AI for human assessment: What do professional assessors need?

RIKU ARAKAWA*, Carnegie Mellon University, USA ACES, Inc., Japan

HIROMU YAKURA*, University of Tsukuba / National Institute of Advanced Industrial Science and Technology (AIST), Japan

We present our case study that aims to help professional assessors make decisions in human assessment, in which they conduct interviews with assessees and evaluate their suitability for certain job roles. Our workshop with two industrial assessors revealed that a computational system that can extract nonverbal cues of assesses from interview videos would be beneficial to assessors in terms of supporting their decision making. In response, we developed such a system based on an unsupervised anomaly detection algorithm using multimodal behavioral features such as facial keypoints, pose, head pose, and gaze. Moreover, we enabled the system to output how much each feature contributed to the outlierness of the detected cues with the purpose of enhancing its interpretability. We then conducted a preliminary study to examine the validity of the system's output by using 20 actual assessment interview videos and involving the two assessors. The results suggested the advantages of using unsupervised anomaly detection in an interpretable manner by illustrating the informativeness of its outputs for assessors. Our approach, which builds on top of the idea of separation of observation and interpretation in human-AI teaming, will facilitate human decision making in highly contextual domains, such as human assessment, while keeping their trust in the system.

CCS Concepts: • **Human-centered computing → Empirical studies in interaction design**; **Empirical studies in collaborative and social computing**; Visualization application domains.

Additional Key Words and Phrases: human-AI collaboration, human assessment, behavior analysis

ACM Reference Format:

Riku Arakawa and Hiromu Yakura. 2022. AI for human assessment: What do professional assessors need?. In *TRAIT '22: ACM CHI Workshop on Trust and Reliance in AI-Human Teams, April 30, 2022, New Orleans, LA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Human assessment is a process aimed to evaluate and make decisions on candidates regarding their suitability for certain types of employment [17]. It has been playing an important role in human resource development, especially for management jobs [8], where the candidates' skills as a manager in their organizations are examined. The assessment process involves multiple methods such as job-related simulations, interviews, and psychological tests. Among them, interviews are commonly used and viewed as a reliable method [14] in which professional assessors conduct short interviews with candidate assessees.

However, the current workflow of interviews in human assessment contains some troubles. For example, assessors often need to review the interviews that are video-recorded to manually check them in detail before making final decisions. In addition, it is pointed out that errors tend to occur due to assessors' subjectivity [14], which in turn suggests the demand for enhanced training programs of the assessors.

Given the development of techniques for human behavior analysis during a conversation [2, 15], we speculated that developing computation support is possible to facilitate the assessment process. However, it remains to be explored what are appropriate systems for achieving such human-AI collaboration in human assessment. For example, as Arrieta

*These authors contributed equally and are ordered alphabetically.

2022. Manuscript submitted to ACM

et al. discussed [5], it is obviously not a good idea to develop a black-box prediction system that just outputs a score for each candidate based on the recorded video of their interview. This is because it would be hard for assessors to rely on the system's output without any explanation, especially in a field like human assessment where they need to make a highly complicated and sensitive decision [12]. In other words, we need to design a human-AI relationship where assessors' trust in AI systems can be nurtured and to develop a plausible system that helps them through the decision-making process.

In this paper, we report our story of designing computational support for human assessment through a workshop involving professional assessors in which we elicited several system requirements. Based on the requirements, we developed a proof-of-concept system that helps assessors by extracting nonverbal behavioral cues from interview videos and presenting them to assessors in an interpretable manner. We also summarize the lessons we learned from a feasibility evaluation using the system, which highlights the effectiveness of our system design, that is, the separation of observation and interpretation based on unsupervised anomaly detection in constructing human-AI collaboration.

2 SYSTEM REQUIREMENTS

To identify the requirements for the supporting tool in human assessment, we first conducted a workshop with assessors from a Japanese human-assessment company. This company conducts approximately 2,000 assessment sessions annually with their clients, mainly to check the suitability as company managers of the employees at their client enterprises. The workshop was informally organized and went on to cover diverse aspects in developing such a system through conversation. We involved two proficient assessors, who regularly manage and educate other assessors as well as conducting assessments. The workshop lasted for approximately three hours in total.

