Human Factors and Ergonomics in Design, Vol. 210, 2026, 65–74
https://doi.org/10.54941/ahfe1007426

Cross-Cultural Impressions and Cue Weighting of Avatars for Emergency Announcements: Evidence From China and Japan

Liwen Zhang[1], Hidekazu Takahashi[1], Toru Nakata[1,2], and Toshikatsu Kato[1]

[1]Graduate School of Science and Engineering, Chuo University, Tokyo, Japan
[2]National Institute of Advanced Industrial Science and Technology, Tokyo, Japan

ABSTRACT

Avatars are increasingly used for serious emergency announcements, but cross-cultural differences in impressions and cue weighting remain unclear. We compared participants from China (n = 40) and Japan (n = 40) evaluating four emergency-announcement avatars. In Part 1, participants watched avatar-based announcement videos and completed comprehension checks; understanding was generally high, while the Chinese participants scored higher than the Japanese with avatar-dependent differences. In Part 2, participants rated the avatars using static reference images (for identification) and reported the importance of six cues (clothes, gender, voice, style, speed, race). Mixed-design analyses showed that cross-cultural differences were most robust for authority, whereas trustworthiness and affability showed weaker effects. The Chinese participants assigned higher overall cue importance than the Japanese, and authority cue-importance patterns differed by country. A supplementary dlib-based facial landmark profile provided descriptive context for avatar-specific China–Japan gaps. These findings highlight the need for country aware validation in designing serious-communication avatars.

Keywords: Serious communication, Emergency announcements, Cross-cultural comparison, Authority impression, Cue importance

INTRODUCTION

Avatars are increasingly used to deliver information in serious communication settings, where recipients must rapidly judge whether a message is reliable and actionable (Stieglitz et al., 2022). Such judgments are partly shaped by facial appearance; face evaluation has been described by fundamental dimensions such as trustworthiness and dominance/competence (Oosterhof and Todorov, 2008). For applied emergency announcements, understanding how users form impressions of an announcing avatar is important because these impressions may influence perceived legitimacy and acceptance of instructions (Hancock et al., 2011).

Impression formation is not purely stimulus-driven: the cues people rely on can vary across countries. This raises the possibility that identical avatar

Received February 18, 2026; Revised April 2, 2026; Accepted April 17, 2026; Available online July 20, 2026

designs may elicit different evaluations across audiences, particularly for authority related impressions. To examine this, we compare participants from China (*CN*) and Japan (*JP*) evaluating the same set of emergency-announcement avatars (*A1–A4*). *Part 1* assesses comprehension from avatar-based videos, and *Part 2* evaluates *Trustworthiness (Tw)*, *Affability (Fa)*, and *Authority (Au)* from static images together with perceived cue importance.

This paper makes three contributions:

1. We quantify cross-national differences in comprehension and in *Tw/Fa/Au* ratings for the same avatars.
2. We analyze cue-importance patterns and cue rating associations to characterize cross cultural cue weighting, with a focus on *Authority*.
3. We provide descriptive 'Dlib' based facial-geometry profiles of the stimuli to contextualize stimulus-level differences (without inferential modeling).

RELATED WORK

Research on face perception shows that people rapidly infer social traits from facial appearance, and these impressions can be captured by a small set of underlying dimensions (e.g., trustworthiness/valence and dominance/competence) (Oosterhof and Todorov, 2008). In parallel, human agent interaction studies indicate that an embodied agent's visual design (e.g., age/gender cues, realism) can systematically shift users' evaluations and acceptance, including trust-related judgments in applied domains such as healthcare communication (ter Stal et al., 2020).

However, cue utilization differs across countries. For example, Japanese and U.S. observers weight eye versus mouth information differently when interpreting faces (Yuki et al., 2007), suggesting that identical avatar appearances may elicit different impressions across audiences. Building on these lines of work, our study connects outcome differences (*Tw/Fa/Au*) with participants' reported cue importance and cue–rating associations within each country to provide more interpretable design guidance. From a cue based judgment perspective (i.e., people infer traits by weighting multiple observable cues), cross-cultural differences may arise not only in outcomes (*Tw/Fa/Au*) but also in the cues people emphasize.

METHOD

Participants

A total of 80 participants took part in the study: 40 from China (*CN*) and 40 from Japan (*JP*). All responses were complete and included in the analysis. Basic demographics (e.g., age and gender) were collected for descriptive reporting. All participants provided informed consent prior to participation, and the study followed relevant institutional guidelines.

