A No Free Lunch Theorem for Human-AI Collaboration

Kenny Peng[1], **Nikhil Garg**[1], **Jon Kleinberg**[2]
[1]Cornell Tech
[2]Cornell University
klp98@cornell.edu, ngarg@cornell.edu, kleinberg@cornell.edu

Abstract

The gold standard in human-AI collaboration is *complementarity*—when combined performance exceeds both the human and algorithm alone. We investigate this challenge in binary classification settings where the goal is to maximize 0-1 accuracy. Given two or more agents who can make calibrated probabilistic predictions, we show a "No Free Lunch"-style result. Any deterministic collaboration strategy (a function mapping calibrated probabilities into binary classifications) that does not essentially always defer to the same agent will sometimes perform worse than the *least accurate* agent. In other words, complementarity cannot be achieved "for free." The result does suggest one model of collaboration with guarantees, where one agent identifies "obvious" errors of the other agent. We also use the result to understand the necessary conditions enabling the success of other collaboration techniques, providing guidance to human-AI collaboration.

1 Introduction

Many important decisions depend—in large part—on prediction. Doctors decide if a patient should undergo a procedure by predicting if the operation will succeed. Judges decide whether or not to grant bail by predicting if the defendant will reoffend. Loan officials decide whether or not to offer a loan by predicting if the loan will be repaid. In all of these tasks, algorithmic predictions are now commonly incorporated into the decision-making process. Still, humans remain a central part of each of these settings, and often have the final say. The hope of human-AI collaboration is that a human and an algorithm can leverage their unique strengths to make predictions that are more accurate than either alone. This standard has been called *complementarity*.

The present work investigates the conditions under which complementarity can be guaranteed when making binary classifications, where the goal is to maximize 0-1 accuracy (minimize the number of misclassifications). We consider a setup in which two or more agents can make calibrated probabilistic predictions on a shared task. These predictions may differ—for example, due to differences in the information available to each agent. Each agent can use their probabilistic predictions to make binary classifications, each achieving some level of accuracy. We ask: Is there a way to combine each agent's calibrated predictions to produce binary classifications that are guaranteed to be at least as accurate, and sometimes *more accurate*, than every individual agent? In other words, if agents are calibrated, can we achieve complementarity "for free"?

Our main result answers this question, mostly in the negative. In fact, Theorem 1 shows that it is difficult to ensure a much lower bar: producing binary classifications that are always at least as accurate as *the worst* individual agent. As we will show—except under very narrow circumstances—the only way to guarantee that a collaboration does not perform worse than the worst agent is to always defer to a single agent (in which case the standard is trivially met). In other words, substantive collaboration in our setup must at times come at a significant cost. There is no "free lunch."

Let us be more concrete, if still somewhat informal. Consider n agents, who, given an input x, produce calibrated probabilistic predictions $P_1(x), P_2(x), \cdots, P_n(x) \in [0,1]$. By calibrated, we mean that among inputs for which an agent predicts a positive label with probability p, the true proportion of positive labels is in fact p. For agent k, given their calibrated prediction $P_k(x)$, the optimal 0-1 classification to maximize accuracy is obtained by a threshold rule: predict 1 if $P_k(x) > 0.5$ and 0 if $P_k(x) < 0.5$. Using $\lfloor \cdot \rceil$ to denote the rounding operator, the optimal classification is $\lfloor P_k(x) \rceil$. In this way, each individual agent can achieve some level of accuracy on their own. A collaboration strategy is a way to combine the calibrated predictions $P_1(x), P_2(x), \cdots, P_n(x)$ into a 0-1 classification. Therefore, a collaboration strategy is defined as a function $\mathcal{C} : [0,1]^n \to \{0,1\}$. There are a number of intuitive collaboration strategies: one approach is to round the average of predicted probabilities; another is to take a majority vote of each agent's classifications; yet another is to defer to the classification of the most confident agent (i.e., the agent whose probabilistic prediction is furthest from $\frac{1}{2}$). Any given collaboration strategy also achieves an accuracy, which can be compared to the accuracy of individual agents.

We call a collaboration strategy $\mathcal{C}$ **reliable** if it is always at least as accurate as the *least accurate* agent. We call $\mathcal{C}$ **non-collaborative** if there exists $k \in [n]$ such that for all $(p_1, p_2, \cdots, p_n) \in (0,1)^n$, $\mathcal{C}(p_1, p_2, \cdots, p_n) = \lfloor p_k \rceil$. In other words, $\mathcal{C}$ is non-collaborative if and only if it always defers to the same agent, except in the special case when an-

other agent is certain in their prediction (i.e., predicts exactly 0 or 1). We may then state our main result.

Theorem 1. *Every reliable collaboration strategy is non-collaborative.*

Consequently, essentially any collaboration strategy that can sometimes achieve complementarity must at other times perform worse than *all* agents. For example, it shows that none of the collaboration strategies described above—averaging probabilities, majority vote, deferring to the most confident agent—are reliable; each of these collaboration strategies sometimes perform worse than the least accurate agent. Theorem 1 is not, however, a statement about these particular collaboration strategies—rather, it applies to the entire space of collaboration strategies $\mathcal{C} : [0,1]^n \to \{0,1\}$. In this way, Theorem 1 is a "No Free Lunch"-style result (Wolpert and Macready 1997). Wolpert and Macready showed that in optimization problems, "if an algorithm performs well on a certain class of problems then it necessarily pays for that with degraded performance on the set of all remaining problems." Like Wolpert and Macready's classical result, our result implies that further structure in the collaboration setting must be leveraged or assumed to obtain guarantees. (There is also a sense in which our result is reminiscent of Arrow's Impossibility Theorem (Arrow 1950), which shows that the only way to aggregate votes while satisfying certain axioms is by deferring to a single dictator. However, our results fundamentally differ, in that Arrow focuses on aggregating ranked preference lists, whereas we focus on aggregating continuous predictions.)

