When AI Influence Helps or Hurts: Guarantees for Ranking-Based Human–AI Aggregation

Jonas Karge[1], Roy Ferguson[2,3], Daniel Grimaldi[2,3], Jonas Haldimann[2,3], Ruvarashe Madzime[4,3], and Thomas Meyer[2,3]

[1] TU Dresden, Dresden, Germany
[2] University of Cape Town, Cape Town, South Africa
[3] Centre for Artificial Intelligence Research (CAIR), South Africa
[4] University of the Western Cape, Cape Town, South Africa

Abstract. We present, to the best of our knowledge, the first formal model of ranking-based human–AI aggregation. In our model, both human and AI judgments are represented by ranking functions, and we study how AI advice affects the epistemic quality of a collective decision. To obtain probabilistic guarantees, we introduce a top-choice abstraction that translates ranking aggregation into a plurality-based voting problem. This allows us to model a specific interaction setting that we call *AI-follow assistance with concurrent panel revision*. Building on this connection, we derive new probabilistic guarantees tailored to our setting. These results provide a formal account of when AI influence can be epistemically harmful by inducing dependence among human agents and when it can be epistemically beneficial by improving the probability that the true world is selected collectively.

Keywords: Human–AI Interaction · Ranking Aggregation · Epistemic Social Choice · Belief Change.

Human–AI interaction (HAI) studies how people and AI systems work together on decision-making problems. Central questions include how to design interaction patterns that support human performance, when control should remain with the human or be delegated to the AI, and which risks arise when humans rely on AI advice [18]. A particular focus in the HAI literature is the structure of the interaction itself, since the timing, format, and sequencing of AI advice can shape human judgment [6]. At the same time, the empirical evidence is mixed: a recent meta-analysis found that human–AI combinations are often better than humans alone, but on average still fail to outperform the better of humans or AI alone, with particularly weak results in decision-making tasks [21]. This makes it important to study not only whether AI advice is accurate in isolation, but also when its influence on a human group becomes detrimental to collective decision-making. Two important questions in this context are, first, how humans and AI systems represent their beliefs about the decision problem and, second, how we can evaluate the correctness of their joint decision.

In this paper, we use *ranking functions* as a joint representation for both human and AI judgments in order to address the first question. Ranking functions

have attracted attention since Spohn introduced them as a qualitative counterpart to probability theory [19]. They map possible worlds to natural numbers, where higher numbers represent greater implausibility and worlds of rank 0 are considered most plausible [19]. Thereby, they provide a numerical representation of plausibility over possible worlds and, in contrast to bare orderings, support arithmetic operations. This makes them a natural formal language for representing, comparing, revising, and aggregating structured beliefs in human–AI settings.

In the present paper, however, we focus on a top-choice abstraction of these rankings, that is, on the worlds each agent ranks most plausibly. To address the second question, we translate the agents' ranking-based beliefs into a voting problem and study the resulting aggregation *epistemically*, in the spirit of the *Condorcet Jury Theorem* (CJT) [16]. Because the original CJT relies on assumptions that rarely hold in practice, we model AI influence through a common-source dependence structure and derive a new plurality-based probabilistic guarantee for our ranking setting.

Recent work from several areas motivates our approach, but these strands have largely developed separately. As a representation of graded belief, we build on ranking functions in the sense of Spohn [19,20] and Huber [12], and use them as a shared epistemic language for both human and AI judgments. On the aggregation side, our work is closest to the *epistemic* view of belief merging, where the central question is whether aggregation identifies an unknown true world rather than merely producing a reasonable compromise [5]. Our probabilistic analysis is conceptually informed by recent voting-theoretic work on dependence and collective reliability, in particular the opinion-leader model of [14], while extending that perspective to our plurality-based top-choice setting. Finally, our setting is motivated by the human–AI interaction literature, which emphasizes that AI advice can induce anchoring and thereby create dependence between human judgments [6,1], while hybrid human–AI systems need not automatically outperform their components [17]. Our contribution is to bring these strands together in a first formal model of ranking-based human–AI aggregation in which AI advice influences human beliefs before aggregation.

Contributions. This paper makes two contributions. First, to the best of our knowledge, we introduce the first formal model of a ranking-based human–AI interaction scenario. Second, by translating the aggregation problem into a voting framework, we show how existing probabilistic guarantees can be used to study when AI influence becomes epistemically beneficial or harmful.

1 Ranking Functions

We model an agent's belief state by means of a ranking function, following the general framework of *ranking functions*, pioneered by Spohn [19,20]. Conceptually, agents hold beliefs about propositions or hypotheses. For aggregation, however, we represent each agent's belief state by a pointwise ranking over a

finite set of possible worlds, from which proposition-level ranks are induced. Intuitively, a ranking function assigns to each possible world a degree of disbelief: lower values indicate greater plausibility, while higher values indicate stronger disbelief. Let Ω be a finite set of possible worlds. A *ranking function* is a map $\kappa : \Omega \to \mathbb{N}_0 \cup \{\infty\}$ with $\min_{\omega \in \Omega} \kappa(\omega) = 0$. Thus, worlds of rank 0 are maximally plausible, higher ranks indicate increasing implausibility, and ∞ rules out a world.

