How Different Explanation Conceptualisations Influence Trust in AI-Assisted Credibility Assessments

Saumya Pareek | Niels van Berkel | Eduardo Velloso | Jorge Goncalves

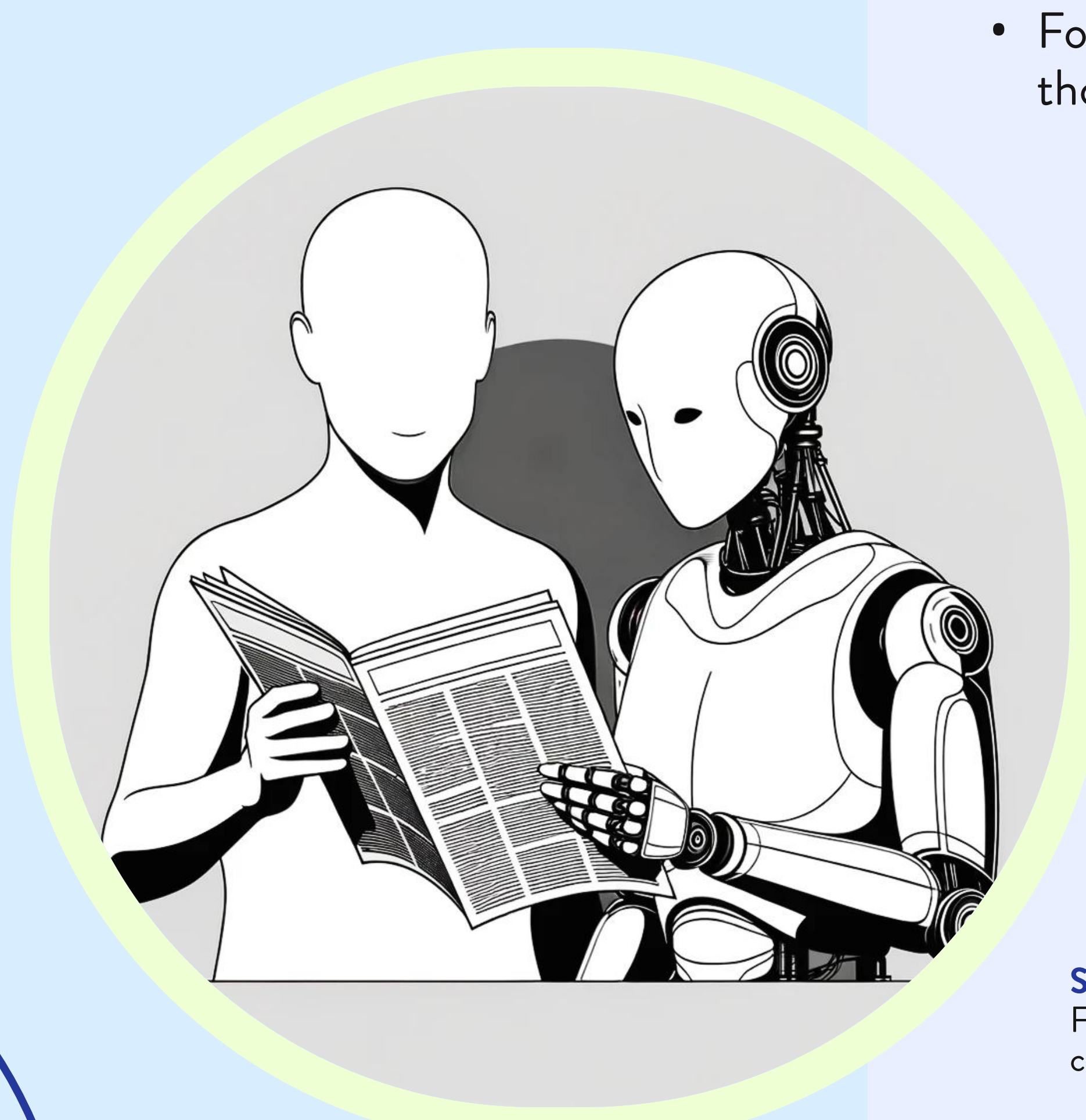
The University of Melbourne · The University of Sydney · Aalborg Universitet

