Appraisal-Driven Facial Actions as Building Blocks for Emotion Inference

Scherer, Klaus R.; Mortillaro, Marcello; Rotondi, Irène; Sergi, Ilaria; Trznadel, Stéphanie

Journal of Personality and Social Psychology
2018, Vol. 114, No. 3, 358–379

0022-3514/18/$12.00 http://dx.doi.org/10.1037/pspa0000107

Appraisal-Driven Facial Actions as Building Blocks for Emotion Inference

Klaus R. Scherer
University of Geneva and University of Munich

Marcello Mortillaro, Irene Rotondi, Ilaria Sergi, and Stéphanie Trznadel
University of Geneva

Although research on facial emotion recognition abounds, there has been little attention on the nature of the underlying mechanisms. In this article, using a "reverse engineering" approach, we suggest that emotion inference from facial expression mirrors the expression process. As a strong case can be made for an appraisal theory account of emotional expression, which holds that appraisal results directly determine the nature of facial muscle actions, we claim that observers first detect specific appraisals from different facial muscle actions and then use implicit inference rules to categorize and name specific emotions. We report three experiments in which, guided by theoretical predictions and past empirical evidence, we systematically manipulated specific facial action units individually and in different configurations via synthesized avatar expressions. Large, diverse groups of participants judged the resulting videos for the underlying appraisals and/or the ensuing emotions. The results confirm that participants can infer targeted appraisals and emotions from synthesized facial actions based on appraisal predictions. We also report evidence that the ability to correctly interpret the synthesized stimuli is highly correlated with emotion recognition ability as part of emotional competence. We conclude by highlighting the importance of adopting a theory-based experimental approach in future research, focusing on the dynamic unfolding of facial expressions of emotion.

Keywords: appraisal inference, appraisal theories of emotion, emotion recognition, emotional competence, facial expression

Supplemental materials: http://dx.doi.org/10.1037/pspa0000107.supp

How well can we recognize emotions from facial expressions? This question, first posed by Charles Darwin in 1872, has been empirically examined from the beginning of the 20th century. A large number of studies have shown that observers can reliably recognize a number of major emotions from prototypical facial expressions with above-chance accuracy. Although there are individual differences and confusions between emotion categories, overall people do remarkably well. In a comprehensive review of this literature, Scherer, Clark-Polner, and Mortillaro (2011) computed mean accuracy rates of more than 70% for static expressions (photos) and of more than 60% for dynamic expressions (video) of six major emotions for Western encoders and decoders. Recognition rates across Western and Non-Western cultures were lower but still much above chance level.

Research in this tradition starts from the assumption that specific emotions lawfully produce corresponding facial expression configurations and that the latter allow observers to recognize the respective emotions. The theoretical bases for this assumption have been quite diverse, including basic or discrete emotion models, dimensional models, adaptational models, circuit models, motivational models, and appraisal models (see Scherer, 2001, for a survey). In recent years, this basic assumption has come under attack claiming that there is no incontrovertible evidence that emotions produce specific facial expressions nor that emotions can be reliably "recognized" by observers in different cultures (see chapters 6, 7, 18, 22, 24 in Fernández-Dols & Russell, 2017). However, there is massive evidence, reviewed later in this section, that emotions often *do* produce specific facial expressions that *can* be reliably "recognized" by observers under certain conditions. To reasonably discuss how often and when this is the case, one needs to agree on a definition of emotion and on a stringent theoretical framework specifying the underlying mechanisms. These central issues are but rarely addressed in research on the expression and recognition of emotion, which is unfortunate because in the absence of precise theoretical notions about lawful mechanisms it is impossible to specify precise hypotheses that can be empirically tested.

Research on the mechanisms of facial emotion perception has mostly targeted the neural structures and pathways underlying the recognition process, including deficits in individuals with autism, schizophrenia, or a variety of neurological disorders (Adolphs,

Klaus R. Scherer, Department of Psychology, University of Geneva and Department of Psychology, University of Munich; Marcello Mortillaro, Irene Rotondi, Ilaria Sergi, and Stéphanie Trznadel, Swiss Center for Affective Sciences, University of Geneva.

Stéphanie Trznadel is now at the Wyss Foundation, Geneva.

This work was supported by funds granted to K. R. Scherer under Swiss National Science Foundation Grant 100014-122491 and European Research Council (ERC) Advanced Grant PROPEREMO (230331).

None of the data appearing in the manuscript have been disseminated in print or on the web. Selected parts of the data have been shown at conferences in more general talks describing the research of the first two authors.

Correspondence concerning this article should be addressed to Klaus R. Scherer, Department of Psychology, University of Geneva, Boulevard du Pont-d'Arve, 40, CH-1211 Geneva, Switzerland. E-mail: klaus.scherer@unige.ch

2002; Meaux & Vuilleumier, 2016). In addition, many researchers, especially in the area of automatic emotion classification, have been interested in whether the nature of the recognition process is more holistic, based on the facial configuration as a Gestalt, or more analytical, focusing on specific features (Tanaka, Kaiser, Butler, & Le Grand, 2012). Another research direction has examined context effects or social functions of facial emotion recognition (Hareli & Hess, 2012). Yet another issue investigated is whether judges infer discrete emotions, affective dimensions, or both (Mendolia, 2007). Although this research has yielded important information, it has not yet led to the identification of the fundamental mechanisms involved. Many authors in this domain seem to assume, explicitly or implicitly, that the underlying process consists of *template matching*; that is, perceivers, consciously or not, match the expressions to a set of (innate or learnt) emotion expression templates and then attribute the respective label. However, the nature of these templates is unclear, as is how many there are and how they originated. Other authors privilege feature-based processing, but, again, the nature of these features, how many there are, and how they are configured is unclear. Unfortunately, none of these theoretical accounts allows formulating precise hypotheses that can be experimentally examined.

In this article, we directly address this important issue. We suggest a reverse engineering approach to this question and argue that the emotion recognition or inference mechanism mirrors the externalization, or production mechanism. More concretely, we claim that expression is lawfully driven by the result of appraisal processes and that in attempting to recognize an emotion, we infer the nature of these appraisal results and attribute emotion descriptors that best capture the emoter's appraisal-driven reactions. To examine these claims empirically, we propose a new research paradigm that focuses on experimentally producing the theoretically predicted facial configurations resulting from specific appraisal results and assessing observers' inferences. We aim to determine to what extent judges associate specific expressive patterns with specific appraisals and specific emotions. To this effect, we do not use actors' interpretations of prototypical emotion expressions (as in most of the previous literature) but synthetic expressive configurations based on theoretical considerations and empirical evidence on the role of appraisal in shaping facial expressions in dynamic emotion processes.

We will use the notion of facial *action units* (AUs; coordinated innervations of facial muscle groups) as specified by the Facial Action Coding System (FACS) proposed by Ekman and Friesen (1978; Ekman, Friesen, & Hager, 2002) to characterize these expressions. To facilitate the somewhat complex specifications of these AUs, the Appendix contains a list with natural language glosses of the muscle groups as well as a photo illustration. Specifically, we predict that the observer uses specific facial movements to infer the results of specific appraisal checks, such as goal conduciveness or coping potential. Implicit attribution rules about appraisal configurations are subsequently used to categorize and label specific emotions or affective states. We hypothesize that the inference rules from facial AUs to specific appraisals on the perception side match the production rules on the production side. We describe one laboratory study and two web experiments with large survey panels in which specific facial AUs are manipulated individually and in combination via synthesized avatar expressions to examine the extent to which participants can infer the predicted appraisals and what emotion labels are attributed to specific appraisal configurations. Although evidence for these elements does not necessarily support the larger claim for an appraisal-based recognition mechanism, it seems to be an essential element to establish its plausibility.

Theoretical and Empirical Background

Component Process Model

Given the central role of the expression mechanism in this article, we first need to outline our theory on the architecture of facial expression and the available evidence. We base our claims on the Component Process Model of emotion (CPM; for detailed descriptions, see Scherer, 1984, 1986, 2001, 2009), which is part of the appraisal theory tradition (see Ellsworth & Scherer, 2003; Scherer, Schorr, & Johnstone, 2001). The CPM assumes that emotions are brief episodic processes during which several organismic subsystems temporarily work together in synchrony, driven by the appraisal of events that are highly relevant for an individual. These appraisals generate motivational effects accompanied by changes in expression, autonomic physiology, and feeling. As emotions are phylogenetically functional, changes in facial and vocal expression allow observers to infer appraisal results and the emotions that are generated in consequence. The production side of the proposed mechanisms is graphically depicted in Figure 1. The central assumption is that emotion episodes are triggered by appraisal (which can occur at multiple levels of cognitive processing, from automatic template matching to complex analytic reasoning) of events, situations, and behaviors (by oneself and others) that are of central significance for an organism's well-being, given their potential consequences and the resulting need to urgently react to the situation. In contrast to other appraisal theories, the CPM is based on a sequential-cumulative mechanism, as shown in Figure 1. The appraisal criteria are evaluated one after another (sequence of appraisal checks) in that each subsequent check builds on the outcome of the preceding check and further differentiates and elaborates on the meaning and significance of the event for the organism and the potential response options: Is the event novel (sudden, unpredictable) or familiar? Is it intrinsically pleasant or unpleasant? Does it help or hinder the attainment of relevant goals? Does the organism have sufficient coping potential (e.g., physical or social power) to deal with the consequences? To what extent is it socially or morally acceptable (which has important implications for social responses)? The cumulative outcome of this sequential appraisal process determines the specific nature of the resulting emotional episode. It is essential to note that, as shown in Figure 1, the CPM assumes that during this process, the result of each appraisal check can, in many cases, already have efferent effects on the preparation of action tendencies (including physiological and motor-expressive responses), which accounts for the dynamic nature of the unfolding emotion episode that involves all of the different components (see Scherer, 2013b). These ongoing changes are continuously integrated and represented by somatosensory centers in the brain, giving rise to nonverbal feelings or qualia that may, depending on the degree of component synchronization, be temporally segmented, categorized, and labeled with emotion words or expressions. It is important to note that the

