【Case Study】

Honorable Case Study Recognition

AI for human assessment: What do professional assessors need?

Riku Arakawa[†]

Carnegie Mellon University

Hiromu Yakura[†]

University of Tsukuba / AIST

[†] Equal contribution

In collaboration with ACES Inc.

CHI23
Hamburg, Germany | Hybrid
April 23-28, 2023
reCHInnecting

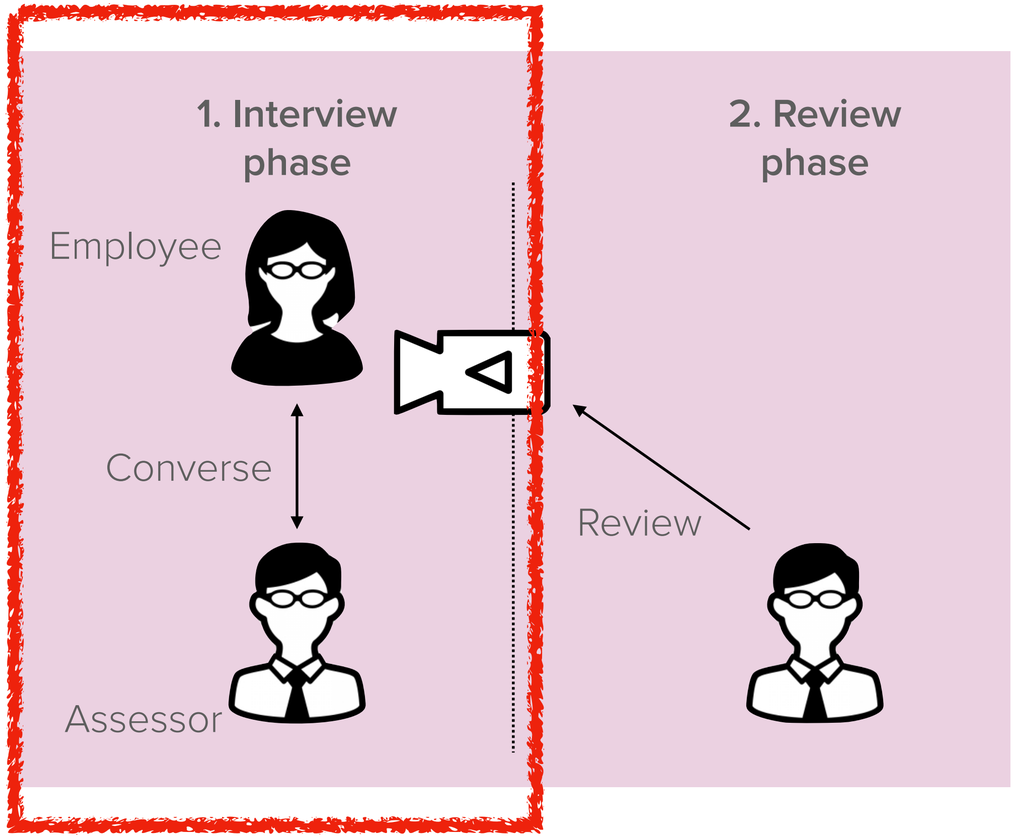
Human assessment

Evaluate candidates regarding their suitability for certain types of employment, mostly through interviews by professional assessors

Human assessment

Human assessment typically consists of two phases:

