RESEARCH PROPOSAL:
HUMAN-AI RANKING AGGREGATION

Jonas Karge

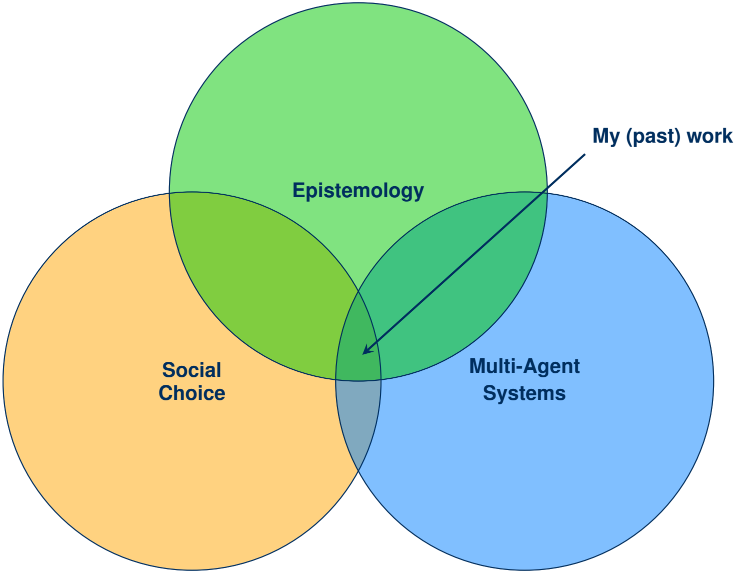
Computational Logic Group

School of Embedded Composite Artificial Intelligence (SECAI)
TU Dresden

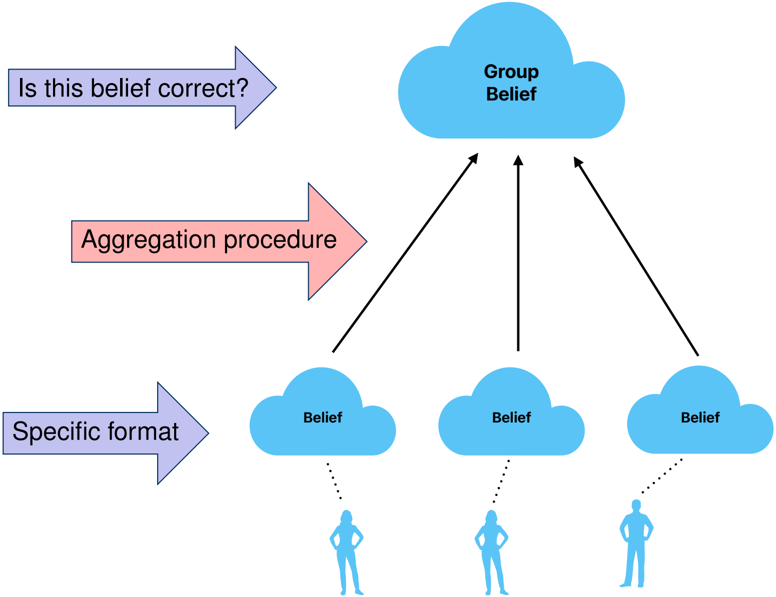
Cape Town, CAIR, March 25th, 2026

Academic Background

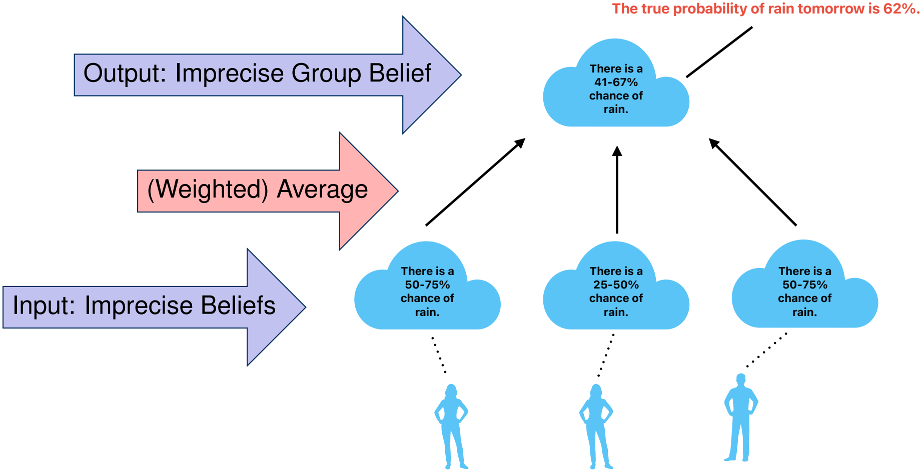
- B.A. Philosophy / French, University of Tübingen
- M.A. Logic, Leipzig University
- Graduate Exchange Program, Ohio University
- PhD Computer Science (Supervisor: Sebastian Rudolph), TU Dresden


