Emotion Prediction with Weighted Appraisal Models – Validating a Psychological Theory of Affect

Laura S. F. Israel and Felix D. Schönbrodt

***Abstract*—Appraisal theories are a prominent approach for the explanation and prediction of emotions. According to these theories, the subjective perception of an emotion results from a series of specific event evaluations. To validate and extend one of the most known representatives of appraisal theory, the Component Process Model by Klaus Scherer, we implemented four computational appraisal models that predicted emotion labels based on prototype similarity calculations. Different weighting algorithms, mapping the models' input to a distinct emotion label, were integrated in the models. We evaluated the plausibility of the models' structure by assessing their predictive power and comparing their performance to a baseline model and a highly predictive machine learning algorithm. Model parameters were estimated from empirical data and validated out-of-sample. All models were notably better than the baseline model and able to explain part of the variance in the emotion labels. The preferred model, yielding a relatively high performance and stable parameter estimations, was able to predict a correct emotion label with an accuracy of 40.2% and a correct emotion family with an accuracy of 76.9%. The weighting algorithm of this favored model corresponds to the weighting complexity implied by the Component Process Model, but uses differing weighting parameters.**

***Index Terms*—Affective computing, appraisal theory, Component Process Model, emotion, predictive models.**

This is an edited manuscript accepted for publication in the Journal IEEE Transactions on Affective Computing. The manuscript will undergo copyediting, typesetting, and review of resulting proof before it is published in its final form.
Please cite this preprint as:
Israel, L. S. F., & Schönbrodt, F. D. (2019). Emotion Prediction with Weighted Appraisal Models—Validating a Psychological Theory of Affect. *IEEE Transactions on Affective Computing*, 1–1. https://doi.org/10.1109/TAFFC.2019.2940937

I. Introduction

Since the 1990s a variety of computational emotion models have been implemented, creating an interdisciplinary field between psychology and computer science. This development has not only been driven by its numerous new applications in artificial intelligence, robotics and human-computer interaction, but also by its contribution to basic emotion research [1]. Computational affect modelling provides a framework to test psychological emotion theories and elaborate their structure. Furthermore, mathematical implementations of cognitive models can help to consolidate and extend verbal theories that often lack formality and explicitness. In the present paper, we therefore used a computational emotion model to extend and validate one of the most prominent approaches for the explanation of affect – the appraisal theories of emotion (see [2] for an overview), specifically, the Component Process Model by Scherer [3].

As emotions are subject to many interdisciplinary fields of research, many differing conceptualizations of emotions can be found. Most theorists though recognize that emotions are multi-componential, integrating different elements such as somatic and motor functions, motivation, cognition and often feeling, the component describing the subjective emotional experience of a person [4]. How these components interact and which role they play in the causation of emotions is heavily debated. An early exploration of the emergence of affect by James [5] defines emotion as the perception of bodily changes that arises as response to the environment. This strict exclusion of the cognitive component in the emotion causation process has since been challenged. Schachter and Singer [6], for example, expanded James' [5] theory by proposing a two-step procedure, in which a stimulus generates an unspecific physical state of arousal but a second cognitive elaboration is needed to interpret the arousal state and label it correctly. Appraisal theories of emotion go even further, apprehending the cognitive evaluation of a stimulus as the trigger of emotions, influencing all of the other components (e.g. [3], [7]–[9]). Appraisal is generally understood as the process of assessing the relevance of a stimulus for one's own welfare regarding personal needs, values, attachments, beliefs and goals, though the presumed number and content of appraisal dimensions vary between theorists [2]. An emotion or emotion family can then be

This research was funded by a grant of the German Research Foundation to Felix Schönbrodt (DFG SCHO 1334/4-1).

described as a function of a distinct appraisal pattern – several of these appraisal profiles for specific emotions have been proposed in the literature [3], [9]–[11]. Consequently, an emotion is not supposed to be elicited by the stimulus itself (contrary to James' theory [5]), but by its meaning for the individual [12]. This holds significant explanatory power, as it can account for the fact that the same stimulus can evoke completely different emotional reactions between individuals or even within the same person on different occasions.

Despite the popularity of this cognitive approach to emotions and the strong commonalities between appraisal theories, there is some disagreement concerning the content of the appraisals and how they are mapped onto emotion categories [2]. Several empirical studies have been conducted to test the theoretical predictions made by appraisal theories (see [13] for a review), but as they were only able to systematically vary few appraisal dimensions at once, other methods need to be applied to further investigate these models as a whole. Here, computational emotion models, specifying which emotional reaction an individual will experience once a specific appraisal pattern is present, can help determine the plausibility of appraisal dimensions and the suspected mapping algorithms. In the past, several models were successfully implemented mapping appraisal profiles either onto distinct emotions labels (e.g. AR [14]) or dimensional representations of affect (e.g. WASABI [15]). Some of those adapted the appraisal profiles proposed by Scherer [3] (e.g. PEACTIDM [16]) others built on the work of Ortony, Clore and Collins [17] (e.g. AR [14]). Most of these models serve to create intelligent agents that act autonomously in simulated environments. To validate the underlying theory though, the model's behavior has to be contrasted with empirical data. The computational appraisal model, formalizing the junction between emotion and cognition, should be able to predict the emotional experience of an individual correctly, otherwise the model may be insufficient or inappropriate to describe the emotion formation process. Such an approach was first put into practice with the Geneva Expert System on Emotions (GENESE) by Scherer [18]. In this framework participants were asked to recall an emotional episode from their past and answer a questionnaire intended to measure 11 different appraisal dimensions. The expert system then calculates the similarity to theoretically derived appraisal patterns representing different prototypical emotions by Euclidean distance and makes guesses about the emotional state recalled by the participant. Subsequently, the predictions are validated by the participant as correctly or incorrectly describing the perceived emotion. In this experimental setup, the system was able to predict an appropriate emotion term in 77.9% of the cases. But the post hoc verification of the prediction might have demand characteristics and might have urged participants to accept an emotion label when they themselves had no clear judgment about their state. Consequently, a new system, the Geneva Emotion Analyst (GEA) [19], was introduced. GEA asks users to label the reported emotion episode before the systems diagnosis is made so that an exact match or mismatch can be determined. In 51% of the cases the first guess of the GEA system matched one of the emotion labels given by the participant. GEA also operates by calculating the distances between users' appraisal ratings and appraisal prototypes, but further incorporates a weighting algorithm that takes into account that some appraisal dimensions might be more important for emotion formation than others.

The described GEA and GENESE system proceed in a classical deductive manner – making predictions about the participants emotional state based strictly on theoretical assumptions. Through deductive reasoning we imply, that if our premises (i.e. our model assumptions) are true then our inferences (i.e. our predictions) must be necessarily true as well [20]. In this manner, the assumed structure of the model can be validated by its predictive accuracy. In the present paper, we want to extend this modelling idea with a more inductive approach. In inductive reasoning premises are based on statistical data such as observed frequencies of a specific feature in a sample. Therefore, every inference that is drawn goes beyond what is logically included in the premise [20]. This entails some uncertainty as not all inferences necessarily need to be valid, but it allows to generate new premises (i.e. model assumptions) that can be validated subsequently. As for the present study, we implemented four affect-derivation models based on the theory of Scherer's Component Process Model [3]. Similar to predecessor systems, all four models are able to predict an emotion term by calculating similarities between an appraisal profile and several emotion prototypes, but apply different kind of weighting algorithms in the appraisal-emotion mapping process. In contrast to earlier models, we also used empirical data to inductively elaborate the models by estimating the appraisal profiles of the emotion prototypes as well as the different appraisal weightings instead of using only theoretically derived parameters. We then validated and compared the models by evaluating their predictive out-of-sample performance. By integrating theory-based as well as data-driven information in computational emotion models and by systematically varying their internal structure (weighting), we hope to engage in the theory formation process and further the understanding of the appraisal-emotion mapping process.

