{S}}_{h'}] < \mathbb{E}[u(0,Y) \mid B=b', Z \in \tilde{\mathcal{S}}_{h'}] \tag{44}$$

In this case, since $b' < \hat{b}_{h'}$, the distribution of positive decisions in $\hat{\pi}$ may decrease for h, b compared to π, *i.e.*,

$$0 \leq P_\pi(T=1 \mid H=h, B=b) - P_{\hat{\pi}}(T=1 \mid H=h, B=b)$$

Combining Eqs. 44, 40 and 41 and using the law of total expectation, we obtain

$$\begin{aligned}
&\mathbb{E}_\pi[u(T,Y) \mid H=h', B=b'] - \mathbb{E}_{\hat{\pi}}[u(T,Y) \mid H=h', B=b'] \\
&\qquad \leq (\mathbb{E}[u(1,Y) \mid H=h', B=b'] - \mathbb{E}[u(0,Y) \mid H=h', B=b']) \cdot 1 \\
&\qquad = (1-\beta_{(h',b')})(\mathbb{E}[u(1,Y) \mid B=b', Z \in \tilde{\mathcal{S}}_{h'}] - \mathbb{E}[u(0,Y) \mid B=b', Z \in \tilde{\mathcal{S}}_{h'}]) \\
&\qquad + \beta_{(h',b')}(\mathbb{E}[u(1,Y) \mid H=h', B=b'] - \mathbb{E}_Y[u(0,Y) \mid H=h', B=b']) \\
&\qquad \leq \beta_{(h',b')}(u(1,1)-u(0,1)),
\end{aligned} \tag{45}$$

where again $\beta_{(h',b')} = P(Z \notin \tilde{\mathcal{S}}_{h'} | H=h', B=b')$.

Now, for a fixed $h' \in \mathcal{H}$, since $|\tilde{\mathcal{S}}_{h'}| \geq (1-\alpha/2)|\mathcal{S}_{h'}|$, we know that $0 \leq \sum_{b \in \mathcal{B}} \beta_{(h',b)} \leq \alpha/2$. Hence, combining Eqs. 34, 43 and 45 from the three cases above, we have that

$$\begin{aligned}
&\mathbb{E}_B[\mathbb{E}_\pi[u(T,Y) \mid H=h', B=b']] - \mathbb{E}_B[\mathbb{E}_{\hat{\pi}}[u(T,Y) \mid H=h', B=b']] \\
&\qquad = \mathbb{E}_B[\mathbb{E}_\pi[u(T,Y) \mid H=h', B=b'] - \mathbb{E}_{\hat{\pi}}[u(T,Y) \mid H=h', B=b']] \\
&\leq \max\{\alpha(u(1,1)-u(1,0)+u(0,0)-u(0,1)) + \frac{\alpha}{2}\cdot(u(0,0)-u(1,0)),\ \frac{\alpha}{2}\cdot(u(1,1)-u(0,1))\} \\
&\qquad \leq \alpha \cdot (u(1,1)-u(0,1) + \frac{3}{2}\cdot(u(0,0)-u(1,0))).
\end{aligned}$$

Finally, since by assumption π is optimal, *i.e.*, $\mathbb{E}_\pi[u(T,Y)] = \mathbb{E}_{\pi^*}[u(T,Y)] = \max_{\pi' \in \Pi(H,B)} \mathbb{E}_{\pi'}[u(T,Y)]$, we can conclude by the law of total expectation that

$$\begin{aligned}
\mathbb{E}_{\pi^*}[u(T,Y)] &= \mathbb{E}_H \mathbb{E}_B[\mathbb{E}_{Y,\, T \mid \pi}[u(T,Y) \mid H, B]] \\
&\leq \mathbb{E}_{\hat{\pi}}[u(T,Y)] + \alpha \cdot (u(1,1)-u(0,1) + \frac{3}{2}\cdot(u(0,0)-u(1,0))).
\end{aligned}$$

This concludes the proof.