2.1 How the assessment is conducted

First of all, we asked the assessors about the overview of their assessment routine to get ourselves familiarized with the process. As follows, they described the details of the process consisting of two sessions: an interview phase and a review phase.

In the interview phase, an assessor plays a certain role and seeks to evaluate how their assessee behaves in the given scenario through a one-on-one interview. For example, to examine their skills as a manager, the assessor plays a role of a subordinate who is not satisfied with their current job. Then, the assessee is asked to address the issue during the interview as the subordinate's manager. To profoundly examine assessees' behaviors, assessors are required to act well as the given role and to strategically behave on the spot to simulate plausible situations that are difficult for the assessees. This interview phase usually lasts for approximately 10–15 minutes and is video-recorded so that the assessors can review it later in the review phase.

In the review phase, the assessors play back the recorded video of the interview phase for each assessee. This review phase usually lasts for 30 minutes. Here, they inspect the assessee's behaviors, both verbal and nonverbal ones, and try to find cues for evaluating them which they might have missed during the interview phase. Then, they make a decision on the assessee's skills and suitability for certain jobs (e.g., "This candidate is B+, suitable for being a manager to some degree, but has some room for improvement.") Moreover, this review phase sometimes involves other assessors who independently check the video in order to validate the final decision. The two assessors who attended this workshop often do it since they are senior to other assessors.

2.2 What support assessors need

Then, we asked them about their ideas about how companies can help them to facilitate their assessment process. Interestingly, both of them agreed that the final decision should be made by humans, not by computers. This is because they thought computers would be incapable of making accurate decisions due to the complex nature of human assessment, and thus, they would ignore the output of computers when they have different opinions. In other words, such a system would likely result in nothing additive for their assessment. Moreover, if the system outputs only the final decision, it will easily cause a critical problem when their client asks for the reasons behind it, which assessors will not be able to explain.

On the other hand, they expected computers to help them reduce the time required to watch the full length of the recorded video. In particular, although the review phase is important to find cues that influence the final decision, watching the video of sessions they actually participated in is time-consuming and mental-demanding. In this sense, they agreed that systems which automatically extract cues for their decision-making would be beneficial in terms of reducing the time spent in the review phase.

2.3 What cues will be helpful

Next, in order to enable computers to detect such insightful cues from the recorded videos of the interview session, we sought to elucidate their characteristics. The two assessors agreed that they mainly try to check assessees' nonverbal behaviors (e.g., their body movement) in the review phase. This is because nonverbal behaviors are relatively implicit compared to verbal information such that their implications can be sometimes missed during the interview phase. In response to this finding, we introduced some of the recent works in human behavior analysis to them and explained that such behaviors can be captured by computers to some extent. We first showed demo videos of several computer vision techniques that digitize our behaviors, such as body pose estimation or facial keypoints detection [10]. Then, we introduced works that estimate human behavioral features based on digitized behaviors, such as attention estimation based on human head poses [4], nod detection based on facial keypoints [13]. In addition, to provide better images of what computers can do for human assessment, we introduced a work by Sanchez-Cortes et al. [16], in which they utilized digitized behaviors to detect emergent leaders in a group discussion. Their work introduced several hand-crafted features such as the number of segments that the amount of one's body movement exceeded a certain threshold, and applied a rule-based method using the features to calculate scores for each candidate.

Although we had anticipated that such works (i.e., scoring based on some rules using human behavioral features such as nodding) could be helpful in our case as well, the two assessors both disagreed with it after contemplation, mentioning several reasons. First, they clarified that their assessment process does not involve explicitly counting such behavioral features, nor do they believe it is meaningful. They explained this reason with a simplified example; people who often nod are not necessarily suitable for managers. Rather, they try to base their impression and judgment, which originally come from the interview phase, on those objective signals within context. At this point, they mentioned that they often focus on scenes where the assesses showed unseen behaviors, such as sudden use of big gestures, because inferring what caused such changes could provide them insights that form the assessors' impression and judgment.

