When Are Two Lists Better Than One?: Benefits and Harms in Joint Decision-Making

Kate Donahue[1]*, **Sreenivas Gollapudi**[2], **Kostas Kollias**[2]

[1]Cornell University
[2]Google
kdonahue@cs.cornell.edu, sgollapu@google.com, kostaskollias@google.com

Abstract

Historically, much of machine learning research has focused on the performance of the algorithm alone, but recently more attention has been focused on optimizing joint human-algorithm performance. Here, we analyze a specific type of human-algorithm collaboration where the algorithm has access to a set of n items, and presents a subset of size k to the human, who selects a final item from among those k. This scenario could model content recommendation, route planning, or any type of labeling task. Because both the human and algorithm have imperfect, noisy information about the true ordering of items, the key question is: which value of k maximizes the probability that the best item will be ultimately selected? For k=1, performance is optimized by the algorithm acting alone, and for k=n it is optimized by the human acting alone. Surprisingly, we show that for multiple of noise models, it is optimal to set k in [2, n-1] - that is, there are strict benefits to collaborating, even when the human and algorithm have equal accuracy separately. We demonstrate this theoretically for the Mallows model and experimentally for the Random Utilities models of noisy permutations. However, we show this pattern is *reversed* when the human is anchored on the algorithm's presented ordering - the joint system always has strictly worse performance. We extend these results to the case where the human and algorithm differ in their accuracy levels, showing that there always exist regimes where a more accurate agent would strictly benefit from collaborating with a less accurate one, but these regimes are asymmetric between the human and the algorithm's accuracy.

1 Introduction

Consider the following motivating example:

> Alice is a doctor trying to classify a scan with one of n different labels. Based on her professional expertise and relevant medical information she has access to, she is able to make some ranking over which of these labels is most likely to be accurate. However, she is not perfect, and sometimes picks the wrong label. She decides to use a machine learning algorithm as a tool. The algorithm similarly has a goal of maximizing the probability of picking the correct label. However, the algorithm and Alice rely on somewhat different information sources in making their predictions: vast troves of data for the algorithm, and personal conversations with the patient for the human, for example. Because of this, their rankings over the true labels will often differ slightly. The algorithm communicates its knowledge by presenting its top k labels to Alice, who picks her top label among those that are presented. For what settings and what values of k will Alice and the algorithm working together have a higher chance of picking the right label?

If the algorithm were able to tell Alice exactly which label she should pick ($k = 1$), then this problem would simply reduce to that of building a highly accurate machine learning system. However, in the medical prediction setting, it is unrealistic to assume that the algorithm can force Alice to pick a particular label. If the algorithm presented all of the items to Alice ($k = n$), then this would be equivalent to Alice solving the task herself. In the case where n is large, considering each possible label may be infeasible. However, even if Alice could consider all n items herself, we will show that there are often settings where allowing the algorithm to narrow the set of items to k strictly increases the probability of picking the correct item.

In human-algorithm collaboration more generally, often the algorithm can provide assistance, but the human makes the final decision. This is the case in other settings as well: a diner trying to find the best restaurant, a driver trying to find the best route, or a teacher trying to find the best pedagogical method. This framework requires a shift in thinking: rather than focus on optimizing the performance of the algorithm alone, the goal is to build an algorithm that maximizes the performance of the human-algorithm system.

In this paper, we will focus on the role of the noise distributions that govern the human and algorithm. In particular, we will be interested in how *independent* these are. In particular, we will be interested in how strongly the human's permutation is affected by the algorithm's prediction, or the strength of *anchoring*. In this paper, we will explore different models of noisy predictions, and give theoretical and empirical results describing when the joint human-algorithm system has a higher chance of picking the best item.

In Section 2, we describe the theoretical model that we will explore and in Section 3 we connect our model to related works. Section 4 considers the case where both the

[*] Work done while at Google.

human and the algorithm have identical accuracy rates, and gives theoretical proofs for conditions where there are strict *benefits* and strict *harms* to using a joint human-algorithm system with the Mallows model, a model of noisy permutations over an ordering. This section also shows empirically that these results hold much more broadly, including for the Random Utilities Model. Next, Section 5 explores the case where the human and algorithm can differ in their accuracy rates, focusing on the case with exactly 3 items, of which the algorithm selects 2 to be presented. In this setting, we show that there is always a regime where a more accurate player can strictly improve their accuracy by joining with a less accurate partner. However, we show that this pattern is asymmetric between the human and the algorithm: the human has a much wider range of algorithmic accuracy rates that it is willing to partner with. Finally, Section 6 concludes and discusses avenues for future work.

2 Models and Notation

2.1 Human-algorithm Collaboration Model

We assume that there are n items $\{x_1, \dots x_n\}$, and that the goal is to pick item x_1. Each item could represent labels for categorical prediction, news articles of varying relevance, or driving directions with variable levels of traffic, for example. There are two actors: the first (A) narrows the items from n total items to a top $k < n$ items which are presented to the second actor (H), which picks a single item among them. One consistent assumption we will make is that the second actor H is not able to directly access or choose from the full set of items: this could be, for example, because $k << n$ and H is bandwidth-limited in how many items it can consider. This model is quite broad: the two actors could be interacting recommendation algorithms, for example, or sequential levels of decision-making among human committees. However, the motivating example we will focus on in this paper is when the first actor A is an algorithm and the second actor H is a human. This setting naturally fits with the assumption that H is bandwidth-limited, and also motivates the assumption that A and H have differing orders for the items, drawn from potentially differing sources of knowledge, but are unable to directly communicate that knowledge to each other. This formulation also allows us to connect with the extensive literature on human-algorithm collaboration, which we discuss further in Section 3.

