First Steps Towards Human-AI Ranking 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 a first formal model of ranking-based human–AI aggregation. In our model, we represent both human and AI judgments by ranking functions over possible worlds and 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 voting problem and connects our setting to a recent model from the voting literature. This allows us to model a specific interaction setting that we refer to as *AI-follow assistance with concurrent panel revision*. The resulting analysis yields a formal account of when AI influence can be epistemically harmful by inducing dependence among human agents.

Keywords

Human–AI Interaction, Ranking Aggregation, Epistemic Social Choice, Belief Change

1. Introduction

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 [1]. 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 [2]. 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 [3]. 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 [4]. They map possible worlds to natural numbers, where higher numbers represent greater implausibility and worlds of rank 0 are considered most plausible [4]. 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, a common model of collective decision-making. This allows us to evaluate the collective belief *epistemically* through the so-called *Condorcet Jury Theorem* (CJT), a classical result in voting theory that provides probabilistic guarantees for the correctness of a collective decision [5]. Since the original CJT relies on assumptions about the voting process that rarely hold in practice, we instead use a recent generalization [6] that, in particular, allows us to quantify the dependence induced by the AI model, which is a major risk in human–AI interaction settings.

SKILLED-LLMs 2026: Joint Workshop on Statistics and Knowledge Integration for Logic, Learning, Ethical Decisions, and LLMs, July 18, 2026, Lisbon, Portugal

jonas.karge@tu-dresden.de (J. Karge); roy.ferguson@uct.ac.za (R. Ferguson); dgrimaldi@dc.uba.ar (D. Grimaldi); jonas@haldimann.de (J. Haldimann); madzimeruvarashe@gmail.com (R. Madzime); tmeyer@airu.org.za (T. Meyer)

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 harmful.

2. Related Work

Our work sits at the intersection of ranking-based belief representation, epistemic aggregation, truth-tracking voting theory, and human–AI interaction. The key gap is that these strands have largely developed separately. In particular, while there is already a rich theory of probabilistic truth-tracking for voting, noisy rankings, and belief merging, these ideas have not yet been brought together in a formal model of ranking-based human–AI aggregation in which AI advice influences human judgments before aggregation. As a first step, we address this gap through a top-choice abstraction that allows us to import probabilistic guarantees from the voting literature, while leaving the extension to full ranking aggregation for future work.

A first relevant line of work concerns ranking functions themselves. Spohn (1988, 2012) develops ranking functions as a formal theory of graded belief in which each possible world receives a degree of disbelief and the most plausible worlds receive rank 0. A ranking on worlds can then be extended to a ranking on all formulas built from these worlds. Huber (2006) recovers a ranking on worlds from a ranking on more complex formulas. We build on this tradition by using ranking functions as the shared epistemic language for both human and AI judgments.

A second relevant line of work comes from belief merging. Classical work by Everaere et al. (2007) studies how several epistemic states can be combined under rationality and fairness constraints. More closely related to our concerns, Everaere et al. (2010) develop an explicitly epistemic view of merging, in which the merged outcome is evaluated by whether it identifies an unknown true world. This connects belief merging to truth-tracking questions of the kind we study here, and it has recently been generalized to a less restrictive voting setting [11]. Relatedly, this has been extended to the aggregation of numerical estimates [12]. Our work continues this epistemic line, but shifts attention to a new dependence source: the case in which an AI system influences the agents' beliefs.

A third line of work comes from voting theory, where the main probabilistic tool used in this paper originates. Caragiannis et al. (2016) study the probability that a voting rule recovers the correct underlying ranking when voters have imperfect information. Procaccia et al. (2015) show that truth-tracking guarantees can persist even when voters are not fully independent, for instance because information flows through a social network. Most directly relevant for us is the recent result by Karge et al. (2024), who prove correctness guarantees for a voting model with an opinion leader, i.e. an external factor that induces correlation in the electorate.

Finally, our work relates to the growing literature on human–AI collaboration. Recent studies show that when people see AI advice early, they may anchor on it and shift their judgment toward the AI's suggestion [15, 16]. In our setting, this matters because such anchoring creates dependence between human judgments and thereby threatens the epistemic benefit of group aggregation. Recent theoretical work sharpens this concern: Peng et al. (2025) show that hybrid human–AI systems do not automatically outperform their components and may even perform worse than the least accurate contributor. From a more normative perspective, Almada (2019) argues that decision-support systems should preserve the human's capacity to question and contest AI advice. What this literature still lacks, however, is a formal model of how AI influence propagates through a group and affects the quality of collective aggregation.

3. 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 [4, 7]. 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

$$\kappa : \Omega \to \mathbb{N}_0 \cup \{\infty\} \tag{1}$$

such that

$$\min_{\omega \in \Omega} \kappa(\omega) = 0. \tag{2}$$

Thus, at least one world is assigned rank 0 and is treated as maximally plausible. Worlds with positive rank are disbelieved to some degree, and worlds with rank ∞ are ruled out.

