The Cognitive Emotion Process:

Examining Appraisal Theory using

Theoretical Modeling and Machine Learning

Laura Israel

Inauguraldissertation

The Cognitive Emotion Process:

Examining Appraisal Theory using

Theoretical Modeling and Machine Learning

Inauguraldissertation

zur Erlangung des Doktorgrades der Philosophie an der

Ludwig-Maximilians-Universität München

vorgelegt von Laura Israel

aus Karlsruhe

München, 2020

Erstgutachter: Felix Schönbrodt

Zweitgutachter: Markus Bühner

Datum der mündlichen Prüfung: 10.07.2020

Abstract

Different theories of emotions have been introduced since the 19th century. Even though a large number of apparent differences between these theories exist, there is a broad consensus today that emotions consist of multiple components such as cognition, physiology, motivation, and subjectively perceived feeling. Appraisal theories of emotions, such as the *Component Process Model* (CPM) by Klaus Scherer, emphasize that the cognitive evaluation of a stimulus or event is the driving component of the emotion process. It is believed to cause changes in all other components and hence to differentiate emotion states. To test the CPM and gain more insights into the multi-componential emotion process, the present thesis examines two emotion sub-processes – the link between the cognitive and the feeling component (study 1) and the link between the cognitive and the physiological component (study 2) – by using different predictive modeling approaches.

In study 1, four theoretically informed models were implemented. The models use a weighted distance metric based on an emotion prototype approach to predict the perceived emotion of participants from self-reported cognitive appraisals. Moreover, they incorporate different weighting functions with weighting parameters that were either derived from theory or estimated from empirical data. The results substantiate the examined link based on the predictive performance of the models. In line with the CPM, the preferred model weighted the appraisal evaluations differently in the distance metric. However, the data-derived weighting parameters of this model deviate from theoretically proposed ones.

Study 2 analyzed the link between cognition and physiology by predicting self-reported appraisal dimensions from a large set of physiological features (calculated from different physiological responses to emotional videos) using different linear and non-linear machine learning algorithms. Based on the predictive performance of the models, the study is able to confirm that most cognitive evaluations were interlinked with different physiological responses. The comparison of the different algorithms and the application of methods for interpretable machine learning showed that the relation between these two components is best represented by a non-linear model and that the studied link seems to vary among physiological signals and cognitive dimensions.

Both studies substantiate the assumption that the cognitive appraisal process is interlinked with physiology and subjective feelings, accentuating the relevance of cognition in emotion as assumed in appraisal theory. They also demonstrate how computational emotion modeling can be applied in basic research on emotions.

Authorship

The two presented studies were conducted and written by the author of this thesis, while Felix Schönbrodt acted as the supervising author. As the studies were published or submitted for publication in consultation with the co-author, the pronoun *we* is used in all references to the studies, as well as in the studies themselves.

The data set used in study 1 (chapter 2) was provided by Klaus Scherer and collected by Scherer and Meuleman (2013): Scherer, K. R., & Meuleman, B. (2013). Human Emotion Experiences Can Be Predicted on Theoretical Grounds: Evidence from Verbal Labeling. *PLOS ONE*, *8*(3), e58166. https://doi.org/10.1371/journal.pone.0058166.

Funding

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

Table of Contents

List of Abbreviations

ALE ..... *Accumulated Local Effects*
CPM ..... *Component Process Model*
DE ..... *Differential Evolution*
EDA ..... *Electrodermal Activity*
EFA ..... *Exploratory Factor Analysis*
EMG ..... *Electromyography*
FL ..... *Featureless Learner*
GAQ ..... *Geneva Appraisal Questionnaire*
GEA ..... *Geneva Emotion Analyst*
GENESE ..... *Geneva Expert System on Emotions*
HRV ..... *Heart Rate Variability*
ICC ..... *Intraclass Correlation*
LASSO ..... *Lasso (Least Absolute Shrinkage and Selection Operator) Regression*
MMCE ..... *Mean Misclassification Error*
OSF ..... *Open Science Framework*
RF ..... *Random Forest*
RSS ..... *Residual Sum of Squares*
SEC ..... *Stimulus Evaluation Check*
SVM ..... *Support Vector Machine*
WGC ..... *Weighted Guess Classifier*

List of Tables

Chapter 1

Chapter 2

Chapter 4

List of Figures

Chapter 1

Chapter 3

Chapter 4

Chapter 5

1 General Introduction

1.1 Introduction

The human language enables us to refer to objects and concepts even when we are lacking a concise definition for them, as Putnam's (1975) semantic theory about the meaning of words describes. This allows us to talk about emotions even though most of us have a rather implicit understanding of what an emotion is without a concrete formalization of the phenomenon. While we do not depend on a profound understanding of emotions in our everyday social interactions, a deeper insight into affective processes and their mechanisms is highly relevant to many fields of research. Whether it is to find out how emotions influence learning or decision making (e.g., Dirkx, 2008), how emotions can be regulated in the context of mental disorders (e.g., Amstadter, 2008) or which role they play in human-computer interaction (e.g., Beale & Peter, 2008), all of these questions seek to understand the emotion process and its regularities on different levels. Varying emotion theories have been introduced since the 19th century. Though these theories deviate from each other, there is a broad consensus today that emotions are multidimensional in the sense that they do not only concern how we think or feel, how we act, or how our body changes physiologically, but that emotions are a complex integration of different components. Disagreement exists about the specific number and identity of the components as well as the order in which they are addressed (for an overview, see chapter 1.2. or Moors, 2009). When trying to empirically study and understand this multi-componential emotion process holistically, one reasonable approach is to analyze the interrelations between each of the components separately and integrate the findings into a global emotion model afterward. Following this rationale, the present thesis investigates two emotion sub-processes, the link between cognition and the subjective feeling of a person (study 1) as well as the relation between cognition and physiology (study 2), by using different predictive modeling approaches. Cognition as the central initiating component of emotions has been proposed by a group of emotion theories that are collectively referred to as *appraisal theories* (e.g., Arnold, 1960; Frijda, 1986; Lazarus, 1991; Ortony, Clore, & Collins, 1988; Scherer, 1984; Smith & Ellsworth, 1985). They assume that a stimulus is evaluated on multiple emotion-relevant dimensions and that the resulting appraisal patterns affect all other engaged components like motor-functions, the autonomous nervous system, motivation, as well as the perceived feeling. This cognitive-focused view of the emotion process builds the theoretical framework of the present thesis. The two studies are aiming to contribute to the understanding

of the multi-componential emotion process by examining whether the assumed links can be substantiated (and consequently the assumptions made by appraisal theories) and by evaluating how the relationship between the components might look like on an algorithmic level. On the methodological side, the thesis demonstrates how different forms of predictive modeling – computational emotion models based on theoretical assumptions and exploratory machine learning models – can be utilized in basic emotion research.

In the following chapters, a more thorough discussion of the theoretical discourse about emotions and their multidimensionality is presented, paying particular attention to the emotion processes proposed by the appraisal theory. In this context, the two analyzed sub-processes will be reviewed as well as the different modeling approaches. Subsequently, the two studies are presented and their results are discussed and combined.

1.2 Emotions as Multi-Componential Processes

Despite the limited tangibility of emotions, first attempts to describe them have been made as early as the 4th century bc by Aristotle. He understood *pathe* (sing. *pathos)*, as he referred to emotions, as the internal responses of a living being to its environment similarly to perception (Schmitter, 2016). Darwin, who engaged in the research of emotions during the 19th century, still considered emotions as passive reflex-like processes (Oatley, Keltner, & Jenkins, 2014). Within the same period, James (1884) developed one of the first profound theories about emotions and their emergence.[1] He viewed them to be embodied processes in the sense that an emotion is the subjective perception of bodily changes that arises in response to the environment. Therefore, he believed that when individuals meet a bear in the woods, they feel fear because they perceive that they tremble and their heart races. The emotion process as described by this theory hence compromises two distinct components – a physiological component and a feeling component that entails what is consciously perceived about the emotion process.

W. James' (1884) physiological theory of emotions faced a lot of criticism. Cannon (1927), for example, noted that a separation of organs from the autonomous nervous system does not alter emotional behavior and also that visceral changes are not specific to any emotions (e.g., heart acceleration occurs in states of both anger and rage). The latter problem of specificity was addressed by Schachter (1964; see also Schachter & Singer, 1962) due to the

[1] A similar theory was simultaneously developed by Lange (1887). Hence, the theory is often referred to as the James-Lange theory.

introduction of an additional cognitive component. He proposed that after the physiological reaction to a stimulus in the form of physical arousal, a cognitive interpretation of the bodily changes in the context of previous experience occurs that then determines which emotion is felt. In regards to the bear scenario, Schachter's (1964) theory implicates that the encounter with the dangerous animal first leads to physical arousal and a subsequent cognitive interpretation. Within this cognitive processing step, the physical arousal might be attributed to the bear. Because the latter represents a potential threat, the perceived arousal might then be labeled with the emotion term *fear*. However, the same physiological arousal could also lead to a totally different emotion when accompanied by a different cognitive attribution (e.g., physical arousal induced by a surprise party might lead to a feeling of joy instead). Thus, Schachter (1964) for the first time introduced cognition as a central element within the emotion elicitation process. His three-componential model holds explanatory power to some degree as it is able to invalidate Cannon‘s (1927) second criticism by explaining why a specific physiological response can be accompanied by different feelings. Schachter and Singer (1962) also found empirical evidence for their assumption in a study in which the artificial induction of arousal by injections of adrenalin was interpreted differently depending on the emotions displayed by a bystander. Both emotion theories, W. James' (1884) and Schachter's (1964), fail to explain though why a physiological response is triggered in the first place – they do not compromise a specific mechanism that determines which kind of stimulus leads to arousal and which stimulus does not (Moors, 2009).

With the introduction of appraisal theories (e.g., Arnold, 1960; Frijda, 1986; Lazarus, 1991; Ortony et al., 1988; Scherer, 1984; Smith & Ellsworth, 1985) the cognitive component was moved to the beginning of the emotion process. This reorganization of the components closed the gap between stimulus and physiology, enabling not only an explanation of why certain stimuli lead to a response but also for the observation that inter-individual and intra-individual differences exist in this context. Appraisal theorists suggest that the stimulus itself is cognitively appraised and that this evaluation affects all subsequent components. Consequently, a stimulus like the bear in the woods might result in physical arousal and the subjective feeling of fear because the bear is appraised as being highly relevant and as an endangerment to the current goals of the individual. When encountering a bear in the zoo though, the same stimulus could lead to a totally different affective response for the same individual as the cognitive evaluation of relevance and goal endangerment could differ in this context.

The primary cognitive component in appraisal theories does not only trigger physiological changes and changes in perceived feeling, but it also affects a motivational component handling action tendencies and action readiness as well as an expression component for expressive and instrumental behavior (Moors, Ellsworth, Scherer, & Frijda, 2013). The emotion elicitation process is hence understood as an integration of five different components. Critics of the cognitive approach to emotions remark that cognition cannot be a necessary condition to emotions (Zajonc, 1980) as empirical studies have demonstrated that affective responses can be elicited even when stimuli are presented subliminally (Kunst-Wilson & Zajonc, 1980). However, appraisal theorists do not necessarily equate cognition with conscious cognition anymore as the appraisals are believed to be processed in an automated and hence subconscious fashion to some extent (e.g., Scherer, 2001).

As each stimulus is assumed to be evaluated on a number of different appraisal criteria in appraisal theory (see chapter 1.3 for a thorough discussion of this topic), the potentially endless number of resulting appraisal patterns also leads to a very large space of different emotion states (e.g., Scherer, 2001). In contrast, affect program theories (e.g., Ekman, 1992; Panksepp, 2005) believe in very few specific emotion categories, also called basic emotions. These emotion categories are connected to distinct neuronal circuits that control specific physiological and behavioral schemes as well as the subjective emotional experience. This theoretical approach differs from the other models as it shifts the focus to the neurobiological basis of the emotion process. Transferred from computational science, Marr (2010) suggests three levels on which an information processing procedure has to be described to fully understand it. While the input and the output of the process of interest are described on the functional level, the algorithmic level is concerned with the mechanisms that translate the input into the output. Lastly, the process can be described on its implementation level by specifying how the mechanisms as well as the input and the output are realized on a physical level. The previously discussed emotion theories mainly focus on the algorithmic level of the emotion process by trying to formalize and describe how a stimulus (i.e., the input of the emotion elicitation process) results in an affective response like a feeling of joy or physical arousal (i.e., the outputs of the emotion process), whereas affect program theories are rather concerned with the implementation level (Moors, 2009). The latter are, however, not fully incompatible with the idea of appraisal theories. Ekman (1992), for example, also believes that appraisals are a trigger of affect programs. The difference rather lies in the subsequent changes in the other components that occur either flexible and continuously with each appraisal evaluation (as in appraisal theory) or in form of fixed schemes controlled by distinct neuronal circuits (as in

affect program theories). Transferred to W. James' (1884) example of the bear encounter, the bear (through appraisal or another mechanism) hence triggers the affect program of fear that can be biologically based or learned through previous experience and that is processed within a distinctive neuronal circuit. The triggered affect program of fear then leads to prototypical changes in the other components such as an increase in heart rate, the subjective perception of fear, and behavioral changes that prepare flight.

In strong contrast to affect program theorists who assume that basic emotions have a specific neurobiological embedding, Russell (2003) believes that basic emotion categories are mere folk concepts that have no use in the scientific description of emotions. His emotion model introduces a new component called *core affect*. Core affect is defined as a neurophysiological state that integrates the two dimensions pleasantness and arousal. Similar to appraisal theories, an endless number of emotional states can be embedded in this two-dimensional space of core affect. Even though core affect is not directed at any specific stimuli in the environment, a specification of the core affect can take place by a cognitive interpretation. Hence, a cognitive appraisal component is also included in Russell's (2003) model, but it is no longer a precondition for emotions. Rather than identifying a single component as the central element of emotion differentiation, he views emotions to be a collection of potentially independent components that can be labeled with a prototypical emotion term when consciously observed by the individual. This means that when a component pattern occurs that matches a prototypical emotion episode built on previous experience, the pattern is determined to be an instance of this category. Therefore, fixed patterns for different emotions do not exist, but the emotion categories are constructed by the individual. A similar constructivist theory of emotions has been proposed by Barrett (2006). When surprised by a bear in the woods, these theories would assume that the core affect of the individual encountering the animal would shift – probably to a state of higher arousal. If the states of all components together (such as core affect, physiology, cognition, and behavior) are recognized as being similar to an emotion episode prototypical for the constructed emotion of fear, the episode is labeled accordingly as being fear.

The comparison of different emotion theories demonstrates quite clearly that to this day no uniform model for emotions exists. Even though there is a high agreement that the emotion process compromises a set of different components, theories differ when it comes to the exact number and identity of the relevant components, the order of the components within the emotion elicitation process, and the way changes occur in the included components (flexible or controlled by fixed emotion schemes). Another central question that has become prevalent in

the discourse during the cognitivist revolution of psychology in the 1960s is which role cognition plays in the emotion process (Scarantino & de Sousa, 2018). Most of the discussed theories (except for W. James', 1884) acknowledge that a cognitive component is somehow involved in the emotion process, but whether cognition is a necessary condition for emotions and hence the primary element of the emotion process is debated. Nonetheless, the agreements as well as dissimilarities between the different emotion theories can guide emotion research – an area of research that has previously been described as a "very confused and confusing field of study" (p. 2) by Ortony et al. (1988). For faster scientific progress, Moors (2009) has called for a shift of focus from superficial theoretical disagreements to those that are more fundamental. Following this recommendation, the present work aims to analyze the crucial question of the role of cognition within the multi-componential emotional process.

1.3 Cognition and Emotion in Appraisal Theory

When approaching the question which role cognition plays in emotion, a working hypothesis or rather a model to be tested is needed. In terms of model validation, which is usually understood as the process of determining how well a model represents the real world (Sornette et al., 2007), concrete and strong model assumptions are needed. Appraisal theories of emotions do not only hold comprehensive explanatory power, as demonstrated in the last chapter, but many of them also make very specific claims about parts of the emotion elicitation process. Naturally, this does not mean that less formalized or vague theories cannot be true but falsifying their assumptions becomes harder. One of the most prominent appraisal theories, the *Component Process Model* (CPM), was developed by Scherer (1984, 2001, 2009). His model compromises a very precise description of the suspected appraisal process as well as assumptions about interactions of cognition with other components. Based on its strong formalization and the resulting validation characteristics, the CPM was chosen as the theoretical basis for the present thesis.

As other appraisal theorists, Scherer (2001, 2009) comprehends emotions as an integration of five sub-components: A cognitive component that regulates the appraisal process; a physiological component connected to efferent changes in the autonomous nervous system such as respiratory or cardiovascular changes; an expression component controlling motor expressions such as gestures, mimic, and voice; a motivational component that can initiate action tendencies; as well as a feeling component that compromises the subjective perception and potentially the verbal labeling of an emotion. Specifically, he defines an emotion to be an episode in which synchronized and interrelated changes occur in all (or most) of these assumed

subsystems that are triggered by the evaluation of an external stimulus as being highly relevant to major goals and concerns of the individual. During this crucial evaluation process, that is controlled by the cognitive component, the stimulus is appraised on several different dimensions that Scherer (2001) calls *stimulus evaluation checks* (hereafter, these checks will be referred to as appraisal dimensions). His proposed 16 appraisal dimensions are further subdivided into four classes of information that determine how relevant an event is for the individual (relevance detection), which consequences an event has and how these will affect the individual (implication assessment), how well the individual can cope with potential consequences (coping potential determination), and how important the event is in regards to the individual self-concept and social norms (normative significance). The outcomes of these dimensions are believed to be highly subjective and to be depending exclusively on the individuals' personal perception of the stimulus. In contrast to some other appraisal theories (e.g., Lazarus, 1991) that assume the outcome of appraisals to be partly categorical, Scherer (2001) postulates that appraisals are evaluated on a continuous scale with a potentially infinite value range. He further claims that the appraisal process is iterative and that the 16 dimensions are appraised in a specific order. See Table 1 for a short description of all appraisal dimensions in their assumed order of occurrence. The proposed sequentiality of appraisals, which is unique to Scherer's (2001) appraisal theory, is thought to ensure the economy of the cognitive appraisal process. He assumes that all appraisals incorporated in relevance detection, such as *suddenness*, *pleasantness*, and *goal/need importance*, are rather low-level and hence fast mechanisms that fall back on attention, memory, as well as motivational processes. Appraisals appearing later on in the process are thought to be more complex cognitive evaluations that are consequently costlier and require functions like reasoning and the evaluation of self-image. The first appraisals determining the relevance of a stimulus to the individual, therefore, act as a filter that decides whether further expensive processing of the stimulus is needed. Only when a certain threshold is surpassed with these appraisals, additional processing through other appraisals is initiated.

Like other appraisal theorists, Scherer (2001) also assumes that appraisals can be processed in an unconscious and automatic fashion. He differs between three processing levels on which each appraisal can be evaluated. There is a sensory-motor level at which the appraisal mechanisms are mainly genetic and based on functions like pattern matching. There is a schematic processing level where the appraisal evaluation falls back on learned schemes. While both of the previous levels are believed to function automatically, stimuli appraised on the third level, the conceptual level, are processed via highly cortical and propositional-symbolic

Table 1

Description of the 16 Appraisal Dimensions as Proposed by the CPM in the Assumed Order of their Occurrence (Scherer, 2001)

Appraisal Objective	Appraisal Dimension	Appraisal Description
Relevance detection	Suddenness	Abruptness of a stimulus
	Familiarity	Degree of familiarity of a stimulus
	Predictability	Predictability of the occurrence of a stimulus
	Intrinsic pleasantness	Pleasantness of a stimulus independent of the momentary state of the individual
	Goal/need importance	Relevance of a stimulus for the momentary hierarchy of goals and needs
Implication assessment	Cause: Agent	Causal attribution of an event to an agent
	Cause: Motive	Inferences about motives or intentions of an agent
	Outcome probability	Likelihood with which certain consequences are expected
	Discrepancy from expectation	Degree to which a stimulus is consistent with the individual's expectations
	Conduciveness	Conduciveness of a stimulus to help reach current goals
	Urgency	Urgency of adaptive actions in response to a stimulus
Coping potential determination	Control	Extent to which a stimulus can be controlled by animate agents
	Power	Power of the individual to exert control or to recruit other individuals to help
	Adjustment	Ability to adjust and cope with the consequences of a stimulus
Normative Significance	Internal standards	Extent to which a stimulus exceeds internal standards such as self-image or personal moral code
	External standards	Compatibility of a stimulus with norms of a salient reference group in terms of desirability and obligatory conduct

mechanisms (i.e., logic-based reasoning in the broadest sense) that require consciousness. Each of the 16 appraisal dimensions can hence be processed on all three levels which are believed to continuously interact and thereby induce top-down and bottom-up effects (Scherer, 2009).

Scherer (2001) outlines the interaction of all five emotion subsystems in his *componential patterning theory*. According to the latter, all emotion components are highly interrelated and multidirectional. As shown in Figure 1, the cognitive appraisal is the initiator of changes in all other subsystems though. This means that the outcome of every single appraisal leads to variations in all other components and modifies changes induced by previous appraisal evaluations. Scherer (2001) illustrates this process with the following example: The detection of a novel stimulus will produce an orientation response such as a heart rate and skin conduction increase in the physiological component, postural changes in the expression component, changes in goal priority assignment in the motivational subsystem, and an increase in alertness and attention in the feeling component. Only milliseconds after these adaptions, the *intrinsic pleasantness* appraisal determines the evaluated stimulus to be unpleasant. Following this appraisal outcome, a stronger heart rate increase in the physiological component occurs as a defense response, a tendency of avoidance is initiated in the motivational subsystem, motor behavior to turn the individual's body away from the unpleasant stimulus is prepared, and a negative feeling is perceived. Similarly, all subsequent appraisal dimensions will continuously alter the other four subcomponents (i.e., physiology, motivation, motor expressions, and subjective feeling). Consequently, an emotion such as fear, that is defined by a specific pattern of component states, can only occur when preceded by a distinct appraisal pattern. As changes in the non-cognitive components are thought to feed back into cognitive elements that are accessed during the appraisal procedure (i.e., attention, memory, reasoning, and self-image), reciprocal relationships between the cognitive component and the non-cognitive components are assumed (see Figure 1). The appraisal procedure is, however, the initiating component of an emotional episode and the primary cause of changes in other components.

As Scherer (2001) regards emotions to be a stream of continuous changes in different subcomponents, he rejects the idea of a limited number of distinct emotions connected to fixed affect programs as assumed by Ekman (1992) or Panksepp (2005). Instead, a potentially huge number of different emotion states results from the combination of the 16 appraisal dimensions. Scherer (2001) acknowledges, however, that some appraisal patterns might form more frequently than others. He refers to these more common and prototypical emotion patterns, which are those for which specific verbal labels exist, as modal emotions (i.e., enjoyment/

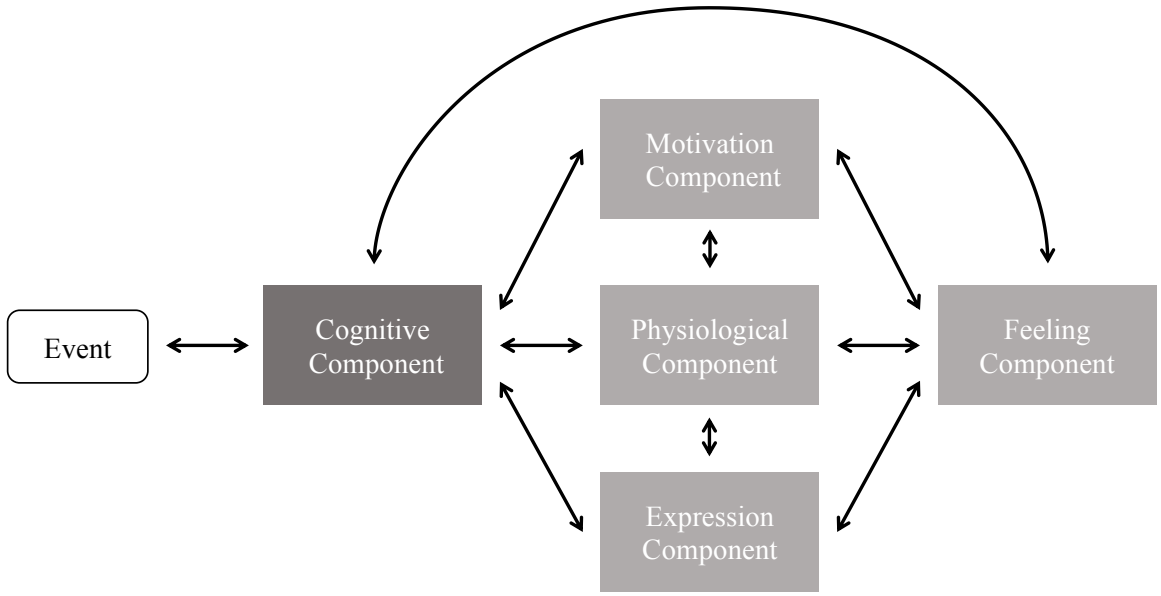

Figure 1. Components of the CPM and their interactions. Adapted from Scherer (2009).

happiness, elation/joy, displeasure/disgust, contempt/scorn, despair, sadness/dejection, anxiety/ worry, fear, irritation/cold anger, rage/hot anger, boredom/indifference, shame, guilt, pride).

Hence, the appraisal process is the main element of the multi-componential emotion process that differentiates between different emotions and initiates all changes in other components. From this appraisal hypothesis, it can be derived that the changes in other components such as the subjective feeling or physiological responses should be predictable from the appraisal patterns or, conversely, that the appraisal patterns should be predictable from respective changes in other components.[2] The present thesis uses these assumptions to investigate the appraisal hypothesis by modeling the link between appraisal and the subjective feeling as well as appraisal and physiology using two different predictive modeling approaches.

1.4 Modeling the Multi-Componential Emotion Process

In the following, the two relations of interest will be discussed – the appraisal-feeling link analyzed in study 1 (link A) of this thesis as well as the appraisal-physiology link analyzed in study 2 (link B). Figure 2 shows the previously discussed CPM model where the two

[2] Note that the CPM (as well as other appraisal theories) imply a causality (appraisals patterns initially cause changes in other components) that cannot be validated with the design used in the current thesis. As the design of both presented studies is non-directional, both findings (appraisals predict changes in other components vs. changes in other components predict appraisals) can be used to substantiate cognitive theories of emotions.

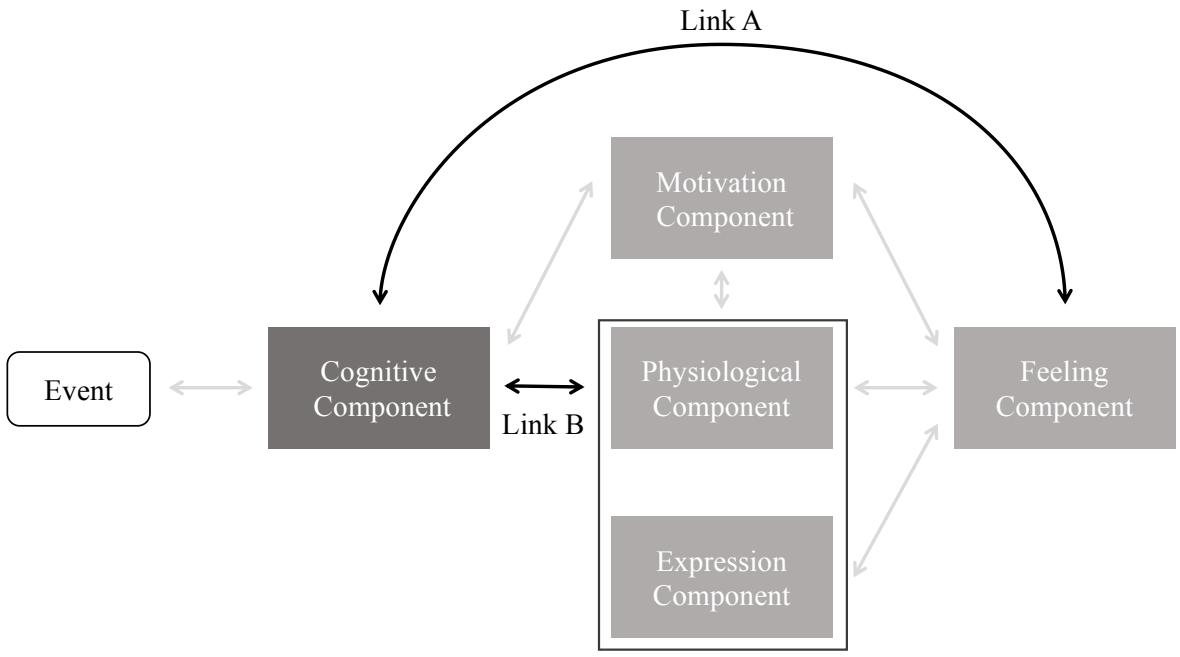

Figure 2. The CPM with the links between components that are studied in study 1 (link A) and study 2 (link B).

examined interrelations are identified. Based on the number of theoretical assumptions and the degree of formalization of the two links, two different predictive modeling approaches will be introduced – theoretically informed modeling used in study 1 (modeling approach A) as well as uninformed modeling with machine learning algorithms applied in study 2 (modeling approach B).

1.4.1 Link A: The Appraisal-Feeling Link

A logical implication of the component patterning theory is that modal emotions (i.e., the prototypical emotion states that can be verbally labeled by a person) should be predictable from appraisal patterns. In 1981, Scherer for the first time introduced prototypical appraisal patterns that he assumed to be connected to different modal emotions. These theoretical prototypes have since been elaborated and refined based on his and his colleague's research (Scherer, 1984, 2001; Scherer & Meuleman, 2013). In Table 2, Scherer's (2001) prototypes (i.e., appraisal values that are thought to lead to the outcome of the respective modal emotion) are exemplarily presented for the emotions fear and happiness. The prototypes indicate, for example, that an experienced emotion labeled with the word *fear* follows an event that is appraised to be unfamiliar (*familiarity* = low) and unpleasant (*intrinsic pleasantness* = low), and that is moreover obstructive to one's current goals (*conduciveness* = obstruct). An emotion

labeled with the term *happiness*, on the other hand, has been appraised as very pleasant (*intrinsic pleasantness* = high) and in line with current goals and needs (*conduciveness* = consonant). The proposed prototypical appraisal outcomes also include *open* parameters. The happiness prototype, for example, has an *open* value for the dimension *familiarity*. As Scherer (2001) explains, this means that the modal emotion is compatible with all potential outcomes of this specific appraisal. Hence, the respective appraisal is not relevant for the modal emotion as it cannot be used to differentiate the emotion category from others. In relation to the example of the happiness prototype, this means that happiness can arise from an event that is appraised as very familiar or very unfamiliar to the individual.

Besides the emotion prototypes, Scherer (2001) also makes assumptions about the algorithmic level of the appraisal-to-feeling process. He indicates that the proposed appraisal dimensions are not equally important in the prediction of the emotion prototypes, but that some

Table 2

Prototypical Appraisal Outcomes for the Modal Emotions Fear and Happiness as Proposed by Scherer (2001)

Appraisal Dimension	Fear	Happiness
Suddenness	High	Low
Familiarity	Low	Open
Predictability	Low	Medium
Intrinsic pleasantness	Low	High
Goal/need importance	High	Medium
Cause: Agent	Oth/nat	Open
Cause: Motive	Open	Intent
Outcome probability	High	Very High
Discrepancy from expectation	Dissonant	Consonant
Conduciveness	Obstruct	High
Urgency	Very High	Very Low
Control	Open	Open
Power	Very Low	Open
Adjustment	Low	High
Internal standards	Open	Open
External standards	Open	Open

Note: oth = other, nat = natural, intent = intentional.

dimensions contribute more strongly to the emotion differentiation process. During the appraisal-to-feeling calculation, the appraisal dimensions are thought to be integrated through a predetermined weighting function. As with the emotion prototypes, Scherer and Meuleman (2013) also introduced theoretically derived appraisal weights that reflect the assumed importance of each appraisal in the emotion differentiation process. Another implication about the algorithmic level is the assumed sequentiality and temporal order of the appraisals in which more expensive appraisals are processed after fast and less costly appraisals.