We will use π^a, π^h to denote the orderings of the algorithm and human over the n items, with $\pi^a_i = x_j$ meaning that the algorithm ranks item x_j in the ith place. We will use $\pi^a_{[k]}$ to denote the k items that the algorithm ranks first (and thus presents to the human) and $\pi^a_{-[n-k]}$ to denote the $n-k$ items that the algorithm ranks last (and fails to present to the human). Both π^a, π^h are random variables drawn from distributions $\pi^a \sim \mathcal{D}^a, \pi^h \sim \mathcal{D}^h$. We will often refer to the joint human-algorithm system as the *combined system.*

The distributions $\mathcal{D}^a, \mathcal{D}^h$ may be independent: this could reflect the case where both the human and algorithm come up with orderings separately, and then the algorithm presents a set of items for the human to pick between, where the human picks the best item according to their previously-determined ranking. We refer to the case of independent orderings as the *unanchored* case. Alternatively, the distributions $\mathcal{D}^a, \mathcal{D}^h$ may be correlated. In particular, we will compare the *unanchored* case with that of *anchored* ordering. In this setting, the algorithm draws an ordering $\pi^a \sim \mathcal{D}^a$, which then becomes the central ranking for the human – we will describe what this means technically for different noise models in the next section. This models settings where the algorithm presents a *ordering* of items to the human, rather than a set, which strongly biases the human. The anchored setting implies a strong degree of correlation between the human and the algorithm's ordering. We will relax this correlation with the *semi-anchored setting*, where the algorithm's ordering π^a influences the human's ordering π^h, but less strongly in the anchored setting. In Section 4 we present theoretical results for the anchored and unanchored case, as well as experimental results for the semi-anchored case, which we formalize further.

2.2 Noise Models

In this section, we introduce the noise models we will use for $\mathcal{D}^a, \mathcal{D}^h$, which governs how the algorithm and human respectively arrive at noisy permutations over each of the n items. Both of these noise models are standard in the literature, which is what prompted us to consider them in our paper.

Mallows Model The first is the Mallows model, which has been used extensively as a model of permutations (Mallows 1957). The model has two components: a central ordering π^* (here, assumed to be the "correct" ordering $\{x_1, x_2, \dots x_n\}$) and an accuracy parameter $\phi > 0$, where higher ϕ means that the distribution more frequently returns orderings that are close to the central ordering π^*. The probability of any permutation π occurring is given by $\frac{1}{Z} \cdot \exp\left(-\phi \cdot d(\pi^*, \pi)\right)$ where Z is a normalizing constant $\sum_{\pi' \in P} \exp\left(-\phi \cdot d(\pi^*, \pi)\right)$ involving a sum over the set all permutations P and $d(\pi^*, \pi)$ is a distance metric between permutations. In this work, we will use Kendall-Tau distance, which is standard. In particular, the Kendall-Tau distance is equivalent to the number of *inversions* in π. An inversion occurs when element x_i is ranked above x_j in the true ordering π^*, but is ranked below x_j in π. This can be roughly thought of as the number of "pairwise errors" π makes in ordering each of the elements. In the Mallows model, we model anchoring through $\mathcal{D}^h$ having the central ordering $\mathcal{D}\left(\pi^* = \pi^a\right), \pi^a \sim \mathcal{D}^a$. In this way, the human takes the algorithm's presented ordering as the "true" ordering and draws permutations centered on it. In the *unanchored* setting the human draws their permutation from a Mallows distribution centered at the correct ordering $\mathcal{D}\left(\pi^* = \{x_1, x_2, \dots x_n\}\right)$.

Random Utility Model The Random Utility Model (RUM) has similarly been extensively used as a model of permutations (Thurstone 1927). In this model, item i has some true value μ_i, where we assume μ_i is descending in i. The human and algorithm only have access to noisy estimates of these values, $\hat{X}^a_i \sim \mathcal{D}(\mu_i, \sigma^2)$ for some distribution

$\mathcal{D}$ with variance σ^2 (often assumed to be Gaussian, which we will use in this paper). These noisy estimates are then used to produce an order π^a, π^h in descending order of the values $\{\hat{X}_i^a\}, \{\hat{X}_i^h\}$. In RUM, we model anchoring through $\hat{X}_i^h \sim \mathcal{D}(\mu_j, \sigma_h^2)$, for where j is the index of item i in the algorithm's permutation π^a. We model the semi-anchored case by $\hat{X}_i^h \sim \mathcal{D}(w_a \cdot \mu_j + (1 - w_a) \cdot \mu_i, \sigma_h^2)$, where w_a is a weight parameter indicating how much the algorithm's ordering anchors the human's permutation, and j is the index of item i in the algorithm's permutation π^a.

3 Related Work

Studying human-algorithm collaboration is a large, rapidly-growing, and highly interdisciplinary area of research. Some veins of research are more ethnographic, studying how people use algorithmic input in their decision-making (Lebovitz, Levina, and Lifshitz-Assaf 2021; Lebovitz, Lifshitz-Assaf, and Levina 2020; Beede et al. 2020; Yang et al. 2018; Okolo et al. 2021). Other avenues work on developing ML tools designed to work with humans, such as in medical settings (Raghu et al. 2018) or child welfare phone screenings (Chouldechova et al. 2018). Finally, and most closely related to this paper, some works develop theoretical models to analyze human-algorithm systems, such as (Rastogi et al. 2022; Cowgill and Stevenson 2020; Bansal et al. 2021a; Steyvers et al. 2022; Madras, Pitassi, and Zemel 2018). Bansal et al. (2021b) proposes the notion of *complementarity*, which is achieved when a human-algorithm system together has performance that is strictly better than either the human or the algorithm could achieve along. (Steyvers et al. 2022) uses a Bayesian framework to model human-algorithmic complementarity, while (Donahue, Chouldechova, and Kenthapadi 2022) studies the interaction between complementarity and fairness in joint human-algorithm decision systems, and (Rastogi et al. 2022) provides a taxonomy of how humans and algorithms might collaborate. (Kleinberg and Raghavan 2021) is structurally similar to ours in that it uses the Mallows model and RUM model to give theoretical guarantees for performance related to rankings of items. However, its setting is human-algorithm *competition* rather than *cooperation*, where the question is whether it is better to rely on an algorithmic tool or more noisy humans to rank job candidates.