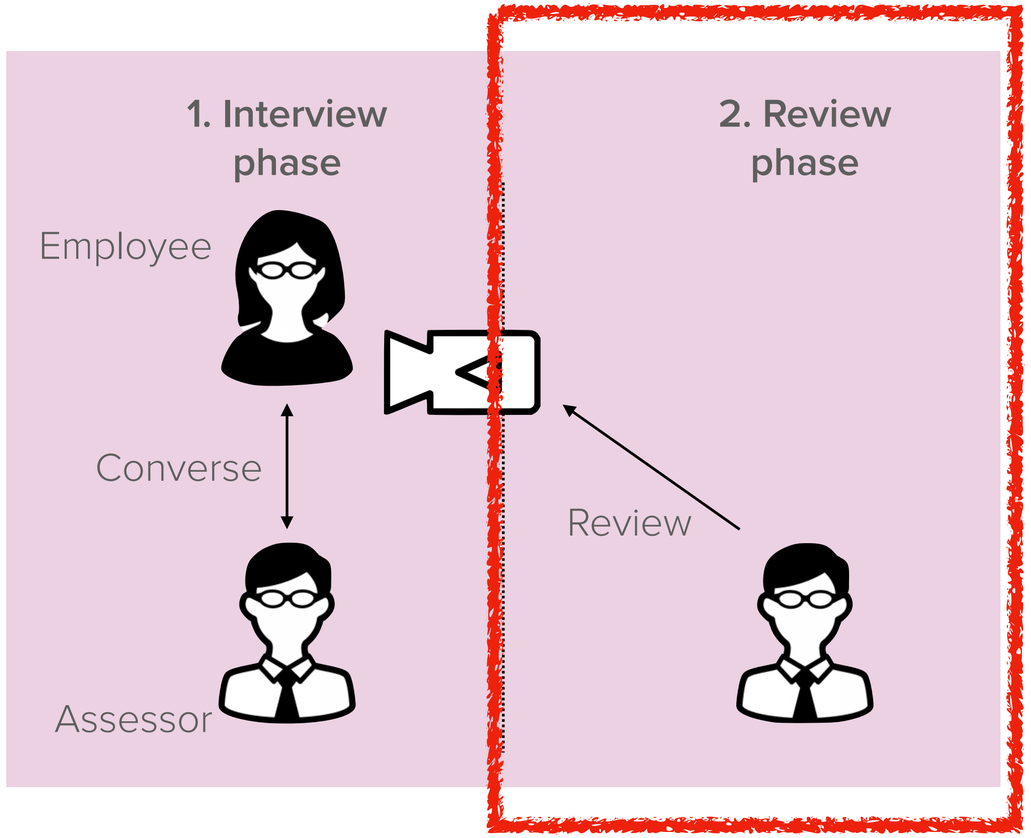
1. **Interview phase**
   - An assessor plays a certain role in an one-on-one interview.
   - The conversation is video-recorded.
2. **Review phase**
   - The assessor playbacks the recorded video.
   - The video is used to find verbal and nonverbal cues for evaluating the employee as a manager.



Human assessment

Human assessment typically consists of two phases:

1. Interview phase
   - An assessor plays a certain role in an one-on-one interview.
   - The conversation is video-recorded.
2. **Review phase**
   - The assessor playbacks the recorded video.
   - The video is used to find verbal and nonverbal cues for evaluating the employee as a manager.



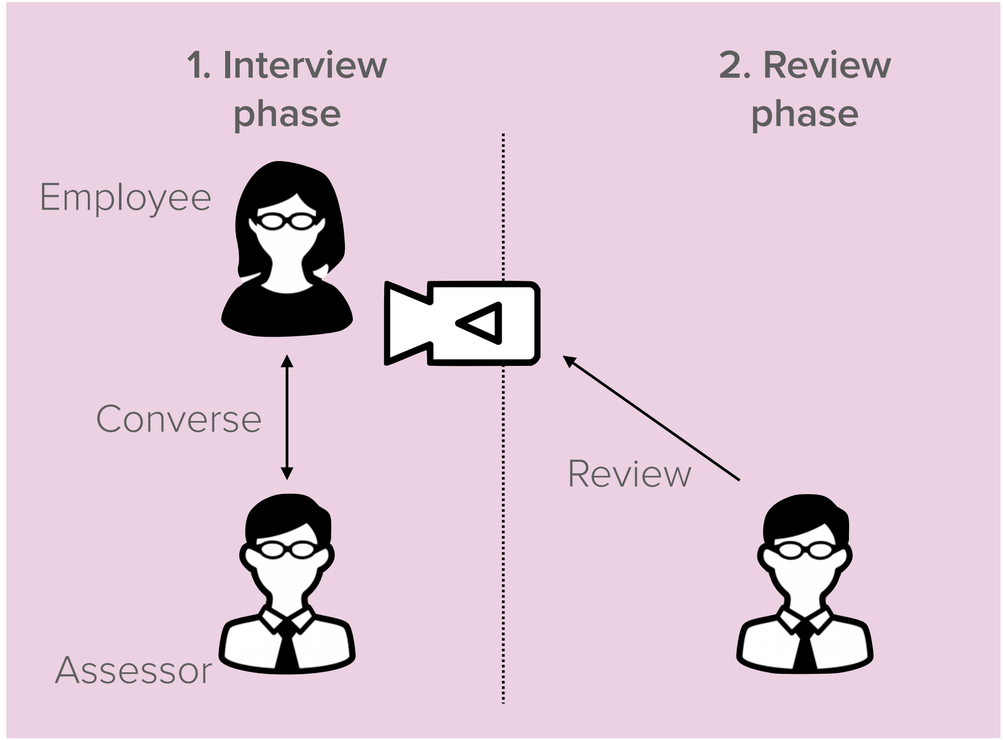
How can we support human assessment with AIs?