We represent each agent $i \in \{1, \ldots, n\}$ by a ranking function $\kappa_i : \Omega \to \mathbb{N}_0 \cup \{\infty\}$, and write $E = (\kappa_1, \ldots, \kappa_n)$ for the resulting profile. An aggregation operator Δ maps E to a collective ranking $\kappa_{\text{agg}} = \Delta(E)$. As a simple baseline inspired by the merging literature [4], we consider *sum pooling*. For any world-score function $f : \Omega \to \mathbb{N}_0 \cup \{\infty\}$ with finite minimum, let $\text{norm}(f)(\omega) := f(\omega) - \min_{\omega' \in \Omega} f(\omega')$. Then $\text{norm}(f)$ is again a ranking function. The raw sum-pooled score of a world ω is $s^E_\Sigma(\omega) := \sum_{i=1}^n \kappa_i(\omega)$, and the corresponding collective ranking is $\kappa_\Sigma := \text{norm}(s^E_\Sigma)$.

2 Human–AI Interaction

In human–AI settings, collective judgments depend not only on the available information but also on how AI advice enters the decision process. The same aggregation rule may therefore behave differently under different interaction patterns. Here, "AI" is understood broadly as a task-specific decision aid whose output can be represented in the same ranking-based language as the humans' beliefs. A natural example is diagnostic decision support, where a system returns a ranked list of candidate diagnoses [10].

A useful distinction is between *AI-first* and *AI-follow* assistance [6]. Under AI-first assistance, the AI's recommendation is available from the outset and may shape the human's initial judgment. Under AI-follow assistance, the human first forms an independent preliminary judgment and only then receives the AI recommendation. This distinction matters because a central epistemic risk in human–AI interaction is *anchoring bias*: humans may shift their judgments too strongly toward the AI's suggestion, thereby introducing dependence between their judgments [1]. More generally, hybrid human–AI systems need not automatically outperform their components and may even be epistemically detrimental [17].

In this paper, we focus on *AI-follow with concurrent panel revision*. A panel consists of several human agents and one AI system. Each human first forms an initial ranking κ_i^0; the AI then provides its ranking κ_{AI}; finally, the humans revise in parallel, yielding updated rankings κ_i^1, and only these revised human rankings are aggregated. We choose this setting for two reasons. First, because the humans form κ_i^0 before seeing the AI's advice, it provides a clear unassisted baseline against which AI assistance can be evaluated. Second, concurrent revision makes the dependence structure tractable: all humans react to the same AI signal, but we do not also have to model sequential deliberation among the humans. This

makes it possible to study formally when AI influence is epistemically beneficial and when it becomes detrimental.

3 A Top-Choice Bridge to Opinion Leader Influence

In this section, we connect our ranking-based framework to the opinion-leader model of Karge et al. [14], which provides a recent generalization of the Condorcet Jury Theorem. Whereas their original framework is stated for approval voting, we reformulate the connection for our plurality-based top-choice setting and derive the corresponding probabilistic guarantees. This provides a direct bridge to the AI-follow interaction pattern with concurrent panel revision. Intuitively, this voting model considers a finite set of voters, or agents, denoted by $\mathcal{N} = \{a_1, \ldots, a_n\}$, and a finite set of alternatives, or worlds, $W = \{\omega_1, \ldots, \omega_m\}$. Under plurality voting, each agent a_i casts exactly one vote for one world in W. For each world ω_j, let $p_i^{\omega_j}$ denote the probability that agent a_i votes for ω_j. Since exactly one world, denoted by ω_*, is assumed to be the true world, the probability $p_i^{\omega_*}$ measures the probability that agent a_i votes correctly. The plurality winner is the world that receives strictly more votes than any other world, and we count the outcome as a success whenever ω_* is the plurality winner.

To model dependence between agents, this framework is based on the *opinion leader* (OL) model, one of the classical dependence models in the voting literature. The opinion leader represents an external influence on the election that does not itself take part in the voting, but publicly selects a single world and affects the electorate through a uniform influence parameter π, that is, the probability that an agent copies the OL's top choice. Moreover, the OL has its own competence parameter $\hat{p}$, namely the probability that its top choice is the true world. Thus, each agent either copies the opinion leader's top choice or keeps her own private top choice.

A top-choice ranking abstraction. At this stage, we do not yet model the full probabilistic revision of complete rankings. Instead, we work with a top-choice abstraction. This keeps the ranking framework in the background while giving us a clean way to model AI-induced dependence and to derive probabilistic bounds on the collective decision's correctness by translating the problem of ranking aggregation into a voting problem. Each human agent i first forms an initial ranking κ_i^0, and the AI provides its own ranking κ_{AI}. We focus on the top-ranked worlds:

$$\tau_i^0 := \arg\min_{\omega \in \Omega} \kappa_i^0(\omega), \qquad \tau_{AI} := \arg\min_{\omega \in \Omega} \kappa_{AI}(\omega), \tag{1}$$

assuming for simplicity that these top worlds are unique, i.e., every agent assigns rank 0 to exactly one world. Thus, τ_i^0 and τ_{AI} denote the worlds to which the human agent and the AI, respectively, assign the lowest rank.

To formally mirror the opinion leader model, we define indicator variables for each $\omega \in \Omega$:

$$X_i^\omega := \mathbf{1}[\tau_i^0 = \omega], \qquad X_{AI}^\omega := \mathbf{1}[\tau_{AI} = \omega]. \tag{2}$$

Here, an indicator variable takes value 1 if the stated condition is satisfied and value 0 otherwise. Thus, X_i^ω records whether agent i's private, i.e. uninfluenced, top choice is ω, and X_{AI}^ω records whether the AI's public top choice is ω. This gives us simple competence parameters at the top-choice level:

$$p_i := \mathbb{P}(\tau_i^0 = \omega_*) = \mathbb{P}(X_i^{\omega_*} = 1), \qquad \hat{p} := \mathbb{P}(\tau_{AI} = \omega_*) = \mathbb{P}(X_{AI}^{\omega_*} = 1), \tag{3}$$

and the average human competence $\bar{p} := \frac{1}{n}\sum_{i=1}^n p_i$.

