Supplementary Online Material (SOM)

Content

1. Pretest for selecting TV commercial stimuli

An online pretest was conducted for selecting suitable TV commercials, targeted to elicit different emotional appraisal. A set of 71 commercials was preselected by the experimenters that were supposed to represent a wide range of possible emotional responses to TV ads.

A sample of n=356 respondents, with an average age of 42 years (ranging from 16 to 74, SD=13.1) and 170 female respondents (48%), were recruited from the German GfK Online Panel and completed an online questionnaire of on average 15 minutes duration. Each respondent was exposed to a subset of 10 randomly selected TV commercials. Immediately after each commercial, respondents assessed the commercial with respect to four appraisal dimensions using 21-point bipolar scales: predictability (related to novelty in the sense of degree of expectedness), valence (intrinsic pleasantness and/or goal/value compatibility), confusion (in the sense of lack of understanding or control of what is happening), and funniness (as a control factor). Overall, each commercial was rated by at least 42 and at most 53 respondents.

Means and standard deviations for the four dimensions were calculated. We ranked the spots based on their means and divided them into six intensity categories based on their percentile (10^{th}, 25^{th}, 50^{th}, 75^{th}, and 90^{th}). Then, the final set of 14 TV commercials has been selected in which (a) the six intensity categories for each dimension were approximately equally distributed, with at least two TV commercials representing the two extremes on each appraisal dimension, and (b) pair-wise correlations between the four appraisal dimensions were minimized. Table S1 summarizes the average ratings of each selected TV commercial for all appraisal dimensions.

	Predictability (1=extremely predictable to 21=extremely unpredictable)	**Valence** (1=extremely unpleasant to 21=extremely pleasant)	**Confusion** (1=extremely confusing to 21=extremely clear)	**Funniness** (1=extremely serious to 21=extremely funny)
1. Commercial in matter-of-fact tone for probiotic yoghurt	5.57	11.26	16.17	9.09
2. Commercial for body care brand showing happy family life and care scenes	4.50	16.62	18.58	10.72
3. Feel-good commercial for a caffeinated soft drink showing several examples that there is good in the world	15.50	16.02	14.98	12.86
4. Cat litter commercial with appearance of a cute kitten	8.64	16.88	17.46	12.60
5. Commercial for kitchen wipes consisting of a series of video clips showing funny mishaps	7.69	15.88	17.47	17.02
6. Car commercial with funny punchline involving three famous sports celebrities	12.36	17.64	17.64	18.32
7. Supermarket chain commercial with funny punchline	10.33	17.73	18.39	17.45
8. Toothpaste commercial depicting gum bleeding and tooth loss	8.00	10.02	16.78	8.80
9. Campaign ad to raise awareness for climate change revealing a drowned polar bear that initially looked like an iceberg	18.32	11.98	13.20	4.86
10. Anti-smoking commercial, woman with gruesome mouth cancer speaks out against smoking, in English language	11.94	2.39	11.37	2.41
11. Campaign ad against sugared soft-drinks, depicting a large amount of fatty tissue being poured from can into a glass and drunk, and spatting over a plate, in English language	14.56	6.24	9.18	9.76
12. Commercial for caffeinated drink with very sudden appearance of a screaming monster	18.92	6.42	8.10	11.68
13. Low-suspense commercial for Chinese cellphone service provider, in Chinese language	14.64	8.16	5.28	10.70
14. Japanese commercial for milk, with people moving in bizarre ways, in Japanese language	17.70	10.52	5.90	15.38

Table S1: Average ratings of the final selection of TV commercials per appraisal dimension.

2. Advertising test

The advertising test itself was conducted in a test studio in Munich. Participants were recruited from the local subject pool of the test studio provider. All participants were native German speakers or spoke German fluently at a level of native speakers. In total, 240 participants completed the study, five of which withheld permission to use their data for analysis, so both their questionnaire answers and their recordings were deleted. The remaining 235 participants were on average 43 years old, with a range from 18 to 78 (SD = 14.7), and 115 were female (49%). The interviews took 45 minutes on average, and participants received a show-up fee of 10 €.