[1] Arakawa and Yakura, REsCUE: A framework for REal-time feedback on behavioral CUEs using multimodal anomaly detection, CHI'19

[2] Arakawa and Yakura, INWARD: A Computer-Supported Tool for Video-Reflection Improves Efficiency and Effectiveness in Executive Coaching, CHI'20

Initial workshop

- We conducted a workshop with 2 professional assessors:
- Difficulties
  - The review phase is time-consuming ⏳
  - Assessors’ subjectivity can lead to a wrong decision 🧐

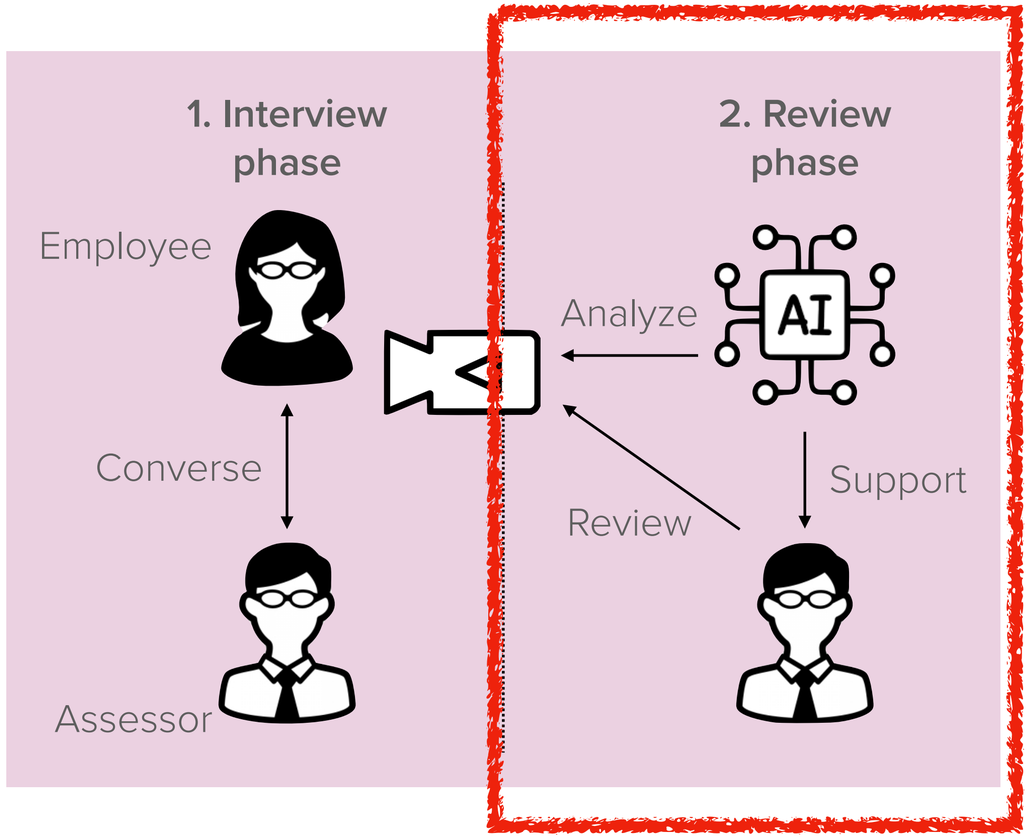


Initial workshop

- We conducted a workshop with 2 professional assessors:
- Difficulties
  - The review phase is time-consuming ⏳
  - Assessors' subjectivity can lead to a wrong decision 🧐

⬇

Q: How can AI systems support professional assessors' decision-making in review-phase?