AI-follow assistance. We now model AI-follow assistance. After seeing the AI's advice, each agent updates her top choice. To match the model in [14], we first assume a common influence level $\pi \in [0, 1]$. Agent i copies the AI's top choice with probability π, and otherwise keeps her own initial top choice:

$$\tau_i^1 = \begin{cases} \tau_{AI} & \text{with probability } \pi, \\ \tau_i^0 & \text{with probability } 1 - \pi. \end{cases} \tag{4}$$

Equivalently, the final vote indicator for world ω is

$$V_i^\omega := \mathbf{1}[\tau_i^1 = \omega], \tag{5}$$

that is, V_i^ω records whether ω becomes agent i's final top-ranked world after revision. Thus, in the language of ranking functions, $V_i^\omega = 1$ exactly when ω receives rank 0 in the agent's post-revision top-choice abstraction. Hence, for every $\omega \in \Omega$,

$$\mathbb{P}(V_i^\omega = 1) = \pi\,\mathbb{P}(X_{AI}^\omega = 1) + (1 - \pi)\,\mathbb{P}(X_i^\omega = 1). \tag{6}$$

That is, the probability that an agent votes for world ω is a probabilistic mixture of two cases: with probability π, the agent copies the AI's top choice, and with probability $1 - \pi$, the agent keeps her own private top choice. We aggregate the revised top choices by vote totals. For each world $\omega \in \Omega$, define

$$V_{\text{total}}^\omega := \sum_{i=1}^n V_i^\omega. \tag{7}$$

Thus, V_{total}^ω is the total number of agents who rank ω first after the revision step. The correct world wins whenever

$$V_{\text{total}}^{\omega_*} > \max_{\omega \neq \omega_*} V_{\text{total}}^\omega. \tag{8}$$

Intuitively, we count how often each world is top ranked by an agent, and the winning world is the one that is top ranked more often than any other world.

To return to the language of ranking aggregation, we induce a collective ranking from these vote scores by defining

$$\kappa_{\text{vote}}(\omega) := \max_{\omega' \in \Omega} V_{\text{total}}^{\omega'} - V_{\text{total}}^\omega. \tag{9}$$

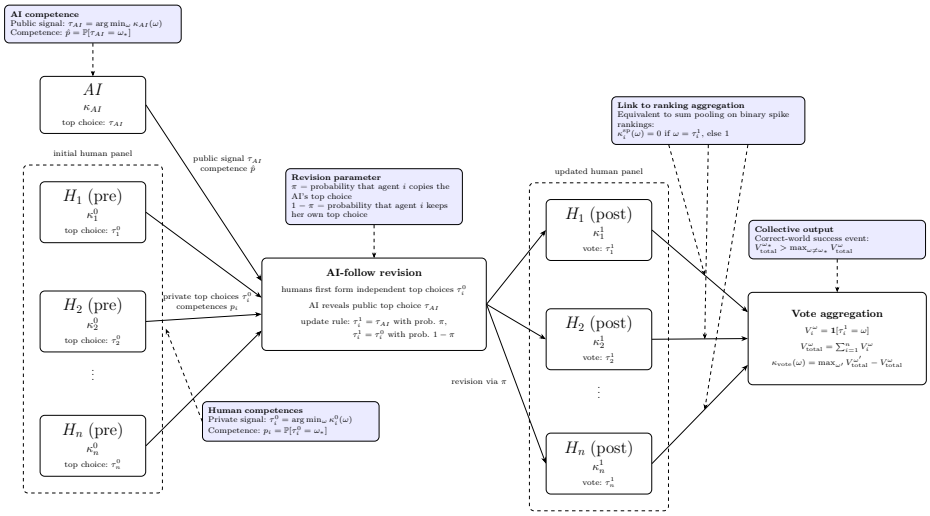

Fig. 1: Overview of the top-choice bridge from ranking aggregation to the opinion leader model. Human agents first form private rankings κ_i^0, from which only their top worlds τ_i^0 are used. The AI provides its own ranking κ_{AI} and public top choice τ_{AI}. During the AI-follow revision step, each agent either copies the AI's top choice with probability π or keeps her own top choice with probability $1-\pi$, yielding revised top choices τ_i^1. These revised top choices are then aggregated into vote totals V_{total}^ω, which induce a collective ranking κ_{vote}.

Thus, a world receives a lower collective rank exactly when it receives more votes, and the worlds with rank 0 are precisely the vote winners. Although our probabilistic analysis operates only at the top-choice level, this construction still yields a collective ranking over all worlds. This construction also links the voting model back to our earlier aggregation rules. For each agent i, define the binary top-choice ranking

$$\kappa_i^{\text{sp}}(\omega) := \begin{cases} 0 & \text{if } \omega = \tau_i^1, \\ 1 & \text{otherwise.} \end{cases} \tag{10}$$

If we apply sum pooling to these binary rankings, the resulting raw score of world ω is

$$s_\Sigma(\omega) = \sum_{i=1}^n \kappa_i^{\text{sp}}(\omega) = n - V_{\text{total}}^\omega. \tag{11}$$

Hence, minimizing the sum-pooled score is equivalent to maximizing the vote score. Thus, the vote-induced ranking κ_{vote} coincides with sum pooling on the revised top-choice rankings. To provide the reader a more detailed understanding of the full pipeline, we illustrate the full formal process in Figure 1.