Data collection setup

Participants were invited to the test location in Munich and seated in front of a computer screen in individual test rooms together with an interviewer. A full HD webcam was positioned on top of the screen, speakers were placed left and right of the screen. For optimal lighting conditions, soft-light lamps were placed left and right of the screen, and curtains in front of the windows prevented direct sunlight and shadows on participants' faces (see Figure S1).

Figure S1: Data collection set up for the studio test.

The interviewer adjusted the position of the webcam so that participants' faces were in the middle of the webcam picture. Participants were instructed that webcam recordings of their faces will be made. However, they were initially not informed about the true purpose of those recordings. Instead, they were told that the recording would be used to analyze eye movements. This was done to reduce expectancy effects and potential attempts to consciously control one's facial expressions. Participants were asked to turn their phones to silent mode, not to eat, chew gum or talk, and at least during recordings not to drink anything and let hands rest on the table. This was done to minimize facial movements unrelated to emotional response to the advertising stimuli, and to avoid covering of the face with one's hands.

Although conducted as a central location test, the study was designed as an online interview. The interviewer typed in the respondent ID and filled out initial questions about whether the participant wore glasses or a beard, and then handed over to the participant for self-administration. The interviewer remained in the room for questions or technical difficulties but was seated at a separate table with the back to the participants to give them some privacy and not to interfere during the recordings.

Data collection procedure

The questionnaire started with questions about demographics, lifestyle and attitudes towards advertising. Then participants were shown the first of overall 17 TV commercials. Each commercial was preceded by two seconds of black screen for potential post hoc baseline corrections and followed by three seconds of black screen to capture potential after effects. During exposure to each TV commercial, the webcam recorded the face of the respondent synchronously with the ad including the black screens. Then they were asked to provide information about their emotional response to the commercial. For that, participants were assigned to one of two conditions, one-fourth of participants to the emotion judgment condition, and three-fourths to the appraisal judgment condition. In the emotion judgment condition, participants were asked to indicate their emotions using the Geneva Emotion Wheel (GEW; see Scherer, 2005; Scherer, Shuman, Fontaine & Soriano, 2013), by selecting the emotions (maximally 2) they felt during the commercial (out of 20 emotion terms plus the two options "other" and "none") and rate their intensity. In the appraisal judgement condition, participants were asked to rate their emotional appraisals for the commercial on five dimensions[1] and to evaluate to what extent they were impressed by the commercial and how much they liked it from not at all to very much. The first spot was a positive, pleasant, albeit not particularly funny TV commercial and was considered a practice trial. Participants were invited after the ratings to ask the interviewer questions if anything was unclear. The next two spots were considered dummy spots, used to familiarize the respondents with the recording situation; they were always shown on second and third position. The remaining 14 spots were shown in random order. They represent the core set of stimuli that were selected based on the pretest results.

After all TV commercial had been shown for a first time, half of the commercials from the core set was randomly selected for second exposure that was followed by detailed scene-to-scene ratings. Participants were shown, on one screen, all scenes from the TV commercial they had just seen (between 10 and 18 scenes, depending on TV commercial). In the emotion judgment condition, participants were asked for each emotion they had indicated after first exposure, to select the scenes during which they felt that emotion, and subsequently, to select the one scene during which the intensity of that emotion was strongest. In the appraisal judgement condition, participants were asked to now rate each scene on one randomly selected core appraisal dimensions (Predictability / Novelty, Pleasantness / Valence, or Control / Confusion). Before the seven TV commercials selected for second exposure, the practice trial spot was also shown for a second time to get familiar with the new type of questions.

After the exposure of the commercials, participants filled out the Berkeley Expressivity Questionnaire (Gross & John, 1997) and then reached the final screen of the questionnaire, informing them about the purpose to the face recordings namely to analyze facial expressions. It was also explained why they were not informed about the true purpose in advance and that they could now withdraw permission to use their recordings. In that case, they were supposed to tell this to the interviewer. In case that the permission was denied, both face recordings and questionnaire data of that respondent were deleted from the dataset.

The study procedure fully adheres to the ICC/ESOMAR International Code on Market, Opinion and Social Research and Data Analytics (ESOMAR, 2016) that represents the global industry standard for self-regulation, undersigned by all ESOMAR member companies.