Stimuli

The stimulus set comprised four short emergency announcement videos featuring four avatars (A1–A4). Each video presented a hotel-related emergency scenario (typhoon, earthquake, or tsunami) and delivered concise

safety instructions (Table 1). Message scripts were matched in length and instruction style and were localized by language (Chinese for CN; Japanese for JP) using semantically equivalent wording. Video presentation parameters (e.g., resolution, audio level, and duration) were kept consistent across stimuli.

Table 1: Overview of the four emergency announcement videos and stimulus avatars (A1–A4).

Avatar	Scenario	Key Instructions (Ultra-Brief)
 *A1*	Typhoon	*Use emergency stairs (no elevator); follow staff to high-ground park; wear shoes; carry phone/valuables/water.*
 *A2*	Earthquake	*Protect yourself during shaking; after shaking, use emergency stairs (no elevator); evacuate to courtyard via lobby.*
 *A3*	Typhoon	*Stay indoors; keep away from windows; if needed, use emergency stairs to indoor evacuation route; prepare flashlight.*
 *A4*	Tsunami advisory	*Avoid coast; if shaking, go upstairs via stairs (no elevator); use mountain-side stairs to high-ground park plaza.*

Procedure

Participants watched four emergency announcement videos (*A1–A4*) and completed a comprehension check after each video. They then rated the corresponding static avatar images on *Tw, Fa,* and *Au*, the order of *A1–A4* was randomized.

Measures

Comprehension Check (Part 1 Video Based)

After each announcement video (*A1–A4*), participants completed a comprehension check consisting of 10 candidate statements and were instructed to select five statements that were consistent with the video content. Scoring was computed as an accuracy score across all 10 statements (0–10): one point was awarded for each statement that was correctly handled, i.e., selecting a true statement or not selecting a false statement. Thus, the score reflects overall comprehension accuracy rather than only the number of correct selections.

Impression Questionnaire (Part 2 Image Based)

In *Part 2*, participants rated each avatar while viewing a static reference image extracted from the corresponding *Part 1* video. The image served only to indicate

which avatar was being evaluated. Participants were instructed to base their judgments on the immediately preceding video (including audio delivery such as voice and speech speed), rather than on the static image alone. Participants evaluated each avatar (A*1–A4*) using a static image on three impression dimensions: *Trustworthiness (Tw)*, *Affability (Fa)*, and *Authority (Au)*. Each dimension consisted of 7 items (21 items in total per avatar). All items were rated on a 5 point Likert scale (1 = strongly disagree, 5 = strongly agree).

The *Tw* and *Au* items were adapted from established source credibility measures that operationalize trustworthiness and competence/authority related evaluations of a communicator, while *Fa* items were adapted from likability/affiliation measures (Ohanian, 1990). For each dimension, item scores were averaged to form a composite score; internal consistency was assessed using Cronbach's α (reported in Supplementary Materials). To explore how participants formed these impressions, we additionally collected self-reported cue importance ratings as a proxy for the cues participants believed they relied on when judging *Tw/Fa/Au*.

Cue importance was rated on a 1–5 scale for six cues (clothes, gender, voice, style, speech speed, race) after each avatar evaluation, and was collected separately for Tw, Fa, and Au. Cue labels were presented in the participant's native language with semantically matched wording.

Table 2: Impression dimensions and example items used in Part 2 (1–5 response scale).

Impression Dimension	An Example out of the 7 Questions
Trustworthiness (Tw)	*"This avatar appears trustworthy."*
Affability/Approachability (Fa)	*"This avatar feels easy to approach."*
Authority (Au)	*"This avatar appears authoritative."*

Dlib-Based Stimulus-Level Cue Profiles (Descriptive)

We extracted 68 facial landmarks from each stimulus avatar image using the *dlib* landmark detector, and derived eight geometric cue indices (e.g., eye/eyebrow, nose, mouth, jaw/chin related measures). Values were z-standardized across *A1–A4* to highlight *relative* differences among the four stimuli. Because the stimulus set was small (n = 4), these cues were reported descriptively (no inferential modeling) to contextualize the impression and cue-importance results (King, 2009; Kazemi and Sullivan, 2014).

Data Analysis

In Part 1, comprehension accuracy scores (0–10) were summarized by country and avatar and analyzed using mixed-design ANOVA (country: between-subject; avatar: within-subject). In Part 2, *Tw/Fa/Au* item responses (5-point Likert) were averaged within each dimension (7 items per dimension) to obtain dimension scores for each avatar. We then compared *Tw/Fa/Au* by country and avatar using mixed-design ANOVA, and conducted cue analyses as described. Statistical significance was assessed at $\alpha = 0.05$. When sphericity was violated, Greenhouse–Geisser corrections were applied for within-subject effects.