II. The Component Process Model (CPM)

Scherer's [3] CPM, the theoretical basis of our models, considers emotions as an "episode of interrelated, synchronized changes in the state of all or most of the five subsystems in response to the evaluation of an external or internal stimulus event as relevant to major concerns of the organism" (p. 93). Each stimulus event is evaluated by a number of criteria, the so-called Stimulus Evaluation Checks (SECs). Scherer proposes 16 of such appraisal dimensions organized in four major classes that determine (1) the relevance of an event to the organism, (2) the implications of an event for personal goals and well-being, (3) the ability to cope and adjust to potential or real consequences of the event and (4) the importance of an event regarding self-concept or social norms (see [3] for a detailed description of the 16 appraisal dimensions). How each dimension is appraised is highly dependent on individual and situational aspects such as motivation, cultural imprint or social pressure. From the interaction of all 16 appraisal dimensions a virtually infinite emotion space arises. Scherer [3] therefore rejects the assumption of a limited number of discrete emotion categories made by many other emotion theorist (e.g. [21]).

Nonetheless, he recognizes that certain appraisal combinations occur more frequently and universally than others. Scherer [3] calls these states, that are usually labelled with a short verbal expression, *modal emotions*. For the 13 modal emotions pleasure, joy, pride, irritation, rage, contempt, disgust, guilt, shame, anxiety, fear, sadness and despair, he proposes theoretically derived appraisal patterns representing the prototypical level of each appraisal dimension for each modal emotion. These prototypes, adapted over the years [3], [19], [22], also include open parameters indicating that a specific dimension might be irrelevant for a modal emotion or that many values are compatible with the affect [23]. Overall, the theoretical prototypes show moderate correlations to appraisal means found in empirical data [19]. During the appraisal process the evaluated dimensions are integrated by a weighting function, that considers each of the 16 appraisal dimensions to be differently important in the affect-centered rating of a situation [23]. For this weighting algorithm, theoretically derived parameters have been proposed as well [19].

III. Extending the CPM

The described appraisal structure was adapted in our four models. The models predict an emotion label from the set of 13 modal emotions by calculating the distance between an empirical appraisal profile, containing ratings for the 16 appraisal dimensions, to 13 emotion prototypes within a 16-dimensional appraisal space. They then return the emotion label indicating which prototype shows the highest resemblance to the empirical vector. In each of the models though, we implemented a different weighting of the appraisal dimensions. As in the GENESE system, the first emotion model (M1) did not use a weighting – all appraisal dimensions were considered to be equally important in the emotion class determination. The second model (M2) and the third model (M3), similar to the GEA system, included 16 parameters, one for each appraisal dimension. This weighting algorithm implies that some appraisal dimensions could be generally more important in the identification of an emotion than others (e.g. the valence of a stimulus being more important than its familiarity), across all emotions. In the fourth model (M4), we implemented a separate weighting parameter for each of the 16 appraisal dimensions within each of the 13 emotion prototypes, resulting in 208 parameters. This more complex weighting allows each appraisal dimension to be differently relevant for each of the modal emotions. This means, for example, that for most emotions such as joy, anger or sadness it could be irrelevant who caused a situation, as all of these emotions can be triggered by one's own actions as well as by actions of others. But for emotions such as guilt or shame, that are more often elicited by one's own actions, the appraisal might be highly relevant. Support for this view also comes from empirical research. For different emotion classes, Ellsworth and Smith [9] identified differing subsets of appraisals, that were predictive for the specific emotion, implying that appraisals might be unequally important within different emotion classes. This assumption, although not expressed in the CPM, does not contradict Scherer's [3] model, as the open parameters he included in the theoretical prototypes can be understood in the same way: If an emotion prototype is compatible with several different levels of an appraisal dimension (as implied by an open parameter in Scherer's prototypes), then this dimension is not relevant for the specific emotion, as it cannot be used to differentiate this emotion from others. This should be reflected in a low weight of the appraisal dimension within the emotion prototype. If this assumption is correct, the more complex weighting algorithm should result in a better performance compared to the 16-dimensional or equal weighting scheme.

While M2 used the theoretically derived weighting parameters [19], parameters in M3 and M4 were estimated from empirical data. By comparing the predictive power of these four differently weighted models, we hope to evaluate if the weighting proposed by the CPM as well as the proposed weighting parameters are appropriate or whether a different kind of mapping algorithm yields a better predictive performance. Also, to evaluate the predictive performance of our models, we compared them to a naive classifier that randomly guesses classes weighted by their frequency in the data set (baseline model) as well as to a Random Forest machine learning model that should be able to yield a very high prediction performance by considering all potential interactions, presenting an upper level of performance that can be reached with the used data set.

As the theoretical prototype profiles show only moderate correlations to the ones found in empirical studies, it seems plausible that the 208 parameters can't be fully deduced from theoretical assumption about the appraisal process. We therefore decided to derive the prototypes directly from an empirical data set that was collected with the GEA system by Scherer and Meuleman [19]. Prototype theory, first introduced by Rosch in 1975, defines the prototype of a category as a reference point for classification based on representativeness [24]. As we describe each emotion category on 16 continuous dimensions (i.e. each dimension can be described by a distribution function), we can assess the most representative instance for each modal emotion by finding the mean of each appraisal dimension in a representative sample. This data-driven approach on a large data set should hence lead to a better prototype assessment and consequently to a better performance than an exclusively theoretical approach. The estimation of the appraisal weights (16 parameters for M3 and 16*13=208 parameters for M4), required a more complex estimation algorithm. We used a genetic optimization method to determine the weighting parameters that would maximize the models' predictive performance.

To summarize, we combine different modelling approaches to validate the CPM and expand its theoretical assumptions: (1) By contrasting our models' predictions with an empirical ground truth, we can assess their predictive power and consequently the plausibility of the underlying theory. If emotions arise from the cognitive evaluations of the 16 dimensions proposed by the CPM, our computational models should be able to predict the correct emotion labels to some degree. With the performance level attained, we can further investigate, whether the appraisal dimensions proposed by the CPM are sufficient to predict the subjective feeling (emotion label) of participants correctly. (2) The systematic variation of the weightings between the different models enables us to inspect whether the weighting algorithm implied by the CPM is

valid or whether different weighting parameters (generated from empirical data), a more complex or even no weighting at all yields a better performance.

IV. Method

In the following section, we describe in more detail the data set that was used in the present study as well as the mathematical implementation of the found appraisal models, the parameter calculation as well as evaluation of the model performance. All corresponding R scripts are provided in our electronic appendix.

A. Dataset

For the estimation of the model parameters as well as for the out-of-sample validation of the resulting models, a data set by Scherer and Meuleman [19] was used. The data was collected via the freely accessible GEA system on the website of the Swiss Center for Affective Sciences[1] over the duration of eight years. The questionnaire implemented in the GEA system is publicly available as the Geneva Appraisal Questionnaire (GAQ) [25] and was specifically developed to assess the results of an appraisal process during an emotional episode through memory and verbal report. In the online questionnaire, participants were asked to recall an emotional episode from their past. After describing the recalled situation, subjects were asked to name the perceived emotion by choosing one or two matching terms from a list of 13 emotions consisting of pleasure, joy, pride, irritation, rage, contempt, disgust, guilt, shame, anxiety, fear, sadness and despair. Participants could also indicate that none of the emotion terms described how they felt. Subsequently, a set of 25 questions was presented that was constructed to assess the appraisal dimensions of Scherer's CPM. Each item, measuring the presence of a specific appraisal during the emotional episode, was rated on a 5-point scale reaching from *not at all* to *extremely* or could be labelled as *not applicable* to the situation. Further information about contextual factors was collected as well, which is not relevant for the present study.

The dataset included 6809 reported emotional episodes. 218 of these observations had to be dismissed because participants did not report any specific emotion label and were therefore lacking a ground truth. The final sample (n = 6591) consisted of 4419 female and 2171 male raters (sex and age of one participant was missing). The majority of participants, about 59 % (n = 3900), were between 20 and 40 years. About 23 % (n = 1483) were in the age group between 12 and 20 years and around 18 % (n = 1207) were older than 40 years. As the questionnaire could be completed in three different languages, the dataset included 625 German, 3015 English and 2951 French speaking participants. 72% of the participants (n = 4720) selected two emotion labels to describe the reported episode, while only 28 % (n = 1871) identified the reported emotion using one single label.