Following Huber's proposition-level formulation, we extend κ from worlds to propositions. Let $\mathcal{A}$ be a field of subsets of Ω, that is, a family of propositions closed under complement, finite unions, and intersections [8]. For each proposition $A \in \mathcal{A}$, we define

$$\kappa(A) := \min_{\omega \in A} \kappa(\omega), \qquad \kappa(\emptyset) := \infty. \tag{3}$$

In words, a proposition is as plausible as its most plausible way of being true. This immediately yields the familiar properties [8]:

$$\kappa(A \cup B) = \min\{\kappa(A), \kappa(B)\} \tag{4}$$

and, whenever $A \subseteq B$,

$$\kappa(B) \leq \kappa(A). \tag{5}$$

Example 3.1 (Running diagnosis example)**.** Let

$$\Omega = \{\omega_{\mathit{flu}}, \omega_{\mathit{cold}}, \omega_{\mathit{pneumonia}}\}$$

be a diagnostic state space, and suppose

$$\kappa(\omega_{\mathit{flu}}) = 0, \quad \kappa(\omega_{\mathit{cold}}) = 1, \quad \kappa(\omega_{\mathit{pneumonia}}) = 5.$$

Then flu is most plausible, cold is less plausible, and pneumonia is highly implausible. At the proposition level,

$$\kappa(\{\omega_{\mathit{cold}}, \omega_{\mathit{pneumonia}}\}) = 1 \quad \text{and} \quad \kappa(\{\omega_{\mathit{flu}}, \omega_{\mathit{cold}}\}) = 0,$$

because propositions inherit the rank of their most plausible world.

3.1. Aggregating Multiple Ranking Functions

We now move from individual belief states to collective ones. We represent each agent's state for aggregation purposes by a pointwise ranking over Ω. Thus, suppose that each agent $i \in \{1, \ldots, n\}$ reports a ranking function

$$\kappa_i : \Omega \to \mathbb{N}_0 \cup \{\infty\}. \tag{6}$$

The input profile is the tuple

$$E = (\kappa_1, \ldots, \kappa_n). \tag{7}$$

An aggregation operator Δ maps this profile to a collective ranking function:

$$\kappa_{\text{agg}} = \Delta(\kappa_1, \ldots, \kappa_n). \tag{8}$$

This means that we pool the agents' scores world by world and thereby obtain a group-level belief state. As a simple baseline inspired by the merging literature [9], we consider *sum pooling*. Since pooling may

produce a raw world-score function whose minimum is not yet 0, we normalize the output. Given any pooled world-score function $f : \Omega \to \mathbb{N}_0 \cup \{\infty\}$ with $\min_{\omega\in\Omega} f(\omega) < \infty$, define

$$\mathrm{norm}(f)(\omega) := f(\omega) - \min_{\omega'\in\Omega} f(\omega'). \tag{9}$$

Then $\min_{\omega\in\Omega} \mathrm{norm}(f)(\omega) = 0$, so $\mathrm{norm}(f)$ is again a valid ranking function. Given a profile $E = (\kappa_1, \ldots, \kappa_n)$, define the raw sum-pooled world-score function by

$$s^E_\Sigma(\omega) := \sum_{i=1}^{n} \kappa_i(\omega). \tag{10}$$

The corresponding normalized collective ranking is

$$\kappa_\Sigma := \mathrm{norm}(s^E_\Sigma). \tag{11}$$

Example 3.2 (Three-agent aggregation)**.** Consider again the diagnostic space

$$\Omega = \{\omega_{flu}, \omega_{cold}, \omega_{pneumonia}\},$$

and suppose three agents report the following world-rankings:

World ω	Agent 1	Agent 2	Agent 3
ω_{flu}	0	3	3
ω_{cold}	4	0	0
$\omega_{pneumonia}$	6	5	5

Sum pooling yields $s^E_\Sigma = (6, 4, 16) \rightsquigarrow \kappa_\Sigma = (2, 0, 12)$. So ω_{cold} is the top-ranked collective diagnosis.

Motivation for Ranking Functions in Human–AI Interaction. We conclude this section by briefly explaining why ranking functions are a useful representation in a human–AI interaction setting.

Ranking functions are well-established representations of epistemic states in belief change and nonmonotonic reasoning [4, 19, 20, 21]. They can furthermore be viewed as numerical implementations of total preorders, which are another common way of representing epistemic states in nonmonotonic reasoning and belief change [22, 23, 24]. Compared with bare orderings, however, ranking functions are more expressive and support arithmetic operations, which lead to useful formal properties in some settings [25, 26].

A further reason to work with ranking functions is their close connection to the belief revision literature. Ranking functions are not only static representations of plausibility; they also support well-studied update and revision operations [27]. This makes them attractive in settings where beliefs may change in response to new information. In particular, they provide a natural formal language for studying how epistemic states can be represented, compared, and aggregated when agents express structured judgments over a space of possible worlds.

4. Human–AI Interaction

The aggregation framework introduced above does not operate in a vacuum. In human–AI settings, belief states are shaped not only by the information available to the agents, but also by the way that information is presented, timed, and exchanged. For this reason, the same aggregation operator may behave very differently depending on the interaction pattern through which human and artificial agents arrive at their beliefs. In the present setting, "AI" should be understood broadly as a task-specific algorithmic decision aid. Depending on the application, this may include statistical predictors, language-model-based systems, or broader decision support systems that integrate several computational components into a human workflow and that contribute structured epistemic input to the joint decision process.

A particularly relevant class of cases arises in diagnostic decision support. Here an AI system may process patient information and return a ranked list of candidate diagnoses, thereby producing an output that can be represented in the same ranking-based language as the human agents' beliefs. As a motivating example, *Isabel* [28] is an online diagnostic decision support system that helps clinicians construct or broaden a differential diagnosis list based on clinical features [29].

4.1. Interaction Patterns

Recent work in the human–AI interaction literature has proposed taxonomies of the main ways in which AI systems enter human decision-making workflows.