Epistemology
My (past) work
Social Choice
Multi-Agent Systems

High Level View



Example: Imprecise Opinion Pooling



Output: Imprecise Group Belief

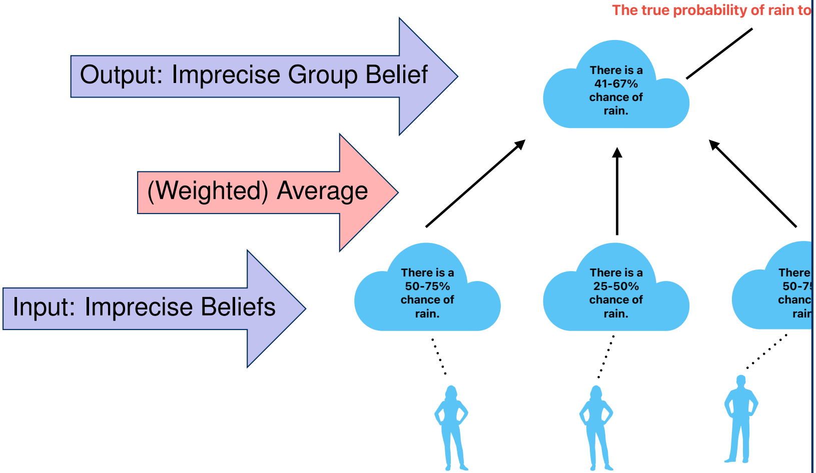
(Weighted) Average

Input: Imprecise Beliefs

The true probability of rain tomorrow is 62%.

There is a 41-67% chance of rain.
There is a 50-75% chance of rain.
There is a 25-50% chance of rain.
There is a 50-75% chance of rain.

Example: Imprecise Opinion Pooling



Output: Imprecise Group Belief

(Weighted) Average

Input: Imprecise Beliefs

The true probability of rain to

There is a 41-67% chance of rain.

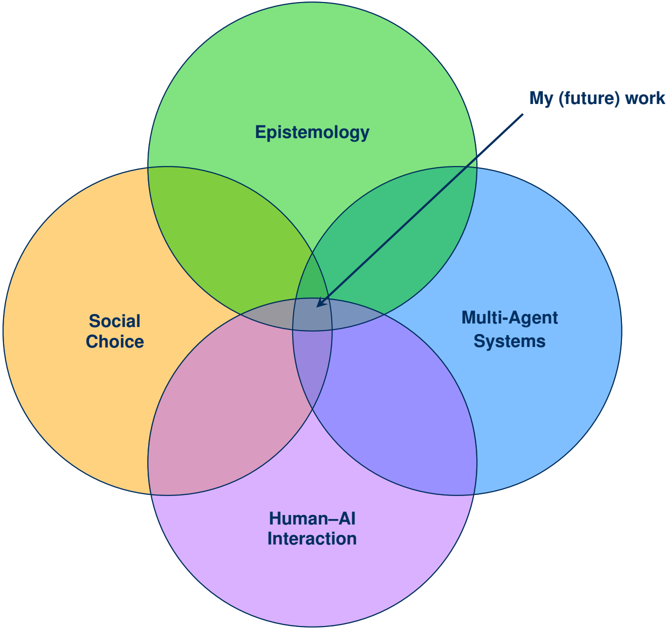
There is a 50-75% chance of rain.

There is a 25-50% chance of rain.