1.4.2 Modeling Approach A: Theoretically Informed Modeling

The CPM (Scherer, 2001, 2009) provides elaborated assumptions about the appraisal-feeling link. When strong hypotheses (i.e., model assumptions) are given, a theoretically informed model can be applied. The general idea of such a modeling approach is to formalize and implement a verbal theory into a computational model that operates on the respective inputs, generates the respective outputs, and uses the theoretically assumed algorithms to transform the input into the output. Following this logic in study 1, the assumptions about the appraisal-feeling link made by the theory were implemented in four computational models that produce emotion categories (i.e., emotion labels) in response to appraisal patterns in the way that is assumed by the theory. Based on an empirical data set in which appraisal patterns as well as verbal emotion labels were assessed via self-report in response to an emotional episode experienced in the past, the implemented theoretical models were used to predict emotion terms from the empirically assessed appraisal patterns. These predictions were subsequently compared to the emotion labels given by the participants (i.e., a ground truth), assessing the predictive accuracy of the models. The predictive performance can then be used as a measure for the validity of the theoretical assumptions realized in the models. If the model assumptions are true, the models should be able to predict the empirically assessed emotion labels correctly to some degree.[3] If the theory underlying the model is incorrect or imprecise, the predictive power should be low. Using a theoretically informed modeling approach to analyze the appraisal-feeling link, therefore, allows validating the concrete theoretical assumptions made by the CPM.

[3] As both the models' input (i.e., the appraisal patterns) as well as their ground truth (i.e., the emotion labels) were assessed by questionnaire, measurement error is most likely present in both variables which consequently afflicts the models' accuracy. Therefore, even if the implemented model assumptions are correct, a perfect performance can never be reached.

Theoretical modeling frequently goes beyond the validation of theoretical assumptions by extending and refining what is implied by a theory. The latter is due to the fact that most verbal theories lack the needed formalization for a mathematical implementation, irrespective of their specificity (Marsella, Gratch, & Petta, 2010). Hence, the theoretical modeling process can reveal hidden assumptions and complexities as well as gaps in the theoretical framework (Marsella et al., 2010). With respect to the appraisal-feeling link, input and output of the analyzed process are clearly defined but how the appraisal patterns are exactly transformed into the emotion label outcomes is not – except for the weighting and the order of the different appraisal dimensions. Consequently, this information gap in the algorithmic level of the theory has to be closed. Following the hypothesis that each modal emotion is connected to a distinct prototypical appraisal pattern, the appraisal-feeling relation in study 1 was realized as a decision rule based on a weighted distance metric between a new appraisal pattern and prototypical appraisal patterns associated with different modal emotions. Based on the calculated distances to all prototypical emotions, the models predict the label of the emotion prototype with the smallest distance (i.e., the highest similarity) to the input appraisal pattern. Hence, each appraisal pattern can be pictured as a point in a 16-dimensional space in which its proximity to other patterns can be determined. Visualizing this concept, Figure 3 shows a two-dimensional scaling of the 13 emotion prototypes used in study 1 as well as an empirically assessed appraisal pattern (INP) from the used data set. The preferred computational model in study 1 predicted the emotion label *fear* for this appraisal pattern as it showed the lowest distance, and hence the highest similarity, to the fear emotion prototype. The implementation of the appraisal-feeling link based on distance measures to emotion prototypes has been done before by Scherer (1993) and Scherer and Meuleman (2013).

Another advantage of theoretical models is that their internal structure can be varied and different model implementations realizing different model assumptions can hence be contrasted with respect to their predictive accuracy and validity. Therefore, we varied the weighting algorithms within the described weighted distance decision rule to test different weighting functions against each other. Four models were implemented to examine whether no differential weighting of the appraisal dimensions (as it has been realized in an expert system by Scherer, 1993), the theoretical weighting parameters for the 16 appraisal dimensions proposed by Scherer and Meuleman (2013), 16 weighting parameters generated from the empirical data set using a genetic optimization method, or a more complex weighting algorithm with 208 parameters also generated with an optimization approach yielded the best out-of-sample performance. As we also generated the emotion prototypes from the empirical data set (and

contrasted them with the theoretical prototypes from Scherer, 2001), the modeling approach of study 1 can be described as a hybrid of theoretically informed and exploratory data-driven methods.

Lastly, it has to be noted that the models in study 1 are mere structural models of the appraisal-feeling link proposed by the CPM (Scherer, 2001, 2009) that do not regard the assumed temporal characteristics of the appraisal process (i.e., the temporal order of the appraisal dimensions). The distance calculation from empirical appraisal patterns to emotion prototypes does not hold any temporal constraints. In the context of a simple accuracy assessment of the models, the temporal dimension of the appraisal procedure has no relevance. However, an app is provided in chapter 3 that visualizes the temporal changes of prototype similarity for any potential appraisal pattern if the assumed temporal order is taken into account.

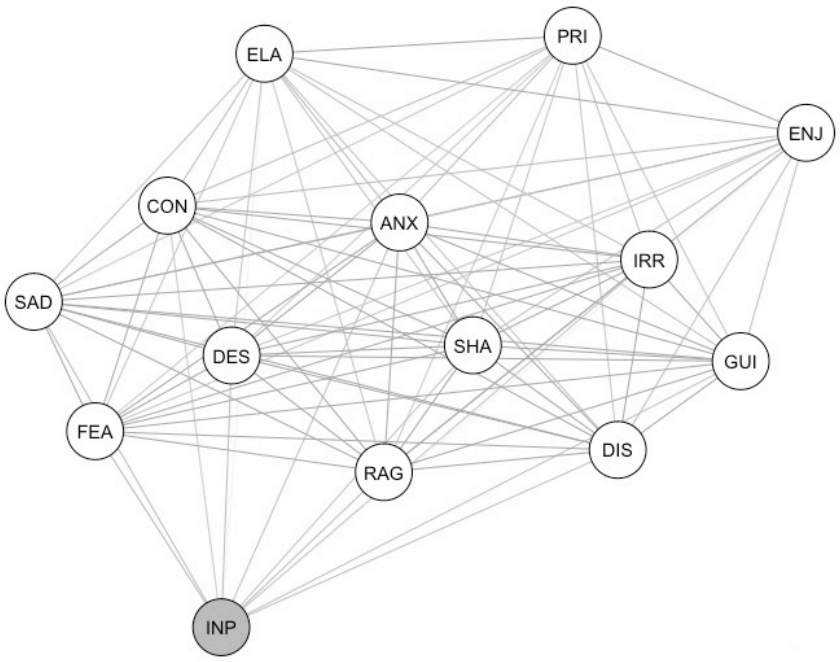

Figure 3. Two-dimensional scaling of the 13 emotion prototypes used in study 1 (SAD = sadness, FEA = fear, CON = contempt, DES = despair, RAG = rage, SHA = shame, DIS = disgust, GUI = guilt, IRR = irritation; ANX = anxiety, ELA = elation, ENJ = enjoyment, PRI = pride) as well as an empirically assessed appraisal input pattern (INP) that was identified as fear by the preferred model in study 1. Note that this is a force embedded layout in which not all distances are displayed spatially correct.

1.4.3 Link B: The Appraisal-Physiology Link

In contrast to the appraisal-feeling link, the theoretical basis concerning the relation between the appraisal component and the physiological component is rather sparse. Again, the input of the appraisal-to-physiology process (i.e., the appraisal dimensions) is clearly defined. Concerning the output of the process, it is possible to imagine a very large set of physiological variables that are potentially affected by the appraisal outcomes during an emotional episode such as cardiovascular, respiratory, electrodermal, muscular or intestinal responses. For ten appraisal dimensions, Scherer (2009) proposes theoretically derived responses in the physiological component connected to these appraisals (in Table 3, an excerpt from these predictions for the appraisal dimensions *pleasantness* and *conduciveness* is presented). As these predictions are derived from theoretical considerations (no detailed information on how they were developed is given), they have to be viewed as very uncertain and rather speculative. For the other six appraisal dimensions, no information is provided on how the appraisal-physiology link might look like. Empirical research on the effect of appraisals on certain physiological signals has been conducted partially. Most of the studies in this field have some serious shortcomings though, including very small sample sizes, the application of rather restricted and outdated statistical methods, or experimental designs in which only specific appraisals were able to be examined (see study 2 in chapter 4 for a more thorough discussion of the empirical research conducted in this field).

The theoretical predictions by Scherer (2009) are also very limited as they only refer to two possible appraisal manifestations – a high and a low evaluation of the respective appraisal (pleasant vs. unpleasant and conducive vs. obstructive for the appraisals presented in Table 3). As the appraisal dimensions are assumed to be continuous though, there is no information about the effect of different outcomes or continuous changes of the appraisals on the physiology of an individual. Generally, no assumptions about the algorithmic level of the appraisal-to-physiology process are given that describe how an appraisal pattern is translated to changes in the physiological component. The only assumption is that changes in appraisals should result in continuous changes in the physiological component (in contrast to affect program theories which assume that hard-wired physiological patterns occur when an emotion is triggered; see Ekman, 1992; Panksepp, 2005). How these changes take place, whether certain appraisals are more important for the induction of physiological responses or by what type of function the appraisal outcome is translated to the physiological component (e.g., linear, polynomial or exponential) is not defined.

Table 3

Extract from Scherer's (2009) Proposed Effects of High and Low Intrinsic Pleasantness and Low and High Conduciveness Evaluations on the Physiological Component

Appraisal Dimension	Appraisal Evaluation	Proposed Physiological Outcome
Intrinsic pleasantness	Pleasant	Sensitization, inhalation, heart rate deceleration, salvation, pupillary dilatation, lids up, open mouth and nostrils, lips part und corners pulled upwards, gaze directed …
	Unpleasant	Defense response, heart rate acceleration, increase in skin conductance level, decrease in salvation, pupillary constriction, slight muscle tonus increase, brow lowering, lid tightening, nose wrinkling, upper lip raising, lip corner depression, chin raise, lip press, nostril compression, tongue thrust, gaze aversion …
Conduciveness	Conducive	Trophotropic shift, decrease in respiration rate, slight heart rate decrease, bronchial constriction, increase in gastrointestinal motility, relaxation of sphincters, decrease in general muscle tonus, relaxation of facial muscle tone …
	Obstructive	Ergotropic shift, preparation for action, corticosteroid and catecholamine, particularly adrenaline secretion, deeper and faster respiration, increase in heart rate and heart stroke volume, vasoconstriction in skin, gastrointestinal tract and sexual organs …

1.4.4 Modeling Approach B: Machine Learning

As discussed in the previous chapter, the theoretical framework for the appraisal-physiology link is less profound than the assumptions made for the appraisal-feeling path. A theoretical modeling approach as used in study 1 is therefore not applicable. In contexts like these, more exploratory analyses can be used to generate new information for theory development. Therefore, an exploratory machine learning approach was applied for the analysis of the appraisal-physiology path in study 2. Instead of theoretically determining the relation of

interest (as in study 1), the used machine learning algorithms are able to acquire the relationship between input and output autonomously. As we do not have strong assumptions on how the appraisals relate to physiological variables, this approach allows generating the algorithmic level of the appraisal-to-physiology process from empirical data. Depending on the machine learning algorithm employed, complex interactions of a large number of predictors and non-linearities can be reflected. Due to their complexity, machine learning algorithms often have high predictive power. On the downside, the high model complexity often leads to reduced comprehensibility and interpretability which is why many of these models are also identified as black-box models. Nevertheless, different methods have been introduced over the years summarized under the term *interpretable machine learning* that allow approximating aspects of the learned model structure (for an overview, see Molnar, 2019).

In study 2, different physiological channels (electromyography, skin conductance, and heart rate variability) were assessed in response to emotional video sequences. As in study 1, the appraisal dimensions were assessed retrospectively (but immediately after the evaluated event) via self-report. 134 features characterizing the different physiological signals were calculated. Subsequently, different machine learning models (a lasso regression model, a random forest, as well as a support vector machine) were trained to learn the relations between the physiological features (input) and the assessed appraisal dimensions (output). By examining whether the appraisals can be predicted from the physiological features, the appraisal-physiology link can be verified. If the appraisals are connected to the considered physiological signals, a sufficiently complex model should be able to predict the appraisals to some degree.[4] By using different types of methods for interpretable machine learning and by comparing the performance of different machine learning models (i.e., linear and non-linear algorithms), it can be further examined how the algorithmic link between appraisal and physiology might look like.

Even though the CPM implies that the appraisals initiate the changes in the physiological component, study 2 models this relation reversed by using the physiological signals to predict the appraisal dimensions. Due to the non-directional experimental design in study 2, the causality of the appraisal-physiology link cannot be tested. Hence, the relation was modeled conversely, as this approach has several advantages. Because one single feature cannot

[4] As in study 1, the presence of measurement error in the self-reported appraisals as well as in the physiological features has to be considered. This means that a perfect predictive performance is very unlikely even when the link between an appraisal and the features exists and a very high model complexity is given.

exhaustively describe a physiological channel (e.g., an electromyographic signal can be described by different amplitude and frequency measures that potentially assess different aspects of the time signal), each physiological signal has to be described by a broad set of different features. Therefore, when using the appraisals to predict changes in physiology, different models would have to be trained for each of the 134 features. This procedure would strongly increase the number of analyzed models which would consequently complicate the interpretation of the results and proliferate the computational costs. Moreover, the modeled relationship between appraisals and physiology would have to be interpreted individually for each of the 134 models. The reversed modeling though, using the physiological features to predict the appraisal dimensions, allowed the construction of several blocked importance measures that quantify the relevance of all features belonging to a physiological channel (e.g., all skin conductance features) in the prediction. Hence, the aggregated effect of the physiological channels can be investigated which is much more informative from a practical standpoint.

In the following, the three machine learning algorithms used in study 2 are presented.

1.4.4.1 Lasso Regression

The *lasso* (least absolute shrinkage and selection operator) regression, as outlined by G. James, Witten, Hastie, and Tibshirani (2013), is a regularized linear model that performs a variable selection by shrinking the regression coefficients of predictors that explain little variance to zero. The variable selection (i.e., shrinkage of coefficients) reduces variance and prevents from overfitting the model to the data. Consequently, the out-of-sample performance of the model can be improved – most notably in models with a large number of variables. To achieve the latter, the estimation function of the linear model is extended by a penalty term that is determined by the tuning parameter λ (i.e., penalty weight) and the number and absolute height of the β-coefficients in the model. This estimation function, where n is the number of samples and p the number of variables, is minimized to find the β-coefficients of the lasso model:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p}|\beta_j| \quad (1)$$

To determine the best value for the tuning parameter λ, a grid of λ values is chosen and the cross-validation error for each of the assigned values of λ is calculated. Subsequently, the value is selected for which the cross-validation error is smallest (G. James et al., 2013).

1.4.4.2 Random Forest

As described by G. James et al. (2013), the *random forest* is a tree-based machine learning algorithm that is able to represent complex interactions and non-linearities. It uses recursive binary splitting to grow large decision-trees on n_{tree} training samples. The training samples are built by bootstrapping which means that n_{obs} observations are randomly drawn from the original data set with replacement (where n_{obs} is the number of observations in the original data set). During the tree-building process, each time a split is made a random sample of m predictors is chosen from the whole set of p predictors. The size m of the considered subset is usually defined to be $m = \sqrt{p}$. From the random subset only one variable is picked, namely, the one that splits the predictor space in a way that leads to the greatest possible reduction of residual sum of squares (RSS) in the resulting regions (i.e., terminal nodes or leaves). In this manner, the predictor space is further divided into different regions until a minimum number of observations in each region is reached. For a new test observation, each tree predicts the mean across all training observations that are assigned to the same region. The predictions of all n_{tree} trees are subsequently averaged. Note that the procedure differs when classification instead of regression trees are applied. In this case, the predictors and splits are chosen based on the mean misclassification error (MMCE). Instead of averaging the observations, the most frequent class in each region is predicted. Lastly, a majority vote over all n_{tree} trees is returned.

1.4.4.3 Support Vector Machine

Like the random forest, the *support vector machine* is a machine learning algorithm that can be applied to classification and regression problems. As it is only applied as a regression model in the present thesis, only this application context will be addressed. As described by Smola and Schölkopf (2004), the β-coefficients of a linear function are minimized in support vector regressions (or more specifically, the Euclidean norm of the β-coefficients is minimized). In this estimation process, a margin of tolerance is established and only deviations larger than ε (i.e., margin tolerance parameter) are considered in the estimation function. In addition, a penalty term is added to the estimation function that is determined by a constant $\lambda > 0$ and the slack variables ξ_i and ξ_i^* which indicate the residuals of the observations y_i from the tolerance margin (where ξ_i is a positive deviation from the margin and ξ_i^* is a negative). The constant λ hence defines the trade-off between the flatness of the linear function and the strength of deviation from the tolerance margin:

$$\frac{1}{2}\|\beta\|^2 + \lambda \sum_{i=1}^{n}(\xi_i + \xi_i^*) \quad (2)$$

By using different types of kernel functions (e.g., polynomial or radial basis functions) the dimensionality of the feature space can be increased so that non-linear relations can be modeled with the support vector machine as well (Smola & Schölkopf, 2004).

1.5 References

Amstadter, A. (2008). Emotion regulation and anxiety disorders. *Journal of Anxiety Disorders*, *22*(2), 211–221. https://doi.org/10.1016/j.janxdis.2007.02.004

Arnold, M. B. (1960). *Emotion and personality*. New York: Columbia University Press.

Barrett, L. F. (2006). Solving the Emotion Paradox: Categorization and the Experience of Emotion. *Personality and Social Psychology Review*, *10*(1), 20–46. https://doi.org/10.1207/s15327957pspr1001_2

Beale, R., & Peter, C. (2008). The Role of Affect and Emotion in HCI. In C. Peter & R. Beale (Eds.), *Affect and Emotion in Human-Computer Interaction: From Theory to Applications* (pp. 1–11). https://doi.org/10.1007/978-3-540-85099-1_1

Cannon, W. B. (1927). The James-Lange Theory of Emotions: A Critical Examination and an Alternative Theory. *The American Journal of Psychology*, *39*(1/4), 106–124. https://doi.org/10.2307/1415404

Dirkx, J. M. (2008). The meaning and role of emotions in adult learning. *New Directions for Adult and Continuing Education*, *120*, 7–18. https://doi.org/10.1002/ace.311

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, *6*(3–4), 169–200. https://doi.org/10.1080/02699939208411068

Frijda, N. H. (1986). *The emotions*. Cambridge: Cambridge University Press.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. https://doi.org/10.1007/978-1-4614-7138-7

James, W. (1884). WHAT IS AN EMOTION? *Mind*, *9*(34), 188–205. https://doi.org/10.1093/mind/os-IX.34.188

Kunst-Wilson, W., & Zajonc, R. (1980). Affective discrimination of stimuli that cannot be recognized. *Science*, *207*(4430), 557–558. https://doi.org/10.1126/science.7352271

Lange, C. G. (1887). *Über Gemütsbewegungen (Org. Om Sindsbevægelser)*. Leipzig: Thomas Theodor.

Lazarus, R. S. (1991). *Emotion and Adaptation*. New York: Oxford University Press.

Marr, D. (2010). *Vision: A computational investigation into the human representation and processing of visual information*. Cambridge, Mass: MIT Press.

Marsella, S., Gratch, J., & Petta, P. (2010). Computational Models of Emotion. In K. R. Scherer, T. Bänzinger, & E. B. Roesch (Eds.), *A blueprint for an affectively competent agent: Cross-fertilization between Emotion Psychology, Affective Neuroscience, and Affective Computing* (pp. 21–41). Oxford: Oxford University Press.

Molnar, C. (2019). Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. Retrieved from https://christophm.github.io/interpretable-ml-book/

Moors, A. (2009). Theories of emotion causation: A review. *Cognition & Emotion*, *23*(4), 625–662. https://doi.org/10.1080/02699930802645739

Moors, A., Ellsworth, P. C., Scherer, K. R., & Frijda, N. H. (2013). Appraisal Theories of Emotion: State of the Art and Future Development. *Emotion Review*, *5*(2), 119–124. https://doi.org/10.1177/1754073912468165

Oatley, K., Keltner, D., & Jenkins, J. M. (2014). *Understanding emotions* (Third edition). Hoboken, NJ: Wiley.

Ortony, A., Clore, G., & Collins, A. (1988). *The Cognitive Structure of Emotions*. Cambridge: Cambridge University Press.

Panksepp, J. (2005). *Affective neuroscience: The foundations of human and animal emotions*. Oxford: Oxford University Press.

Putnam, H. (1975). *The Meaning of "Meaning."* Retrieved from http://hdl.handle.net/11299/185225

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, *110*(1), 145–172. https://doi.org/10.1037/0033-295X.110.1.145

Scarantino, A., & de Sousa, R. (2018). Emotion. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2018). Retrieved from https://plato.stanford.edu/archives/win2018/entries/emotion/

Schachter, S. (1964). The Interaction of Cognitive and Physiological Determinants of Emotional State. In *Advances in Experimental Social Psychology* (Vol. 1, pp. 49–80). https://doi.org/10.1016/S0065-2601(08)60048-9

Schachter, S., & Singer, J. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological Review*, *69*(5), 379–399. https://doi.org/10.1037/h0046234

Scherer, K. R. (1981). Wider die Vernachlässigung der Emotion in der Psychologie [On the neglect of emotion in psychology]. In W. Michaelis (Ed.), *Bericht über den 32. Kongress*

der Deutschen Gesellschaft für Psychologie in Zürich 1980 (pp. 304–317). Göttingen: Hogrefe.

Scherer, K. R. (1984). On the nature and function of emotion: A component process approach. In K. R. Scherer & P. Ekman (Eds.), *Approaches to Emotion* (pp. 293–317). Hillsdale, NJ: Erlbaum.

Scherer, K. R. (1993). Studying the emotion-antecedent appraisal process: An expert system approach. *Cognition & Emotion*, *7*(3–4), 325–355. https://doi.org/10.1080/02699939308409192

Scherer, K. R. (2001). Appraisal considered as a process of multilevel sequential checking. In K. R. Scherer, A. Schorr, & J. Johnstone (Eds.), *Appraisal processes in emotion* (pp. 92–120). New York: Oxford University Press.

Scherer, K. R. (2009). The dynamic architecture of emotion: Evidence for the component process model. *Cognition & Emotion*, *23*(7), 1307–1351. https://doi.org/10.1080/02699930902928969

Scherer, K. R., & Meuleman, B. (2013). Human Emotion Experiences Can Be Predicted on Theoretical Grounds: Evidence from Verbal Labeling. *PLOS ONE*, *8*(3), e58166. https://doi.org/10.1371/journal.pone.0058166

Schmitter, A. M. (2016). 17th and 18th Century Theories of Emotions. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2016). Retrieved from https://plato.stanford.edu/archives/win2016/entries/emotions-17th18th/

Smith, C. A., & Ellsworth, P. C. (1985). Patterns of cognitive appraisal in emotion. *Journal of Personality and Social Psychology*, *48*(4), 813–838. https://doi.org/10.1037/0022-3514.48.4.813

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, *14*(3), 199–222. https://doi.org/10.1023/B:STCO.0000035301.49549.88

Sornette, D., Davis, A. B., Ide, K., Vixie, K. R., Pisarenko, V., & Kamm, J. R. (2007). Algorithm for model validation: Theory and applications. *Proceedings of the National Academy of Sciences*, *104*(16), 6562–6567. https://doi.org/10.1073/pnas.0611677104

Zajonc, R. (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist*, *35*(2), 151–175. https://doi.org/10.1037/0003-066X.35.2.151

2 Study 1: Emotion Prediction with Weighted Appraisal Models

This paper is reprinted from 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*. https://doi.org/10.1109/TAFFC.2019.2940937 and was funded by a grant of the German Research Foundation to Felix Schönbrodt (DFG SCHO 1334/4-1). The data set used in the present study was provided by Scherer and Meuleman (2013).

2.1 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 predict emotion labels based on prototype similarity calculations. Different weighting algorithms, mapping the models' input to a distinct emotion label, were integrated into 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.

2.2 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 (Marsella, Gratch, & Petta, 2010). Computational affect modeling 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 – appraisal theories of emotion (for an overview see Moors, Ellsworth, Scherer, & Frijda, 2013), specifically, the *Component Process Model* (CPM) by Scherer (1984, 2001, 2009).

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 (Moors, 2009). 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 (1884) defines emotion as the perception of bodily changes that arises as a response to the environment. This strict exclusion of the cognitive component in the emotion causation process has since been challenged. Schachter and Singer (1962), for example, expanded James' (1884) 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 by apprehending the cognitive evaluation of a stimulus as the trigger of emotions, influencing all of the other components (e.g., Roseman, 2001; Scherer, 2001; Smith & Lazarus, 1990; Smith & Ellsworth, 1985). 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 (Moors et al., 2013). An emotion or emotion family can then be described as a function of a distinct appraisal pattern – several of these appraisal profiles for specific emotions have been proposed in the literature (Frijda, 1986; Roseman, 1984; Scherer, 2001; Smith & Ellsworth, 1985). Consequently, an emotion is not supposed to be elicited by the stimulus itself (contrary to the theory of James, 1884) but by its meaning for the individual (Moors, 2010). 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 (Moors et al., 2013). Several empirical studies have been conducted to test the theoretical predictions made by appraisal

theories (for a review, see Scherer, 2009), 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 that map appraisal profiles either onto distinct emotions labels (e.g., AR by Elliott, 1992) or dimensional representations of affect (e.g., WASABI by Becker-Asano, 2008). Some of those adapted the appraisal profiles proposed by Scherer (2001; e.g., PEACTIDM by Marinier, Laird, & Lewis, 2009), while others built on the work of Ortony, Clore, and Collins (1988; e.g., AR by Elliott, 1992). 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 (1993b). 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 that represent 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 had demand characteristics and hence could 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; Scherer & Meuleman, 2013), was introduced. GEA asks users to label the reported emotion episode before the system's 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 participant's 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 (Douven, 2017). 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 modeling 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 (Douven, 2017). This entails some uncertainty as not all inferences necessarily need to be valid, but it allows us 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 CPM. 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 kinds of weighting algorithms in the appraisal-feeling 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 weights 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.

2.3 The Component Process Model (CPM)

Scherer's (2001) theory, 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 (for a detailed description of the 16 appraisal dimensions, see Scherer, 2001). 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 (2001), therefore, rejects the assumption of a limited number of discrete emotion categories made by many other emotion theorists (e.g., Ekman, 1992). Nonetheless, he recognizes that certain appraisal combinations occur more frequently and universally than others. Scherer (2001) calls these states, that are usually labeled 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 (Scherer, 1984, 2001; Scherer & Meuleman, 2013), also include open parameters, indicating that a specific dimension might be irrelevant or that many different values are compatible with the respective modal emotion (Sander, Grandjean, & Scherer, 2005). Overall, the theoretical prototypes show moderate correlations to appraisal means found in empirical data (Scherer & Meuleman, 2013). 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 (Sander et al., 2005). For this weighting algorithm, theoretically derived parameters have been proposed as well (Scherer & Meuleman, 2013).

2.4 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, and 13 emotion prototypes within a 16-dimensional appraisal space. They then return the emotion label of the prototype that 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) included 16 parameters (one for each appraisal dimension) similar to the GEA system. This weighting algorithm implies that across all emotions some appraisal dimensions could be generally more important in the identification of an emotion than others (e.g., the valence of a stimulus could be more important than its familiarity). 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, Smith and Ellsworth (1985) 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 explicitly expressed in the CPM, does not contradict Scherer's (2001) 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 [2001] 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 (Scherer & Meuleman, 2013), 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 baseline model that randomly guesses classes weighted by their frequency in the data set (*weighted guess classifier*; WGC) as well as to a *random forest* (RF) 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 cannot be fully deduced from theoretical assumptions about the appraisal process. Therefore, we decided to derive the prototypes directly from an empirical data set that was collected with the GEA system by Scherer and Meuleman (2013). Prototype theory, first introduced by Rosch in 1983, defines the prototype of a category as a reference point for classification based on representativeness. 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 modeling 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.

2.5 Method

Our electronic appendix, including all corresponding R scripts and further supporting information, is provided via our Open Science Framework (OSF) repository at https://osf.io/te4z3/.

2.5.1 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 (2013) was used. The data was collected via the freely accessible GEA system on the website of the Swiss Center for Affective Sciences[5] over the duration of eight years. The questionnaire implemented in the GEA system is publicly available as the *Geneva Appraisal Questionnaire* (GAQ; Geneva Emotion Research Group, 2002) and was specifically developed to assess the results of an appraisal process during

[5] https://www.unige.ch/cisa/research/materials-and-online-research/online-research/

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 proposed by Scherer (2001). 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 labeled 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 were 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.

2.5.2 Data Preprocessing

For the further use in our emotion models, we aggregated the 25 appraisal items to the 16 appraisal dimensions proposed by the CPM (Scherer, 2001) by calculating mean values 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 (Sander et al., 2005), 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 labeled 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 package* by Mahto, 2018). 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 (Lunardon, Menardi, & Torelli, 2014). The oversampling as well as all further analyses and implementations were conducted in R (Version 3.4.2; R Core Team, 2018). Using the first emotion label as class label again, we randomly sampled instances from the data set with the *upSample* function from the *caret* package (Kuhn, 2008) 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.

2.5.3 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 dimensions separately. They thus give different importance to the dimensions 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 ith emotion prototype,

p_{ij} is the prototype value of the jth appraisal dimension of the ith emotion prototype,

e_j is the empirical value of the jth appraisal dimension,

w_j is the appraisal weight given to the jth appraisal dimension,

w_{ij} is the appraisal weight given to the jth appraisal dimension of the ith 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.

2.5.4 Estimation of Model Parameters

2.5.4.1 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 labeled with the according emotion term were aggregated. Episodes labeled 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 across all episodes labeled 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:

$$p_{ij} = \frac{\sum_{k=1}^{n_i} r_{ijk}}{n_i} \quad (4)$$

where,

p_{ij} is the prototype value of the jth appraisal dimension of the ith emotion prototype,

r_{ijk} is the kth rating of the jth appraisal dimension that was labeled with the ith emotion class,

n_i is the number of episodes labeled with the ith 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 (2001). 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.

2.5.4.2 Theoretical Appraisal Importance

The weighting parameters (w_j) for model M2 were derived from the theoretical weights used by Scherer and Meuleman (2013). The authors actually present a numerical weighting parameter for each of the items used in the GAQ. 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.

2.5.4.3 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 across 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 (1997). DE is a global optimization algorithm suited for high-dimensional, non-linear problems that do not require an either continuous or differentiable function. Like 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 (Ardia, Mullen, Peterson, & Ulrich, 2016). The parallel search within a whole population of parameter configurations helps to avoid local minima which makes DE superior to many direct search methods (Storn & Price, 1997). To conduct the optimization, the *DEoptim* package (Ardia et al., 2016) 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 (Ardia et al., 2016), was set to 0.9. Storn and Price (1997) 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. (2016) 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 solutions 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 across 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 a robustness measure. Additionally, we wanted to contrast the optimized parameters of M3 to the theoretical weights by Scherer and Meuleman (2013) 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.