A.4 Proof of Theorem 8

If f_B is $\alpha/2$-multicalibrated with respect to $\{\mathcal{S}_h\}_{h \in \mathcal{H}}$, then, by definition, for any $h \in \mathcal{H}$, there exists $\tilde{\mathcal{S}}_h \subset \mathcal{S}_h$ with $|\mathcal{S}| \geq (1-\alpha/2)|\mathcal{S}_h|$ such that, for any $b \in [0,1]$, it holds that

$$|P(Y=1 \mid f_B(Z)=b, Z \in \tilde{\mathcal{S}}_h) - b| \leq \alpha/2.$$

This directly implies that, for any $h', h'' \in \mathcal{H}$ and $b', b'' \in [0,1]$, we have that

$$P(Y=1 \mid f_B(Z)=b', Z \in \tilde{\mathcal{S}}_{h'}) - b' - P(Y=1 \mid f_B(Z)=b'', Z \in \tilde{\mathcal{S}}_{h''}) - b'' \leq \alpha \quad (46)$$

and, using linearity of expectation, we further have that

$$P(Y=1 \mid f_B(Z)=b', Z \in \tilde{\mathcal{S}}_{h'}) - P(Y=1 \mid f_B(Z)=b'', Z \in \tilde{\mathcal{S}}_{h''}) \leq \alpha + b' - b'', \quad (47)$$

showing that, whenever $b' \leq b''$, the α-alignment condition is met. This proves that f_B is α-aligned with respect to f_H.

Finally, if f_B is $\alpha/2$-multicalibrated with respect to $\{\mathcal{S}_h\}_{h\in\mathcal{H}}$, then, it is $\alpha/2$-calibrated with respect to any of the sets $\mathcal{S}_h$. Since $\mathcal{Z} = \cup_{h\in\mathcal{H}}\mathcal{S}_h$, this implies that f_B is $\alpha/2$-calibrated with respect to $\mathcal{Z}$. This concludes the proof.

A.5 Proof of Proposition 1

Given a discretization parameter λ, Algorithm 1 works with a discretized notion of α-multicalibration, namely (α,λ)-multicalibration:

Definition 10. *Let $\mathcal{C} \subseteq 2^{\mathcal{Z}}$ be a collection of subsets of $\mathcal{Z}$. For any $\alpha, \lambda > 0$, confidence function $f_B : \mathcal{Z} \to [0,1]$ is (α,λ)-multicalibrated with respect to $\mathcal{C}$ if, for all $S \in \mathcal{C}$, $b \in \Lambda[0,1]$, and all $S_{h,\lambda(b)}(g)$ such that $|\mathcal{S}_{h,\lambda(b)}| \geq \alpha\lambda|\mathcal{S}_h|$, it holds that*

$$|\mathbb{E}[f_B(X,H) - P(Y=1 \mid X,H) \mid (X,H) \in \mathcal{S}_{h,\lambda(b)}]| \leq \alpha\,. \quad (48)$$

Here, we can analogously define a discretized notion of α-alignment, namely (α,λ)-alignment.

Definition 11. *For $\alpha, \lambda > 0$, a confidence function $f_B : \mathcal{Z} \to [0,1]$ is (α,λ)-aligned with respect to f_H if, for all $h', h'' \in \mathcal{H}$, $h' \leq h''$, and all $b', b'' \in \Lambda[0,1]$, $b' \leq b''$, with $|\mathcal{S}_{h',\lambda(b')}| > \alpha/2 \cdot \lambda|S_{h'}|$ and $|\mathcal{S}_{h'',\lambda(b'')}| > \alpha/2 \cdot \lambda|S_{h''}|$, we have*

$$P(Y=1 \mid (X,H) \in \mathcal{S}_{h',\lambda(b')}) - P(Y=1 \mid (X,H) \in \mathcal{S}_{h'',\lambda(b'')}) \leq \alpha\,. \quad (49)$$

In what follows, we first show that (α,λ)-multicalibration with respect to $\{\mathcal{S}_h\}_{h\in\mathcal{H}}$ implies $(2\alpha+\lambda,\lambda)$-alignment with respect to f_H.

Theorem 12. *For $\alpha, \lambda > 0$, if f_B is (α,λ)-multicalibrated with respect to $\{\mathcal{S}_h\}_{h\in\mathcal{H}}$, then f_B is $(2\alpha+\lambda,\lambda)$-aligned with respect to f_H .*

Proof. If f_B is (α,λ)-multicalibrated with respect to $\{S_h\}_{h\in\mathcal{H}}$, then, by definition, for all $h \in \mathcal{H}$, $b \in \Lambda[0,1]$, and all $\mathcal{S}_{h,\lambda(b)}$ such that $|\mathcal{S}_{h,\lambda(b)}| \geq \alpha \cdot \lambda|\mathcal{S}_h|$, it holds that

$$|\mathbb{E}[f_B(X,H) - P(Y=1 \mid X,H) \mid (X,H) \in \mathcal{S}_{h,\lambda(b)}]| \leq \alpha. \quad (50)$$