Secondly, the assessors were skeptical about the accuracy of the captured behavioral features (e.g., attention level, nodding). In detail, although they found the digitized behaviors shown in the demo videos precise to some extent, they questioned the validity of heuristics subsequently used to estimate those features (e.g., time window and angular threshold of the head pose used to detect nodding). Moreover, they were concerned about individual differences; such heuristics must be dependent on each assessee in order to precisely capture the features. In sum, it would be hard for the assessors to trust a system based on such behavioral features (e.g., the number of noddings) due to the limited validity of the algorithms used to estimate them. Then, the system would result in being excluded from their process of decision making [19].

2.4 Our conclusion

From the above discussion, we agreed that developing a system that outputs some indices about assessees based on behavioral features is not an optimal approach from the perspective of human-AI teaming. In contrast, we paid attention to the assessors' idea that extracting informative scenes based on assesses' nonverbal behaviors would be helpful to review the interview phase efficiently. This approach represents the separation of observation and judgment; computers *watch* the whole video on behalf of humans and humans *make decisions* by checking scenes extracted by the computers. Its benefit lies in the design that human assessors can make final decisions within the context of human-to-human communication, which is still hard for computers to deal with [9]. As a result, it would minimize the risk of losing the assessors' trust in the system due to uninterpretable decisions or unreliable indices made by the system.

In fact, the efficiency of such design (i.e., separation of observation and judgment) in reflecting videos has been confirmed in the field of executive coaching [3]. They applied unsupervised anomaly detection [18] to the time-series data of the digitized human behaviors (e.g., body pose, facial keypoints) of coachees in coaching sessions. They found that presenting scenes where coachees' nonverbal behaviors changed in comparison to other scenes enables professional coaches to reflect on the coaching sessions efficiently. Since the algorithm to detect such scenes is based on unsupervised learning and does not involve pre-defined rules, it can mitigate biases that may arise in designing heuristics to estimate behavioral features (e.g., attention, nodding). Given the concern on such biases as the assessors mentioned in Section 2.3, we expected that utilizing unsupervised learning is crucially helpful when we seek to achieve human-AI teaming in human assessment without losing assessors' trust.

To conclude, the assessors and the authors agreed to introduce the design of separating observation and judgment to facilitate their assessment in the review phase. More specifically, we anticipated that the system would extract several scenes from the interview videos based on unsupervised anomaly detection and the assessors would mainly review the scenes to make decisions on the assessees efficiently.

3 FEASIBILITY EVALUATION

We then conducted a preliminary study to examine whether such an approach is feasible. Here, as a first step, we sought to measure the accuracy of the unsupervised anomaly detection algorithm in terms of how its outputs are actually informative to assessors.

3.1 Material

We prepared 20 videos of the assessment interviews that had been conducted at the company before this study. Neither of the two assessors participated in the interview sessions, and they were asked to assume the situation that they would be reviewing the videos independently, as we mentioned in Section 2.1. Thus, they reviewed the videos by focusing on verbal and nonverbal behaviors to find cues for evaluating assessees, as usual. These assessment interviews were conducted online due to the COVID-19 situation, and each video was approximately 10-minutes long.

3.2 Algorithm Detail

The algorithm we used is adopted from [2], which is unsupervised anomaly detection with multimodal signals as input, as mentioned in Section 2.4. Consulting with the assessors, we chose the following four nonverbal behavioral data as input to be used: facial keypoints, body pose, head pose, and gaze. As mentioned in Section 2.3, each modality is accurately digitized from videos using recent computer vision techniques. We used AlphaPose [10] to get facial keypoints and body pose. Based on the estimated facial keypoints, the head pose was calculated by solving a perspective-n-point problem. Finally, based on the facial keypoints and head pose, the gaze was estimated based on the RTGENE model [11].