[1] The core appraisal dimensions were Novelty / Predictability, from very predictable to very unpredictable, Pleasantness, from very unpleasant to very pleasant, and Control / Confusion, from very confusing to very clear, on 11-point scales from -5 to +5. In the same way, two additional dimensions were rated, funniness from very serious to very funny, and moral acceptance from very offensive to very acceptable.

3. Selection of recording sections for FACS coding

Overall, the video material collected in Study 2 amounted to around 37.5 hours. As FACS coding is very time-consuming and effortful, we focused on sections in the recordings where changes in facial expressions could actually be observed. To identify such sections, a two-step selection procedure was applied:

Step 1: Automatic preselection of sections based on the following criteria:

(a) For each commercial, scenes were identified that received extreme ratings by the respondents (that is, -4 or lower or +4 or higher on the respective appraisal dimension) in the scene-to-scene ratings. The relevant section in the face recording was defined as the period from 2 seconds before and 6 seconds after the timestamp of a scene that received peak ratings. The procedure resulted in a set of 781 snippets.

(b) In addition, we identified key scenes in the TV commercials that we considered most likely to evoke a discernible facial expression in viewers. Overall, 20 key scenes were identified (e.g., sudden appearance of a monster on the screen, person in ad starting to speak Chinese, cute kitten shown for the first time in ad, etc.). Again, the relevant section in the face recording was defined as the period from approximately 2 seconds before and 6 seconds after the key scene – times could vary to avoid cutting sections at strange points in time. Thus, start and end times were adjusted to the beginning and end of scenes in the TV commercials. Minimum duration of the section was 6 seconds, maximum duration 10 seconds. This procedure resulted in the definition 4621 snippets, of which 1000 were randomly selected.

Step 2: Manual selection of the final snippet set for coding based on the following procedure:

In a second step, the 781 snippets were assigned randomly to three psychology graduate students who watched them and selected those in which a facial movement could be observed. Movement was defined as the presence of at least one AU with a discernible apex (that is, facial expression had lower intensity, or no expression was visible, before and after) that could also be at low intensity. Snippets with head movements alone were not eligible, neither were snippets with blinks only or expressions while talking. Overall, 408 snippets with discernable movements with apex were selected for FACS coding and observer appraisal rating. In these snippets, 155 different participants from the sample are depicted.

4. FACS coding

The 408 snippets were distributed among five certified FACS coders, based on an agreement on contract data processing obliging the coders to strict adherence to data protection laws, including strict confidentiality, access control to the computers on which the coding was done, and elimination of all copies of the recordings after coding was completed.

In a previous study (Seuss et al., 2019), inter-coder reliabilities for these coders had been determined at between 0.72 and 0.76. Each coder received a different subset that contained between 60 and 114 snippets. Coders received a basic payment per hour plus a variable hourly bonus that was dependent on professional experience and the previously determined reliability scores. Overall, it took 146.5 hours to code the 408 snippets.

Coding instructions followed the FACS manual (Ekman, Friesen & Hager, 2002). All facial AUs were dynamically coded. Each AU was coded in three phases: Onset phase starting with the frame where the first appearance change associated with the AU is observed, apex phase beginning at the frame where all appearance changes have reached a plateau or peak where no further increase is noticed,

and offset phase starting at the frame where the first evidence of a decrease in intensity is observed until disappearance of the AU or a new onset.

Additionally, intensity was scored at apex. To increase reliability between coders, three levels of intensity were used instead of five. These levels are A (small action, corresponding to a and b in FACS manual), B (moderate to strong action, corresponding to c in FACS manual), and C (estimated maximum action, corresponding to d and e in FACS manual). Asymmetry is also scored at apex only. Unilateral AUs occurring only at one side of the face are scored with the letter L (left side) or R (right side). Asymmetric AUs occurring on both sides but in different intensities are scored with the letters L & R plus a number indicating degree of asymmetry: 1 = small, 2 = medium, 3 = strong. The video annotation tool used for coding was ANVIL (Kipp, 2001).

5. Observer appraisal rating procedure

The 408 video snippets selected from the face recordings of participants viewing TV commercials were shown to 12 psychology students for appraisal rating. For data protection reasons, the students were invited to come to one of the authors' offices and rate the videos on computers provided there. Additionally, students signed confidentiality agreements by which they confirmed not to share any information about the videos they were exposed to, to ensure the privacy of the recorded respondents. After a short introductory training session, the computers in the author's offices remained available for a period of four days so students could rate the snippets at their own speed.