[1] https://www.unige.ch/fapse/emotion/demo/TestAnalyst/GERG/apache/htdocs/index.php

B. Data Pre-processing

For the further use in our emotion models, we aggregated the 25 appraisal items to the 16 appraisal dimensions proposed by the CPM [3] by calculating means for the dimensions measured with more than one item. Additionally, we normalized the data to a range from 0 to 1. All *not applicable* answers were set to missing (about 12% of the dataset). As imputations of the missing cases would contradict the theoretical assumption that some appraisal dimensions might be completely irrelevant for certain emotions [23], missing values were kept in the dataset. Instead, we handled missing data in our emotion models by pairwise deletion. For all episodes with more than one emotion label, we randomized the order of the emotion terms, as it was not clear how the order was achieved within the GEA system. For the episodes labelled with only one emotion term the second emotion label was set to *Undetermined*.

For the out-of-sample validation of the emotion models the dataset was split into two subsets by stratified sampling (using the *stratified* function from the *splitstackshape* [26]). Using the first emotion labels as strata, a training set holding 50% of the data (n = 3296) and a test set holding the other half (n = 3295) were created. As the emotion categories in the training set (as well as is the whole data set) were rather unbalanced, with some emotions (such as contempt or disgust) being underrepresented, we used an oversampling algorithm to create an additional balanced training set to use in the optimization of the model parameters. This is a crucial step as unbalanced datasets in supervised classification tasks can lead to the overpowering of prevalent classes and ignorance of rare ones [27]. The oversampling as well as all further analyses and implementations were conducted in R (Version 3.4.2) [28]. Using again the first emotion label as class label, we randomly sampled instances from the data set with the *upSample* function from the *caret* package [29], so that all emotion categories would have the same frequency as the largest class in the data set. The resulting oversampled training set consisted of 8034 instances, 618 for each emotion category.

C. Model Implementations

To make predictions about an emotional state, the models (M1, M2, M3 and M4) take an input vector containing the numerical ratings of the 16 appraisal dimensions for that specific state. By calculating the sum of squared differences, the distance between this input vector and 13 emotion prototypes, which represent the mean level of an appraisal dimension within a specific emotion category in the original (unbalanced) training set, is determined. Appraisal dimensions that are missing in the input vector are not considered in the distance calculation. This means that dimensions marked as irrelevant or not applicable by the participant are excluded. While M1 does not include a weighting, M2 and M3 weighted each of the 16 appraisal dimension separately. Thus, giving the dimensions different importance during the distance calculation. In M4, each of the appraisal dimensions within each emotion category is weighted differently. Each weight therefore represents the appraisal

dimensions relative importance within a specific emotion category. Consequently, each of the 13 resulting distance scores in M4 is obtained with a different weighting algorithm, leading to different maximum distances. To compare the scores, each value is normalized to a range between 0 and 1. To obtain a consistent metric for all four models, score normalization was also implemented in the other two models. The normalized distances are subsequently reversed to similarity scores (s_i). Hence, larger values indicate a higher similarity to a prototype. The similarity metrics of the four models are calculated by the following formulas:

$$M1: s_i = 1 - \frac{\sum_{j \notin Q} (p_{ij} - e_j)^2}{\sum_{j \notin Q} 1} \quad (1)$$

$$M2, M3: s_i = 1 - \frac{\sum_{j \notin Q} (w_j (p_{ij} - e_j))^2}{\sum_{j \notin Q} w_j^2} \quad (2)$$

$$M4: s_i = 1 - \frac{\sum_{j \notin Q} (w_{ij} (p_{ij} - e_j))^2}{\sum_{j \notin Q} w_{ij}^2} \quad (3)$$

where
s_i is the similarity to the i[th] emotion prototype,
p_{ij} is the prototype value of the j[th] appraisal dimension of the i[th] emotion prototype,
e_j is the empirical value of the j[th] appraisal dimension,
w_j is the appraisal weight given to the j[th] appraisal dimension,
w_{ij} is the appraisal weight given to the j[th] appraisal dimension of the i[th] emotion prototype,
Q is the set holding the indices of missing values in the empirical vector.

Based on the resulting similarities (s_i) the models make a prediction, returning the emotion with the highest resemblance to the input vector (i.e. the smallest normalized distance between input and prototype). By comparing the models' predictions with the actual emotion labels, the classification performance can be obtained to evaluate their predictive power.

D. Estimation of Model Parameters

1) Emotion prototypes

The emotion prototypes (p_{ij}) used in all four models were calculated from the empirical data contained in the (unbalanced) training set. For each emotion prototype, consisting of 16 prototypical appraisal values, episodes labelled with the according emotion term were aggregated. Episodes labelled with two emotion terms were included in the prototype calculations of both emotion categories. For each of the 13 emotions, the mean level of each of the 16 appraisal dimensions was calculated over all episodes labelled with the respective emotion category – resulting in a 13 x 16 prototype appraisal matrix. Each prototype within this matrix was calculated by the following formula on the unbalanced training set:

$$p_{ij} = \frac{\sum_{k=1}^{n_i} r_{ijk}}{n_i}$$

where
p_{ij} is the prototype value of the j[th] appraisal dimension of the i[th] emotion prototype,
r_{ijk} is the k[th] rating of the j[th] appraisal dimension that was labelled with the i[th] emotion class,
n_i is the number of episodes labeled with the i[th] emotion class.

The number of observations included in the prototype calculation ranged from $n = 81$ (Contempt) to $n = 992$ (Sadness), where cases with two labels counted for both prototypes. To assess the resemblance between the newly calculated prototypes and the theory, we calculated Pearson correlations between the 13 empirical assessed prototypes and the theoretical prototypes proposed by Scherer [7] (Table 5.4.). The latter are reported as categorical variables and were translated to continuous values for this purpose. Also, a mean correlation across all prototypes was calculated by Fisher's Z-transforming the correlation coefficients, computing the mean and transforming the value back to a correlation coefficient.

2) Theoretical Appraisal Importance

The weighting parameters (w_j) for model M2 were derived from the theoretical weights used by Scherer and Meuleman [19]. The authors actually present a numerical weighting parameter for each of the items used in the Geneva appraisal questionnaire. As the items were aggregated to build the 16 dimensions proposed by the CPM, we also averaged the weighting parameters to obtain one weight for each of the 16 appraisal dimensions.

3) Optimization of Appraisal Importance

A genetic algorithm was used to find the 16 or 208 appraisal weights that would minimize the predictive error of M3 and M4. Two objective functions (i.e. the functions to be minimized during the optimization processes) were defined that determine the mean misclassification error (MMCE) of the respective model over all observations of the balanced training set with the previously calculated prototypes p_{ij} and the 16 appraisal weights w_j or the 208 appraisal weights w_{ij} as free parameters. The optimizations were conducted using the *Differential Evolution* (DE) algorithm introduced by Storn and Price [30]. DE is a global optimization algorithm suited for high-dimensional, non-linear problems without requiring an either continuous or differentiable function. As other genetic algorithms, DE uses biology-inspired processes such as mutation, crossover and selection on a population to iteratively minimize or maximize the objective-function over successive generations [31]. The parallel search within a whole population of parameter configurations helps to avoid local minima, which makes DE superior to many direct search methods [30]. To conduct the optimization the *DEoptim* package [31] was used. The bounds of each parameter were set to 0.000001 (lower bound) and 10 (upper bound). To speed up the optimization process and to prevent misconvergence, the default settings of *DEoptim* were adapted. The step tolerance (*steptol*) was set to 200 and the relative convergence tolerance (*reltol*) to 0.001, which means that the optimization converges if there is no parameter configuration that decreases the MMCE by at least 0.001 after 200 populations. Additionally, the crossover rate

(*CR*), influencing the number of mutated values in the parameter configuration of a new population [29], was set to 0.9. Storn and Price [30] recommend using a higher *CR* of 0.9 or 1 to speed up convergence. Finally, the differential weighting factor (*F*) that is used to create new parameter configurations in the mutation process was set to 0.7, as Ardia et al. [31] suggest to lower or higher F a little (default setting is 0.8) to prevent misconvergence. By default, the population size *NP* is set to 10**p* (where *p* is the number of parameters), which means that *DEoptim* optimizes 160 potential solution for M3 and 2080 solutions for M4 in parallel.