One important pattern is *AI-First Assistance*. In this setup, the decision problem and the AI's recommendation are presented at the same time, or the AI's output is available from the beginning of the task. The human may then accept, modify, or override the AI's recommendation. This is a natural pattern for highly integrated support tools, but it also creates a direct route for the AI system to shape the human's initial hypothesis generation [16]. A second important pattern is *AI-Follow Assistance*. Here the human first forms an independent preliminary judgment, and only afterward receives the AI's recommendation. The AI output can then be used as a second opinion, a check against omissions, or a prompt for reassessment [16].

Beyond these basic patterns, the literature also distinguishes more interactive forms of assistance. Under *Request-Driven Assistance*, the human actively decides when to consult the AI system, for example by clicking a button to request help. Under *Secondary Assistance*, the AI does not provide a direct answer, but rather supporting information such as risk profiles or auxiliary evidence that the human must interpret. Under *AI-Guided Dialogic Engagement*, the AI and the human exchange information over several turns. Under *User-Guided Interactive Adjustments*, the user changes inputs, assumptions, or constraints in order to observe how the AI's output responds [16]. In the present paper we focus on the simpler static cases first.

4.2. Risks and Challenges

Human–AI collaboration is often attractive because it seems to promise a combination of complementary strengths. Yet mixed panels also introduce distinctive epistemic risks.

- A central concern is *anchoring bias*. When AI assistance is available from the outset, human agents may anchor too strongly on the AI's suggestion [15]. In the present framework, this means that the human ranking may no longer be treated as independent.
- A second concern is the broader possibility that human–AI collaboration may fail to produce any epistemic gain at all. Recent theoretical work has shown that, without further structural assumptions, hybrid human–AI systems need not outperform their components and can even perform worse than the least accurate individual contributor [17].
- A third concern is that the collaborative integration of AI output is often far less structured than the output format itself. Even when an AI system produces a simple ranked list of candidate diagnoses, the way in which humans interpret, revise, defer to, or ignore that ranking can vary dramatically across interaction settings [30].

In what follows, we address these issues by embedding one specific interaction pattern, namely, *AI-follow with concurrent panel revision*, into an existing voting framework with opinion leader influence, a classical dependence model from the voting literature. This move gives us a simple and structured account of how information flows through the panel. It also lets us model the dependence between the AI and the human agents explicitly. Finally, it allows us to draw on established guarantees from the opinion leader model to quantify the probability that the collective ranking places the true world at the top, and thus to study the epistemic gain of AI assistance under different parameter settings.

4.3. A First Model: AI-Follow with Concurrent Panel Revision

We begin with a deliberately simple setting. A panel consists of several human agents and one AI system. Each human agent first forms an initial ranking κ_i^0 on her own. The AI then provides its ranking κ_{AI}. After seeing this advice, the human agents revise their rankings in parallel, which yields updated rankings κ_i^1. We then aggregate only these revised human rankings into a collective outcome.

We choose this setting for two reasons. First, it gives us a clear human baseline. Because the human agents form κ_i^0 before they see the AI's advice, we can compare the assisted panel with the unassisted one. This lets us ask the central question to this work: when does AI assistance decrease the epistemic quality of the collective judgement?

Second, this interaction pattern permits to formalize the dependence structure. Since all human agents revise at the same time after receiving the same AI signal, we can study how AI advice affects their rankings without also having to model sequential deliberation among the humans. That makes concurrent revision a good starting point for a first formal analysis.

5. A Top-Choice Bridge to Opinion Leader Influence

In this section, we connect our ranking-based framework to the voting model recently proposed by Karge et al. [6]. Their model provide a generalization of the Condorcet Jury Theorem, providing probabilistic bounds on the collective belief tracking the truth. 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 $\mathcal{N} = \{a_1, \ldots, a_n\}$, where each agent a_i may approve any number of alternatives, or worlds, from a finite set $W = \{\omega_1, \ldots, \omega_m\}$. For each world ω_j, agent a_i is assumed to approve ω_j with some probability $p_i^{\omega_j}$. Since exactly one alternative is assumed to represent the ground truth, denoted by ω_*, the probability $p_i^{\omega_*}$ is referred to as the agent's competence. The alternative that receives the most approvals wins the *approval vote* and thereby represents the collective opinion. If ω_* receives strictly more approvals than any competitor, we count this as a success.

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 approves alternatives and affects the electorate through a uniform influence parameter π, that is, the probability that an agent copies the OL's choice. Moreover, the OL has its own competence parameter $\hat{p}$, that is, the probability that it approves the true world. Thus, each agent may either copy the opinion leader's belief or vote according to her own belief.

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 import the bounds from the opinion leader setting 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:

$$\tau_i^0 := \arg\min_{\omega \in \Omega} \kappa_i^0(\omega), \qquad \tau_{AI} := \arg\min_{\omega \in \Omega} \kappa_{AI}(\omega), \tag{12}$$

assuming for simplicity that these top worlds are unique, i.e., every agent assignes 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; here, $\arg\min$ returns a world attaining the minimum rank.

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

$$X_i^\omega := \mathbf{1}[\tau_i^0 = \omega], \qquad X_{AI}^\omega := \mathbf{1}[\tau_{AI} = \omega]. \tag{13}$$

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, that is, 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:

$$p_i := \mathbb{P}(\tau_i^0 = \omega_*) = \mathbb{P}(X_i^{\omega_*} = 1), \qquad \hat{p} := \mathbb{P}(\tau_{AI} = \omega_*) = \mathbb{P}(X_{AI}^{\omega_*} = 1), \tag{14}$$

and the average human competence

$$\bar{p} := \frac{1}{n} \sum_{i=1}^{n} p_i. \tag{15}$$

Intuitively, p_i and $\hat{p}$ represent the probability that the human agent, respectively the AI, ranks the true world first.