The ratings were done with a proprietary computerized annotation software, illustrated by a screenshot in Figure S2.

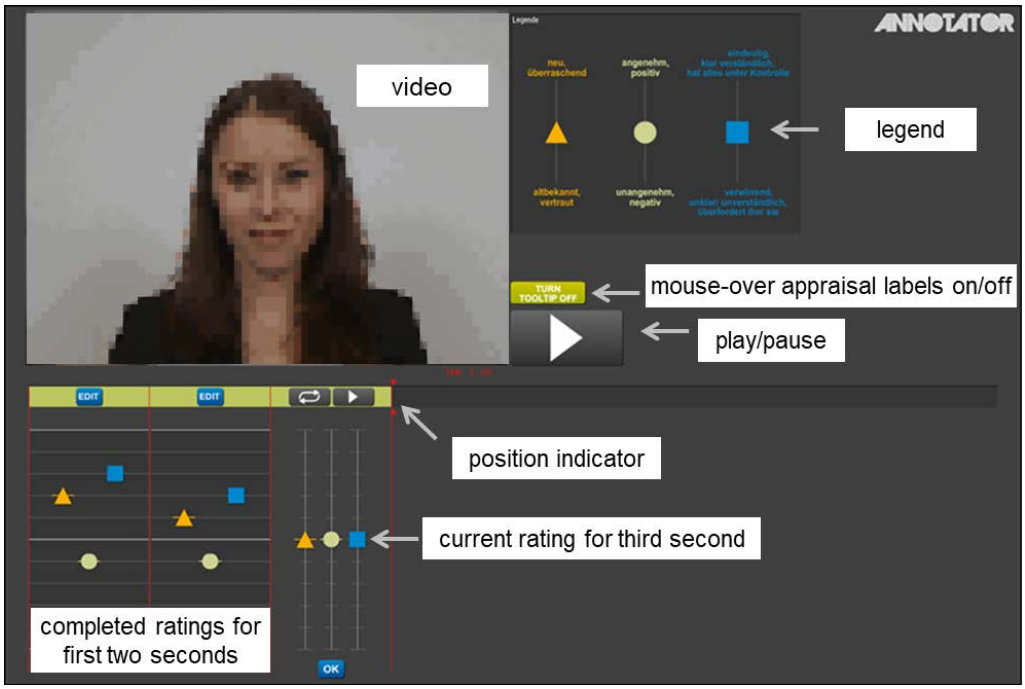

Figure S2: Screenshot of software used for second-by-second observer ratings of perceived appraisals. Please note: For this publication, the person's face has been pixelated for GDPR reasons.

Snippets were presented in individually randomized orders. Each snippet first had to be watched once from start to end. The students could replay the snippet as often as they wanted. After that, they played it second by second. For each second, they were asked to judge the experience of the respondent from her or his face by providing ratings for the three appraisal dimensions Novelty,

Pleasantness, and Control on bipolar 11-point scales. Each second could also be replayed if needed. Previous ratings could be revised.

Students received a flat payment of 13 € per hour, plus a variable bonus of up to 4 € that was derived from the correlation of their ratings with the average of the other students. The bonuses ranged from 2.43 € to 3.12 €.

6. Reliability of appraisal ratings

Number of ratings and stability of the mean rating

To evaluate whether 12 raters are sufficient to produce stable average ratings we applied a bootstrap approach. To this end, for each second and appraisal dimension the real average rating was computed and compared to a bootstrapped mean rating. The following procedure was used:

- From the individual ratings for each second, a random sample (with replacement) of the same length was drawn. For each rated second, the average of the randomly drawn ratings was computed (i.e., the bootstrapped mean).
- This was repeated R=5000 times (i.e., for each second 5000 samples are drawn and for each of the 5000 samples a mean is computed).
- Using the 5000 means per second, the 95% confidence interval was computed per second.
- It was checked if the real mean is within the confidence interval.
- Using t-tests (two-sided, paired) it was checked if the average of the bootstrapped means is significantly different from the real mean.