Note, however, that Theorem 1 leaves a small window for collaboration: if a (calibrated) agent is certain, it is clearly possible (and optimal) to defer to that agent. This points to a possible model of successful human-AI collaboration, in which one agent is only in charge of identifying "obvious" errors of the other agent. In contrast, it is not sufficient, however, for an agent to be very confident in their prediction (i.e., predicting close to probability 0 or 1). The agent must be certain. At a high level, the reason is because an agent's predicted probability is not (generally) calibrated after conditioning on the other agents' predictions. The exception is when an agent is fully certain, in which case no outside information can alter this certainty.

Many results in computer science and economics—experts, ensemble prediction methods, Condorcet's Jury Theorem, and recent human-AI collaboration techniques—demonstrate that collaboration in prediction is often possible when further structure is imposed. Theorem 1 can be used to shed light on what conditions enable these approaches to succeed. In Section 3, we survey this literature, identifying two conditions that distinguish these approaches from our setting, both of which enables success: independence in predictions, or (learned) knowledge of the joint distribution of agent predictions and outcomes. Neither condition is satisfied in the setup of Theorem 1. Our result thus suggests that conditions like these are necessary to ensure successful collaboration.

In particular, our setup is reminiscent of typical implementations of human-AI collaboration where humans are shown a probabilistic algorithmic prediction to incorporate into their decision (e.g., Shin, Han, and Rhee (2021); Cabrera, Perer, and Hong (2023)). In these situations, neither independence nor knowledge of the joint distribution is guaranteed. Existing empirical evidence suggests that these kinds of implementations do not typically result in complementarity, even in laboratory settings (e.g., Green and Chen (2019); Lai and Tan (2019); Kiani et al. (2020)). (Vaccaro, Almaatouq, and Malone 2024) provide a meta-analysis demonstrating that most studies of human-AI collaboration do not document complementarity. It should be noted that in real-world settings, human behavioral biases and cognitive limitations further complicate human-AI collaboration (Buçinca, Malaya, and Gajos 2021; Bhatt et al. 2021; Bondi et al. 2022). Theorem 1 suggests that even should humans not suffer from these limitations, there remain fundamental challenges in combining expertise in prediction problems.

To prove Theorem 1 (which we do in Section 4), it suffices to show that for any collaboration strategy $\mathcal{C} : [0,1]^n \to \{0,1\}$ that is not non-collaborative, it is possible to construct a setting in which $\mathcal{C}$ is less accurate than all individual agents. The high-level approach is to use the violation of the non-collaborative condition to construct a set of inputs on which we know the collaboration strategy's behavior: specifically, for every $k \in [n]$, there must exist $(p_1, p_2, \cdots, p_n)$ that $\mathcal{C}$ classifies differently than k. We use this to construct a setting in which $\mathcal{C}$ performs at least as bad as every agent and strictly worse than k. After doing this for each k, we can then "glue together" the resulting settings to obtain the desired adversarial example. Intuitively, our result and proof leverage the insight that what matters for collaboration is the *joint* distribution of agent predictions: conditional on the predictions of other agents, when is an agent accurate? Despite calibration implying that each agent has a strong sense of when they are more or less confident, calibration alone does not reveal sufficient information about interactions between agent predictions—more is needed to know when to defer to one agent or another.

2 A No Free Lunch Theorem

A binary classification problem can be represented as a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ where $\mathcal{X}$ is the input space. A **classifier** is a function $\hat{Y} : \mathcal{X} \to \{0,1\}$. The 0-1 **accuracy** of a classifier is given by

$$\mathbb{E}_{(X,Y)\sim\mathcal{D}} |\hat{Y}(X) - Y|. \tag{1}$$

A **predictor** is a function $P : \mathcal{X} \to [0,1]$. A predictor P is **calibrated** on $\mathcal{D}$ if

$$\Pr_{(X,Y)\sim\mathcal{D}}[Y = 1 \mid P(X) = p] = p \tag{2}$$

for all $p \in [0,1]$ (more precisely, for all $p \in \text{Image}(P)$).

We may now define collaboration settings.

Definition 1. *A **collaboration setting** is an ordered tuple*

$$S = (\mathcal{D}, P_1, \cdots, P_n), \tag{3}$$

where $\mathcal{D}$ is a probability distribution over $\mathcal{X} \times \{0,1\}$ and $P_1, \cdots, P_n$ are each calibrated predictors on $\mathcal{D}$.

The predictor P_i induces the classifier $\hat{Y}_i : x \mapsto \lfloor P_i(x) \rceil$, where $\lfloor \cdot \rceil$ is the rounding operator (without loss of generality, set $\lfloor 0.5 \rceil = 1$). $\hat{Y}_i$ achieves 0-1 accuracy

$$\mathrm{acc}_i(S) := \mathbb{E}_{(X,Y)\sim\mathcal{D}}[\max\{P_i(X), 1 - P_i(X)\}]. \quad (4)$$

A collaboration strategy is a way to combine the predicted probabilities of each agent to make a classification.

Definition 2. *A **collaboration strategy** is a deterministic function* $\mathcal{C} : [0,1]^n \to \{0,1\}$.

Here, $\mathcal{C}$ should be interpreted as a function that takes in n predicted probabilities and returns a 0-1 classification. Given a collaboration setting $S = (\mathcal{D}, P_1, \cdots, P_n)$, a collaboration strategy $\mathcal{C}$ induces the classifier

$$\hat{Y}_{\mathcal{C}} : \mathcal{X} \to \{0,1\}, \quad x \mapsto \mathcal{C}(P_1(x), \cdots, P_n(x)). \quad (5)$$

Let $\mathrm{acc}_{\mathcal{C}}(S)$ denote the 0-1 accuracy of $\hat{Y}_{\mathcal{C}}$ on $\mathcal{D}$. In other words, for each $x \in \mathcal{X}$, each agent i produces a predicted probability $P_i(x)$. The collaboration strategy aggregates these predictions into a binary classification.

Definition 3. *A collaboration strategy* $\mathcal{C} : [0,1]^n \to \{0,1\}$ *is **reliable** if* $\mathrm{acc}_{\mathcal{C}}(S) \geq \min_{i\in[n]} \mathrm{acc}_i(S)$ *for all collaboration settings* S.