System requirements

- The assessors were **skeptical about AI-based end-to-end decision making** because human assessment should consider various factors specific to each employee.
  - They are highly human-contextual and difficult to be captured by computers.

System requirements

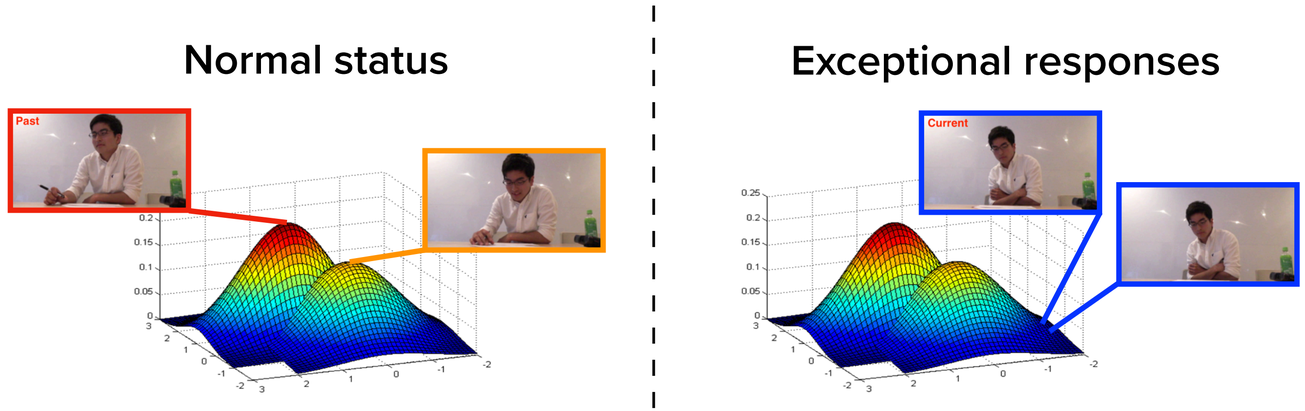
- The assessors were **skeptical about AI-based end-to-end decision making** because human assessment should consider various factors specific to each employee.
  - They are highly human-contextual and difficult to be captured by computers.
- The assessors **expected AI systems to help them not miss important behavior cues due to their subjectivity or mental demands.**
  - Then, the assessors can revise their judgment by taking the contextual meaning of such AI-detected cues into consideration.

▼

Hypothesis: Separating observation (by AI) and judgment (by professionals)