1. Research Gap

- More and more approaches to detect misinformation are being automated to deal with scale. **But how can we make end-users trust these automated credibility decisions?**
- Explanations help foster trust in AI systems. But past research on AI-based credibility systems either offered no explanations, or explanations that are overly technical and model-centric. These approaches do not assist users in forming mental models of the AI's decision-making, an aspect crucial to collaborative-decision making.
- To empower individuals to trust AI-based credibility indicators, it is thus **imperative to design explanations that possess a strong undertone of human reasoning and convey a model's decision in terms of how humans construct and revise theories.**

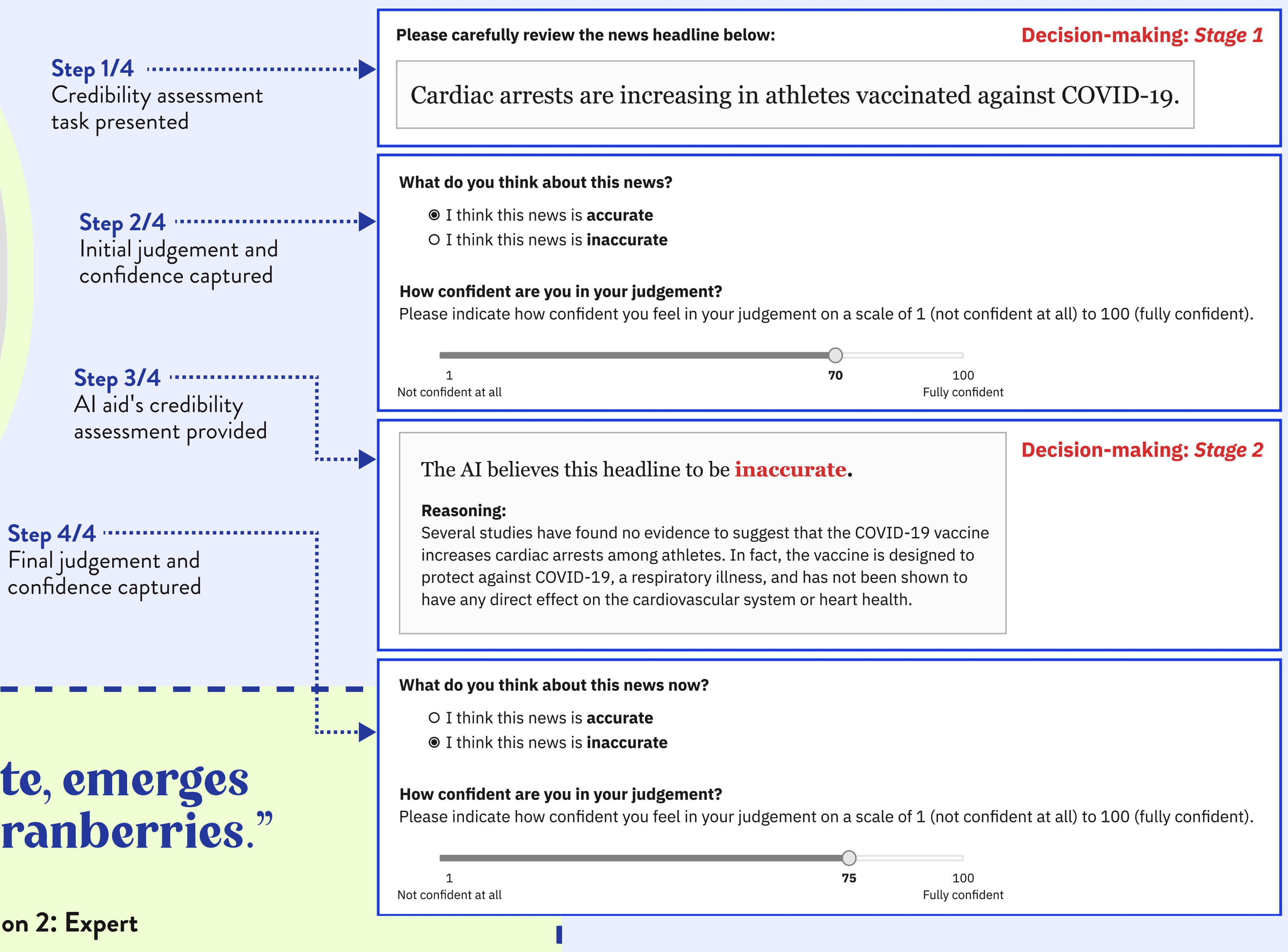
2. Design Opportunity

- Jaccard & Jacoby outline four approaches, called **Conceptualisation Validations**, that humans use when assessing the worth of a new concept/information:
  - **Consensual/peer:** what level of acceptance or consensus does the claim receive from the masses?
  - **Expert:** do experts with relevant knowledge endorse the claim?
  - **Internal/logical:** is the claim free from logical inconsistencies?
  - **Empirical:** what empirical evidence exists to support the claim?
- **Overarching question:** **How can AI explanations designed using different conceptualisation validations (CVs) shape users' decision-making and reliance on AI during collaborative credibility assessment?**