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 [6], 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:

$$\tau_i^1 = \begin{cases} \tau_{AI} & \text{with probability } \pi, \\ \tau_i^0 & \text{with probability } 1 - \pi. \end{cases} \tag{16}$$

Equivalently, the final vote indicator for world ω is

$$V_i^\omega := \mathbf{1}[\tau_i^1 = \omega], \tag{17}$$

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 $\omega \in \Omega$,

$$\mathbb{P}(V_i^\omega = 1) = \pi\, \mathbb{P}(X_{AI}^\omega = 1) + (1 - \pi)\, \mathbb{P}(X_i^\omega = 1). \tag{18}$$

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

$$V_{\text{total}}^\omega := \sum_{i=1}^{n} V_i^\omega. \tag{19}$$

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

The correct world wins whenever

$$V_{\text{total}}^{\omega_*} > \max_{\omega \neq \omega_*} V_{\text{total}}^\omega. \tag{20}$$

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

$$\kappa_{\text{vote}}(\omega) := \max_{\omega' \in \Omega} V_{\text{total}}^{\omega'} - V_{\text{total}}^\omega. \tag{21}$$

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

$$\kappa_i^{\mathrm{sp}}(\omega) := \begin{cases} 0 & \text{if } \omega = \tau_i^1, \\ 1 & \text{otherwise.} \end{cases} \tag{22}$$

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

$$s_\Sigma(\omega) = \sum_{i=1}^{n} \kappa_i^{\mathrm{sp}}(\omega) = n - V_{\mathrm{total}}^{\omega}. \tag{23}$$

Hence, minimizing the sum-pooled score is equivalent to maximizing the vote score. In this special case, the vote-induced ranking κ_{vote} coincides with sum pooling on the revised top-choice rankings.

Example 5.1 (From post-revision rankings to a collective ranking)**.** Consider again the diagnostic space

$$\Omega = \{\omega_{flu}, \omega_{cold}, \omega_{pneumonia}\},$$

and suppose that after the revision step, four human agents have the following rankings:

World ω	Agent 1	Agent 2	Agent 3	Agent 4
ω_{flu}	0	2	0	0
ω_{cold}	1	0	2	1
$\omega_{pneumonia}$	4	3	5	3

Post-revision rankings κ_i^1.

The corresponding top worlds are

$$\tau_1^1 = \omega_{flu}, \qquad \tau_2^1 = \omega_{cold}, \qquad \tau_3^1 = \omega_{flu}, \qquad \tau_4^1 = \omega_{flu}.$$

Hence the vote totals are

$$V_{\mathrm{total}}^{\omega_{flu}} = 3, \qquad V_{\mathrm{total}}^{\omega_{cold}} = 1, \qquad V_{\mathrm{total}}^{\omega_{pneumonia}} = 0.$$

Using κ_{vote}, we obtain

$$\kappa_{\mathrm{vote}}(\omega_{flu}) = 0, \qquad \kappa_{\mathrm{vote}}(\omega_{cold}) = 2, \qquad \kappa_{\mathrm{vote}}(\omega_{pneumonia}) = 3.$$

Thus, the induced collective ranking is

$$\omega_{flu} \prec \omega_{cold} \prec \omega_{pneumonia}.$$

So, even though only the top worlds of the post-revision rankings enter the voting step, the resulting vote totals still induce a full collective ranking over all worlds.

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. When working with ranking functions, worlds with rank 0 have a special role as they represent the worlds believed to be most plausible. In the context of belief revision, the worlds with rank 0 correspond to the models of the agent's belief set. Analogously, in the context of defeasible reasoning, formulas with rank 0 are the formulas an agent considers plausible if no other context information is given. Therefore, for now, we assume that agents exchange their beliefs about which worlds have rank 0. Intuitively, this seems plausible for short interactions, as it corresponds to an agent telling others what she thinks is correct, without also communicating what she might believe under other circumstances. This assumption also aligns well with belief aggregation by voting, where everyone votes for the preferred choice or choices while ignoring the remainder of the preference ranking.