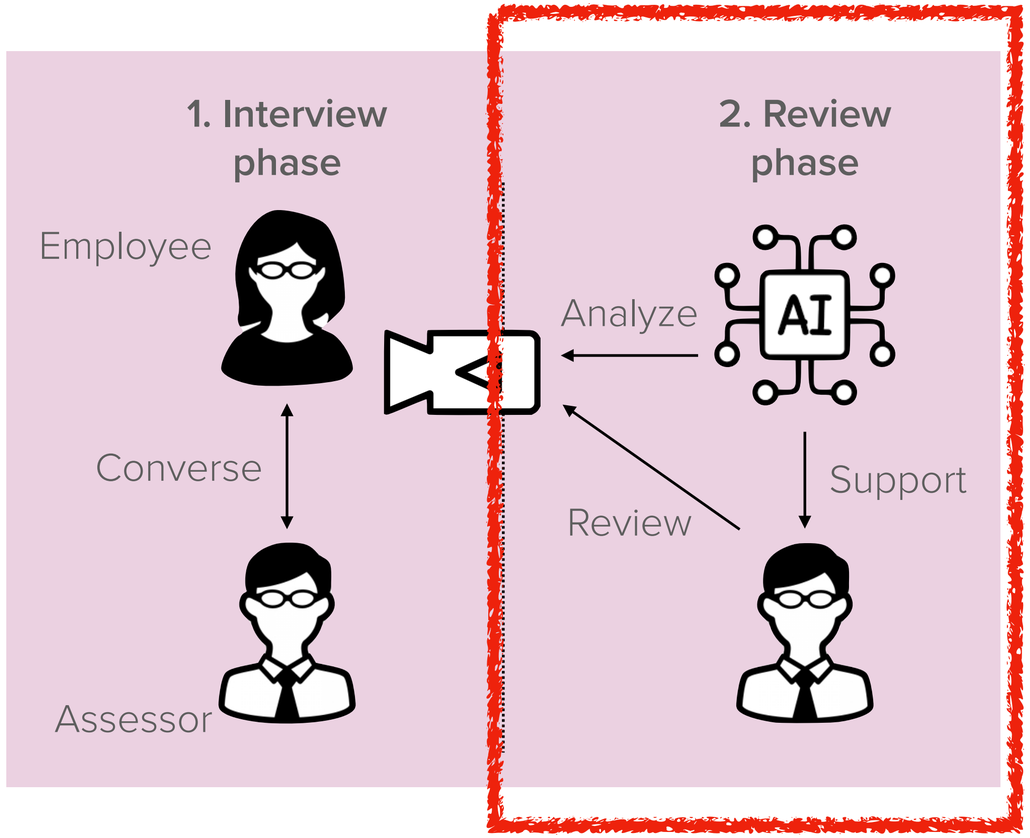
Feasibility study

- We adopted nonverbal behavior analysis algorithm, REsCUE [1] used in executive coaching.
  - It can extract anomalous cues of people in conversation.
  - It provides clear visualization of the cues based on GMM.



Feasibility study

- 20 interview videos
- Two assessors annotated important scenes manually
- Our algorithm also extracts anomaly scenes



Findings

- We examined the agreement between the algorithm and the assessors

  and found that the algorithm does not completely replicate their annotation.

  - The discrepancy was attributed to both **false-positive detection** and assessors' subjectivity.