One related area of research is "conformal prediction" where the goal is to optimize the subset that the algorithm presents to the human, such as in (Straitouri et al. 2022; Wang, Joachims, and Rodriguez 2022; Angelopoulos et al. 2020; Vovk, Gammerman, and Shafer 2005; Babbar, Bhatt, and Weller 2022; Straitouri and Rodriguez 2023). This formulation is structurally similar to ours, but often takes a different approach (e.g. optimizing the subset given some prediction of how the human will pick among them). Another related area is "learning to defer", where an algorithmic tool learns whether to allow a human (out of potentially multiple different humans) to make the final decision, or to make the prediction itself (e.g. (Hemmer et al. 2022; Madras, Pitassi, and Zemel 2018; Raghu et al. 2019)). Finally, a third related area is multi-stage screening or pipelines, where each stage narrows down the set of items further (e.g. (Blum, Stangl, and Vakilian 2022; Wang and Joachims 2023; Dwork, Ilvento, and Jagadeesan 2020; Bower et al. 2022)). (Hron et al. 2021) specifically studies the case with multiple imperfect nominators who each suggest an action to a ranker, who picks among them (and explores how to optimize this setting).

Some papers study how humans rely on algorithmic predictions - for example, (De-Arteaga, Fogliato, and Chouldechova 2020) empirically studies a real-life setting where the algorithm occasionally provided incorrect predictions and explores how the human decision-maker is able to overrule its predictions, while (Benz and Rodriguez 2023) studies under what circumstances providing confidence scores helps humans to more accurately decide when to rely on algorithmic predictions. (Mclaughlin and Spiess 2023) studies a case where the human decision-maker views the algorithm's recommendation as the "default" - similar to our "anchoring" setting, while (Vasconcelos et al. 2023) studies how explanations can reduce the impact of anchoring, and (Fogliato et al. 2022) empirically studies the impact of anchoring in a medical setting. (Rambachan et al. 2021) studies how to identify human errors in labels from observational data, while (Alur et al. 2023) explores how an algorithmic system can detect when a human actor has access to different sources of information than the algorithm itself. Also in a medical setting, (Cabitza, Campagner, and Sconfienza 2021) studies how "interaction protocols" with doctors and algorithmic tools can affect overall accuracy. (Chen et al. 2023) empirically explores how human rely on their intuition along with algorithmic explanations in making decisions. (Mozannar et al. 2023) explores a setting where an LLM is making recommendations of code snippets to programmers, with the goal of making recommendations that are likely to be accepted. Related to complementarity, (Guszcza et al. 2022) describes the principles of "hybrid intelligence" necessary for optimizing human-algorithm collaboration.

There has also been a series of work looking more specifically at human-algorithm collaboration in bandit settings. Gao et al. (2021) learns from batched historical human data to develop an algorithm that assigns each task at test time to either itself or a human. Chan et al. (2019) studies a setting where the human is simultaneously learning which option is best for them. However, their framework allows the algorithm to overrule the human, which makes sense in many settings, but is not reasonable in some settings like as our motivating medical example. Bordt and Von Luxburg (2022) formalizes the problem as a two-player setting where both the human and algorithm take actions that affect the reward both experience. (Agarwal and Brown 2022) and (Agarwal and Brown 2023) study the case where a "menu" of k arms out of n are presented to the human, who selects a final one based on a preference model. This setting differs from ours in the model of human preferences over items, as well as the goal of optimizing for the algorithm's overall regret. (Yao et al. 2023) studies a related setting where multiple content creators each recommend a top k set of items to humans, who pick among those k according to a RUM - key differences are that content creators are competing with each other and also learning their own utility functions over time. (Tian et al. 2023) considers the case where the human's mental model of the algorithm is changing over time, and models this as a dynamical system.

Additionally, some work has used the framework of the human as the final decision-maker and studied how to disclose information so as to incentivize them to take the "right" action. Immorlica et al. (2018) studies how to match the best regret in a setting where myopic humans pull the final arm. Hu et al. (2022) studies a related problem with combinatorial bandits, where the goal is to select a subset of the total arms to pull. Bastani et al. (2022) investigates a more applied setting where each human is a potential customer who will become disengaged and leave if they are suggested products (arms) that are a sufficiently poor fit. Kannan et al. (2017) looks at a similar model of sellers considering sequential clients, specifically investigating questions of fairness. In general, these works differ from ours in that they assume a new human arrives at each time step, and so the algorithm is able to selectively disclose information to them.

4 Impact of Anchoring on Joint Performance

In this section, we explore the impact of anchoring on the performance of the joint system. Our goal is *complementarity* as defined in (Bansal et al. 2021b): when the joint system has a higher chance of picking the best item than either the human or algorithm alone. In particular, we will show that complementarity is impossible for anchored orderings, no matter what number of k items are or the relative accuracy levels of the human and algorithm ϕ^h, ϕ^a. By contrast, we will show that complementarity is possible with unanchored orderings even when the human and algorithm have equal accuracy rates, so long as the number as presented items $k = 2$. The first subsection describes theoretical tools that hold for all probability distributions, the next two subsections gives theoretical results for the Mallows model distribution, while the last subsection extends these results experimentally for the RUM, including the semi-anchored setting.