2.5.4.4 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, the predicted emotion class 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 classes 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 true positive examples over all positive labeled examples) to assess the models' performance (Bekkar, Djemaa, & Alitouche, 2013).[6]

To contrast the models' classification performance with a naive baseline model, we also reported the performance of the WGC that randomly predicts classes dependent on their relative frequency in the data set. As another benchmark, we conducted a RF classification using the 16 appraisal dimensions as features.[7] We chose the ranger learner from the *ranger* package (Wright & Ziegler, 2017) with hyperparameters set to default. The model computation was conducted within the *mlr* framework by Bischl et al. (2016). 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 RF 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 labeled 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 (2013) 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 (Bella, Hernandez-Orallo, & Ramirez-Quintana, 2009). We therefore calculated two-way intraclass correlations (*ICC*)

[6] 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.

[7] 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 & Scherer, 2013). The results of this benchmark experiment can be found in the electronic appendix.

between the real class proportions in the data and the class proportions in the predictions of the models.

2.6 Results

2.6.1 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 = .47 to the appraisal values of the prototypes proposed by Scherer (2001; see Table 1).

Table 1

Pearson Correlations of the Appraisal Dimensions of the Prototypes Calculated from the Data Set and the Theoretical Prototypes from Scherer (2001)

Emotion Prototype	*r*
Pleasure	.44
Joy	.56
Disgust	.48
Sadness	.57
Despair	.64
Anxiety	.57
Fear	.73
Irritation	.34
Rage	.60
Shame	.06
Guilt	.07
Pride	.42
Contempt	.31

2.6.2 Emotion Classification

The WGC 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 2 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 WGC. 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 (2013), 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 WGC baseline model. Nevertheless, M2 was outperformed by the unweighted M1, which reached higher scores in all classes except for despair, irritation, and contempt as well as a higher overall accuracy.

The DE 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)[8] 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 for all emotions except for rage and irritation.

The DE 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 WGC, M1, M2, as well as M3. But the class-wise precision scores show that M4 actually yielded worse precisions than the simpler

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

M3 in all classes except 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 RF showed an overall better performance than all other models. The class-wise precision scores ranged from 14.8% for contempt to 78.0% for joy. The RF outperformed M1 and M3 for 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.

Table 2

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 the WGC model are equivalent to those of a random model without weighting of class frequencies.

.

Pearson's correlations 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: $r(11) = .83$, $p < .001$; M2: $r(11) = .92$, $p < .001$; M3: $r(11) = .87$, $p < .001$; M4: $r(11) = .86$, $p < .001$) as well as for the RF ($r(11) = .78$, $p = .002$).

2.6.3 Emotion Family Classification

In the classification of the four emotion families, the naive WGC showed an overall accuracy of 43.6% 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 3 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

Table 3

Percentage Precision Scores of the Four Emotion Families 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 the WGC model are equivalent to those of a random model without weighting of class frequencies.

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 for which both models performed equally well. In the three other classes, M4 reached also lower precision scores than the unweighted M1.

Finally, the RF 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 RF also yielded higher precisions for the happiness, anger, and the shame/guilt family, but was surpassed by M1 and M3 for the disgust family.

2.6.4 Model Calibration

With an *ICC* of .317 (p = .134, *CI* [-.259, .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 -.129 (p = .67, *CI* [-.619, .433]. With an *ICC* of .411 (p = .072, *CI* [-.156, .774]), M3 yielded a slightly higher calibration than M1. M4 reached a moderate *ICC* of .705 (p = .002, *CI* [.277, .900]). The RF classifier showed an even higher *ICC* of .808 (p < .001, *CI* [.484, .937]). Naturally, the model with the highest *ICC* was the WGC, reproducing the class probability distribution of the data set perfectly with an *ICC* of .997 (p < .001, *CI* [.989, .999]).

2.6.5 Appraisal Weights

Table 4 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 (2013) was modest ($r(14)$ = .30, p = .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.

2.7 Discussion

In the present study, we used a predictive modeling approach to validate and extend 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 were implemented in the

Table 4

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

four models to assess their plausibility by comparing their performances. To generate new information, parameters within these models, including the emotion prototypes and the weighting parameters (for M3 and M4), were generated from empirical data and contrasted with theoretical assumptions from the literature.

All four theoretical models performed notably better than the baseline model (WGC), 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 set 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 skewness. 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" labeled 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 additional 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 (2013) 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 led 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 theoretically derived parameters of M2; consequently, 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 DE optimizations, showed a low mean variance which indicates good stability of the optimization results and suggests 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 of whether a stimulus is likely to result in pleasure or pain (Sander et al., 2005), is also included in other appraisal theories (e.g., Frijda, 1986; Smith & Ellsworth, 1985). The very importance of pleasantness in the emergence of emotions is also reflected in other emotion models such as Russell's (2003) theory of *core affect*. He describes affect 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 electronic 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 et al. (2005) 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 (2000) links urgency to the dimension of activation or arousal, which has been identified as the second of two relevant dimensions by Russell (2003). 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 (Fontaine, Scherer, Roesch,

& Ellsworth, 2007). The attained weighting can also be compared to other instantiations of appraisal models. Lazarus' (1991) *cognitive-motivational-relational theory*, 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., Frijda, 1986; Smith & Ellsworth, 1985). Besides the four dimensions included in the CPM, Lazarus' model additionally contains the dimensions *goal content* and *future expectation*. The former appraisal, which is also included in the appraisal theory of Roseman (1984), 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 (Ortony et al., 1988), 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 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 the number and identity of dimensions between appraisal theories are 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 (Moors, 2009; Scherer, 1999). Furthermore, theorists differ in their view on parsimonious modeling, where some try to include only sufficient or typical appraisals, while others focus on completeness (Moors, 2009; Scherer, 1999). When comparing the present results to other appraisal theories, it is also important to remark that most theories do not make particular assumptions on how the appraisals are aggregated during the emotion emergence process (i.e., they do not 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 dimension 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%. However, the precision analysis 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 labeled 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 overall accuracy, higher precision rates for most classes, and a better calibration, M3 can be preferred over the unweighted M1 model and M2 with the theoretically 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 (Sander et al., 2005; Scherer & Meuleman, 2013) leads to a considerable improvement. The big gap in performance between M2 and M3 suggest though that the 16 theoretical weighting parameters do not 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 overall accuracy, the precision rates drop due to 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 WGC, 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 the CPM. 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 Ellsworth and Smith (1988) that reported differing appraisal importances between emotion classes.

We additionally included the RF 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 RF 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 regard to our computational emotion models, this means that there is 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 labels. Particularly the usage of self-report for the measurement of appraisals has been criticized (e.g., Davidson, 1992), 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 (Sander et al., 2005). It can be questioned whether appraisal dimensions driven by more basal cognitive functions are actually consciously accessible and consequently, whether these constructs can actually be measured adequately using subjective self-reports. Many theorists recognize this limitation of self-assessed appraisals (Frijda, 1993; Lazarus, 1991; Scherer, 1993a). Scherer (1993a) 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-awareness, is common to many fields of psychology. In the past, studies have tried to detect physiological markers of different appraisal dimensions (for an overview, see Scherer, 2009), 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 (Scherer & Meuleman, 2013). Scherer (1993a) expresses his hope that the 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. The main problem of relying 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 by using more reliable measurement techniques when they are made available.

In summary, the computational modeling approach used in the present study lends some support to psychological appraisal theories of emotions 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 suggests 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, which 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 appraisals 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 the appraisal process. Also, due to its many advantages, the application of computational emotion modeling as a way of validating and extending hypotheses generated based on empirical research or theory should be integrated more strongly in the theory development process.

2.8 References

Ardia, D., Mullen, K. M., Peterson, B. G., & Ulrich, J. (2016). *DEoptim: Differential Evolution in R*. Retrieved from https://CRAN.R-project.org/package=DEoptim

Becker-Asano, C. (2008). *WASABI: Affect simulation for Agents with Believable Interactivity.* (Doctoral dissertation). Faculty of Technology, University of Bielefeld, Bielefeld.

Bekkar, M., Djemaa, H. K., & Alitouche, T. A. (2013). Evaluation Measures for Models Assessment over Imbalanced Data Sets. *Journal of Information Engineering and Applications*, *3*(10), 27–38.

Bella, A., Hernandez-Orallo, J., & Ramirez-Quintana, M. J. (2009). Calibration of machine learning models. In E. S. Olivas, J. D. M. Guerrero, M. Martinez-Sober, J. R. Magdalena-Benedito, & A. J. Serrano López (Eds.), *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques* (pp. 128–146). Hershey, PA, USA: IGI Global.

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., … Jones, Z. M. (2016). mlr: Machine Learning in R. *Journal of Machine Learning Research*, *17*(170), 1–5.

Davidson, R. J. (1992). Prolegomenon to the structure of emotion: Gleanings from neuropsychology. *Cognition and Emotion*, *6*(3–4), 245–268. https://doi.org/10.1080/02699939208411071

Douven, I. (2017). Abduction. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Retrieved from https://plato.stanford.edu/archives/sum2017/entries/abduction/

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, *6*(3–4), 169–200. https://doi.org/10.1080/02699939208411068

Elliott, C. D. (1992). *The Affective Reasoner: A process model of emotions in a multi-agent system* (Doctoral dissertation). Institute for Learning Sciences, Northwestern University, Evanston, Illinois.

Ellsworth, P. C., & Smith, C. A. (1988). Shades of Joy: Patterns of Appraisal Differentiating Pleasant Emotions. *Cognition & Emotion*, *2*(4), 301–331. https://doi.org/10.1080/02699938808412702

Fontaine, J. R. J., Scherer, K. R., Roesch, E. B., & Ellsworth, P. C. (2007). The World of Emotions is not Two-Dimensional. *Psychological Science*, *18*(12), 1050–1057. https://doi.org/10.1111/j.1467-9280.2007.02024.x

Frijda, N. H. (1986). *The emotions*. Cambridge: Cambridge University Press.

Frijda, N. H. (1993). The place of appraisal in emotion. *Cognition & Emotion*, *7*(3–4), 357–387. https://doi.org/10.1080/02699939308409193

Geneva Emotion Research Group. (2002). *Geneva Appraisal Questionnaire (GAQ)*. Retrieved from https://www.unige.ch/cisa/files/3414/6658/8818/GAQ_English_0.pdf

James, W. (1884). WHAT IS AN EMOTION? *Mind*, *9*(34), 188–205. https://doi.org/10.1093/mind/os-IX.34.188

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, *28*(5), 1–26.

Lazarus, R. S. (1991). *Emotion and Adaptation*. New York: Oxford University Press.

Lunardon, N., Menardi, G., & Torelli, N. (2014). ROSE: a package for binary imbalanced learning. *R Journal*, *6*(1), 79–89.

Mahto, A. (2018). *splitstackshape: Stack and Reshape Datasets After Splitting Concatenated Values*. Retrieved from https://CRAN.R-project.org/package=splitstackshape

Marinier, R. P., Laird, J. E., & Lewis, R. L. (2009). A computational unification of cognitive behavior and emotion. *Cognitive Systems Research*, *10*(1), 48–69. https://doi.org/10.1016/j.cogsys.2008.03.004

Marsella, S., Gratch, J., & Petta, P. (2010). Computational Models of Emotion. In K. R. Scherer, T. Bänzinger, & E. B. Roesch (Eds.), *A blueprint for an affectively competent agent: Cross-fertilization between Emotion Psychology, Affective Neuroscience, and Affective Computing* (pp. 21–41). Oxford: Oxford University Press.

Meuleman, B., & Scherer, K. R. (2013). Nonlinear Appraisal Modeling: An Application of Machine Learning to the Study of Emotion Production. *IEEE Transactions on Affective Computing*, *4*(4), 398–411. https://doi.org/10.1109/T-AFFC.2013.25

Moors, A. (2009). Theories of emotion causation: A review. *Cognition & Emotion*, *23*(4), 625–662. https://doi.org/10.1080/02699930802645739

Moors, A. (2010). Automatic Constructive Appraisal as a Candidate Cause of Emotion. *Emotion Review*, *2*(2), 139–156. https://doi.org/10.1177/1754073909351755

Moors, A., Ellsworth, P. C., Scherer, K. R., & Frijda, N. H. (2013). Appraisal Theories of Emotion: State of the Art and Future Development. *Emotion Review*, *5*(2), 119–124. https://doi.org/10.1177/1754073912468165

Ortony, A., Clore, G., & Collins, A. (1988). *The Cognitive Structure of Emotions*. Cambridge: Cambridge University Press.

R Core Team. (2018). *R: A Language and Environment for Statistical Computing*. Retrieved from https://www.R-project.org/

Rosch, E. (1983). Prototype classification and logical classification: The two systems. In E. Scholnick (Ed.), *New trends in conceptual representation: Challenges to Piaget's theory* (pp. 73–86). Hillsdale, NJ: Erlbaum.

Roseman, I. J. (1984). Cognitive Determinants of Emotion: A Structural Theory. *Personality and Social Psychology Review*, *5*, 11–36.

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

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, *110*(1), 145–172. https://doi.org/10.1037/0033-295X.110.1.145

Sander, D., Grandjean, D., & Scherer, K. R. (2005). A systems approach to appraisal mechanisms in emotion. *Neural Networks*, *18*(4), 317–352. https://doi.org/10.1016/j.neunet.2005.03.001

Schachter, S., & Singer, J. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological Review*, *69*(5), 379–399. https://doi.org/10.1037/h0046234

Scherer, K. R. (1984). On the nature and function of emotion: A component process approach. In K. R. Scherer & P. Ekman (Eds.), *Approaches to Emotion* (pp. 293–317). Hillsdale, NJ: Erlbaum.

Scherer, K. R. (1993a). Neuroscience projections to current debates in emotion psychology. *Cognition & Emotion*, *7*(1), 1–41. https://doi.org/10.1080/02699939308409174

Scherer, K. R. (1993b). Studying the emotion-antecedent appraisal process: An expert system approach. *Cognition & Emotion*, *7*(3–4), 325–355. https://doi.org/10.1080/02699939308409192

Scherer, K. R. (1999). Appraisal Theory. In T. Dalgleish & M. J. Power (Eds.), *Handbook of Cognition and Emotion* (pp. 637–663). https://doi.org/10.1002/0470013494.ch30

Scherer, K. R. (2000). Psychological Models of Emotion. In J. Borod (Ed.), *The Neuropsychology of Emotion* (pp. 137–162). Oxford and New York: Oxford University Press.

Scherer, K. R. (2001). Appraisal considered as a process of multilevel sequential checking. In K. R. Scherer, A. Schorr, & J. Johnstone (Eds.), *Appraisal processes in emotion* (pp. 92–120). New York: Oxford University Press.

Scherer, K. R. (2009). The dynamic architecture of emotion: Evidence for the component process model. *Cognition & Emotion*, *23*(7), 1307–1351. https://doi.org/10.1080/02699930902928969

Scherer, K. R., & Meuleman, B. (2013). Human Emotion Experiences Can Be Predicted on Theoretical Grounds: Evidence from Verbal Labeling. *PLOS ONE*, *8*(3), e58166. https://doi.org/10.1371/journal.pone.0058166

Smith, C. A., & Ellsworth, P. C. (1985). Patterns of cognitive appraisal in emotion. *Journal of Personality and Social Psychology*, *48*(4), 813–838. https://doi.org/10.1037/0022-3514.48.4.813

Smith, C. A., & Lazarus, R. S. (1990). Emotion and Adaptation. In L. A. Pervin (Ed.), *Handbook of personality: Theory and Research* (pp. 609–637). New York: The Guilford Press.

Storn, R., & Price, K. (1997). Differential Evolution – A Simple and Efficient Heuristic for global Optimization over Continuous Spaces. *Journal of Global Optimization*, *11*(4), 341–359.

Wright, M. N., & Ziegler, A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, *77*(1), 1–17. https://doi.org/10.18637/jss.v077.i01

3 The APPraisal App

3.1 The App

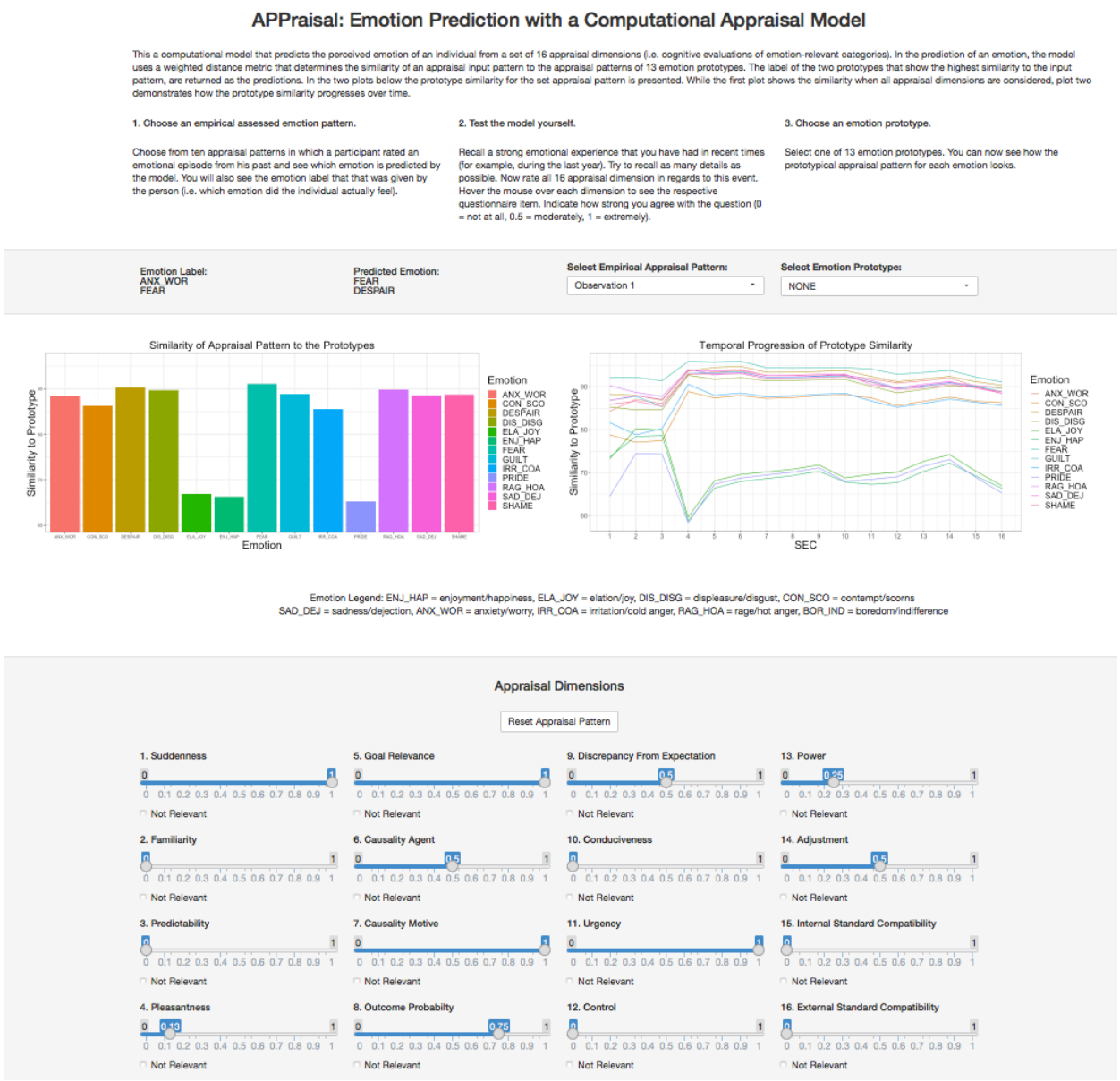

Figure 1. Screenshot of the APPraisal app interface.

Figure 1 shows a screenshot of the APPraisal app interface. The app can be accessed at https://laura-israel.shinyapps.io/appraisal/. APPraisal was built using the *shiny* package by Chang, Cheng, Allaire, Xie, and McPherson (2019) which provides a developing framework for web applications in R (R Core Team, 2018).

3.2 Concept and Description

The idea of this app is to provide a tool for the application of the preferred model M3 (i.e., the model with the 16 optimized appraisal weights) of study 1. The app enables to predict emotions from entered appraisal ratings and flexibly visualizes the model's outcome and its predictions. For this purpose, the app features an interface where the 16 appraisal dimensions used in study 1 can be rated. For each dimension, an exemplary item from the GAQ (Geneva Emotion Research Group, 2002) is presented to the user (when the mouse hovers over the respective appraisal slider). As feedback to the rated appraisals, two emotion terms are predicted by the model (i.e., the two emotion categories whose prototypes show the highest similarity/smallest distance to the input appraisal pattern).

Besides the two predicted emotion labels, the app also provides two visualizations. The first graphic shows the similarity of the entered appraisal pattern to all 13 emotion prototypes, thereby illustrating the logic of the prototype approach of study 1 (i.e., the prototype that is most similar to the input pattern is predicted by the model). See study 1 for detailed information on the prototype calculation. The second graphic demonstrates how the similarity to the 13 prototypes progresses and changes when the appraisals are processed sequentially in the order that is assumed by Scherer (2001, 2009). As discussed in chapter 1.4.2, the distance metric implemented in the model M3 does not hold any temporal constraints but calculates the distance for all appraisal dimensions simultaneously. The model is therefore not a process model but a structural model of the appraisal process. In the second graphic of the app though, the process is visualized by plotting the prototype similarity of the appraisal input pattern for each of the 16 appraisal dimensions successively. At each appraisal step, only the present appraisal and all previous dimensions are included in the distance calculation (e.g., at step one, the similarity between the input and the prototypes is calculated only for the *suddenness* dimension, while the similarity determination in step two is based on the dimensions *suddenness* and *familiarity*. For that purpose, the similarity formula of M3 (see chapter 2.5.3) was extended by the appraisal step index *n*:

$$s_{in} = 1 - \frac{\sum_{j \notin Q}^{n} (w_j (p_{ij} - e_j))^2}{\sum_{j \notin Q}^{n} w_j^2} \quad (1)$$

where,

s_{in} is the similarity to the ith emotion prototype at the nth appraisal step,

p_{ij} is the prototype value of the jth appraisal dimension of the ith emotion prototype,

e_j is the empirical value of the jth 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.

Besides the option to rate the appraisal dimensions and test the model themselves, the app also provides two additional functions. First, the user can select one of ten empirical appraisal patterns. These appraisal patterns are randomly sampled observations from the data set of Scherer and Meuleman (2013) used in study 1. When an empirical observation is selected, the appraisal interface is updated with the respective appraisal ratings and the model predicts two emotion labels. In addition, the user can also see the true emotion labels given by the participant, which allows contrasting the emotion label with the predictions of the model. The second function allows to select from 13 emotion prototypes. If an emotion prototype is selected, the appraisal interface is again updated with the prototypical appraisal pattern (calculated from the empirical data in study 1) for the respective emotion category. This feature can hence be used to examine the similarity between the prototypes, as the app then visualizes the resemblance between the selected prototype and all other prototypes.

3.3 Discussion

Apart from the mere visualization of the model's predictions, the app also provides further insights into the structure of model M3. In Figure 2, a screenshot from the APPraisal app is presented with *Observation 7* selected from the empirical appraisal patterns. The respective appraisal pattern was labeled as enjoyment/happiness and elation/joy by the participant (see *Emotion Label* field). The first graphic of Figure 2 shows that this appraisal pattern has a high similarity to all three positive emotion prototypes (i.e., elation/joy, enjoyment/happiness as well as pride) and a substantially lower similarity to the remaining negative emotion classes (i.e., despair, fear, guilt, shame, displeasure/disgust, contempt/scorn, sadness/dejection, anxiety/worry, irritation/cold anger, rage/hot anger, boredom/indifference). A reversed pattern can be found in Figure 3, which shows a screenshot with *Observation 1* selected as the empirical appraisal pattern. This pattern was labeled with the emotion terms anxiety/worry and fear and again, the first graphic indicates a very high similarity to all emotion prototypes with a negative valence, but a substantially lower similarity to the three positive emotion classes. Hence, the model seems to be very good in differentiating between positive and negative emotion classes (i.e., in the differentiation of valence) but less powerful in distinguishing emotion categories within these two groups. This observation is in line with the

findings of study 1 that the emotion families were much more predictable than the individual emotion classes, with the happiness family that includes all three positive emotions yielding the highest precision (92.0%). As in Figure 2, all three positive emotions are very similar to the input, the resulting similarity pattern used in the prediction is not clear-cut. Consequently, the first predicted emotion label pride does not match the emotion label given by the participant, but the second label enjoyment/happiness does. As only the first prediction of the model was used in study 1, this outcome would have been labeled as inaccurate in the evaluation of the emotion classification. This clearly demonstrates how the model's performance is affected.

The lack of differentiability between emotion classes has to be explained by the similarity of their prototypes. It can be assumed that all prototypes of emotions with a negative valence (or all emotions with a positive valence) must be very similar to each other. This presumption can be confirmed by the examination of the prototype function in the app. When selecting an arbitrary prototype from the *Emotion Prototype* section, a very high similarity to all prototypes of the same valence is shown. This can be observed for all 13 available emotion prototypes. The temporal progression of the prototype similarity suggests something similar. The second graph in Figures 2 and 3 demonstrates that the temporal progressions of the similarity metric also resemble each other for both negative and positive emotions which indicates that the respective prototypes are in close proximity. The observation of prototype similarity lends support for the assumptions discussed in study 1, that the lack of clarity in the emotion labels available to the participants might have led to an increase of measurement error in the labels which subsequently would have influenced the differentiability of the prototypes. It is also plausible that the inclusion of observations that were labeled with two different emotion labels in the prototype calculation contributed to the assimilation of the prototypes – especially of those that frequently occur together (as emotions of the same emotional valence do).

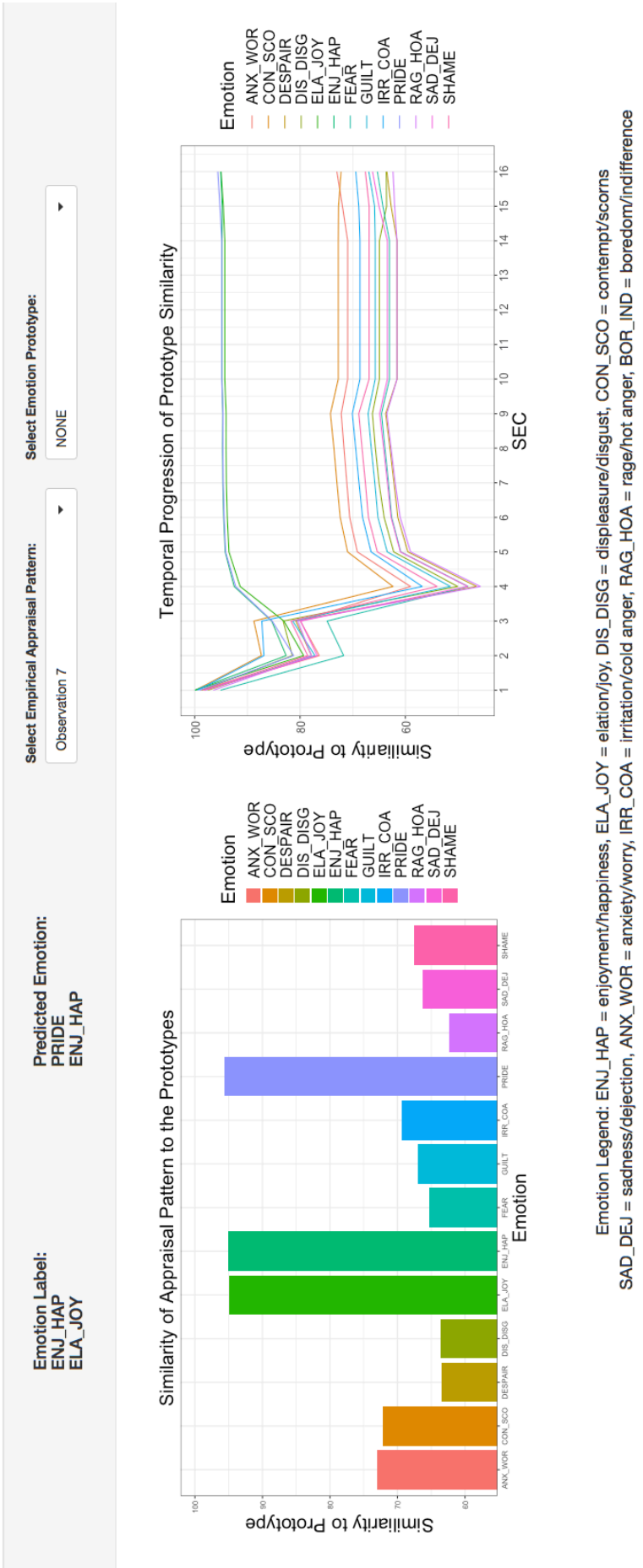

Figure 2. Screenshot from the APPraisal app (https://laura-israel.shinyapps.io/appraisal/) with observation 7 selected from the empirical input patterns.

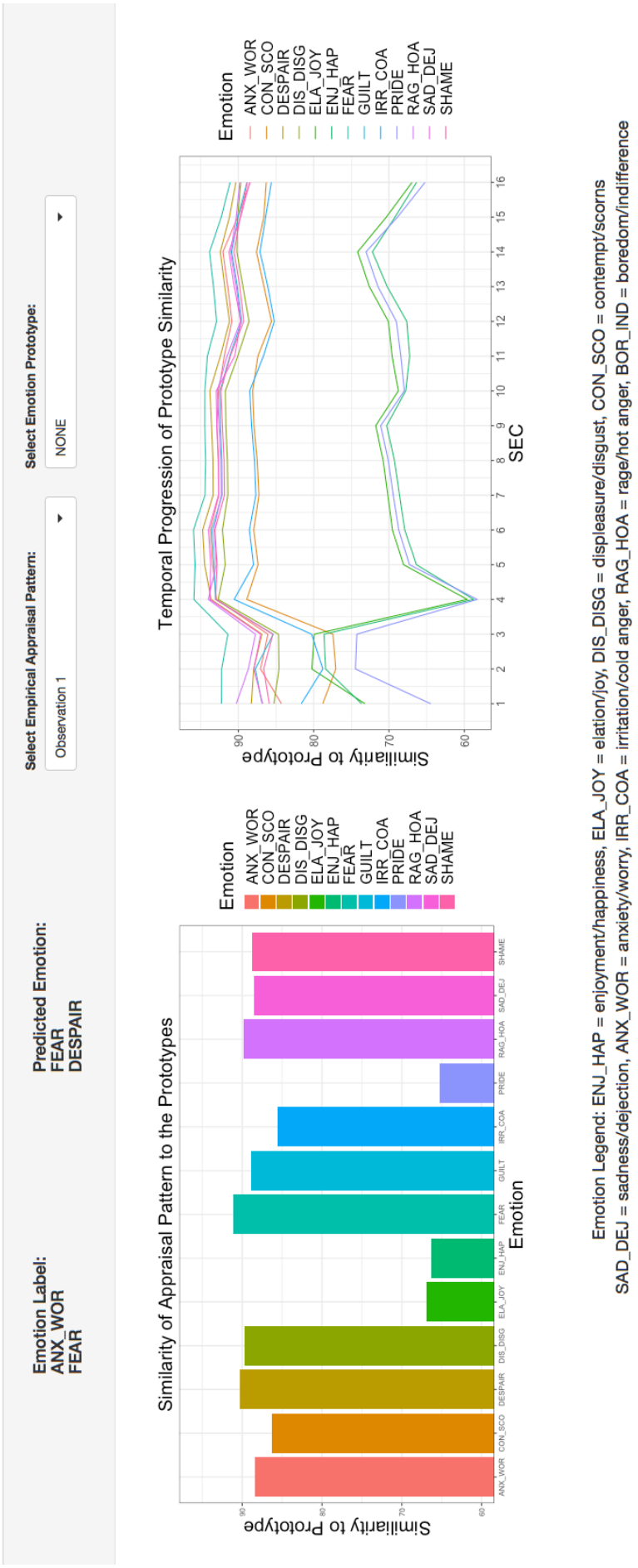

Figure 3. Screenshot from the APPraisal app (https://laura-israel.shinyapps.io/appraisal/) with observation 1 selected from the empirical input patterns.