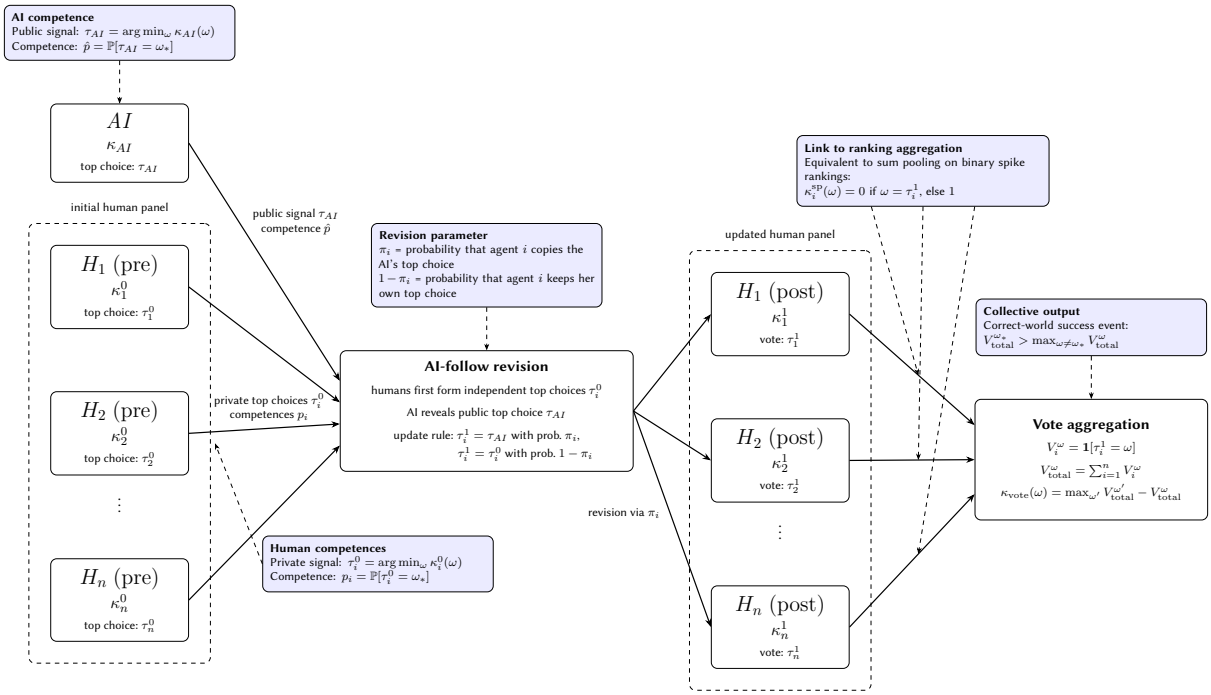

Figure 1: Overview of the top-choice bridge from ranking aggregation to the opinion leader model. Human agents first form private rankings κ_i^0, from which only their top worlds τ_i^0 are used. The AI provides its own ranking κ_{AI} and public top choice τ_{AI}. During the AI-follow revision step, each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

5.1. Vote Aggregation and Sum Pooling

We briefly illustrate the connection between voting and sum pooling. Once we collapse each revised ranking to its top world, each agent contributes a very simple *binary top-choice ranking*: the world she votes for receives score 0, and every other world receives score 1. If we now sum these binary rankings across agents, the total score of a world counts how many agents *did not* vote for it. This is why

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

Here, $V_{\mathrm{total}}^\omega$ is the number of votes that world ω receives, so $n - V_{\mathrm{total}}^\omega$ is the number of agents who did not place ω at the top. As a result, a world gets a low sum-pooled score exactly when it gets many votes. Minimizing the sum-pooled score is therefore the same as maximizing the vote total.

Example 5.2. Suppose $n = 3$ and the revised top choices are

$$\tau_1^1 = \omega_{flu}, \qquad \tau_2^1 = \omega_{cold}, \qquad \tau_3^1 = \omega_{flu}.$$

Then the vote totals are

$$V_{\mathrm{total}}^{\omega_{flu}} = 2, \qquad V_{\mathrm{total}}^{\omega_{cold}} = 1, \qquad V_{\mathrm{total}}^{\omega_{pneumonia}} = 0.$$

The corresponding binary top-choice rankings are

$$\kappa_1^{\mathrm{sp}} = (0, 1, 1), \qquad \kappa_2^{\mathrm{sp}} = (1, 0, 1), \qquad \kappa_3^{\mathrm{sp}} = (0, 1, 1),$$

where the coordinates are ordered as $(\omega_{flu}, \omega_{cold}, \omega_{pneumonia})$. Summing these vectors gives

$$s_\Sigma = (1, 2, 3).$$

So flu has the lowest sum-pooled score, cold comes next, and pneumonia comes last. This is exactly the same ordering we get from the vote totals $(2, 1, 0)$: flu is best because it has the most votes, cold is second, and pneumonia is last.

5.2. Probabilistic Guarantees

In this section, we state the lower bound derived in Karge et al. (2024) and outline the underlying assumptions on the competence and dependence structure necessary to apply these bounds.

More specifically, this lower bound relies on three assumptions, paraphrased from Karge et al. (2024) in the notation of our top-choice framework:

- First, the human agents must be *independent at the private-signal level*: conditional on the true world ω_*, the initial top choices $\tau_1^0, \dots, \tau_n^0$ are independent. Equivalently, the corresponding variables X_i^ω are independent once the true world is fixed.
- Second, the panel must be *reliable on average*. This is what the parameter $\Delta p > 0$ captures: for every competing world $\omega \neq \omega_*$, the average probability that the humans place ω_* first exceeds the average probability that they place ω first by at least Δp. In this sense, Δp measures the panel's built-in margin in favor of the true world before the AI's advice is revealed.
- Third, the AI's influence must not be too strong. Since π is the probability that an agent replaces her own initial top choice τ_i^0 by the AI's top choice τ_{AI}, the opinion leader model requires

$$\pi < \frac{\Delta p}{\Delta p + 1},$$

  which means that the AI may shape the panel's revisions, but not so strongly that it overwhelms the panel's own average tendency to favor the true world.

A Bound on the Correct Top Choice. Adapting the bound of Karge et al. (2024) to our top-choice bridge, we can bound the probability that the correct world is top-ranked in the induced collective ranking by

$$\mathbb{P}\big(\omega_* \text{ wins}\big) \;\geq\; \hat{p}\Big(1 - (m-1)e^{-\frac{n}{2}\Delta p^2(1-\pi)^2}\Big) + (1-\hat{p})\Big(1 - (m-1)e^{-\frac{n}{2}(\Delta p(1-\pi)-\pi)^2}\Big). \tag{24}$$

This expression gives a direct worst-case estimate of epistemic gain under different parameter settings. In particular, it makes transparent how success depends on the size of the panel n, the average reliability gap Δp, the AI's competence $\hat{p}$, and the strength of AI influence π.