In words, a collaboration strategy is reliable if and only if it always performs at least as well as the *least accurate* agent.

Definition 4. *A collaboration strategy* $\mathcal{C} : [0,1]^n \to \{0,1\}$ *is **non-collaborative** if there exists* $k \in [n]$ *and* $\alpha \in \{0,1\}$ *such that*

$$\mathcal{C}(p_1, p_2, \cdots, p_n) = \begin{cases} \lfloor p_k \rceil & p_k \neq \frac{1}{2} \\ \alpha & p_k = \frac{1}{2} \end{cases} \quad (6)$$

for all $(p_1, p_2, \cdots, p_n) \in (0,1)^n$.

A collaboration strategy is non-collaborative if it essentially always defers to the classification of a single agent $k \in [n]$. The definition admits two exceptions. First, when $p_k = \frac{1}{2}$, the collaboration may select either 0 or 1, but must always select the same such value; this choice has no effect on the accuracy of the classifier, so this exception can be essentially ignored. Second, the collaboration strategy need not select $\lfloor p_k \rceil$ when $(p_1, p_2, \cdots, p_n) \notin (0,1)^n$—i.e., when $p_i \in \{0,1\}$ for some $i \in [n]$; this exception is somewhat more interesting, and we will return to it after stating our main result.

Theorem 1. *Every reliable collaboration strategy is non-collaborative.*

Theorem 1 implies that the search for "free" complementarity in human-AI collaboration is futile. Any collaboration strategy that is reliable (performs no worse than the *worst* agent) is non-collaborative (essentially always defers to the same one agent).

While Theorem 1 implies that *any* reliable collaboration strategy generally must defer to the same agent, it allows for one exception. In particular, taking note of Definition 4, if some other agent is entirely confident in their prediction ($p_i \in \{0,1\}$ for some i), it is not necessary to always defer to the same agent k. Indeed, it is always "reliable" (and, in fact, optimal) to defer to the prediction of an agent who is entirely confident, since agents are calibrated—guaranteeing that on the set of points for which the agent predicts probability 1, all points are truly positive. This suggests a simple collaboration strategy: always defer to a fixed agent, except when another agent is entirely confident. For example, either the human or algorithm can be assigned "primary" decision-making power, and the other party can be tasked only with overriding obvious mistakes. For example, in sports, "robot referees" are sometimes delegated cases where the algorithm is certain, such as in making line calls in tennis or offside calls in soccer.

This exception illustrates an underlying intuition behind Theorem 1: conditional on other agents' predictions, a given agent's prediction is no longer calibrated. Only in the specific case when an agent is *certain* in their prediction, can they be confident. Otherwise, even if an agent predicts probability 0.95, for example, deferring to that agent is not guaranteed to be a good idea in all situations—given the predictions of other agents, and conditional on the choice of deferring to the agent, the prediction of 0.95 may be far from calibrated.

3 Implications for Human-AI Collaboration

In this section, we use Theorem 1 to better understand the conditions that enable effective human-AI collaboration. We begin by discussing numerous settings in ML, human-AI collaboration, and beyond, in which collaboration has been shown to be possible (often, with guarantees). We then identify two common features of these "success stories"—features which are not present in the setup of Theorem 1. We then argue that common implementations of human-AI collaboration also lack these features, and suggest a path forward. Like Wolpert and Macready's "No Free Lunch Theorem," a primary use of Theorem 1 is in clarifying the additional structure needed to ensure successful prediction.

3.1 Successful Collaborations

There are many lines of work that are fundamentally about collaboration in classification tasks, each of which—unlike us—obtain positive results. For example, the machine learning literature is ripe with such results. Combining expert predictions is a basic problem in online learning theory (see Blum (2005) for an overview). There, it has been shown, for example, that expert predictions can be combined to perform better than the best linear combination of experts (Littlestone, Long, and Warmuth 1991). The idea of mixing expert predictions is also seen in the literature on ensemble classifiers, including boosting methods and random forests. Pivoting, Condorcet's jury theorem (de Condorcet 1785) provides a collaboration-based view of voting theory: when individual jurors are biased towards the correct decision, the majority vote is more accurate than individual votes. Finally, recent work on human-AI collaboration has presented numerous approaches to achieving complementarity (Madras, Pitassi, and Zemel 2018; Donahue, Chouldechova, and Kenthapadi 2022; Alur, Raghavan, and Shah 2024). Why do

each these methods work, and why do they have guarantees? We suggest two distinct features behind these successful collaborations.

Leveraging Independence: Condorcet's Jury Theorem, Wisdom of Crowds, Random Forests. Condorcet's jury theorem (de Condorcet 1785) states that when individual jurors are biased towards the correct decision (i.e., vote in that direction independently with probability $p > \frac{1}{2}$), the majority vote is more accurate than any individual vote. Indeed, the idea of aggregating independent signals appears in a much broader literature studying information aggregation and the "wisdom of crowds." A key distinction between Theorem 1 and these settings is a lack of "independence" in our setting; agents need not make predictions "independently" in our setup. This also appears to be the key distinction between our setting and that of majority vote ensemble classifiers in ML such as random forests, which rely on some amount of independence in how decision trees are constructed (Breiman 2001). Independence cannot be guaranteed in human-AI collaboration in this way.

Leveraging Learning: Experts, Boosting. Theorem 1 contrasts with the longstanding machine literature in machine learning that shows how multiple "expert" predictions can be effectively combined (Blum 2005). In fact, the idea of combining expert advice has formed a fruitful baseline intuition for how to build effective ML algorithms more broadly, such as in boosting (Schapire et al. 1999). A crucial aspect of these methods is the process of *learning* which experts to trust, and when and how much to trust them. (Note that while both random forests and boosting are considered ensemble methods—methods that combine predictions—the former relies on independence of predictions and the latter on learning joint behavior.) This learning process is absent in the setup of Theorem 1. While agents have strong information about their own predictions (calibration), they do not know direct information about joint behavior

3.2 Human-AI Collaboration

We have described two general approaches to ensuring effective collaboration: independence and learning. Since we cannot generally ensure that a human and algorithm produce independent estimates, we focus on the potential of the latter approach.