These behavioral data are calculated every frame and then the anomaly detection model processes its time-series data. Specifically, the Gaussian Mixture Model (GMM) is fitted to the distribution of the frames in an online unsupervised learning manner. Then, whenever a new data comes, the model outputs its outlierness based on the parameters obtained using the previous data while updating the parameters to fit the distribution including the new data. In practice, the model processes the data window by window (i.e., data observed within a specific duration) and outputs the outlierness on a batch basis. Based on the sequence of the outlierness, we can identify anomaly scenes in the interview video that are likely to be informative to assessors as nonverbal cues. In addition, the model is also capable of identifying the most anomalous and representative frame within each window. For more details, please refer to [2].

Moreover, in this study, we extended their algorithm to enhance the interpretability of its output. That is, we enabled the GMM model to output how each behavioral modality contributed to the estimated outlierness of the window. This contribution of each modality is calculated as the change in the likelihood when we overwrite the parameters of the GMM model to ignore the corresponding feature in the estimation of the outlierness. If the outlierness decreased largely, it implies that the cause of the outlier was likely due to the ignored feature, and vice versa. In this way, we can identify the modality that most contributed to the change by comparing the decrease of the outlierness. We anticipated that this extension could provide the assessors with further capabilities to interpret the output of the model. This would meet the common practice in constructing better AI systems that recommends providing the clear attribution of their outputs to the corresponding inputs [1].

3.3 Procedure

This study involved the two professional assessors who participated in the workshop Section 2. We asked them to independently review the 20 videos we prepared in Section 3.1 by focusing on the assessees' nonverbal behaviors. Then, they were requested to list the top 10 scenes for each video that are important to assess the assessee. When they extracted such a scene, they were also asked to describe which behavior modality they based their thoughts on in a text, e.g., "At this moment, the assessee is making up his smile a little too much [facial keypoints]" and "Her eyes are scurrying and she is restless [gaze]." Finally, we asked them to write evaluations on the assessees as they usually do.

At the same time, the algorithm described in Section 3.2 processed each of the 20 videos. The window size was set to be 15 seconds. An example of the algorithm's output, i.e., time-series likelihood of scenes, is shown in Figure 1. Based on the likelihood values, the algorithm then extracted the top 10 anomalous scenes according to the order of magnitude of the likelihood. Finally, we compared the scenes extracted by the algorithm with those extracted by the two professional assessors.

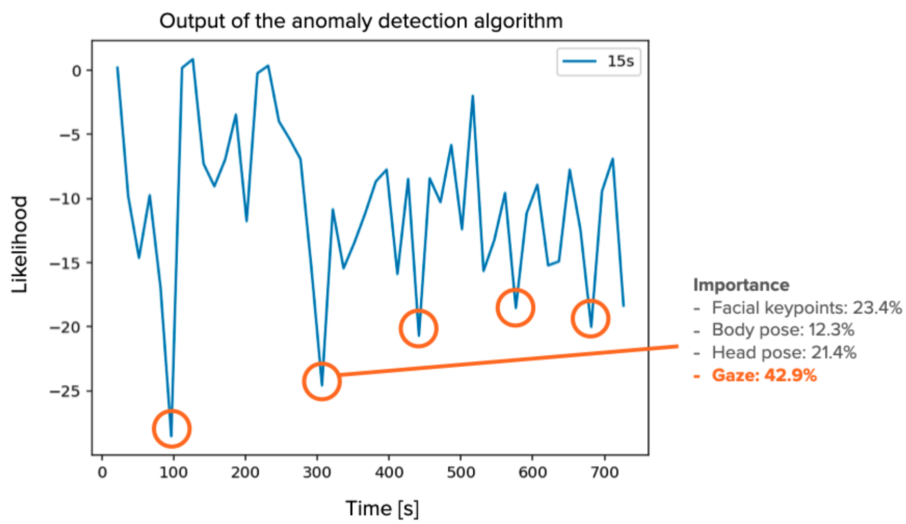

Fig. 1. Example output of the anomaly detection algorithm applied to one interview video. It calculates time-series likelihoods of each scene (15-seconds window) based on multimodal behavioral data. For each detected scene, the algorithm outputs the importance of each modality. For example, the detected scene around 300 seconds most likely comes from the gaze data.