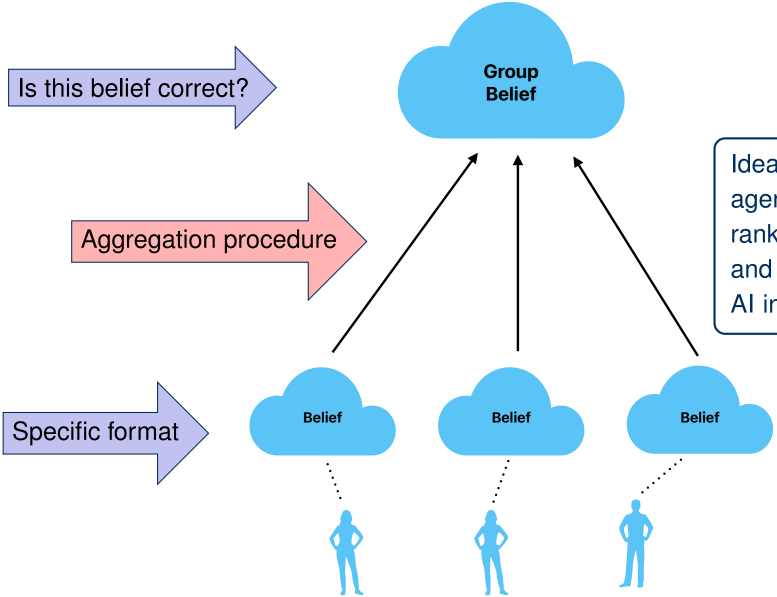
There 50-7 chanc rair

Questions:

- What is a suitable aggregation procedure?
- What is the probability for the group belief to contain the correct value?
- Under what conditions does this probability converge?


My (future) work
Epistemology
Social Choice
Multi-Agent Systems
Human–AI Interaction

High Level View



Is this belief correct?

Aggregation procedure

Specific format

Idea: Represent agent beliefs through ranking functions and consider human-AI interaction panels.

Section 1: Ranking Functions

Introduction: Ranking Functions[1]

Definition (Ranking function): Let Ω be a finite set of possible worlds. A **ranking function** is a map $\kappa : \Omega \to \mathbb{N}_0 \cup \{\infty\}$ such that

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

Intuitively, $\kappa(\omega)$ represents the degree of **disbelief** associated with world ω.

- $\kappa(\omega) = 0$: The world ω is considered maximally plausible.
- $\kappa(\omega) > 0$: The world ω is disbelieved to some degree.
- $\kappa(\omega) = \infty$: The world ω is impossible.

[1] Notation and terminology are adapted from Huber, Franz. ”Ranking Functions and Rankings on Languages.” **Artificial Intelligence** (2006).

Example: Ranking Functions

Example (Medical Diagnosis): Consider a simplified diagnostic scenario with three possible worlds representing diseases: $\Omega = \{\omega_{flu}, \omega_{cold}, \omega_{pneumonia}\}$. An agent might assign the following ranks based on patient symptoms:

$$\kappa_{agent}(\omega) = \begin{cases} 0 & \text{if } \omega = \omega_{flu} \\ 1 & \text{if } \omega = \omega_{cold} \\ 5 & \text{if } \omega = \omega_{pneumonia} \end{cases}$$

Here, the agent considers the Flu to be the most plausible diagnosis (rank 0), a Cold slightly surprising (rank 1), and Pneumonia highly unlikely (rank 5).

Ranking Functions for Multiple Agents

⇒ Next, we want to aggregate the beliefs of multiple agents into a collective belief state.

Object of belief: belief states, i.e. ranking functions

$$\kappa_i : \Omega \to \mathbb{N}_0 \cup \{\infty\},$$

where $\kappa_i(\omega)$ is agent i's degree of disbelief for outcome ω.

Ranking Function Aggregation. Each agent i reports a belief state κ_i.

The **input profile** is

$$E = (\kappa_1, \ldots, \kappa_n),$$

and we aggregate these belief states via an **aggregation (pooling) operator** Δ:

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

In practice, this means that we pool the agents' scores **world-by-world** to obtain a collective belief state.

Base-case pooling functions: sum vs. max[1]

Definition (Sum and Max Pooling):

- **Sum pooling:**

$$\kappa_\Sigma(\omega) = \mathrm{norm}\Big(\sum_{i=1}^{n} \kappa_i(\omega)\Big).$$

  A world is pushed down if **many** agents find it implausible.

- **Max pooling:**

$$\kappa_{\max}(\omega) = \mathrm{norm}\Big(\max_{i=1,\dots,n} \kappa_i(\omega)\Big).$$

  A world is pushed down as soon as **one** agent finds it very implausible.

For any pooled world-score function $f : \Omega \to \mathbb{N}_0 \cup \{\infty\}$ with $\min_{\omega\in\Omega} f(\omega) < \infty$, define

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

Then $\min_{\omega\in\Omega} \mathrm{norm}(f)(\omega) = 0$, so $\mathrm{norm}(f)$ is a valid ranking function.