3. Methodology

- We conducted a survey-based study with 320 participants where **participants collaboratively assessed news credibility with a simulated AI.**
- **We showed participants both factual and fake news headlines**, each joined by an AI-based credibility indicator and an explanation whose presence and type varied between treatments.
- Our experimental design manipulated the following: ***AI judgement*** (i.e. agreeing or disagreeing with the user's assessment), ***Scientificness*** of the headline (i.e. scientific or non-scientific in nature), ***Political Congruence*** of the headline with participant's beliefs (i.e. congruent, incongruent, or non-political), ***Explanation Conceptualisation Validation (CV)*** (i.e. Control [no explanation], Consensual, Expert, Internal, or Empirical).
- For each headline, we measured participants' **credibility judgements** and their **confidence** in those judgements twice — once before and once after displaying the indicator.



Step 1/4 Credibility assessment task presented

Please carefully review the news headline below: **Decision-making: *Stage 1***

Cardiac arrests are increasing in athletes vaccinated against COVID-19.

Step 2/4 Initial judgement and confidence captured

What do you think about this news?

● I think this news is **accurate**
○ I think this news is **inaccurate**

How confident are you in your judgement?
Please indicate how confident you feel in your judgement on a scale of 1 (not confident at all) to 100 (fully confident).

1 Not confident at all — **70** — 100 Fully confident

Step 3/4 AI aid's credibility assessment provided

Decision-making: *Stage 2*

The AI believes this headline to be **inaccurate.**

Reasoning:
Several studies have found no evidence to suggest that the COVID-19 vaccine increases cardiac arrests among athletes. In fact, the vaccine is designed to protect against COVID-19, a respiratory illness, and has not been shown to have any direct effect on the cardiovascular system or heart health.

Step 4/4 Final judgement and confidence captured

What do you think about this news now?

○ I think this news is **accurate**
● I think this news is **inaccurate**

How confident are you in your judgement?
Please indicate how confident you feel in your judgement on a scale of 1 (not confident at all) to 100 (fully confident).

1 Not confident at all — **75** — 100 Fully confident

"Wisconsin, the dairy state, emerges as the top producer of cranberries."

Explanation 1: Consensual/peer

Of the individuals taking this survey with you, 71% have rated it as accurate while 29% have rated it as inaccurate.

Explanation 2: Expert

Several experts in the agricultural industry have long acknowledged Wisconsin's dual status as both a dairy and cranberry powerhouse, ranking at the top.

Explanation 3: Internal/logical

Cranberries flourish in Wisconsin, given its unique geography and climate. Cranberries thrive in wet, acidic soil, and Wisconsin's northern regions are dotted with thousands of shallow, marshy bogs that provide the perfect growing conditions.