3.4 Findings

Before examining the output of the algorithm, we first inspected the annotation of the professional assessors. Interestingly, we found that the cues each assessor regarded as important diverged from those of each other to some extent. Specifically, the proportion of cues that both assessors listed in the annotation data was 52.5% among all the listed cues. However, their overall evaluations for the 20 assessees were consistent.

From this result, we can infer that, even though the scenes the assessors had focused on were varied, they had something common in what they read from the cues, resulting in consistent evaluations. It also implies that detecting such cues would be intractable for supervised learning because we will have difficulty in constructing clear criteria about which cues should be detected, hindering the preparation of training data. This finding guided us to reframe what professional assessors expect AI systems to do. In detail, considering that each assessor focuses on different cues in videos, we found that it is not meaningful nor practical to detect cues that are completely coincident with their annotation. Rather, since the review phase is prepared to find some missed cues (See Section 2.1), finding cues that do not match the annotation data but are informative by using computers can be valuable.

We then evaluated the agreement between the outputs of the algorithm and the annotation data. As a result, we confirmed that approximately 38.0% of the cues the assessors regarded as important had been detected as anomalous by our algorithm. In addition, when we set the algorithm to enumerate the top 15 anomalous scenes, we found that the value of recall increased to 51.0%, which was almost equivalent to the agreement rate between the two assessors. This result supports the comments of the assessors (Section 2.3) and the conclusion we reached in the workshop (Section 2.4); scenes detected by applying unsupervised anomaly detection to digitized behaviors can serve as a basis for assessors' decision making.

To confirm this, we also qualitatively examined the detected cues with the guidance of the assessors regarding whether or not they are informative. We all first found that our algorithm worked well to detect important scenes even though the modality that the algorithm focused on can differ from that of assessors. For instance, the algorithm detected a scene that one of the professional assessors marked as important, in which an assessee froze for a moment to figure out the best response to an assessor's critical question. In this case, the algorithm displayed that the facial keypoints most contributed to the outlierness of the scene. This allowed us to infer that the algorithm detected the change in the keypoints around the assessee's mouse, which was caused by stopping the utterance, and the change in the keypoints around the assessee's eyes, which was caused by shutting the eyes tightly to ponder.

This room for interpretation provided by the algorithm used was further beneficial when it detected a scene that was originally not marked as important by the professional assessors. For example, the algorithm detected a scene in which an assessee was emphasizing the words that had been repeatedly used to tell their dissenting opinion. Here, the algorithm pointed out the change in the assessee's facial keypoints; indeed, the scene illustrated the figure of the assessee getting so close to the camera to emphasize their words that half of their keypoints were disappeared. Interestingly, the overall evaluation by the professional assessors for the assessee coincided with the detected scene. Specifically, they remarked that the assessee "was highly persistent" and "showed impatience, especially during the first half." After reflecting on this specific scene with the assessor, we agreed that this scene also has informative cues for the final judgment which the assessors had not noticed in the review phase. Given this case, we also agreed that, if the algorithm had not been designed to output such interpretable results, we might have ignored it because of our incapability of explaining its reason. In other words, we would likely assume that such output is a false-positive case of a black-box system.

Furthermore, we found that the algorithm yielded some (actual) false-positive results, such as a scene in which an assessee stopped their words for a moment to hold in a burp. The scene was detected from the changes in the facial keypoints of the assessee but is a physiological phenomenon that is apparently not informative for the assessment. Still, the transparent mechanism of the algorithm allowed us to infer the reason behind the detection, which prevents the deterioration of the trust in the algorithm from assessors.

In short, the performance of the algorithm was favorably received by the professional assessors. Particularly, they found that the design of separating observation and judgment that we reached a consensus on in Section 2.4 would facilitate their assessment, even though the algorithm does not completely replicate their annotation. We infer that there are two factors that helped the assessors regard the design as acceptable. First, the design allows them to take into consideration various factors specific to each assessee that are highly human-contextual and difficult to be captured by computers. Second, the design guides the assessors to infer the reason behind the detection, rather than excluding the algorithm from their process of decision making.