This directly implies that, for all $h', h'' \in \mathcal{H}, b', b'' \in \Lambda[0,1]$ with $|\mathcal{S}_{h',\lambda(b')}| \geq \alpha \cdot \lambda|S_{h'}|$ and $|\mathcal{S}_{h'',\lambda(b'')}| \geq \alpha \cdot \lambda|S_{h''}|$, it holds that

$$\begin{aligned}&\mathbb{E}[f_B(X,H) - P(Y=1 \mid X,H) \mid (X,H) \in \mathcal{S}_{h'',\lambda(b'')}] \\ &\quad - \mathbb{E}[f_B(X,H) - P(Y=1 \mid X,H) \mid (X,H) \in \mathcal{S}_{h',\lambda(b')}] \leq 2\alpha\end{aligned} \quad (51)$$

and, using the linearity of expectation, we have that

$$\begin{aligned}&P(Y=1 \mid (X,H) \in \mathcal{S}_{h',\lambda(b')}) - P(Y=1 \mid (X,H) \in \mathcal{S}_{h'',\lambda(b'')}) \\ \leq& 2\alpha + \mathbb{E}[f_B(X,H) \mid (X,H) \in \mathcal{S}_{h',\lambda(b')}] - \mathbb{E}[f_B(X,H) \mid (X,H) \in \mathcal{S}_{h'',\lambda(b'')}].\end{aligned} \quad (52)$$

Whenever $b' \leq b''$, due to the λ-discretization, we have that

$$\mathbb{E}[f_B(X,H) \mid (X,H) \in \mathcal{S}_{h',\lambda(b')}] - \mathbb{E}[f_B(X,H) \mid (X,H) \in \mathcal{S}_{h'',\lambda(b'')}] \leq \lambda \quad (53)$$

Hence, we have shown that if f_B is α-multicalibrated, then for all $h', h'' \in \mathcal{H}, b', b'' \in \Lambda[0,1]$ with $|\mathcal{S}_{h',\lambda(b')}| \geq \alpha \cdot \lambda|S_{h'}|$ and $|\mathcal{S}_{h'',\lambda(b'')}| \geq \alpha \cdot \lambda|S_{h''}|$, we have

$$P(Y=1 \mid (X,H) \in \mathcal{S}_{h',\lambda(b')}) - P(Y=1 \mid (X,H) \in \mathcal{S}_{h'',\lambda(b'')}) \leq 2\alpha + \lambda\,. \quad (54)$$

Further, note that $(2\alpha+\lambda)/2 \cdot \lambda > \alpha \cdot \lambda$ as $\lambda > 0$. This concludes the proof. □

Next, we show that, if f_B is (α, λ)-aligned, then $f_{B,\lambda}$ is α-aligned with respect to f_H.

Theorem 13. *For $\alpha, \lambda > 0$, if f_B is (α, λ)-aligned with respect to f_H, then $f_{B,\lambda}$ is α-aligned with respect to f_H.*

Proof. The proof is similar to the proof of Lemma 1 in Hébert-Johnson et al. [11]. Consider all $\mathcal{S}_{h,\lambda(b)}$ such that $|\mathcal{S}_{h,\lambda(b)}| < \alpha\lambda|\mathcal{S}_h|$. By the λ-discretization, there are at most $1/\lambda$ such sets, thus, the cardinality of their union is at most $1/\lambda\alpha\lambda|\mathcal{S}_h| = \alpha|\mathcal{S}_h|$. Hence, for all $h \in \mathcal{H}$, there exists a subset $\tilde{\mathcal{S}}_h \subset \mathcal{S}_h$ with $|\tilde{\mathcal{S}}_h| \geq (1-\alpha)|\mathcal{S}_h|$ such that, for all $h', h'' \in \mathcal{H}$, with $h' \leq h''$, and all $b', b'' \in \Lambda[0,1]$, with $b' \leq b''$, it holds that

$$P(Y = 1 \mid (X,H) \in \mathcal{S}_{h',\lambda(b')} \cap \tilde{\mathcal{S}}_{h'}) - P(Y = 1 \mid (X,H) \in \mathcal{S}_{h'',\lambda(b'')} \cap \tilde{\mathcal{S}}_{h''}) \leq \alpha\,. \tag{55}$$