RESULTS

Comprehension Check (Part 1)

Comprehension accuracy (0–10) was generally high (Fig. 1). A mixed-design ANOVA (country: between-subject; avatar: within-subject) showed significant main effects of country, $F(1,78) = 74.85$, $p < .001$, $\eta p^2 = .490$, and avatar, $F(3,234) = 13.87$, $p < .001$ (*Greenhouse–Geisser corrected;* $\approx = .878$), $\eta p^2 = .151$, as well as a significant country × avatar interaction, $F(3,234) = 6.93$, $p < .001$, $\eta p^2 = .082$. Descriptively, *CN* participants showed higher comprehension than *JP* across avatars, with the *CN–JP* gap varying by stimulus (largest for *A1/A4*; smallest for *A2*). Follow-up simple-effects tests (*CN* vs. *JP* within each avatar, Holm-corrected) showed higher comprehension scores for *CN* across all four avatars (all $p < .05$), with the largest gaps for *A1* and *A4* and the smallest gap for *A2*.

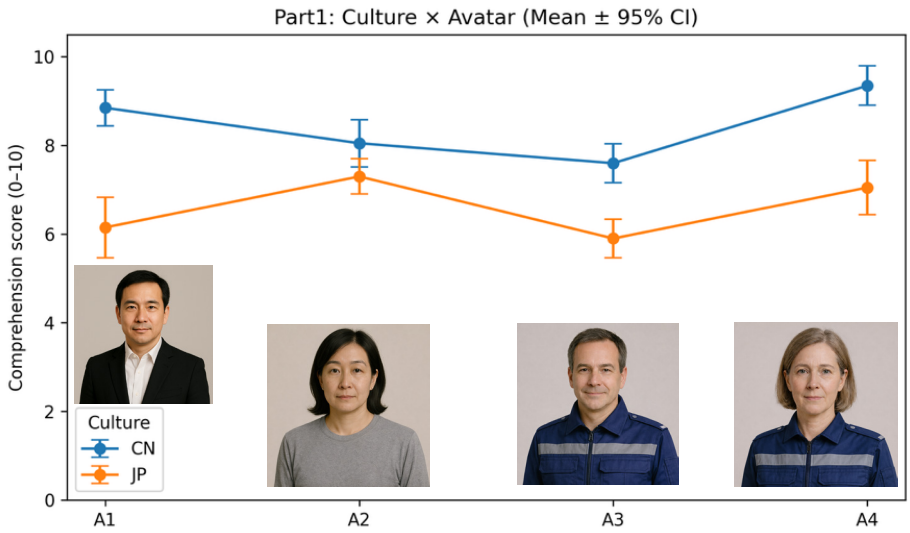

Figure 1: Comprehension accuracy by avatar (Mean ± 95% CI).

Impression Ratings *(Part 2: Tw, Fa, Au)*

For *Tw*, the mixed-design ANOVA showed a trend toward a country effect, of country, $F(1,78) = 3.76$, $p = .056$, $\eta p^2 = .046$, with no significant effects of avatar or interaction (*ps* ≥.43). For *Fa*, the main effect of avatar was significant, $F(3,234) = 2.86$, $p = .037$, $\eta p^2 = .035$, while the country main effect was marginal, $F(1,78) = 3.68$, $p = .059$, $\eta p^2 = .045$, and the interaction was not significant ($p = .359$). For *Au*, significant main effects were found for country, $F(1,78) = 6.88$, $p = .010$, $\eta p^2 = .081$, and avatar, $F(3,234) = 5.55$, $p = .001$, $\eta p^2 = .066$, with no interaction, $F(3,234) = 1.88$, $p = .134$, $\eta p^2 = .024$. Overall, *Au* ratings tended to be higher in *CN* than in *JP* (Fig. 2). Exploratory within-country pairwise comparisons (Holm-corrected) suggested that the avatar main effects for *Au* and *Fa* were primarily driven by the *CN* group (notably *A2* being lower in *Au* and *A3* being lower in *Fa*), whereas no reliable avatar differences were observed in the *JP* group.