This Bootstrap analysis showed that for all rated seconds the real average ratings are within the bootstrapped confidence interval. In addition, the t-tests show that bootstrapped means did not significantly differ from the real means for the vast majority of the observations. Only in very few cases (4% for Valence, 3 % for Control, 2% for Novelty), the means differ significantly. For those few seconds where the bootstrapped means and the real means differ significantly, the (absolute) difference between the real and the bootstrapped mean are on average extremely low:

Valence:	0.0018
Control:	0.0017
Novelty:	0.0016

Inter-rater reliability

To evaluate inter-rater reliability, we computed several measures proposed in the literature.

	Valence	Control	Novelty
Average of correlations between one rater's rating and the average of all other raters' ratings	.85	.65	.71
Average of all pairwise Spearman correlations (Spearman, 1904)	.72	.45	.46
Intraclass correlation coefficient (ICC, Shrout & Fleiss, 1979)	.73	.43	.51
Kendall's W (Siegel & Castellan, 1988)	.74	.49	.49
Average of pairwise weighted Cohen's Kappa (Cohen, 1968)	.64	.37	.44
Fleiss' Kappa (Fleiss, 1971) for three categories (negative/neutral/positive)	.55	.38	.41
Krippendorff's Alpha (Krippendorff, 1970)	.71	.42	.43

Table S2: Overview of different inter-rater reliability measures.

Given the structure of the data, the correlations between ratings (correlation of individual raters with average of others and pairwise Spearman), the ICC, Kendall's W as well as the average weighted Cohen's Kappa are most suitable, all indicating similar inter-rater reliability.

7. Data cleaning and preparations

Appraisal ratings are available on a second-by-second basis, whereas AU data (both from automatic detection and manual FACS coding) exist for every video frame in 25 Hz resolution. To align the two data sources, AU data are averaged to 1 Hz. Thus, the dataset contains, for each second, mean appraisal ratings from the students as criterion, and AU data averaged across frames for each second as predictors.

The AU detection algorithm is designed to produce AU intensities in the interval [0,1] using soft constraints introduced during the state estimation process. The modeling of these soft constraints is detailed in Hassan et al., 2018. In some rare cases, however, it can happen that the AU intensities are slightly below zero or above one. For consistency, we left-censored the data at zero and right-censored at 1.

The aim of the appraisal inference is to predict appraisals from AU data. On the one hand, however, not only facial expressions but also other behaviors visible to the raters such as posture changes can influence their appraisal ratings. On the other hand, not every AU expresses an emotion but can represent, for instance, facial tics, habits or blinks. To this end, all observations have been removed where (a) FACS coders did not detect an AU but appraisals ratings made by the judges are different from zero, and (b) AUs have been detected by the FACS coders but appraisal ratings made by the judges are not different from zero. Keeping these observations would add only noise to either the target variable or the predictors.

To reduce the complexity of the prediction problems, we transformed Control and Novelty into unipolar prediction problems. As, relatively speaking, fewer cases of positive Control ratings (i.e., 13.4%) and negative Novelty ratings (i.e., 18.5%) were observed, we decided to focus on the more frequent ratings of negative Control – that is, to what extent a person is confused or unable to cope

with the situation (i.e., lack of Control) – and positive Novelty – that is, to what extent a person experiences the situation as novel. To this end, we censored the data and set all positive ratings for Control and negative ratings for Novelty to zero.

In addition, we transformed Valence ratings from a bipolar (positive – neutral – negative) into two separate unipolar variables (neutral to positive and neutral to negative) and created a variable for positive Valence by setting all negative rating values to zero, and a variable for negative Valence by setting all positive rating values to zero.

The current system is intended as an extension of an earlier development of the research group, a system to directly infer emotional Valence from appearance changes in the face (Garbas, Ruf, Unfried & Dieckmann, 2013). This Valence detection system infers positive and negative Valence separately. To this end, all snippets have been analyzed frame by frame by the Valence detection system and the two scores for positive Valence and negative Valence are added to the dataset as predictors, also averaged across video frames for each second.

For calibration, all AU intensities delivered by the AU detection module are individually calibrated to control for individual habits and characteristics. This is done by AU-wise subtraction of the lowest AU intensity measured after the first half second from each AU intensity.