4.1 Preliminary Definitions and Tools

First, this subsection describes preliminary tools we will need in order to prove the anchoring results in later sections. Note that every result in this subsection holds for all distributions of human and algorithmic permutations $\mathcal{D}^h, \mathcal{D}^a$, and regardless of the level of anchoring. However, we will find these tools useful for analysis in later subsections with more specific assumptions on $\mathcal{D}^h, \mathcal{D}^a$.

First, Definitions 1 defines "good events" where the joint human-algorithm system picks the best arm, where the algorithm alone would not have, and Definition 2 defines"bad events", where the joint system fails to pick the best arm, where the algorithm alone would have. Note that these could be identically defined with respect to when the human would have picked the best arm. However, defining events relative to the algorithm will make later proofs technically simpler.

Definition 1. *A "good event" is a pair of permutations ρ^a, ρ^h where the joint human-algorithm system selects the best arm x_1 when the algorithm alone would not have picked it. The "good event" occurs when in one of two cases holds:*

1. *The algorithm does not rank x_1 first but includes it in the k items it presents, while the human ranks item x_1 first ($\rho_1^a \neq x_1, x_1 \in \rho_{[k]}^a, \rho_1^h = x_1$)*
2. *Identical to case 1, but instead the human ranks x_1 in position $m \geq 2$, and the algorithm removes all of the items the human had ranked before it ($\rho_1^a \neq x_1, x_1 \in \rho_{[k]}^a, \rho_m^h = x_1, \rho_{[m-1]}^h \subseteq \rho_{-[n-k]}^a$)*

Definition 2. *A "bad event" is a pair of permutations π^a, π^h where the joint human-algorithm system* fails *to pick the best arm, where the algorithm alone would have picked it.*

A "bad event" occurs when the algorithm ranks x_1 first, but the human does not ($\pi_1^a = x_1, \pi_1^h \neq x_1$) and it is not *the case that the human ranks x_1 in position m, and the algorithm removes all of the items the human had ranked before it (*not *that $\pi_1^a \in \pi_k^a, \pi_m^h = x_1, \pi_{[m-1]}^h \subseteq \pi_{-[n-k]}^a$).*

Complementarity occurs whenever the total probability of "good events" is greater than the total probability of "bad events".

Lemma 1 states that there exists a bijective mapping between "good events" and "bad events" - that is, for every "good event" there is a unique corresponding "bad event". As an immediate corollary, we see that there must be equal numbers of good and bad events. These results show the importance of the probability distributions $\mathcal{D}^a, \mathcal{D}^h$: given a uniform distribution over permutations, the good events and bad events are equally likely, so any complementarity must be driven by certain permutations being more likely than others.

Lemma 1. *For any human algorithm system with $k < n$, there is a bijective mapping between "good events" and "bad events".*

Corollary. *There are equal numbers of "good events" and "bad events".*

While the full proof of Lemma 1 is deferred to the full version[1], the relevant bijective mapping will be useful for later analysis. We define it as "best-item-mapping", a function mapping from "good events" to "bad events" by swapping the indices of the best item x_1 and whichever item x_j that the algorithm had ranked first instead of x_1.

Definition 3 (Best-item mapping)**.** *Take any pair of orderings ρ^a, ρ^h such that*

$$\rho_1^a = x_j \quad \rho_i^a = x_1 \quad \rho_m^h = x_1 \quad \rho_\ell^h = x_j$$

for $x_j \neq x_1$. Then, we construct the new orderings π^a, π^h by flipping the location of items x_1, x_j, keeping all other items in the same location:

$$\pi_1^a = x_1 \quad \pi_i^a = x_j \quad \pi_m^h = x_j \quad \pi_\ell^h = x_1$$

[1] https://arxiv.org/abs/2308.11721

4.2 Anchoring Always Causes Worse Performance

The preliminary results for “good events” and “bad events” in the previous subsection hold for all probability distributions $\mathcal{D}^a, \mathcal{D}^h$ and all types of anchoring between these distributions. In this and the next subsection, we will focus on the Mallows model and give conditions such that the joint system will perform strictly worse or better than human or algorithm alone.

Theorem 1 below, begins by showing that when anchoring is present, the joint system always has strictly worse accuracy than the algorithm alone - no matter how many items are presented k or the relative accuracy rates of the human and algorithm ϕ^a, ϕ^h. This is a quite general impossibility result, indicating that a wide range of conditions lead to undesirable performance.

Theorem 1. *In the anchored setting with Mallows model distributions for permutations, the probability of picking the best arm strictly* decreases *in the joint human-algorithm system, as compared to the algorithm alone. This holds for any $k < n$, no matter the accuracy rates for the algorithm and human ϕ^a, ϕ^a.*

While we defer a full proof of Theorem 1 to the full version, we give an informal proof sketch below:

Proof sketch. This proof uses the best-item mapping from Definition 3. In particular, we take any “good event”, apply the best-item mapping, and show that the corresponding “bad event” is strictly more likely than the “good event”.

Given the Mallows model, a permutation π is more likely if they involve fewer *inversions* (instances where $i < j$ but $\pi_i < \pi_j$: a lower-valued item is ranked above a higher-valued item). Best-item mapping works by flipping the rank of the best item x_1 and x_j, defined as whichever item the algorithm ranked first in the “good event”. This mapping changes the relative ranking of x_1 and x_j, but also the pairwise ranking of every item that is in between x_j and x_1. The full proof proves that this process always strictly *decreases* the total number of inversions in the algorithm’s ranking, relative to the “good event”.