We repeated the optimization process several times (10 times for M3 and 5 times for M4) with different random seeds, reporting the parameter configuration with the best out-of-sample performance (highest mean precision over all 13 emotion classes; see next paragraph for a description of the performance measures) as well as the mean variance of the parameter solutions as robustness measure. Additionally, we wanted to contrast the optimized parameters of M3 to the theoretical weights by Scherer and Meuleman [19] that we used in model M2. To this end, we report the Pearson correlation between the theoretical weights and the best parameter configuration of M3.

E. Model Validation

The four models with the theoretically and empirically generated parameters (p_{ij}, w_j and w_{ij}) were validated on the hold-out test set. For each of the models' predictions, we determined whether the predicted emotion class matched the given emotion label or, if two labels were present, matched either of the two labels. As the overall accuracy (or MMCE) can be a misleading performance indicator for unbalanced data sets (as more weight is put on frequent than on rare classes) and because we also wanted to analyze the performance for each emotion class separately, we additionally reported class-wise precision scores (number of real positive examples over all positive labelled examples) to assess the models' performance [32].[2]

To contrast the models' classification performance with a naive classifier, we also reported the performance of a weighted guess model that randomly predicts classes dependent on their relative frequency in the data set. As another benchmark, we conducted a Random Forest classification using the 16 appraisal dimensions as features.[3] We choose the ranger learner from the *ranger* package [34] with hyperparameters set to default. The model computation was conducted within the *mlr* framework by Bischl et al. [35]. As the model is not able to handle missing data, we recoded the 16 appraisal dimensions to factors and included missing values as an additional level. Thereby, we were able to train the Random Forest on the whole oversampled training set and validate it on the entire hold-out test set. Supervised black box models are able to learn data inherent structures by labelled instances. Their high predictive power comes at the cost of their interpretability. The model can be seen as a conservative upper limit of performance that can be reached with the present input variables, as the variance that is not explained by the model is rather due to incomplete input information or measurement error than insufficient model complexity.

Previous analyses by Scherer and Meuleman [19] had shown that the 13 emotion classes cluster into four emotion families: The *happiness* family with pleasure, joy and pride, the *anger* family including irritation, rage, contempt and disgust, the *distress* family including anxiety, fear, sadness and despair as well as the *shame and guilt* family. Because of this finding and the close resemblance of the emotion terms, which might make it difficult for participants to differentiate between the labels, we also assessed the classification performance for the four emotion families.

Next to classical performance measures, we also wanted to test how well each model was calibrated. Decalibration in discrete classification tasks is present, when a model predicts classes in proportions that do not match the original class distribution [36]. We therefore calculated two-way intraclass correlations (ICCs) between the real class proportions in the data and the class proportions in the predictions of the models.

V. Results

In the following results section, we report the results of our parameter calculation as well as the performance of the four implemented appraisal models in comparison to the naive baseline and machine learning model.

A. Prototypes

The prototypes (p_{ij}) for the 13 modal emotions calculated from the unbalanced training set can be found in the electronic appendix. The appraisal values of the newly attained prototypes showed a mean correlation of $r = 0.47$ to the appraisal values of the prototypes proposed by Scherer [3] (Table I).

TABLE I
PEARSON CORRELATIONS OF THE APPRAISAL DIMENSIONS BETWEEN THE PROTOTYPES CALCULATED FROM THE DATA SET AND THE THEORETICAL PROTOTYPES FROM SCHERER [7].

Emotion Prototype	r
Pleasure	0.44
Joy	0.56
Disgust	0.48
Sadness	0.57
Despair	0.64
Anxiety	0.57
Fear	0.73
Irritation	0.34
Rage	0.60
Shame	0.06
Guilt	0.07

[2] Because the present task is a multi-label as well as a multi-class classification problem and due to further characteristics of the data, no further performance measures were applicable.

[3] We compared different machine learning algorithms, finding that the tree based approach worked best with this type of data (which is in line with the findings of Meuleman and Scherer [33]). The results of this benchmark experiment can be found in the electronic appendix.

Pride	0.42
Contempt	0.31

B. Emotion classification

The weighted guess baseline model showed an overall accuracy of 17.9% in the classification of the 13 emotions on the test set. The class-wise precision (see Table II for all precision scores) of this naive model ranged from 2.0% (contempt) to 30.5% (sadness).

The first model without any weighting (M1) yielded an overall accuracy of 37.1 % on the test set that was considerably higher than the overall accuracy of the weighted guess model. The class-wise precision varied widely with scores ranging from 3.7 % (contempt) to 82.7 % (joy). For all 13 emotion categories, the classification performance of M1 was notably higher than the performance of the baseline model.

The second model (M2) using the theoretical weights by Scherer and Meuleman [19] showed an overall accuracy of 27.1%. Again, the precision scores differed strongly between classes, ranging from 4.2% (contempt) to 61.8% (sadness). All class-wise precision scores were higher than the precision scores yielded by the weighted guess baseline model. Nevertheless, M2 was outperformed by the unweighted M1, that reached higher scores in all classes except for despair, irritation and contempt as well as a higher overall accuracy.

The Differential Evolution optimization for the 16 parameters of M3 was repeated using 10 random seeds. The parameter configurations over the 10 replications showed a mean variance of 1.09 (range = 0.12–4.36)[4] with some parameters, such as the weight for the pleasantness appraisal, being estimated more robustly than others. The best solution (yielding the highest out-of-sample mean precision) converged after 534 iterations (populations) with an in-sample accuracy of 42.2%. The out-of-sample accuracy on the validation test set reached 40.2% and was higher than the overall accuracy of the baseline model, M1 and M2. The class-wise precision scores, ranging from 4.3% (contempt) to 81.6% (joy), exceeded all precision scores of the baseline model. In 10 of the 13 emotion classes M3 reached a higher precision than the unweighted M1. For the emotions pleasure, joy and rage though, M1 yielded slightly better values. M3 also outperformed M2 in 11 of the 13 emotion classes, yielding higher scores except for the emotions rage and irritation.

The Differential Evolution optimization for the 208 parameters of M4 was repeated five times using different random seeds. The parameter configurations showed a variance of 5.03 (range = 0.11–18.24) across optimization repetitions. This is substantially higher than the variation of parameters in M3, which points towards a strong instability in the optimization. Again, some of the 208 parameters were estimated robustly over the iterations, while some showed a very high variance. The parameter solution with the best out-of-sample performance converged after 1635 iterations at an in-sample accuracy of 45.3%. On the validation test set, the model showed an out-of-sample accuracy of 43.2% that outperformed the weighted guess classifier, M1, M2 as well as M3. But the class-wise precision scores show that M4 actually yielded worse precisions than the simpler M3 in all classes expect for two (despair and guilt). Furthermore, it outperformed the unweighted M1 in only five cases (despair, anxiety, shame, guilt and contempt) and the theoretical weighted M3 in only 7 of the 14 classes (pleasure, joy, despair anxiety, fear, shame and guilt). Still, the precision scores of M4 were higher than the ones of the baseline model for all emotion classes.

With an out-of-sample accuracy of 52.3%, the Random Forest showed overall a better performance than all other models. The class-wise precision scores ranged from 14.8 % for contempt to 78.0% for joy. The Random Forest outperformed M1 and M3 in 9 of the 13 classes. Only for the classes joy, sadness, rage and pride, M1 and M3 showed a better performance. M2 was outperformed in all cases except for sadness and rage. Again, all precision values were notably higher than the scores of the baseline model.

Pearson's correlation between class frequency in the test set and the precision scores revealed significant positive relations between class size and predictive performance for all four models (M1: $r(11) = 0.83$, $p < 0.001$; M2: $r(11) = 0.92$, $p < 0.001$; M3: $r(11) = 0.87$, $p < 0.001$; M4: $r(11) = 0.86$, $p < 0.001$) as well as for the Random Forest ($r(11) = 0.78$, $p = 0.002$).