[1] Inspired by: Everaere, Patricia, Sébastien Konieczny, and Pierre Marquis. ”The strategy-proofness landscape of merging.” Journal of Artificial Intelligence Research (2007).

Example (Three agents, same diagnosis space): Let $\Omega = \{\omega_{flu}, \omega_{cold}, \omega_{pneumonia}\}$. Three agents report the following belief states (ranking functions):

World ω	$\kappa_1(\omega)$	$\kappa_2(\omega)$	$\kappa_3(\omega)$
ω_{flu}	0	3	3
ω_{cold}	4	0	0
$\omega_{pneumonia}$	6	5	5

Sum pooling.

$$\begin{aligned} \kappa_\Sigma(\omega_{flu}) &= 6 \\ \kappa_\Sigma(\omega_{cold}) &= 4 \\ \kappa_\Sigma(\omega_{pneumonia}) &= 16 \end{aligned}$$

$$\mathrm{norm}(\kappa_\Sigma) = (2, 0, 12)$$

so the group's top diagnosis is ω_{cold}.

Max pooling.

$$\begin{aligned} \kappa_{\max}(\omega_{flu}) &= 3 \\ \kappa_{\max}(\omega_{cold}) &= 4 \\ \kappa_{\max}(\omega_{pneumonia}) &= 6 \end{aligned}$$

$$\mathrm{norm}(\kappa_{\max}) = (0, 1, 3)$$

so the group's top diagnosis is ω_{flu}.

Setion 2: Human-AI Interaction

What Does ”AI” Mean in Human-AI Interaction?

In the HAI literature, ”AI” typically denotes **task-specific algorithmic systems** acting as decision aids or collaborative partners:

Common Conceptualizations of AI:

- **Statistical Predictors:** Traditional machine learning models (e.g., Bayesian networks, random forests) that output probabilities, or risk scores.
- **Generative AI & LLMs:** Acting as ”reasoning engines” or synthesizers that process natural language (e.g., parsing patient history, suggesting treatments).
- **Decision Support Systems (DSS):** System that aids management in decision-making by gathering, analyzing, and visualizing large amounts of data.

AI in Practice: Diagnostic Decision Support Systems

AI systems that generate ranked lists of diagnoses are already actively used in clinical environments known as **Diagnostic Clinical Decision Support Systems (CDSS)**.

Real-World Example: Isabel Pro

- **What it is:** A widely deployed diagnostic CDSS used by physicians and healthcare institutions globally to reduce diagnostic errors.
- **How it works:** It extracts clinical features (symptoms, age, gender, medical history) directly from a patient's Electronic Health Record using natural language processing.
- **The Output:** It calculates the probabilities of various diseases and outputs a ranked diagnosis list.

Insights from Isabel Pro

Feedback from physicians using Isabel Pro illustrates how different **Human–AI interaction patterns** naturally emerge in clinical practice.[1]

”[I use] it as a ”second check”, just to make certain I haven’t forgotten something.” – Dr. G. Dhaliwal, UCSF.
Pattern: The human forms an initial ranking first, then requests the AI’s ranking to catch omissions.

”I open up the interface while I am talking to the patient... enter the data, and then discuss the diagnostic suggestions...” – Dr. L. Sprecher, Mayo Health System.
Pattern: The AI’s ranking is generated concurrently during data collection, actively shaping the doctor’s real-time hypothesis generation.

”Isabel ensures that I consider the broadest of differential diagnoses and prevents diagnostic tunnel vision.” – Dr. A. Winrow, Kingston Hospital.
Pattern: The AI introduces highly plausible but rare diagnoses.

[1] https://www.isabelhealthcare.com/customer-satisfaction/testimonials

Human–AI Interaction Patterns

To understand how human and artificial agents can jointly arrive at a decision, we need to structure the **interaction patterns** that govern their collaboration.[1]

1. **AI-First Assistance (Concurrent)**
   - The decision-making problem and the AI's predicted outcome are displayed simultaneously. The user can either accept or override the AI's advice.
2. **AI-Follow Assistance (Sequential)**
   - The user forms an independent preliminary prediction first. Only then is the AI's recommendation presented for comparison and potential reassessment.
3. **Request-Driven Assistance**
   - The user actively controls **when** they want to receive AI assistance (e.g., clicking a button to ask for help), fostering a stronger sense of human agency.

[1] Gomez, Catalina, et al. ”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)

Human–AI Interaction Patterns (cont.)

4. Secondary Assistance

- The AI offers supplementary information (e.g., risk profiles) rather than a direct solution. The human must interpret this data to solve the primary task.