Example 5.3 (Computing the bound from the model parameters)**.** Suppose there are $m = 3$ candidate worlds and $n = 50$ human agents. Let $\Delta p = 0.3$, so that before seeing the AI's advice, the human panel is on average at least 0.3 more likely to place the true world at the top than any competing world. Let $\hat{p} = 0.2$, meaning that the AI places the true world first with probability 0.2, and let $\pi = 0.1$, so that each human copies the AI's top choice with probability 0.1. Then Equation (24) yields the lower bound

$$\mathbb{P}\big(\omega_* \text{ wins}\big) \;\geq\; 0.1585.$$

This should be interpreted as a *worst-case lower bound*: it guarantees only that, under the assumptions of the model, the probability that the induced collective ranking places the true world strictly at the top is at least 0.1585.

We illustrate the potential detrimental effect of *anchoring bias* in Figure 2. In the figure, we fix the number of worlds m, the number of agents n, and the AI's competence $\hat{p}$, while varying the human reliability gap Δp on the x-axis and the AI influence parameter π across the different curves. Intuitively, the bound improves when the panel is larger or when the humans have a stronger built-in margin in favor of the true world, that is, when n or Δp increases. Conversely, stronger AI influence π weakens the guarantee when the AI is not sufficiently reliable. In Figure 2, where $\hat{p} = 0.2$, we see that, all else being equal, larger values of π yield uniformly lower success guarantees. In this parameter setting, stronger AI influence is therefore epistemically detrimental to collective accuracy.

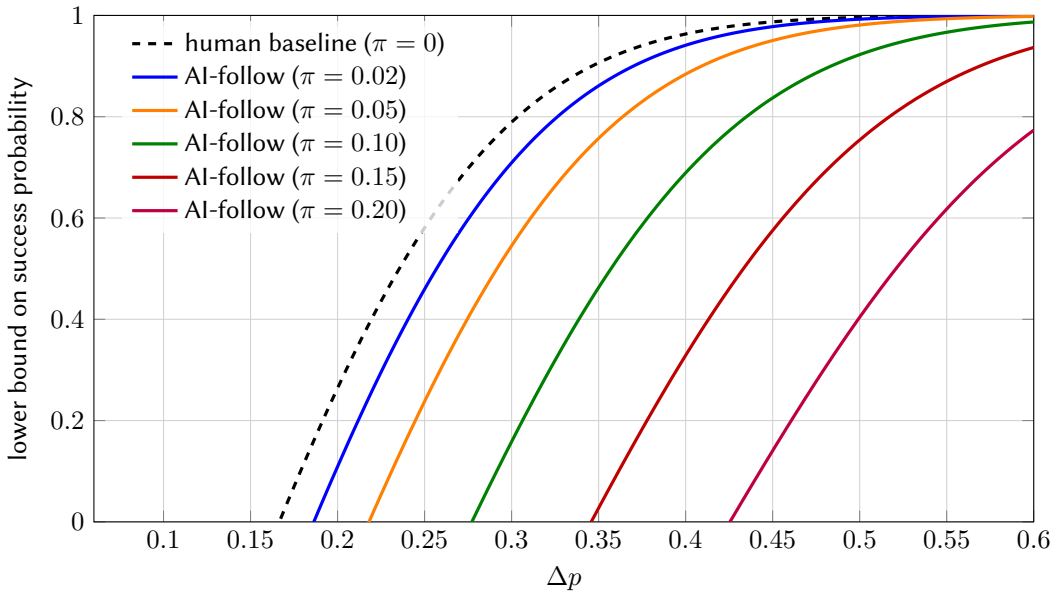

Figure 2: Lower bounds on success probability as a function of Δp for different levels of AI influence. We fix $m = 3$, $n = 50$, and $\hat{p} = 0.2$. The dashed black curve shows the standalone human baseline. Higher values of π lower the guarantee, and this effect becomes more pronounced as AI influence increases.

Epistemically Beneficial versus Epistemically Detrimental. Under these assumptions, Equation (24) gives a lower bound on the probability that the correct world wins after revision and aggregation. In our setting, we read this as a lower bound on the probability that the induced collective ranking places the true world at the top. However, as Figure 3 shows, even when the AI is always correct ($\hat{p} = 1$), increasing AI influence (π) still lowers the guaranteed success probability. This feature is inherited from the way Equation (24) is derived in Karge et al. (2024), and should therefore be interpreted as a property of the imported voting bound rather than as a general claim about ranking-based human–AI aggregation. This nevertheless provides a useful first benchmark: it shows how harmful AI-induced dependence can be studied formally and identifies a special case of our framework in which sum-based aggregation inherits established probabilistic guarantees.[1]

6. Outlook: Belief-Change Perspectives on Panel Revision

As we have indicated at the end of Section 3, ranking functions connect naturally to belief change theory. This is particularly relevant for our setting because the central interaction step, namely AI-follow revision in Figure 1, is 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 simple top-choice copying mechanism. A natural next step is therefore to analyze the revision process itself by means of belief change operators over sets of possible worlds [31, 32].

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 $\circ$ [33] can be used to express the choice between retaining the original top world and switching to the top world suggested by the AI. In this sense, the present update rule can be read schematically as either preserving the current top choice, $\tau_i^0 \circ \tau_{AI} = \tau_i^0$, or adopting the AI's top choice, $\tau_i^0 \circ \tau_{AI} = \tau_{AI}$. 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.

[1] More precisely, in the derivation of Equation (24), the opinion leader is treated in a worst-case way that effectively allows it to approve all incorrect worlds as well. As a result, even when $\hat{p} = 1$, larger values of π still weaken the bound, because stronger copying makes it harder to separate the correct world from its competitors in the voting scores.