3.4 References

Chang, W., Cheng, J., Allaire, J. J., Xie, Y., & McPherson, J. (2019). *shiny: Web Application Framework for R*. Retrieved from https://CRAN.R-project.org/package=shiny

Geneva Emotion Research Group. (2002). *Geneva Appraisal Questionnaire (GAQ)*. Retrieved from https://www.unige.ch/cisa/files/3414/6658/8818/GAQ_English_0.pdf

R Core Team. (2018). *R: A Language and Environment for Statistical Computing*. Retrieved from https://www.R-project.org/

Scherer, K. R. (2001). Appraisal considered as a process of multilevel sequential checking. In K. R. Scherer, A. Schorr, & J. Johnstone (Eds.), *Appraisal processes in emotion* (pp. 92–120). New York: Oxford University Press.

Scherer, K. R. (2009). The dynamic architecture of emotion: Evidence for the component process model. *Cognition & Emotion*, *23*(7), 1307–1351. https://doi.org/10.1080/02699930902928969

Scherer, K. R., & Meuleman, B. (2013). Human Emotion Experiences Can Be Predicted on Theoretical Grounds: Evidence from Verbal Labeling. *PLOS ONE*, *8*(3), e58166. https://doi.org/10.1371/journal.pone.0058166

4 Study 2: Predicting Affective Appraisal from Physiology

A slightly altered version of this paper is published as Israel, L. S. F., & Schönbrodt, F. D. (2020). Predicting Affective Appraisals from Facial Expressions and Physiology using Machine Learning. Behavior Research Methods. https://doi.org/10.3758/s13428-020-01435-y. It was funded by a grant of the German Research Foundation to Felix Schönbrodt (DFG SCHO 1334/4-1).

4.1 Abstract

The present study explored the interrelations between a broad set of appraisal ratings and five physiological signals, including facial electromyography, electrodermal activity, and heart rate variability, that were assessed in 157 participants watching 10 emotionally charged videos. 134 features were extracted from the physiological data and a benchmark comparing different kinds of machine learning algorithms was conducted to test how well the appraisal dimensions can be predicted from these features. For 13 out of 21 appraisals, a robust positive R^2 was attained, indicating that the dimensions are actually related to the considered physiological channels. The highest R^2 (.407) was reached for the appraisal dimension *intrinsic pleasantness*. Moreover, the comparison of linear and non-linear algorithms and the inspection of the links between the appraisals to single physiological features using *Accumulated Local Effects* (ALE) plots indicates that the relationship between physiology and appraisals is non-linear. By constructing different importance measures for the assessed physiological channels, we could show that for the 13 predictable appraisals the five channels explained different amounts of variance and that only a few blocks incrementally explained variance beyond the other physiological channels.

4.2 Introduction

The cognitivist revolution during the 1960s, an intellectual movement replacing behaviorism that had dominated psychology in the first half of the 20th century, also lead to new progressions in affective science (Scarantino & de Sousa, 2018). Lead by Arnold (1960) and Lazarus (1966), the emotion formation process, neglected in earlier behavioristic approaches to emotions, came to the focus of research and formed the basis for the new tradition of appraisal theories. These conceive emotions as an evaluative process in which the meaning of a stimulus to the individual is determined – the relevance of a stimulus for one's well-being is appraised in respect to personal values, needs, attachments, and goals (Moors, Ellsworth, Scherer, &

Frijda, 2013). In contrast to other conceptualizations of the emotion process (e.g., Schachter & Singer, 1962), appraisal theorists place this cognitive component at the beginning of an emotional episode, resulting in bodily, motor, and motivational changes and potentially in the subjective perception of a feeling (Moors, 2009). An emotion is hence understood as a multi-componential process, integrating the cognitive appraisal with its subsequent constituents. To understand the complex emergence of emotions, a lot of research has been conducted to learn how these components interact with each other. The main focus has been to understand how specific appraisal patterns map onto the subjective perception of emotions. Prototypical appraisal patterns for different emotion classes have been derived from theoretical assumptions (e.g., Frijda, 1986; Roseman, 1984; Scherer, 2001; Smith & Ellsworth, 1985) as well as from empirical data (e.g., Israel & Schönbrodt, 2019; Meuleman & Scherer, 2013). Another important objective is to specify the link between appraisal and physiology, showing how different appraisal outcomes lead to changes in the motor system or the autonomic nervous system (ANS).

Furthering our knowledge on the connection between cognition and the body in affective states is not only fundamental to understand emotions as a whole but could also help to develop better tools to measure the cognitive appraisal process. To the present day, the majority of research on this topic has to rely on the use of questionnaires (e.g., Meuleman & Scherer, 2013; Scherer, 1993b, 1997; Scherer & Meuleman, 2013). Using this type of assessment, only constant appraisal ratings can be obtained that cannot depict potential changes in appraisal during an emotional situation. Further, the appraisal process is always evaluated in retrospect, often with a large temporal distance to the event of interest (e.g., Geneva Emotion Research Group, 2002), which potentially affects the reliability of the ratings. This demonstrates the need for the development of more indirect continuous measurement tools in the future, which can be realized by studying how physiology relates to self-reported appraisals.

4.3 The Link between Appraisal and Physiology

The *Component Process Model* (CPM) by Scherer (1984, 2001, 2009), one of the best-known realizations of the appraisal theory, assumes 16 different appraisal dimensions. For ten of these dimensions, Scherer (2009) makes elaborate predictions on how they relate to response patterns in the physiological component. He predicts, for example, that in the evaluation of the *intrinsic pleasantness* of a stimulus, a higher pleasantness leads to physiological changes such as heart rate deceleration, pupillary dilatation, and parted lips with pulled up corners, while an unpleasant stimulus should result in an opposite reaction with a heart rate acceleration, pupillary

constriction, and lip corner depression. As these theoretical predictions are rather speculative, different studies have tried to investigate these theoretical links in experimental settings. Van Reekum et al. (2004) induced pleasant and unpleasant as well as goal conducive and goal constructive events in a computer game while measuring several physiological reactions. A higher skin conductance response for pleasant compared to unpleasant events was found, and obstructive events led to higher skin conductance, a stronger increase in heart rate variability, and higher puls transit times compared to conducive events. Aue and Scherer (2008) varied the same two appraisal dimensions in a performance task in which pleasant and unpleasant pictures were presented. During the task, pictures would increase or decrease in size, where an increase of a pleasant stimulus was considered as goal conducive and a decrease of the same picture as goal obstructive (the converse logic was applied to unpleasant pictures). The authors reported an increase in heart rate and higher activity of the zygomaticus major muscle for pleasant as well as higher corrugator muscle activity for unpleasant pictures. Higher zygomaticus response, higher heart rate, and higher skin conductance was found for the conducive conditions and higher corrugator activity for the obstructive ones. Similar studies that induced appraisal outcomes in an experimental setting have been conducted by Aue, Flykt, and Scherer (2007), Delplanque et al. (2009), Gentsch, Grandjean, and Scherer (2013), Kreibig, Gendolla, and Scherer (2012) as well as Lanctôt and Hess (2007).

Even though studies like these provide important insights into the relationship between appraisal and physiology, only very few appraisals could be tested at a time. As the majority of these studies also used very small sample sizes, the reliability of their results can be questioned. Moreover, there was little control whether the experimental conditions actually induced the respective appraisals, as a specific stimulus might not be pleasant, relevant or goal conducive to all participants depending on their personal context. Another important downside of the experimental induction of appraisals is that not all dimensions can be analyzed, as some appraisals like compatibility with self-image and internal norms (an appraisal that has been proposed within Scherer's [2009] CPM) can hardly be induced in an experimental setting.

Altogether, there are rather incomplete theoretical assumptions as well as a lack of empirical evidence on the relations between appraisal and physiology. For many appraisal dimensions, we have no predictions at all about their relation to physiology (neither from theory nor from empirical studies). In fields of research where a strong theoretical background is missing, exploratory methods can be very useful to generate new knowledge and fill in the gaps.

4.4 Exploring the Physiology-Appraisal Link

The goal of the present study is to take a more holistic approach to investigate the interrelations between a whole set of appraisals and measured physiological reactions by applying exploratory and data-driven methods based on machine learning on a larger sample. Machine learning modeling with features extracted from physiological data has gained popularity not only in the field of medical diagnostics (Magoulas & Prentza, 2001) but has also been applied in emotion recognition (for an overview, see Jerritta, Murugappan, Nagarajan, & Wan, 2011). Studies focusing on the latter induce emotional states using auditory, visual, or audio-visual material during which different physiological signals are assessed and let participants name their perceived emotional state afterward. Subsequently, different features characterizing the signals are extracted from the data and used to predict the emotional output using different machine learning algorithms. The evaluation of these models can then tell how well emotion categories can be predicted from this kind of data and validate the assumed link between the perceived feeling and bodily responses during an emotional situation. Furthermore, it can be assessed which features are most important in predicting an emotion category.

To establish the link between physiological responses and appraisal, the same approach can be applied. For this purpose, we presented emotionally charged video material to participants while measuring their *heart rate variability* (HRV), *electrodermal activity* (EDA) and surface *electromyography* (EMG) on three facial sites – the *zygomaticus major* site, the *corrugator supercilii* site, and the *frontalis* muscle site. All five channels have been identified as affect related and have been used in the prediction of emotions before (e.g., Haag, Goronzy, Schaich, & Williams, 2004; Kim & Andre, 2008; Rigas, Katsis, Ganiatsas, & Fotiadis, 2007). The three measured EMG sites are physiologically connected to the motions of smiling (zygomaticus major), frowning (corrugator supercilii), the raise of eyebrows, indicating expressions of surprise (frontalis; Murata, Saito, Schug, Ogawa, & Kameda, 2016), as well as many other facial expressions. They are known to enable the identification of the valence of a stimulus as well as the detection of mental stress (Egger, Ley, & Hanke, 2019). The CPM marks several facial responses as outcomes of specific appraisals (for a detailed description, see Table 1 in Scherer & Ellgring, 2007), and the discussed empirical studies substantiate this interrelation (Aue et al., 2007; Aue & Scherer, 2008; van Reekum et al., 2004). EDA, the measure of skin conductivity, is also known to be related to affective reactions, especially eccrine glands measured on the palms that decrease during relaxation and increase during phases of exertion (Egger et al., 2019). A link between EDA and different appraisals such as *conduciveness*, *goal relevance*, *novelty*, and *pleasantness* of stimuli has been reported in several empirical studies

as well (Aue & Scherer, 2008; Scherer, 2009; van Reekum et al., 2004). As changes in heartbeat are modulated by the sympathetic and parasympathetic system (Rainville, Bechara, Naqvi, & Damasio, 2006), HRV, which measures changes in beat-to-beat intervals, has been used effectively for the detection of emotional arousal (Egger et al., 2019). Several theoretical relations between electrocardiographic features and appraisals have been predicted by the CPM, also implying a connection between the cognitive evaluation of a stimulus and heart rate (Scherer, 2009). Consequently, all physiological measures collected in the present study are closely interlinked with affect and are presumably predictive for different appraisal outcomes.

After the measurement of the physiological responses to each video, we assessed 15 different appraisal dimensions that have been proposed by the CPM: *suddenness* (How sudden does an event occur?), *familiarity* (How familiar is the event?), *predictability* (How predictable was the occurrence of an event?), *intrinsic pleasantness* (How pleasant was an event?), *goal/need importance* (How relevant is an event for the achievement of current goals?), *cause agent* (Who or what caused an event?), *cause motive* (Was an event caused intentionally?), *outcome probability* (Can potential consequences of an event be determined?)*, discrepancy from expectation* (Did an event contradict previously built expectations?), *conduciveness* (Does an event help to attain personal goals?)*, urgency* (Is it urgent to react to an event?), *control* (Can the outcomes of an event be controlled?)*, adjustment* (Is it possible to adjust to the outcomes of an event?), *compatibility with external* and *internal standards* (Is an event compatible with social norms and laws or self-image?). See Scherer (2001) for a more thorough description of the appraisals. For the assessment of these appraisal dimensions a modified version of the *Geneva Appraisal Questionnaire* (GAQ; Geneva Emotion Research Group, 2002) was used. We extracted 134 features from the five assessed physiological channels and predicted each appraisal dimension using a tree-based, a linear and a kernel-based machine learning model, reporting the overall cross-validated model performances for each dimension. If a link between the measured physiological signals and an appraisal dimension exists, an adequate model should be able to predict the appraisal outcome to some degree. We also constructed two different importance measures depicting the significance of each of the five physiological channels in the appraisal predictions and exemplarily analyzed the type of relationship between the appraisal dimensions and selected features.

With this data-driven approach, we are, in contrast to earlier studies, able to investigate a whole set of appraisals at once and also do not rely on uncertain appraisal inductions. We are able to analyze the appraisal-physiology link for several dimensions that have not been tested empirically yet – many of which cannot be tested in a classical experimental design. In addition,

we consider not only non-linear relations in our data but can also account for complex interactions. Moreover, as all performances and importance measures are obtained validated on out-of-sample data, our results and the derived conclusions can be considered as more robust against overfitting and therefore as more generalizable. With the exploratory analysis of the appraisal-physiology link, we hope to generate new knowledge in a rather fragmented section of emotion research.

4.5 Method

Reproducible scripts, open data, and open materials (including codebooks and video stimuli) are provided via our Open Science Framework (OSF) repository at https://osf.io/cbhfq/.

4.5.1 Participants

172 participants were recruited for the present study that either received a payment or a participation certificate. The sample size was based on available funding. As each participant viewed and rated 10 videos, 1720 observations resulted from this data collection. Due to technical problems such as signal interruption and corrupted files, that lead to the missing of one or more of the physiological signals (EMG, EDA or HRV data), several observations and participants had to be excluded. The final sample consisted of 157 participants (female = 95) and 1556 observations. The majority of subjects were psychology students at the Ludwig Maximilian University of Munich with an average age of 25.47 (range = 19-62).

4.5.2 Stimulus Material

To produce different appraisal outcomes and physiological reactions, emotional video sequences were used to induce various emotional states. Videos marked with a Creative Common CC-BY license, that allows modification and redistribution of the content, were gathered during an extensive online web search on the video-sharing service YouTube (YouTube, n.d.). Videos were selected by their potential emotional effect on the viewer, covering the four basic emotions fear, sadness, disgust, and joy. To control for culture and language effects, only German or language-free videos were included. Video sequences were cut to not exceed a maximum length of 30 s. In an online study, a selection of 20 videos was pretested. The videos were presented in a randomized order to 28 participants (female = 17). They were asked to label the videos with emotion terms, rate the intensity of their emotional experience during the observation, and answer a questionnaire constructed to assess the 16

appraisal dimensions implied by the CPM (see chapter 4.5.4 for a detailed description of the questionnaire). In total, 211 video ratings were collected in the pretest with between 7-15 ratings per video. To predict the appraisal dimensions from the physiological data, the ratings of each appraisal should show a sufficient amount of variance. In addition, the video content should be intensive enough to elicit a measurable physiological reaction. Based on these two criteria, a set of 8 videos was selected, showing both high variance in the appraisal ratings and high affective intensity. Even though all positive videos were rated as less intense and showed lower appraisal variance, two positive videos were also included to balance out the valence of the data set. Overall, 10 emotional videos with a mean length of 24.8 s (range = 10.5–30.5) were included. All videos are provided via our electronic appendix on our OSF repository.

4.5.3 Apparatus

For the measurement of the EMG and EDA signals, pre-gelled disposable electrodes with a .8 cm Ag/AgCl detection surface were used. For common-mode rejection, all sites were measured using a bipolar recording scheme. EMG electrode placement for corrugator, frontalis, zygomaticus and ground electrode was conducted following the guidelines by Fridlund and Cacioppo (1986). Electrodes for the bipolar skin conductance measurement were placed on the thenar and hypothenar eminences of the non-dominant hand of the participants (Fowles et al., 1981). A fixture on the non-dominant hand was conducted to prevent any interference with the electrodermal measurement during the tasks. The skin was prepared by cleaning the measurement sites with alcohol wipes (70% Isopropanol) and applying an abrasive electrode gel to lower the skin impedance.

For data collection, a Biopac BioNomadix MP160 data acquisition system with two wireless 2-channel EMG transmitters and one wireless PPG and EDA transmitter was used (Kremer, Mullins, Macy, Findlay, & Peterlin, 2019). Channel calibration and data acquisition were conducted using the corresponding software Acqknowledge (Version 5.0.2; Kremer et al., 2019). In accordance with the *Nyquist Theorem*, which indicates that a sinusoid signal should be sampled at least at twice its frequency for correct reconstruction, signals were sampled at a frequency of 1000 Hz (De Luca, 2003). For the HRV measurement, a Polar H10 heart rate sensor as well as a Polar V800 heart rate monitor was used, which have been proven to be consistent with measures derived from an electrocardiogram system (Giles, Draper, & Neil, 2016). The experimental program to present the videos and assess the subsequent rating of the appraisal dimensions was implemented using the E-Prime 2.0 software (Schneider, Eschman, & Zuccolotto, 2012). To synchronize the physiological data collected with AqcKnowledge and

the videos presented in E-Prime, the Observer XT (Version 14.1.1121; Zimmerman, Bolhuis, Willemsen, Meyer, & Noldus, 2009), a software for behavioral coding and event logging, was used to control and integrate both data streams. The preliminary questionnaire sent to the participants was provided via the survey framework FormR (Arslan, Tata, & Walther, 2018).

4.5.4 Procedure

Each participant received a randomized code consisting of four numerals to use as identification throughout the two-part study. First, participants completed an online questionnaire from home. In this preliminary survey, subjects were informed about the study and gave their consent to participate and to publish of their fully anonymized data. Subsequently, all relevant demographic information and further variables not included in the present study (e.g., personality, motives, emotional sensitivity[9]) were collected. For the second part of the study, each participant was invited to a laboratory. After receiving a brief introduction, the subject was asked to put on the Polar strap with the heart rate sensor. The investigator then prepared the subject's skin, applied the electrodes as described, and affixed the two EMG transmitters to the head and the EDA transmitter to the wrist of the non-dominant hand of the participant.

Before starting the testing, a calibration of the EMG and EDA transmitters was conducted, during which the transmitter leads were connected to the electrodes. Participants were instructed to do different facial movements to test if contractions would result in peaks in the respective signals. During this test phase, the investigator avoided using any emotion-related terms like *smiling* or *frowning* to bias the subject as little as possible. If a reliable signal was detected, the participant was seated in front of a computer screen and the heart rate measurement and the experimental program was started. To prevent subjects from feeling observed, the investigator monitored the physiological signal from a separated area during the following testing, intervening only if noise occurred or when electrodes needed to be reattached. Subjects were advised to place their non-dominant hand with the EDA transmitter on the table and move this hand as little as possible, answering and navigating through the study using their dominant hand on a keyboard in front of them. The participants followed a standardized instruction provided to them on screen, starting with a baseline measurement of two minutes, in which participants were instructed to close their eyes and relax. Afterward, the ten videos were presented in randomized order, each followed by a questionnaire for the assessment of the

[9] For the full set of assessed variables, see the codebook of our preliminary questionnaire at our OSF repository.

appraisal dimensions. In addition, subjects were asked to label the emotion they felt during the video and answered items relating to their immersion during the viewing of the video – these ratings had no relevance to the present study.

The presented appraisal questionnaire was based on the German version of the GAQ (Geneva Emotion Research Group, 2002). The GAQ was developed to assess through recall and verbal report as much information as possible about the appraisal process during an emotional episode. The original questionnaire, consisting of 26 items, asks to recall an arbitrary moment in the past when an intense emotion was experienced and rate the respective experience on the 16 appraisal dimensions of the CPM (e.g., *At the time of experiencing the emotion, did you think that the event happened very suddenly and abruptly?)*. For the purpose of the present study, one item for each of the appraisal dimensions was selected from the questionnaire and slightly altered to fit the video rating context (e.g., *Did you think that the events in the video happened very suddenly and abruptly?*). Only the dimension *Cause Agent*, that identifies who the agent of an evaluated event is, was assessed using three different items, identifying whether the protagonist of a video, a person different from the protagonist, or natural forces caused the events. Furthermore, we constructed an additional item for each of the four dimensions *goal/need importance*, *conduciveness*, *urgency*, and *adjustment*, that asked the participant to rate the respective dimension from the perspective of the protagonist of the video (e.g., *Can you live with, and adjust to, the consequences of the displayed events? Do you think that the protagonist can live with, and adjust to, the consequences of the events?*). As the participant's goals and actions were probably not strongly affected by the passive viewing of the mostly fictional video content, we suspected that for these dimensions, the assumed effect on the protagonist (e.g., the potential outcome of the event to the character) might be more relevant to the emotional evaluation of the video then the evaluation of the effect on oneself – especially if the viewer feels strongly involved. The dimension *power*, that evaluates the degree in which the rater can influence a situation himself, was excluded from the questionnaire as participants could obviously not influence the outcome of the videos – therefore, this appraisal was not meaningful. All items were rated on a 5-point scale ranging from *not at all*, *moderately* to *extremely*. In addition, participants were able to indicate that a question did not apply to the content of the video.

All items of the appraisal questionnaire (the original German ones as well as their English translation) and the respective appraisal dimensions can be found in the codebook of our data set in our electronic appendix.

4.5.5 Data Preprocessing

The preprocessing and all further analyses were conducted in R (Version 3.4.2; R Core Team, 2018). For each participant, the physiological signals (EMG, EDA, HRV) during the viewing of each video were extracted using the E-Prime timestamps, indicating the onset and offset of each video during the experiment. All data points assessed during other phases of the experiment were discarded except for the baseline measurement. To determine the noise contamination in the EMG data, frequency spectra were calculated using the *spec* function from the *seewave* package (Sueur, Aubin, & Simonis, 2008). The signals showed high noise contamination due to movement artifacts in the frequency range below 40 Hz as well as electromagnetic noise at 50 Hz. Therefore, a Butterworth high-pass filter with a cut-off frequency of 40 Hz was applied using the *highpass* function from the *biosignalEMG* package (Guerrero & Macias-Diaz, 2018). To filter out electromagnetic noise, a notch filter with a width of .5 Hz was applied at the respective frequency using the *bwfilter* function from the *seewave* package (Sueur et al., 2008). In line with the recommendations of Fridlund and Cacioppo (1986), we also applied a low-pass filter at 250 Hz using the *lowpass* function from the *biosignalEMG* package (Guerrero & Macias-Diaz, 2018). In addition, a baseline correction using the mean level of activation during the baseline measurement was applied to the EMG channels using the *dcbiasremoval* function from the *biosignalEMG* package (Guerrero & Macias-Diaz, 2018). As some residues of movement artifacts remained in the data and because these artifacts might influence features based on the amplitude of the signal, we added two more robust amplitude features containing a 20% trimming of the signal (see next section) to the feature set. To remove the tonic level from the EDA signal, a high pass filter at .5 Hz was applied to the data, as has been recommended by Braithwaite, Watson, Jones, and Rowe (2013), using again the *bwfilter* from the *seewave* package (Sueur et al., 2008).

4.5.6 Physiological Features

For the description of the different physiological signals, several sets of features were implemented. For the characterization of the EMG signals time and frequency domain, 32 different features were calculated (see Table 1 for an overview of all features). The specific computation of these features is based on the formulas provided by Phinyomark, Limsakul, and Phukpattaranont (2009) and Phinyomark, Phukpattaranont, and Limsakul (2012). Where necessary, features were normalized to make them independent from the length of the time series. While most of these features are used for the characterization of time series data in

Table 1

Features Extracted from EMG, EDA and HRV Channels

Features	EMG	EDA	HRV
Mean absolute value	X	X	
20% trimmed mean value	X	X	
Mean absolute value attenuated with a moving-window-20%-trimmed-mean filter	X	X	
Simple square integral	X	X	
Variance	X	X	
Absolute value of the 3rd – 5th spectral movement	X	X	
1st – 4th order autoregressive coefficients	X	X	
Root mean square	X	X	
Log detector	X	X	
Percentage waveform length	X	X	
Average amplitude change	X	X	
Difference absolute standard deviation value	X	X	
Percentage zero-crossings	X	X	
Percentage zero-crossings (.005 mv threshold)	X		
Percentage slope sign changes	X	X	
Myopuls percentage	X	X	
Percentage Wilson amplitude	X		
Median frequency of the amplitude spectrum	X	X	
Mean frequency of the amplitude spectrum	X	X	
Median frequency of the frequency spectrum	X	X	
Mean frequency of the frequency spectrum	X	X	
Peak frequency	X	X	
Mean power	X	X	
Total power	X	X	
1st – 3rd Spectral Movement	X	X	
Standard deviation of RR intervals			X
Root mean square of RR intervals			X
Percentage of successive RR intervals differing more than 50 ms			X
Ratio of the power of the low and high-frequency bands			X
Triangular interpolation of the discrete distribution of the RR intervals			X
Ratio of the standard deviation along the identity line and the standard deviation of the perpendicular axis of the Poincaré plot			X
Total number of RR intervals divided by the number of intervals in the modal bin			X
Total number of relative RR intervals divided by the number of intervals in the modal bin			X

general, some of them are more specifically applied to EMG data. As only the percentage Wilson amplitude and the zero-crossing percentage (with the .005 mV threshold) yielded zero variance on the EDA data though, all other features were deemed as appropriate to describe the skin conductance signal as well. For the analysis of the HRV data, we implemented a different set of features based on the recommendations of Vollmer (2015). Overall, 134 features were calculated – 32 for each of the EMG channels, 31 for the EDA data, and 8 for the heart rate variability data. See the R scripts provided in our electronic appendix for a formal description of the feature set.

4.5.7 Machine Learning Modeling

4.5.7.1 Benchmark

Most appraisal dimensions were assessed by a single item in our questionnaire. For the dimensions assessed with more than one item, we calculated inter-item correlations. As all correlations were low (all $r < .4$), we refrained from aggregating the items and included each of them as a separate appraisal dimension (for a similar approach, see Scherer and Meuleman, 2013). All negative poled items were reversed. For each of the 21 appraisal dimensions, we constructed a regression task using the 134 physiological features as predictors. In each task, we excluded all observations with a missing rating (*does not apply* answer) in the respective appraisal dimension. Hence, the different tasks compromised data sets of different sizes that ranged from n = 1556 for *pleasantness* to n = 948 for *internal standards* ($M = 1337.6$). For each of the 21 tasks, a benchmark experiment was conducted that compared a baseline model, a featureless learner (FL) that predicted the mean, to a random forest model (RF), a lasso regression model (LASSO), and a support vector machine (SVM) using the *mlr* package (Bischl et al., 2016). For all models, the default hyperparameter settings were used. To evaluate the models' performances, we conducted a 20 x 5 cross-validation and report the aggregated R^2. As our data set contained several observations per subject, we blocked the samples by subject within each fold to take into account the nested structure of the data. As the preprocessing of the physiological data might not be sufficient to fully eliminate artifacts in our data and because the linear model and the SVM used in the benchmark might be affected by outliers caused by such artifacts, we added an additional preprocessing step for these two models (LASSO and SVM). First, an outlier analysis was conducted on the 134 features, eliminating all values that were more than three standard deviations away from the mean of the feature. These missing values were subsequently imputed within each fold by using random numbers drawn from the

remaining empirical distribution of the feature. The RF model that reached the highest performance for all appraisal dimensions was selected for all further analyses. To determine for which appraisal dimensions the RF was able to robustly reach a positive R^2 and hence was able to explain variance in the appraisals, we looked at the variation of R^2 scores within the 100 cross-validation folds. To consider an appraisal as robustly predictable, we determined that at least 85% of the attained R^2 values should be positive (i.e., the 15% quantile should lie above 0).

4.5.7.2 Blocked Feature Importance

In a second step, we analyzed how strong the physiological channels contributed to the prediction of the appraisal dimensions that attained a positive R^2 in the previous analysis. We, therefore, constructed two blocked permutation importance measures also based on the R^2 that can quantify the impact of each of the five physiological signals (zygomaticus, corrugator, frontalis, EDA, and HRV) summarizing all features of the respective channel.

The first channel-based importance measure, R_B^2, aims to quantify how well a physiological channel can predict an appraisal dimension in general. To this end, we selected only the features calculated from the physiological channel of interest (e.g., all corrugator features) and trained the RF model on 60% of the data using only the selected feature subset. Subsequently, the R^2 was assessed on the remaining 40% test sample. The performance was calculated 100 times using different random splits and averaged subsequently (in order to avoid too small and unstable hold-out test sets, we chose a 40% test set, instead of the previously applied 20% test set):

$$R_B^2 = \frac{\sum_{i=1}^{100} R_{B,i}^2}{100} \quad (1)$$

where,

B is the block that contains all variables of the physiological channel of interest,

$R_{B,i}^2$ is the out-of-sample R^2 of the model trained with only the variables of B in the i^{th} repetition.

R_B^2 shows how much variance can be explained by the variable block in the absence of any other information and hence can be considered as a kind of "main effect" of the physiological channel, representing the overall variance that can be explained by the predictors of the channels and all interactions within the feature block.

The second channel-based importance measure, ΔR_B^2, aims to quantify the variance that can be uniquely explained by the channel beyond all other channels. For the computation, we again randomly split the data set in a training set holding 60% of the data and a test set holding the remaining 40%. First, the RF is trained with all the available features and the out-of-sample R^2 is assessed. In a second step, the out-of-sample performance of the model trained with all features that do not belong to the physiological channel of interest (e.g., all frontalis, zygomaticus, EDA, and HRV features but not the corrugator features) is assessed. To quantify the importance of the variable block of interest, the difference between the two R^2 is calculated. For a more robust assessment, the calculation is again repeated over 100 iterations and aggregated subsequently, as shown in the following formula:

$$\Delta R_B^2 = \frac{\sum_{i=1}^{100}(R_i^2 - R_{\neg B,i}^2)}{100} \quad (2)$$

where,
B is the block that contains all variables of the physiological channel of interest,
R_i^2 is the out-of-sample R^2 of the model trained with all features in the ith repetition,
$R_{\neg B,i}^2$ is the out-of-sample R^2 of the model trained without the variables of block B in the ith repetition.

As the second model is trained and validated with all features except for the variable block of interest, $R_{\neg B,i}^2$ represents the variance that can be explained by all other variables as well as all their interactions. The difference in R^2 between the complete model and the partial model consequently represents the variance that can be explained by the block of interest (as well as its interactions with other blocks) beyond all other variables. ΔR_B^2, hence, represents the incremental variance that is uniquely explained by the physiological channel, while R_B^2 also compromises the shared variance that can also be explained by other blocks. A similar importance calculation has been recommended by Yarkoni and Westfall (2017). For the calculation of both importance measures, observations were again blocked for subjects. In addition, we again applied a robustness measure by only reporting the importance of dimensions for which the attained R_B^2 or ΔR_B^2 were positive in at least 85% of the iterations.

4.5.7.3 Accumulated Local Effects Plots

As the R^2 feature importance only gives information about the relevance of the feature blocks but not about the direction and type of the relations between the appraisals and the

physiological channel, we also report *Accumulated Local Effects* (ALE) plots that visualize for given values of the feature the effect on the prediction of the outcome variable (i.e., appraisal dimension; Molnar, 2019). As this additional step was conducted to gain more insight into the machine learning models, we focussed on features that are easy to interpret from a mathematical as well as from a physiological perspective. The most straight forward interpretation can be attained by looking at features describing the amplitude height (i.e., mean absolute value, simple squared integral, root mean squared signal, absolute value of the 3rd – 5th spectral movement, and log detector), as these are clearly associated to muscle contraction for EMG (Day, 2002) and sympathetic activity or arousal for EDA (Benedek & Kaernbach, 2010). We also considered all time-domain HRV features as all of them describe the amount of variability in subsequent heartbeat intervals, excluding the high and low-frequency band ratio as well as the non-linear measure based on the Poincaré plot. We calculated the feature importance for each amplitude related as well as the HRV features and selected the one with the highest robust importance (yielding a positive importance in at least 85 of 100 iterations) for each of the appraisals that yielded a sufficient overall performance. To this end, a feature-based importance measure similar to the R_B^2 was used, calculating the R^2 for a RF model with only the feature of interest as a predictor. To prevent overfitting in these single-feature models, we restricted the tree depth of the RF to three. We report the ALE plots of the best feature within each appraisal dimension using the *iml* package (Molnar, Bischl, & Casalicchio, 2018). The plots were again calculated from the RF model with only the respective feature as a predictor and the tree depth restricted to three. To prevent extrapolation in regions of sparse data of the feature, we only plotted data within the 5% and 95% quantile of the feature.