Learning enables collaboration by understanding the joint behavior of agents. Theorem 1 itself illustrates this point in one setting: when an agent is entirely certain in their prediction. In such a setting, the relevant "joint" information is fully understood: regardless of what the other agents predict, it is safe to defer to the agent. This connects more generally to recent work showing that complementarity is achievable exactly when there are subsets of the domain in which each agent has an advantage (Donahue, Chouldechova, and Kenthapadi 2022). The approach of "overriding only when an agent is certain" can be viewed in this framing, in which there is a designated region in which one agent has a clear advantage (due to their full certainty), but otherwise, the other agent is deferred to. However, as Theorem 1 implies, the regions in which each agent has an advantage cannot in general be determined *a priori* from only their predictions in each region (even if calibrated probabilities intuitively give a measure of effectiveness). Thus, implementations of such collaboration models must reason more directly with whether or not one agent is more equipped to handle a given subset of the feature space. One way in which to do this is by performing additional training using joint information. Indeed, idea has been taken up by "learning to defer" approaches (e.g., Madras, Pitassi, and Zemel (2018); Mozannar, Satyanarayan, and Sontag (2022)). Similarly, Alur, Raghavan, and Shah (2024) introduce a method to identify subsets of inputs that are indistinguishable to an agent, in which case signal from the prediction of the other agent can then be leveraged to improve predictions overall. These approaches require data from the joint distribution of agent predictions and outcomes. Theorem 1 suggests that such data is necessary to obtain guarantees.

4 Proof of Theorem 1

Before proceeding to the proof of Theorem 1, we begin by establishing some basic language and tools with which to analyze and construct collaboration settings.

4.1 Preliminaries

Correctness and Agreement. We first introduce basic language to describe the performance of agents and collaboration strategies.

Definition 5. *For a collaboration setting $(\mathcal{D}, P_1, \cdots, P_n)$ and $x \in \mathcal{X}$, we say that*

- *agent i is **correct** on x if*

$$\hat{Y}_i(x) = \left\lfloor \Pr_{(X,Y)\sim\mathcal{D}}[Y = 1 \mid X = x] \right\rceil$$

- *agent i is **incorrect** on x if*

$$\hat{Y}_i(x) \neq \left\lfloor \Pr_{(X,Y)\sim\mathcal{D}}[Y = 1 \mid X = x] \right\rceil$$

- *agents i and j **agree** on x if*

$$\hat{Y}_i(x) = \hat{Y}_j(x)$$

- *agents i and j **disagree** on x if*

$$\hat{Y}_i(x) \neq \hat{Y}_j(x).$$

For each of these statements, we can replace i or j with a collaboration strategy $\mathcal{C}$.

For example, if $\Pr_{(X,Y)\sim\mathcal{D}}[Y = 1 \mid X = x] = 0.75$, then i is correct on x if and only if $\hat{Y}_i(x) = 1$. (This is the correct classification to maximize 0-1 accuracy.) Using the language established in Definition 5, we can make some simple observations about accuracies. For example, if i is correct on x whenever j is correct on x, this implies that $\mathrm{acc}_i(S) \geq \mathrm{acc}_j(S)$. If there furthermore is some x for which i is correct but j is incorrect, and where $\Pr_{(X,Y)\sim\mathcal{D}}[X = x] \neq 0$, this implies that $\mathrm{acc}_i(S) > \mathrm{acc}_j(S)$.

Combining collaboration settings. Having established some basic language with which to describe and analyze the performance of experts and collaboration strategies on a collaboration setting, we now establish a basic tool for constructing collaboration settings "piece by piece."

Proposition 6. ***Linear combinations of settings.*** *Consider ℓ collaboration settings $S_1, \cdots, S_\ell$. Then for all $(\lambda_1, \cdots, \lambda_\ell) \in \Delta^\ell$, there exists a collaboration setting S such that*

$$\operatorname{acc}_i(S) = \sum_{m=1}^{\ell} \lambda_m \operatorname{acc}_i(S_m) \tag{7}$$

$$\operatorname{acc}_{\mathcal{C}}(S) = \sum_{m=1}^{\ell} \lambda_m \operatorname{acc}_{\mathcal{C}}(S_m) \tag{8}$$

for all $i \in [n]$ and collaboration strategies $\mathcal{C}$.

Proof. Let $S_m = (\mathcal{D}_m, P_{1,m}, \cdots, P_{n,m})$ for $m \in [\ell]$, where $\mathcal{D}_m$ is a distribution over $\mathcal{X}_m \times \{0,1\}$. Then consider the collaboration setting $S = (\mathcal{D}, P_1, \cdots, P_n)$, where $\mathcal{D}$ is a distribution over $\mathcal{X} \times \{0,1\}$ for $\mathcal{X} := \bigcup_{m=1}^{\ell} \{(m,x) : x \in \mathcal{X}_m\}$. Specifically, we define $\mathcal{D}$ such that

$$\Pr_{(X,Y)\sim\mathcal{D}}[X = (m,x)] = \lambda_m \Pr_{(X_m,Y_m)\sim\mathcal{D}_m}[X_m = x] \tag{9}$$

and

$$\Pr_{(X,Y)\sim\mathcal{D}}[Y = 1 | X = (m,x)] \tag{10}$$

$$= \Pr_{(X_m,Y_m)\sim\mathcal{D}_m}[Y_m = 1 | X_m = x]. \tag{11}$$

$\mathcal{D}$ is essentially the distribution obtained by first randomly sampling $m \in [\ell]$ with probability λ_ℓ and then sampling from $\mathcal{D}_m$. Now define P_i such that $P_i(m,x) = P_{i,m}(x)$ for all $x \in \mathcal{X}_m$. P_i is calibrated since $P_{i,m}$ is calibrated for all $m \in [\ell]$. By inspection, S satisfies (7) and (8). □

Building Calibrated Predictors from Partitions. Finally, given a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$, we show how partitions of the input space $\mathcal{X}$ induce calibrated predictors. Indeed, let $\mathcal{A}_i$ be a partition of $\mathcal{X}$. Then $\mathcal{A}_i$ induces a calibrated predictor P_i, where for $x \in A \in \mathcal{A}_i$,

$$P_i(x) := \Pr_{(X,Y)\sim\mathcal{D}}[Y = 1 \,|\, X \in A]. \tag{12}$$

Here P_i is the Bayes-optimal predictor given a "coarsening" of the input space into the partitions $\mathcal{X}$. In this way, a collaboration setting may also be identified by an ordered tuple $(\mathcal{D}, \mathcal{A}_1, \cdots, \mathcal{A}_n)$. This approach is central to the subsequent proofs.