8. Comparison of methods for appraisal inference

Based on 1Hz raw data, i.e. without

"[…]

- Truncation of outliers in AU intensity estimates data to the interval [0,1].
- Removal of observations with non-AU-related appraisal assessments as well as observations with non-appraisal related AUs, such as facial tics.
- Transformation of all appraisal data into two-class data, i.e., transforming the Valence data into two two-class problems, positive Valence vs. non-positive Valence, and negative Valence vs. non-negative Valence; for the remaining appraisal dimensions, we focused on identifying signs of Novelty (vs. non-Novelty), and signs of Confusion/negative Control (vs. non-negative Control).

Several methods for appraisal inference have been compared using GNU-R:

- OLS Regression, full model
- OLS Regression, nested model using stepwise variable selection
- Random Forrest (using randomForest function in the randomForest package)
- Regression Tree (using the rpart function in the rpart package)
- Support Vector Regressions (RBF-kernel, optimized gamma (10^{-6} – 10^{-2}) and C (10^{0} – 10^{2}) using svm function in the e1071)
- Single-hidden-layer neural network (optimized layer size (0 – 10) and decay (0.01, 0.1, 0.5, or 1) using nnet function in the nnet package)
- Multi-layer perceptron (two layers which size 12 and 6 respectively. More layers and different sizes didn't change the results significantly, using mlp function from RSNNS package)
- MARS, pruning method = cross validation (5 folds, 10 repetitions, using earth function from earth package)

Input variables:

- Positive and negative classifier from SHORE
- Detected AU intensities

Criterion for comparison is Pearson correlation coefficient.

Results

	OLS	OLS stepw.	RF	Tree	SVR	nnet	MLP	MARS
control	.42	.42	.46	.45	.42	.45	.25	.36
valence	.73	.73	.75	.69	.64	.74	.7	.65
novelty	.26	.26	.27	.16	.37	.27	.11	.11

Summary

- Random Forrest outperforms OLS regression marginally for all three appraisal dimensions
- Regression Tree outperforms OLS regression marginally for control
- SVM outperforms OLS regression for Novelty
- Neural Net outperforms OLS regression marginally for all three appraisal dimensions

Since performance gains (correlation for RF predictions and ground truth is only 0.01 – 0.04 higher than for OLS predictions) are only marginal, we opted for a OLS regression since a OLS regression is more efficient (w.r.t runtime), easier to apply to new data, and much more transparent than the other methods as regression coefficients (and their signs) are directly interpretable.

The only exception is the control dimension inferred by SVR where an increase in correlation of 0.11 could be achieved. Nevertheless, for consistency of an interpretable approach we chose OLS regression also for the novelty dimension.

9. Description of the best models

The following tables contain the AUs and the corresponding coefficients from OLS estimation of the optimal model for each appraisal dimension. Since the models were derived in a 10-fold cross validation resulting in 10 different vectors of coefficients for each appraisal dimension, the coefficients reported below have been averaged across the folds.

Regression weights for lack of control, confusion	
Intercept	.018
pos.shore	-.0004
neg.shore	.0004
AU01_InnerBrowRaiser	.0525
AU02_OuterBrowRaiser	.1924
AU04_BrowLowerer	.4144
AU06_CheekRaiser	-.0555
AU07_LidTightener	.1504
AU09_NoseWrinkler	-.4083
AU10_UpperLipRaiser	.1007
AU13_SharpLipPuller	.0804
AU14_Dimpler	-.0316
AU15_LipCornerDepressor	.0757
AU16_LowerLipDepressor	.1288
AU20_LipStretcher	-.1132
AU23_LipTightener	-.0623
AU24_LipPresser	-.6012
AU25_LipsPart	-.0555
AU43_EyesClosed	.0528

Regression weights for novelty	
Intercept	.0121
pos.shore	-.0003
AU01_InnerBrowRaiser	.4242
AU02_OuterBrowRaiser	.3792
AU07_LidTightener	-.1425
AU10_UpperLipRaiser	.0857
AU11_NasolabialDeepener	.0696
AU16_LowerLipDepressor	.0911
AU24_LipPresser	.356
AU25_LipsPart	.1069
AU26_JawDrop	-.2587
AU43_EyesClosed	-.0447