Now, we are working on deploying the algorithm into the real workflow of the assessment. Through such a deployment, we will evaluate how our system influence assessors' final decision. While we believe that assessors can make better overall decisions by benefiting our system in a trustful manner, we acknowledge that the deployment of AI systems does not always improve the entire performance of a human-AI team [7, 20]. Given that, we want to conduct an in-depth evaluation to figure out the best practice in bringing AI systems into human assessment, which may cover not only interface designs but also operational aspects.

4 LESSONS LEARNED

In this paper, we presented our case study conducted with professional assessors where we elucidated what they need from computer systems to realize human-AI teaming in human assessment, a domain entwined with highly human contexts. Based on their demands drawn in the workshop, we developed a system based on unsupervised anomaly detection that can detect informative nonverbal cues from interview videos without relying on any heuristics. As a first step of supporting the assessors with this system, we evaluated how reliable its outputs are in terms of accuracy by comparing them with the annotations done by two professional assessors. As a result, our algorithm can detect not only scenes that the assessors had focused on for their assessment but also those that had been not noticed but informative, while with a capability of presenting them in an interpretable manner. The results confirmed the efficacy of our design, namely, the separation of observation and interpretation in developing a supporting system for human assessment. We are currently conducting further verification of this system by refining its interface and evaluating the efficiency made possible by the system through a quantitative and qualitative analysis.

If we had started developing an algorithm to capture informative cues in the paradigm of supervised classification (e.g., detecting specific human behavioral features such as nodding), it would not have worked effectively in concert with professional assessors. The reason is validated by our findings of the study (Section 3.4). First, it was suggested that different assessors look at different cues while having the same assessment result. This inconsistency and unclear boundaries of classes make it difficult to prepare a dataset to train a model. Moreover, even if we could train such a model based on supervised learning, it would lack transparency and validity in its output, hindering the assessors from constructing a mental model about the behavior of the AI model. As we see in Section 2.3, this would result in the failure of establishing a trust in human-AI relationships [6], leading the assessors to ignore the output of the model due to false positives. Given these points, we conclude that the separation of observation and interpretation made possible by unsupervised anomaly detection will be a promising approach to building human-AI teaming, especially in highly contextual domains that inevitably require a human decision.

Acknowledgment

We would like to thank the support of LEAD CREATE Inc., a Japanese human-assessment company. This work was supported in part by JST ACT-X Grant Number JPMJAX200R and JSPS KAKENHI Grant Numbers JP21J20353.

REFERENCES

[1] Saleema Amershi, Daniel S. Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi T. Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 ACM CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, Article 3, 19 pages. https://doi.org/10.1145/3290605.3300233

[2] Riku Arakawa and Hiromu Yakura. 2019. REsCUE: A framework for REal-time feedback on behavioral CUEs using multimodal anomaly detection. In *Proceedings of the 2019 ACM CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 572. https://doi.org/10.1145/3290605.3300802

[3] Riku Arakawa and Hiromu Yakura. 2020. INWARD: A Computer-Supported Tool for Video-Reflection Improves Efficiency and Effectiveness in Executive Coaching. In *Proceedings of the 2020 ACM CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 1–13. https://doi.org/10.1145/3313831.3376703

[4] Riku Arakawa and Hiromu Yakura. 2021. Mindless Attractor: A False-Positive Resistant Intervention for Drawing Attention Using Auditory Perturbation. In *Proceedings of the 2021 ACM CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 99:1–99:15. https://doi.org/10.1145/3411764.3445339

[5] Alejandro Barredo Arrieta, Natalia Díaz Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Information Fusion* 58 (2020), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[6] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S. Lasecki, Daniel S. Weld, and Eric Horvitz. 2019. Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. *Proceedings of the 7th AAAI Conference on Human Computation and Crowdsourcing*, 2–11.