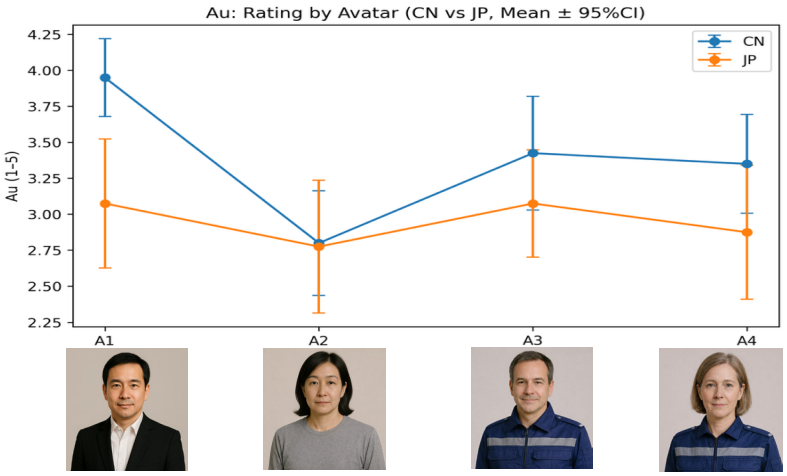

Figure 2: Authority ratings by avatar (Mean ± 95% CI). Mean Authority (Au) ratings (1–5) for each avatar (A1–A4) by country with 95% confidence intervals.

Cue Importance and Cue–Rating Associations

Cue importance ratings (1–5) were analyzed for six cues (clothes, gender, voice, style, speed, race). These cue-importance ratings were used as a self report proxy for participants' cue utilization/weighting during impression formation (Blackhurst et al., 2024). We treated the cues as potential inputs to *Tw/Fa/Au* judgments (rather than one-to-one determinants), and examined whether cross-cultural differences in cue weighting aligned with differences in impression ratings. Across *Tw, Fa*, and *Au*, mixed-design ANOVAs showed a strong main effect of country (all $p < .001$; $\eta p^2 \approx .27 - .28$), indicating higher overall cue importance in *CN* than *JP*. For *Au*, significant main effects of country, $F(1,78) = 28.63$, $p < .001$, $\eta p^2 = .268$, and cue, $F(5,390) = 5.61$, $p < .001$, $\eta p^2 = .067$, and a country × cue interaction, $F(5,390) = 5.39$, $p < .001$, $\eta p^2 = .065$, indicated cross-cultural differences in which cues were emphasized (Fig. 3).

To relate self reported cue weighting to impression outcomes, we computed within country participant level Spearman correlations between cue importance ratings and impression ratings using participant level averages across A1–A4 (*CN/JP*; $n = 40$ each). Due to space limitations, Fig. 4 visualizes the correlation matrix for *JP* (*Tw/Fa/Au* blocks separated by vertical lines); the *CN* matrix showed similarly modest associations and is not shown. Overall, correlations were small and dimension-specific: *Tw* exhibited relatively clearer cue-specific variation than *Fa*, whereas *Au* showed less regular cue associations, consistent with more context-dependent cue-to-authority mapping. Given the small number of stimuli ($n = 4$), these correlation results are reported as exploratory. We visualize *Au* because it exhibited the most robust cross-cultural divergence (the *CN–JP* gap was reliable only for *A1* after Holm-corrected *CN–JP* comparisons within each avatar), whereas *Tw* and *Fa* showed smaller and less stable differences and are omitted due to space constraints.

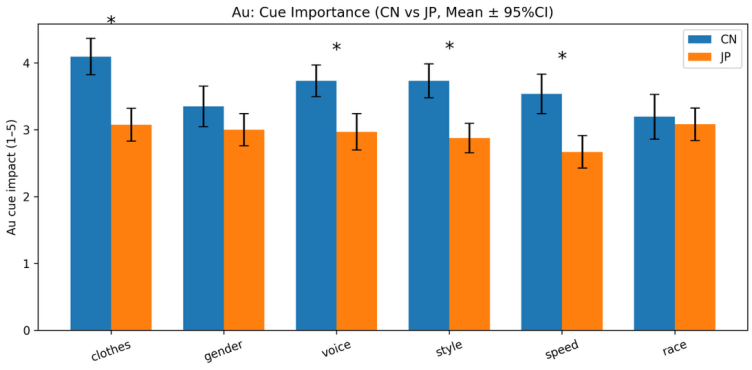

Figure 3: Cue importance for authority (Mean ± 95% CI). Mean cue-importance ratings (1–5) for Au by country (clothes, gender, voice, style, speed, race) with 95% confidence intervals.