5. AI-Guided Dialogic Engagement

- The AI facilitates a conversational, turn-taking exchange. The AI requests specific constraints or attributes, and the user provides them until a candidate solution is found.

6. User-Guided Interactive Adjustments

- The user manipulates inputs or assumptions to observe how the AI's outcome changes, reversing the information flow.

Risks and Challenges in Human–AI Collaboration

While mixed Human–AI panels offer great potential, their **asymmetric roles** and **structured information flows** introduce significant epistemic risks:

- **Anchoring Bias:** Concurrent AI assistance can cause human agents to heavily anchor to the AI's prediction.
- **Epistemic Performance Gap:** The 'No Free Lunch' theorem for Human-AI collaboration shows that without careful design, hybrid teams do not just fail to achieve synergy. They can actually perform strictly worse than the least accurate individual agent.[a]
- **Information Flow:** Even when an AI system provides a clean, easily interpretable output, such as a simple ranked list of diagnoses, the actual **collaborative integration** of that information remains highly unstructured.
- **Cross-Format Aggregation:** Inputs rarely arrive in a unified format.

[a] Peng, Kenny, Nikhil Garg, and Jon Kleinberg. "A no free lunch theorem for human-ai collaboration." Proceedings of the AAAI Conference on Artificial Intelligence. 2025.

Formalizing Basic Human–AI Panels

As a first step, we aim to bridge formal aggregation with HAI patterns by focusing on a restricted, well-defined setting:

Scope and Modeling Assumptions:

- Isolate the simplest, static interaction paradigms (AI-First and AI-Follow).
- Require all agents to report their belief states via ranking functions.
- Explicitly model the **human's** anchoring bias (interaction-induced correlation) when exposed to AI assistance.
- Analyze configurations where multiple human agents interact with a single AI model.

Ultimately, we want to derive conditional statements of the following form:

Target Guarantee:
”If the panel utilizes interaction pattern X and aggregation operator Δ (subject to constraints Y), the collective decision mathematically guarantees an epistemic benefit over standalone baselines.”

Section 3: Human-AI Ranking Aggregation

Brief Motivation: Why Rankings May Be Safer

Example (Probabilities vs. Rankings): Suppose the true diagnosis is **Flu**.

1. Probabilistic Pooling

$$P_{H1} = (\text{Flu: } 0.60,\ \text{Cold: } 0.40)$$
$$P_{H2} = (\text{Flu: } 0.55,\ \text{Cold: } 0.45)$$
$$P_{AI} = (\text{Flu: } 0.01,\ \text{Cold: } 0.99)$$

Simple average $\bar{P}$:

$$\bar{P}(\text{Flu}) = 0.386, \quad \bar{P}(\text{Cold}) = 0.613$$

Result: Fails

2. Ranking Pooling

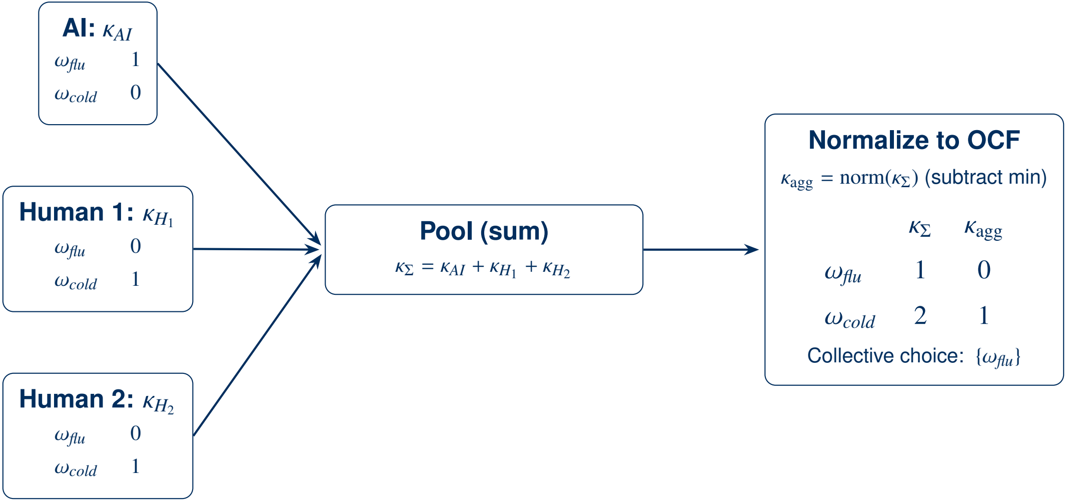
$$\kappa_{H1} = (\text{Flu: } 0,\ \text{Cold: } 1)$$
$$\kappa_{H2} = (\text{Flu: } 0,\ \text{Cold: } 1)$$
$$\kappa_{AI} = (\text{Flu: } 1,\ \text{Cold: } 0)$$