Motivation for Top-Choice Rankings. As outlined above, in ranking functions, worlds of rank 0 are the most plausible and, in belief-revision terms, correspond

to the agent's current belief set. We therefore focus on the worlds assigned rank 0, i.e., the agent's top choices. For short interactions, this is a natural abstraction: agents need only communicate what they currently take to be correct, rather than their full ranking. It also matches plurality-style aggregation, where each agent votes only for her top-ranked world.

4 Probabilistic Top-Choice Guarantees

In this section, we derive probabilistic guarantees for the top-choice aggregation model introduced above. More specifically, we study a mixed human–AI panel in which agents are represented by their top-ranked worlds and the collective outcome is obtained by sum pooling, equivalently, plurality aggregation over revised top choices.

Proof idea. The proof proceeds in five steps and follows the overall strategy of the opinion-leader proof in [14], while adapting it to our plurality-based top-choice model and refined competitor-wise bound. We first fix a competitor $\omega_\dagger$ and express the pairwise plurality margin between ω_* and $\omega_\dagger$ as a sum of agent-level random variables. We then condition on the AI top choice, compute the corresponding conditional expectation, and show that this mean margin remains positive. Next, we apply Hoeffding's inequality to the sum of these independent bounded random variables. Finally, we use a union bound over all competitors and average over the AI signal.

Throughout this section, we condition on the event that the true world is $\omega_* \in \Omega$, and we continue to use the notation introduced in Section 3. In particular, τ_i^0 and τ_{AI} denote the human and AI top choices, X_i^ω and X_{AI}^ω the corresponding indicator variables, and V_i^ω together with V_{total}^ω the post-revision vote indicators and total vote scores. To prepare the proof, we introduce the competence profiles

$$p_i^\omega := \mathbb{P}(\tau_i^0 = \omega) = \mathbb{P}(X_i^\omega = 1), \qquad \bar{p}^\omega := \frac{1}{n}\sum_{i=1}^{n} p_i^\omega,$$

for every agent $i \in \mathcal{N}$ and world $\omega \in \Omega$. This extends the notation from Section 3, where $p_i = p_i^{\omega_*}$ and $\bar{p} = \bar{p}^{\omega_*}$. Moreover, for every competitor $\omega_\dagger \in \Omega \setminus \{\omega_*\}$, let

$$\Delta p^{\omega_\dagger} := \bar{p}^{\omega_*} - \bar{p}^{\omega_\dagger}.$$

Thus, $\Delta p^{\omega_\dagger}$ is the human panel's pairwise reliability gap between the true world and the competitor $\omega_\dagger$. For the AI, define $\hat{p}^\omega := \mathbb{P}(\tau_{AI} = \omega) = \mathbb{P}(X_{AI}^\omega = 1)$ for every $\omega \in \Omega$. In particular, $\hat{p} := \hat{p}^{\omega_*} = \mathbb{P}(\tau_{AI} = \omega_*) = \mathbb{P}(X_{AI}^{\omega_*} = 1)$.

For each agent $i \in \mathcal{N}$, let $C_i \in \{0, 1\}$ be the copying variable, where $C_i = 1$ means that agent i copies the AI's public top choice and $C_i = 0$ means that agent i keeps her private top choice. Then

$$\tau_i^1 = \begin{cases} \tau_{AI}, & \text{if } C_i = 1, \\ \tau_i^0, & \text{if } C_i = 0, \end{cases} \tag{12}$$

and, equivalently,

$$V_i^{\omega} = C_i\, X_{AI}^{\omega} + (1 - C_i)\, X_i^{\omega} \qquad \text{for every } \omega \in \Omega. \tag{13}$$

Recall that the total plurality score of world ω is $V_{\text{total}}^{\omega} := \sum_{i=1}^{n} V_i^{\omega}$. Accordingly, the probability of strict collective success is written directly as

$$\mathbb{P}\left(V_{\text{total}}^{\omega_*} > \max_{\omega \neq \omega_*} V_{\text{total}}^{\omega} \right).$$

Probabilistic Assumptions. Throughout the proof, we make the following assumptions on the dependence structure and competence profiles of the agents:

(A1) Conditional on the true world ω_*, the private top choices τ_i^0 are independent across all agents $i \in \mathcal{N}$, and the AI top choice τ_{AI} is independent of them.
(A2) Conditional on ω_* and τ_{AI}, the copying variables C_i, $i \in \mathcal{N}$, are i.i.d. Bernoulli random variables with common parameter π, and they are independent of the private top choices τ_i^0.
(A3) For every competitor $\omega_\dagger \in \Omega \setminus \{\omega_*\}$,

$$(1 - \pi)\Delta p^{\omega_\dagger} - \pi > 0. \tag{14}$$

Assumption (14) says that even in the worst case where the AI backs a specific competitor $\omega_\dagger$, the human panel's residual private advantage for the true world remains positive after AI influence is taken into account.