Next, we consider the human’s permutation. Best-item mapping also flips the indices of x_1, x_j in the human’s permutation. However, in the anchored setting the human’s distribution $\mathcal{D}^h$ is defined relative to the algorithm’s presented permutation. Therefore, flipping the indices of x_1, x_j for the algorithm is equivalent to relabeling the items, meaning that the human’s “good event” permutation is exactly as likely as the human’s “bad event” ordering, given the changed permutation. Because of this, our results hold no matter the accuracy rates of the human and algorithm ϕ^h, ϕ^a. □

4.3 Strictly Better Performance is Always Achievable Without Anchoring

In the previous section, we showed that complementarity is impossible in the anchored setting under a wide range of conditions. In this section, we will give specific conditions for when complementarity is achievable in the *unanchored* setting: specifically, whenever the human and algorithm have equal accuracy rates $\phi^a = \phi^h$ and the algorithm presents $k = 2$ items. We consider this setting particularly important because it is extremely achievable: even if the human is very bandwidth limited, it is extremely reasonable to assume that they are able to consider a finalist set of 2 items to pick between.

Theorem 2. *In the unanchored setting with permutations governed by the Mallows model, the probability of picking the best arm strictly* increases *in the joint human-algorithm system when exactly 2 items are presented ($k = 2$) and $\phi^a = \phi^h$.*

While we will again defer a full proof to the full version, we will offer a proof sketch:

Proof sketch. Similar to Theorem 1, we use the best-item mapping to map between good and bad events. However, we show that in the unanchored setting, this mapping always results in a “bad event” that is equally or less likely than the corresponding “good event”.

First, we consider the algorithm’s permutations. Here, we show that best-item mapping actually *decreases* the total number of inversions by exactly one, making the “bad event” ordering for the algorithm strictly *more* likely. Decreasing the number of inversions is the *opposite* of the overall goal of this proof; the requirement that $k = 2$ is what upper bounds this number of inversions by exactly 1.

However, we show that this effect is counteracted by the human’s permutation. In the unanchored setting, the human’s permutation is completely independent of the algorithm’s permutation, so the analysis is much more involved than in Theorem 1. Specifically, we consider each “good event” case in Definition 1 and show that best-item mapping always *increases* the total number of inversions by at least one.

Because the human and algorithm are assumed to have equal accuracy rates, the increase in inversions from the human’s permutations cancels out the decrease in inversions from the algorithm’s permutations, showing that the “bad event” is no more likely than the corresponding “good event”.

The proof concludes by constructing an example where the “good event” is *strictly* more likely than the “bad event”, showing that the total probability of “good events” is strictly more likely than the total probability of “bad events”. □

Finally, we wish to comment briefly on the permutation distributions $\mathcal{D}^a, \mathcal{D}^h$. Both the statements of Theorem 1 and Theorem 2 are specific to the Mallows model. However, the proof technique relies very weakly on the Mallows assumption. Specifically, the only property that is necessary is that the best-item mapping in Definition 3 weakly decreases (for anchored) or increases (for unanchored) the probability of permutations occurring. For Mallows model, this is satisfied because the probability of a permutation occurring is governed by the number of inversions present. Other probability distributions satisfying this property would show identical properties to those proven in Theorems 1 and 2.

4.4 Numerical Extensions and Partial Anchoring

In this subsection, we extend the results of the previous subsections in two ways. First, we consider the Random Utility Model, another commonly used model of noisy permutations over items. Secondly, we model cases where the human is influenced by the algorithm's presented ranking of items, but not completely anchored on it - the semi-anchored case. Specifically, we model the semi-anchored case as there human draws their mean from a noise distribution with mean $w_a \cdot \mu_j + (1 - w_a) \cdot \mu_i$, where w_a is a weight parameter indicating how much the algorithm's ordering anchors the human's permutation, and j is the index of item i in the algorithm's permutation π^a. In this way, $w_a = 0$ reflects the unanchored case, while $w_a = 1$ reflects the anchored case.

Figure 1 demonstrates numerical simulations for the RUM, given decreasing weight w_a. Note that the x-axis gives k number of items presented: $k = 1$ is the accuracy of the algorithm alone, while $k = 5$ gives the accuracy of the human considering all items (but potentially anchored on the algorithm's ordering).

The top figure has $w_a = 1$, reflecting complete anchoring. In this case, we see accuracy is maximized at $k = 1$, which is when the algorithm acts alone. This result matches with Theorem 1's findings for the Mallows model: in a completely anchored setting, complementarity is impossible. Note that Figure 1's demonstrates even stronger results: that the accuracy of the joint system is *decreasing* in k the number of items presented.

The bottom figure has $w_a = 0$ (no anchoring). In this case, we note that accuracy is identical at $k = 1, k = n = 5$: the human and algorithm have equal accuracy in these plots and are independent, so they each have equal accuracy when acting alone. Here, we see that the expected accuracy at $k = 2$ is greater than the accuracy at $k = 1, k = 5$, again matching the results from Theorem 2 for the Mallows model. However, we again see stronger results in Figure 1, which shows that for the given parameters the joint system exhibits complementarity for all $k \in [2, n - 1]$.

Finally, the middle figure describes cases when the human is partially anchored on the algorithm and exhibits results intermediate to the top and bottom figures. Specifically, it seems like complementarity occurs whenever k is "sufficiently small" so the benefits of having the human's ranking outweighs the harms of anchoring.