The λ-discretization sets all values of $(x,h) \in \mathcal{S}_{h',\lambda(b')}$ to $f_{B,\lambda}(x,h) = \mathbb{E}[f_B(X,H) \mid f_B(X,H) \in \lambda(b')]$. Note that, for $(x,h) \in \mathcal{S}_{h',\lambda(b')}$, $f_{B,\lambda}(x,h) \in \lambda(b')$ and for $(x,h) \in \mathcal{S}_{h'',\lambda(b'')}$, $f_{B,\lambda}(x,h) \in \lambda(b'')$, so it still holds that $\mathbb{E}[f_B(X,H) \mid f_B(X,H) \in \lambda(b')] \leq \mathbb{E}[f_B(X,H) \mid f_B(X,H) \in \lambda(b'')]$. Thus, using Eq. 55, we have that

$$\begin{aligned} &P(Y = 1 \mid f_B(X,H) = \mathbb{E}[f_B(X,H) \mid (X,H) \in \lambda(b')], (X,H) \in \tilde{\mathcal{S}}_{h'}) \\ &- P(Y = 1 \mid f_B(X,H) = \mathbb{E}[f_B(X,H) \mid (X,H) \in \lambda(b'')], (X,H) \in \tilde{\mathcal{S}}_{h''}) \leq \alpha \end{aligned} \tag{56}$$

This concludes the proof. □

Finally, using Theorems 12 and 13, it readily follows that, given a parameter α', the discretized confidence function $f_{B,\lambda}$ returned by Algorithm 1 satisfies $(2\alpha' + \lambda)$-aligned calibration with respect to f_H.

A.6 Proof Theorem 9

We structure the proof in three parts. We first explain the calibration guarantee that UMD provides and how it relates to human-aligned calibration. Then, we derive a lower bound on the size of the subsets $\mathcal{D} \cap \mathcal{S}_h$ so that the discretized confidence function $f_{B,\lambda}$ satisfies α-aligned calibration with respect to f_H with high probability. Finally, building on this result, we derive an upper bound on $|\mathcal{D}|$ so that $f_{B,\lambda}$ satisfies α-aligned calibration with high probability as long as there exists $\gamma > 0$ so that $P((X,H) \in \mathcal{S}_h) \geq \gamma$ for all $h \in \mathcal{H}$.

Conditional Calibration implies Human-Aligned Calibration. Running UMD on a dataset $\mathcal{D} \in (\mathcal{Z} \times \mathcal{Y})^n$, where each datapoint is sampled from $P^{\mathcal{M}}$, guarantees (α, ξ)-conditional calibration, a PAC-style calibration guarantee [12]. Given a dataset $\mathcal{D}$, a confidence function f_B satisfies (α, ξ)-conditional calibration if, with probability at least $1 - \xi$ over the randomness in $\mathcal{D}$,

$$\forall b \in [0,1], \quad |P(Y = 1 | f_B(X,H) = b) - b| \leq \alpha\,.$$

This stands in contrast to the definition of α-calibration, which requires only that the confidence $f_B(X,H)$ is at most α away from the true probability for $1 - \alpha$ fraction of $\mathcal{Z}$.

Similarly, using an union bound over all $h \in \mathcal{H}$, $(\alpha/2, \xi/|\mathcal{H}|)$-conditional calibration of f_B on each $\mathcal{S}_h$, $h \in \mathcal{H}$, implies that, with probability at least $1 - \xi$ over the randomness in $\mathcal{D}$, f_B satisfies that

$$\forall h \in \mathcal{H}, \quad \forall b \in [0,1], \quad |P(Y = 1 | f_B(X,H) = b, H = h) - b| \leq \alpha/2\,. \tag{57}$$

Hence, analogously to the proof of Theorem 8, this implies that, with probability at least $1 - \xi$ over the randomness in $\mathcal{D}$, f_B also satisfies that

$$\begin{aligned} &\forall h, h' \in \mathcal{H}, h \leq h', \ \ \forall b, b' \in \mathcal{G}, b \leq b', \\ &\qquad P(Y = 1 | f_B(X,H) = b, H = h) - P(Y = 1 | f_B(X,H) = b', H = h') \leq \alpha\,. \end{aligned} \tag{58}$$

In summary, from Eqs. 57 and 58, we can conclude that $(\alpha/2, \xi/|\mathcal{H}|)$-conditional calibration of f_B on each $\mathcal{S}_h$, $h \in \mathcal{H}$, implies that, with probability at least $1 - \xi$, f_B satisfies α-aligned calibration, where, for all $h \in \mathcal{H}$, we have that $\tilde{\mathcal{S}}_h = \mathcal{S}_h$.