Summation & Normalization:

$$\text{norm}(\Sigma\kappa) = (\text{Flu: } 0,\ \text{Cold: } 1)$$

Result: Succeeds

Toy Example: Safe Rankings



AI: κ_{AI}

ω_{flu}	1
ω_{cold}	0

Human 1: κ_{H_1}

ω_{flu}	0
ω_{cold}	1

Human 2: κ_{H_2}

ω_{flu}	0
ω_{cold}	1

Pool (sum)

$\kappa_{\Sigma} = \kappa_{AI} + \kappa_{H_1} + \kappa_{H_2}$

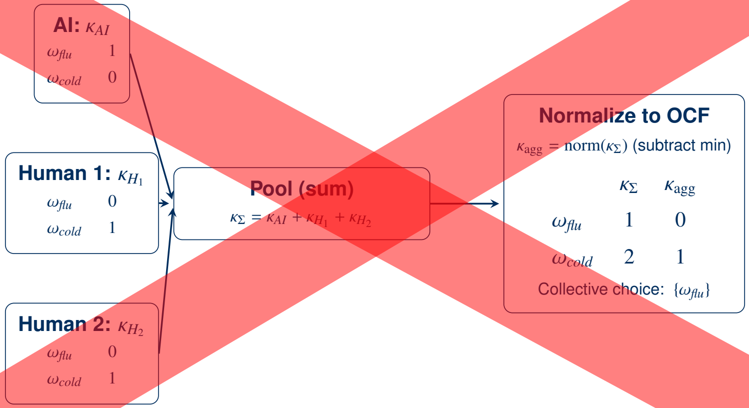
Normalize to OCF

$\kappa_{\text{agg}} = \text{norm}(\kappa_{\Sigma})$ (subtract min)

	κ_{Σ}	κ_{agg}
ω_{flu}	1	0
ω_{cold}	2	1

Collective choice: $\{\omega_{flu}\}$

Toy Example: Safe Rankings



AI: κ_{AI}

ω_{flu}	1
ω_{cold}	0

Human 1: κ_{H_1}

ω_{flu}	0
ω_{cold}	1

Human 2: κ_{H_2}

ω_{flu}	0
ω_{cold}	1

Pool (sum)

$\kappa_\Sigma = \kappa_{AI} + \kappa_{H_1} + \kappa_{H_2}$

Normalize to OCF

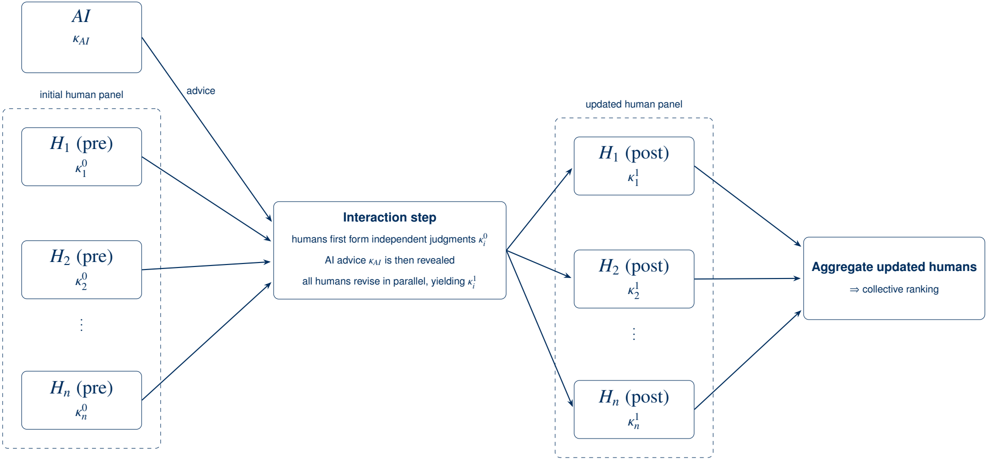
$\kappa_{\text{agg}} = \text{norm}(\kappa_\Sigma)$ (subtract min)