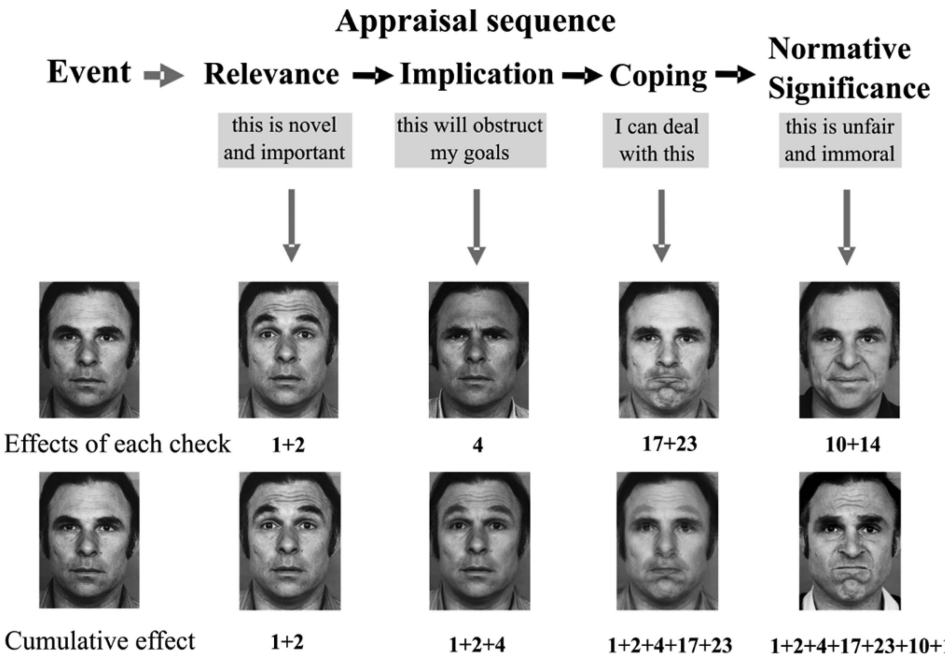

Figure 1. Cumulative sequential appraisal patterning as part of the Component Process Model (Scherer, 2001, 2009). Cumulative effects were generated by additive morphing of the action unit specific photos. Adapted from Figure 19.1 in Scherer et al., 2017.

architecture of the CPM implies that there is a near infinity of different emotion processes and that, in consequence, the final stage of categorizing and labeling is necessarily a highly reductive and impoverished description of the underlying dynamic unfolding of component synchronization. The validity of this assumption is bolstered by empirical findings showing the ubiquity of mixed or blended emotions (e.g., Scherer & Ceschi, 1997; Scherer & Meuleman, 2013).

The central assumption of the CPM that is of particular importance for the current article is that, in most cases, the results of each individual appraisal check sequentially drive the dynamics and configuration of the facial expression of emotion and that the sequence and pattern of movements of the facial musculature allow direct diagnosis of the underlying appraisal process and the resulting nature of the emotion episode (see Scherer, 1992; Scherer & Ellgring, 2007; Scherer, Mortillaro, & Mehu, 2013, for further details). Although this model is largely compatible with several alternative models proposed in the literature (e.g., Frijda & Tcherkassof, 1997), it also makes specific predictions about the effects of the results of certain appraisal checks on the autonomic and somatic nervous systems, indicating exactly which physiological changes and which motor expression features are expected (see Table 1 in Scherer, 2009). These predictions are based on specific motivational and behavioral tendencies expected to enable the adaptive response demanded by the result of a particular stimulus evaluation check. In socially living species, adaptive responses are required not only for internal physical regulation and motor action, but also for interaction and communication with others.

Predictions for appraisal-driven facial expression. Originally, specific predictions for facial expression were elaborated on the basis of several classes of determinants: (a) the effects of typical physiological changes, (b) the preparation of specific instrumental motor actions such as searching for information or approach/avoidance behaviors, and (c) the production of signals to communicate with conspecifics (see Lee, Susskind, & Anderson, 2013; Scherer, 1984, 1992, 2001). As the muscles in the face and vocal tract serve many different functions in particular situations, such predictions can serve only as approximate guidelines. Table 1 provides an illustrative example for facial movements predicted to be triggered in the sequential order of the outcomes of individual appraisal checks. The complete set of CPM predictions (following several revisions; described in Kaiser & Wehrle, 2001; Scherer & Ellgring, 2007; Scherer et al., 2013; and Sergi, Fiorentini, Trznadel, & Scherer, 2016) is provided in Table S1 of the supplementary online material (SOM), together with pertinent empirical evidence (see Appendix A1 in SOM).

Empirical evidence for the CPM predictions on facial patterning. Here we briefly review the evidence for the proposed appraisal patterning mechanism published to date. As the rapidly changing cognitive appraisal processes cannot be assessed directly in an objective fashion and as self-report is limited because individuals are often not conscious of their own appraisals, researchers have used indirect methods to determine the relationship between appraisals and different types of facial expressions. Three major paradigms have been used: (a) having people recall their emotional experiences and act out their expressions, (b) asking encoders (often actors) to produce facial expressions of different emotions, which can then be systematically analyzed for the appraisals that tend to produce the respective emotion, and (c) experimentally manipulating a person's appraisals and measuring the resulting expression.

Table 1
Illustration of CPM Facial Action Unit (AU) Predictions for Fear

Cumulative sequence of appraisal	Appraisal checks	CPM predictions for AUs generated by specific appraisal results	Appraisal results predicted for fear	AUs predicted to be produced by individual appraisal result
1	Novelty			
	Sudden/unpredictable	1, 2, 4, 5, 7, 26, 38	Very high	1, 2, 4, 5, 7, 26, 38
	Familiar/predictable	—	Not applicable	
2	Intrinsic Pleasantness			
	Pleasant	5, 26, 38 or 12, 25	Open	
	Unpleasant	4, 7, 9, 10, 15, 17, 24, 39; or 16, 19, 25, 26	Open	
3	Goal/Need Significance			
	Conduciveness	12, 25	Not applicable	
	Obstructiveness	4, 7, 23, 17	Very high	4, 7, 23, 17
4	Coping Potential			
	High power/control	4, 5 (or 7), 23, 25 (or 23, 24)	Not applicable	
	Low power/control	15, 25, 26, 41, 43 (or 1, 2, 5, 26, 20)	Very high	1, 2, 5, 15, 20, 25, 26, 41
	CPM predictions of AUs that could potentially occur for the emotion of fear as based on the accumulation of the effects of the pertinent appraisals			1, 2, 4, 5, 7, 15, 17, 20, 23, 25, 26, 38, 41, 43

Note. Column 1 and 2: Major appraisal checks postulated by the CPM (except self/norm compatibility) and the respective alternative outcomes; Column 3: The Action units (AUs) predicted as potential expressions for the respective alternative results; Column 4: The degree of pertinence of the specific appraisal outcome (high or very high) for the elicitation of fear ("Open – both outcome alternatives of a check can occur); Column 5: The resulting AUs (from Column 3), expected to occur in the sequence shown in column 1. AU legend: 1 inner brow raiser, 2 outer brow raiser, 4 brow lowerer, 5 upper lid raiser, 7 lid tightener, 15 lip corner depressor, 17 chin raiser, 20 lip stretcher, 23 lip tightener, 25 lips part, 26 jaw drop, 38 nostril dilator, 41 lids droop, 43 eye closure.

Emotion recall. Smith (1989) asked participants to rate their appraisals during recalled experiences of different emotions and subsequently pose the facial expression they would have shown, showing that the eyebrow frown reflects unpleasantness and that smiling indicates pleasantness.

Emotion portrayals. Banse and Scherer (1996) asked professional actors to portray the expressions of 14 major emotions. The combinations of AUs consistently used by a large proportion of the actors can then be interpreted for the appraisal patterns that are expected to generate the respective emotions. Using this corpus, Scherer and Ellgring (2007) showed that there was little evidence for emotion-specific prototypical affect programs. Rather, they concluded that the results were consistent with many of the CPM predictions for dynamic configurations of appraisal-driven adaptive facial actions: AUs 1 and 2 (inner and outer brow raiser) and AU 5 (upper lid raiser) occurring mostly in response to novelty and lack of control; AU 4 (brow lowerer) indicating appraisals of unexpectedness, discrepancy, and goal obstructiveness; AUs 6 (cheek raiser) and 12 (lip corner puller) signaling appraisals of intrinsic pleasantness and goal conduciveness; and AU 9 (nose wrinkler) and 10 (upper lip raiser) indicating unpleasantness appraisals. More recently, realistic enactments of emotional reactions from professional actors using Stanislavski or method acting techniques were obtained (Scherer & Bänziger, 2010) and used to investigate the specific AU configurations employed by the actors to portray 18 different emotions, including different members of the same emotion family such as irritation and anger (Mehu & Scherer, 2015).

For the present article, we extended this approach to review the data on *actor portrayals of emotion* currently available in the literature and to determine the proportion of actors using certain AUs to portray major emotions for which there are empirically supported predictions on the appraisal results that tend to elicit the respective emotion. We identified nine empirical articles that provide relevant data for this issue (Campos, Shiota, Keltner, Gonzaga, & Goetz, 2013; Carroll & Russell, 1997; Du, Tao, & Martinez, 2014; Galati, Scherer, & Ricci-Bitti, 1997; Gosselin, Kirouac, & Doré, 1995; Krumhuber & Scherer, 2011; Mehu & Scherer, 2015; Mortillaro, Mehu, & Scherer, 2011). The results of our secondary analysis are documented in Appendix A in the supplemental online materials (SOM) and summarized in Table S1 (SOM).

Appraisal induction. Kaiser, Wehrle, and Schmidt (1998) used a gaming paradigm to analyze the relation between facial expression, emotion-antecedent appraisal, and subjective feeling, showing that the appraisal patterning approach performed better than an emotion prototype approach in explaining the occurrence of specific AUs (see Kaiser & Wehrle, 2001, p. 294). Smith (1989, p. 342) reviewed early facial electromyography (EMG) research that justifies the plausibility of expecting specific muscle responses to stimuli that are likely to elicit certain appraisals. Empirically, Pope and Smith (1994) found a positive relationship between the pleasantness of an imagined scenario and activity at the zygomaticus major site, whereas activity at the corrugator supercilii site was an indicator of goal obstacles.