[7] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S. Weld, Walter S. Lasecki, and Eric Horvitz. 2019. Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*. AAAI Press, Menlo Park, CA, 2429–2437. https://doi.org/10.1609/aaai.v33i01.33012429

[8] Douglas W. Bray and Donald L. Grant. 1966. The Assessment Center in the Measurement of Potential for Business Management. *Psychological Monographs: General and Applied* 80, 17 (1966), 1–27. https://doi.org/10.1037/h0093895

[9] Development, Concepts and Doctrine Centre. 2018. *Joint Concept Note 1/18: Human-Machine Teaming*. Technical Report. Ministry of Defence, UK.

[10] Haoshu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. 2017. RMPE: Regional Multi-person Pose Estimation. In *Proceedings of the 2017 IEEE International Conference on Computer Vision*. IEEE, New York, NY, 2353–2362. https://doi.org/10.1109/ICCV.2017.256

[11] Tobias Fischer, Hyung Jin Chang, and Yiannis Demiris. 2018. RT-GENE: Real-Time Eye Gaze Estimation in Natural Environments. In *Proceedings of the 15th European Conference on Computer Vision*, Vol. 11214. Springer, Cham, Switzerland, 339–357. https://doi.org/10.1007/978-3-030-01249-6_21

[12] Ella Glikson and Anita Williams Woolley. 2020. Human Trust in Artificial Intelligence: Review of Empirical Research. *Academy of Management Annals* 14, 2 (2020), 627–660. https://doi.org/10.5465/annals.2018.0057

[13] Kiyosu Maeda, Riku Arakawa, and Jun Rekimoto. 2022. CalmResponses: Displaying Collective Audience Reactions in Remote Communication. *arXiv* abs/2204.02308 (2022). https://doi.org/10.48550/arXiv.2204.02308

[14] Chris Roberts, Tyler Clark, Annette Burgess, Michael Frommer, Marcia Grant, and Karyn Mossman. 2014. The Validity of a Behavioural Multiple-Mini-Interview within an Assessment Centre for Selection into Specialty Training. *BMC Medical Education* 14, 1 (2014), 1–11. https://doi.org/10.1186/1472-6920-14-169

[15] Ognjen Rudovic, Mihalis A. Nicolaou, and Vladimir Pavlovic. 2017. Machine Learning Methods for Social Signal Processing. In *Social Signal Processing*, Judee K. Burgoon, Nadia Magnenat-Thalmann, Maja Pantic, and Alessandro Vinciarelli (Eds.). Cambridge University Press, Cambridge, UK, 234–254. https://doi.org/10.1017/9781316676202.018

[16] Dairazalia Sanchez-Cortes, Oya Aran, Marianne Schmid Mast, and Daniel Gatica-Perez. 2012. A Nonverbal Behavior Approach to Identify Emergent Leaders in Small Groups. *IEEE Transactions on Multimedia* 14, 3-2 (2012), 816–832. https://doi.org/10.1109/TMM.2011.2181941

[17] Riccardo Sartori, Arianna Costantini, and Andrea Ceschi. 2020. Psychological Assessment in Human Resource Management: Discrepancies between Theory and Practice and Two Examples of Integration. *Personnel Review* ahead-of-print (2020), 1–15. https://doi.org/10.1108/pr-05-2019-0281

[18] Kenji Yamanishi, Jun'ichi Takeuchi, Graham J. Williams, and Peter Milne. 2004. On-Line Unsupervised Outlier Detection Using Finite Mixtures with Discounting Learning Algorithms. *Data Mining and Knowledge Discovery* 8, 3 (2004), 275–300. https://doi.org/10.1023/B:DAMI.0000023676.72185.7c

[19] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Jianlong Zhou, and Fang Chen. 2019. Do I Trust My Machine Teammate?: An Investigation from Perception to Decision. In *Proceedings of the 24th ACM International Conference on Intelligent User Interfaces*. ACM, New York, NY, 460–468. https://doi.org/10.1145/3301275.3302277

[20] Guanglu Zhang, Ayush Raina, Jonathan Cagan, and Christopher McComb. 2021. A cautionary tale about the impact of AI on human design teams. *Design Studies* 72 (2021), 100990. https://doi.org/10.1016/j.destud.2021.100990