4.6 Results

Descriptive statistics (mean and standard deviation) of the 21 assessed appraisal dimensions and the ten videos as well as the sample sizes of the appraisal subsets used the different appraisal prediction models can be found in the electronic appendix. Figure 1 shows the predictive performance of the three machine learning models (RF, LASSO, SVM) and the baseline model (FL) for the 21 assessed appraisal dimensions sorted by the maximum averaged R^2. The featureless baseline model, predicting the mean of the respective appraisal, naturally reached an R^2 of around 0 for all dimensions. The tree-based RF model yielded the best performance for all 21 appraisal dimensions, while the SVM performed consistently worse than the RF across all appraisal dimensions and also worse than the LASSO except for the *internal standards* and *adjustment (protagonist)* appraisals. Consequently, the RF was considered as the

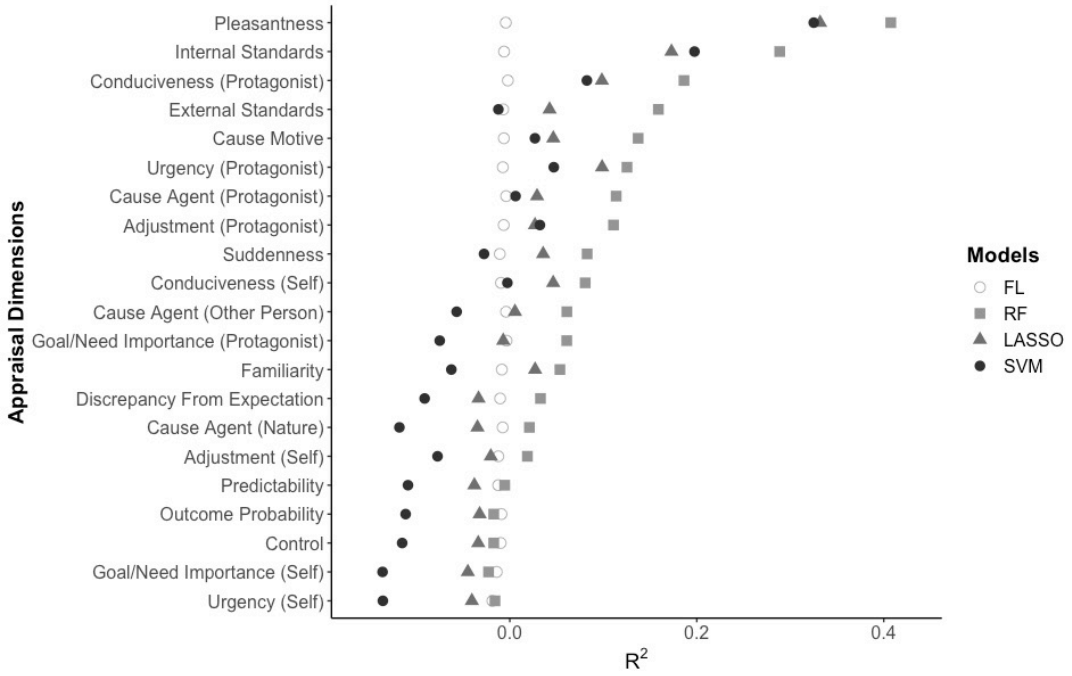

Figure 1. R^2 of the featureless learner (FL), the random forest (RF), the lasso regression (LASSO), and the support vector machine (SVM) for the 21 appraisal dimensions averaged over the 20 x 5 cross-validation folds. Appraisal dimensions are sorted by their overall performance.

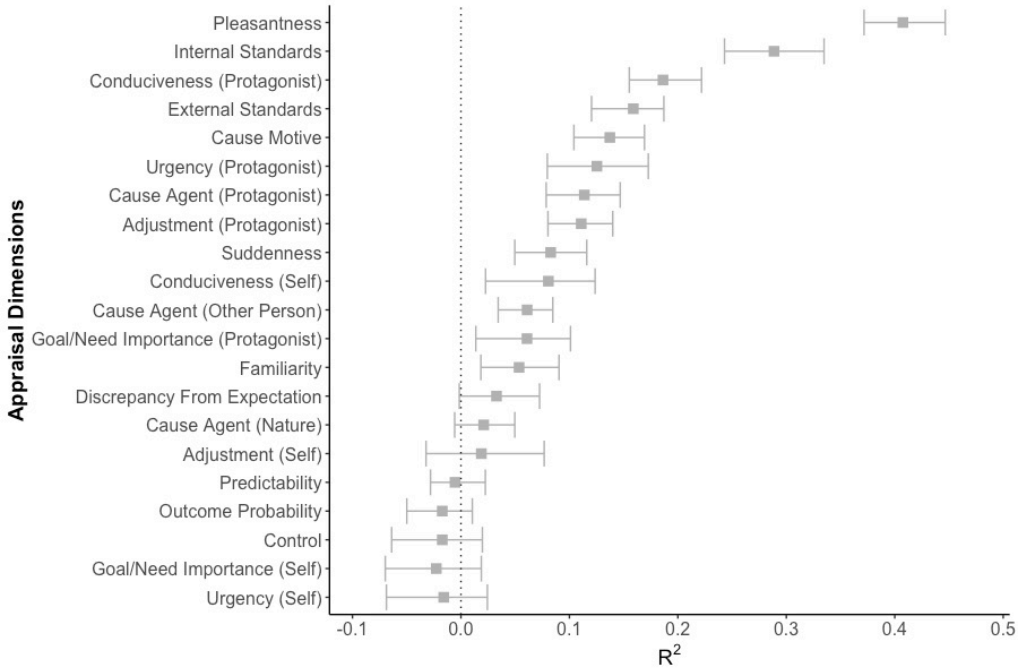

Figure 2. R^2 of the random forest (RF) for the 21 appraisal dimensions with error bars indicating the 15% and the 85% quantile of the reached R^2 within the 20 x 5 cross-validation folds. Appraisal dimensions are sorted by their overall performance.

superior model in this context and was used for all further analyses. The RFs performance varied strongly between the appraisal dimensions, ranging from -.016 to .407 with *pleasantness* (R^2 = .407) and *internal standards* (R^2 = .289) yielding the highest performance and *predictability, outcome probability, control, goal/need importance (self),* and *urgency (self)* the worst performance with a negative R^2. To rule out that the differences in the reached performance were simply due to the different sample sizes between the appraisal dimensions, we calculated a Pearson correlation between the maximal reached R^2 and the sample sizes used for each model – no significant relation was detected ($r(19) = -.077$, $p = .739$).

The inspection of the performance variation within the folds of the RF model (Figure 2) showed that in addition to the five dimensions yielding an overall negative R^2, *discrepancy from expectation* (R^2 = .033, 15% quantile= - .002), *cause agent* (*nature* ; R^2 = .021, 15% quantile = -.006) and *adjustment* (*self* ; R^2 = .019, 15% quantile = -.032) also yielded a negative performance in at least 15% of the folds. Consequently, we considered these dimensions as not robustly predictable and excluded them from the further analysis as well.

Figure 3 shows the blocked importance measures of the different physiological channels for the appraisal dimensions for which a sufficient overall R^2 was attained. For the first importance measure, R_B^2, the zygomaticus and corrugator channels overall seemed to contribute similarly to the prediction (M_{zyg} = .110, M_{corr} = .108). Frontalis, EDA, and HRV performed worse, with HRV having the smallest overall importance (M_{front} = .084, M_{EDA} = .085, M_{HRV} = .044). In 7 out of 13 appraisal dimensions, the zygomaticus channel showed the highest importance value, only yielding no importance for *cause agent (other person).* The corrugator channel yielded the highest importance for the other six appraisals but did not explain any variance for the *familiarity* appraisal. The frontalis channel did not attain a robust positive R_B^2 for the *conduciveness (self),* the *cause agent (other person),* and the *familiarity* appraisal, while the EDA channel yielded no robust importance for *goal/need importance (protagonist)* and *familiarity*. The HRV channel robustly explained variance for only 7 of the 13 dimensions, contributing nothing to the prediction of *cause agent (protagonist)*, *adjustment (protagonist)*, *conduciveness (self), cause agent (other person)*, *goal/need importance (protagonist)*, and *familiarity*. Naturally, with the decrease in overall R^2, the reached R_B^2 decreased as well.

In the second importance analysis, the ΔR_B^2 that represents the uniquely explained variance of the variable block and its interactions, the zygomaticus channel reached the highest importance across appraisals compared to the other physiological channels (M_{zyg} = .012, M_{corr} = .004, M_{front} = .001, M_{EDA} = .002, M_{HRV} = .003). The zygomaticus uniquely explained variance for the appraisals *pleasantness*, *internal standards*, *conduciveness (protagonist)*, *external*

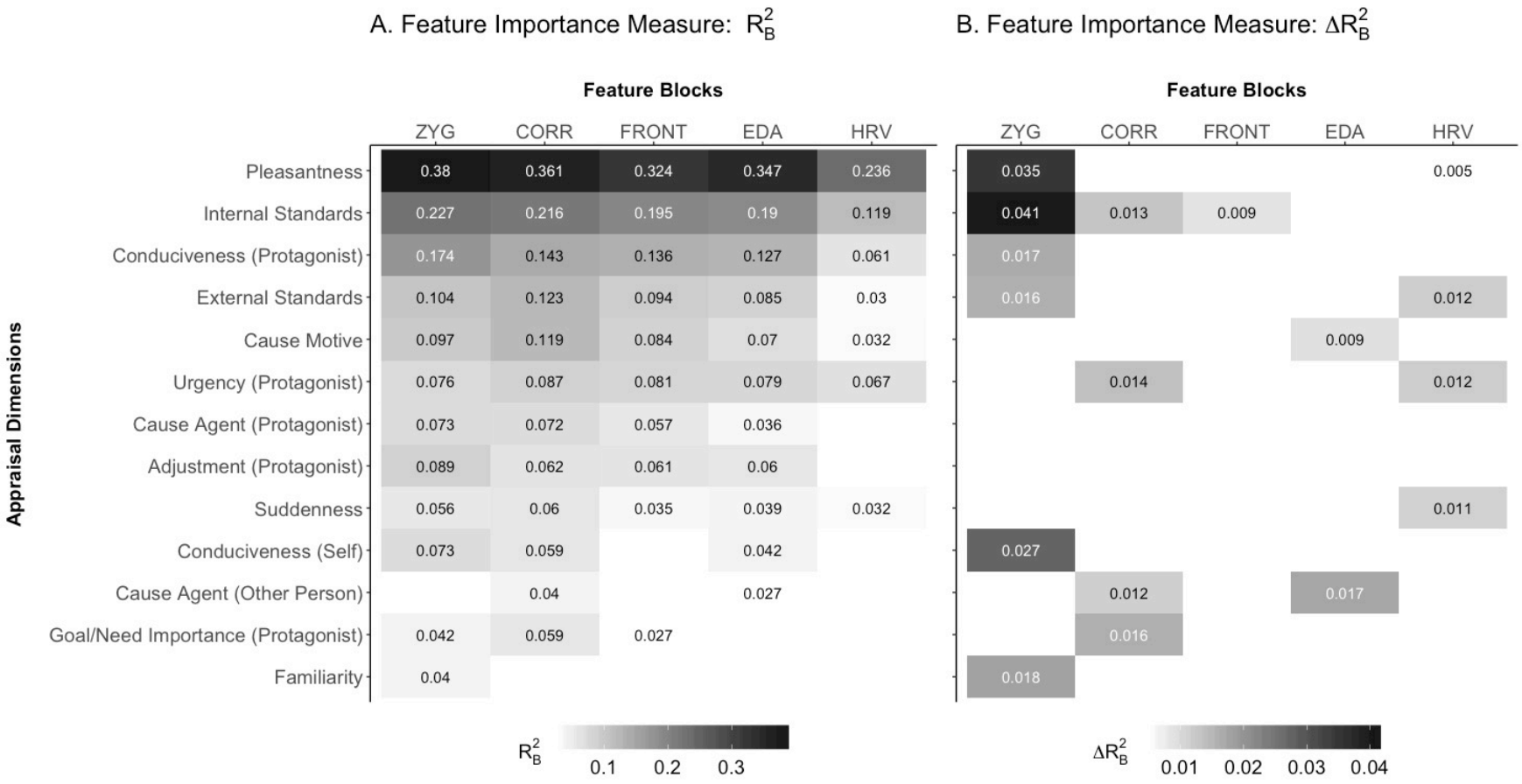

Figure 3. Blocked importance measures (A: R_B^2 *and* B: ΔR_B^2) of the five variable blocks (zygomaticus, corrugator, frontalis, EDA, and HRV) for the 13 appraisal dimensions that robustly yielded a positive overall R^2. All importance measures with more than 15% negative or zero values over the 100 iterations are omitted. Appraisal dimensions are sorted by their overall performance.

standards, *conduciveness (self)*, and *familiarity*, while the corrugator channel explained incremental variance for the *internal standards*, *urgency (protagonist)*, *cause agent (other person)*, and *goal/need importance (protagonist)* appraisal. The frontalis channel only reached a robust positive importance for the *internal standards* dimension and the EDA channel for *cause motive* and *cause agent (other person)*. Even though the HRV block seemed to have a rather low overall contribution (R_B^2) compared to the other physiological channels, it actually explained variance beyond the other blocks for four appraisals including *pleasantness*, *external standards*, *urgency (protagonist)*, and *suddenness*.

For 5 of the 13 dimensions (i.e., *cause motive*, *urgency [protagonist]*, *suddenness*, *cause agent [other person]*, and *familiarity*), no feature with a robust positive importance could be detected. Hence, these dimensions were excluded from the ALE plots. For the remaining eight appraisal dimensions, seven zygomaticus amplitude features and one corrugator amplitude feature were selected (see Figure 4). All features showed a positive feature importance and hence were able to explain variance in the respective appraisal (M = .044, range = .017-.084). *Internal standards*, *conduciveness (protagonist; self)*, *external standards*, *cause agent (pro-*

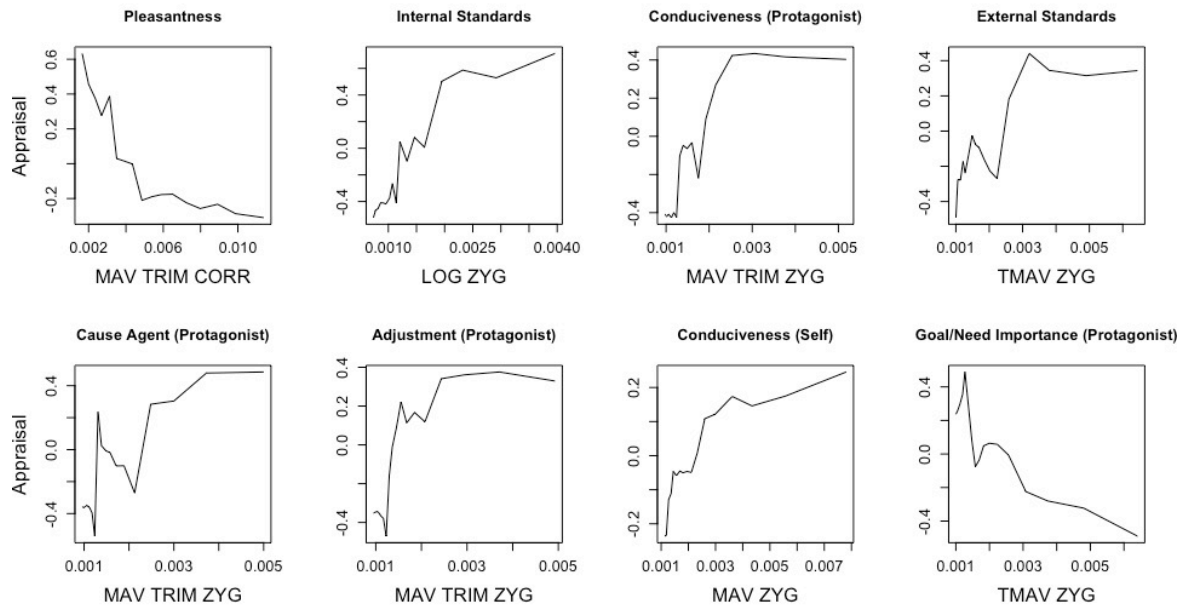

Figure 4. ALE plots for the seven appraisal dimensions for which a feature with a robust positive importance was detected. MAV: Mean absolute amplitude. MAV TRIM: 20 % trimmed mean absolute amplitude. TMAV: Mean absolute value attenuated with a moving-window-20%-trimmed-mean filter. LOG: e to the power of the mean logarithm of the absolute signal.

tagonist), and *adjustment (protagonist)* all showed a tendency towards a positive relationship with the zygomaticus amplitude (i.e., higher ratings of the respective appraisal were related with a higher zygomaticus amplitude). The appraisal *goal/need importance (protagonist),* on the other hand, showed a negative relation with the feature indicating zygomaticus amplitude height. Lastly, the *pleasantness* appraisal showed a negative relation with the corrugator amplitude. For all ALE plots, the type of link can be described as mostly non-linear.

4.7 Discussion

The present study aimed at exploring how different physiological channels relate to the appraisal dimensions of the CPM (Scherer, 2009) by validating whether the dimension can be predicted using features extracted from the respective physiological signals. The appraisals were assessed by questionnaire after presenting subjects different emotional video sequences during which the activation of different facial muscles, EDA, and HRV were collected. We compared three different machine learning models – a linear, a tree-based, and a kernel-based algorithm – to a baseline model, evaluating which type of model was most appropriate to represent the internal structure of the data. Moreover, we analyzed the relevance of each physiological channel by constructing two different blocked importance measures. Finally, we took a further step towards making the machine learning models interpretable by looking at ALE plots that depict the relation between an appraisal and a single physiological feature.

The benchmark comparing the predictive performance of the RF, the LASSO, and the SVM model showed that for 13 out of 21 appraisal dimensions a robust R^2 was attained. Hence, it can be concluded that the dimensions *discrepancy from expectation*, *cause agent (nature)*, *adjustment (self)*, *predictability*, *urgency (self)*, *outcome probability*, *control*, and *goal/need importance (self)* were physiologically related to neither the activity of the zygomaticus, the corrugator, and the frontalis, nor to EDA or HRV. The theoretical predictions made by the CPM (Scherer, 2009) are to some degree incongruent to these results, as it was theoretically assumed that the *control* appraisal would be related to the activity of different facial muscles such as zygomaticus and corrugator and the *predictability* appraisal to all five assessed channels. We were not able to empirically substantiate these relations in the setting of the present study, where emotions were induced by watching videos. Further, it was noticeable, that the *adjustment*, *urgency*, and *goal/need importance* dimensions were predictable, reaching a substantially higher R^2 than the baseline model when appraised from the perspective of video protagonist. This suggests that the appraisals might be related to the assessed physiological channels, but that in the passive viewing of a video sequence the appraisal attribution to the protagonist could

be more decisive. This would mean that for the affective evaluation of a passively experienced event it is more important whether one feels that the protagonist of the event can adjust to the consequences, has to react urgently, or is influenced strongly by the events, rather than the appraisal of those dimensions from one's own perspective. The fact that we were able to predict from the physiological features whether an event was caused by the protagonist or by a different person in the video plot (*cause agent [protagonist]* and *cause agent [other person]* appraisals) but not if the event was caused by natural forces or chance (*cause agent [nature]* appraisal), could mean that the three items (intended to measure a single appraisal or construct) actually constitute separate appraisal dimensions – an assumption that is also supported by the insufficient correlations of the items. Alternatively, the appraisal outcome, indicating that an event was caused by nature rather than by a person, might affect different physiological components that were not considered in the present study.

For the 13 dimensions for which a robust positive R^2 was attained, the RF performed consistently better than the LASSO and the SVM. This comparison clearly shows that the relations between the physiological features and the appraisal dimensions cannot be sufficiently represented by a linear model, but are probably highly non-linear. This assumption is also supported by the single-feature ALE plots, which also showed non-linear links between appraisal and physiology. Evidence for the non-linear relationship between physiological features and the valence and arousal evaluation of an event has been demonstrated by Russo, Vempala, and Sandstrom (2013). The authors showed that both dimensions can be predicted with a cross-validated R^2 of 62.4% (valence) or 82.8% (arousal) from physiological features extracted from EDA, HRV, facial EMG, and the respiration rate of a person when using non-linear neuronal networks. The predictability decreases though when a simpler linear model was applied (valence: $R^2 = 53.3\%$; arousal: $R^2 = 59.3\%$). Hence, a linear model does not seem to provide sufficient complexity to fully display the link between appraisal and physiology. The usage of linear models for better interpretability and the linear phrasing of relations derived from theory or empirical studies (e.g., Scherer, 2009), therefore, probably constitutes a simplification or could even be misleading. Even though the used SVM model is also able to represent non-linear relations, it performed substantially worse than the RF. This finding could be explained by the SVM's sensitivity to outliers (Wen, Hao, & Yang, 2010). Although we conducted a rather strict outlier exclusion beforehand, artifacts might not have been fully eliminated. Another explanation might also be that we used default choices for the kernel function, that was set to a radial basis function, and other hyperparameters. Nonetheless, the

high performance of the RF model demonstrates the superiority of a non-linear approach in this context.

The out-of-sample R^2 of those dimensions that were robustly predictable varied strongly, ranging from R^2 = .054 for *familiarity* to R^2 = .407 for *pleasantness.* Especially for the dimensions in the lower end of this range, the assessed five physiological measures are probably not sufficient to fully explain their variance. It is likely that those appraisals affect further aspects of physiology that are consequently needed to fully predict them. The reliability of our items is unknown, but our single item measures clearly limit the maximally attainable R^2. Moreover, based on the already mentioned debate on how well automatically processed appraisals can actually be assessed via self-report (Davidson, 1992; Scherer, 1993a, 2005), the measurement by questionnaire might more generally be a cause for increased measurement error in the appraisal data. We nonetheless tried to assess the appraisal process in a less retrospect way compared to the original GAQ (Geneva Emotion Research Group, 2002) by asking participants to rate the appraisal dimensions immediately after the emotional video was viewed in a controlled laboratory setting, hoping to minimize potential measurement error and retrospective biases as far as possible. Due to artifacts and noise, that cannot be fully prevented, measurement error was of course also present in our physiological features to some extent. Considering these assumptions, the achieved performances seem reasonable.

The first blocked importance measure, the R_B^2, that was implemented to assess how much variance the variables of each channel and their interactions can explain within the 13 appraisals with a sufficient overall R^2, showed that the zygomaticus and corrugator channels contributed similarly to the appraisal prediction and overall seemed to be most important. On average, the frontalis and EDA channels explained less variance as the zygomaticus and corrugator, while the HRV seemed to be the least relevant channel. For the channels that yielded a robust positive importance, it can be assumed that a relation between the respective appraisal and the physiological channel exists. Some of these links have already been made by theory or empirical work – others are somewhat contradictory to previous findings. Scherer's (2009) theoretical assumptions for *pleasantness, suddenness, familiarity, conduciveness*, and *goal/need importance* entail all physiological channels, predicting modifications in facial expressions, skin conductance, as well as cardiovascular changes. These predictions are only partially in line with our findings. All five channels yielded a robust positive importance for the *pleasantness,* the *conduciveness (protagonist)*, and the *suddenness* appraisal, hence, all channels were connected to these three appraisals. For *goal need/importance (protagonist)* though, variance was robustly explained by only the three EMG channels. A relation between

the appraisal and EDA or HRV was consequently not confirmed within the present context. In addition, *familiarity* seemed to be related to only the zygomaticus channel in our study. Previous empirical research on the physiological changes connected to the *pleasantness* appraisal also demonstrated relations to zygomaticus (Aue & Scherer, 2008; Lanctôt & Hess, 2007), corrugator (Delplanque et al., 2009; Lanctôt & Hess, 2007) and frontalis activity (Aue & Scherer, 2008; Delplanque et al., 2009) as well as to changes in EDA (van Reekum et al., 2004) and HRV (Delplanque et al., 2009). Van Reekum et al. (2004), on the other hand, were not able to find any effect of *pleasantness* on either frontalis activity or on HRV. The authors cast doubt whether *pleasantness* is at all relevant in affect-related physiology or whether the dimension influences the ANS. Our results though demonstrate that the evaluation of the intrinsic pleasantness of an event is related to changes in facial EMG as well as to HRV and hence has an impact on the ANS. A more plausible explanation, that is also recognized by the authors, is that the experimental induction of an appraisal by using games or other stimuli is not always effective. Another problem could be the authors' use of a linear MANOVA model to analyze these relations, as we clearly demonstrated that the link between *pleasantness* and physiological features is represented substantially better by a non-linear model. For the *conduciveness* appraisal, the impact on corrugator activity (Aue et al., 2007; Aue & Scherer, 2008; Gentsch et al., 2013; Lanctôt & Hess, 2007), zygomaticus activity (Aue et al., 2007; Aue & Scherer, 2008; Lanctôt & Hess, 2007), EDA (Aue & Scherer, 2008; van Reekum et al., 2004), and HRV (van Reekum et al., 2004) has also been demonstrated in several empirical studies. Van Reekum et al. (2004) who also studied the impact of *conduciveness* on the frontalis muscle were again not able to determine a significant effect. Even though this finding could also be explained by the already mentioned potential weaknesses of their design and statistical analysis as well as by their very small sample size (n = 33), it is worth mentioning that the frontalis block in our study did also not explain any variance for the *conduciveness (self)* dimension that was evaluated from the participants' own perspective but a relatively high importance when evaluated from the perspective of the video protagonist – the same was true for the HRV block. Lastly, the found link between the *goal/need importance (protagonist)* appraisal to zygomaticus and corrugator activity was also confirmed in an empirical study by Aue et al. (2007). Kreibig, Gendolla, and Scherer (2012) reported a medium effect of EDA on *goal/need importance* which we could not replicate in our study though. For the remaining seven appraisal dimensions, no studies have been conducted to our knowledge. Even though the CPM by Scherer (2009) additionally makes predictions for the *external* and *internal standards* dimensions, the physiological channels analyzed in the present study are not considered as potential outputs.

Therefore, we were able to demonstrate for the first time that the dimensions *internal* and *external standards*, *cause motive*, and *urgency (protagonist)* are also related to changes in facial EMG, EDA, and HRV and that *cause agent (protagonist)* and *adjustment (protagonist)* are related to facial EMG and HRV. Lastly, we could demonstrate that the *cause agent (other person)* appraisal is linked to corrugator activity as well as to HRV.

With the ΔR_B^2 blocked importance measure, we additionally analyzed how much incremental variance a block can explain beyond the other considered blocks. This analysis adds to the question of whether a dimension has a unique contribution to the prediction of an appraisal dimension rather than whether the dimension is related to it at all. Therefore, the results are less relevant for the basic research on the physiology of appraisals but can be used when the most economic modeling of an appraisal physiology link is the goal. The importance measure shows that for each dimension between one to five channels do not explain incremental variance, which means that the respective channel can be compensated by the other four channels in the model and that excluding the channel from the complete model would not lead to a loss in performance. For *cause agent (protagonist)* and *adjustment (protagonist)*, for example, the variance explained by each of the five physiological blocks could also be explained by the other four channels in the model. Moreover, for only 17 of the 65 measures (5 channels x 13 appraisals), a robust positive channel importance was attained, which means that in only 17 cases a channel was able to explain variance beyond the other predictors in the appraisal model. This shows that the channels must be correlated to some degree. For 8 of the 13 dimensions, either the zygomaticus or the corrugator block could be removed if all other dimensions are considered, as in these dimensions either of the two physiological channels yielded no robust positive importance. The zygomaticus channel seems to hold a higher share of incremental variance overall, even though both channels, zygomaticus and corrugator, were able to explain a comparable amount of variance in the appraisals in the first importance analysis. Moreover, the frontalis dimension, which also achieved an overall substantial R_B^2 (M_{front} = .084), could actually be removed for all appraisals except for *internal standards* without a loss in performance if the other four blocks were included in the model. Similarly, the EDA block could be excluded for all considered dimensions except for two. Interestingly, although the HRV block explained less variance (R_B^2) compared to the other physiological signals (M_{HRV} = .044), it actually uniquely explained variance for four dimensions and should therefore not be excluded when modeling the respective appraisals. For the EMG measures, a correlation between two blocks, which leads to shared variance and hence to their interchangeability, could also be caused by crosstalk between facial muscles and not necessarily

has to implicate a true relation – especially for the frontalis and corrugator muscles that are in close proximity to each other, this has to be considered.

In our last analysis, we specifically looked at the type and direction of the relation between each appraisal and the most important amplitude or HRV feature of the respective dimension. The complexity of machine learning models that can account for high-order interactions and non-linearity is one of the main benefits of these models but also constitutes an obvious downside – their interpretability. ALE plots are one approach to increase interpretability by visualizing the influence of a single feature on the prediction of a model. For eight appraisal dimensions, an interpretable feature with a robust positive importance measure was detected. With the resulting eight ALE plots, we were again able to replicate some findings of previous empirical research. Like Aue and Scherer (2008), we found a negative link between corrugator and *pleasantness* – a result that is also in line with the theoretical assumptions by Scherer (2009). We further found a positive relation between both *conduciveness* dimensions (*protagonist and self*) and the zygomaticus activity, which has also been reported by previous studies (Aue et al., 2007; Aue & Scherer, 2008). The finding that *goal/need importance (protagonist)* is negatively related to the activity of the zygomaticus is partially congruent to the findings of Aue et al. (2007) that reported a lower zygomaticus activity related to stimuli of cultural threat used to induce goal relevance. The authors also reported an increasing zygomaticus activity to stimuli depicting biological threat, though, which contradicts our results. As the used sample in this study was rather small (n = 42) and as only linear relations were considered, our results might be more reliable. However, it is also possible that the induced goal importance scenarios in the study actually constitute two different appraisal dimensions, producing different results. The remaining ALE plots suggest that zygomaticus activity overall increased if events were rated as more compatible with *internal* and *external standards,* when the protagonist was thought to be able to adjust well to the consequences of shown events (*adjustment [protagonist]*), and when the protagonist of the video was identified as the cause of events (*cause agent [protagonist]*). The ALE plots showed mostly non-linear relationships, which indicates again that the use of linear models and the subsequent linear interpretation of the resulting relations might be misleading.

4.8 Limitations

The present study holds several limitations. Even though our video selection tried to cover a broad range of emotions and potentially initiated appraisals, the specific selection might not have induced the full range in all appraisal dimensions. Moreover, as we measured each

appraisal dimension with a single item, we have to assume rather low reliability of our measurements which probably affected the reached R^2 in our study. As many appraisal dimensions are thought to be processed at least partially in an automated fashion, appraisal critics and appraisal theorists alike question whether the appraisal process can be accessed exhaustively via self-report alone (Davidson, 1992; Scherer, 1993a, 2005). Hence, the general reliance on self-reported data for the assessment of the appraisals probably contributes to measurement error in our data as well. It is an obvious paradox that when trying to find a way to assess the appraisal process (or any other contents of cognition) in a more objective indirect way (e.g., based on measures like EMG or by neuroscientific approaches) research will not get around asking participants about their inner states. Even when inducing appraisals in an experimental context, we should somehow verify how an event is actually evaluated. This validity problem is unfortunately not fully solvable with currently available measurement tools and the reliability they provide. Measurement error in the physiological channels due to artifacts, noise, and crosstalk is also not fully avoidable, even with a thorough preprocessing. Consequently, the model performance in our study could also be limited by afflicted physiological features. Potential crosstalk between EMG regions might have also affected the results of our second importance measure by decreasing the incrementally explained variance of some physiological channels. Moreover, because we were only able to assess the appraisal ratings once by self-report (not continuously), we had to aggregate the continuously assessed physiological measurements on video-level as well. Both measures hence rather depict a summary of appraisal and physiology during the video – the respective information loss most likely also affected the reached performance levels. Lastly, as the video selection in our pretest was also based on emotional intensity (eight of the ten videos were rated as intense), the results might be restricted to more intense emotion episodes.