More recently, a number of dedicated studies have been published in the emotion research literature in which different strategies were used to experimentally manipulate appraisals through task design and consequently to measure facial EMG responses in different regions of the face (forehead, brow, eyes, cheeks; Aue, Flykt, & Scherer, 2007; Aue & Scherer, 2008; Delplanque et al., 2009; Gentsch, Grandjean, & Scherer, 2015; Lanctôt & Hess, 2007: Van Peer, Grandjean, & Scherer, 2014; Van Peer, Grandjean, & Scherer, 2016). The largely convergent results of these studies suggest that individual appraisals produce specific facial muscle innervations. In particular, as predicted, AU 4 (corrugator)

is activated by appraisal of intrinsic unpleasantness, goal obstructiveness, and high power; AU 12 (zygomaticus) by intrinsic unpleasantness and goal obstructiveness; AUs 17 and/or 20 (mentalis/risorius) by intrinsic unpleasantness and goal obstructiveness; and AUs 1 and 2 (frontalis) by novelty. Four of these seven studies also measured the precise timing of AU onset and confirmed the CPM predictions about the sequential occurrence and cumulative effects of different appraisal outcomes as reflected in facial expression. Gentsch et al. (2015, Figure 11) provide a summary of the chronography of appraisal sequences and highlight the important role of interaction effects between manipulated appraisal checks of sequential accumulation. We conclude that there is now substantial evidence for the production aspect of the claim that the results of individual appraisal checks directly affect the facial musculature, producing specific innervation patterns in the predicted muscle groups.

Extending the Model to Emotion Perception/Recognition

The CPM assumes that one of the functions of expressing appraisal results in the face, body, and voice is the social communication of the event evaluation and the resulting action tendencies. In the course of evolution, the development of mechanisms to reliably communicate reactions to behaviors and events, as well as the consequent action tendencies, has greatly facilitated efficient social interaction, for example, threat signals to avoid fighting (Mehu & Scherer, 2012; Mortillaro, Mehu, & Scherer, 2013). For this mechanism to operate smoothly, observers must be able to accurately decode appraisal results and action tendencies. The CPM assumes that the AUs that carry the information necessary to convey specific appraisals and behavioral intentions on the perception/inference side are the same as those involved on the production side (as discussed in the previous section; (see also Mortillaro, Meuleman, & Scherer, 2012, for an application of this model to affective computing). As an extension to the CPM, the senior author of this article has developed the Tripartite Emotion Expression and Perception model (TEEP; Scherer, 2013a) on the basis of early suggestions by Bühler (1934) and Brunswik (1956; lens model). Figure 2 illustrates a specific adaptation of the TEEP for the facial expressions of appraisals. The left, distal side, of the model reflects the production relationships reviewed in the preceding section for both theory and empirical evidence. The middle part of the model concerns the transmission of distal cues (AUs) in the face via the visual communication channel, producing proximal percepts about the respective facial movements in the observer (here described by normal language glosses for the respective facial movements). The rightmost part of the graph concerns the inference from the proximal percepts to the presumably underlying appraisals or action tendencies. Given that emotion communication by and large works well, we assume that observers can indeed identify the appraisals that tend to produce specific AUs, as described in the previous section.

The aim of the empirical part of this paper is to experimentally test this hypothesis by asking the question: "Which appraisals and action tendencies do naïve observers infer from the presence of specific AUs or AU combinations in the face?" We will first review some preliminary studies in this domain before describing the specific aims pursued in the research program proposed and the methods used in the three studies reported here.

One possibility to estimate the rules used to interpret facial expressions is to explicitly ask raters about the kind of inferences they make about underlying appraisal processes from specific facial features or feature combinations. In a pioneering study, Frijda and Philipszoon (1963) showed that smiles and frowns are associated with pleasantness and unpleasantness, respectively, and that widened eyes and narrowed or closed eyes are associated with high and low levels of attention, respectively. Scherer and Grandjean (2008) asked observers to judge photos of facial emotion

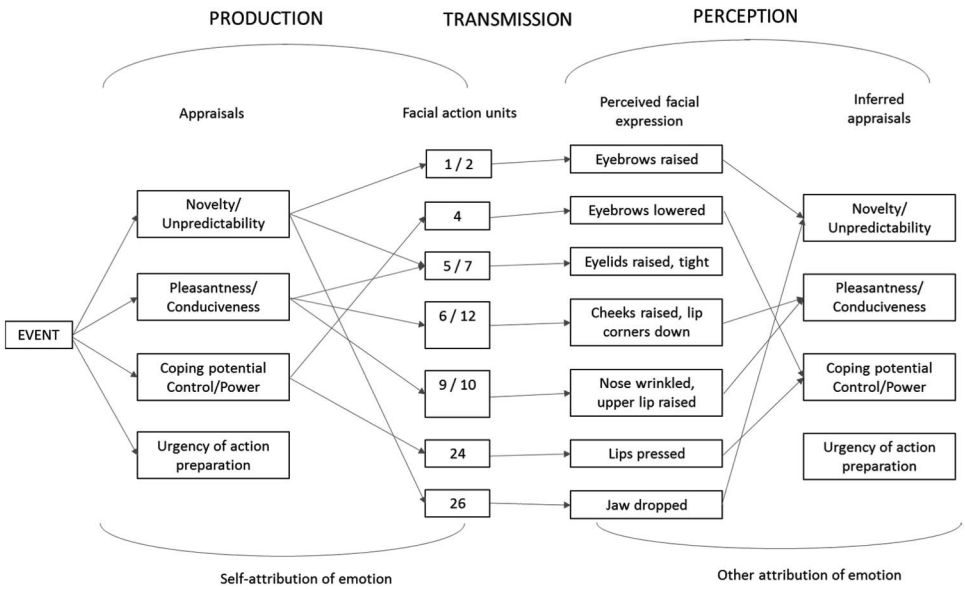

Figure 2. The Tripartite Emotion Expression and Perception model (adapted from Scherer, 2013a).

expressions on different criteria. Emotion categories and appraisals were judged significantly more accurately and confidently than were messages or action tendencies. Shuman, Clark-Polner, Meuleman, Sander, and Scherer (2017) used more ecologically valid, dynamic, and multimodal stimuli and an alternative response measure to examine the inferences from facial expression. The results confirmed that observers can reliably infer multiple types of information (subjective feelings, appraisals, action tendencies, and social messages) from complex emotion expressions. These approaches provide only indirect evidence because of the nature of the stimuli used: photos or videos of prototypical emotion portrayals by actors. In an important review paper, Krumhuber, Kappas, and Manstead (2013) have shown the central importance of studying the *dynamics* of the facial expression of emotion. Thus, future research should use dynamic expression stimuli.

Furthermore, although actor enactments are useful research tools, the testing of specific, theoretically derived hypotheses requires experimental manipulation of the different elements of expression (e.g., action units). This is particularly true for the CPM's claim of sequential unfolding of appraisals from basic to more complex criteria and the assumption that this sequential unfolding will affect the sequence of facial actions generated by the appraisal results. A promising approach consists of systematically manipulating specific facial movements by animation of synthetic faces. Wehrle, Kaiser, Schmidt, and Scherer (2000) used both static and dynamic images of schematically synthesized facial expressions to assess whether judges can correctly recognize emotions that are exclusively based on theoretically predicted configurations of AUs. Recognition rates for the synthetic expressions were far above chance, and the confusion patterns were comparable to those obtained with posed photos. In addition, dynamic presentation increased overall recognition accuracy and reduced confusions between unrelated emotions.

The rapidly increasing sophistication of facial animation in movie productions and consequently in affective computing research has led to the appearance of a number of software packages for facial synthesis. Our group developed FACSGen, a computer program for creating realistic 3D facial expressions in avatar faces from individual FACS AUs (Roesch et al., 2011). FACSGen provides researchers with full control over all major facial AU movements, allowing the user to specify onset, duration, and intensity of each AU independently. Thus, FACSGen allows the creation of videos that show the activation of single AUs or combinations of different AUs (including prototypical emotion configurations, as proposed by Ekman et al., 2002) at different intensities and speeds. Another important feature is the ability to use a photofit procedure to model the avatar face on the photo of a real person.

Krumhuber, Tamarit, Roesch, and Scherer (2012) compared FACSGen to other, mostly template based, facial synthesis software packages and validated the instrument by showing (a) that trained FACS coders correctly coded the synthesized AUs, and (b) that emotional expressions generated with FACSGen convey affective meaning that is reliably recognized by lay participants. The mean recognition rate of 72% was high and comparable to those previously reported with human faces.

This study confirmed the utility of using the FACSGen tool to produce experimentally manipulated synthetic avatar animations of emotion specific AU combinations to study appraisal inferences by naïve observers. In this article, we focus on the inferences observers draw from individual AUs and AU combinations synthesized according to theoretical CPM predictions. Here we report three studies in this line: Study 1 examined (a) the ability of judges to recognize FACSGen-generated prototypical emotion expressions (as proposed by Ekman et al., 2002) and the role of single AUs in emotion inference and (b) the specificity of single AUs in signaling different appraisal dimensions. Study 2 focused on the relative effect of single AUs in frequently considered pairwise combinations and the nature of valence halo effects in appraisal judgment (as found in Sergi et al., 2016), using a tightly controlled design. In Study 3, we investigated (a) the ability of judges to infer precisely defined appraisal dimensions from specific large AU combinations generated by FACSGen based on the updated CPM predictions listed in Table 4 and (b) the attribution of different emotion words to synthetic emotion expressions theoretically composed from these appraisal-specific AU combinations according to CPM predictions (using a forced-choice answer format).

Study 1

Research Questions and Experimental Design

FACSGen technology is ideally used to investigate the type of appraisal and emotional inferences that naïve raters usually make from specific facial AUs and prototypical emotion configurations. In a preliminary study (Sergi et al., 2016), we investigated the appraisal inferences that judges make for systematically manipulated AU combinations in animated avatars. From specific CPM predictions, we specified AU combinations as probable outcomes of appraisal results on eight dimensions. Dynamic facial expressions displaying each combination at three levels of intensity were created with FACSGen. Fifteen judges rated the resulting 126 videos separately on each of the eight dimensions. Contrary to expectation (given the independent ratings of each scale) we found a strong halo effect of valence (as represented by the rating scales of intrinsic unpleasantness/pleasantness and goal conducive/obstructiveness) on all other appraisal ratings (e.g., AU combinations rated as indicating a high coping potential/power appraisal were also judged as very high on pleasantness and goal conduciveness). To control for this valence halo, we partialed out a valence superfactor in further analyses of the remaining dimensions. The results can be summarized as follows: (a) There was excellent agreement between participants' appraisal judgments, (b) appraisal ratings varied systematically between AU combinations, (c) most of the predicted AU-appraisal associations were confirmed by significant results, and (d) only a few AU combinations *uniquely* discriminated the predicted appraisals (suggesting that complex interactions between many AUs are needed to disambiguate meaning).