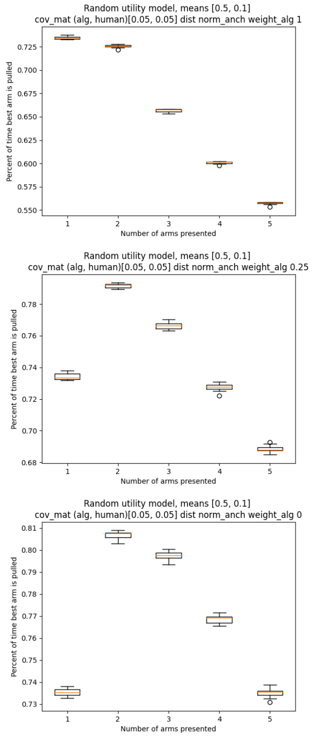

Figure 1: 10 trials, $5 \cdot 10^4$ simulations. RUM $\mathcal{D}_1^a = N(\mu = 0.5, \sigma^2 = 0.05), \mathcal{D}_{i>1}^a = N(\mu = 0.1, \sigma^2 = 0.05)$.

5 Asymmetric Complementarity Zones Without Anchoring

In previous sections, Theorem 1 showed that complementarity is impossible for the Mallows model, regardless of the levels of accuracy for the human and algorithm, while Theorem 2 showed that complementarity is possible in the unanchored case with identical accuracy rates $\phi_h = \phi_a$ with $k = 2$. In this section, we will further explore the unanchored setting, but allowing accuracy rates to differ. Specifically, we will show that there always exist regions of complementarity: cases where a more accurate agent would strictly increase its accuracy by collaborating with a less accurate partner. However, these regions are *asymmetric*: it is more likely that a more accurate human would gain from collaborating than a more accurate algorithm.

5.1 Provable Benefits From Joining With a Less Accurate Partner

Throughout this section, we will model the algorithm and human permutations as coming from a Mallows model. For analytical tractability, our theoretical results will focus on the case with $n = 3, k = 2$.

First, Lemma 2 shows that, no matter how accurate the algorithm is, there always exists a (slightly) more accurate human such that the joint system is strictly more accurate than either (achieving complementarity).

Lemma 2 (More accurate human)**.** *Consider $n = 3, k = 2$ where the human and algorithm both have unanchored Mallows models with $\phi_a \neq \phi_h$. Then, there exists a region of complementarity where a more accurate human obtains higher accuracy when collaborating with a less accurate algorithm. Specifically, for all $\phi_a > 0$, so long as $\phi_h \in [\phi_a, \min(1.3 \cdot \phi_a, \phi_a + 0.3)]$ the joint system has better performance than either the human alone or algorithm alone.*

For context, a Mallows model with $n = 3$ recovers the correct permutation $[x_1, x_2, x_3]$ with probability 48% of the time with $\phi = 1$ and 57% of the time with $\phi = 1.3$, so the regions in Lemma 2 represent moderate but meaningful differences in accuracy levels.

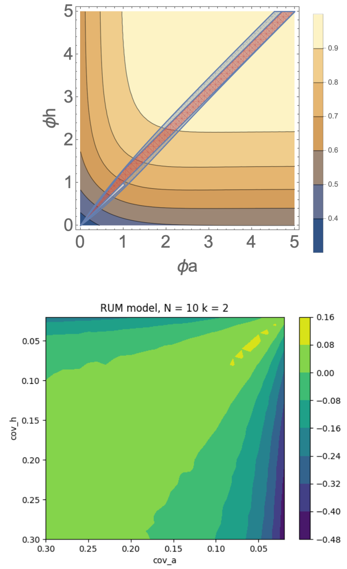

Figure 2: Above: contour plot showing increase or decrease in accuracy over maximum of human, algorithm accuracy given a Mallows model distribution with $n = 3, k = 2$. Positive regions indicate complementarity (blue). Below: *RUM* with *Normal* distribution with $n = 10, k = 2$. x and y axes show increasing accuracy (decreasing variance).

Next, Lemma 3 gives a corresponding result for when the algorithm is more accurate than the human. However, this region differs substantially from that in Lemma 2: it is substantially narrower, indicating a much smaller range where complementarity is possible.

Lemma 3 (More accurate algorithm)**.** *However, the roles of the human and algorithm are not symmetric: for the same setting as in Lemma 2, the zone of complementarity is much narrower. Specifically, complementarity is possible for* $\phi_a \in [\phi_h, \phi_h \cdot (1 + 0.01)$*, for all* $\phi_h \leq 1$*, but is never possible for any* $\phi_a \geq \phi_h + 0.15$ *for* $\phi_a \geq 1$.

These results are illustrated in Figure 2 (top). The contour plot gives the accuracy of the joint human-algorithm system, which is strictly increasing in ϕ^a, ϕ^h. Overlaid in blue is the analytically derived region of complementarity. The regions derived in Lemmas 2 and 3 are overlaid in red and white, respectively. Note that the red region encompasses almost all of the zone of complementarity, while the white region is comparatively minuscule.

Lemma 4 explains these results: for this setting, the performance of the joint system is always higher when the more accurate actor is the human, rather than the algorithm. For intuition for this asymmetry, consider the marginal impact of a more accurate algorithm - it will be slightly more likely to include the best item x_1 among the $k = 2$ it presents. However, once the algorithm is sufficiently accurate, it will almost always present x_1, so increasing accuracy will have diminishing returns. A more accurate human will be more likely to select the best item x_1, given that it is presented - which will more directly make the joint human-algorithm system more accurate.

This explains why the region of complementarity is larger when the human is the more accurate one - the human's accuracy more directly increases the accuracy of the joint system, which outperforms the more accurate actor (here, the human) for a wider range of accuracy differentials.

Lemma 4. *Given any two sets of Mallows accuracies* $\phi_1 > \phi_2$*, for* $n = 3, k = 2$*, the joint system always has strictly higher accuracy whenever* $\phi_a = \phi_1 > \phi_h = \phi_2$.