4.9 Conclusion

In summary, we were able to investigate the connection of several physiological measures to a broad set of appraisal dimensions by using a data-driven machine learning approach. The results of the present study are based on a substantially higher sample size than most of the discussed research on this topic and all findings were additionally validated on hold-out data as well as checked for robustness. We were able to replicate some findings of previous research. Also, we were able to investigate the appraisal-physiology link for six dimensions (*internal standards*, *external standards*, *cause motive*, *urgency*, *cause agent* and *adjustment*) that have not been empirically (or theoretically) analyzed yet – probably due to the fact that

these dimensions are difficult to test using the appraisal induction designs typically applied in this field of research. Moreover, our results indicate that the links between physiology and affect related appraisal are non-linear and that future studies should refrain from using simple linear models as the results might be misleading. With these new insights, we hope to extend the knowledge base on the appraisal-physiology relation and facilitate further research on this topic.

By analyzing additional physiological channels and their links to appraisals, future research should be able to increase the predictability of appraisal dimensions even more. Overall, the fact that cognitive categories such as the perceived compatibility of an event with laws and social norms (*external standards* dimensions) can be predicted (at least to some degree) by physiological measures is impressive. The results lend support for cognitive theories of emotions like the CPM (Scherer, 2009), that assume that emotions are not simply the subjective perception of a bodily response to a stimulus but that the cognitive evaluation of our environment is the central element in a multi-componential emotion process.

4.10 References

Arnold, M. B. (1960). *Emotion and personality*. New York: Columbia University Press.

Arslan, R. C., Tata, C., & Walther, M. P. (2018). formr: A study framework allowing for automated feedback generation and complex longitudinal experience sampling studies using R (Version v0.18.3). https://doi.org/10.5281/zenodo.3229668

Aue, T., Flykt, A., & Scherer, K. R. (2007). First evidence for differential and sequential efferent effects of stimulus relevance and goal conduciveness appraisal. *Biological Psychology*, *74*(3), 347–357. https://doi.org/10.1016/j.biopsycho.2006.09.001

Aue, T., & Scherer, K. R. (2008). Appraisal-driven somatovisceral response patterning: Effects of intrinsic pleasantness and goal conduciveness. *Biological Psychology*, *79*(2), 158–164. https://doi.org/10.1016/j.biopsycho.2008.04.004

Benedek, M., & Kaernbach, C. (2010). A continuous measure of phasic electrodermal activity. *Journal of Neuroscience Methods*, *190*(1), 80–91. https://doi.org/10.1016/j.jneumeth.2010.04.028

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., … Jones, Z. M. (2016). mlr: Machine Learning in R. *Journal of Machine Learning Research*, *17*(170), 1–5.

Braithwaite, J. J., Watson, D. G., Jones, R., & Rowe, M. (2013). *A Guide for Analysing Electrodermal Activity (EDA) & Skin Conductance Responses (SCRs) for Psychological Experiments*. Birmingham: University of Birmingham.

Davidson, R. J. (1992). Prolegomenon to the structure of emotion: Gleanings from neuropsychology. *Cognition and Emotion*, *6*(3–4), 245–268. https://doi.org/10.1080/02699939208411071

Day, S. (2002). *Important factors in surface EMG measurement* [Technical Report]. Retrieved from http://www. andrewsterian.com/courses/214/EMG\ measurement\ and\ recording.pdf

De Luca, G. (2003). *Fundamental concepts in EMG signal acquisition*. Retrieved from https://www.delsys.com/downloads/TUTORIAL/fundamental-concepts-in-emg-signal-acquisition.pdf

Delplanque, S., Grandjean, D., Chrea, C., Coppin, G., Aymard, L., Cayeux, I., … Scherer, K. R. (2009). Sequential unfolding of novelty and pleasantness appraisals of odors: Evidence

from facial electromyography and autonomic reactions. *Emotion*, *9*(3), 316–328. https://doi.org/10.1037/a0015369

Egger, M., Ley, M., & Hanke, S. (2019). Emotion Recognition from Physiological Signal Analysis: A Review. *Electronic Notes in Theoretical Computer Science*, *343*, 35–55. https://doi.org/10.1016/j.entcs.2019.04.009

Fowles, D. C., Christie, M. J., Edelberg, R., Grings, W. W., Lykken, D. T., & Venables, P. H. (1981). Publication Recommendations for Electrodermal Measurements. *Psychophysiology*, *18*(3), 232–239. https://doi.org/10.1111/j.1469-8986.1981.tb03024.x

Fridlund, A. J., & Cacioppo, J. T. (1986). Guidelines for Human Electromyographic Research. *Psychophysiology*, *23*(5), 567–589. https://doi.org/10.1111/j.1469-8986.1986.tb00676.x

Frijda, N. H. (1986). *The emotions*. Cambridge: Cambridge University Press.

Geneva Emotion Research Group. (2002). *Geneva Appraisal Questionnaire (GAQ)*. Retrieved from https://www.unige.ch/cisa/files/3414/6658/8818/GAQ_English_0.pdf

Gentsch, K., Grandjean, D., & Scherer, K. R. (2013). Temporal dynamics of event-related potentials related to goal conduciveness and power appraisals. *Psychophysiology*, *50*(10), 1010–1022. https://doi.org/10.1111/psyp.12079

Giles, D., Draper, N., & Neil, W. (2016). Validity of the Polar V800 heart rate monitor to measure RR intervals at rest. *European Journal of Applied Physiology*, *116*(3), 563–571. https://doi.org/10.1007/s00421-015-3303-9

Guerrero, J. A., & Macias-Diaz, J. E. (2018). *BiosignalEMG: Tools for Electromyogram Signals (EMG) Analysis*. Retrieved from https://CRAN.R-project.org/package=biosignalEMG

Haag, A., Goronzy, S., Schaich, P., & Williams, J. (2004). Emotion Recognition Using Bio-sensors: First Steps towards an Automatic System. In E. André, L. Dybkjær, W. Minker, & P. Heisterkamp (Eds.), *Affective Dialogue Systems* (Vol. 3068, pp. 36–48). https://doi.org/10.1007/978-3-540-24842-2_4

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*. https://doi.org/10.1109/TAFFC.2019.2940937

Jerritta, S., Murugappan, M., Nagarajan, R., & Wan, K. (2011). Physiological signals based human emotion Recognition: A review. *2011 IEEE 7th International Colloquium on Signal Processing and Its Applications*, 410–415. https://doi.org/10.1109/CSPA.2011.5759912

Kim, J., & Andre, E. (2008). Emotion recognition based on physiological changes in music listening. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *30*(12), 2067–2083. https://doi.org/10.1109/TPAMI.2008.26

Kreibig, S. D., Gendolla, G. H. E., & Scherer, K. R. (2012). Goal relevance and goal conduciveness appraisals lead to differential autonomic reactivity in emotional responding to performance feedback. *Biological Psychology*, *91*(3), 365–375. https://doi.org/10.1016/j.biopsycho.2012.08.007

Kremer, J. M., Mullins, M., Macy, A., Findlay, F., & Peterlin, E. (2019). *AcqKnowledge 5 Software Guide For Life Science Research Applications – Data Acquisition and Analysis with Biopac Hardware Systems*. Biopac Systems, Inc.

Lanctôt, N., & Hess, U. (2007). The timing of appraisals. *Emotion*, *7*(1), 207–212. https://doi.org/10.1037/1528-3542.7.1.207

Lazarus, R. S. (1966). *Physchological stress and the coping process*. New York: McGraw-Hill.

Magoulas, G. D., & Prentza, A. (2001). Machine Learning in Medical Applications. In G. Paliouras, V. Karkaletsis, & C. D. Spyropoulos (Eds.), *Machine Learning and Its Applications* (Vol. 2049, pp. 300–307). https://doi.org/10.1007/3-540-44673-7_19

Meuleman, B., & Scherer, K. R. (2013). Nonlinear Appraisal Modeling: An Application of Machine Learning to the Study of Emotion Production. *IEEE Transactions on Affective Computing*, *4*(4), 398–411. https://doi.org/10.1109/T-AFFC.2013.25

Molnar, C. (2019). Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. Retrieved from https://christophm.github.io/interpretable-ml-book/

Molnar, C., Bischl, B., & Casalicchio, G. (2018). iml: An R package for Interpretable Machine Learning. *JOSS*, *3*(26), 786. https://doi.org/10.21105/joss.00786

Moors, A. (2009). Theories of emotion causation: A review. *Cognition & Emotion*, *23*(4), 625–662. https://doi.org/10.1080/02699930802645739

Moors, A., Ellsworth, P. C., Scherer, K. R., & Frijda, N. H. (2013). Appraisal Theories of Emotion: State of the Art and Future Development. *Emotion Review*, *5*(2), 119–124. https://doi.org/10.1177/1754073912468165

Murata, A., Saito, H., Schug, J., Ogawa, K., & Kameda, T. (2016). Spontaneous Facial Mimicry Is Enhanced by the Goal of Inferring Emotional States: Evidence for Moderation of "Automatic" Mimicry by Higher Cognitive Processes. *PLOS ONE*, *11*(4), e0153128. https://doi.org/10.1371/journal.pone.0153128

Phinyomark, A., Limsakul, C., & Phukpattaranont, P. (2009). A Novel Feature Extraction for Robust EMG Pattern Recognition. *Journal of Computer Science*, *1*(1), 71–81.

Phinyomark, A., Phukpattaranont, P., & Limsakul, C. (2012). Feature reduction and selection for EMG signal classification. *Expert Systems with Applications*, *39*(8), 7420–7431. https://doi.org/10.1016/j.eswa.2012.01.102

R Core Team. (2018). *R: A Language and Environment for Statistical Computing*. Retrieved from https://www.R-project.org/

Rainville, P., Bechara, A., Naqvi, N., & Damasio, A. R. (2006). Basic emotions are associated with distinct patterns of cardiorespiratory activity. *International Journal of Psychophysiology*, *61*(1), 5–18. https://doi.org/10.1016/j.ijpsycho.2005.10.024

Rigas, G., Katsis, C. D., Ganiatsas, G., & Fotiadis, D. I. (2007). A User Independent, Biosignal Based, Emotion Recognition Method. In C. Conati, K. McCoy, & G. Paliouras (Eds.), *User Modeling 2007* (Vol. 4511, pp. 314–318). https://doi.org/10.1007/978-3-540-73078-1_36

Roseman, I. J. (1984). Cognitive Determinants of Emotion: A Structural Theory. *Personality and Social Psychology Review*, *5*, 11–36.

Russo, F. A., Vempala, N. N., & Sandstrom, G. M. (2013). Predicting musically induced emotions from physiological inputs: Linear and neural network models. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00468

Scarantino, A., & de Sousa, R. (2018). Emotion. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2018). Retrieved from https://plato.stanford.edu/archives/win2018/entries/emotion/

Schachter, S., & Singer, J. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological Review*, *69*(5), 379–399. https://doi.org/10.1037/h0046234

Scherer, K. R. (1984). On the nature and function of emotion: A component process approach. In K. R. Scherer & P. Ekman (Eds.), *Approaches to Emotion* (pp. 293–317). Hillsdale, NJ: Erlbaum.

Scherer, K. R. (1993a). Neuroscience projections to current debates in emotion psychology. *Cognition & Emotion*, *7*(1), 1–41. https://doi.org/10.1080/02699939308409174

Scherer, K. R. (1993b). Studying the emotion-antecedent appraisal process: An expert system approach. *Cognition & Emotion*, *7*(3–4), 325–355. https://doi.org/10.1080/02699939308409192

Scherer, K. R. (1997). Profiles of Emotion-antecedent Appraisal: Testing Theoretical Predictions across Cultures. *Cognition & Emotion*, *11*(2), 113–150. https://doi.org/10.1080/026999397379962

Scherer, K. R. (2001). Appraisal considered as a process of multilevel sequential checking. In K. R. Scherer, A. Schorr, & J. Johnstone (Eds.), *Appraisal processes in emotion* (pp. 92–120). New York: Oxford University Press.

Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social Science Information*, *44*(4), 695–729. https://doi.org/10.1177/0539018405058216

Scherer, K. R. (2009). The dynamic architecture of emotion: Evidence for the component process model. *Cognition & Emotion*, *23*(7), 1307–1351. https://doi.org/10.1080/02699930902928969

Scherer, K. R., & Ellgring, H. (2007). Are facial expressions of emotion produced by categorical affect programs or dynamically driven by appraisal? *Emotion*, *7*(1), 113–130. https://doi.org/10.1037/1528-3542.7.1.113

Scherer, K. R., & Meuleman, B. (2013). Human Emotion Experiences Can Be Predicted on Theoretical Grounds: Evidence from Verbal Labeling. *PLOS ONE*, *8*(3), e58166. https://doi.org/10.1371/journal.pone.0058166

Schneider, W., Eschman, A., & Zuccolotto, A. (2012). *E-Prime 2.0*. Pittsburgh: Psychology Software Tools, Inc.

Smith, C. A., & Ellsworth, P. C. (1985). Patterns of cognitive appraisal in emotion. *Journal of Personality and Social Psychology*, *48*(4), 813–838. https://doi.org/10.1037/0022-3514.48.4.813

Sueur, J., Aubin, T., & Simonis, C. (2008). Seewave: A free modular tool for sound analysis and synthesis. *Bioacoustics*, *18*, 213–226.

van Reekum, C., Johnstone, T., Banse, R., Etter, A., Wehrle, T., & Scherer, K. R. (2004). Psychophysiological responses to appraisal dimensions in a computer game. *Cognition & Emotion*, *18*(5), 663–688. https://doi.org/10.1080/02699930341000167

Vollmer, M. (2015). A robust, simple and reliable measure of heart rate variability using relative RR intervals. *2015 Computing in Cardiology Conference*, 609–612. https://doi.org/10.1109/CIC.2015.7410984

Wen, W., Hao, Z., & Yang, X. (2010). Robust least squares support vector machine based on recursive outlier elimination. *Soft Computing*, *14*(11), 1241–1251. https://doi.org/10.1007/s00500-009-0535-9

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

YouTube. (n.d.). Retrieved from https://youtu.be/

Zimmerman, P. H., Bolhuis, J. E., Willemsen, A., Meyer, E. S., & Noldus, L. P. J. J. (2009). The Observer XT: A tool for the integration and synchronization of multimodal signals. *Behavior Research Methods*, *41*(3), 731–735. https://doi.org/10.3758/BRM.41.3.731

5 General Discussion

The present thesis examined two paths of the Component Process Model (CPM), an appraisal emotion theory developed by Scherer (1984, 2001, 2009). Both presented studies focused on the role of the cognitive component within the proposed multi-componential emotion process. Study 1 analyzed the connection between cognitive appraisals and the subjective feeling of an individual (link A) by using theoretically informed computational models combined with parameter estimations from empirical data, while study 2 investigated the link between appraisal and physiological responses (link B) using different machine learning algorithms.

5.1 Link A: The Appraisal-Feeling Link

5.1.1 Results

The results of study 1 demonstrate that the link between the evaluation of emotion-relevant appraisal dimensions and the perceived feeling during a retrospectively evaluated emotional episode exists. All four implemented models (M1-M4), that predicted emotion classes using a decision rule based on a prototype similarity metric, as well as the examined *random forest* (RF) machine learning algorithm predicted the perceived feeling of participants from the assessed appraisal patterns better than the naive baseline model. Regardless of their algorithmic implementation, all models were able to explain variance in the labeled subjective feeling based on the self-reported appraisal patterns. This finding aligns with previously conducted studies, like the ones by Scherer (1993) and Scherer and Meuleman (2013), that were also able to predict emotion labels from appraisal ratings assessed via questionnaire. The predictive performance of the theoretically informed models varied though depending on their implemented weighting mechanism. The preferred model M3 (evaluated based on the overall predictive accuracy, the emotion class- and family-wise precision scores, the model calibration and the parsimony of the model) weighted the 16 assessed appraisal dimensions differently strong in the similarity metric and used weighting parameters that were attained by an optimization procedure from empirical data. The superior performance of M3 affirms the idea formulated by Scherer (2001) that the appraisal dimensions are differently important in the appraisal-to-feeling process. Moreover, it indicates that an equal weighting of the appraisals (as implemented in M1), a much more complex weighting (as implemented in M4) or a weighting with the theoretically derived weighting parameters proposed by Scherer and Meuleman (2013;

as implemented in M2) might not display the algorithmic level of this process comparably well. The comparison of the preferred model M3 with a RF machine learning algorithm showed though, that the latter performed better for the majority of the emotion classes and emotion families. The machine learning model was chosen as one upper boundary to demonstrate which predictive performance can be reached if the model complexity is increased. The results show that the complex structure of the RF can explain more variance in the emotion labels than the optimized prototype approach of M3.

5.1.2 Differentiability of the Emotion Prototypes

The presented APPraisal app visualizes M3's predictions and the respective prototype similarities for different appraisal patterns. The app shows that the empirical prototypes (attained from the empirical data set in study 1) are very similar for emotion classes of the same valence (i.e., positive or negative emotions). The similarity of the prototypes and the resulting lack of differentiability between these classes might be one reason for the weaker performance of the theoretically informed model M3 in comparison to the RF. Several aspects were discussed that might have affected the prototype calculation such as the lack of clarity in the used emotion labels, the predefined set of emotion terms participants had to choose from as well as the way the prototypes were calculated from the empirical data. While the first two problems indicate a more general measurement problem that would have also affected the performance of the machine learning model (and will be discussed more detailed in chapter 5.3), a problem with the calculation of the prototypes would only be relevant for the theoretical models. For the calculation of the prototypes, the appraisal patterns for each observation labeled with the respective emotion were averaged. As most of the observations (72%) were labeled with two emotion terms though, the majority of observations were included in the calculation of two different emotion prototypes, which potentially led to a blending and converging of prototypes – especially for those emotions that often occur simultaneously (it is plausible that emotions of the same valence occur more frequently together such as sadness and anger or happiness and pride). An approach to prevent this kind of merging and improve the prototypes' differentiability in future studies would be to include only instances that were labeled with a single and therefore explicit label in the prototype calculation. In the present work, we have specifically refrained from doing so as the observations with a single label were rather rare (28%) in the used data set, which would have resulted in too few observations for the prototype calculation of many of the emotions – the number of single label observations for the 13 emotion classes ranged from 5 to 281 with an average of $M = 70.38$.

5.1.3 Comparison of the Prototype Approach and the Random Forest Algorithm

A further comparison of the M3 model and the RF algorithm is rather difficult as the two models differ very strongly in their mathematical implementation. One major difference can be highlighted though. In Figure 1A, the concept of the emotion classification in M3 is visualized for an example with three hypothetical prototypes and two hypothetical appraisal dimensions. The two-dimensional predictor space is divided into three areas so that the boundaries lie exactly between the three prototypes. Each new observation is then classified with the label of the prototype in the closest proximity (i.e., the prototype area in which the observation falls into). When the value in one or both appraisal dimensions increases or decreases so that the observation moves away from the prototype, the distance to the respective prototype increases monotonously. When the observation is consequently moved out of the prototype area, it is classified as a different emotion. In the classification with the RF, as it has been described in chapter 1.4.4.2, another scenario is possible. As shown in Figure 1B, the predictor space is divided using the appraisal dimensions so that the greatest possible reduction of the mean misclassification error (MMCE) is achieved. This procedure is repeated until the stopping criterion is reached and the most frequent class in the resulting areas is predicted. As a consequence, the predictor space can be split many times, potentially leading to different areas associated with the same emotion label. This can happen for a single tree as illustrated in Figure 1B (were two areas have formed that are associated with emotion E1), but also when the majority votes over a whole set of trees are considered. A new observation is again classified by the area it falls into (e.g., as emotion E1), but when the value in one or both appraisal dimensions increases or decreases so that the observation is moved into another area of the predictor space, the same emotion might be classified again (i.e., again as emotion E1). Transferred to the idea of emotion prototypes, this would mean that different prototypes for the same emotion could exist. It would be conceivable, for example, that the emotion pride is prototypically connected to events that are caused solely by oneself (hence, indicating a high value in the *cause agent* dimension), such as the achievement of a university degree, but also to events which one did not cause (i.e., indicating a low value in the *cause agent* dimension), such as the professional career of a partner or the achievement of one's child. Hence, two different prototypical values of *cause agent* would be connected to the same emotion (i.e., two different prototypical appraisal patterns for pride would exist). As the RF is able to represent such multiple separated classification areas and showed a higher performance for most emotion classes in study 1, this representation of the predictor space might be more accurate. The potential existence of multiple prototypes for each emotion could be another reason why the

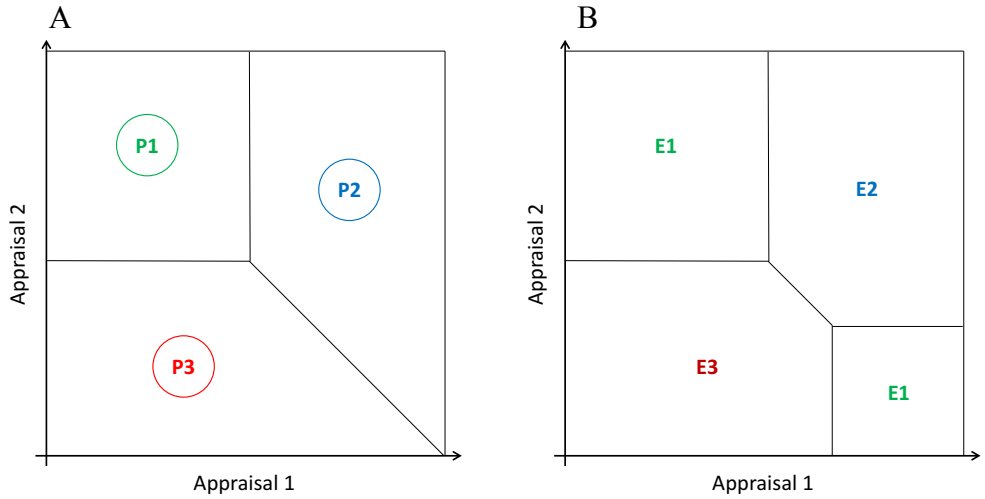

Figure 1. Two examples that demonstrate how the predictor space can be divided when using A) the prototype similarity approach applied in study 1 and B) a single tree from the random forest (RF) algorithm, where Appraisal 1 and 2 are two hypothetical appraisal dimensions, P1-P3 are three hypothetical prototypes and E1-E3 are three hypothetical emotion classes.

calculated prototypes in study 1 (that were averaged over all instances with the same emotion label) did not differentiate very well. To test this hypothesis, future research could examine if the predictive performance can be improved by finding different clusters within observations labeled with the same emotion (e.g., all pride observations) and subsequently generate multiple prototypes from the attained clusters. However, different emotion clusters could also be an indicator of measurement error, as participants might label their emotional states "incorrect" due to a lack of self-awareness or terminological confusion.[10]

Despite the possible disparities between the two models it must also be stressed that the difference in performance between M3 and the RF model is not very large – with the M3 model even reaching a slightly better performance for four emotion classes and one emotion family. As discussed in study 1, it is however striking that even with the very complex machine learning model an accuracy of over 52.3% for emotion classes and 80.8% for emotion families could not be exceeded, indicating that a much better performance can probably not be reached due to measurement error in the self-reported appraisal dimensions and emotion labels. It is possible,

[10] As the relationship between a verbal emotion label and a specific emotional state is learned rather implicitly, there are no explicitly correct labels for certain component patterns. Therefore, the term *incorrect* indicates that the label is usually connected to a different affective state and hence appraisal pattern.

however, that the performance of the RF could still be increased slightly by tuning its hyperparameters, as the optimal values for such parameters depend on the used data set (Probst, Wright, & Boulesteix, 2019) and we only used default settings.

5.1.4 Appraisal Dimensions and their Relevance

The weights of model M3 were attained using a genetic optimization algorithm. They indicate that all 16 appraisal dimensions contributed (to varying degrees) to the prediction of the emotion labels. This finding, as thoroughly discussed in study 1, challenges other appraisal theories that assume a substantially smaller set of appraisals (e.g., Lazarus, 1991; Ortony, Clore, & Collins, 1988; Roseman, 1984). Instead, it would be conceivable that the expansion of the appraisal set could further increase the performance of the model. Rather than fixating on one single appraisal theory, future research should combine the appraisals proposed by different theorists to determine a potential bigger set of relevant dimensions from empirical data. Moreover, while we aggregated the items in study 1 to 16 appraisal dimensions as indicated by the GAQ (Geneva Emotion Research Group, 2002), we refrained from doing so in study 2 using each item as a separate appraisal dimension because of the low absolute inter-item correlations (all $r < .4$). In study 1 though, as the goal was to reproduce the theory as closely as possible to test its plausibility, the 16 appraisal dimensions were maintained, even though the inter-item correlations of the aggregated dimensions were often low as well ($M = .36$, range = .01–.89). This indicates that at least some of the items rather represent dimensions of their own and that the aggregation of these uncorrelated items might have contributed to measurement error.

With a correlation of $r = .30$, the new optimized weights of M3 deviate demonstrably from the weights proposed by Scherer and Meuleman (2013) implemented in model M2. Even though model M3 performed substantially better than the M2, we remarked that the relative height of the attained appraisal weights should be interpreted with caution as the latter are highly dependent on the used data set, the mathematical realization of the distance metric and the used emotion prototypes. Based on the previously discussed potential downsides of the prototype calculation, it has to be emphasized even more, that the weighting parameters should be validated in different contexts. A good starting point would be to construct a suitable importance measure to quantify the relevance of the appraisal features in the RF model – examining how relevant the different appraisal dimensions are when predicting emotion labels using a different model with diverging characteristics.

5.1.5 Comparison of the Weighting Algorithm of Model M3 and Model M4

Model M3 was previously referred to as the preferred model of study 1 but was not identified as the superior or best model, as the comparison of M3 and M4 led to a rather ambiguous picture. Even though M4 yielded a higher overall accuracy and a better calibration to the class frequencies, the class-wise precision scores were lower than for M3 for most emotion classes and families. Based on the criterion of parsimony and interpretability, we hence preferred model M3 as the less complex model with robustly estimated (and therefore interpretable) weighting parameters. From the different performance indicators considered in study 1, it can be concluded that M4 clearly has different prediction characteristics than M3 but cannot be identified as the worse model explicitly. Hence, the preference for M3 should not prevent future research to further look into the idea of differently weighted appraisal dimensions within different emotions. Besides some empirical evidence pointing in this direction (Ellsworth & Smith, 1988), Fernando, Kashima, and Laham (2017) introduced the idea of variable appraisal set models. The theory assumes that each emotion is elicited by a different set of appraisal dimensions. Though these sets may overlap to some degree, not all appraisals should be relevant to all emotions. The implementation of M4 is equivalent to this idea, as each appraisal weight for each emotion could have been shrunken – potentially even leading to the full elimination of an appraisal. While Scherer's (2001) theory does not include an explicit description of this concept, the open parameters in his emotion prototypes have the same meaning.

5.1.6 Theoretical Prototypes

Since it would have gone beyond the scope of study 1, the predictive performance of the models with the theoretical prototypes proposed by Scherer (2001) was not evaluated. For a systematic comparison with the models M1-M4, all four models would have to be implemented with the empirical as well as the theoretical prototypes. For a quick (and computationally less costly) examination, however, the performance of model M2 with the theoretical prototypes instead of the empirical prototypes can be considered. With an overall accuracy of 27.3% for emotion classes and 60.6% for emotion families as well as precision scores ranging from 5.9% to 61.6%, the model yielded a very similar performance to M2 in study 1 (class accuracy = 27.1%, family accuracy = 62.4%, precision range = 4.2% – 61.8%). This is interesting because the reported mean correlation of $r = .47$ between the theoretical and empirical prototypes indicates that the prototypes deviate to some degree. The two-dimensional scaling of both prototypes, the theoretical (white nodes) and empirical ones (grey nodes), in

Figure 2 demonstrates that for most emotion categories the two prototypes are similar (i.e., in close proximity to each other), but that the theoretical prototypes rather represent extreme values within the appraisal space compared to the empirically assessed prototypes. Hence, the theoretical prototypes seem to be more consistent with the concept of stereotypes applied in social science (Judd & Park, 1993) than with the prototype definition of Rosch (1983): Judd and Park (1993) define a stereotype as a set of beliefs about the attributes (here, appraisal values) of a certain group (here, emotions) that do not necessarily have to be accurate but seek to display whether the attribute is more or less prevalent compared to another group. Hence, the main goal of stereotypes is rather to accentuate (sometimes exaggerate) differences between groups, than to describe the groups representatively.

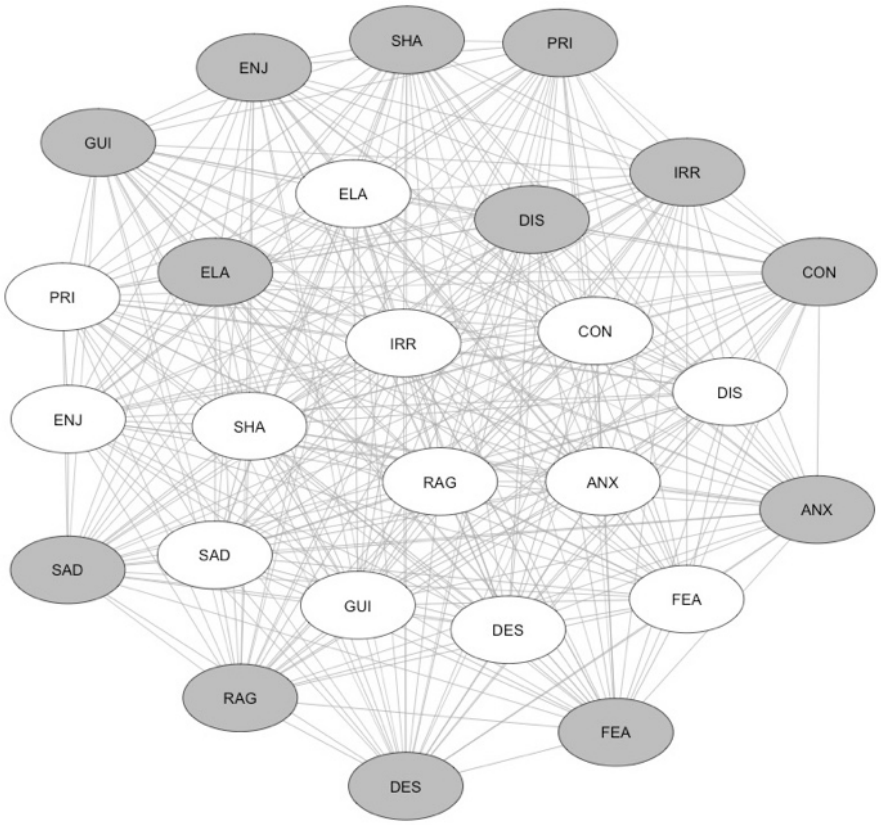

Figure 2. Two-dimensional scaling of the 13 theoretical prototypes proposed by Scherer (2001; grey) as well the emotion prototypes used in study 1 (white; SAD = sadness, FEA = fear, CON = contempt, DES = despair, RAG = rage, SHA = shame, DIS = disgust, GUI = guilt, IRR = irritation, ANX = anxiety, ELA = elation, ENJ = enjoyment, PRI = pride). Note that this is a force embedded layout in which not all distances are displayed spatially correct.