However, we felt that it would be important to demonstrate the ecological validity of such synthetic stimuli more directly. In part A of Study 1 we wanted (a) to confirm that synthetic versions of prototypical AU configurations suggested by Ekman and collaborators (2002; see also Table S1 in SOM) would be appropriately recognized by human judges and (b) to consequently obtain similar emotion ratings for a large number of single AUs (including some that are rarely studied) to determine the extent to which judges use specific single AUs in emotion inference. We were specifically

interested to see whether some single AUs might even have sufficient signal value for recognition. In Part B, the aim was to study the extent to which judges use individual AUs to infer a number of major appraisal (as Sergi et al., 2016, mostly studied AU combinations).

Method

Participants. Participants were 57 French-speaking students of the University of Geneva (50.8% females, 18 to 33 years, $M = 26.0 \pm 4.5$). The study was approved by the University of Geneva ethics board. Written consent was obtained from participants before the experiment, and they were paid 15 CHF for their participation. Twenty-nine students participated in Part A (emotion ratings) and 28 in Part B (appraisal ratings). The data for one participant in Part B were removed from the analyses because the person exclusively used the extreme points of the rating scale (−100 or + 100).

Stimulus design and production. A total of 128 videos were created by using FACSGen 2.0 Animation software (Krumhuber et al., 2012; Roesch et al., 2011). In this set of 25 single AUs, the combination AU 1 + 2 and six emotion prototype configurations (i.e., anger, disgust, fear, surprise, sadness, happiness) were used (see Table S2 in SOM). The latter were derived from the prototypes defined by Ekman and colleagues (Ekman & Friesen, 1978; Ekman et al., 2002; see also Table S1 in SOM). Two different encoders/avatars were used (one male, one female) and two intensity levels were implemented (50% and 75% of the maximum possible intensity). Illustrative examples of the synthetic expressions are shown in Figure 3.

We generated 128 video clips lasting 2 s in which the dynamic expression was synthesized at a frame rate of 25 images per second. Single AUs and AU combinations unfolded linearly starting with a neutral face and reached the apex after 660 ms (onset duration), the apex lasted 670 ms (apex duration), and then the expression returned to neutral in 670 ms (offset duration). Durations were chosen after an informal investigation of the genuineness of the facial expression unfolding. All videos were synthesized at a frame rate of 25 images per second and were rendered in color, with the same viewpoint, camera focal length, and central lightning. Stimuli measured 800 × 1,200 pixels and were displayed on a black background.

Rating scales. Participants in Part A performed "emotion ratings." They were asked to rate each video stimulus on each of six emotion scales (anger, disgust, fear, surprise, sadness, happiness) for the extent to which they thought the encoder might have felt the respective emotion. In Part B, "appraisal ratings," participants were asked to imagine for each video the event that might have caused the facial expression shown and were asked to rate the extent to which the encoder might have performed each of the following six appraisal checks for the event that might justify the facial expression: Unexpectedness, Pleasantness, Unpleasantness, Power, Agency, and Norm Compatibility (i.e., the extent to which the event was unexpected, pleasant, unpleasant, controllable by the person, caused by another person, and conforming to social norms; see Table S3 in SOM). We used separate scales for the two poles of the valence-related appraisal dimension to test the symmetry of inference. Agency and Norm Compatibility were included to examine whether there were any AUs that would be systematically used to infer these appraisals (even though, from the point of production, there is little likelihood of facial responses to such appraisal results and no predictions have been made in the CPM).

Figure 3. Study 1: Examples of action unit (AU) combinations. Left to right: Happiness, Anger, AU 1 + 2 (75% intensity), AU 5 (75% intensity). See the online article for the color version of this figure.

Procedure. Participants arrived individually at the laboratory and were seated in front of a 17" color video monitor in one of four computer workstations that were visually isolated from each other. After signing the consent form and filling out a short questionnaire about demographic data, participants were told that they were going to see videos of facial expressions and answer questions. The experimenter explained that the videos were obtained by recording people in realistic situations and were subsequently converted to synthesized stimuli (i.e., avatars) to avoid confounding variables such as identity, gender, and so forth. The experiment was run on E-Prime 2.0 software (Psychology Software Tools, Inc., 2012). Written detailed instructions, different for each experimental condition, were presented on the screen. Subsequently, participants started the experiment with two training items to familiarize themselves with the task.

At the beginning of each trial, a fixation cross was presented for 500 ms at the center of the screen followed by the video stimulus. Participants were allowed to watch the video as many times as they wanted. For each question, participants were asked to move a slider on a continuous scale between *not at all* (−100) and *very much* (+100). At the end of the experiment, they were debriefed and paid for their participation. The experiment lasted approximately 60 min.

Results

Part A: Emotion ratings.

Rater consistency. We calculated Cronbach's alpha for each emotion scale. Total reliability was excellent for all scales (Anger, 0.96; Happiness, 0.97; Disgust, 0.91; Sadness, 0.93; Fear, 0.90; Surprise, 0.95).

Participant gender effects. For each emotion scale, a repeated-measures analysis of variance (ANOVA) was performed with Intensity, Encoder/Avatar gender, and AUs as within-subject factors, and Participant Gender (Male, Female) as the between-subjects factor. The main effect of Participant Gender was not significant for any of the six scales. For the Sadness scale, there were two significant interactions: AU × Gender, $F(31, 837) = 1.810$, $p = .005$, $\eta^2 = 0.033$; Intensity × AU × Gender, $F(31, 837) = 1.511$, $p = .037$, $\eta^2 = 0.011$. For the Fear scale, there was a significant interaction of Intensity × AU × Gender, $F(31,$

837) = 1.517, $p = .036$, $\eta^2 = 0.014$. Given the minor effects of Participant Gender, we decided to drop this factor from further analyses.

Overall ANOVA. We then performed a repeated-measures ANOVA with Emotion scale (six levels), Intensity (two levels), Encoder/Avatar gender (two levels), and AUs (32 levels) as within-subject factors. The results (using the Greenhouse-Geisser criterion) showed significant main effects for Emotion, $F(4, 112.3) = 13.014$, $p < .001$, $\eta^2 = 0.317$; Intensity, $F(1, 28) = 7.01$, $p = .013$, $\eta^2 = 0.2$; and AUs, $F(31, 10.1) = 11.753$, $p < .001$, $\eta^2 = 0.296$; they also showed significant interaction effects for Emotion × Intensity, $F(4.5, 125.2) = 4.829$, $p = .001$, $\eta^2 = 0.147$; Emotion × Avatar Gender, $F(3.9, 108) = 3.549$, $p = .01$, $\eta^2 = 0.112$; Emotion × AUs, $F(11.7, 328.5) = 40.182$, $p < .001$, $\eta^2 = 0.589$; and Emotion × Intensity × AUs, $F(18.5, 519.1) = 1.846$, $p = .017$, $\eta^2 = 0.062$. The main effects are of little interest in the current context, as are the interaction effects not involving AUs. Of central interest is the Emotion × AUs interaction, which shows a massive effect size. The interaction Emotion × Intensity × AUs, because of higher intensity tending to increase ratings for particular AUs (for similar results, see Wingenbach, Ashwin, & Brosnan, 2016), also reaches significance, but has a very small effect size. In consequence, we focus on the differential patterns of AU ratings for the six emotions. The profiles of means for the individual AU ratings for the six emotions, with the ratings ranked by descending size, are reported in Table S4 in SOM. The significance of the differences between the respective numbers is difficult to establish, as the large number of levels on the AU factor does not allow the application of the Bonferroni post hoc procedure, which is usually advised for repeated-measures ANOVAs. In consequence, we used a procedure that is regularly applied in statistical analyses of electroencephalogram and functional MRI data in which many different measurement points are used. The procedure consists of applying a k-means cluster analysis that aims to partition *n* observations into a given number of k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. Our purpose was to separate the AU profiles for each emotion into three subgroups that can be ranked, on the basis of the mean ratings (see Table S4) into three categories—top, middle, and low. The objective is to identify those AUs the presence of which has the strongest impact on the inference of a specific emotion (*top*, receiving high positive ratings), those that are unrelated to the emotions (*middle*, receiving ratings in the center of he scale) or those which rarely, if ever, give rise to inferring the emotion (*low*, strong negative ratings). Based on the results of the k-cluster analysis (See Table S5A in SOM), we measured the significance of the differences between the means of the three groups by one-way ANOVAs (for reasons of economy of space, and given the very high level of significance of the overall AU factor reported above, the detailed results are not shown here).

This procedure provided a principled manner to determine, for each emotion, the group of AUs that significantly contributed to the inference of the specific emotion. It is important to underline that three groups were chosen to increase the level of discrimination and be sure to maximize the chances of finding the AUs that carry most information even when presented in isolation. We included the specific AU configurations, intended to represent specific emotions, in the ranked lists (see Table S2 in SOM)—but not in the k-means cluster analysis—to allow a direct comparison between the relative effects of individual AUs and the prototypical AU combinations.

In what follows, we only discuss the top groups as these results are most pertinent to our research question. The AUs in those groups are shown below, with their respective verbal descriptors and the mean ratings in parentheses:

Anger: AU4 (Brow lowering; 47.8), AU9 (26.4), AU24 (Lip pressor; 8.9), AU10 (Upper lip raiser; 5.4).

Happiness: AU12 (Lip corner puller; 38.4), AU13 (Cheek puffer; 50.8), bilateral AU 14 (Dimpler; 30.0), unilateral AU14 (8.8).

Disgust: AU9 (Nose wrinkler; 32.0), AU10 (Upper lip raiser; 32.4), AU11 (Nasolabial furrow deepener; 7.9).

Sadness: AU15 (Lip corner depressor; 28.7), AU1 (Inner brow raiser; −0.8), AU17 (Chin raiser; −1.9).

Surprise: two clusters, (a) AU27 (Mouth stretch; 67.1), AU5 (Upper lid raiser; 37.8), AU26 (Jaw drop; 23.2), the combination AU1 + 2 (inner and outer Brow raiser; 31.4), and (b) individual AU2 (28.7) individual AU1 (−2.5).