TABLE II
PERCENTAGE PRECISION SCORES OF THE 13 EMOTION CLASSES FOR M1 WITH NO WEIGHTING, M2 WITH THE 16 THEORETICAL WEIGHTS, M3 WITH THE 16 OPTIMIZED WEIGHTS, M4 WITH 208 OPTIMIZED WEIGHTS, WEIGHTED GUESS CLASSIFIER (WGC), AND RANDOM FOREST (RF) CLASSIFIER.

Emotion	N[a]	M1	M2	M3	M4	WGC[b]	RF
Pleasure	363	44.7	21.4	43.7	37.3	11.0	51.4
Joy	719	82.7	49.6	81.6	75.2	21.8	78.0
Disgust	163	12.9	11.5	15.0	7.3	5.0	20.0
Sadness	1006	64.1	61.8	69.2	55.1	30.5	55.8
Despair	431	25.9	28.5	28.2	33.3	13.1	31.5
Anxiety	667	32.5	28.1	43.5	34.3	20.2	50.4
Fear	579	37.0	34.7	38.8	36.1	17.6	42.4
Irritation	320	26.5	27.9	26.7	22.1	9.7	32.5
Rage	633	43.9	42.3	42.4	37.4	19.2	41.9
Shame	189	9.1	6.7	20.0	9.3	5.7	33.3
Guilt	226	15.3	7.1	15.5	15.7	6.9	32.7
Pride	300	36.9	32.9	38.6	25.5	9.1	35.7
Contempt	67	3.7	4.2	4.3	3.8	2.0	14.8

Note: N = Sample size of the emotion classes in the validation test set. [a] Note that the class sample sizes do not add up to the total sample size of the test set, as many observations have two class labels. [b] The precision scores of WGC model are equivalent to those of a random model without weighting of class frequencies.

C. Emotion family classification

In the classification of the four emotion families, the naive weighted guess classifier showed an overall accuracy of 43.6%

[4] With parameters constrained between 0.000001 and 10, the maximum variance possible was 25.

on the test set. The class-wise precision scores ranged from 11.9% for the shame/guilt family to 62.1% for the disgust family (see Table III for precision scores of all models).

M1 with no weighting algorithm showed an overall higher accuracy of 73.9% on the test set. All precision scores, ranging from 24.5% (shame/guilt) to 90.1% (happiness), were considerably higher than the scores of the naive baseline model.

Model M2 with the theoretically derived weighting parameters yielded an overall lower accuracy of 62.4%. The precision scores of the emotion families were higher than the ones of the baseline model, but worse than the precisions of M1 for all classes.

M3 with the 16 optimized appraisal weights reached a higher out-of-sample accuracy (76.9%) than M1 and showed higher precision scores for all emotion families except for anger. The precision scores ranged from 27.7% for shame/guilt to 92.0% for happiness.

With an overall out-of-sample accuracy of 71.9%, the complex weighted model M4 with the 208 optimized parameters performed again better than the baseline model, but showed a lower accuracy than M1 and M3. The class-wise precision scores ranging from 21.6% (shame/guilt) to 92.0% (happiness) were again lower than the precision scores of the simpler M3 for all classes except for happiness, where both models performed equally well. In the three other classes, M4 reached also lower precision scores than the unweighted M1.

Finally, the Random Forest classifier again showed an overall higher out-of-sample accuracy than the other models (80.8%). With precision scores ranging from 37.5% (shame/guilt) to 94.3% (happiness), the Random Forest also yielded higher precisions for the happiness, anger and shame/guilt family, but was surpassed by M1 and M3 for the disgust family.

TABLE III
PERCENTAGE PRECISION SCORES OF THE 4 EMOTION CLASSES FOR M1 WITH NO WEIGHTING, M2 WITH THE 16 THEORETICAL WEIGHTS, M3 WITH 16 WEIGHTS, M4 WITH 208 WEIGHTS, WEIGHTED GUESS CLASSIFIER (WGC), AND THE RANDOM FOREST (RF).

Emotion family	N[a]	M1	M2	M3	M4	WGC[b]	RF
Happiness	953	90.1	64.7	92.0	92.0	28.9	94.3
Anger	981	54.0	49.7	53.6	49.8	29.8	60.5
Disgust	2048	86.2	83.5	86.3	84.6	62.2	85.0
Shame/Guilt	393	24.5	18.8	27.7	21.6	11.9	37.5

Note: N = Sample size of the emotion classes in the validation test set. [a] Note that the class sample sizes do not add up to the total sample size of the test set, as many observations have two class labels. [b] The precision scores of WGC model are equivalent to those of a random model without weighting of class frequencies.

D. Model calibration

With an *ICC* of 0.317 (p = 0.134, *CI* [-0.259, 0.727]), the class probability distribution of M1 showed a poor consistency with the actual class probabilities in the data. M2 had a worse *ICC* of -0.129 (p = 0.67, CI [-0.619, 0.433]. With an *ICC* of 0.411 (p = 0.072, *CI* [-0.156, 0.774]), M3 yielded a slightly higher calibration than M1. M4 reached a moderate *ICC* of 0.705 (p = 0.002, *CI* [0.277, 0.900]). The Random Forest classifier showed an even higher ICC of 0.808 (p < 0.001, *CI* [0.484, 0.937]). Naturally, the model with the highest ICC was the weighted guess classifier, reproducing the class probability distribution of the data set perfectly with an *ICC* of 0.997 (p < 0.001, *CI* [0.989, 0.999]).

E. Appraisal weights

Table IV shows the parameter configuration (w_j) of M3 that yielded the best out-of-sample performance. The 16 optimized weighting parameters ranged from 2.53 (outcome probability) to 9.71 (intrinsic pleasantness). The Pearson correlation between the optimized weights and the theoretical weights reported by Scherer and Meuleman [19] was modest (r(14) = 0.30, p = 0.26). The theoretical weights (w_j) as well as the 208 parameters (w_{ij}) for M3 can be found in the electronic appendix. As many of the parameters of M3 showed a rather high variance (which indicates that the optimization results are lacking robustness), we caution against interpreting these parameters.

TABLE IV
16 APPRAISAL WEIGHTS OF THE DIFFERENTIAL EVOLUTION OPTIMIZATION OF M3 WITH THE BEST OUT-OF-SAMPLE PERFORMANCE.

Appraisal dimension	Weights w_j
Intrinsic pleasantness	9.71
Urgency	7.94
Goal/need Relevance	7.68
Internal standards	6.89
Power	6.12
External standards	5.89
Adjustment	5.82
Suddenness	5.45
Familiarity	5.42
Predictability	5.17
Conduciveness	4.47
Control	4.01
Cause: Agent	3.52
Discrepancy from expectation	3.05
Cause: Motive	3.01
Outcome probability	2.53

VI. Discussion

In the present study, we used a predictive modelling approach to validate and extent the CPM model, an appraisal emotion theory, by assessing the emotion prediction accuracy of four computational emotion models. The models used ratings of 16 appraisal dimensions assessed in an online questionnaire to predict an emotion term by calculating the similarities between the ratings and 13 emotion prototypes. Different weighting algorithms (mapping of appraisals to emotions) were implemented in the four models to assess their plausibility by comparing the models' performances. To generate new

information, parameters within these models, including the emotion prototypes as well as the weighting parameters (for M3 and M4), were generated from the empirical data and contrasted with theoretical assumptions from the literature.