4.2 Main Proof

In the remainder of this section, we prove Theorem 1 in full. We first rewrite Theorem 1 in an equivalent formulation.

Theorem 1. *For a collaboration strategy $\mathcal{C}$, $\operatorname{acc}_{\mathcal{C}}(S) \geq \min_{i\in[n]} \operatorname{acc}_i(S)$ for all collaboration settings S if and only if there exists $k \in [n]$ and $\alpha \in \{0,1\}$ such that for all $(p_1, p_2, \cdots, p_n) \in (0,1)^n$:*

(i) If $p_k \neq \frac{1}{2}$, then $\mathcal{C}(p_1, p_2, \cdots, p_n) = \lfloor p_k \rceil$.
(ii) If $p_k = \frac{1}{2}$, then $\mathcal{C}(p_1, p_2, \cdots, p_n) = \alpha$.

The high level plan for the proof is to first show that there must exist k such that condition (i) holds. Then, given that there must exist k such that (i) holds, we show that it must be the case that (ii) also holds (for the same k). The heart of the proof is in showing the first step, formalized in the proposition below.

Proposition 7. *For a collaboration strategy $\mathcal{C}$, if $\operatorname{acc}_{\mathcal{C}}(S) \geq \min_{i\in[n]} \operatorname{acc}_i(S)$ for all collaboration settings S, then there must exist $k \in [n]$ such that for all $(p_1, p_2, \cdots, p_n) \in (0,1)^n$ where $p_k \neq \frac{1}{2}$, $\mathcal{C}(p_1, p_2, \cdots, p_n) = \lfloor p_k \rceil$.*

To show Proposition 7, suppose for sake of contradiction that there does not exist $k \in [n]$ satisfying this property. This means that for every $k \in [n]$, there must exist some tuple $(p_1, p_2, \cdots, p_n) \in (0,1)^n$ where $p_k \neq \frac{1}{2}$, such that $\mathcal{C}(p_1, p_2, \cdots, p_n) \neq \lfloor p_k \rceil$. We show in Lemma 8 below that the existence of such a tuple implies that there is a collaboration setting S_k such that $\operatorname{acc}_k(S_k) > \operatorname{acc}_{\mathcal{C}}(S_k)$ and $\operatorname{acc}_i(S_k) \geq \operatorname{acc}_{\mathcal{C}}(S_k)$ for all $i \in [n] \setminus \{k\}$. Since we can construct such an S_k for all $k \in [n]$, Proposition 6 implies the existence of a collaboration setting S such that

$$\operatorname{acc}_{\mathcal{C}}(S) = \sum_{k=1}^{n} \frac{1}{n} \operatorname{acc}_{\mathcal{C}}(S_k) < \sum_{k=1}^{n} \frac{1}{n} \operatorname{acc}_i(S_k) = \operatorname{acc}_i(S) \tag{13}$$

for all $i \in [n]$, providing the desired contradiction. Therefore, it suffices to show Lemma 8.

Lemma 8. *Consider a collaboration strategy $\mathcal{C}$. Suppose that there exists a tuple $(p_1, p_2, \cdots, p_n) \in (0,1)^n$ where $p_k \neq \frac{1}{2}$ and $\mathcal{C}(p_1, p_2, \cdots, p_n) \neq \lfloor p_k \rceil$. Then there exists a collaboration setting S_k such that*

(i) $\operatorname{acc}_k(S_k) > \operatorname{acc}_{\mathcal{C}}(S_k)$,
(ii) $\operatorname{acc}_i(S_k) \geq \operatorname{acc}_{\mathcal{C}}(S_k)$ for all $i \neq k$.

Proof. We first define the collaboration setting $S_k(\mathcal{D}, \mathcal{A}_1, \cdots, \mathcal{A}_n)$. Define the input space $\mathcal{X} = \{0, 1, \cdots, n\}$, and for all $i \in [n]$, let $\mathcal{A}_i$ be the partition comprising the set $\{0, i\}$ together with the singleton sets $\{j\}$ for $j \notin \{0, i\}$. ($\mathcal{X}$ and $\mathcal{A}_i$ are depicted in Figure 1.) Define a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ in the following manner. Set

$$\Pr_{(X,Y)\sim\mathcal{D}}[Y = 1 \,|\, X = x] = \begin{cases} 1 - \mathcal{C}(p_1, \cdots, p_n) & x = 0 \\ \mathcal{C}(p_1, \cdots, p_n) & x \in [n]. \end{cases} \tag{14}$$

Furthermore, when $\mathcal{C}(p_1, \cdots, p_n) = 0$, set

$$\Pr_{(X,Y)\sim\mathcal{D}}[X = x] = \begin{cases} \frac{1}{1+\sum_{j\in[n]}(1-p_j)/p_j} & x = 0 \\ \frac{(1-p_i)/p_i}{1+\sum_{j\in[n]}(1-p_j)/p_j} & x \in [n]. \end{cases} \tag{15}$$

When $\mathcal{C}(p_1, \cdots, p_n) = 1$, set

$$\Pr_{(X,Y)\sim\mathcal{D}}[X = x] = \begin{cases} \frac{1}{1+\sum_{j\in[n]} p_j/(1-p_j)} & x = 0 \\ \frac{p_i/(1-p_i)}{1+\sum_{j\in[n]} p_j/(1-p_j)} & x \in [n]. \end{cases} \tag{16}$$