Lower bound on $|\mathcal{D} \cap \mathcal{S}_h|$ to achieve conditional calibration with UMD. Running UMD on each partition $\mathcal{D} \cap \mathcal{S}_h$ of $\mathcal{D}$ induced by $h \in \mathcal{H}$ achieves $(\alpha/2, \xi/|\mathcal{H}|)$-conditional calibration as long as each subset $\mathcal{D} \cap \mathcal{S}_h$ of the data is large enough. More specifically, the following lower bound on the size of the subsets $\mathcal{D} \cap \mathcal{S}_h$ readily follows from Theorem 3 in Gupta et al. [12].

Lemma 4. *The discretized confidence function $f_{B,\lambda}$ returned by $|\mathcal{H}|$ instances of UMD, one per $\mathcal{S}_h$, is $(\alpha/2, \xi/|\mathcal{H}|)$-conditional calibrated on $\mathcal{S}_h$ for any $\xi \in (0,1)$ if*

$$|\mathcal{D} \cap \mathcal{S}_h| \geq n_{min} := \left(\frac{2 \log\left(\frac{2|\mathcal{H}|}{\xi} \cdot \left\lceil \frac{1}{\lambda} \right\rceil \right)}{\alpha^2} + 2 \right) \cdot \left\lceil \frac{1}{\lambda} \right\rceil \tag{59}$$

Proof. Let B denote the number of bins in UMD. Theorem 3 in Gupta et al. [12] states that, if $f_B(X, H)$ is absolutely continuous with respect to the Lebesgue measure[11] and $|\mathcal{D} \cap \mathcal{S}_h| \geq 2B$, then the discretized confidence function output by UMD is (ϵ, ξ')-conditionally calibrated for any $\xi' \in (0,1)$ and

$$\epsilon = \sqrt{\frac{\log(2B/\xi')}{2(\lfloor |\mathcal{D} \cap \mathcal{S}_h|/B \rfloor - 1)}}. \tag{60}$$

Then, for a given α, setting $\epsilon = \alpha/2$, $B = \lceil 1/\lambda \rceil$ and $\xi' = \xi/|\mathcal{H}|$, we can solve Eq. 60 for the lower bound on $|\mathcal{D} \cap \mathcal{S}_h| \geq n_{\text{min}}$ with n_{min} as defined in Eq. 59. □

Upper bound on $|\mathcal{D}|$ to achieve conditional calibration with UMD. Suppose $P((X,H) \in \mathcal{S}_h) \geq \gamma$ for all $h \in \mathcal{H}$. When $|\mathcal{H}| \geq 2$, we give an upper bound on $|\mathcal{D}|$ so that with high probability $|\mathcal{D} \cap \mathcal{S}_h| \geq n_{\text{min}}$ for all $h \in \mathcal{H}$.

In the process of sampling $\mathcal{D} \in (\mathcal{Z} \times \mathcal{Y})^n$ from $P^{\mathcal{M}}$, let $R_i^{(h)} = 1$ denote the event that the i-th datapoint (x_i, h_i, y_i) has confidence value h, *i.e.*, $h_i = h$. Then, we can express $|\mathcal{D} \cap \mathcal{S}_h|$ in terms of random variable $R^{(h)}$, defined as

$$R^{(h)} = \sum_{i=1}^{|\mathcal{D}|} R_i^{(h)}. \tag{61}$$

Since $R_i^{(h)}$ is a Bernoulli-distributed variable with $P(R_i^{(h)}) = P((X,H) \in \mathcal{S}_h)$, the expected value of $R^{(h)}$ is $\mu(h) := \mathbb{E}[R^{(h)}] = P((X,H) \in \mathcal{S}_h) \cdot |\mathcal{D}| \geq \gamma \cdot |\mathcal{D}|$.