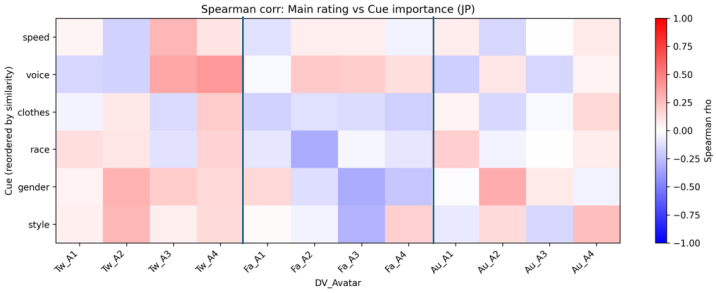

Figure 4: Spearman correlations between cue importance and impression ratings of the Japanese participants. Vertical lines indicate Tw/Fa/Au blocks.

Dlib-Based Facial Geometry Profiles

Fig. 5 summarizes stimulus-level dlib based facial geometry profiles (z-scores) for *A1–A4* alongside the *CN–JP* gaps in *Part 2* ratings (King, 2009; Kazemi and Sullivan, 2014), providing descriptive context. The profiles indicate distinct structural differences across avatars (e.g., mouth/jaw width and nose-related geometry), which may help interpret stimulus dependent cross cultural gaps. Given the small stimulus set ($n = 4$), these results are reported descriptively without inferential claims.

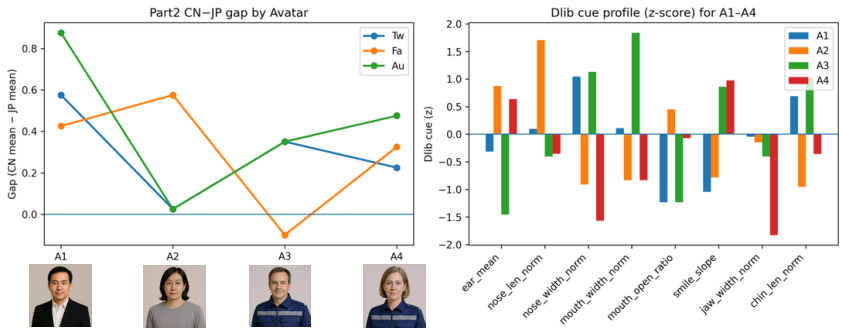

Figure 5: CN–JP rating gaps (Part 2) and dlib facial-geometry profiles (A1–A4). Left: mean gaps (CN – JP) for Tw/Fa/Au by avatar. Right: eight dlib-derived geometric indices from 68 landmarks, z-standardized across A1–A4.

DISCUSSION

Key Findings

Across both countries, comprehension was generally high, indicating that participants understood the emergency announcements. The cross-national pattern suggests that *Authority* impressions in serious communication may be especially sensitive to cultural interpretation, whereas *Trustworthiness* and *Affability* may depend on multiple cues that can partially compensate for one another. The cue importance results further imply that participants may adopt different cue weighting strategies across countries, which can lead to divergent evaluations even when the avatar appearance is held constant. These findings motivate cue aware design considerations for emergency style avatar communication.

Why Authority Was More Country Sensitive

Authority judgments in serious communication may rely strongly on culturally grounded expectations about "proper" communicators. This is consistent with the significant country main effect on *Au* (with a non-significant interaction), suggesting a stable cultural shift in *Au* ratings across avatars (Fig. 2). In contrast, *Tw* and *Fa* showed only marginal country effects, implying that trust and interpersonal warmth were less sensitive to cultural background under the present stimuli.

One plausible explanation is that authority judgments in emergency communication rely on culturally learned norms about legitimacy and role appropriate demeanour. In Japan, social expectations for "proper" official communication may emphasize contextual appropriateness and restraint, whereas Chinese participants may attend more to explicit, reportable cues that signal competence or leadership. This interpretation is consistent with the stronger country effect on cue importance (*CN* > *JP*) and the country × cue interaction for *Au*. In addition, the modest cue–rating correlations in *JP* suggest that self-reported cue weighting may not fully capture implicit integration processes; future work combining self-report with behavioral measures (e.g., eye-tracking or response times) could clarify how cues are used during impression formation.

Interpreting Cue-Importance Differences

A notable result is the strong main effect of country on cue importance (*CN* > *JP*) across dimensions. Moreover, the significant country × cue interaction for *Au* (Fig. 3) suggests that countries differ not only in how much they rely on explicit cues, but also in which cues are emphasized when judging authority. The modest cue–rating correlations observed in *JP* (Fig. 4) further imply that self-reported cue weighting may translate less directly into final impressions in this group.