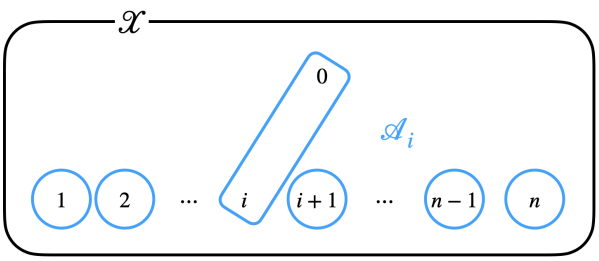

Figure 1: An illustration of a collaboration setting constructed in the proof of Proposition 7: the input space $\mathcal{X} = \{0, 1, 2, \cdots, n\}$ and the partition $\mathcal{A}_i$ (comprised of $\{0, i\}$ and the remaining singleton sets $\{j\}$ for $i \notin \{0, i\}$). In this setting, agent i is always correct for the inputs $x \notin \{0, i\}$, since $P_i(x)$ is exactly $\Pr[Y = 1|X = x]$ on these points. The full collaboration setting used in Proposition 7 consists of combining n such settings—one for each agent $k \in [n]$. Each such setting S_k is constructed such that the collaboration strategy $\mathcal{C}$ performs strictly worse than agent k and no better than the remaining agents.

This completes the construction of the collaboration setting S_k. S_k satisfies the key property that for all $i \in [n]$, $\Pr_{(X,Y)\sim\mathcal{D}}[Y = 1 \,|\, X \in \{0, i\}] = p_i$. This implies that $P_i(0) = P_i(i) = p_i$, which further implies a central feature of this construction: that

$$\hat{Y}_{\mathcal{C}}(0) = \mathcal{C}(P_1(0), \cdots, P_n(0)) = \mathcal{C}(p_1, \cdots, p_n). \quad (17)$$

Meanwhile, by construction (14), $\Pr[Y = 1 \,|\, X = 0] = 1 - \mathcal{C}(p_1, \cdots, p_n)$. Therefore, $\mathcal{C}$ is incorrect on 0 (using the terminology established in Definition 5). On the other hand, for all $i \in [n]$, agent i is correct on all $x \in [n] \setminus \{i\}$ since agent i is correct on all singletons in $\mathcal{A}_i$. The correctness of agents i and the collaboration strategy $\mathcal{C}$ can be summarized as follows:

	$x = 0$	$x = i$	$x \in [n] \setminus i$
agent i	?	?	correct
$\mathcal{C}$	incorrect	?	?

Now note that $\hat{Y}_i(0) = \hat{Y}_i(i)$ and $\Pr[Y = 1 \,|\, X = 0] = 1 - \Pr[Y = 1 \,|\, X = i]$, so i must be correct on either 0 or i. To complete the proof, we would like to show that $\mathrm{acc}_{\mathcal{C}}(S_k) < \mathrm{acc}_k(S_k)$ and $\mathrm{acc}_{\mathcal{C}}(S_k) \le \mathrm{acc}_i(S_k)$ for $i \neq k$. To show these inequalities, since agent i is always correct on $j \notin \{0, i\}$, it suffices to analyze the accuracy of the classifiers on $x = 0$ and $x = i$. This can be handled in two cases:

- When $\mathcal{C}$ agrees with agent i on $x = 0$, agent i is incorrect on $x = 0$ and correct on $x = 1$. In this case, agent i is correct whenever $\mathcal{C}$ is correct, so $\mathrm{acc}_{\mathcal{C}}(S_k) \le \mathrm{acc}_i(S_k)$.
- When $\mathcal{C}$ disagrees with agent i on $x = 0$, agent i is correct on $x = 0$ and incorrect on $x = 1$; $\mathcal{C}$ is incorrect on $x = 0$, and perhaps correct on $x = 1$. Therefore, agent i is at least as accurate than $\mathcal{C}$ on $\{0, i\}$ if $\Pr[X = 0] \ge \Pr[X = i]$, and is strictly more accurate than $\mathcal{C}$ if $\Pr[X = 0] > \Pr[X = i]$.

  If $p_i = \frac{1}{2}$, then $\Pr[X = 0] = \Pr[X = i]$. Therefore, agent i is at least as accurate as $\mathcal{C}$ on $\{0, i\}$, so $\mathrm{acc}_{\mathcal{C}}(S_k) \le \mathrm{acc}_i(S_k)$.

  If $p_i \neq \frac{1}{2}$, since i makes the optimal classification with respect to the set $\{0, i\}$ (recalling the partition $\mathcal{A}_i$), $\Pr[X = 0] > \Pr[X = i]$. Therefore, agent i is strictly more accurate than $\mathcal{C}$ on $\{0, i\}$, so $\mathrm{acc}_{\mathcal{C}}(S_k) < \mathrm{acc}_i(S_k)$.

In every case, $\mathrm{acc}_{\mathcal{C}}(S_k) \le \mathrm{acc}_i(S_k)$. Furthermore, by assumption, k disagrees with $\mathcal{C}$ on $x = 0$ and $p_k \neq \frac{1}{2}$, so $\mathrm{acc}_{\mathcal{C}}(S_k) < \mathrm{acc}_k(S_k)$. □

We now show the second component of the proof, handling the case where $p_k = \frac{1}{2}$.