Let $|\mathcal{D}| = 2 \cdot |\mathcal{H}| \cdot \log(2/\xi) \cdot 1/\gamma \cdot n_{\text{min}}$, observe that in this case

$$P(R^{(h)} \leq n_{\text{min}}) = P\left(R^{(h)} \leq \frac{\gamma}{2|\mathcal{H}| \cdot \log(2/\xi)} \cdot |\mathcal{D}| \right).$$

For $|\mathcal{H}| \geq 2$ and $\xi \in (0,1)$, we have $1/(2|\mathcal{H}| \cdot \log(2/\xi)) \in (0,1)$ and we can use a variation of the Chernoff bound to show

$$\begin{aligned}
P(R^{(h)} \leq n_{\text{min}}) &\leq P\left(R^{(h)} \leq \frac{1}{2|\mathcal{H}| \cdot \log(2/\xi)} \cdot \mu(h) \right) \\
&\leq e^{-\mu(h)\left(\frac{2|\mathcal{H}| \cdot \log(2/\xi) - 1}{2|\mathcal{H}| \cdot \log(2/\xi)}\right)^2 \cdot \frac{1}{2}} \\
&= e^{-\mu(h) \cdot \frac{1}{2} \cdot \left(1 - \frac{1}{|\mathcal{H}| \cdot \log(2/\xi)} + \frac{1}{(2|\mathcal{H}| \cdot \log(2/\xi))^2}\right)} \\
&\leq \frac{\xi}{2} \cdot e^{-|\mathcal{H}| \cdot n_{\text{min}} \cdot \left(\frac{1}{2} - \frac{1}{2|\mathcal{H}| \cdot \log(2/\xi)} + \frac{1}{2(2|\mathcal{H}| \cdot \log(2/\xi))^2}\right)},
\end{aligned}$$

where the first and last inequality results from using $\mu(h) > \gamma \cdot |\mathcal{D}|$. We can now use a union bound to obtain a lower bound on the probability that for any $h \in \mathcal{H}$, $|\mathcal{D} \cap \mathcal{S}_h| \leq n_{\text{min}}$, *i.e.*,

$$P(\exists h \in \mathcal{H} :\ |\mathcal{D} \cap \mathcal{S}_h| \leq n_{\text{min}}) \leq \frac{\xi}{2} \cdot |\mathcal{H}| \cdot e^{-|\mathcal{H}| \cdot n_{\text{min}} \cdot \left(\frac{1}{2} - \frac{1}{2|\mathcal{H}| \cdot \log(2/\xi)} + \frac{1}{2(2|\mathcal{H}| \cdot \log(2/\xi))^2}\right)} \tag{62}$$

One can verify that for $|\mathcal{H}| \geq 2$ and $n_{\text{min}} \geq 1$, we have $P(\exists h \in \mathcal{H} :\ |\mathcal{D} \cap \mathcal{S}_h| \leq n_{\text{min}}) \leq \frac{\xi}{2}$. Hence, if $|\mathcal{D}| = 2 \cdot |\mathcal{H}| \cdot \log(2/\xi) \cdot 1/\gamma \cdot n_{\text{min}}$, then, for all $h \in \mathcal{H}$, $|\mathcal{D} \cap \mathcal{S}_h| \leq n_{\text{min}}$ with probability $1 - \xi/2$.

[11] If f_B is not continuous with respect to the Lebesgue measure (or equivalently put, f_B does not have a probability density function), a randomization trick can be used to ensure that the results of the theorem hold.

Combining this result and Lemma 4, we have that the discretized confidence function $f_{B,\lambda}$ returned by $|\mathcal{H}|$ instances of UMD, one per $\mathcal{S}_h$, is $(\alpha/2, \xi/(2|\mathcal{H}|))$-conditional calibrated on each $\mathcal{S}_h$ with probability at least $1 - \xi/2$ for any $\xi \in (0, 1)$ if

$$|\mathcal{D}| = 2 \cdot |\mathcal{H}| \cdot \frac{\log(2/\xi)}{\gamma} \cdot \left(\frac{2 \log\left(\frac{4|\mathcal{H}|}{\xi} \cdot \left\lceil \frac{1}{\lambda} \right\rceil \right)}{\alpha^2} + 2 \right) \cdot \left\lceil \frac{1}{\lambda} \right\rceil \tag{63}$$

Finally, using a union bound, we can conclude that $f_{B,\lambda}$ achieves α-aligned calibration with respect to f_H with probability at least $1 - \xi$ from

$$|\mathcal{D}| = O\left(|\mathcal{H}| \cdot \frac{\log(|\mathcal{H}|/\xi\lambda)}{\alpha^2 \cdot \lambda \cdot \gamma} \right)$$

samples. This concludes the proof.