All four theoretical models performed notably better than the baseline model (weighted guess classifier), that randomly predicted emotion classes weighted by their frequency in the data set. This shows that the appraisal dimensions, evaluating 16 different emotion-relevant aspects of a situation, are able to explain a part of the variance in the subjective feeling experienced by subjects. By integrating all 16 dimensions equally strong during the classification task, M1 was able to predict one of the given emotion labels correctly in 37.1 % of the cases. The precisions scores of M1 varied strongly between emotions with classes included more frequently in the data being predicted with higher precisions. This observation, that was apparent for all models, is only partly due to the lower baseline probability in smaller classes. It is plausible that the prototypes (p_{ij}) calculated from these small classes are less reliable, as there might be insufficient information to build a prototype and because of the mean's sensitivity to outliers and skewedness. Consequently, the classification performance in classes with poor prototypes drops. When looking at the family classification performance of M1, it can be seen that even though the exact emotion label was found in only a third of the cases, the model actually predicted the correct emotion family in 73.9% with precision rates up to 90.1% (happiness family). As presumed, this high increase in performance might be due to the fact that the emotion labels often were very similar to each other (e.g. pleasure vs. joy or fear vs. anxiety). The lack of clarity in the terminology might lead to a differing understanding of the emotion labels between participants or to randomness in the selection of emotion terms. As a consequence, prototypes calculated from a subset with many "wrongfully" labelled ratings lack the ability to differentiate between emotion classes. Also, many appraisal ratings might not be true instances of the modal emotion they are identified as, because participants are forced into a few distinct emotion classes. Especially when two labels are given, the appraisal patterns rather reflect a blend of two modal emotions or even a separate emotion state. The characteristics of the broader emotion families, might therefore be more stable and better differentiating. As an additionally performance evaluation, we looked at the models' calibration to the class probability distribution in the data, where M1 yielded a poor performance as it was not able to reproduce the true class frequencies.

With an overall accuracy of 27.1% for the emotion classes and 62.4% for the emotion families, model M2 with the 16 theoretical derived weighting parameters, yielded the worst performance of all four CPM models, also showing the worst model calibration. This indicates that the appraisal importance assumed by Scherer and Meuleman [19] seems to be not a very good estimation of the true appraisal importance – at least in the context of the present data and with the current computation of the similarity index. Even the equally weighted (or unweighted) model M1, showed a better overall accuracy as well as higher precision scores for most classes. Furthermore, the implementation of the 16 empirically derived weighting parameters in M3 lead to an overall increase in model performance. M3 reached a substantially higher out-of-sample accuracy of 40.2% than M1 and M2, with higher precision rates for most of the emotion classes. The same pattern was found for the emotion family classification where M3 again reached a higher overall accuracy and higher precision rates. The difference in performance between M2 and M3 is also in line with the finding that the optimized parameters of M3 did not show a substantial correlation with the theoretical derived parameters of M2, subsequently the parameters differed strongly. Even though smaller classes were oversampled in the balanced training set, precision differences between smaller and larger classes remained. Again, this suggests that performance differences between classes could be due to insufficient information in the prototype calculation. The ICC between the model's class distribution and the true class distribution showed a slightly better model calibration than M1. The weighting parameter configuration, assessed across repeated Differential Evolution optimizations, showed a low mean variance which indicates a good stability of the optimization results and suggest that the found parameters reflect the global minimum of the objective function. Within M3, the appraisal dimension *pleasantness* received by far the highest weight ($w = 9.71$) for the emotion classification. Intrinsic pleasantness, the basal evaluation whether a stimulus is likely to result in pleasure or pain [23], is also included in other appraisal theories (e.g. [9], [10]). The very importance of pleasantness in the emergence of emotions is also reflected in other emotion models, such as Russell's [37] theory of Core Affect. He describes emotions as an integral blend of two dimensions, arousal (activation vs. deactivation) and valence (pleasure vs. displeasure) of a stimulus. It is plausible that the valence of a stimulus is a strong predictor, as it clearly separates the emotions space into positive and negative emotions. This can be seen in the prototype values for pleasantness (see appendix), as all positive emotions (happiness, joy and pride) showed very high pleasantness while all negatives emotions showed a very low pleasantness prototype. The second highest appraisal weight was placed on the dimension *urgency* ($w = 7.94$). Sander, Grandjean and Scherer [23] describe urgency as the appraisal that determines if an event endangers high priority goals or needs and if the organism has to react quickly or flee. Hence, a high rating of urgency should lead to an immediate increase in action readiness and response of the automatic nervous system. Scherer [3] links urgency to the dimension of activation or arousal, that has been identified as the second of two relevant dimensions by Russell [37]. Both dimensions together, are able to perfectly separate negative and positive emotions with joy, happiness and pride having very high prototype values for pleasantness as well as low prototype values for urgency, while the other negative emotions have very low values in the pleasantness dimensions and higher values in urgency. But it is obvious that the two dimensions are not sufficient to differentiate between all the thirteen emotions categories. Another argument against the two-dimensional approach to emotions is the fact, that none of the remaining 14 appraisals were shrunken down to a weight of 0. In fact, further dimensions such as *goal/need relevance* ($w = 7.68$) as well as *internal standards* ($w = 6.89$) yielded considerably high weights, while *outcome probability* obtained the lowest value with $w = 2.53$. This indicates that all 16 appraisal dimensions contributed to the emotion determination to some degree, which

supports the belief that two dimensions are not sufficient to represent and describe emotional states properly [38]. The attained weighting can also be compared to other instantiations of appraisal models. Lazarus' cognitive-motivational-relational theory [39], for example, includes only six dimensions, four of which are also present in the CPM (goal/need relevance, conduciveness, cause and power). The weighting parameters show though, that the additional parameters not included in Lazarus' simpler model such as urgency (w = 7.94) or internal standards (w = 6.89) also seem to contribute strongly to the prediction of emotions. Especially, the absence of the pleasantness appraisal in his model seems striking, as this appraisal yielded the highest weight (w = 9.71) in our model and is included in many other appraisal theories (e.g. [9], [10]). Besides the four dimensions included in the CPM, Lazarus' model additionally contains the dimension *goal content* and *future expectation*. The former appraisal, which is also included in the appraisal theory of Roseman [11], defines the current type of goal being at stake, while the latter evaluates whether one thinks an event will work out favorably in the future. Both dimensions could potentially explain additional variance in the emotion classification. Another appraisal theory, the OCC model [17], reduces the evaluation process to only three main appraisal domains: The evaluation of events in terms of their desirability, the rating of actions as praise or blameworthy as well as the appraising of objects as either appealing or unappealing. These three dimensions are presented by the appraisals conduciveness, compatibility of internal and external standards as well as pleasantness in the CPM. Again, our results indicate though that these three dimensions are not sufficient enough to differentiate between all 13 emotion classes used in the present study. It has to be remarked though, that the differences in number and identity of dimensions between appraisal theories is mainly due to the number of emotions a model aims to explain – when trying to predict only four emotion classes such as joy, anger, fear and disgust, one obviously does not need as many predictors as a model trying to explain a broader range of emotions [4]. Furthermore, theorist differ in their view on parsimonious modelling, where some try to include only sufficient or typical appraisals, while others focus on completeness [4]. When comparing the present results to other appraisal theories, it is also important remark that most theories don't make particular assumptions how the appraisals are aggregated during the emotion emergence process (i.e. they don't make any comments on the importance of different appraisals). The comparison between M1 and M3 though clearly shows that an equal weighting of appraisals restrains the model performance.

Model M4 used a more complex weighting algorithm than M3 with a separate weighting not only for each appraisal dimension but also for each appraisal dimensions within each of the 13 modal emotions. The application of the 208 weights resulted in a slightly higher out-of-sample accuracy of 43.2%. The precision analysis though showed that M4 actually yielded lower precision rates than M3 for most emotion classes and even some lower precisions rates than M1. This apparent paradox – the model with the higher accuracy actually showing a poorer class-wise predictive performance – can be explained by the classification behavior of M4 as well as the calculation of the precision scores. M4 very frequently predicts the classes that are prevalent in the data set such as sadness, joy, fear and rage. This better calibration to the class probability distribution in the data also shows in the higher ICC score of the model. In the more frequently predicted classes, M4 classifies more cases correctly than the two other models (leading to a higher overall accuracy) but also produces way more false positives. As the precision score is the proportion of correctly classified instances in all as positive labelled observations, the precision scores of M4 are lower for these emotion categories even though more instances were classified correctly. The same pattern was present for the emotion families, where M4 showed a poorer performance in three out of four classes. The 208 parameters obtained by the optimization showed a notably higher variation than the parameters of M3 with some parameters yielding almost diametrical values over the five optimization repetitions. This indicates that the optimizations which all stopped at a similar in-sample accuracy found different equivalent parameter configurations. Hence, no global optimum was found and the parameters should not be interpreted.