Theorem 1. *Under Assumptions (A1)–(A3),*

$$\begin{aligned}\mathbb{P}\left(V_{\text{total}}^{\omega_*} > \max_{\omega \neq \omega_*} V_{\text{total}}^{\omega} \right) \geq 1 - \sum_{\omega_\dagger \neq \omega_*} \Big[& \hat{p} \exp\Big(-\frac{n}{2} \big((1-\pi)\Delta p^{\omega_\dagger} + \pi \big)^2 \Big) \\ & + \hat{p}^{\omega_\dagger} \exp\Big(-\frac{n}{2} \big((1-\pi)\Delta p^{\omega_\dagger} - \pi \big)^2 \Big) \\ & + \big(1 - \hat{p} - \hat{p}^{\omega_\dagger} \big) \exp\Big(-\frac{n}{2} \big((1-\pi)\Delta p^{\omega_\dagger} \big)^2 \Big) \Big].\end{aligned}$$

Proof. **Step 1: Fix a competitor and define pairwise plurality margin.** Fix $\omega_\dagger \in \Omega \setminus \{\omega_*\}$. For each agent $i \in \mathcal{N}$, define

$$D_i := V_i^{\omega_*} - V_i^{\omega_\dagger}. \tag{15}$$

Thus, V_i^{ω} indicates whether agent i's revised vote goes to ω, while D_i records agent i's pairwise contribution in the comparison between the true world ω_* and the competitor $\omega_\dagger$. Since each agent has exactly one revised top choice, we have

$$D_i \in \{-1, 0, 1\} \subseteq [-1, 1]. \tag{16}$$

In particular, $D_i = 1$ if agent i votes for ω_*, $D_i = -1$ if she votes for $\omega_\dagger$, and $D_i = 0$ otherwise.

$$\sum_{i \in \mathcal{N}} D_i = V^{\omega_*} - V^{\omega_\dagger}. \tag{17}$$

Hence the true world defeats $\omega_\dagger$ exactly if

$$\sum_{i\in\mathcal{N}} D_i > 0.$$

Step 2: Condition on a fixed AI top choice. Fix $\omega_j \in \Omega$ and condition on the event $\tau_{AI} = \omega_j$. By (13), for every $\omega \in \Omega$,

$$V_i^{\omega} = C_i\,\mathbf{1}[\omega_j = \omega] + (1 - C_i)\,X_i^{\omega}.$$

In particular,

$$V_i^{\omega_*} = C_i\,\mathbf{1}[\omega_j = \omega_*] + (1 - C_i)\,X_i^{\omega_*},$$

and

$$V_i^{\omega_\dagger} = C_i\,\mathbf{1}[\omega_j = \omega_\dagger] + (1 - C_i)\,X_i^{\omega_\dagger}.$$

Subtracting these two expressions and using the definition of D_i, we obtain

$$\begin{aligned} D_i &= V_i^{\omega_*} - V_i^{\omega_\dagger} \\ &= \Big(C_i\,\mathbf{1}[\omega_j = \omega_*] + (1 - C_i)\,X_i^{\omega_*}\Big) - \Big(C_i\,\mathbf{1}[\omega_j = \omega_\dagger] + (1 - C_i)\,X_i^{\omega_\dagger}\Big) \\ &= C_i\big(\mathbf{1}[\omega_j = \omega_*] - \mathbf{1}[\omega_j = \omega_\dagger]\big) + (1 - C_i)\big(X_i^{\omega_*} - X_i^{\omega_\dagger}\big). \end{aligned} \tag{18}$$

Intuitively, this expression says that agent i's pairwise contribution between ω_* and $\omega_\dagger$ is determined by the AI's choice with probability π, namely when $C_i = 1$, and by the agent's own private top choice with probability $1 - \pi$, namely when $C_i = 0$.

By Assumptions (A1) and (A2), the random variables D_i, for $i \in \mathcal{N}$, are independent conditional on $\tau_{AI} = \omega_j$. Indeed, once the AI signal is fixed, each D_i depends only on the pair (C_i, τ_i^0), and these pairs are independent across agents.

Step 3: Compute the conditional expectation.

Using (18), together with Assumption (A2), which gives

$$\mathbb{E}[C_i \mid \tau_{AI} = \omega_j] = \pi \qquad \text{and} \qquad \mathbb{E}[1 - C_i \mid \tau_{AI} = \omega_j] = 1 - \pi,$$

we obtain, by linearity of expectation,

$$\mathbb{E}[D_i \mid \tau_{AI} = \omega_j] = \pi\big(\mathbf{1}[\omega_j = \omega_*] - \mathbf{1}[\omega_j = \omega_\dagger]\big) + (1 - \pi)\mathbb{E}\big[X_i^{\omega_*} - X_i^{\omega_\dagger} \mid \tau_{AI} = \omega_j\big]\,. \tag{19}$$

By Assumption (A1), that is, by the conditional independence of τ_{AI} and τ_i^0, this simplifies to

$$\mathbb{E}[D_i \mid \tau_{AI} = \omega_j] = \pi\big(\mathbf{1}[\omega_j = \omega_*] - \mathbf{1}[\omega_j = \omega_\dagger]\big) + (1 - \pi)\big(p_i^{\omega_*} - p_i^{\omega_\dagger}\big). \tag{20}$$

Now define the average conditional margin

$$\mu_j := \frac{1}{n}\sum_{i\in\mathcal{N}} \mathbb{E}[D_i \mid \tau_{AI} = \omega_j]\,. \tag{21}$$

Then

$$\mu_j = \pi\big(\mathbf{1}[\omega_j = \omega_*] - \mathbf{1}[\omega_j = \omega_\dagger]\big) + (1-\pi)\Delta p^{\omega_\dagger}. \tag{22}$$