Proposition 9. *Consider a collaboration strategy $\mathcal{C}$ such that there exists $k \in [n]$ such that for all $(p_1, p_2, \cdots, p_n) \in (0, 1)^n$ where $p_k \neq \frac{1}{2}$, $\mathcal{C}(p_1, p_2, \cdots, p_n) = \lfloor p_k \rceil$. Then, if $\mathrm{acc}_{\mathcal{C}}(S) \ge \min_{i\in[n]} \mathrm{acc}_i(S)$ for all collaboration settings S, there must exist $\alpha \in \{0, 1\}$ such that for all $(p_1, p_2, \cdots, p_n) \in (0, 1)^n$ where $p_k = \frac{1}{2}$,*

$$\mathcal{C}(p_1, p_2, \cdots, p_n) = \alpha. \quad (18)$$

Proof. We first establish a collaboration setting $S_1 = (\mathcal{D}, \mathcal{A}_1, \cdots, \mathcal{A}_n)$ such that $\mathrm{acc}_{\mathcal{C}}(S_1) = \mathrm{acc}_k(S_1)$ and $\mathrm{acc}_{\mathcal{C}}(S_1) < \mathrm{acc}_i(S_1)$ for all $i \neq k$. Set $\mathcal{X} = \{0, 1\}$ and choose $\mathcal{D}$ where

$$\Pr[X = 0] = 1/3, \qquad \Pr[Y = 1 \,|\, X = 0] = \epsilon,$$
$$\Pr[X = 1] = 2/3, \qquad \Pr[Y = 1 \,|\, X = 1] = 1 - \epsilon,$$

where $\epsilon < \frac{1}{2}$. Now set $\mathcal{A}_k = \{\{0, 1\}\}$ and $\mathcal{A}_i = \{\{0\}, \{1\}\}$ for all $i \neq k$. Then $P_k(0) = P_k(1) = \frac{2}{3} - \frac{\epsilon}{3}$, while $P_i(0) = \epsilon$ and $P_i(1) = 1 - \epsilon$ for $i \neq k$. Then $\hat{Y}_{\mathcal{C}}(0) = \hat{Y}_k(0)$ and $\hat{Y}_{\mathcal{C}}(1) = \hat{Y}_k(1)$ since $(P_1(0), \cdots, P_n(0)), (P_1(1), \cdots, P_n(1)) \in (0, 1)^n$. Then observe that $\mathrm{acc}_k(S_1) = \mathrm{acc}_{\mathcal{C}}(S_1) = \frac{2}{3} - \frac{\epsilon}{3}$ (since $\mathcal{C}$ always defers to agent k's classification) and $\mathrm{acc}_i(S_1) = 1 - \epsilon$ for $i \neq k$. For $\epsilon < \frac{1}{2}, \frac{2}{3} - \frac{\epsilon}{3} < 1 - \epsilon$, giving the desired conclusion.

More substantively, we now establish a collaboration setting S_2 such that $\mathrm{acc}_{\mathcal{C}}(S_2) < \mathrm{acc}_k(S_2)$. If there does not exist α such that $\mathcal{C}(p_1, p_2, \cdots, p_n) = \alpha$ for all $(p_1, p_2, \cdots, p_n) \in (0, 1)^n$ where $p_k = \frac{1}{2}$, then there must exist two tuples $(p_1, \cdots, p_n), (q_1, \cdots, q_n) \in (0, 1)^n$ such that $p_k = q_k = \frac{1}{2}$ and

$$\mathcal{C}(p_1, \cdots, p_n) = 1 \quad (19)$$
$$\mathcal{C}(q_1, \cdots, q_n) = 0. \quad (20)$$

Now consider

$$\mathcal{X} = \{(0, i) : i \in \{0, \cdots, n\}\} \cup \{(1, i) : i \in \{0, \cdots, n\}\}. \quad (21)$$

Choose $\mathcal{D}$ such that $\Pr[X = x]$ is equal to

$$\begin{cases} \frac{1}{2+\sum_{j\in[n]} p_j/(1-p_j)+(1-q_j)/q_j} & x \in \{(0,0), (1,0)\} \\ \frac{p_i/(1-p_i)}{2+\sum_{j\in[n]} p_j/(1-p_j)+(1-q_j)/q_j} & x = (0, i) \\ \frac{(1-q_i)/q_i}{2+\sum_{j\in[n]} p_j/(1-p_j)+(1-q_j)/q_j} & x = (1, i) \end{cases} \quad (22)$$

and

$$\Pr[Y=1|X=x]=\begin{cases}0 & \text{for } x=(0,0)\\ 1 & \text{for } x=(0,i):i\in[n]\\ 1 & \text{for } x=(1,0)\\ 0 & \text{for } x=(1,i):i\in[n]\end{cases}. \tag{23}$$

Then

$$\Pr[Y=1\,|\,X\in\{(0,0),(0,i)\}]=p_i \tag{24}$$

$$\Pr[Y=1\,|\,X\in\{(1,0),(1,i)\}]=q_i. \tag{25}$$

Take $\mathcal{A}_k$ to be the partition with $\{(0,0),(1,0)\}$ and the remaining singleton sets. Take $\mathcal{A}_i$ for $i\neq k$ to be any partition with $\{(0,0),(0,i)\}$ and $\{(1,0),(1,i)\}$. Then consider $S_2=(\mathcal{D},\mathcal{A}_1,\cdots,\mathcal{A}_n)$. Then

$$P_k((0,0))=\frac{1}{2},\quad P_i((0,i))=p_i, \tag{26}$$

$$P_k((1,0))=\frac{1}{2},\quad P_i((1,i))=q_i, \tag{27}$$

so

$$\hat{Y}_{\mathcal{C}}(0,0)=\mathcal{C}(p_1,\cdots,p_n)=1=1-\Pr[Y=1\,|\,X=(0,0)], \tag{28}$$

so $\mathcal{C}$ is incorrect on $(0,0)$. Similarly, $\mathcal{C}$ is also incorrect on $(1,0)$. Meanwhile, k is correct on either $(0,0)$ or $(1,0)$, and is correct on all $(0,i)$ and $(1,i)$ for $i\neq 0$. Therefore, $\mathrm{acc}_{\mathcal{C}}(S_2)<\mathrm{acc}_k(S_2)$.