Explanation 4: Empirical

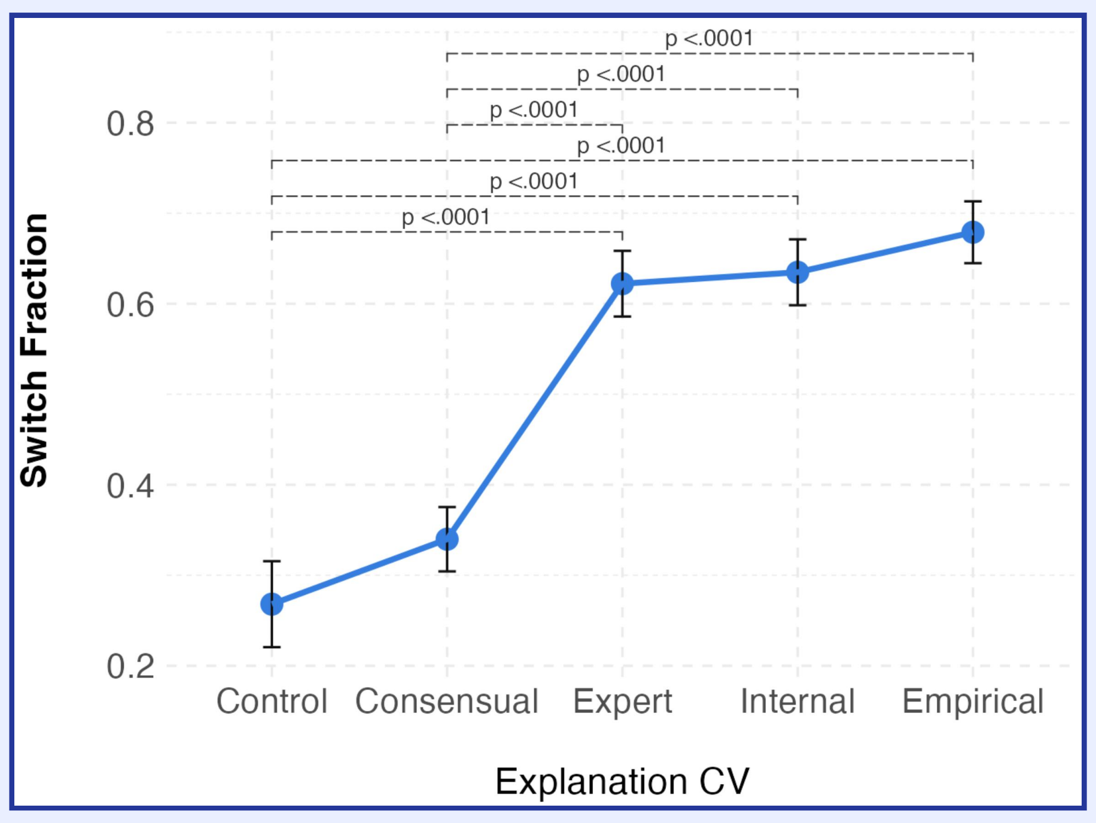
Wisconsin produced a record-breaking 5.38 million barrels of cranberries in 2020. This represents over 60% of the total US cranberry production for the year, solidifying Wisconsin's position as the nation's top cranberry producer.

4. Measures

Reliance measured using:

$$\text{Switch Fraction} = \frac{\textit{Number of instances where a user changes their judgement}}{\textit{Total number of instances with a disagreement with the AI}}$$

5. Findings



> *"The explanations helped me trust the AI more. It was nice to see how it came up with the answers. It felt **transparent and honest**. [...] It showed me that it was not trying to fool me or hide anything." (P62).*

1. Providing **explanations led to higher reliance compared no explanations.**
2. **We observed substantially different reliance on the same AI judgement based on the explanation accompanying it.** Consensual explanations were the least effective piece of information supplied. In contrast, Expert, Internal, and Empirical explanations were almost twice as effective, despite lacking external sources to corroborate their claims.
3. **Explanations were highly effective irrespective of the AI's correctness** — participants could not detect when they were being guided towards the truth.
4. Headline ***scientificness*** **and** ***political congruence*** **did not influence switching behaviour,** individuals aligned their judgement with the AI for both attitude-affirming and challenging headlines.
5. We observed both **automation bias and aversion**:
   a. Participants with higher trust in AI relied more on its judgements and perceived it as superior.
   b. Others were reluctant to trust the AI irrespective of its accuracy, embracing their initial (in)correct beliefs, mirroring the `boomerang effect' observed in traditional corrections to misinformation.

6. Key Learnings & Takeaways

Explanation Framing

The framing of explanations matters — identical AI judgements explained with different CVs led to different reliance.

Duality of Explanations & Over-reliance

Consider the dual nature of explanations when designing AI-based credibility systems — they can guide users towards the truth and also away from it. Our **participants exhibited unwarranted reliance on the AI irrespective of its accuracy,** because they believed it to be comprehensive and making accurate judgements, as suggested by our qualitative results.

Sources & Accuracy Metrics

Explanations **did not cite any sources, or provide AI accuracy metrics,** yet we observed over-reliance — would providing corroborating sources or AI performance metrics promote more careful scrutiny of explanations?

Critical Information Assessment

Nevertheless, the inclusion of even Consensual **explanations motivated individuals to pause and critically re-examine headlines,** promoting a more deliberative information-assessment habit.