By contrasting the four models M1, M2, M3 and M4, we wanted to test the plausibility of their underlying weighting algorithms. With a higher over-all accuracy, higher precision rates for most classes and a better calibration, M3 can be preferred over the unweighted M1 model and M2 with the theoretical derived parameters. Even though the increase in performance between M1 and M3 is not massive, the differential weighting of the 16 appraisal dimensions as it has been proposed in the literature [19], [23] leads to a considerable improvement. The big gap in performance between M2 and M3, suggest though that the 16 theoretical weighting parameters don't seem to be a good representation of appraisal importance within the used data set. A more ambiguous picture emerges when M3 is compared to the more complex weighted model M4. Even though M4 yields a higher over-all accuracy, the precision rates drop due its strong calibration to the few large classes in the sample. Despite the better calibration of M4 (higher ICC), a good estimation of the class distribution cannot be a stand-alone criterion for model performance as the weighted guess classifier, the naive baseline model, satisfied this aspect perfectly. A clear detriment of M4 is that the weighting parameters in the model are not interpretable due to the missing stability of the optimization results. Under the principle of parsimony, which recommends choosing the simpler and interpretable model, we would therefore favor M3, the model that is implied by Scherer's theory. Also, from a perspective of cognitive economy, the complex weighting of M4 might be too costly for a highly automated process like emotion formation. This preference contradicts the findings of Ellsworth and Smith [40], that believe that appraisal importance differs between emotion classes.

We additionally included the Random Forest model to see what an uninformed black box model could derive from the data. As expected, the model showed an overall good performance, yielding higher accuracies and higher precision scores for many emotion classes and emotion families. The Random Forest also showed a good calibration to the class frequencies in the data. Nonetheless, there was still variation in the emotion labels that could not be explained by the model as 47.7% of the emotion classes and 19.2% of the emotion families

were classified incorrectly. This shows that even with a more elaborate structure, there is an upper boundary of model performance that probably cannot be exceeded with the present data. With regards to our computational emotion models, this means that there is a limited scope for further model improvement. Instead, it seems likely that the appraisal ratings in the present data set are not sufficient to explain all variance in the subjective feeling of the participants. There could be further appraisal dimensions necessary to clearly distinguish between all 13 emotion classes, but it is also plausible that the models' performances are impaired by measurement error in appraisal ratings or emotion labelling. Particularly the usage of self-report for the measurement of appraisals has been criticized (e.g. [41]), as it relies on information that is consciously accessible and can be verbalized easily. Therefore, the method might not be suitable to assess automatic and subconscious processes. The CPM actually implies that the 16 appraisal dimensions rely on different cognitive functions some of which are more basal and automatic like memory- and attention-driven processes whereas others also engage higher cognitive functions like reasoning and evaluation of self-image [23]. It can be questioned whether appraisal dimensions driven by more basal cognitive functions are actually consciously accessible and consequently, whether we can measure these constructs adequately using subjective self-reports. Many theorists recognize this limitation of self-assessed appraisals ([39], [42], [43]). Scherer [43] himself states that it is unlikely that all appraisal processes are consciously accessible and easy to verbalize – specifically those processed subcortically. He believes that some subliminal processes can be reconstructed from memory, but that many self-reported ratings are more likely constructed by using established schemata of emotions and prototypes for certain event types. If participants use these rather heuristic methods for the evaluation of some dimensions, ratings have to be affected by measurement error to some degree. This measurement problem, relying on introspection for the assessment of cognitive and psychological processes, many of which being at least partly subconscious or not accessible due to a lack of self-knowledge, is common to many fields of psychology. In the past, studies have tried to detect physiological markers of different appraisal dimensions (see [13] for an overview), which could help to develop a more objective operationalization of the appraisal process. Unfortunately, these studies were only able to manipulate a few appraisal dimensions at a time (but never the complete set of appraisals) and even though there is some knowledge about physiological feedback related to specific appraisals, it is very difficult to assess an underlying appraisal dimension in an experimental setting [19]. Scherer [43] expresses his hope that technological progression of neuroscientific methods will someday enable us to map different contents of processing (not only cognitive processes) in the brain. But until this or other methodological developments enable a more objective measurement of the appraisals, studies on this topic will continue to rely on self-reported ratings. In further research, the subjective measurements of appraisals might be improved though, by using more direct and less retrospective evaluations of an event. Asking participants to rate an event immediately after they experienced it, could make the appraisal evaluation more accessible, but the main problem, the reliance on introspection, will remain nonetheless. This important limitation of the present study, the reliability of the appraisal measurements, has to be kept in mind when interpreting the results. Not only has this limitation an influence on the upper performance that can be reached with the present models, but it will also affect the estimated model parameters. We therefore cautioned against generalizing the found parameters and further urge to validate the weights on different types of data sets – not only changing the appraised contexts, but also using more reliable measurement techniques, when they are made available.

In summary, the computational modelling approach used in the present study lends some support to the psychological appraisal theories of emotion and the CPM. Using the 16 appraisal dimensions proposed by the latter, we were able to predict emotions given by subjective self-report much more frequently than simply by chance. The comparison of the four weighting algorithms also suggest that the 16 appraisal dimensions contribute differently strong to the emotion classification process. Even though this is also in line with the model assumptions, the weighting parameters of the preferred model that were attained by optimization deviate from the theoretical weights. As the new parameters have been derived inductively from the data and due to the limitations of the present data set, further research has to be conducted to validate these findings in different contexts. As the ratings of appraisal by self-report are very likely afflicted by a high measurement error, future research needs to focus on the development of more objective assessments of appraisal. Also, due to its many advantages, the application of computational emotion modelling as a way of validating and extending hypotheses generated by empirical research or theory, should be integrated more strongly in the theory development process.

Electronic Appendix

The electronic appendix and further supporting information are provided via the Open Science Framework (OSF) at https://osf.io/te4z3/.

Acknowledgement

We thank Prof. Dr. Klaus Scherer for providing us the data for this paper and for his valuable commentary in the initial stage of this research project.

References

[1] S. Marsella, J. Gratch, and P. Petta, 'Computational Models of Emotion', in *A blueprint for an affectively competent agent: Cross-fertilization between Emotion Psychology, Affective Neuroscience, and Affective Computing*, K. R. Scherer, T. Bänzinger, and E. Roesch, Eds. Oxford: Oxford University Press, 2010, pp. 21–41.

[2] A. Moors, P. C. Ellsworth, K. R. Scherer, and N. H. Frijda, 'Appraisal Theories of Emotion: State of the Art and Future Development', *Emot. Rev.*, vol. 5, no. 2, pp. 119–124, Apr. 2013.

[3] K. R. Scherer, 'Appraisal considered as a process of multilevel sequential checking', in *Appraisal processes in emotion*, K. R. Scherer, A. Schorr, and J. Johnstone, Eds. New York: Oxford University Press, 2001, pp. 92–120.

[4] A. Moors, 'Theories of emotion causation: A review', *Cogn. Emot.*, vol. 23, no. 4, pp. 625–662, Jun. 2009.

[5] W. James, 'WHAT IS AN EMOTION ?', *Mind*, vol. os-IX, no. 34, pp. 188–205, Apr. 1884.

[6] S. Schachter and J. Singer, ‘Cognitive, social, and physiological determinants of emotional state.’, *Psychol. Rev.*, vol. 69, no. 5, pp. 379–399, Sep. 1962.

[7] I. J. Roseman, ‘A model of appraisal in the emotion system: Integrating theory, research, and applications’, in *Appraisal Processes in Emotions*, K. R. Scherer, A. Schorr, and J. Johnstone, Eds. New York: Oxford University Press, 2001, pp. 3–34.

[8] C. A. Smith and R. S. Lazarus, ‘Emotion and Adaptation’, in *Handbook of personality: Theory and Research*, L. A. Pervin, Ed. New York: The Guilford Press, 1990, pp. 609–637.