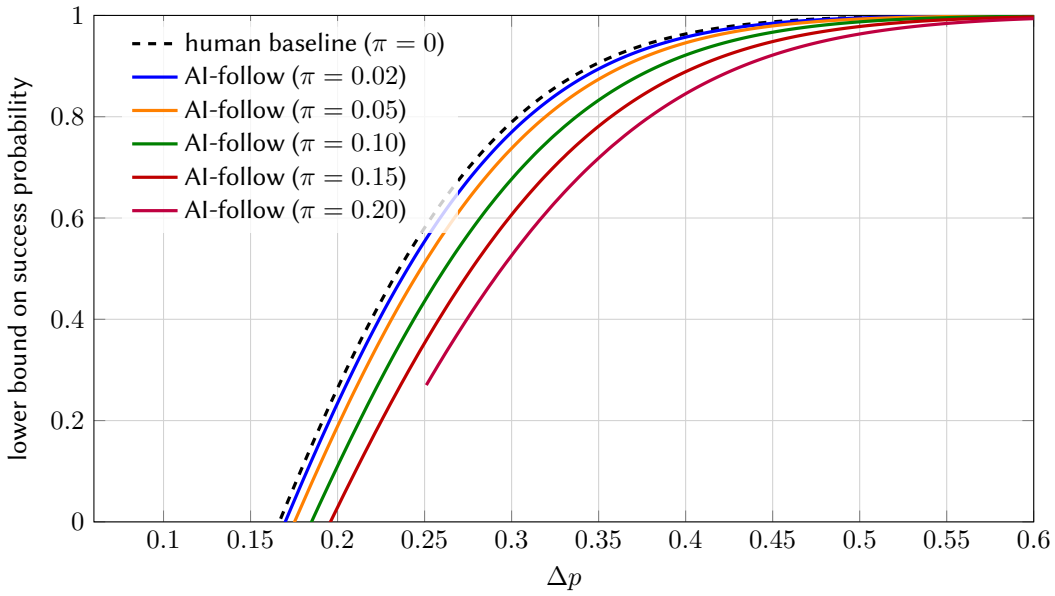

Figure 3: Lower bounds on success probability as a function of Δp for different levels of AI influence. We fix $m = 3$, $n = 50$, and $\hat{p} = 1$. The rightmost curve is truncated on the left because, for $\pi = 0.20$, the bound is only valid for $\Delta p > \pi/(1 - \pi) = 0.25$.

A natural generalization to consider is the case where we 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. Following this generalization, one can investigate other non-prioritized belief change operators, in which the new information does not necessarily take priority over the old, and which model more complex enrichment of the human agents' decisions in light of the AI's advice, for instance while accounting for doubt [34] or relevance-sensitive attitudes [35]. Moreover, it may be possible to model both the transition to H_i (pre) and the subsequent revision by AI advice as iterative belief change processes [23]. Lastly, it would be of interest to investigate how such richer credibility-limited revision operators relate to the probabilistic guarantees appearing in the present paper.

7. Conclusion

Summary. In this work, we presented a first formal model of a ranking-based human–AI aggregation scenario. More specifically, we studied the interaction pattern of AI-follow assistance with concurrent panel revision, in which human agents first form their beliefs independently and 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 voting framework. This allowed us to import probabilistic guarantees from the opinion leader model and to derive a first formal account of when AI influence can be epistemically harmful through the dependence induced by anchoring bias. 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.

Future Work. There are several natural directions for future work. First, the current top-choice bridge should be extended to full ranking functions, so that probabilistic guarantees can be studied beyond the top-ranked world. Second, the panel revision step should be modeled in a richer way by using explicit belief-revision operators rather than a simple copying parameter. Third, it would be important to move beyond the current worst-case perspective and to identify conditions under which AI influence is not only detrimental, but epistemically beneficial. More generally, future work

should clarify how richer models of revision and aggregation interact with probabilistic truth-tracking guarantees in human–AI settings.

8. Acknowledgments

This work is partly supported by BMFTR (Federal Ministry of Research, Technology and Space) in DAAD project 57616814 (SECAI, School of Embedded Composite AI, https://secai.org/) as part of the program Konrad Zuse Schools of Excellence in Artificial Intelligence. This work is based on the research supported in part by the National Research Foundation of South Africa (REFERENCE NO: SAI240823262612). This work was also supported in part by funding from the Mastercard Foundation Scholars Program.

Declaration on Generative AI

During the preparation of this work, the author used GenAI models for the following purposes: grammar and spelling checks; improvements to writing style; formatting assistance; and improvements to notation. After using these tool, the author(s) reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] B. Shneiderman, Human-centered artificial intelligence: Reliable, safe & trustworthy, International Journal of Human–Computer Interaction 36 (2020) 495–504.

[2] C. Gomez, S. M. Cho, S. Ke, C.-M. Huang, M. Unberath, 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 (2025) 1521066.

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

[4] W. Spohn, 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, Springer, 1988, pp. 105–134.

[5] M. J. A. N. C. CondorcetMarquis de Condorcet, Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix, Imprimerie Royale, Paris, 1785.

[6] J. Karge, J.-M. Burkhardt, S. Rudolph, D. Rusovac, 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, 2024, pp. 983–991.

[7] W. Spohn, The laws of belief: Ranking theory and its philosophical applications, Oxford University Press, 2012.

[8] F. Huber, Ranking functions and rankings on languages, Artificial Intelligence 170 (2006) 462–471.

[9] P. Everaere, S. Konieczny, P. Marquis, The strategy-proofness landscape of merging, Journal of Artificial Intelligence Research 28 (2007) 49–105.