Hence there are three cases:

$$\mu_{\omega_*} = (1-\pi)\Delta p^{\omega_\dagger} + \pi, \qquad \textit{AI selects the true world} \tag{23}$$
$$\mu_{\omega_\dagger} = (1-\pi)\Delta p^{\omega_\dagger} - \pi, \qquad \textit{AI selects the competitor} \tag{24}$$
$$\mu_j = (1-\pi)\Delta p^{\omega_\dagger}, \quad \omega_j \in \Omega \setminus \{\omega_*, \omega_\dagger\}, \quad \textit{AI selects another wrong world} \tag{25}$$

By Assumption (A3), even the worst-case mean (24) is strictly positive. Thus,

$$\mathbb{E}\left[\sum_{i\in\mathcal{N}} D_i \,\middle|\, \tau_{AI} = \omega_j\right] = n\,\mu_j > 0 \qquad \text{for every } \omega_j \in \Omega. \tag{26}$$

Step 4: Apply Hoeffding's inequality pairwise.

Recall Hoeffding's inequality [11]: if $X_1, \ldots, X_n$ are independent random variables satisfying $l_i \le X_i \le u_i$ almost surely, then for $X := \sum_{i=1}^n X_i$ and every $t > 0$,

$$\mathbb{P}(X - \mathbb{E}(X) \le -t) \le \exp\!\left(-\frac{2t^2}{\sum_{i=1}^n (u_i - l_i)^2}\right).$$

We apply this inequality conditional on $\tau_{AI} = \omega_j$, with $X_i = D_i$, $l_i = -1$, $u_i = 1$, and $t = n\mu_j$. By Step 2, the random variables D_i, $i \in \mathcal{N}$, are independent conditional on $\tau_{AI} = \omega_j$, and by (16) they satisfy $-1 \le D_i \le 1$.

Intuitively, we now bound the probability of *pairwise failure* against $\omega_\dagger$: since strict success requires $V_{\text{total}}^{\omega_*} > V_{\text{total}}^{\omega_\dagger}$, failure is the event $V_{\text{total}}^{\omega_*} \le V_{\text{total}}^{\omega_\dagger}$, where ties already count against success. Therefore,

$$\begin{aligned}
\mathbb{P}\big(V_{\text{total}}^{\omega_*} \le V_{\text{total}}^{\omega_\dagger} \,\big|\, \tau_{AI} = \omega_j\big) &= \mathbb{P}\left(\sum_{i\in\mathcal{N}} D_i \le 0 \,\middle|\, \tau_{AI} = \omega_j\right) \\
&= \mathbb{P}\left(\sum_{i\in\mathcal{N}} D_i - \mathbb{E}\left[\sum_{i\in\mathcal{N}} D_i \,\middle|\, \tau_{AI} = \omega_j\right] \le -n\mu_j \,\middle|\, \tau_{AI} = \omega_j\right) \\
&\le \exp\!\left(-\frac{2n^2(\mu_j)^2}{\sum_{i\in\mathcal{N}}(1-(-1))^2}\right) \\
&= \exp\!\left(-\frac{n}{2}(\mu_j)^2\right).
\end{aligned} \tag{27}$$

Step 5: Pass from one competitor to all competitors and average over the AI signal.

For each competitor $\omega_\dagger \in \Omega \setminus \{\omega_*\}$, Step 4 yields

$$\mathbb{P}\big(V_{\text{total}}^{\omega_*} \le V_{\text{total}}^{\omega_\dagger} \,\big|\, \tau_{AI} = \omega_j\big) \le \exp\!\Big(-\frac{n}{2}\big(\mu_j^{\omega_\dagger}\big)^2\Big), \tag{28}$$

where we now restore the competitor index in $\mu_j^{\omega_\dagger}$.

Hence, conditional on $\tau_{AI} = \omega_j$, the probability that the true world fails to beat at least one competitor is

$$\mathbb{P}\left(\bigcup_{\omega_\dagger \neq \omega_*} \{V_{\text{total}}^{\omega_*} \leq V_{\text{total}}^{\omega_\dagger}\} \,\middle|\, \tau_{AI} = \omega_j \right).$$

By the union bound and (28),

$$\mathbb{P}\left(\bigcup_{\omega_\dagger \neq \omega_*} \{V_{\text{total}}^{\omega_*} \leq V_{\text{total}}^{\omega_\dagger}\} \,\middle|\, \tau_{AI} = \omega_j \right) \leq \sum_{\omega_\dagger \neq \omega_*} \exp\Big(-\frac{n}{2}\big(\mu_j^{\omega_\dagger}\big)^2\Big). \tag{29}$$

Taking expectations over τ_{AI} gives

$$\begin{aligned}
&\mathbb{P}\left(\bigcup_{\omega_\dagger \neq \omega_*} \{V_{\text{total}}^{\omega_*} \leq V_{\text{total}}^{\omega_\dagger}\} \right) \\
&\quad = \sum_{\omega_j \in \Omega} \hat{p}^{\omega_j}\, \mathbb{P}\left(\bigcup_{\omega_\dagger \neq \omega_*} \{V_{\text{total}}^{\omega_*} \leq V_{\text{total}}^{\omega_\dagger}\} \,\middle|\, \tau_{AI} = \omega_j \right) \\
&\quad \leq \sum_{\omega_j \in \Omega} \hat{p}^{\omega_j} \sum_{\omega_\dagger \neq \omega_*} \exp\Big(-\frac{n}{2}\big(\mu_j^{\omega_\dagger}\big)^2\Big) \\
&\quad = \sum_{\omega_\dagger \neq \omega_*} \sum_{\omega_j \in \Omega} \hat{p}^{\omega_j} \exp\Big(-\frac{n}{2}\big(\mu_j^{\omega_\dagger}\big)^2\Big).
\end{aligned} \tag{30}$$