[9] C. A. Smith and P. C. Ellsworth, ‘Patterns of cognitive appraisal in emotion.’, *J. Pers. Soc. Psychol.*, vol. 48, no. 4, pp. 813–838, May 1985.

[10] N. H. Frijda, *The emotions*. Cambridge: Cambridge University Press, 1986.

[11] I. Roseman, ‘Cognitive Determinants of Emotion: A Structural Theory’, *Rev Soc Psychol*, vol. 5, pp. 11–36, Jan. 1984.

[12] A. Moors, ‘Automatic Constructive Appraisal as a Candidate Cause of Emotion’, *Emot. Rev.*, vol. 2, no. 2, pp. 139–156, Apr. 2010.

[13] K. R. Scherer, ‘The dynamic architecture of emotion: Evidence for the component process model’, *Cogn. Emot.*, vol. 23, no. 7, pp. 1307–1351, Nov. 2009.

[14] C. D. Elliott, ‘The Affective Reasoner: A process model of emotions in a multi-agent system’, Institute for Learning Sciences, Northwestern University, 1992.

[15] C. Becker-Asano, ‘WASABI: Affect simulation for Agents with Believable Interactivity.’, Faculty of Technology, University of Bielefeld, 2008.

[16] R. P. Marinier, J. E. Laird, and R. L. Lewis, ‘A computational unification of cognitive behavior and emotion’, *Cogn. Syst. Res.*, vol. 10, no. 1, pp. 48–69, Mar. 2009.

[17] A. Ortony, G. Clore, and A. Collins, *The Cognitive Structure of Emotions*. Cambridge: Cambridge University Press, 1988.

[18] K. R. Scherer, ‘Studying the emotion-antecedent appraisal process: An expert system approach’, *Cogn. Emot.*, vol. 7, no. 3–4, pp. 325–355, May 1993.

[19] K. R. Scherer and B. Meuleman, ‘Human Emotion Experiences Can Be Predicted on Theoretical Grounds: Evidence from Verbal Labeling’, *PLoS ONE*, vol. 8, no. 3, p. e58166, Mar. 2013.

[20] I. Douven, ‘Abduction’, *The Stanford Encyclopedia of Philosophy*. 2017.

[21] P. Ekman, ‘An argument for basic emotions’, *Cogn. Emot.*, vol. 6, no. 3–4, pp. 169–200, May 1992.

[22] K. R. Scherer, ‘On the nature and function of emotion: A component process approach’, in *Approaches to Emotion*, K. R. Scherer and P. Ekman, Eds. Hillsdale, NJ: Erlbaum, 1984, pp. 293–317.

[23] D. Sander, D. Grandjean, and K. R. Scherer, ‘A systems approach to appraisal mechanisms in emotion’, *Neural Netw.*, vol. 18, no. 4, pp. 317–352, May 2005.

[24] E. Rosch, ‘Prototype classification and logical classification: The two systems’, in *New trends in conceptual representation: Challenges to Piaget’s theory*, E. Scholnick, Ed. Hillsdale, NJ: Erlbaum, 1983, pp. 73–86.

[25] Geneva Emotion Research Group, ‘Geneva Emotion Research Group (2002), “Geneva Appraisal Questionnaire (GAQ): Format, Development, and Utilization”, Version 3.0 (August), http://www.affective-sciences.org/system/files/page/2636/GAQ_English.PDF.’ 2002.

[26] A. Mahto, *splitstackshape: Stack and Reshape Datasets After Splitting Concatenated Values*. 2018.

[27] N. Lunardon, G. Menardi, and N. Torelli, ‘ROSE: a package for binary imbalanced learning’, *R J.*, vol. 6, no. 1, pp. 79–89, Jun. 2014.

[28] R Core Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2018.

[29] M. Kuhn, ‘Building Predictive Models in R Using the caret Package’, *J. Stat. Softw.*, vol. 28, no. 5, pp. 1–26, Nov. 2008.

[30] R. Storn and K. Price, ‘Differential Evolution – A Simple and Efficient Heuristic for global Optimization over Continuous Spaces’, *J. Glob. Optim.*, vol. 11, no. 4, pp. 341–359, Dec. 1997.

[31] D. Ardia, K. M. Mullen, B. G. Peterson, and J. Ulrich, *DEoptim: Differential Evolution in R*. 2016.

[32] M. Bekkar, H. K. Djemaa, and T. A. Alitouche, ‘Evaluation Measures for Models Assessment over Imbalanced Data Sets’, *J. Inf. Eng. Appl.*, vol. 3, no. 10, pp. 27–38, Jan. 2013.

[33] B. Meuleman and K. Scherer, ‘Nonlinear Appraisal Modeling: An Application of Machine Learning to the Study of Emotion Production’, *IEEE Trans. Affect. Comput.*, vol. 4, no. 4, pp. 398–411, Oct. 2013.

[34] M. N. Wright and A. Ziegler, ‘ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R’, *J. Stat. Softw.*, vol. 77, no. 1, pp. 1–17, Mar. 2017.

[35] B. Bischl *et al.*, ‘mlr: Machine Learning in R’, *J. Mach. Learn. Res.*, vol. 17, no. 170, pp. 1–5, Sep. 2016.

[36] A. Bella, J. Hernandez-Orallo, and M. J. Ramirez-Quintana, ‘Calibration of machine learning models’, in *Handbook of Research on Machine Learning Applications*, 2009, pp. 128–146.

[37] J. A. Russell, ‘Core affect and the psychological construction of emotion.’, *Psychol. Rev.*, vol. 110, no. 1, pp. 145–172, Jan. 2003.

[38] J. R. J. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth, ‘The World of Emotions is not Two-Dimensional’, *Psychol. Sci.*, vol. 18, no. 12, pp. 1050–1057, Dec. 2007.

[39] R. S. Lazarus, *Emotion and Adaptation*. New York: Oxford University Press, 1991.

[40] P. C. Ellsworth and C. A. Smith, ‘Shades of Joy: Patterns of Appraisal Differentiating Pleasant Emotions’, *Cogn. Emot.*, vol. 2, no. 4, pp. 301–331, Oct. 1988.

[41] R. J. Davidson, ‘Prolegomenon to the structure of emotion: Gleanings from neuropsychology’, *Cogn. Emot.*, vol. 6, no. 3–4, pp. 245–268, May 1992.

[42] N. H. Frijda, ‘The place of appraisal in emotion’, *Cogn. Emot.*, vol. 7, no. 3–4, pp. 357–387, May 1993.

[43] K. R. Scherer, ‘Neuroscience projections to current debates in emotion psychology’, *Cogn. Emot.*, vol. 7, no. 1, pp. 1–41, Jan. 1993.

Laura S. F. Israel was born in Karlsruhe, Baden-Württemberg, Germany in 1989. She received a B.A. in history and cognitive science as well as a M.Sc. in cognitive science at the Albert Ludwig University in Freiburg in 2016. She is currently pursuing a Ph.D. degree in psychology at the Ludwig Maximilian university in Munich.

From 2013 to 2017, she worked first as a Research Assistant and later as Research Associate at the Center for Cognitive Science in Freiburg focusing on psycholinguistic research – specifically on verbal humor. Since 2017, she is involved in a project on automatic emotion classification from video sequences funded by the German Research Foundation. In addition to computational modelling of emotions and emotion prediction, she is also interested in the physiological measurement of emotions.

Felix D. Schönbrodt received his diploma in psychology from Saarland University, Germany, in 2006 and his doctoral degree from Humboldt University Berlin, Germany, in 2010. He currently is a lecturer at the psychological department of the Ludwig Maximilian University Munich, Germany.

His research interests include implicit and explicit motives, quantitative methods, and issues revolving open science and the replicability of research. One special focus is to provide statistical packages in R and interactive statistical web apps which can be used for teaching and for an enhanced understanding and usage of quantitative methods. Felix Schönbrodt is an initiator of the "Commitment to Research Transparency" (http://www.researchtransparency.org).

Dr. Schönbrodt is a member of the German Psychological Society and received the Leamer-Rosenthal Prize for Open Social Science in 2016.