5.2 Numerical Extensions

Finally, Figure 2 (below) extends these results numerically. This figure extends the theoretical results in multiple ways: first, it show $n = 10$, which means substantially more items are presented than in the top figure. Holding $k = 2$, this means that the algorithm has a "harder" job to identify the $k = 2$ best arm. Secondly, the bottom figure shows the Random Utility Model of permutations, where greater accuracy levels are reflected by smaller standard deviations in noise. Similar to Section 4, we include this to show study how our theoretical results for the Mallows model extend to those RUM.

Note that even in this setting, we see qualitatively similar results to the top figure: there always exists a region of complementarity: specifically, in regions of low accuracy for the algorithm and human (bottom left of the figure) this region is largest, and this region roughly extends as the human and algorithm accuracy increase (diagonally to the upper right). However, we note that this region of complementarity is *asymmetric*: a more accurate human is more likely to benefit from partnering with a less accurate algorithm. That is visually apparent from how much further the zone of complementarity extends up the y axis (human covariance). Again, this is because increases in the accuracy of the human more directly increase the accuracy of the joint human-algorithm system.

6 Discussion and Future Work

In this paper, we have proposed a model of human-algorithm collaboration where neither the human or algorithm has ultimate say, but where they successively filter the set of n items down to k and finally a single choice. We focus on how the noise distributions $\mathcal{D}^a, \mathcal{D}^h$ influence whether the combined system has a higher chance of picking the best (correct) item. Future work extend our results to a wider range of noise model. Other interesting extensions could consider more complex models of human-algorithm collaboration - for example, cases where the human and algorithm can "vote" on the ordering of items, or other models of interaction. Additionally, they could explore cases where either the human or the algorithm is inherently biased - for example, when the algorithm has a central distribution that does that rank the best item first.

References

Agarwal, A.; and Brown, W. 2022. Diversified Recommendations for Agents with Adaptive Preferences. *ArXiv*, abs/2210.07773.

Agarwal, A.; and Brown, W. 2023. Online Recommendations for Agents with Discounted Adaptive Preferences. *arXiv preprint arXiv:2302.06014*.

Alur, R.; Laine, L.; Li, D. K.; Raghavan, M.; Shah, D.; and Shung, D. 2023. Auditing for Human Expertise. *arXiv preprint arXiv:2306.01646*.

Angelopoulos, A.; Bates, S.; Malik, J.; and Jordan, M. I. 2020. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*.

Babbar, V.; Bhatt, U.; and Weller, A. 2022. On the Utility of Prediction Sets in Human-AI Teams. arXiv:2205.01411.

Bansal, G.; Nushi, B.; Kamar, E.; Horvitz, E.; and Weld, D. S. 2021a. Is the Most Accurate AI the Best Teammate? Optimizing AI for Teamwork. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35.

Bansal, G.; Wu, T.; Zhou, J.; Fok, R.; Nushi, B.; Kamar, E.; Ribeiro, M. T.; and Weld, D. 2021b. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16.

Bastani, H.; Harsha, P.; Perakis, G.; and Singhvi, D. 2022. Learning personalized product recommendations with customer disengagement. *Manufacturing & Service Operations Management*, 24(4): 2010–2028.

Beede, E.; Baylor, E.; Hersch, F.; Iurchenko, A.; Wilcox, L.; Ruamviboonsuk, P.; and Vardoulakis, L. M. 2020. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–12.

Benz, N. L. C.; and Rodriguez, M. G. 2023. Human-Aligned Calibration for AI-Assisted Decision Making. arXiv:2306.00074.

Blum, A.; Stangl, K.; and Vakilian, A. 2022. Multi stage screening: Enforcing fairness and maximizing efficiency in a pre-existing pipeline. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1178–1193.

Bordt, S.; and Von Luxburg, U. 2022. A Bandit Model for Human-Machine Decision Making with Private Information and Opacity. In *International Conference on Artificial Intelligence and Statistics*, 7300–7319. PMLR.

Bower, A.; Lum, K.; Lazovich, T.; Yee, K.; and Belli, L. 2022. Random Isn't Always Fair: Candidate Set Imbalance and Exposure Inequality in Recommender Systems. *arXiv preprint arXiv:2209.05000*.

Cabitza, F.; Campagner, A.; and Sconfienza, L. M. 2021. Studying human-AI collaboration protocols: the case of the Kasparov's law in radiological double reading. *Health information science and systems*, 9: 1–20.

Chan, L.; Hadfield-Menell, D.; Srinivasa, S.; and Dragan, A. 2019. The assistive multi-armed bandit. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 354–363. IEEE.

Chen, V.; Liao, Q. V.; Vaughan, J. W.; and Bansal, G. 2023. Understanding the role of human intuition on reliance in human-AI decision-making with explanations. *arXiv preprint arXiv:2301.07255*.

Chouldechova, A.; Benavides-Prado, D.; Fialko, O.; and Vaithianathan, R. 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*, 134–148. PMLR.

Cowgill, B.; and Stevenson, M. T. 2020. Algorithmic social engineering. In *AEA Papers and Proceedings*, volume 110, 96–100.

De-Arteaga, M.; Fogliato, R.; and Chouldechova, A. 2020. A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–12.

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.

Dwork, C.; Ilvento, C.; and Jagadeesan, M. 2020. Individual fairness in pipelines. *arXiv preprint arXiv:2004.05167*.

Fogliato, R.; Chappidi, S.; Lungren, M.; Fisher, P.; Wilson, D.; Fitzke, M.; Parkinson, M.; Horvitz, E.; Inkpen, K.; and Nushi, B. 2022. Who Goes First? Influences of Human-AI Workflow on Decision Making in Clinical Imaging. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 1362–1374. New York, NY, USA: Association for Computing Machinery. ISBN 9781450393522.