Now split the inner sum according to the three cases (23)–(25):

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{\omega_\dagger \neq \omega_*} \{V_{\text{total}}^{\omega_*} \leq V_{\text{total}}^{\omega_\dagger}\} \right) \leq \sum_{\omega_\dagger \neq \omega_*} \Big[& \hat{p} \exp\Big(-\frac{n}{2}\big((1-\pi)\Delta p^{\omega_\dagger} + \pi\big)^2\Big) \\
& + \hat{p}^{\omega_\dagger} \exp\Big(-\frac{n}{2}\big((1-\pi)\Delta p^{\omega_\dagger} - \pi\big)^2\Big) \\
& + \big(1 - \hat{p} - \hat{p}^{\omega_\dagger}\big) \exp\Big(-\frac{n}{2}\big((1-\pi)\Delta p^{\omega_\dagger}\big)^2\Big) \Big].
\end{aligned} \tag{31}$$

Taking complements yields the bound stated in Theorem 1.

This expression gives a direct lower bound on collective success under the new plurality theorem. In particular, it makes transparent how the guarantee depends on the panel size n, the competitor-wise human reliability gaps $\Delta p^{\omega_\dagger}$, the full AI competence profile $(\hat{p}^{\omega})_{\omega \in \Omega}$, and the strength of AI influence π.

Example 1. Suppose there are $m = 3$ candidate worlds and $n = 50$ human agents. Let $\Delta p^{\omega_\dagger} = 0.3$ for each competitor $\omega_\dagger \neq \omega_*$, so that before seeing the AI's advice, the human panel is on average at least 0.3 more likely to place the

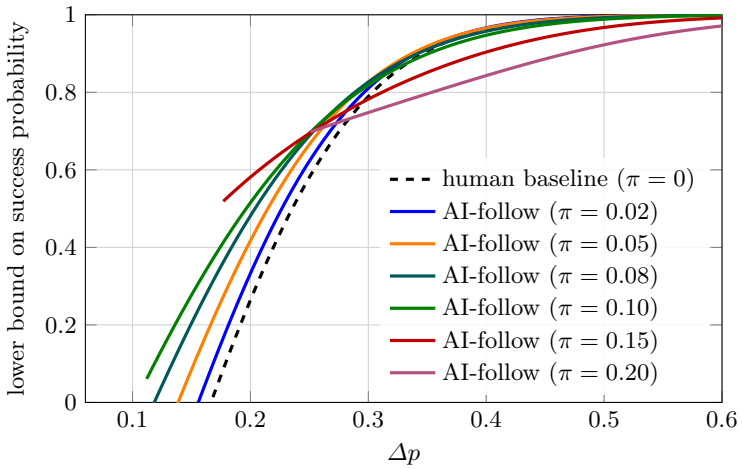

Fig. 2: Lower bounds from Theorem 1 as a function of Δp for different influence levels π, in a parameter setting where moderate AI influence can be beneficial.

true world at the top than either competing world. Suppose further that the AI competence profile is given by $\hat{p} = \hat{p}^{\omega_*} = 0.8, \hat{p}^{\omega_\dagger} = 0.1$for each $\omega_\dagger \neq \omega_*$, and let $\pi = 0.05$, so that each human copies the AI's top choice with probability 0.05. Then Theorem 1 yields $\mathbb{P}(V_{\text{total}}^{\omega_*} > \max_{\omega \neq \omega_*} V_{\text{total}}^{\omega}) \geq 0.8268$. For comparison, the unassisted human baseline ($\pi = 0$) yields the lower bound success probability of 0.7892. In this parameter setting, the new bound therefore certifies that moderate AI influence is epistemically beneficial.

We further illustrate the lower bound from Theorem 1 in Figure 2. The figure plots the guaranteed success probability as a function of the human reliability gap Δp for different influence levels π. To make the example concrete, we fix $m = 3$, $n = 50$, $\hat{p} = 0.8$, and $\hat{p}^{\omega_\dagger} = 0.1$ for each competitor $\omega_\dagger \neq \omega_*$. The dashed black curve shows the unassisted human baseline ($\pi = 0$). In this parameter setting, moderate AI influence can be epistemically beneficial: some assisted curves lie above the baseline because the AI selects the true world with high probability. As required by Assumption (A3), each curve is shown only on the range where $(1 - \pi)\Delta p^{\omega_\dagger} - \pi > 0$.

5 Conclusion

In this work, we introduced, to the best of our knowledge, the first formal model of a ranking-based human–AI aggregation scenario. We focused on AI-follow assistance with concurrent panel revision, in which human agents first form beliefs independently and only then may revise them after the AI's opinion is revealed. To analyze this setting probabilistically, we worked with a top-choice abstraction of ranking functions and translated the resulting aggregation problem into

a plurality-based voting framework. This allowed us to connect our model to the opinion-leader literature while deriving new probabilistic guarantees tailored to our setting. In particular, our results yield theoretical conditions under which AI influence preserves a positive truth-tracking margin and provide lower bounds on the probability that the true world is selected collectively. In this way, the paper gives a first formal account of when AI-induced dependence, such as anchoring, can be epistemically harmful and when it can be beneficial. At the same time, the resulting vote totals still induce a collective ranking over all worlds, even though the current guarantees apply only at the top-choice level.