However, the observation that the theoretical prototypes represent more extreme values might also stem from the way we translated the categorical prototype levels given by Scherer (2001) to numerical ones (ranging from 0 to 1). We converted the category *very high* to 1 and the category *very low* to 0 – these were the most extreme values possible. As it very unlikely that the most extreme values are the prototypical ones in a large population, these values (though they did not occur very frequently in the prototypical appraisal patterns) are probably not very plausible. As the model gives no information on how the categories translate to numerical values though, a different solution was not feasible and possibly not intended by Scherer (2001).

5.2 Link B: The Appraisal-Physiology Link

5.2.1 Results

To analyze the second path of interest – the appraisal-physiology path – an empirical study was conducted in which participants watched emotional videos while different physiological measures were assessed (study 2). Subsequently, subjects rated different appraisal dimensions based on an adapted form of the GAQ (Geneva Emotion Research Group, 2002). As no detailed theoretical assumptions about the connection between appraisals and the assessed physiological channels (i.e., corrugator activity, zygomaticus activity, frontalis activity, EDA and HRV) exist, we calculated 134 features from the five physiological channels and predicted each of the appraisal dimensions by using different machine learning models. The highest predictive performance was again reached by the RF model indicating that a non-linear link most appropriately represents the appraisal-physiology relation. This assumption was also supported by the *Accumulated Local Effect* (ALE) plots which showed non-linear effects of single features on the appraisal outcomes. The two newly constructed importance measures, R_B^2 and ΔR_B^2, showed that the five physiological channels were differently important in the prediction of different appraisal dimensions and only very few blocks actually explained incremental variance. The findings were partly in line with previous empirical findings and theoretical assumptions and added new information for six appraisals whose connection to physiology has not been investigated yet (i.e., *internal standards*, *external standards*, *cause motive*, *urgency [protagonist]*, *cause agent [protagonist]*, *adjustment [protagonist]*).

5.2.2 Comparison of the Linear and Non-Linear Machine Learning Models

The results of the study demonstrate that the link between appraisal and physiology was best represented by the RF model that is able to display complex interactions and non-linearity. The *lasso regression* (LASSO) and the *support vector machine* (SVM), on the other hand, performed substantially worse. Therefore, it was concluded that the appraisal-physiology link is most likely non-linear. Noteworthy though was that the SVM was used with a radial basis kernel function and hence was also able to learn non-linear relations but reached a low performance nonetheless. We first argued that this might be due to its proneness to outliers (Wen, Hao, & Yang, 2010) and the possibility that the outlier analysis was not able to effectively identify all outliers (though it was very conservative). This problem could be addressed in future studies by applying more advanced methods for outlier detection or methods that reduce the effect of outliers. Yang, Huang, Chan, King, and Lyu (2004) propose a two-step procedure to attenuate the effect of outliers in SVM regression used with a non-linear radial basis function. As we described in chapter 1.4.4.3, the constant λ and the slack variables ξ_i and ξ_i^*, which indicate the positive and negative deviation from the tolerance margin, define the penalty term of the SVM estimation function. Consequently, outliers will lead to large values of ξ_i and ξ_i^* which increases the model error. The authors therefore advise training a SVM model with the margin tolerance parameter ε_i (for the non-linear SVM the margin width ε can vary). Afterward, they instruct to identify all data points whose ξ_i or ξ_i^* are larger than a certain threshold $\tau \cdot \varepsilon_i$ as outliers (i.e., all data points that deviate more than τ times from the tolerance margin) and subsequently increase the tolerance parameter ε_i when ξ_i or ξ_i^* is determined as an outlier. By locally increasing the margin width ε_i and training the model again, the effect of the respective outliers is attenuated by reducing the respective slack variables ξ_i or ξ_i^* (i.e., by reducing the deviation from the tolerance margin).

An elaborated method for removing outliers from multidimensional electromyography features based on a k-nearest neighbor algorithm and an Euclidean distance metric was introduced by Marateb, Rojas-Martínez, Mansourian, Merletti, & Mañanas Villanueva (2012). The algorithm determines for each observation the degree to which the data point deviates from its neighbors (i.e., the distance to the closest data points in the feature space) and consequently its degree of outlierness. Subsequently, the distribution of the resulting outlierness values is calculated and the best cutoff point to separate the bulk of data from the outliers is estimated. This method is an additional option to remove artifacts from physiological data (e.g., due to power line interference) and could be a useful addition to noise removal by filtering.

We have argued before that the default setting of the SVM kernel function (which is used to introduce non-linearity to the SVM; see Fröhlich & Zell, 2005), might also have affected the results. Fröhlich and Zell (2005) recommend tuning the parameters of the kernel function to achieve good results. All previously addressed approaches could be used to increase the performance of the SVM in future studies, or vice versa provide an explanation for the reduced predictive power of the model compared to the RF algorithm.

5.2.3 Predictability of Appraisals

For the RF model, we found a robust relation (i.e., a robust positive R^2) between the physiological channels and 13 of the 21 considered appraisals. For six of these dimensions, the study was the first to demonstrate that such a connection exists. Eight appraisal dimensions were not robustly predictable. Hence, it was concluded that these dimensions might be connected to physiological changes that were not assessed in the study. The finding could moreover indicate that the outcomes of different appraisal dimensions affect different sets of physiological channels. Similar to the variable appraisal set theory by Fernando et al. (2017), which assumes that different subsets of appraisals are important for the determination of different feelings or modal emotions, it is also possible that different subsets of appraisals determine the outcome of different physiological responses. This assumption is substantiated by the results of the first importance measure R_B^2 that determines the variance that can be explained by each physiological channel for each appraisal dimension in the absence of other physiological blocks. It is apparent that not all considered channels explained variance for all appraisals. HRV, for example, seemed to be an important predictor for the *pleasantness* appraisal but did not explain any variance for the dimensions *familiarity*, *conduciveness* or *goal/need importance*. Presumably, a different set of physiological variables has to be considered for a further investigation of those dimensions that were not predictable. Kreibig (2010) analyzed a large set of different cardiovascular, respiratory and electrodermal parameters and their connection to different emotions in a review of 134 publications. Besides HRV and EDA, he found that several more cardiovascular measures such as heart rate, forehead temperature, arterial pressure, and stroke volume as well as respiratory measures such as the respiration rate and hyperventilation were affected by different emotional states. Future studies should therefore also consider these physiological markers in the investigation of the appraisal-physiology link, expanding the findings that are provided by the present thesis.

5.2.4 Self and Protagonist Perspective

For the assessment of four appraisal dimensions (*goal/need importance*, *adjustment*, *conduciveness*, and *urgency*), we constructed two items asking the participants to rate the dimensions from the perspective of the perceived protagonist of the video sequence as well as from their own perspective. We assumed that the perception of the protagonist's point of view could be more important in the affective evaluation of a video. Based on the low inter-item correlations we treated these items as separate appraisal dimensions. It could be observed that three of the respective dimensions – *urgency*, *adjustment* and *goal/need importance* – were only predictable when appraised from the protagonist's perspective. Similarly, the *conduciveness* dimension reached a substantially higher R^2 when evaluated from the point of view of the protagonist. This finding together with the low inter-item correlations indicates that the new protagonist items actually constitute separate dimensions and that these had a stronger link to the physiological measures in the study. Generally, the appraisal dimensions proposed by Scherer (2001, 2009) do not really reflect that emotions can also be felt due to empathizing with another person (or a fictive character in a book or movie). The *cause agent* dimension only asks whether another person, oneself or natural forces caused an event, but it does not assess whether one passively or actively participated in a situation and whether the appraisal process refers to one's inner states or to the states one attributes to another person. When individuals feel happy at the end of a romantic movie in which the two protagonists fell in love, it is plausible that the happiness does not result from the fact that they can adjust very well to the consequences of the event (i.e., the happy ending) but from the belief that the protagonists can (a high adjustability is assumed by Scherer's happiness prototype displayed in Table 2 of chapter 1.4.1). This becomes clearer when considering the scenario of a scary movie in which the protagonist is threatened, killed or hurt and individuals experience the emotion fear. Scherer assumes that low adjustability is prototypical for this emotion (see Table 2 of chapter 1.4.1). But again, as individuals are most likely able to adapt to the outcome of the movie (probably as much as they can adapt to the consequences of a romantic movie), it seems obvious that the protagonist cannot adjust very well. Hence, in the case of passive observation and strong identification or empathizing, the appraisal evaluation actually concerns the believes that individuals have about the object of identification. While some appraisal dimensions should not be affected by the passivity or presence in the situation, such as *pleasantness* that is defined as the intrinsic pleasantness of an event independent of the state of an individual (i.e., independent of individual wishes, preferences, goals, etc.), some dimensions should be affected. For these appraisals, an additional set of dimensions has to exist that reflects the states that are ascribed to others. The

prototypical appraisal outcomes for these dimensions might be similar to their egocentric counterparts but could also deviate to some degree.

5.2.5 Feature Set

For the five physiological channels, a broad set of 134 features was constructed to characterize the signals extensively. The features were based on the descriptions of Phinyomark, Limsakul, and Phukpattaranont (2009) and Phinyomark, Phukpattaranont, and Limsakul (2012) for the *electromyography* (EMG) signals and on Vollmer (2015) for the *heart rate reliability* (HRV) data. Only a few of the proposed features were not considered, mostly because they were only applicable for a moving-window analysis approach where features are extracted from consecutive time bins of the signal. As discussed in the study, we also applied the constructed EMG features for the analysis of the *electrodermal activity* (EDA) signal, as most of the features were suitable for time series data in general, omitting only two features that yielded no variance on the EDA data. However, a couple of more specific features for EDA data exist that were not implemented. Shukla, Barreda-Angeles, Oliver, Nandi, and Puig (2019), for example, used a broad set of predictors that also contained features quantifying the rise times of the EDA amplitudes, in addition to frequency and amplitude features similar to the ones used in this work. The inclusion of further features more specific to the characteristics of the EDA signal could potentially increase the predictive power of the models and also have an effect on the EDA importance.

5.2.6 Handling of Correlated Features

A major difficulty with the feature set was the large number of correlated features. Even though many of the features were correlated based on their mathematical similarity, each feature could potentially describe a slightly different aspect of the respective signal and hence explain incremental variance. Moreover, due to crosstalk between EMG regions or even due to the same noise sources in the laboratory environment, features of different physiological channels could also be correlated to some degree. Correlated features do not have to be problematic for machine learning models per se, but the calculation of feature importance measures can be strongly affected (Nicodemus & Malley, 2009). When adding a correlated feature to a model, the importance of the associated features may decrease as the importance is then potentially split between both features (Molnar, 2019). The first attempt to handle this problem was a reduction of the dimensionality of the feature space for each physiological channel by building factor scores using exploratory factor analysis (EFA). Alternatively, we

tried to reduce the number of features by performing a feature selection based on pairwise correlations. As the factors proposed by the EFA were not interpretable and as both approaches (feature aggregation and selection) lead to a substantial loss in performance, the feature reduction was not realized. Instead, two blocked importance measures were constructed that were able to handle the correlative relations between the features in different ways. A feature permutation importance measure for the RF algorithm that is able to handle correlated features has previously been introduced by Strobl, Boulesteix, Kneib, Augustin, and Zeileis (2008). This importance measure is however computational costly and was therefore too time-consuming for the application with 13 models with 134 features each. The importance of each individual feature was moreover not very informative, as we were actually interested in the relevance of each of the five channels – therefore two blocked importance measures were implemented. The R_B^2 measure quantifies the variance that can be explained by each physiological block (containing all features of the respective channel). As the features belonging to other physiological blocks are not included in the RF model from which the importance is attained, correlative relations between the features of the considered block and the features of other physiological blocks do not affect the results. Moreover, as the importance is not evaluated for each feature separately but for the physiological block as a whole, the potential importance splitting between correlated features within the block does not affect the results as only the explained variance across all features is regarded. While the R_B^2 measure circumvents the problem of correlations to features of other blocks by excluding them from the model, the second importance measure ΔR_B^2 quantifies the relation between the blocks by indicating how much variance can be explained by the respective feature block beyond all other features. If a block does not reach a robust importance and hence explains no incremental variance, the features of the block have to be strongly correlated to features of at least one of the other blocks (given that the block is able to explain variance in the respective appraisal dimension to begin with). The ΔR_B^2 measure can however not depict with which blocks variance is shared – for this information a pairwise inspection of the blocks has to be conducted. We constructed the measure to determine if a physiological signal is needed in the prediction of the respective appraisal when finding the most economical model is the goal. But again, the measure cannot determine which set of blocks is sufficient to predict the appraisal without a loss in performance, but can only indicate that a block might be redundant if all other blocks are included. While the R_B^2 measure demonstrated that the different physiological channels contributed differently to the appraisal prediction and that a link between appraisal and physiology did not exist for all considered appraisal dimensions and all blocks in the study, the ΔR_B^2 measure additionally

showed that only a few channels uniquely explained variance confirming the correlative relations between the physiological channels.

5.2.7 Effect of Modeling Direction

The description of the physiological channels with numerous features is also the reason why the appraisal-physiology link was modeled in the reversed direction. In contrast to the causality that is implied by the CPM (Scherer, 2001, 2009), we predicted the appraisal dimensions from physiology instead of predicting changes in the physiological channels from the appraisal outcomes. To implement the theoretical implied modeling direction, a single physiological outcome variable would be needed. As one single feature cannot sufficiently describe the complex amplitude and frequency characteristics of the time-series signals and as the previously described feature-aggregation (EFA) and feature-selection (correlation-based) approaches did not provide a satisfactory solution, this way of modeling the appraisal-physiology link seemed to be not feasible. Moreover, the study design does not allow us to test the causality between appraisal and physiology. Even though the applied methodological approach is sufficient for examining whether a specific appraisal (such as *pleasantness*) relates to a certain physiological channel (such as zygomaticus activity), one aspect of the appraisal-physiology relation was not covered. The model cannot take into account that physiological changes might be caused by an interaction of several appraisals such as *pleasantness* and *suddenness*. Only when the modeling direction is reversed (predicting changes in a physiological channel using all appraisal dimensions as features), the effect of the appraisal interactions on physiological changes can be considered. Even though this aspect is definitely an interesting one when investigating the appraisal-physiology path, the interpretability of such interactions would remain difficult when using machine learning models. While it would be possible to construct importance measures to indicate how much variance is explained by an appraisal alone[11] or by all its interactions with other appraisals[12], the type and directions of the

[11] See the importance measure that was used in the feature selection for the ALE plots in chapter 4.5.7.3. This feature importance measure quantifies the variance that can be explained by a single feature without considering other variables.

[12] This type of importance measure could be created by taking the difference between a classical permutation feature importance as proposed by Molnar (2019) and the single feature importance used in chapter 4.5.7.3. As the classical permutation importance quantifies the variance that can be explained by a single feature and its interactions with all other features in a model, and the single feature importance shows the variance explained

specific interaction effects as well their magnitude could not be derived (information that would be attainable from linear regression for example). Consequently, modeling the appraisal-physiology link in the theoretically proposed direction would probably not have added much value beyond the presented results but could be implemented in future research with more interpretable models.

5.3 The Problem of Measurement Error

The biggest limitation of both studies is the measurement of the relevant variables such as appraisals and emotion labels using questionnaires. Every model, as well as the conclusions deduced from it, can only be as good as the measurement it is based on. The reliability of psychological variables assessed by self-report is a problem in many fields of psychology in which more objective and direct measures cannot be applied. Gnambs (2015), for example, demonstrated that nearly half of the variance in observed scores of personality questionnaires arises from measurement error. Measurement error in this context could, for example, result from inter-individual differences in item interpretation (Gnambs, 2015). In the case of cognitive appraisal, which is thought to be processed at least partially in an automated fashion (Scherer, 2001), the accessibility of the appraisal ratings could be an additional problem. The assessed appraisals have to be understood as an approximation of the appraisal process, given the assumed limited awareness of the process (a more detailed discussion on this topic and criticism on the assessment of appraisals by self-report is presented in study 1 in chapter 2.7). Therefore, substantial measurement error in the appraisal ratings of both studies has to be assumed.

The data used in study 1 were collected with the GEA tool (Scherer & Meuleman, 2013), a freely accessible web tool, over a period of several years. Though the authors excluded a small percentage of participants from the sample due to missing answers and response bias[13], it is not clear if this quality assessment was sufficient. Moreover, participants rated an emotional episode from their past so that the retrospective evaluation might have decreased the accessibility of the appraisal ratings even more. We therefore decreased the temporal distance to the evaluated event in study 2 to counteract this problem. To reduce the length of the testing

without these interactions, the difference would reflect the variance proportion explained by the interactions only. Note that this importance measure would require uncorrelated features though.

[13] Observations were considered as biased when two or less unique answers were given by the participant or when over 70% of the items were answered with the not applicable category.

to an acceptable duration and hence ensure a sufficient sample size, we used a shortened version of the GAQ to measure the appraisal dimensions, assessing most dimensions with a single-item. We also slightly altered the items of the original questionnaire to match the video rating context. Though several cognitive interviews were conducted in the development process of the adapted version, a full analysis of the test quality was not carried out. It must be assumed that these factors all influenced the reliability of the measured appraisal dimensions negatively. Similarly, measurement error has to be present in the emotion labels as well (that were used in study 1). Besides the fact that the emotion terms might have been understood differently (most likely reinforced by the semantic similarity of the emotion terms), participants were also forced to rate their feelings by choosing from a limited list of emotions. Even though they were also able to choose more than one label, this restriction to distinct categories might have also contributed to error in the emotion labels (assuming that a huge space of different emotion states exists and the emotion labels only represent the 13 modal emotions proposed by Scherer, 2001). Even though some improvements can be implemented in future research (e.g., not relying on single item measures), the described problems cannot be fully avoided. In the case of emotion labels, a clear ground truth is needed for the application of predictive models. In the case of appraisals, more objective measurement methods to assess the appraisal procedure are not available yet.

The physiological measures applied in study 2 can be deemed as more objective as they do not depend on self-awareness. At least for facial EMG though, effects of social desirability could have been present, leading participants to mask their facial expressions to some degree. Measurement error was however mainly introduced due to sources of noise and artifacts that were not canceled out in the laboratory environment. Various measures have been taken to limit the influence of these confounding variables as much as possible, such as positioning the experimenter out of sight of participants to decrease social desirability effects as well as using a bipolar recording scheme for EMG and EDA, applying an appropriate data preprocessing and constructing more robust features to reduce the influence of artifacts and noise.

When interpreting the results, it has to be taken into account how the measurement error might have influenced the findings. First of all, the presence of error potentially limited the reached performance of the predictive models of both studies. In study 2, it might have also concealed relationships between physiological variables and appraisals to some degree. However, the found relations (i.e., connections between appraisal dimensions and emotion labels as well as between physiological variables and appraisals) were all attained using cross-validation on large samples and in study 2 with an additional robustness criterion. The reported relations can hence confidently be considered as valid and robust.

5.4 Integrating the Results into a Multi-Componential Emotion Model

Scherer and Moors (2019) describe emotions "as an interface between an organism and its environment, constantly mediating between changing events and social contexts on the one hand and the individual's responses and experiences on the other" (p. 721). With the present thesis, two of the mediating sub-processes of this complex mechanism were investigated with the goal to increase the understanding of how the different components engaged in an emotional episode interact. In the following, the central paths of the multi-componential CPM model (presented in Figure 1 of chapter 1.3) will be addressed and integrated with the findings of the two studies.

5.4.1 Event to Appraisal

The initial path of the CPM is the one that interlinks the environment (i.e., a stimulus or an event) and the cognitive component (i.e., the appraisal process). As the appraisals are derived from cognitive elements such as memory, attention, and self-image, this initial path is highly individual. Due to the introduction of the cognitive component, the model is able to explain why the same stimulus might result in different emotional responses in individuals and within the same individual on different occasions. Unless researchers are able to access and measure all cognitive elements embedded in the appraisal process, the path is difficult to investigate. By using questionnaires such as the GAQ (Geneva Emotion Research Group, 2002), developed to approximate the appraisal process by asking participants to consciously rate the appraisal dimensions, it is possible to examine subsequent paths that connect the cognitive evaluation to other components such as physiology and feeling. It has been pointed out before that in the analysis of these paths we cannot investigate their causality and validate that the appraisal procedure is the initiating component within the examined processes as appraisal outcomes were not systematically induced during the data collection.[14]

Though the present thesis did not explicitly investigate the appraisal process itself, some conclusions about the dimensionality of the cognitive component can be drawn. Concerning the appraisal set proposed by Scherer (2001, 2009), we found all dimensions to be relevant

[14] As we have discussed in study 2, the systematic induction of appraisals poses several problems such as the lack of certainty about whether an experimental condition or stimulus actually lead to the presumed appraisal outcome and the fact that some appraisals simply cannot be induced experimentally. Moreover, as it is theoretically assumed that the appraisal process is highly individual, it is more generally questionable if stimuli can be constructed that universally lead to a specific appraisal rating.

predictors of the subjective feeling experienced by participants and most of the appraisals to be connected to emotion-related physiological responses. Both presented studies concluded though that the appraisal dimensions assumed by the CPM (and also appraisal sets with similar dimensions proposed by other appraisal representatives) are probably insufficient to fully explain variation in the multiple emotion components. Study 2 demonstrated that in the passive viewing of an event, additional appraisal dimensions might be relevant that refer to the states and appraisal evaluations that individuals attribute to others. Within current theories though, the cognitive appraisal process had only been described as an egocentric and self-evaluative procedure. Moreover, we found that some of the items of the GAQ (Geneva Emotion Research Group, 2002) that were constructed to measure the same appraisal most likely constitute separate appraisal dimensions that each explain incremental variance. The reached performance in study 1 also suggested that additional appraisals could improve the differentiability of emotion categories.

5.4.2 Appraisal to Physiology and Expression

As the expression component entails facial, vocal and gestural expressions (Scherer & Moors, 2019) that are also physiologically entangled, we combined the relation of appraisal and the expression component and appraisal and the physiological component to one single path in study 2 (see Figure 2 in chapter 1.4). We investigated physiological changes in HRV, EDA, and EMG, which can also be an indicator for overt mimics (Van Boxtel, 2010), and analyzed the connection between these physiological responses and different appraisals. Besides rather vague theoretical predictions on the relation between appraisal outcomes and physiology by Scherer (2009) and some empirical studies (e.g., Aue & Scherer, 2008; Kreibig, Gendolla, & Scherer, 2012; Lanctôt & Hess, 2007; van Reekum et al., 2004) whose weaknesses have been discussed in study 2, no information was present on the appraisal-to-physiology link except for the assumption that changes would occur continuously.

The comparison of different machine learning algorithms as well as the inspection of single feature ALE plots indicated that the relation between measured physiological channels and the appraisals is best represented by a non-linear model. Increasing ratings of *pleasantness*, for example, were connected to a non-linear decrease in corrugator activity, and a rise in *conduciveness* was linked to a non-linear increase in zygomaticus activity. Moreover, not all investigated appraisals showed a connection to all five physiological channels – some appraisals were not predictable at all, showing no relation to any of the channels, while others were related to only a subset of the investigated physiological responses. Hence, it can be

assumed that different aspects of physiology are linked to different appraisals. Given the assumed causality (i.e., appraisal initiates changes in the physiological component), this would mean that an event that is appraised as being very sudden affects the EDA of a participant, leads to changes in HRV, and an adaption of the individual's mimic. A very familiar event on the other hand also affects the activity of facial muscles but does not lead to changes in EDA or HRV (but maybe to changes in other physiological constituents such as body temperature or heart rate that were not investigated in study 2). The potential effect of interactions of different appraisals on physiology could not be investigated due to the used modeling direction. In Scherer's (2001) description of the appraisal process (see chapter 1.3), he indicates that each appraisal outcome leads to variations in all other components and modifies changes in these components that have been induced by previous appraisals. Though this description rather seems to indicate that the appraisals affect the other components independently from each other, his appraisal prototypes for different modal emotions (presented in Table 2 of chapter 1.4.1) do imply that interactions of appraisals affect the feeling component. Hence, it is plausible that interactions between appraisals could also explain variance in the physiological responses and should therefore also be considered in future investigations of the appraisal-physiology link. For such an analysis though, statistical models with better interpretability should be applied.

Although the results substantiate the link between cognition and physiology as assumed by appraisal theory, it is important to discuss the results' compatibility with other emotion theories as well. As the introduction of different theoretical approaches in chapter 1.2 suggested, a comparison of different models is challenging because they often focus on differing processing levels or diverging temporal stages of the emotion process, frequently leaving important aspects open or unspecified that are more clearly addressed in another theory. An attempt to draw these comparisons is made nevertheless.

The found relationship between cognitive appraisals and physiological responses is obviously incompatible with the outdated view that emotions have no cognitive component and are mere physiological constructs as described by James (1884). Schachter and Singer (1962; also Schachter, 1964) assume that emotions result from an interaction of physiology and cognition. They specifically state that the same physiological arousal can lead to a great diversity of different emotions depending on the cognitive evaluation it is accompanied by. This indicates that the physiological response and the cognitive evaluation must be independent

of each other to some degree.[15] The results of study 2 indicate though that cognition and physiology are interlinked with each other such that changes in one component are clearly accompanied by changes in the other. Moreover, the measured physiological reactions in response to the emotion videos were complex and multidimensional and hence not compatible with the idea of the unidimensional arousal that Schachter and Singer (1962) believe to be the physiological basis of every emotion (a critique that has been voiced early on by Plutchik & Ax, 1967).

Constructivist emotion theories see the ambiguous physiological arousal proposed by Schachter and Singer (1962) as "a historical predecessor of modern-day conception of `core affect´" (MacCormack & Lindquist, 2017, p.2). Such theories view the two physiological dimensions arousal and pleasantness as the primitive affect component that can be specified due to a cognitive processing step (Barrett, 2006; MacCormack & Lindquist, 2017; Russell, 2003). In contrast to Schachter and Singer (1962) though, Russell (2003) and Barrett (2006) presume that cognitive processes can also be the initial cause of a shift in core affect and hence in physiology (though core affect can also be initiated by other non-cognitive processes). Besides the assumption that cognition can be involved though, the mechanism is not further discussed. The found relation between cognitive appraisal and physiology is therefore potentially also compatible with this conceptualization of the emotion process.

Affect program theories such as the ones by Ekman (1992) and Panksepp (2005) propose a limited number of basic emotions and specific physiological patterns that function on distinct neuronal circuits occurring once an appropriate trigger is present. Even though these theories also recognize that appraisal can be a trigger for affect programs (e.g., Ekman, 1999), the assumption of a very limited number of prototypical physiological schemes does not seem to fit the results of study 2, where we showed that changes in individual appraisals are associated with varying changes in the physiological signals. This rather indicates that shifts in the appraisal dimensions are accompanied by a flexible adaption of the physiological subcomponents and consequently a very large variety of different physiological states. However, Scherer (2009) stresses that affect program theorists more recently moved to assign

[15] It has to be noted, though, that the authors do not elaborate the mechanism that leads to the emergence of the physiological arousal – a major shortcoming of the theory (as was discussed in chapter 1.2). As it is not proclaimed how the physiological arousal emerges, it cannot be precluded completely that a correlation between the cognitive and the physiological changes of some degree could exist.

higher flexibility to affect programs and also recognize more complex emotion categories besides the proposed basic ones.

The previous comparison shows that the direct link between cognitive elements and physiology is implicit in many emotion theories. It would be helpful if this connection was therefore more explicitly emphasized in other emotion models to reflect the results of this and other studies on the relation of cognition and physiology in emotion, but also to mitigate superficial differences between the emotion theories.

5.4.3 Appraisal to Feeling

The appraisal-feeling link, indicating which appraisal outcomes lead to which subjectively perceived feeling (and consequently to which verbal emotion label), was examined with different theoretically informed emotion models in study 1. The study demonstrated that verbal emotion labels can be predicted from self-rated appraisal dimensions to some degree, substantiating the appraisal-feeling link and verifying that the appraisals proposed by Scherer (2001, 2009) are indeed relevant predictors for the subjective feeling of an individual. Concerning the algorithmic level of this link, the study validated the idea that the appraisal dimensions are differently important in emotion prediction. Whether the appraisal weights are the same for all emotions (as in the preferred model M3) or whether they vary between them (as in M4) should be further investigated. The latter, though not the preferred model solution in the study, would be in line with the idea of the variable appraisal set theory (Fernando et al., 2017). We have argued in study 1 that the complex weighting algorithm might be too costly for a fast functioning process like emotion elicitation. This presumption though takes as the basis that each appraisal is evaluated and weighted and subsequently integrated. It could also be possible that only a subset of appraisals is processed to begin with, which would potentially even decrease the cognitive costs of the process. The prototype approach realized in the preferred model M3 seems to be an appropriate realization of the appraisal-to-feeling link, yielding an only slightly worse performance than the machine learning model. The comparison to the RF indicates though that an implementation of multiple prototype clusters for each emotion could be superior – an idea that has not been discussed in appraisal theory yet.

In regards to other emotion theories, study 1 again clearly demonstrated that cognition is a central element of emotion elicitation and hence irreconcilable with a non-cognitive emotion model as the one by James (1884). Schachter and Singer (1962) assume that emotions arise from the interaction of cognition and physiological arousal and that cognition alone is insufficient for the elicitation of an emotion. Nonetheless, we were able to predict emotion

labels from cognitive evaluations alone. As we have mentioned in study 1 though, the *urgency* appraisal dimension has previously been linked to physiological arousal (Scherer, 2000) and we also found that *urgency* was an important predictor of the participants' feelings (indicated by a high appraisal weight). This could be seen as partially compatible with the assumptions of Schachter and Singer (1962), that arousal is a predictor of emotions, given the assumption that a strong correlation between the *urgency* appraisal and physiological arousal actually exists.[16] However, it could also be compatible with the indirect path between appraisal and feeling that is mediated by the physiological component in the CPM model (see Figure 1 in chapter 1.3). The hypothetical strong relationship between the cognitive *urgency* dimensions and physiological arousal would simultaneously contradict the implicit assumption of the independence of cognition and physiology in Schachter and Singer's (1962) model.

Moors (2009) sees a big difference between the appraisal and constructivist theories (such as the one by Barrett, 2006) in the way they conceptualize the formation of the link between appraisals and emotion categories – while constructivists view this link to be learned and hence more individual, appraisal theorists believe that the algorithmic level of this link is fixed and hence the same for all individuals. With the prototypes and weighting parameters calculated from empirical data, we were able to reach a good predictive performance over a sample of 6591 participants, which indicates that the algorithmic level of the link has to be similar between individuals. If the link between appraisals and perceived feeling would be completely individual with a great variance between participants, a very low predictive power would be expected with a generic model. It would be interesting though to analyze how strong the model parameters would deviate between participants when attained from large within-subject samples, and if the predictive performance would increase by applying personalized prediction models. Even though Scherer (2009) points out several differences between appraisal and constructivist theories, mainly criticizing the idea of core affect being two-dimensional, the idea of prototypical component patterns in constructivist theories seems to be similar to the implementation of prototypical appraisal patterns in the present study (as well as to the approaches of Scherer, 1993; Scherer & Meuleman, 2013). The theories rather seem to differ concerning the origin of the prototypes as being biologically defined or constructed individually based on previous experience and culture. In regards to the criticisms of core affect, the used

[16] In study 2, we found that the HRV block explained variance in the urgency appraisal. As HRV is interpreted as an indicator of physiological arousal (Egger, Ley, & Hanke, 2019), it can be assumed that a connection between urgency and arousal exists to some degree.

data set in study 1 contained two appraisal dimensions that describe valence and arousal (*pleasantness* and *urgency*). Though both dimensions turned out to be important predictors of the perceived feeling of participants (i.e., attained high weighting parameters), the two dimensions alone were apparently not sufficient to differentiate between emotions, which underpins Scherer's (2009) critique (see also Fontaine, Scherer, Roesch, & Ellsworth, 2007).