	κ_Σ	κ_{agg}
ω_{flu}	1	0
ω_{cold}	2	1

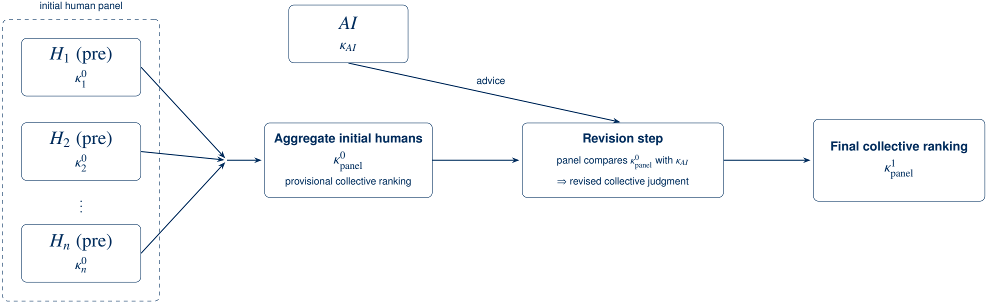
Collective choice: $\{\omega_{flu}\}$

AI-Follow Assistance

AI-follow assistance with concurrent panel revision



AI-follow assistance at the panel level



initial human panel

H_1 (pre) κ_1^0

H_2 (pre) κ_2^0

$\vdots$

H_n (pre) κ_n^0

AI κ_{AI}

advice

Aggregate initial humans

κ^0_{panel}

provisional collective ranking

Revision step

panel compares κ^0_{panel} with κ_{AI}

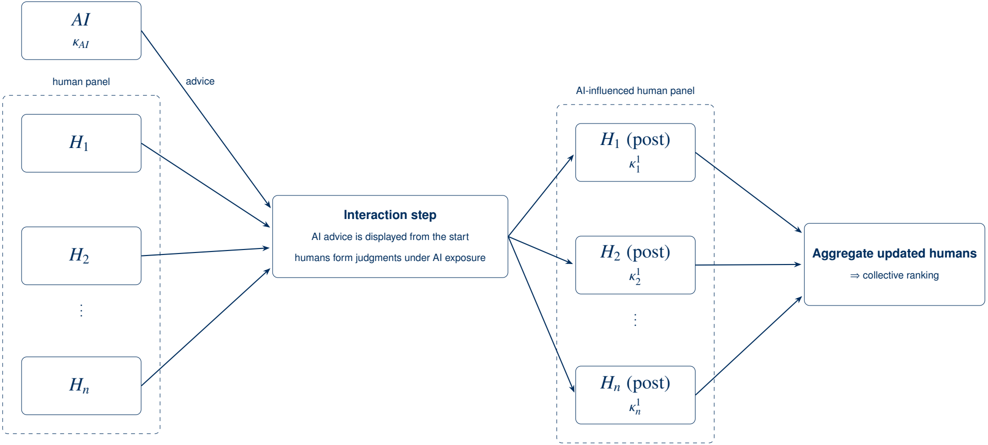
⇒ revised collective judgment

Final collective ranking

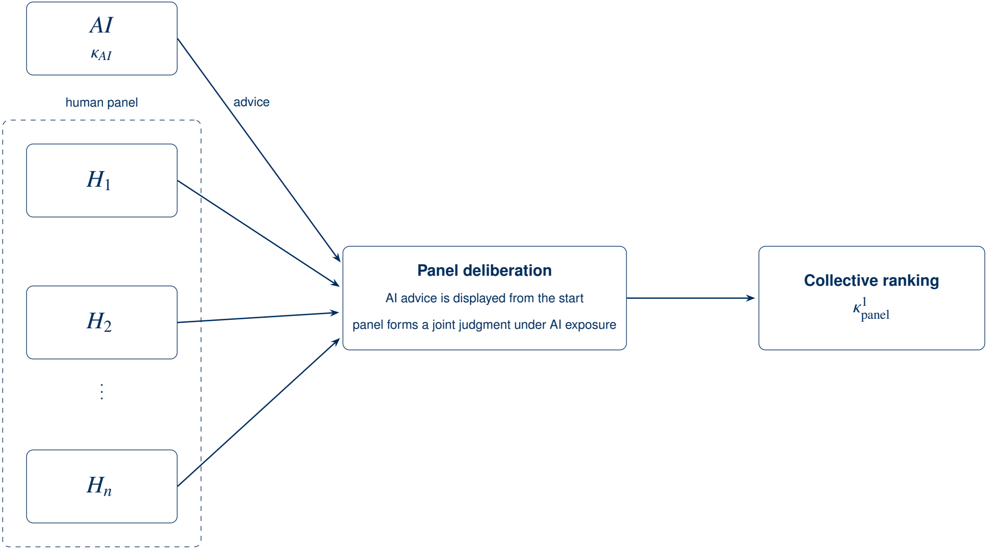
κ^1_{panel}

AI-First Assistance

AI-first assistance with concurrent panel formation

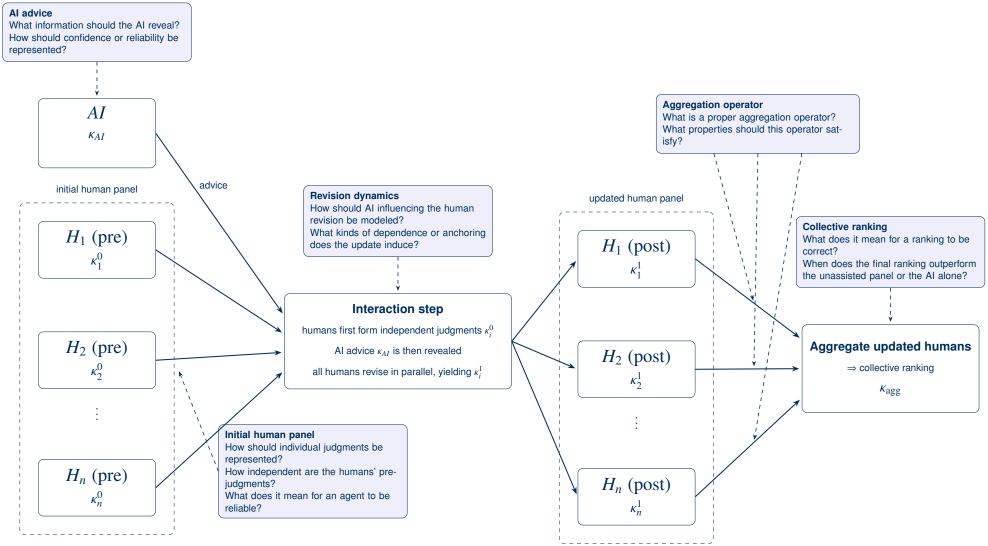

AI
κ_{AI}
advice
human panel
H_1
H_2
$\vdots$
H_n
Interaction step
AI advice is displayed from the start
humans form judgments under AI exposure
AI-influenced human panel
H_1 (post)
κ_1^1
H_2 (post)
κ_2^1
$\vdots$
H_n (post)
κ_n^1
Aggregate updated humans
$\Rightarrow$ collective ranking

AI-first assistance with panel-level deliberation


AI
κ_{AI}
human panel
advice
H_1
H_2
⋮
H_n
Panel deliberation
AI advice is displayed from the start
panel forms a joint judgment under AI exposure
Collective ranking
κ^1_{panel}

Questions to ask: AI-follow assistance with concurrent panel revision

AI-follow assistance with concurrent panel revision



AI advice
What information should the AI reveal?