A natural next step is to replace the simple copying mechanism studied here by richer models of belief revision. This is particularly relevant because the central interaction step in our framework, namely AI-follow revision in Figure 1, is itself a revision step: each human agent moves from an initial state H_i (pre) to a revised state H_i (post) after receiving the AI's advice. In the present paper, we model this transition only through a top-choice copying rule. A more refined account could instead analyze the revision process by means of belief change operators over sets of possible worlds [8,15]. Under the current top-choice abstraction, each agent's pre-revision state is represented by a single top world τ_i^0, and the AI contributes a single top world τ_{AI}. In this simplified setting, a credibility-limited revision operator $\circ$ [9] can be used to express the choice between retaining the original top world and switching to the top world suggested by the AI. Under the present top-choice abstraction, any belief change can only amount to replacing the agent's original top world by the AI's top world. A natural generalization is to allow a top-choice set of worlds rather than a single top-choice world. For standard revision operators, this is equivalent to allowing an arbitrary formula of the language to represent the agent's beliefs. This in turn opens the way to non-prioritized belief change operators, in which the new information need not take priority over the old, and which may capture richer ways in which human agents respond to AI advice, for instance while accounting for doubt [7] or relevance-sensitive attitudes [2]. It may also be possible to model both the transition to H_i (pre) and the subsequent revision by AI advice as iterative belief change processes [3]. More broadly, a central next step is to extend the present top-choice bridge to full ranking aggregation. The current probabilistic guarantees apply only to the event that the true world is placed at the top of the collective outcome, since the analysis collapses each agent's ranking to a single top choice before aggregation. A full-ranking extension would make it possible to investigate truth-tracking beyond the top-ranked world and to ask whether AI influence improves or worsens the quality of the collective ranking as a whole. A further natural extension would be to replace the common influence parameter π by agent-specific parameters π_i, for instance by incorporating average correlation as done recently in [13].

References

1. Carter, L., Liu, D.: How was my performance? exploring the role of anchoring bias in ai-assisted decision making. International Journal of Information Management

82, 102875 (2025)
2. Casini, G., Meyer, T., Varzinczak, I.: Simple conditionals with constrained right weakening. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19. pp. 1632–1638. International Joint Conferences on Artificial Intelligence Organization (7 2019)
3. Darwiche, A., Pearl, J.: On the logic of iterated belief revision. Artif. Intell. **89**(1-2), 1–29 (1997)
4. Everaere, P., Konieczny, S., Marquis, P.: The strategy-proofness landscape of merging. Journal of Artificial Intelligence Research **28**, 49–105 (2007)
5. Everaere, P., Konieczny, S., Marquis, P.: The epistemic view of belief merging: Can we track the truth? In: Proceedings of the 19th European Conference on Artificial Intelligence. pp. 621–626 (2010)
6. Gomez, C., Cho, S.M., Ke, S., Huang, C.M., Unberath, M.: Human-ai collaboration is not very collaborative yet: A taxonomy of interaction patterns in ai-assisted decision making from a systematic review. Frontiers in Computer Science **6**, 1521066 (2025)
7. Grimaldi, D., Martinez, M.V., Rodriguez, R.O.: Moderated revision. International Journal of Approximate Reasoning **166**, 109–126 (2024)
8. Grove, A.: Two modellings for theory change. Journal of Philosophical Logic **17**(2), 157–170 (1988)
9. Hansson, S.O., Fermé, E., Cantwell, J., Falappa, M.: Credibility limited revision. Journal of Symbolic Logic **66**, 1581–1596 (2001)
10. Henderson, E.J., Rubin, G.P.: The utility of an online diagnostic decision support system (isabel) in general practice: a process evaluation. JRSM short reports **4**(5), 1–11 (2013)
11. Hoeffding, W.: Probability inequalities for sums of bounded random variables. Journal of the American statistical association **58**(301), 13–30 (1963)
12. Huber, F.: Ranking functions and rankings on languages. Artificial Intelligence **170**(4-5), 462–471 (2006)
13. Karge, J.: Taming dilation in imprecise pooling. In: International Conference on Principles and Practice of Multi-Agent Systems. pp. 428–443. Springer (2024)
14. Karge, J., Burkhardt, J.M., Rudolph, S., Rusovac, D.: To lead or to be led: a generalized condorcet jury theorem under dependence. In: Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems. pp. 983–991 (2024)
15. Katzuno, H., Mendelson, A.: Propositional knowledge base revision and minimal change. Tech. rep., Technical Report n°3, Department of Computer Science, University of Toronto (1991)
16. Marquis de Condorcet, M.J.A.N.C.: Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix. Imprimerie Royale, Paris (1785)
17. Peng, K., Garg, N., Kleinberg, J.: A no free lunch theorem for human-ai collaboration. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 14369–14376 (2025)
18. Shneiderman, B.: Human-centered artificial intelligence: Reliable, safe & trustworthy. International Journal of Human–Computer Interaction **36**(6), 495–504 (2020)
19. Spohn, W.: Ordinal conditional functions: A dynamic theory of epistemic states. In: Causation in decision, belief change, and statistics: Proceedings of the Irvine Conference on Probability and Causation. pp. 105–134. Springer (1988)
20. Spohn, W.: The laws of belief: Ranking theory and its philosophical applications. Oxford University Press (2012)

21. Vaccaro, M., Almaatouq, A., Malone, T.: When combinations of humans and ai are useful: A systematic review and meta-analysis. Nature Human Behaviour **8**(12), 2293–2303 (2024)