Lastly, the observation that 72% of the participants in the data set of study 1 used two emotion labels to describe their perceived feeling indicates that participants experienced rather complex emotions. This observation questions the idea of a limited number of basic emotions as proposed by affect program theorists (e.g., Ekman, 1992; Panksepp, 2005). As remarked before though, affect program theorists moved to acknowledge more complex emotions as well (Scherer, 2009).

5.5 Conclusion

The present thesis aimed at investigating the role of cognition in the emotion elicitation process. The studies were based on the idea that the multi-componential emotion process has to be broken down to its different processing levels by analyzing the links between all engaged components to attain a holistic understanding of emotions. By using a theoretical modeling approach, we were able to model the link between cognitive appraisals and the verbally labeled feeling of participants. It was demonstrated that emotions can be predicted from cognitive appraisal dimensions substantiating the important role of cognition in emotion differentiation. By applying machine learning models to analyze the relations between appraisals and different physiological responses we were again able to provide evidence for the existence of the respective link. In regards to the algorithmic level of the two paths, assumptions made by appraisal theory were elaborated and extended by comparing different model implementations, deductively generating new model parameters from empirical data and applying methods for the interpretation of black-box models.

Scherer (1999) commented about the research on appraisal theories that "[t]heory development in this area may benefit from efforts to use computer modeling of appraisal theory, helping to test the consistency of predictions, simulate alternative outcomes, and evaluate alternative versions of theories" (p. 655). The two predictive modeling approaches presented in this thesis possess all of these valuable features and demonstrate how computational emotion models can advance emotion research.

5.6 References

Aue, T., & Scherer, K. R. (2008). Appraisal-driven somatovisceral response patterning: Effects of intrinsic pleasantness and goal conduciveness. *Biological Psychology*, *79*(2), 158–164. https://doi.org/10.1016/j.biopsycho.2008.04.004

Barrett, L. F. (2006). Solving the Emotion Paradox: Categorization and the Experience of Emotion. *Personality and Social Psychology Review*, *10*(1), 20–46. https://doi.org/10.1207/s15327957pspr1001_2

Egger, M., Ley, M., & Hanke, S. (2019). Emotion Recognition from Physiological Signal Analysis: A Review. *Electronic Notes in Theoretical Computer Science*, *343*, 35–55. https://doi.org/10.1016/j.entcs.2019.04.009

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, *6*(3–4), 169–200. https://doi.org/10.1080/02699939208411068

Ekman, P. (1999). Basic Emotions. In Tim Dalgleish & M. J. Power (Eds.), *Handbook of Cognition and Emotion* (pp. 45–60). https://doi.org/10.1002/0470013494.ch3

Ellsworth, P. C., & Smith, C. A. (1988). Shades of Joy: Patterns of Appraisal Differentiating Pleasant Emotions. *Cognition & Emotion*, *2*(4), 301–331. https://doi.org/10.1080/02699938808412702

Fernando, J. W., Kashima, Y., & Laham, S. M. (2017). Alternatives to the fixed-set model: A review of appraisal models of emotion. *Cognition and Emotion*, *31*(1), 19–32. https://doi.org/10.1080/02699931.2015.1074548

Fontaine, J. R. J., Scherer, K. R., Roesch, E. B., & Ellsworth, P. C. (2007). The World of Emotions is not Two-Dimensional. *Psychological Science*, *18*(12), 1050–1057. https://doi.org/10.1111/j.1467-9280.2007.02024.x

Fröhlich, H., & Zell, A. (2005). Efficient parameter selection for support vector machines in classification and regression via model-based global optimization. *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, *3*, 1431–1436. https://doi.org/10.1109/IJCNN.2005.1556085

Geneva Emotion Research Group. (2002). *Geneva Appraisal Questionnaire (GAQ)*. Retrieved from https://www.unige.ch/cisa/files/3414/6658/8818/GAQ_English_0.pdf

Gnambs, T. (2015). Facets of measurement error for scores of the Big Five: Three reliability generalizations. *Personality and Individual Differences*, *84*, 84–89. https://doi.org/10.1016/j.paid.2014.08.019

James, W. (1884). WHAT IS AN EMOTION? *Mind*, *9*(34), 188–205. https://doi.org/10.1093/mind/os-IX.34.188

Judd, C. M., & Park, B. (1993). Definition and assessment of accuracy in social stereotypes. *Psychological Review*, *100*(1), 109–128. https://doi.org/10.1037/0033-295X.100.1.109

Kreibig, S. D. (2010). Autonomic nervous system activity in emotion: A review. *Biological Psychology*, *84*(3), 394–421. https://doi.org/10.1016/j.biopsycho.2010.03.010

Kreibig, S. D., Gendolla, G. H. E., & Scherer, K. R. (2012). Goal relevance and goal conduciveness appraisals lead to differential autonomic reactivity in emotional responding to performance feedback. *Biological Psychology*, *91*(3), 365–375. https://doi.org/10.1016/j.biopsycho.2012.08.007

Lanctôt, N., & Hess, U. (2007). The timing of appraisals. *Emotion*, *7*(1), 207–212. https://doi.org/10.1037/1528-3542.7.1.207

Lazarus, R. S. (1991). *Emotion and Adaptation*. New York: Oxford University Press.

MacCormack, J. K., & Lindquist, K. A. (2017). Bodily Contributions to Emotion: Schachter's Legacy for a Psychological Constructionist View on Emotion. *Emotion Review*, *9*(1), 36–45. https://doi.org/10.1177/1754073916639664

Marateb, H. R., Rojas-Martínez, M., Mansourian, M., Merletti, R., & Mañanas Villanueva, M. A. (2012). Outlier detection in high-density surface electromyographic signals. *Medical & Biological Engineering & Computing*, *50*(1), 79–89. https://doi.org/10.1007/s11517-011-0790-7

Molnar, C. (2019). Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. Retrieved from https://christophm.github.io/interpretable-ml-book/

Moors, A. (2009). Theories of emotion causation: A review. *Cognition & Emotion*, *23*(4), 625–662. https://doi.org/10.1080/02699930802645739

Nicodemus, K. K., & Malley, J. D. (2009). Predictor correlation impacts machine learning algorithms: Implications for genomic studies. *Bioinformatics*, *25*(15), 1884–1890. https://doi.org/10.1093/bioinformatics/btp331

Ortony, A., Clore, G., & Collins, A. (1988). *The Cognitive Structure of Emotions*. Cambridge: Cambridge University Press.

Panksepp, J. (2005). *Affective neuroscience: The foundations of human and animal emotions*. Oxford: Oxford University Press.

Phinyomark, A., Limsakul, C., & Phukpattaranont, P. (2009). A Novel Feature Extraction for Robust EMG Pattern Recognition. *Journal of Computer Science*, *1*(1), 71–81.

Phinyomark, A., Phukpattaranont, P., & Limsakul, C. (2012). Feature reduction and selection for EMG signal classification. *Expert Systems with Applications*, *39*(8), 7420–7431. https://doi.org/10.1016/j.eswa.2012.01.102

Plutchik, R., & Ax, A. F. (1967). A Critique of Determinants of Emotional State by Schachter and Singer (1962). *Psychophysiology*, *4*(1), 79–82. https://doi.org/10.1111/j.1469-8986.1967.tb02740.x

Probst, P., Wright, M. N., & Boulesteix, A. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(3). https://doi.org/10.1002/widm.1301

Rosch, E. (1983). Prototype classification and logical classification: The two systems. In E. Scholnick (Ed.), *New trends in conceptual representation: Challenges to Piaget's theory* (pp. 73–86). Hillsdale, NJ: Erlbaum.

Roseman, I. J. (1984). Cognitive Determinants of Emotion: A Structural Theory. *Personality and Social Psychology Review*, *5*, 11–36.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, *110*(1), 145–172. https://doi.org/10.1037/0033-295X.110.1.145

Schachter, S. (1964). The Interaction of Cognitive and Physiological Determinants of Emotional State. In *Advances in Experimental Social Psychology* (Vol. 1, pp. 49–80). https://doi.org/10.1016/S0065-2601(08)60048-9

Schachter, S., & Singer, J. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological Review*, *69*(5), 379–399. https://doi.org/10.1037/h0046234

Scherer, K. R. (1984). On the nature and function of emotion: A component process approach. In K. R. Scherer & P. Ekman (Eds.), *Approaches to Emotion* (pp. 293–317). Hillsdale, NJ: Erlbaum.

Scherer, K. R. (1993). Studying the emotion-antecedent appraisal process: An expert system approach. *Cognition & Emotion*, *7*(3–4), 325–355. https://doi.org/10.1080/02699939308409192

Scherer, K. R. (1999). Appraisal Theory. In T. Dalgleish & M. J. Power (Eds.), *Handbook of Cognition and Emotion* (pp. 637–663). https://doi.org/10.1002/0470013494.ch30

Scherer, K. R. (2000). Psychological Models of Emotion. In J. Borod (Ed.), *The Neuropsychology of Emotion* (pp. 137–162). Oxford and New York: Oxford University Press.

Scherer, K. R. (2001). Appraisal considered as a process of multilevel sequential checking. In K. R. Scherer, A. Schorr, & J. Johnstone (Eds.), *Appraisal processes in emotion* (pp. 92–120). New York: Oxford University Press.

Scherer, K. R. (2009). The dynamic architecture of emotion: Evidence for the component process model. *Cognition & Emotion*, *23*(7), 1307–1351. https://doi.org/10.1080/02699930902928969

Scherer, K. R., & Meuleman, B. (2013). Human Emotion Experiences Can Be Predicted on Theoretical Grounds: Evidence from Verbal Labeling. *PLOS ONE*, *8*(3), e58166. https://doi.org/10.1371/journal.pone.0058166

Scherer, K. R., & Moors, A. (2019). The Emotion Process: Event Appraisal and Component Differentiation. *Annual Review of Psychology*, *70*(1), 719–745. https://doi.org/10.1146/annurev-psych-122216-011854

Shukla, J., Barreda-Angeles, M., Oliver, J., Nandi, G. C., & Puig, D. (2019). Feature Extraction and Selection for Emotion Recognition from Electrodermal Activity. *IEEE Transactions on Affective Computing*, 1–1. https://doi.org/10.1109/TAFFC.2019.2901673

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*(1), 307. https://doi.org/10.1186/1471-2105-9-307

Van Boxtel, A. (2010). Facial EMG as a Tool for Inferring Affective States. In A. J. Spink, F. Grieco, O. E. Krips, L. W. S. Loijens, L. P. J. J. Noldus, & P. H. Zimmerman (Eds.), *Proceedings of Measuring Behavior 2010*. Eindhoven, The Netherlands.

van Reekum, C., Johnstone, T., Banse, R., Etter, A., Wehrle, T., & Scherer, K. R. (2004). Psychophysiological responses to appraisal dimensions in a computer game. *Cognition & Emotion*, *18*(5), 663–688. https://doi.org/10.1080/02699930341000167

Vollmer, M. (2015). A robust, simple and reliable measure of heart rate variability using relative RR intervals. *2015 Computing in Cardiology Conference*, 609–612. https://doi.org/10.1109/CIC.2015.7410984

Wen, W., Hao, Z., & Yang, X. (2010). Robust least squares support vector machine based on recursive outlier elimination. *Soft Computing*, *14*(11), 1241–1251. https://doi.org/10.1007/s00500-009-0535-9

Yang, H., Huang, K., Chan, L., King, I., & Lyu, M. R. (2004). Outliers Treatment in Support Vector Regression for Financial Time Series Prediction. In N. R. Pal, N. Kasabov, R. K. Mudi, S. Pal, & S. K. Parui (Eds.), *Neural Information Processing* (Vol. 3316, pp. 1260–1265). https://doi.org/10.1007/978-3-540-30499-9_196

6 Appendix – German Summary

Trotz einer Vielzahl unterschiedlicher theoretischer Ansätze zur Erklärung von Emotionen, herrscht unter Emotionspsychologen weitgehende Einigkeit darüber, dass sich Emotionen aus multiplen Komponenten wie Kognition, Physiologie, Motivation und dem subjektiv erlebten Gefühl von Personen zusammensetzten. Unterschiede zwischen den Theorien existieren vor allem bezüglich der zeitlichen Reihenfolge, in der die einzelnen Komponenten auftreten und der Art und Weise, wie diese interagieren (für einen Überblick siehe Moors, 2009). Um den multidimensionalen Emotionsprozess zu untersuchen, betrachtet die vorliegende Dissertation zwei Subprozesse, die das Zusammenspiel der kognitiven Komponente mit der physiologischen Komponente (Studie 2) sowie mit der subjektiven Gefühlskomponente (Studie 1) beschreiben. Als theoretische Grundlage dienen der Arbeit sogenannte Appraisal-Theorien (z.B., Arnold, 1960; Frijda, 1986; Lazarus, 1991; Ortony, Clore, & Collins, 1988; Scherer, 1984; Smith & Ellsworth, 1985), insbesondere das *Component Process Model* (CPM) von Scherer (1984, 2001, 2009). Diese gehen davon aus, dass die kognitive Evaluation eines Reizes auf mehreren emotionsrelevanten Bewertungsdimensionen das zentrale Element eines jeden Emotionsprozesses ist und Veränderungen in allen anderen Komponenten durch diese kognitive Evaluation initiiert werden. Die zwei vorgestellten Studien haben das Ziel, die jeweiligen Zusammenhänge zwischen den Emotionskomponenten aufzuzeigen und somit die zentrale Rolle der kognitiven Komponente zu validieren. Darüber hinaus sollen die algorithmischen Eigenschaften der beiden Prozesse analysiert werden. Die Studien greifen dabei auf unterschiedliche prädiktive Modellierungsansätze zurück – theoretisch-informierte Modellierung und uninformierte Machine-Learning-Modelle (sogenannte Black-Box-Modelle).

6.1 Studie 1: Der Appraisal-Gefühl-Pfad

6.1.1 Theorie

Scherers (2001) CPM geht von 16 unterschiedlichen Appraisal-Dimensionen aus, durch die die Relevanz (*relevance detection*) und möglichen Auswirkungen eines Ereignisses (*implication assessment*), das Potential mögliche Folgen zu bewältigen (*coping potential*) sowie die normative Signifikanz (*normative significance*) eines Ereignisses bewertet werden. Er geht weiter davon aus, dass jede der 16 Dimensionen auf einer kontinuierlichen Skala bewertet wird, wodurch eine große Anzahl an unterschiedlichen Bewertungsmustern entstehen

kann und folglich eine große Anzahl unterschiedlicher Emotionszustände. Scherer (2001) nimmt jedoch an, dass manche dieser Appraisal-Kombinationen öfter auftreten als andere. Er bezeichnet diese häufigeren Emotionszustände, die außerdem mit Emotionsbegriffen benannt werden können, als Modalemotionen (*modal emotions*). Aus der Annahme, dass der kognitive Appraisal-Prozess Veränderungen in allen anderen Komponenten initiiert und so zwischen verschiedenen Emotionen differenziert, leitet Scherer (2001) ab, dass subjektiv-erlebte Gefühle (bzw. die Emotionsbegriffe mit denen Personen diese beschreiben) aus Appraisal-Bewertungsmustern vorhersagbar sein sollten. Dabei geht er außerdem davon aus, dass die einzelnen Appraisal-Dimensionen unterschiedlich wichtig für die Differenzierung von erlebten Emotionen sind.

6.1.2 Methode

Aufgrund der relativ detaillierten Annahmen die das CPM bezüglich des Appraisal-Gefühl-Pfads trifft, wurde eine theoretisch-informierte Modellierung für die Untersuchung dieses Zusammenhangs gewählt. Dabei wurden Scherers (2001) theoretische Hypothesen in ein mathematisches Modell übersetzt und anhand ihrer Prädiktionskraft evaluiert. Sollte die Theorie den Prozess korrekt abbilden, so sollte das computationale Modell in der Lage sein, auf Basis von empirisch erfassten Appraisal-Bewertungen, das dazugehörige erlebte Gefühl von ProbandInnen korrekt vorherzusagen.

Die Verrechnung der Appraisal-Bewertungen zur Bestimmung des erlebten Gefühls wurde als eine Distanzmetrik zur verschiedenen, ebenfalls empirisch erfassten Emotionsprototypen umgesetzt. Dabei wird die Ähnlichkeit des empirisch erfassten Appraisal-Musters zu einem prototypischen Appraisal-Muster bestimmt, dass die durchschnittlichen Appraisalbewertungen über Beobachtungen, die mit dem gleichen Emotionsbegriff beschrieben wurden, abbildet. Es wird die Emotion vorhergesagt, deren Emotionsprotyp am ähnlichsten zum jeweiligen Input-Muster ist. Über vier unterschiedliche Modelle dieser Art wurde außerdem die Gewichtung der Appraisal-Dimensionen innerhalb der Distanzberechnung variiert. Dabei wurden unterschiedlich komplexe Gewichtungsfunktionen sowie aus der Theorie abgeleitete und empirisch geschätzte Gewichtungsparameter verglichen. Für die Evaluierung der out-of-sample Modelperformanz sowie die Schätzung der Modellparameter (Prototypen und Gewichtungsparameter) wurde ein Datensatz (n = 6591; weiblich = 4491) von Scherer and Meuleman (2013) verwendet. In der zugehörigen Studie wurden ProbandInnen gebeten sich an eine emotionale Episode aus ihrer Vergangenheit zu erinnern, das jeweils erlebte Gefühl mit Emotionsbegriffen zu beschreiben sowie die 16 Appraisal-Dimensionen des

CPMs mit einem hierfür entwickelten Fragebogen zu bewerten (Geneva Emotion Research Group, 2002).

6.1.3 Ergebnisse und Diskussion

Die Studie zeigt, dass die Verknüpfung zwischen den emotions-relevanten Appraisal-Dimensionen und dem subjektiven Gefühl von ProbandInnen tatsächlich existiert (konsistent zu den Befunden von Scherer, 1993; Scherer & Meuleman, 2013). Unabhängig von ihrem Gewichtungsalgorithmus waren alle vier Modelle in der Lage das erlebte Gefühl (d.h. die Emotionsbegriffe) besser vorherzusagen als ein naives Baseline-Modell. Die Höhe der erreichten Performanz variierte jedoch zwischen den Modellen. Basierend auf der out-of-sample Vorhersagekraft (insgesamt sowie für die einzelnen Emotionsklassen und -familien), der Fähigkeit die Verteilung der Emotionsklassen im Datensatz korrekt wiederzugeben sowie der Sparsamkeit des Models, wurde Modell M3 als das bevorzugte klassifiziert. Dieses Modell gewichtet die 16 Appraisal-Dimensionen unterschiedlich stark so wie es von Scherer (2001) theoretisch angenommen wurde und verwendet Gewichtungsparameter, die mithilfe eines genetischen Algorithmus (*Differential Evolution*; Storn & Price, 1997) aus den empirischen Daten geschätzt wurden. Eine Gleichgewichtung der Appraisal-Dimensionen (Model M1), die Verwendung von aus der Theorie abgeleiteten Gewichtungsparametern (Model M2) sowie ein komplexerer Gewichtungsalgorithmus (Model M4) schienen die Prädiktionsleistung dagegen einzuschränken.

Der Vergleich zu einem baumbasierten Machine-Learning-Algorithmus (*Random Forest*; siehe James, Witten, Hastie, & Tibshirani, 2013) zeigte darüber hinaus, dass die Vorhersagekraft mit einer höheren Modellkomplexität noch etwas gesteigert werden kann. Die leicht höhere Performanz des Random Forests wurde auf dessen Fähigkeit zurückgeführt, mehrere Klassifizierungsräume für die einzelnen Emotionen zu erlernen. Im Rückschluss auf den theoretisch-informierten Prototypenansatz wurde deshalb angenommen, dass die Implementierung mehrerer Prototypen (d.h., mehrerer prototypischer Appraisal-Muster) pro Emotion die Performanz der theoretischen Modelle potentiell verbessern könnte.

Auch mit dem Machine-Learning-Modell konnte jedoch keine perfekte Vorhersage der Emotionsklassen erreicht werden, was unter anderem auf einen hohen Messfehler in den erhobenen Variablen hindeutet. Als eine weitere potentielle Limitation der Studie wurde außerdem die Art der Prototypen-Berechnung angeführt. Die aus dem empirischen Datensatz berechneten Prototypen stellen die mittlere Ausprägung aller Appraisal-Muster dar, die mit der jeweiligen Emotion beschrieben wurden. Da die ProbandInnen jedoch sehr häufig zwei

Emotionsbegriffe zur Beschreibung ihrer Gefühle wählten (in 72% aller Beobachtungen) und solche ambigen Beobachtungen jeweils in die Berechnung zweier unterschiedlicher Emotionsprototypen eingingen, kann dies die Differenzierbarkeit der Prototypen beeinflusst haben.

6.2 Studie 2: Der Appraisal-Physiologie-Pfad

6.2.1 Theorie

Im Gegensatz zum Appraisal-Gefühl-Pfad macht das CPM nur wenige konkrete Annahmen zum Zusammenhang von Appraisal und Physiologie. Für zehn der Appraisal-Dimensionen im CPM formuliert Scherer (2009) Vermutungen über deren Effekt auf die physiologische Komponente (z.B. als angenehm bewerte Reize gehen mit einem Anstieg in der Herzfrequenz einher, während unangenehme Reize zu einer niedrigeren Herzfrequenz führen). Da diese Annahmen recht unkonkret sind und außerdem auf rein theoretischen Überlegungen basieren, sind diese als eher spekulativ einzuordnen. Darüber hinaus gibt es auch einige empirische Studien, die den Zusammenhang zwischen einzelnen Appraisal-Dimensionen und verschiedenen physiologischen Reaktionen untersucht haben (z.B., Aue, Flykt, & Scherer, 2007; Aue & Scherer, 2008; Delplanque et al., 2009; Gentsch, Grandjean, & Scherer, 2013; Kreibig, Gendolla, & Scherer, 2012; Lanctôt & Hess, 2007; van Reekum et al., 2004). Obwohl die Studien einen ersten Einblick in den Appraisal-Physiologie-Pfad bieten, weisen sie jedoch einige Schwachstellen auf (z.B., kleine Stichproben, Fokus auf einige wenige Appraisal-Dimensionen, experimentelle Induktion von Appraisal-Bewertungen ohne ausreichende Kontrolle über tatsächliche Effekte der Experimentalbedingungen).

6.2.2 Methode

Aufgrund der weniger verlässlichen theoretischen sowie empirischen Annahmen über den Appraisal-Physiologie-Pfad, wurde in Studie 2 eine uninformierte Machine-Learning-Modellierung herangezogen. Statt den mathematischen Zusammenhang wie in Studie 1 vorher zu definieren, sind Machine-Learning-Modelle in der Lage die Beziehung zwischen Input (hier Appraisal-Dimensionen) und Output (hier physiologische Reaktionen) selbstständig abzubilden. In der Studie wurden ProbandInnen verschiedene emotionale Videosequenzen vorgespielt, während fünf verschiedenen physiologische Signale erhoben wurden – Elektromyographie an drei Gesichtsmuskeln (zygomaticus major, frontalis, corrugator supercilii), Hautleitfähigkeit und Herzratenvariabilität. Im Anschluss bewerteten die

ProbandInnen wiederrum Appraisal-Dimensionen mithilfe eines Fragebogens, der auf dem *Geneva Appraisal Questionnaire* (GAQ; Geneva Emotion Research Group, 2002) basiert. Insgesamt gingen 1556 Beobachtungen von 157 Versuchspersonen (weiblich = 95) in die nachfolgende Modellierung ein.

Zur Beschreibung der physiologischen Kanäle wurden insgesamt 134 Features berechnet, die die Amplituden- und Frequenzeigenschaften der jeweiligen Signale charakterisierten. Verschiedene (lineare und non-lineare) Machine-Learning-Algorithmen wurden trainiert, um mithilfe der berechneten physiologischen Features die erhobenen Appraisal-Dimensionen vorherzusagen. Zur weiteren Interpretation der Modelle, wurden verschiedene Methoden zur Erhöhung der Interpretierbarkeit von Machine-Learning-Modellen angewendet (Molnar, 2019).

6.2.3 Ergebnisse und Diskussion

Die beste Vorhersage der Appraisal-Dimensionen durch die physiologischen Prädiktoren wurde erneut mit einem baumbasierten Machine-Learning-Algorithmus erzielt. Der Random Forest erreichte ein R^2, das für die einzelnen Appraisal-Dimensionen zwischen -.016 (*urgency*-Dimension) und .407 (*pleasantness*-Dimension) schwankte. Da dieses Modell auch in der Lage ist non-lineare Zusammenhänge abzubilden und eine deutlich bessere Performanz als das lineare Lasso-Modell (siehe James et al., 2013) aufwies, wurde gefolgert, dass der Zusammenhang zwischen den Appraisal-Dimensionen und den physiologischen Kanälen nicht linear ist. Diese Annahme wurde außerdem durch die deskriptive Analyse der Accumulated-Local-Effects-Plots (ALE) unterstützt, die für mögliche Werte eines einzelnen Features den Effekt auf die abhängige Variable (hier die jeweilige Appraisal-Dimension) visualisieren (Molnar, 2019) und die ebenfalls non-lineare Relationen abbildeten. Nicht alle der untersuchten Appraisal-Dimensionen konnten jedoch mit den physiologischen Features robust vorhergesagt werden. Deshalb wurde angenommen, dass diese Appraisal-Dimensionen möglicherweise mit anderen physiologischen Kanälen zusammenhängen, die in der Studie nicht berücksichtig wurden. Durch die Konstruktion zweier verschiedener Importance-Maße für die fünf physiologischen Signale, konnte gezeigt werde, dass auch für die Dimensionen, die vorhergesagt werden konnten nicht alle Blöcke gleich viel Varianz aufklärten. Außerdem wurde gezeigt, dass nur wenige physiologische Kanäle inkrementelle Varianz in den Appraisal-Dimensionen aufklärten, was wahrscheinlich auf korrelative Zusammenhänge zwischen den Features zurückzuführen ist.

Neben den Items des GAQ (Geneva Emotion Research Group, 2002), der auch zur Erfassung der Appraisal-Dimensionen für Studie 1 verwendet wurde, wurden für Studie 2 weitere Items konstruiert, die nach der Evaluation der Appraisal-Dimensionen aus Perspektive des Video-Protagonisten fragten. Dabei wurde für diese Dimensionen ein höheres R^2 erreicht als für die egozentrisch evaluierten Dimensionen. Dies weist darauf hin, dass in passiven Beobachtungssituationen, in denen sich in einen Protagonisten eingefühlt wird, ein weiteres Set an Appraisal-Dimensionen relevant ist.

6.3 Konklusion

Die präsentierten Studien sind in der Lage, den Zusammenhang der kognitiven Evaluation eines Stimulus zum erlebten Gefühl sowie zu physiologischen Reaktionen während einer emotionalen Episode nachzuweisen. Damit bestätigen die Ergebnisse die zentrale Rolle der kognitiven Komponente, wie sie von Scherer (1984, 2001, 2009) angenommen wird. Mithilfe der angewendeten prädiktiven Modellierungsansätze konnten Evidenz für bestehende Annahmen seiner Theorie gesammelt sowie neues Wissen aus empirischen Daten generiert werden.

6.4 Literaturverzeichnis

Arnold, M. B. (1960). *Emotion and personality*. New York: Columbia University Press.

Aue, T., Flykt, A., & Scherer, K. R. (2007). First evidence for differential and sequential efferent effects of stimulus relevance and goal conduciveness appraisal. *Biological Psychology*, *74*(3), 347–357. https://doi.org/10.1016/j.biopsycho.2006.09.001

Aue, T., & Scherer, K. R. (2008). Appraisal-driven somatovisceral response patterning: Effects of intrinsic pleasantness and goal conduciveness. *Biological Psychology*, *79*(2), 158–164. https://doi.org/10.1016/j.biopsycho.2008.04.004

Delplanque, S., Grandjean, D., Chrea, C., Coppin, G., Aymard, L., Cayeux, I., … Scherer, K. R. (2009). Sequential unfolding of novelty and pleasantness appraisals of odors: Evidence from facial electromyography and autonomic reactions. *Emotion*, *9*(3), 316–328. https://doi.org/10.1037/a0015369

Frijda, N. H. (1986). *The emotions*. Cambridge: Cambridge University Press.

Geneva Emotion Research Group. (2002). *Geneva Appraisal Questionnaire (GAQ)*. Retrieved from https://www.unige.ch/cisa/files/3414/6658/8818/GAQ_English_0.pdf

Gentsch, K., Grandjean, D., & Scherer, K. R. (2013). Temporal dynamics of event-related potentials related to goal conduciveness and power appraisals. *Psychophysiology*, *50*(10), 1010–1022. https://doi.org/10.1111/psyp.12079

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. https://doi.org/10.1007/978-1-4614-7138-7

Kreibig, S. D., Gendolla, G. H. E., & Scherer, K. R. (2012). Goal relevance and goal conduciveness appraisals lead to differential autonomic reactivity in emotional responding to performance feedback. *Biological Psychology*, *91*(3), 365–375. https://doi.org/10.1016/j.biopsycho.2012.08.007

Lanctôt, N., & Hess, U. (2007). The timing of appraisals. *Emotion*, *7*(1), 207–212. https://doi.org/10.1037/1528-3542.7.1.207

Lazarus, R. S. (1991). *Emotion and Adaptation*. New York: Oxford University Press.

Molnar, C. (2019). Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. Retrieved from https://christophm.github.io/interpretable-ml-book/

Moors, A. (2009). Theories of emotion causation: A review. *Cognition & Emotion*, *23*(4), 625–662. https://doi.org/10.1080/02699930802645739

Ortony, A., Clore, G., & Collins, A. (1988). *The Cognitive Structure of Emotions*. Cambridge: Cambridge University Press.

Scherer, K. R. (1984). On the nature and function of emotion: A component process approach. In K. R. Scherer & P. Ekman (Eds.), *Approaches to Emotion* (pp. 293–317). Hillsdale, NJ: Erlbaum.

Scherer, K. R. (1993). Studying the emotion-antecedent appraisal process: An expert system approach. *Cognition & Emotion*, *7*(3–4), 325–355. https://doi.org/10.1080/02699939308409192

Scherer, K. R. (2001). Appraisal considered as a process of multilevel sequential checking. In K. R. Scherer, A. Schorr, & J. Johnstone (Eds.), *Appraisal processes in emotion* (pp. 92–120). New York: Oxford University Press.

Scherer, K. R. (2009). The dynamic architecture of emotion: Evidence for the component process model. *Cognition & Emotion*, *23*(7), 1307–1351. https://doi.org/10.1080/02699930902928969

Scherer, K. R., & Meuleman, B. (2013). Human Emotion Experiences Can Be Predicted on Theoretical Grounds: Evidence from Verbal Labeling. *PLOS ONE*, *8*(3), e58166. https://doi.org/10.1371/journal.pone.0058166

Smith, C. A., & Ellsworth, P. C. (1985). Patterns of cognitive appraisal in emotion. *Journal of Personality and Social Psychology*, *48*(4), 813–838. https://doi.org/10.1037/0022-3514.48.4.813

Storn, R., & Price, K. (1997). Differential Evolution – A Simple and Efficient Heuristic for global Optimization over Continuous Spaces. *Journal of Global Optimization*, *11*(4), 341–359.

van Reekum, C., Johnstone, T., Banse, R., Etter, A., Wehrle, T., & Scherer, K. R. (2004). Psychophysiological responses to appraisal dimensions in a computer game. *Cognition & Emotion*, *18*(5), 663–688. https://doi.org/10.1080/02699930341000167