[10] P. Everaere, S. Konieczny, P. Marquis, The epistemic view of belief merging: Can we track the truth?, in: Proceedings of the 19th European Conference on Artificial Intelligence, 2010, pp. 621–626.

[11] J. Karge, S. Rudolph, A generalized Condorcet jury theorem for belief fusion, in: Proceedings of the 19th International Conference on Principles of Knowledge Representation and Reasoning, 2022.

[12] J. Karge, Questions about quantities: Epistemic numerical estimate aggregation, in: Proceedings of the 22nd International Conference on Principles of Knowledge Representation and Reasoning, 2025.

[13] I. Caragiannis, A. D. Procaccia, N. Shah, When do noisy votes reveal the truth?, ACM Transactions on Economics and Computation 4 (2016) 1–30.
[14] A. D. Procaccia, N. Shah, J. Tucker-Foltz, Ranked voting on social networks, in: Proceedings of the 24th International Joint Conference on Artificial Intelligence, 2015.
[15] L. Carter, D. Liu, How was my performance? exploring the role of anchoring bias in ai-assisted decision making, International Journal of Information Management 82 (2025) 102875.
[16] C. Gomez, S. M. Cho, S. Ke, C.-M. Huang, M. Unberath, 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 (2025) 1521066.
[17] K. Peng, N. Garg, J. Kleinberg, A no free lunch theorem for human-ai collaboration, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 39, 2025, pp. 14369–14376.
[18] M. Almada, Human intervention in automated decision making: Toward the construction of contestable systems, in: Proceedings of the 17th International Conference on Artificial Intelligence and Law, ACM, 2019, pp. 1–10.
[19] G. Kern-Isberner, N. Skovgaard-Olsen, W. Spohn, Ranking Theory, 2019. doi:10.7551/mitpress/11252.003.0034.
[20] D. Lehmann, M. Magidor, What does a conditional knowledge base entail?, Artif. Intell. 55 (1992) 1–60. URL: https://doi.org/10.1016/0004-3702(92)90041-U. doi:10.1016/0004-3702(92)90041-U.
[21] G. Kern-Isberner, A thorough axiomatization of a principle of conditional preservation in belief revision, Ann. Math. Artif. Intell. 40 (2004) 127–164. URL: https://doi.org/10.1023/A:1026110129951. doi:10.1023/A:1026110129951.
[22] C. E. Alchourrón, P. Gärdenfors, D. Makinson, On the logic of theory change: Partial meet contraction and revision functions, J. Symb. Log. 50 (1985) 510–530. URL: https://doi.org/10.2307/2274239. doi:10.2307/2274239.
[23] A. Darwiche, J. Pearl, On the logic of iterated belief revision, Artif. Intell. 89 (1997) 1–29. URL: https://doi.org/10.1016/S0004-3702(96)00038-0. doi:10.1016/S0004-3702(96)00038-0.
[24] D. Lehmann, Another perspective on default reasoning, Ann. Math. Artif. Intell. 15 (1995) 61–82. doi:10.1007/BF01535841.
[25] J. P. Haldimann, C. Beierle, G. Kern-Isberner, Epistemic state mappings among ranking functions and total preorders, Journal of Applied Logics 10 (2023) 155–192. URL: https://www.collegepublications.co.uk/ifcolog/?00058.
[26] J. P. Haldimann, C. Beierle, Finest syntax splittings of ranking functions and total preorders on worlds, in: P. Marquis, T. C. Son, G. Kern-Isberner (Eds.), Proceedings of the 20th International Conference on Principles of Knowledge Representation and Reasoning, KR 2023, Rhodes, Greece, September 2-8, 2023, 2023, pp. 747–751. URL: https://doi.org/10.24963/kr.2023/75. doi:10.24963/KR.2023/75.
[27] F. Huber, Belief revision ii: Ranking theory, Philosophy Compass 8 (2013) 613–621. doi:10.1111/phc3.12047.
[28] Isabel Healthcare, Differential diagnosis tool, https://www.isabelhealthcare.com/, n.d. Accessed: 2026-03-30.
[29] E. J. Henderson, G. P. Rubin, The utility of an online diagnostic decision support system (isabel) in general practice: a process evaluation, JRSM short reports 4 (2013) 1–11.
[30] Isabel Healthcare, Isabel testimonials – what clinicians think, https://www.isabelhealthcare.com/customer-satisfaction/testimonials, n.d. Accessed: 2026-03-30.
[31] A. Grove, Two modellings for theory change, Journal of Philosophical Logic 17 (1988) 157–170. URL: http://www.jstor.org/stable/30227207.
[32] H. Katzuno, A. Mendelson, Propositional knowledge base revision and minimal change, Technical Report, Technical Report n°3, Department of Computer Science, University of Toronto, 1991.
[33] S. O. Hansson, E. Fermé, J. Cantwell, M. Falappa, Credibility limited revision, Journal of Symbolic Logic 66 (2001) 1581–1596.
[34] D. Grimaldi, M. V. Martinez, R. O. Rodriguez, Moderated revision, International Journal of Approximate Reasoning 166 (2024) 109–126. URL: https://www.sciencedirect.com/science/article/

pii/S0888613X24000136. doi:https://doi.org/10.1016/j.ijar.2024.109126.
[35] G. Casini, T. Meyer, I. Varzinczak, Simple conditionals with constrained right weakening, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, International Joint Conferences on Artificial Intelligence Organization, 2019, pp. 1632–1638. URL: https://doi.org/10.24963/ijcai.2019/226. doi:10.24963/ijcai.2019/226.