Fear is clearly a special case: Only the prototype AU combination reliably produces fear judgments; even though the cluster analysis separates AU27 and AU5 from other AUS, by looking at the mean profiles none of the individual AUs even come close. This suggests that the recognition of fear requires a complete configuration of the appropriate facial cues, reflecting the sequence of underlying appraisals—something very negative happening unexpectedly and the person lacking sufficient power to control the consequences. For some of the other emotions, with a simpler underlying appraisal structure (e.g., unpleasantness for disgust, pleasantness/conduciveness for joy, novelty/unexpectedness for surprise), individual AUs directly linked to the respective appraisal check may be sufficient for recognition. Another special case is surprise, for which there are two clusters of AUs that obtain high mean ratings. We expected that all these AUs would be part of the same top clusters, but probably the strong role of the combination AU1 + 2 and its components in signaling surprise yielded more the complex result in the cluster analysis.

Table 2 lists a cross-tabulation of the intended and inferred emotions for these special emotion prototype configurations. It shows that participants recognized the intended emotion and gave high ratings to the respective category, except in the case of disgust (the ratings for disgust were generally higher for the intended anger prototype). However, as shown in the disgust column of Table S4 in SOM, the single AUs 9 and 10 by themselves produce

Table 2
Cross-Tabulation of the Intended Emotion Displays and the Inferential Ratings by Participants

	Inferred					
Intended	Anger	Happiness	Disgust	Sad	Fear	Surprise
Anger	**65.2**	−67.9	*24.8*	−70.6	−57.4	−70.6
Happiness	−87.6	**35.4**	−79.1	−81.2	−75.5	−52.0
Disgust	−32.4	−65.5	**−5.3**	−37.2	−52.6	−67.6
Sad	−58.8	−67.8	−64.6	**52.9**	−43.4	−72.6
Fear	−61.9	−71.6	−45.6	−7.1	**52.0**	−26.2
Surprise	−42.6	−46.7	−29.5	−42.4	*28.5*	**79.5**

Note. Bold = intended matches; italics = expected confusions.

very high disgust ratings. The fear label was also applied to some extent to the surprise prototype, probably because of the presence of the AU 1 + 2 combination in both fear and surprise (see Table 1). This confirms our prediction that the FACSGen-generated prototype configurations would be clearly recognized, validating the general approach of synthetically producing different facial expressions.

The next aim of Study 1 was to examine whether there are AUs, constituents of prototypical emotion configurations, that play a determinant role in the signal characteristics of the prototype or that are even sufficient on their own to generate the appropriate inference. For Anger ($M = 65.2$), Sadness ($M = 52.9$), Fear ($M = 52.0$), and Surprise ($M = 79.5$), the prototypes obtain the highest ratings, but for Anger AU 4 and for Surprise AU 27 are close runner-ups, suggesting that these AUs may indeed be almost sufficient to signal the respective emotion by themselves. In the case of Happiness ($M = 35.4$), the single AU 12 and particularly AU 13 perform better than the prototype, suggesting that they carry most of the signal value. For disgust ($M = -5.3$), although the prototype does not seem to work, AUs 9 and 10 each carry substantial signal values and it is possible that the combination of the two could perform even better. Interestingly, AU9 has also high signal value for Anger and it is not exclusive of Disgust; this may explain confusions between these two emotions in perception studies.

A third aim of Study 1 was to explore whether there are single AUs, not currently represented in the prototypes, that elicit powerful specific inferences. Judging from the results shown in Table S4 in SOM, this does not seem to be case, except possibly for (bilateral) AU 14, rarely mentioned in the literature, which may provide a pointer for the presence of happiness, and partly AU9 for Anger. We may conclude that the prototypical emotion AU configurations in FACSGen-generated dynamic expressions are indeed well recognized and thus validate the use of this approach in the systematic study of emotion inference and recognition. The results have yielded important clues to the relative signal value of individual AUs, encouraging further work on the relative role of single AUs in facial communication.

Part B: Appraisal ratings.

Rater consistency. We calculated Cronbach's alpha for each appraisal scale: Unpredictability, 0.85; Pleasantness, 0.96; Unpleasantness, 0.94; Power, 0.78; Agency, 0.79; and Normative Significance, 0.90. Although agreement is excellent for the valence-related appraisals, there is somewhat less agreement on Power and Agency, possibly because these criteria are less frequently applied explicitly in everyday life.

Participant gender effects. For each appraisal scale, a repeated-measures ANOVA was performed with Intensity, Encoder/Avatar gender, and AUs as within-subject factors, and Participant Gender (Male, Female) as the between-subjects factor. The main effect of Participant Gender was never significant nor were any of the interactions involving that factor. Consequently, we dropped the Participant Gender factor from further analyses.

Overall ANOVA. We then performed a repeated-measures ANOVA with Appraisal scale (six levels), Intensity (two levels), Encoder/Avatar gender (two levels), and single AUs (26 levels) as within-subject factors. The results (using the Greenhouse-Geisser criterion) showed significant main effects for Appraisal, $F(2.9, 78.8) = 13.590$, $p < .001$, $\eta^2 = 0.335$; Intensity, $F(1, 27) = 7.845$, $p = .009$, $\eta^2 = 0.225$; and AUs, $F(7.1, 191.5) = 15.987$, $p < .001$, $\eta^2 = 0.372$; they also showed significant interaction effects for Appraisal × Intensity, $F(2.8, 75.5) = 5.977$, $p = .001$, $\eta^2 = 0.181$, and Appraisal × AUs, $F(11.7, 317.2) = 18.382$, $p < .001$, $\eta^2 = 0.405$. Again, the main effects are of little interest in the current context, as are the interaction effects not involving AUs. Therefore, we focus on the Appraisal × AUs interaction, which shows an important effect size. The profiles of mean AU ratings for the six appraisal scales, with the ratings ranked by descending size are reported in Table S6 in SOM. Notably, the above ANOVA was computed for the set of 26 single AUs, as this was the central aim of Part B. However, as the participants in Part B also rated the emotion prototypes analyzed in Part A, Table S6 in SOM shows (in a separate section at the bottom of the table) the means of these prototypes on the six-point-appraisal scale for the sake of comparison. As for the emotion ratings, we applied a k-means cluster analysis to separate the profile into three coherent subgroups—top, middle, and low—and to consequently measure the significance of the differences between the three groups by a one-way ANOVA. Note that here the top group is interpreted as representing a high level on the respective dimension and the bottom group as a low level.

The AUs that were included in the top cluster resulting from the k-cluster analyses for the different appraisals are shown below, with their respective verbal descriptors and the mean ratings:

Unexpectedness: AUs 27 (Mouth stretch; 40.3), AU5 (Upper lid raiser; 28.5).

Pleasantness: AU13 (Cheek puffer; 47.9), AU12 (Lip corner puller; 43.7), AU14 (30.4), and AU18 (Lip puckerer; 27.0).

Unpleasantness:AU9 (Nose wrinkler; 48.5), AU4 (Brow lowerer; 37.3), AU10 (Upper lip raiser; 34.9), AU15 (Lip corner depressor; 28.3).

Power: AU14 (Dimpler; 27.0), AU13 (Cheek puffer; 25.1), AU12 (Lip corner puller; 24.5).

The results concerning these AUs are entirely in line with the predictions of the CPM for the respective appraisal check. Based on the detailed results shown in Table S6 in SOM, we note that for Unexpectedness there are three other predicted AUs that do not make it into the top cluster but reach relatively high ratings: AU26 (Jaw drop; 15.0), the combination AU1 + 2 (inner and outer Brow raiser; 14.6), and individual AU2 (13.3). The AU patterns for high Power basically resemble those of pleasantness, without any distinctive additional AUs that might be specific for high power. On the low Power end, however, we find moderately high ratings for AU27 (Mouth stretch; −15.9) and AU11 (Nasolabial deepener; −11.2). Participants were unable to meaningfully infer Agency or Normative Significance from the single AUs. The distribution of means for Agency is completely atypical and for Normative Significance we find a weak copy of the Pleasantness results. This was expected, as no predictions for these appraisal dimensions are made in CPM, assuming that the cognitive operations involved in these checks do not lead to functional motor movements used for information search or action preparation. We included these dimensions here to empirically confirm this hypothesis indirectly by the fact that there are apparently no useful signals to observers.

The mean ratings for the prototypical emotion AU configurations correspond to expectations: There are high values for Unex-

pectedness on the surprise (M = 55.9) and the fear (M = 46.3) prototypes, and high ratings on happiness (M = 38.9) and low ratings for all negative emotions (Anger, M = −62.4; Fear, M = −52.1; Sadness, M = −50.5; Disgust, M = −45.4) for Pleasantness (the reverse being the case for the Unpleasantness scale). Interestingly, confirming theoretical hypothesis, Surprise was rated nor pleasant (M = −3.0) nor unpleasant (M = −7.9). Low power ratings are found for surprise (M = −31.8) and fear (M = −25.3), as expected, but contrary to expectation, power is not rated highly for the anger prototype (M = 3.1) - rather, there are high power ratings for happiness (M = 24.1), confirming that high power seems to be signaled by valence-related AUs.

We have not systematically compared the patterns of these results with the predictions shown in Table S1 for two reasons: one is that the predictions have been made for the production domain and it is likely that not all of the respective AUs serve as signals in the perception domain; the other is that, as the predictions were intended for AU combinations. In fact, although the results suggest that some AUs (such as 12, 13, 5, 9, 10, 27) may have signal value of their own, it remains an empirical question as to what extent inferences are based on individual AUs or AU combinations. This is the purpose of Study 2.

Study 2

Research Questions and Experimental Design

Whereas Study 1 focused on the role of single AUs and prototypical emotion configurations, in Study 2 we were specifically concerned with pairwise AU combinations that are frequently highlighted in the literature as having a similar function in expression and recognition (e.g., the AU 6 + 12 pair as a reliable indicator of happiness). Furthermore, whereas in Study 1, we asked only for general Pleasantness and Unpleasantness, in Study 2, we wanted to examine the theoretical distinction between different types of valence, in particular intrinsic pleasantness (e.g., sensory pleasure) and goal-related valence, that is, goal conduciveness versus obstructiveness (see Scherer, 2013b, pp. 153–156).