B Multicalibration Algorithm

In this section, we give a high-level description of the post-processing algorithm for multicalibration introduced by Hébert-Johnson et al. [11]. The algorithm works with a discretization of $[0, 1]$ into uniform sized bins of size λ, for a $\lambda > 0$. Formally the λ-discretization of $[0, 1]$, is defined as

Definition 14 (λ-discretization [11])**.** *Let $\lambda > 0$. The λ-discretization of $[0, 1]$, denoted by $\Lambda[0, 1] = \{\frac{\lambda}{2}, \frac{3\lambda}{2}, \ldots, 1 - \frac{\lambda}{2}\}$, is the set of $1/\lambda$ evenly spaced real values over $[0, 1]$. For $b \in \Lambda[0, 1]$, let*

$$\lambda(b) = [b - \lambda/2, v + \lambda/2) \tag{64}$$

be the λ-interval centered around b (except for the final interval, which will be $[1 - \lambda, 1]$).

It starts by partitioning each subspace $\mathcal{S}_h$ into $1/\lambda$ groups $\mathcal{S}_{h,\lambda(b)} = \{(x, h) \in \mathcal{S}_h \mid f_B(x, h) \in \lambda(b)\}$, with $b \in \Lambda[0, 1]$. Then, it repeatedly looks for a large enough group $\mathcal{S}_{h,\lambda(b)}$ such that the absolute difference between the average confidence value $\mathbb{E}[f_B(X, H) \mid (X, H) \in \mathcal{S}_{h,\lambda(b)}]$ and the probability $P(Y = 1 \mid (X, H) \in \mathcal{S}_{h,\lambda(b)})$ is larger than α and, if it finds it, it updates the confidence value $f_B(x, h)$ of each $(x, h) \in \mathcal{S}_{h,\lambda(b)}$ by this difference. Once the algorithm cannot find any more such a group, it returns a discretized confidence function $f_{B,\lambda}(x, h) = \mathbb{E}[f_B(X, H) \mid f_B(X, H) \in \lambda(b)]$, with $b \in \Lambda[0, 1]$ such that $f_B(x, h) \in \lambda(b)$, which is guaranteed to satisfy $(\alpha + \lambda)$-multicalibration.

Algorithm 1 provides a pseudocode implementation of the overall algorithm. Within the implementation, it is worth noting that the expectations and probabilities can be estimated with fresh samples from the distribution or from a fixed dataset using tools from differential privacy and adaptive data analysis, as discussed in Hébert-Johnson et al. [11].

Algorithm 1 Post-processing algorithm for $(\alpha + \lambda)$-multicalibration

Input: confidence function f_B, parameters $\alpha, \lambda > 0$
Output: confidence function $f_{B,\lambda}$
repeat
  updated $\leftarrow$ `false`
  for $\mathcal{S}_h \in \mathcal{C}$ & $b \in \Lambda[0, 1]$ **do**
    $\mathcal{S}_{h,\lambda(b)} \leftarrow \mathcal{S}_h \cap \{(x, h) \in \mathcal{Z} \mid f_B(x, h) \in \lambda(b)\}$
    if $P((X, H) \in \mathcal{S}_{h,\lambda(b)}) < \alpha\lambda \cdot P((X, H) \in \mathcal{S}_h)$ **then**
      continue
    $\bar{b}_{h,\lambda(b)} \leftarrow \mathbb{E}[f_B(X, H) \mid (X, H) \in \mathcal{S}_{h,\lambda(b)}]$
    $r_{h,\lambda(b)} \leftarrow P(Y = 1 \mid (X, H) \in \mathcal{S}_{h,\lambda(b)})$
    if $|r_{h,\lambda(b)} - \bar{b}_{h,\lambda(b)}| > \alpha$ **then**
      updated $\leftarrow$ `true`
      for $(x, h) \in \mathcal{S}_{h,\lambda(b)}$ **do**
        $f_B(x, h) \leftarrow f_B(x, h) + (r_{h,\lambda(b)} - \bar{b}_{h,\lambda(b)})$ {project into $[0, 1]$ if necessary}
until updated $=$ `false`
for $b \in \Lambda[0, 1]$ **do**
  $\bar{b}_{\lambda(b)} \leftarrow \mathbb{E}[f_B(X, H) | f_B(X, H) \in \lambda(b)]$
  for $(x, h) \in \mathcal{Z} : f_B(x, h) \in \lambda(b)$ **do**
    $f_{B,\lambda}(x, h) \leftarrow \bar{b}_{\lambda(b)}$
return $f_{B,\lambda}$