Facial Geometry as an Auxiliary Explanation (Dlib)

To complement the impression and cue-importance results, we used the stimulus-level Dlib profiles (Fig. 5) as an objective reference of facial-geometry differences among *A1–A4*. Although no statistical inference is possible with four stimuli, the profiles help contextualize why the *CN–JP* gaps in *Part 2* varied by avatar. Specifically, avatars showing more pronounced mouth/jaw-related geometry

differences appeared to co-occur descriptively with larger cross-cultural gaps in Authority, whereas patterns for Affability were comparatively less systematic across avatars. These observations are consistent with the cue-importance results, where Authority judgments showed a significant country × CUE interaction, suggesting that cultures may weight available appearance cues differently depending on the avatar. Future work with a larger stimulus set should test these hypothesized links using inferential models.

Limitations and Future Work

The study used only four stimulus avatars, which limits inference about which specific visual features drive cross-cultural differences. Cue importance was measured via self-report and may not fully capture implicit cue usage. Future work should expand the avatar set and systematically manipulate visual and delivery-related cues (e.g., appearance, voice, speaking style/speed), ideally combining ratings with behavioral measures (e.g., eye-tracking) to clarify how cues are used in authority judgments across countries.

CONCLUSION

This study examined cross-cultural evaluations of avatars used for emergency announcements, comparing participants from China (*CN*) and Japan (*JP*) using comprehension checks, impression ratings (*Tw, Fa, Au*), and cue-importance judgments. Overall comprehension was high, and cross-national differences were most consistent for perceived Authority, whereas differences in Trustworthiness and Affability were comparatively weaker. Practically, avatar design for serious communication should prioritize culturally robust authority signaling and consider how users weight visual and delivery cues when forming impressions. Future work should expand the stimulus set and systematically manipulate key cues (e.g., appearance and voice) to establish more generalizable design guidelines. For deployment, authority sensitive cues should be validated separately in each target country, and designers should avoid assuming that a single 'professional-looking' avatar will generalize across cultures.

ACKNOWLEDGMENT

This research was partially supported by the JST-Mirai Program (Japan Science and Technology Agency; PI: Prof. Mihoko Niitsuma, 2022–2024) and by the Joint Research Fund of the Institute of Science and Engineering, Chuo University (PI: Prof. Toshikazu Kato, 2022–2024). The authors also thank Prof. Patrick Rau at Tsinghua University and Mr. Liu Yankuan for their assistance in recruiting Chinese participants and collecting the questionnaire data.

REFERENCES

Blackhurst T, Warmelink L, Roestorf A, Hartley C. The Brunswik Lens Model: A theoretical framework for advancing understanding of deceptive communication in autism. Front Psychol. 2024 Jul 11; 15:1388726. doi: 10.3389/fpsyg.2024.1388726. PMID: 39055993; PMCID: PMC11271661.

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., and Parasuraman, R., "A meta-analysis of factors affecting trust in human-robot interaction," Human Factors, 53(5), 517–527, 2011.

Kazemi, V., and Sullivan, J., "One millisecond face alignment with an ensemble of regression trees," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

King, D. E., "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, 10, 1755–1758, 2009.

Ohanian, R., "Construction and validation of a scale to measure celebrity endorsers' perceived expertise, trustworthiness, and attractiveness," *Journal of Advertising*, 19(3), 39–52, 1990.

Oosterhof, N. N., and Todorov, A., "The functional basis of face evaluation," *Proceedings of the National Academy of Sciences of the United States of America,* 105(32)*,* 11087–11092, 2008.

Stieglitz, S., Hofeditz, L., Brünker, F., Ehnis, C., Mirbabaie, M., and Ross, B., "Design principles for conversational agents to support Emergency Management Agencies," International Journal of Information Management, 63, 102469, 2022. https://doi.org/10.1016/j.ijinfomgt.2021.102469

ter Stal, S., Broekhuis, M., van Velsen, L., Hermens, H., and Tabak, M., "Embodied Conversational Agent Appearance for Health Assessment of Older Adults: Explorative Study," *JMIR Human Factors*, 7(3), e19987, 2020.

Yuki, M., Maddux, W. W., and Masuda, T., "Are the windows to the soul the same in the East and West? Cultural differences in using the eyes and mouth as cues to recognize emotions in Japan and the United States," *Journal of Experimental Social Psychology*, 43(2), 303–311, 2007.