In contrast to the study by Sergi et al. (2016), briefly described in the introduction to Study 1, where the large number of individual combinations made it difficult to determine the specific contribution of specific AUs, we decided to reduce the number of AUs to those AU pairs that are essential for emotion expression (see Scherer et al., 2013, Table 1). As a criterion, we used the presence of a pair of AUs in the CPM predictions and in the list of prototypical elements specified by Ekman and collaborators in the EMFACS coding scheme for basic emotions (Ekman et al., 2002, see Table S1). In addition, we considered the pattern of empirical findings that resulted from our overview of the expression portrayal literature (see the description of production earlier) and which were highlighted in Study 1 (e.g., AUs 6 + 12, AUs 9 + 10). Of particular interest was the exact role of AU 4 in combinations. This resulted in the choice of the following pairs of AUs (appraisal prediction according to Table S1 in parentheses): 1 + 4 (low power), 9 + 10 (intrinsic unpleasantness), 17 + 24 (obstructiveness), 6 + 12 (conduciveness), and 4 + 7 (novelty/unpredictability). We decided to systematically determine the respective role of each AU component of a combination in a 2 × 2 factorial design: presence versus absence of each component in a pair. Although the main effects were expected to show the independent contribution of each AU, the interaction effects should show the additional effect of combining both AUs, allowing us to gain a better understanding of the underlying inference mechanism. In addition, in this study, we wanted to test the importance of the identity and gender of the avatars. Furthermore, we improved the ecological validity of the avatar faces by using more avatar identities and by making them more lifelike and realistic by adding hair and a background setting (for illustrative examples, see Figure 4). Finally, rather than using a small sample of student participants, we decided to use a large, more representative adult survey sample so that we could examine the effects of age and gender. In an attempt to control for the valence superfactor encountered in the study by Sergi et al. (2016), as well as in Study 1, we formed subgroups of participants who were asked to rate the videos on only one of four major appraisal checks: Novelty/Familiarity, Intrinsic Unpleasantness/Pleasantness, Goal conduciveness/obstructiveness, and High versus Low Power/Coping ability. The assumption was that if raters concentrated on one specific appraisal dimension, there would be less danger that valence connotations of the different appraisal dimension labels would affect the judgments on other dimensions (e.g., assuming that goal conduciveness was also pleasant and goal obstructiveness unpleasant).

We further hypothesized that participants who have high emotional competence (EC), being able to accurately interpret the emotions of their interaction partners, should score higher in recognizing appraisals from synthetic faces. In consequence, we also administered a test of emotion recognition ability, predicting that participants with higher scores on this test would also be more likely to correctly identify the appraisal patterns expressed by single AUs and pairwise AU combinations.

Method

Participants. We recruited 156 participants (50.6% females, 18 to 65 years, M = 45.4 ± 12.2) via the Qualtrics survey panel

Figure 4. Study 2: Examples of action unit (AU) combinations used as stimuli, showing the effect of hair and background (stills from the original sequences). Left: AUs 4 + 7 (Unpredictable); Right: AUs 17 + 24 (Goal obstructive). See the online article for the color version of this figure.

(https://www.qualtrics.com/online-sample/) in exchange for monetary compensation. The selection criteria were as follows: native English speakers, Caucasian origin, and over 18 years old.

Stimulus design and production.

Task 1: Appraisal ratings. As in the study by Sergi et al. (2016) and Study 1, we used the FACSGen animation tool (Krumhuber et al., 2012; Roesch et al., 2011) for creating 56 short videos with four different avatar faces that portrayed eight individual FACS AUs (1, 4, 6, 7, 9, 10, 12, 17, 24) and five AU combinations (1 + 4, 4 + 7, 6 + 12, 9 + 10, 17 + 24), in addition to a neutral (no expression) video for each avatar face. The computer-generated expressions showed the full unfolding of single AUs or combinations from the onset over the apex to the offset. The duration of the expression and the unfolding patterns were based on data obtained from the core set of the GEMEP corpus of actor emotion portrayals (Bänziger, Mortillaro, & Scherer, 2012) for which FACS coding by trained observers (including onset, apex, and offset timing) had been obtained (Mehu, Mortillaro, Bänziger, & Scherer, 2012). As the preceding study by Sergi et al. (2016) had shown that the medium intensity of AU expression (60% of maximum) provided adequate and realistic stimuli and we had not found any interaction effects for AUs and intensities of 50% and 75% in Study 1, the 60% intensity was also used to produce the stimuli for the current study. The advantage is that the stimuli appear more natural; note, however, that stronger effects are very likely to be found with more extreme intensities. In other words, we expected our results to be rather on the conservative side.

Task 2: Emotion recognition test. We used the short version of the Geneva Emotion Recognition Test (GERT-S, Schlegel & Scherer, 2016; see also Schlegel & Scherer, 2016), which contains 42 video clips, lasting 1 to 3 s, which were taken from the GEMEP core set. In these video clips, 10 French-Swiss actors (five males, five females) portray 14 different emotions by using facial expressions, gestures and two different meaningless speech-like utterances.

Procedure.

Task 1: Appraisal ratings. Participants were told in the instructions that the purpose of the study was to examine the claim that the face is a "mirror of the soul" and that one can read someone's thoughts from facial expressions. It was explained that the facial expressions of real persons experiencing certain emotions had been videotaped, that these persons had been asked to indicate the type of thought that was most in their mind during the event, and that the videotaped expressions had been transposed to avatar faces via synthesis to protect the identity of the individuals. Participants were informed that they were going to view 60 short videos of avatar expressions and would be asked to judge the type of thought that the individuals represented by the avatars had about the situation they were facing at the time. It was explained that to make this difficult task a bit easier, they were to focus on only one of the different types of thought reported earlier. In four different subgroups, participants were then given one of the following four dimensions to use in their judgments: Novelty (event is sudden, unexpected; or familiar, expected); Pleasantness (situation is unpleasant, disagreeable; or pleasant, agreeable); Goal conduciveness (event is unfavorable, obstructive; or favorable, advantageous); and Power/Coping potential (one is powerless, lacking the resources to cope; or powerful, able to master the consequences).

Ratings of the likelihood that the respective facial expression signaled the given type of thought (e.g., unpleasant, disagreeable vs. pleasant, agreeable) were to be made by sliding a button on a continuous bipolar scale representing the two extremes (from *−5 to + 5,* the 0 point indicating that the respective rating dimension was not at all applicable to a given expression; see Figure S1 in SOM). After the instructions, participants were presented with two practice examples to familiarize themselves with the task. They had the option of reading the instructions again, and when they were ready, they could start the real task. The panel participants were randomly assigned to one of the four between conditions, that is, type of thought. The order of presentation of the videos within each condition was also randomized. Each video could be played only once with no replay option.

Task 2: Emotion recognition test. Participants were instructed that they were going to see a series of short videos in which different actors express various emotions and that they had to guess which emotion was portrayed in the videos. They were also told that the actors' utterances consisted of a series of meaningless syllables, but that they could still recognize the emotional tone of the utterance. After the instructions, participants performed three practice examples to familiarize themselves with the task, and then they could start the main task. The 42 videos were all presented in random order within one single block. They played automatically as soon as the participants gave their response for the previous video, so that each video could be viewed only once. After each video, participants had to choose which emotion was portrayed by the actor by selecting one of 14 emotions. The 14 emotions were presented in a circle, and participants had to click on the button next to the correct emotion word (see Figure S2 in SOM).

These procedures were approved by both the ethical committee of the Department of Psychology at the University of Geneva and the Ethics Department of the European Research Council.

Results

Rater agreement. The data were cleaned before further analysis by removing, separately for each group, participants who had not provided any responses for a large majority of stimuli or who had responded throughout with extreme judgments (using only −5 or 5), resulting in the following *N*: Novelty, 33 (four removed); Pleasantness, 39 (none removed); Goal conduciveness, 38 (one removed); Power, 37 (three removed). Rater agreement for the remaining 145 raters was computed as Cronbach's alpha, yielding the following coefficients: Novelty, .90; Pleasantness, .98; Goal conduciveness, .97; Power, .61. Thus, as in Study 1, there was excellent agreement on the ratings for Novelty, Pleasantness, and Goal conduciveness, but somewhat less agreement on Power, suggesting that the raters found it more difficult to understand this appraisal dimension or to infer it from the facial expression.

ANOVAs. Separately for each of the four groups of raters (each using only one of the four appraisal checks) and separately for each of the five types of AU combinations (1 + 2, 4 + 7, 6 + 12, 9 + 10, 17 + 24), the ratings were analyzed by means of a repeated-measures ANOVA of the 2 × 2 × 4 design with the repeated factors AU A (present, absent), AU B (present, absent),

and Avatar identity (two females and two males). Significant main effects for single AUs can be interpreted as independent contributions of the respective AUs, whereas significant interaction effects reveal the additional effect of combining both AUs. Significant main effects for avatar identity would show the effect of the architecture of the face and, more important, significant interaction effects between avatar identity and the two AU factors would have implications for the generalizability of the AU results. The pertinent results of all 20 separate analyses are shown in Table 3. As there were no significant interaction effects of the avatar identity factor with the two AU factors, only the F, p, and η^2 values for the AU factors and their interaction are shown. Although there were a few main effects for avatar identity (in particular one female and one male face being generally rated as less powerful), these differences are not pertinent to the current discussion, as there were no interaction effects.

We discuss the ANOVA results for the expected patterns separately for the four appraisal dimensions:

The expected effects for Intrinsic Pleasantness and Goal conduciveness/obstructiveness are largely confirmed. However, the theoretically justified distinction between intrinsic pleasantness (e.g., sensory pleasure) and goal-related valence is not readily apparent in the appraisal inferences from facial expression. Both AUs 9 and 10 (linked to the expression of sensory disgust) by themselves produce strong increases in ratings of negative valence for both appraisals, just as in Study 1. However, the effect size for AU 9 is somewhat higher for intrinsic unpleasantness (as predicted on the basis of the evolutionary origin of the nose wrinkle). The question raised in the discussion of this result in Study 1—Would a combination of 9 and 10 increase the signal value?—is answered in the negative by the current results. Rather, the effect is weakened. AUs 17 and 24 do not function well as a pair for negative valence. However, AU 17 on its own is seen as a signal of goal obstructiveness, as predicted (and, to a lesser extent, for unpleasantness). AU 12 is a powerful predictor of positive valence for both valence appraisals. However, the effect size for AU 12 is somewhat higher for conduciveness than for intrinsic pleasantness, again consistent with Study 1 where AU 13 was rated somewhat higher as an indicator for pleasantness than was AU 12. Contrary to expectations based on the classic literature, the addition of AU 6 does not strengthen this effect; rather, it weakens it, mirroring the pattern of findings in Study 1.

As often found in the literature, AU 4 is also a strong predictor of negative valence inferences, suggesting that something unpredicted, possibly implying negative consequences, occurred. This interpretation is supported by the fact that, as predicted, AU 4 leads to an increase in novelty inferences.