How should confidence or reliability be represented?

AI
κ_{AI}

advice

initial human panel

H_1 (pre) κ_1^0

H_2 (pre) κ_2^0

$\vdots$

H_n (pre) κ_n^0

Revision dynamics
How should AI influencing the human revision be modeled?
What kinds of dependence or anchoring does the update induce?

Interaction step

humans first form independent judgments κ_i^0

AI advice κ_{AI} is then revealed

all humans revise in parallel, yielding κ_i^1

Initial human panel
How should individual judgments be represented?
How independent are the humans' pre-judgments?
What does it mean for an agent to be reliable?

updated human panel

H_1 (post) κ_1^1

H_2 (post) κ_2^1

$\vdots$

H_n (post) κ_n^1

Aggregation operator
What is a proper aggregation operator?
What properties should this operator satisfy?

Collective ranking
What does it mean for a ranking to be correct?
When does the final ranking outperform the unassisted panel or the AI alone?

Aggregate updated humans

⇒ collective ranking

κ_{agg}

Next Steps

If you would like to collaborate on this, you can

- think of topics from this proposal that you found interesting;
- drop by my office;
- send an email to `jonas.karge@tu-dresden.de`.

If we are feeling ambitious, we could try to draft a short paper for the Joint Workshop on Statistics and Knowledge Integration for Logic, Learning, Ethical Decisions, and LLMs (SKILLED-LLMs 2026) (co-located with FLOC).