The result follows by applying Proposition 6 with S_1 and S_2, taking λ_1 sufficiently close to 1. In particular, Proposition 6 implies that for all λ, there exists a collaboration setting S such that

$$\mathrm{acc}_i(S)=\lambda\,\mathrm{acc}_i(S_1)+(1-\lambda)\,\mathrm{acc}_i(S_2) \tag{29}$$

for all $i\in[n]$, and

$$\mathrm{acc}_{\mathcal{C}}(S)=\lambda\,\mathrm{acc}_{\mathcal{C}}(S_1)+(1-\lambda)\,\mathrm{acc}_{\mathcal{C}}(S_2). \tag{30}$$

Then, for all $\lambda<1$, since $\mathrm{acc}_k(S_1)=\mathrm{acc}_{\mathcal{C}}(S_1)$ and $\mathrm{acc}_k(S_2)>\mathrm{acc}_{\mathcal{C}}(S_2)$, we have that $\mathrm{acc}_k(S)>\mathrm{acc}_{\mathcal{C}}(S)$. Now, since $\mathrm{acc}_i(S_1)>\mathrm{acc}_{\mathcal{C}}(S_1)$, regardless of $\mathrm{acc}_i(S_2),\mathrm{acc}_{\mathcal{C}}(S_2)$, by taking λ sufficiently close to 1, $\mathrm{acc}_i(S)>\mathrm{acc}_{\mathcal{C}}(S)$. This provides the desired contradiction. □

Finally, recall that Theorem 1 follows directly from sequentially applying Propositions 7 and 9.

5 Conclusion

In this paper, we proved a "No Free Lunch"-style result in human-AI collaboration. In particular, in a classification setting with multiple calibrated agents, we showed that any collaboration strategy that is guaranteed to perform no worse than the *least accurate* agent must essentially always defer to the same agent. The result does, however, imply one successful collaboration strategy: deferring to the same agent except when another agent is fully certain in their prediction. More broadly, Theorem 1 suggests that strong individual information (calibration) is not sufficient to enable collaboration; rather, successful collaboration hinges on learning joint information across agents.

Open Problems. The present result suggests a number of problems for future work. In the binary setting, it is not clear from the proof of Theorem 1 whether or not complementarity can be achieved for loss functions beyond 0-1 accuracy. For example, the baseline of reliability can be guaranteed under ℓ_2 loss by simply predicting the average probability of agents. Moreover, the present result places no restrictions on the distribution over $\mathcal{X}\times\{0,1\}$. In the spirit of the No Free Lunch Theorem, it would be interesting to consider what restrictions on this distribution enable collaboration strategies (and what those collaboration strategies are). Finally, Theorem 1 does not obviously extend to multi-class classification problems. Multi-class problems further opens up the possibility of collaboration strategies that succeed in set prediction (Straitouri et al. 2023; De Toni et al. 2024).

Acknowledgments

We thank Rohan Alur, Kate Donahue, Sophie Greenwood, and Rajiv Movva for valuable discussion and feedback. Nikhil Garg is supported by NSF CAREER IIS-2339427 and Cornell Tech Urban Tech Hub and Amazon research awards.

References

Alur, R.; Raghavan, M.; and Shah, D. 2024. Human expertise in algorithmic prediction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Arrow, K. J. 1950. A difficulty in the concept of social welfare. *Journal of Political Economy*, 58(4): 328–346.

Bhatt, U.; Antorán, J.; Zhang, Y.; Liao, Q. V.; Sattigeri, P.; Fogliato, R.; Melançon, G.; Krishnan, R.; Stanley, J.; Tickoo, O.; et al. 2021. Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 401–413.

Blum, A. 2005. On-line algorithms in machine learning. *Online Algorithms: The State of the Art*, 306–325.

Bondi, E.; Koster, R.; Sheahan, H.; Chadwick, M.; Bachrach, Y.; Cemgil, T.; Paquet, U.; and Dvijotham, K. 2022. Role of human-AI interaction in selective prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 5286–5294.

Breiman, L. 2001. Random forests. *Machine Learning*, 45: 5–32.

Buçinca, Z.; Malaya, M. B.; and Gajos, K. Z. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-computer Interaction*, 5(CSCW1): 1–21.

Cabrera, Á. A.; Perer, A.; and Hong, J. I. 2023. Improving human-AI collaboration with descriptions of AI behavior. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1): 1–21.

de Condorcet, N. 1785. Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix.

De Toni, G.; Okati, N.; Thejaswi, S.; Straitouri, E.; and Gomez-Rodriguez, M. 2024. Towards Human-AI Complementarity with Predictions Sets. *arXiv preprint arXiv:2405.17544*.

Donahue, K.; Chouldechova, A.; and Kenthapadi, K. 2022. Human-algorithm collaboration: Achieving complementarity and avoiding unfairness. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1639–1656.

Green, B.; and Chen, Y. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 1–24.

Kiani, A.; Uyumazturk, B.; Rajpurkar, P.; Wang, A.; Gao, R.; Jones, E.; Yu, Y.; Langlotz, C. P.; Ball, R. L.; Montine, T. J.; et al. 2020. Impact of a deep learning assistant on the histopathologic classification of liver cancer. *NPJ Digital Medicine*, 3(1): 23.

Lai, V.; and Tan, C. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 29–38.

Littlestone, N.; Long, P. M.; and Warmuth, M. K. 1991. On-line learning of linear functions. In *Proceedings of the twenty-third annual ACM Symposium on Theory of Computing*, 465–475.

Madras, D.; Pitassi, T.; and Zemel, R. 2018. Predict responsibly: improving fairness and accuracy by learning to defer. *Advances in Neural Information Processing Systems*, 31.

Mozannar, H.; Satyanarayan, A.; and Sontag, D. 2022. Teaching humans when to defer to a classifier via exemplars. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 5323–5331.

Schapire, R. E.; et al. 1999. A brief introduction to boosting. In *IJCAI*, volume 99, 1401–1406. Citeseer.

Shin, W.; Han, J.; and Rhee, W. 2021. AI-assistance for predictive maintenance of renewable energy systems. *Energy*, 221: 119775.

Straitouri, E.; Wang, L.; Okati, N.; and Rodriguez, M. G. 2023. Improving expert predictions with conformal prediction. In *International Conference on Machine Learning*, 32633–32653. PMLR.

Vaccaro, M.; Almaatouq, A.; and Malone, T. 2024. When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 1–11.

Wolpert, D. H.; and Macready, W. G. 1997. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1): 67–82.