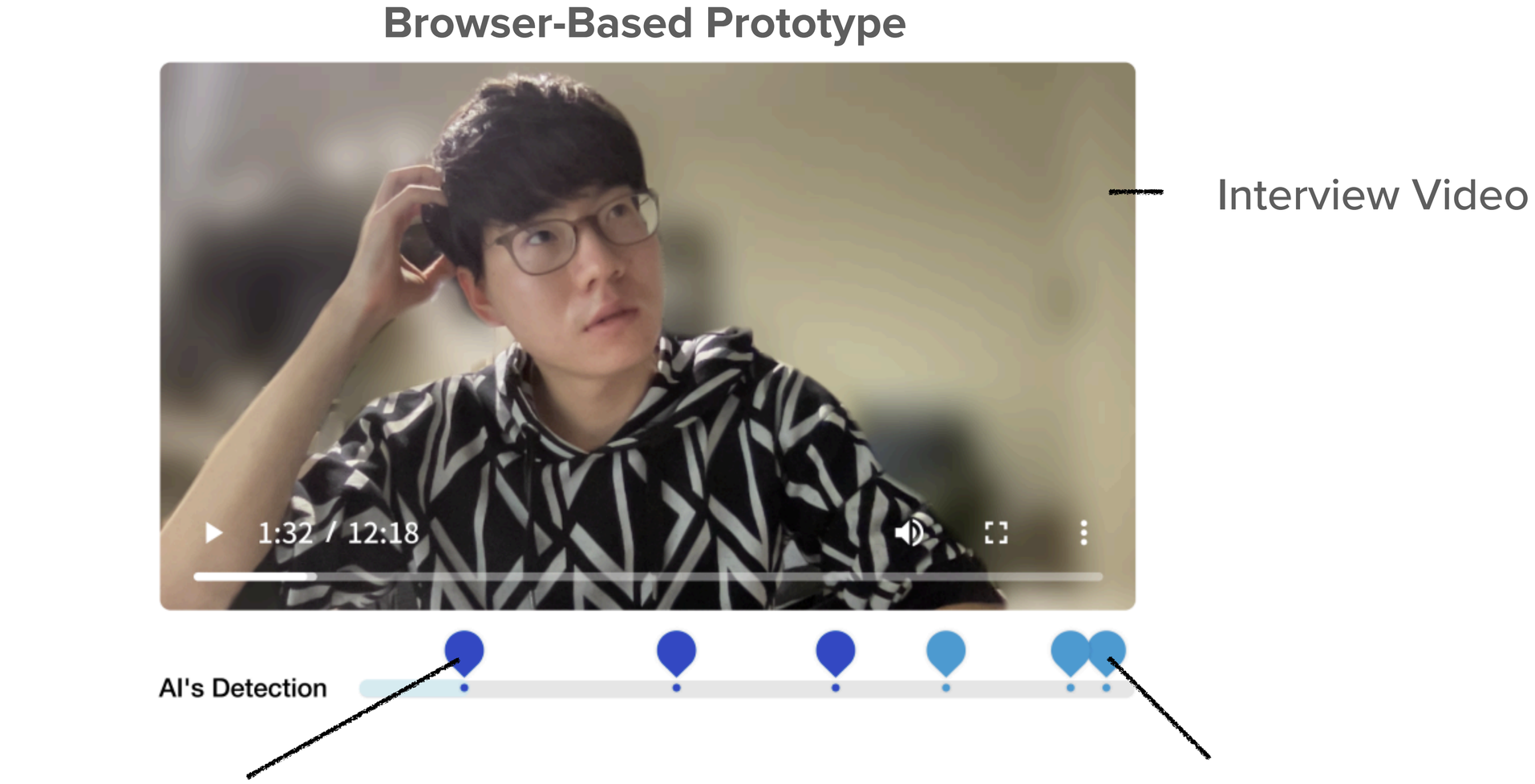
Findings

- We examined the agreement between the algorithm and the assessors and found that the algorithm does not completely replicate their annotation.
  - The discrepancy was attributed to both **false-positive detection** and assessors' subjectivity.
- However, the assessors found that the algorithm would facilitate their assessment.
  - **The interpretable output of the anomaly-detection-based algorithm guided them to infer the reason behind the detection, questioning their decisions.**
  - It helped maintain the assessors' trust in the case of false-positives 🤝

▼

A: The separation contributed to the trust in this highly contextual domain.

Usability study — Prototype



Usability study — Procedure

- 6 professional assessors who had not participated in our first study:
  - 2 junior assessors, 4 senior assessors
- Each assessor reviewed randomly chosen four videos with the prototype.
- We conducted semi-structured interviews after they reviewed all videos to ask about usability of the prototype.

Usability study — Result

- **Deepened quality of assessment**
  - enhanced objectivity (← false-positive)
  - gain confidence (← true-positive)
  - not lose confidence (← false-negative)

  “rethought such cases but could easily resolve the conflict by referring to other signals”

Usability study — Result

- **Deepened quality of assessment**
  - enhanced objectivity (← false-positive)
  - gain confidence (← true-positive)
  - not lose confidence (← false-negative)

  “rethought such cases but could easily resolve the conflict by referring to other signals”

Usability study — Result

- **Deepened quality of assessment**
  - enhanced objectivity (← false-positive)
  - gain confidence (← true-positive)
  - not lose confidence (← false-negative)

  “*rethought such cases but could easily resolve the conflict by referring to other signals*”

Usability study — Result

- **Room for improvement of the prototype**
- **Potential use scenarios**

Please refer to the paper!

Lessons learned

- It is neither recommended nor feasible to train an AI model that replicates assessors' decision-making process.
  - Inevitable inconsistency among their processes

    (= different assessors look at different cues while having the same assessment result)
  - Lack of interpretability and validity in its output.

- Our design of **separating observation and judgment** is a promising approach in such highly contextual domains.
  - Importantly, our goal is not replacing human decision, but helping them.

Lessons learned

- It is neither recommended nor feasible to train an AI model that replicates assessors' decision-making process.
  - Inevitable inconsistency among their processes

    (= different assessors look at different cues while having the same assessment result)
  - Lack of interpretability and validity in its output.

- Our design of **separating observation and judgment** is a promising approach in such highly contextual domains.
  - Importantly, our goal is not replacing human decision, but helping them.

【Case Study】

Also read → REsCUE (CHI'19) and INWARD (CHI'20)

AI for human assessment: What do professional assessors need?

Riku Arakawa[†]

Carnegie Mellon University

Hiromu Yakura[†]

University of Tsukuba / AIST

[†] Equal contribution

In collaboration with ACES Inc.

CHI23
Hamburg, Germany | Hybrid
April 23-28, 2023
reCHInnecting

Limitations and future work

- Larger study
  - findings were obtained from studies with a single assessment company
  - number of professional assessors involved in the study was small

- Effects of AI on the final decision by assessors
  - how the system can further contribute to assessors’ decision-making