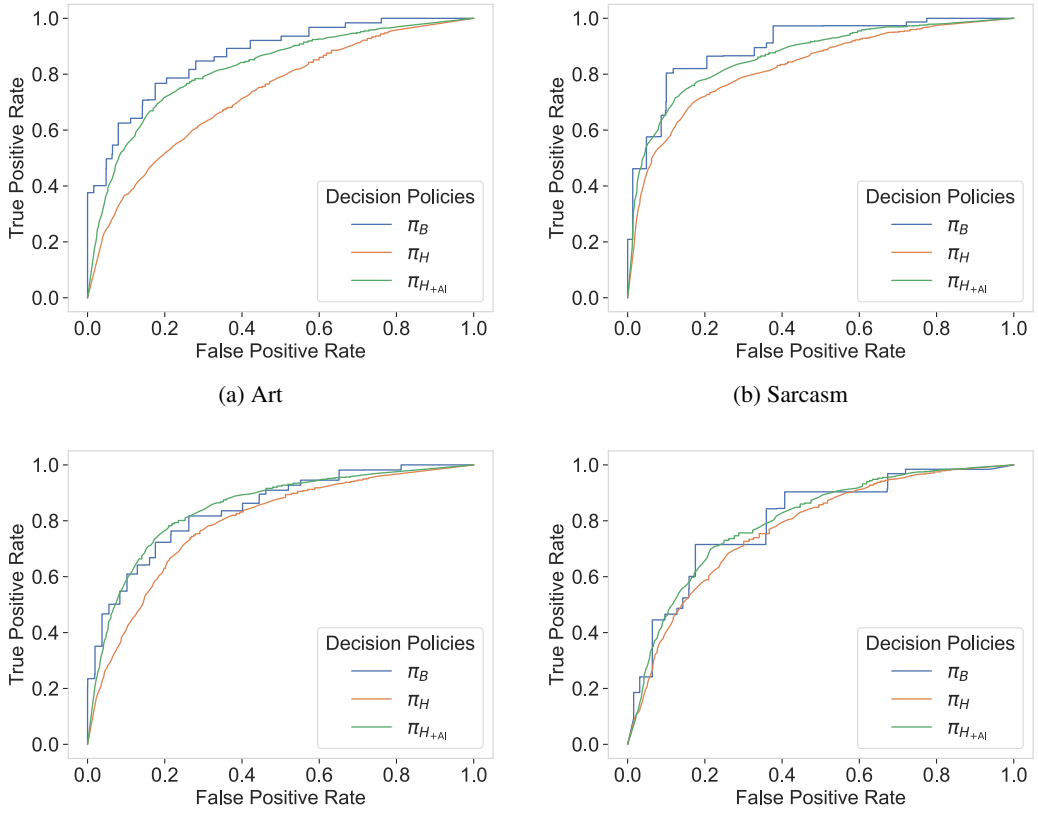
C Additional Details about the Experiments

Transformation of confidence values. The confidence values in the Human-AI Interactions dataset were originally recorded on a scale of $[-1, 1]$, where 1 means complete certainty on the correct true label and -1 means complete certainty on the incorrect label. To better match our theoretical framework, we transform all confidence values to a scale of $[0, 1]$, where 1 means complete certainty that the true label $y = 1$ and 0 means complete certainty that the true label is $y \neq 1$. More formally, let $\hat{b}, \hat{h}, \hat{h}_{\text{+AI}} \in [-1, 1]$ be the original confidence values in the dataset, then we obtain $b \in [0, 1]$ via the following transformation:

$$b = \begin{cases} (\hat{b} + 1)/2 & \text{if } y = 1 \\ 1 - (\hat{b} + 1)/2 & \text{if } y = 0, \end{cases}$$

and analogously for h and $h_{\text{+AI}}$.

Comparing decision policies π_B, π_H and $\pi_{H_{\text{+AI}}}$. Figure 6 shows the ROC curves for the decision policies π_B, π_H and $\pi_{H_{\text{+AI}}}$ in each of the four tasks in the Human-AI Interactions dataset.



(a) Art (b) Sarcasm

(c) Cities (d) Census

Figure 6: ROC curves for the decision policies π_B, π_H and $\pi_{H_{\text{+AI}}}$.