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

Guszcza, J.; Danks, D.; Fox, C. R.; Hammond, K. J.; Ho, D. E.; Imas, A.; Landay, J.; Levi, M.; Logg, J.; Picard, R. W.; et al. 2022. Hybrid Intelligence: A Paradigm for More Responsible Practice. *Available at SSRN*.

Hemmer, P.; Schellhammer, S.; Vössing, M.; Jakubik, J.; and Satzger, G. 2022. Forming effective human-AI teams: building machine learning models that complement the capabilities of multiple experts. *arXiv preprint arXiv:2206.07948*.

Hron, J.; Krauth, K.; Jordan, M.; and Kilbertus, N. 2021. On Component Interactions in Two-Stage Recommender Systems. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 2744–2757. Curran Associates, Inc.

Hu, X.; Ngo, D. D.; Slivkins, A.; and Wu, Z. S. 2022. Incentivizing Combinatorial Bandit Exploration. *arXiv preprint arXiv:2206.00494*.

Immorlica, N.; Mao, J.; Slivkins, A.; and Wu, Z. S. 2018. Incentivizing exploration with selective data disclosure. *arXiv preprint arXiv:1811.06026*.

Kannan, S.; Kearns, M.; Morgenstern, J.; Pai, M.; Roth, A.; Vohra, R.; and Wu, Z. S. 2017. Fairness incentives for myopic agents. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, 369–386.

Kleinberg, J.; and Raghavan, M. 2021. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22): e2018340118.

Lebovitz, S.; Levina, N.; and Lifshitz-Assaf, H. 2021. Is AI ground truth really "true"? The dangers of training and evaluating AI tools based on experts' know-what. *Management Information Systems Quarterly*.

Lebovitz, S.; Lifshitz-Assaf, H.; and Levina, N. 2020. To incorporate or not to incorporate AI for critical judgments: The importance of ambiguity in professionals' judgment process. *Collective Intelligence, The Association for Computing Machinery*.

Madras, D.; Pitassi, T.; and Zemel, R. 2018. Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer. In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Mallows, C. L. 1957. Non-null ranking models. I. *Biometrika*, 44(1/2): 114–130.

Mclaughlin, B.; and Spiess, J. 2023. Algorithmic Assistance with Recommendation-Dependent Preferences. In *Proceedings of the 24th ACM Conference on Economics and Computation*, EC '23, 991. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701047.

Mozannar, H.; Bansal, G.; Fourney, A.; and Horvitz, E. 2023. When to Show a Suggestion? Integrating Human Feedback in AI-Assisted Programming. *arXiv preprint arXiv:2306.04930*.

Okolo, C. T.; Kamath, S.; Dell, N.; and Vashistha, A. 2021. "It cannot do all of my work": Community Health Worker Perceptions of AI-Enabled Mobile Health Applications in Rural India. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–20.

Raghu, M.; Blumer, K.; Corrado, G.; Kleinberg, J.; Obermeyer, Z.; and Mullainathan, S. 2018. The algorithmic automation problem: Prediction, triage, and human effort. *NeurIPS Workshop on Machine Learning for Health (ML4H)*.

Raghu, M.; Blumer, K.; Sayres, R.; Obermeyer, Z.; Kleinberg, B.; Mullainathan, S.; and Kleinberg, J. 2019. Direct uncertainty prediction for medical second opinions. In *International Conference on Machine Learning*, 5281–5290. PMLR.

Rambachan, A.; et al. 2021. Identifying prediction mistakes in observational data. *Harvard University*.

Rastogi, C.; Leqi, L.; Holstein, K.; and Heidari, H. 2022. A Unifying Framework for Combining Complementary Strengths of Humans and ML toward Better Predictive Decision-Making. *arXiv preprint arXiv:2204.10806*.

Steyvers, M.; Tejeda, H.; Kerrigan, G.; and Smyth, P. 2022. Bayesian modeling of human–AI complementarity. *Proceedings of the National Academy of Sciences*, 119(11): e2111547119.

Straitouri, E.; and Rodriguez, M. G. 2023. Designing Decision Support Systems Using Counterfactual Prediction Sets. arXiv:2306.03928.

Straitouri, E.; Wang, L.; Okati, N.; and Rodriguez, M. G. 2022. Provably Improving Expert Predictions with Conformal Prediction. arXiv:2201.12006.

Thurstone, L. L. 1927. A law of comparative judgment. *Psychological review*, 34(4): 273.

Tian, R.; Tomizuka, M.; Dragan, A. D.; and Bajcsy, A. 2023. Towards Modeling and Influencing the Dynamics of Human Learning. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, 350–358.

Vasconcelos, H.; Jörke, M.; Grunde-McLaughlin, M.; Gerstenberg, T.; Bernstein, M. S.; and Krishna, R. 2023. Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1): 1–38.

Vovk, V.; Gammerman, A.; and Shafer, G. 2005. *Algorithmic learning in a random world*, volume 29. Springer.

Wang, L.; and Joachims, T. 2023. Uncertainty Quantification for Fairness in Two-Stage Recommender Systems. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, 940–948.

Wang, L.; Joachims, T.; and Rodriguez, M. G. 2022. Improving screening processes via calibrated subset selection. In *International Conference on Machine Learning*, 22702–22726. PMLR.

Yang, Q.; Scuito, A.; Zimmerman, J.; Forlizzi, J.; and Steinfeld, A. 2018. Investigating how experienced UX designers effectively work with machine learning. In *Proceedings of the 2018 Designing Interactive Systems Conference*, 585–596.

Yao, F.; Li, C.; Nekipelov, D.; Wang, H.; and Xu, H. 2023. How Bad is Top-K Recommendation under Competing Content Creators? *arXiv preprint arXiv:2302.01971*.