As one might have expected on the basis of the low reliability of the Power ratings, there are no strong effects for any of the AUs or AU combinations with the exception of AU 12: Smiling, quite justifiably, is perceived as a sign of being in control of the situation (just as in Study 1).

To analyze individual differences between judges, we computed difference scores between the neutral (no expression) ratings for each avatar and the ratings of the respective AUs and AU combinations, assuming that larger difference scores indicate greater sensitivity for the facial changes. The mean difference scores are shown in Table S7 in SOM. A two-way ANOVA did not show any significant age or gender main effects or interactions.

To examine the hypothesis that participants with higher emotion recognition competence (as measured by the GERT-S) would be better able to infer the appraisal results signaled by different facial AUs, we correlated these difference scores with the GERT scores of the participants. Except for the Power appraisal dimension, there were a large number of significant correlations, providing strong support for the hypothesis. In particular, the most effective AUs (4, 9, and 12) were rated much higher on the predicted appraisal dimensions by high GERT-S scorers (with positive correlations ranging from $r = .40$ to .50, $p < .001$; see Table S7 in SOM).

Table 3
Results of 20 Separate Repeated-Measures ANOVAs for AU Group and Appraisal Dimension

	Appraisals															
	Novelty (Familiarity)				Pleasantness				Goal conduciveness				High power/Coping potential			
AUs	F	p	η^2	Direction	F	p	η^2	Direction	F	p	η^2	Direction	F	p	η^2	Direction
1	1.777	.192	.053		3.845	.057	.092		5.133	.029	.122		.003	.957	0	
4	**19.36**	**<.001**	**.377**	pos	**101.263**	**<.001**	**.727**	neg	**115.632**	**<.001**	**.758**	neg	1.417	.242	.038	
1 and 4	.88	.355	.027		.046	.831	.001		1.404	.244	.037		.471	.497	.013	
9	**17.438**	**<.001**	**.353**	pos	**147.542**	**<.001**	**.795**	neg	**69.027**	**<.001**	**.651**	neg	.158	.693	.004	
10	**15.491**	**<.001**	**.326**	pos	**43.685**	**<.001**	**.535**	neg	**49.557**	**<.001**	**.573**	neg	.38	.541	.01	
9 and 10	2.113	.156	.062		**12.501**	**.001**	**.248**	neg	**18.099**	**<.001**	**.328**	neg	.004	.948	0	
17	.016	.901	0		**9.1**	**.005**	**.193**	neg	**13.964**	**.001**	**.274**	neg	2.709	.108	.07	
24	4.493	.042	.123		4.27	.046	.101		0	.998	0		.707	.406	.019	
17 and 24	4.385	.044	.121		.558	.460	.014		.012	.915	0		.284	.598	.008	
4	**11.441**	**.002**	**.263**	pos	**95.579**	**<.001**	**.716**	neg	**79.121**	**<.001**	**.681**	neg	2.1	.156	.055	
7	2.646	.114	.076		4.462	.041	.105		1.472	.233	.038		.065	.800	.002	
4 and 7	1.165	.289	.035		.574	.454	.015		**12.614**	**.001**	**.254**	neg	.017	.898	0	
6	1.247	.272	.038		.683	.414	.018		2.873	.098	.072		.716	.403	.02	
12	**48.962**	**<.001**	**.605**	neg	**83.238**	**<.001**	**.687**	pos	**122.108**	**<.001**	**.767**	pos	**9.664**	**.004**	**.212**	pos
6 and 12	.653	.425	.02		1.3	.261	.033		2.949	.094	.074		1.625	.211	.043	

Note. Bold text indicates significant p after Benjamini-Hochberg correction, FDR = .05. Direction = positive (pos) or negative (neg) direction of effect. ANOVA = analyses of variance; AU = action unit; FDR = false discovery rate.

Table 4
Revised Predictions of Specific AU Combinations for Different Appraisal Outcomes

Order	Appraisal checks	Outcomes (common language gloss)	Action tendencies	Predicted AU combinations
1	Novelty	Sudden (event occurred suddenly, is new to the person)	Orientation (widen visual field)	AUs 1 + 2 + 5 + 26
		Unpredictable (event requires close attention)	Scrutiny (visual focusing)	AUs 4 + 7
2	Intrinsic pleasantness	Pleasant (pleasant sensation: taste, smell, sight)	Approach (capture and savor)	AUs 6 + 13 + 14 + 23 + 43 + 53
		Unpleasant (unpleasant sensation: taste, smell, sight)	Avoidance (blocking stimulation)	AUs 9 + 10
3	Goal conduciveness	Conducive (just what the person wanted)	Contentment (relaxation)	AUs 5 + 6 + 12 + 25 + 27
		Obstructive (not at all what the person wanted)	Disappointment (tension)	AUs 4 + 11 + 14 + 17 + 24 or AUs 4 + 11 + 15 + 20 + 23
4	Power/Coping potential	Low (feels that nothing much can be done, resignation)	Resignation (accommodation)	AUs 1 + 4 + 16 + 25 + 26 + 43 + 54
		High (feels ready to confront and tackle any obstacles or enemies)	Confrontation (threat, attack)	AUs 7 + 17 + 53 or AUs 4 + 10 + 22 + 27

Note. Column 1: Predicted sequential order of appearance of the individual appraisal checks. Column 2: Major appraisal checks. Column 3: Direction of the outcomes of the checks (everyday language gloss in parentheses, as presented to raters). Column 4: Action tendencies likely to be elicited (accounting for the functional nature of the facial response). Column 5: Predicted action unit (AU) combinations representing the facial actions appropriate to the behavior preparation.

Discussion

In line with predictions, clear evidence from main effects shows that AUs 4 (Brow lowerer), 9 (Nose wrinkler), 10 (Upper lip raiser), 12 (Lip corner puller), and 17 (Chin raiser) strongly determine the ratings of novelty intrinsic unpleasantness/pleasantness and goal-conduciveness/obstructiveness. However, in each case, the single AUs carry most of the information used by participants; adding 1 (Inner brow raiser) or 7 (Lid tightener) to 4, or 6 (cheek raiser) to 12, has almost no effect or even weakens the effect on appraisal inferences. Although the specific appraisal inferences predicted for the AU combinations studied are confirmed, the differentiation between appraisal dimensions seems limited and almost absent for the power dimension. Despite asking participants to rate only one single appraisal dimension, we found similar halos to those reported in Sergi et al. (2016) and found in Study 1. We explain this by the fact that people have inherent implication or association rules in their inference structures, probably often based on ecological correlations. For example, to have high power is pleasant and helpful for reaching one's goals. Thus, in trying to make maximal use of possibly relevant cues for their rating task, individuals apply the implicit associations in their inference. We concluded that, to control for this mechanism, it would be advisable to use a forced-choice recognition task in future research (participants having to decide on the best alternative for the information carried by different AUs and AU combinations).

We did find some significant effects for avatar identity that can be interpreted as preexisting stereotypes about gender in particular (e.g., males being perceived as making high power appraisals more easily) and/or about facial architecture distinguishing different identities (the avatars were produced from photographs of actual persons). However, the fact that there were no significant interaction effects with the AU factors suggests that the findings can be generalized to different types of avatars. Nevertheless, it might be advisable for future research to avoid using gender-typed attributes, particularly salient facial appearances and striking backgrounds that might elicit context stereotypes.

The rater background variables, age and gender, did not reach significance, suggesting a high degree of uniformity of the inference rules. However, as predicted, we found many significant correlations with the scores on test of emotion recognition ability consisting of multimodal video stimuli (actor portrayals). This relationship should be further explored in future research, as it may allow a better understanding of the underlying inference mechanisms.

Study 3

Research Questions and Experimental Design

In the preceding studies, only single AUs and potentially important AU combinations predicted to be part of a complete emotion display were used as stimuli. The results show that some individual AUs orient observer inferences in the direction of particular appraisals, in particular AU 4 and AUs 9 and 10, toward negative valence appraisal, and AU 12 toward positive valence appraisal. However, these individual AUs are rarely specific for appraisal dimensions such as control and power, which are needed to differentiate the whole gamut of human emotions. We also examined pairwise combinations of AUs that have been presented in the literature as powerful indicators of specific emotions. Contrary to expectations, we found that these combinations have generally less signal value than do their individual constituents. The obvious conclusion is that inferences of the results of specific appraisal checks must be based, like emotion inferences, on complex configurations of facial AUs that reflect the large variety of the motor responses activated during emotion episodes. The purpose of Study 3 is to experimentally test precise predictions for specific AU configurations for the inference of both appraisals and emotions. Unlike the rating procedure adopted in Studies 1 and 2, in Study 3 we chose a forced-choice answer format, assuming that this method is more appropriate for the task and likely to reduce the halo effects found in previous studies.

Task 1: Inferring appraisals from complex AU configurations specific to major appraisal checks. Table 4 shows the specific predictions and the 10 corresponding AU combinations used in Study 3. In Task 1 of this study, the 10 synthesized video stimuli representing these combinations were presented in a forced-choice task that asks the participants to choose the best of eight alternatives for each experimental stimulus.

Task 2: Inference of emotions from combined appraisal configurations according to their theoretically defined appraisal patterns. For this task, we hypothesized that a combination of several appraisals as expressed by specific AU combinations would allow the inference of emotion categories. In this task, we used the same AU combinations as in Task 1, but presented them successively in a video animation, in the sequence predicted by the CPM, to form a new, more complex, combination of appraisals corresponding to a given emotion or emotion family. For example, as shown in Table 1, fear is expected to result from the sequential cumulative appraisals of first suddenness, then obstructiveness, and finally low power or coping potential. Using the AU predictions shown in Table 4 we prepared a fear facial synthesis by first animating the combined AUs 1, 2, 5, 26 for the suddenness appraisal, then 4, 11, 15, 20, 23, or 4, 11, 15, 20, 23 for obstructiveness, and finally 16, 25, 26, 43, 54 for low power. The reader can visualize this sequence by using the illustrations in Figure 5 and imagining the sequential dynamic animation of these patterns. In some cases, only one or two components were used to simulate an emotion, for example pleasantness for enjoyment and unpredictability and pleasantness for interest. As shown in Table 4 and Figure 5, we used two alternative options for low and high power.