Regression weights for positive valence	
Intercept	.0649
pos.shore	.0015
neg.shore	-.0001
AU01_InnerBrowRaiser	-.0947
AU02_OuterBrowRaiser	-.0942
AU04_BrowLowerer	-.0384
AU05_UpperLidRaiser	.1451
AU06_CheekRaiser	.1522
AU07_LidTightener	-.1303
AU12_LipCornerPuller	.0571
AU13_SharpLipPuller	.3838
AU14_Dimpler	.1065
AU15_LipCornerDepressor	-.0625
AU16_LowerLipDepressor	-.4074
AU20_LipStretcher	.1737
AU25_LipsPart	.1877
AU26_JawDrop	.3712
AU27_MouthStretch	.2766
AU43_EyesClosed	.1344

Regression weights for negative valence	
Intercept	.0249
pos.shore	-.0002
neg.shore	.0007
AU01_InnerBrowRaiser	.1553
AU04_BrowLowerer	.2267
AU06_CheekRaiser	-.049
AU07_LidTightener	.364
AU09_NoseWrinkler	-.4251
AU10_UpperLipRaiser	.0878
AU12_LipCornerPuller	-.0345
AU13_SharpLipPuller	.0576
AU14_Dimpler	-.0463
AU15_LipCornerDepressor	.1297
AU16_LowerLipDepressor	.238
AU17_ChinRaiser	.0835
AU23_LipTightener	-.0627
AU24_LipPresser	-.5234
AU25_LipsPart	-.1017
AU26_JawDrop	-.1459
AU27_MouthStretch	.1198
AU43_EyesClosed	.0397

10. References

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*, 213–220.

Ekman, P., Friesen, W. V., & Hager, J. C. (2002). *The Facial Action Coding System* (2nd ed.). Salt Lake City, UT: Research Nexus eBook.

ESOMAR. (2016). *ICC/ESOMAR International Code on Market, Opinion and Social Research and Data Analytics*. Retrieved from https://www.esomar.org/what-we-do/code-guidelines.

Fleiss, J. L. (1971) "Measuring nominal scale agreement among many raters." *Psychological Bulletin*, Vol. 76, No. 5 pp. 378–382

Garbas, J. U., Ruf, T., Unfried, M., & Dieckmann, A. (2013). Towards robust real-time valence recognition from facial expressions for market research applications. *Proceedings of the Humaine Association Conference on Affective Computing and Intelligent Interaction*, pp. 570-575.

Gross, J. J., & John, O. P. (1997). Revealing feelings: Facets of emotional expressivity in self-reports, peer ratings, and behavior. *Journal of Personality and Social Psychology, 72*, 435-448.

Hassan, T., Seuss, D., Ernst, A., Garbas, J. (2018). A Kalman filter with state constraints for model-based dynamic facial action unit estimation. *Forum Bildverarbeitung 2018*.

Kipp, M. (2001) Anvil - A generic annotation tool for multimodal dialogue. *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech)*, pp. 1367-1370.

Krippendorff, Klaus (1970). Estimating the reliability, systematic error, and random error of interval data. *Educational and Psychological Measurement, 30* (1), 61–70.

Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social Science Information, 44(4)*, 693-727.

Scherer, K. R., Shuman, V., Fontaine, J. R. J, & Soriano, C. (2013). The GRID meets the Wheel: Assessing emotional feeling via self-report. In J. R. J. Fontaine, K. R. Scherer & C. Soriano (Eds.), *Components of Emotional Meaning: A sourcebook* (pp. 281-298). Oxford: Oxford University Press.

Seuss, D., Dieckmann, A., Hassan, T., Garbas, J.U., Ellgring, J.H., Mortillaro, M., & Scherer, K. (2019). Emotion expression from different angles: A video database for facial expressions of actors shot by a camera array. *Proceedings of the 8th International Conference on Affective Computing and Intelligent Interaction (ACII 2019),* pp. 35-41. Cambridge, United Kingdom.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlation: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420–428.

Siegel, S., & Castellan, N. J., Jr. (1988). Nonparametric statistics for the behavioral sciences (2nd ed.). New York, NY: McGraw-Hill.

Spearman C. (1904). "The proof and measurement of association between two things". American Journal of Psychology. 15 (1): 72–101