The following 10 target emotions were defined, based on the predictions for the corresponding appraisal outcomes (with two alternatives for five emotions, as shown in parentheses):

Surprise	Sudden
Enjoyment	Pleasant
Contentment	Conducive
Disappointment	Obstructive (2)
Interest	Unpredictable + Pleasant
Happiness	Pleasant + Conducive
Sadness	Obstructive + Low Power (2)
Fear	Sudden + Obstructive + Low Power (2)
Anger	Unpredictable + Obstructive + High Power (2)
Disgust	Unpleasant + Obstructive + High Power (2)

We added three additional emotion labels (pride, despair, contempt) to the 10 target emotions on the judgment instrument, for two reasons: to make the task more difficult and to avoid our results being attributed to guessing or elimination (DiGirolamo & Russell, 2017), and to account for the many frequent confusions found in the literature for some of the target emotions, especially Happiness/Pride, Sadness/Despair, Fear/Despair, Fear/Surprise, Anger/Contempt, and Disgust/Contempt (Bänziger & Scherer, 2010, Table 6.1.5; Bänziger et al., 2012, Table 6 and Table C2 in their supplemental materials). We were interested to see whether we would find the same confusion patterns with our synthesized facial expression. Thus, pride shares the same basic appraisal structures happiness (Pleasant + Conducive), contempt is similar to anger and disgust (Obstructive + High Power), and despair has something common with bot sadness and fear—Obstructive + Low Power. In consequence we expect frequent confusions. In addition, we expected that some of the emotion targets that were based only on one single appraisal result (e.g., disappointment following a goal obstruction) would be confused with a more complex emotion (e.g., fear, which also implies obstructiveness of an event). The detailed design for Task 2 is shown in Table S8 in SOM.

Task 3: Test of emotion recognition ability. As in Study 2 we administered the GERT-S to determine whether participants with high emotion recognition inferred the predicted target category with higher accuracy.

Method

Participants. To facilitate comparison, we used a similar sample of participants in a web survey setting to that used in Study 2. We recruited 134 participants (50% females, 20 to 83 years, $M = 41.36 \pm 14.52$) via the Qualtrics survey panel in exchange for monetary compensation. The selection criteria were as follows: native English speakers, Caucasian origin, and over 18 years old. We also requested the Qualtrics panel management to recruit about 50% males and females, as well as an age range split evenly between three age

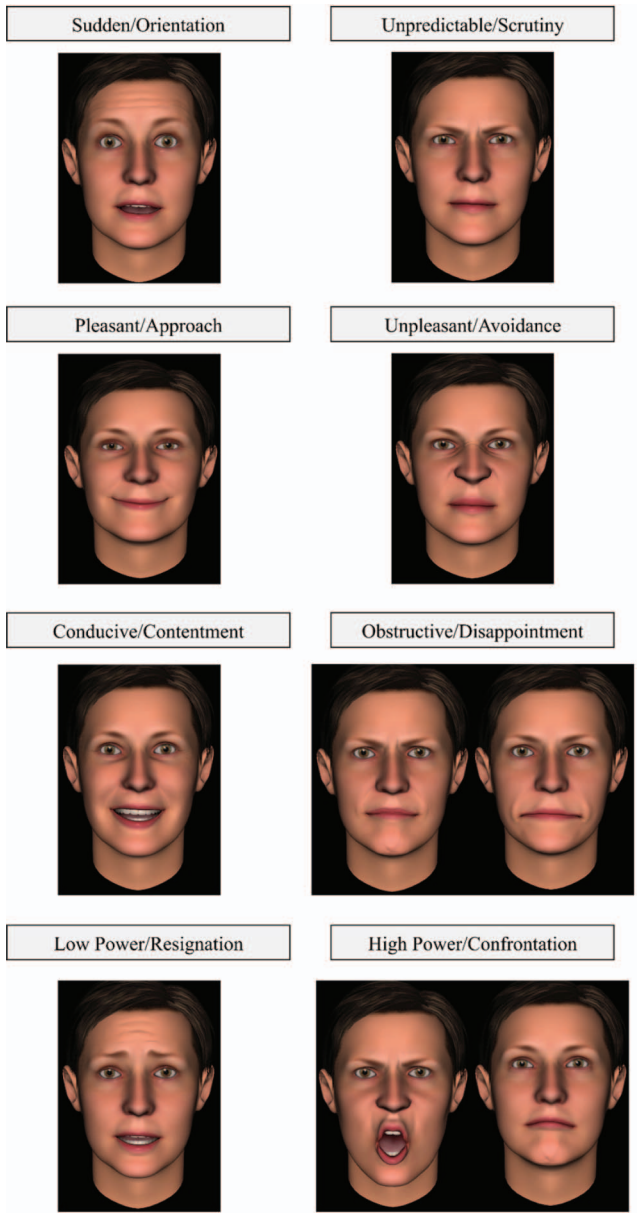

Figure 5. Predicted Action Unit (AU) combinations for the outcomes of major appraisals. See the online article for the color version of this figure.

categories, 18–30, 31–45, and over 45, to produce a sample as representative of the general population as possible.

Stimulus design and production. The study comprised the three different tasks described above with different sets of stimuli for each task. Based on the conclusions concerning avatar identity and background features reached in Study 2, we redesigned the synthesis of the avatars.

Task 1: Appraisal ratings. The stimuli for this task were produced on the basis of two computer-generated avatar faces: one male, one female. For each gender, 10 stimuli were produced by combining different AUs on the basis of CPM predictions of appraisals, as shown in Table 4 and illustrated in Figure 5. The 10 combinations were identical for both genders and comprised between two and six AUs expressed simultaneously. The total duration of each stimulus was 2.96 s and was divided into two parts (see Figure S3 in SOM): (a) expression unfolding: dynamic expression starting from a neutral face and gradually reaching an apex (from 0.00 to 1.00 s), and (b) still apex expression (from 1.00 to 2.96 s). It should be noted that not all AUs in a group combination were presented with the same degree of intensity (FACSGen allows to vary between 0 and 100%). The exact intensity to be used for a particular AU depended on the relative importance (as reflected in the frequency of mention in the literature, see Table S1 in SOM) as well as on the degree of compatibility with other AUs in the same animation). Both pretests and expert judgments were used in making final decisions on the relative intensity of the AUs to be combined.

Task 2: Emotion inference. We used the same stimuli as in Task 1 but combined into theoretically predicted sequences as described above. Separately for each avatar gender, 15 stimulus combinations were created for the 10 target emotions. In addition, three random-sequence filler stimuli were produced, yielding 36 stimuli to be judged. The duration of each video was 2.96 s for the single expressions, 3.96 s for the two-component combinations, and 5.08 s for the three-component combinations. As in Task 1, each video stimulus ended with a 2-s still apex expression.

Task 3: Emotion recognition test. We again administered the GERT-S, as described in Study 2.

Procedure.

Task 1: Appraisal ratings. The instructions for this task were comparable to those in Study 2 except that for each face, the participants were asked to choose the one thought, reaction, or behavioral intention that, in their view, *best explains* the facial expression. Participants had to choose one of eight alternatives (appraisal checks and glosses used for ease of comprehension) on four dimensions:

(1) *Novelty*
Type of event = occurring suddenly versus requiring close attention or scrutiny
(2) *Pleasantness*
Type of bodily = sensation pleasant sensation versus unpleasant sensation
(3) *Goal conduciveness*
Relation to goals = what the person wanted versus not at all what the person wanted
(4) *Power/Coping ability*
Behavioral intention = resignation versus confrontation

After reading the instructions, participants were presented with two training examples (one for each gender) to become familiar with the task. During the main task, the 20 videos were automatically presented in random order, and participants could replay each video only once in case they missed the first presentation. After viewing each video, the participants were presented with the response table from which they could select one option by clicking the appropriate cell (see Figure S4 in SOM). Each option corresponded to one type of appraisal (novelty, pleasantness, goal conduciveness, power) and to one of the two alternatives on each dimension (Option A or B).

Task 2: Emotion inference. In this task, the instructions first explained to the participants that emotions are often generated by complex configurations of several thoughts or evaluations and that the expressions of these thoughts had been arranged in different sequences corresponding to the emergence of different emotions. The participants were instructed that they were going to view 36 videos with two avatar faces expressing these emotions and that they would have to choose the *single most appropriate emotion label* for the expression among 14 different emotion words (we added "relief" for appropriate balance between positive and negative emotions. We also provided the category "None of these" to further discourage random guessing.

As in Task 1, participants started with two practice examples to familiarize themselves with the task, and they had the opportunity to read the instructions a second time before starting the task. In the main task, the 36 videos were automatically played in random order, and participants could replay them once if necessary. They were then presented with a response wheel with the 14 emotion words, from which they had to choose one option by clicking the circle next to the word (see Figure S5 in SOM).

Task 3: Emotion recognition test. The procedure for the administration of the GERT-S was the same as that described for Study 2.

Results and Discussion

Given the forced-choice response format, we determined for each participant and stimulus whether the response chosen corresponded to one of the predicted target emotions. If this was the case, it was considered a hit (in the sense of corresponding to prediction). In this way, we could compute a personal hit rate for each participant that could then be analyzed for individual differences. At the aggregate level, for each of the appraisal and emotion categories intended by the AU combinations chosen for synthesis, the total percentage of participants having chosen the respective category was computed and corrected for the respective chance level. Note that the resulting tables do not correspond to the confusion matrices provided in typical emotion recognition studies because (a) the matrices are not symmetrical and (b) there is no ground truth that is compared one-to-one with a particular response category. As outlined in the introduction, given the fact that the CPM predicts complex blends of emotions based on the combination of appraisals that determine the process, we are interested in the broader range of inferences made by the judges. For this reason, established correction procedures (e.g., the unbiased hit rate; Wagner, 1993) do not apply. However, as we wanted to account for the uneven distribution of response frequencies over the categories in interpreting what we consider "hits" with respect to our predictions, we have corrected the raw percentages by the response-specific chance levels (given the frequency of category use). Both the raw and the corrected accuracy percentages are shown in Tables S9 and S10 in SOM, where we also provide the

chance levels adjusted for the frequency of responses for each category. Thus, for interpretation, one can either use the raw percentages and the adjusted chance levels or the adjusted percentages with the theoretical chance level. In the article, we will only report the corrected percentages, which can be directly compared to the theoretical chance levels.

Task 1: Appraisal judgments. We computed separate response distribution matrices for the male and the female avatar, but as the profile correlations reached an average of .88, we present in Table 5 the overall accuracy matrix for the appraisal judgments for both avatar genders combined. In this case, theoretical chance level amounts to 11.1%, given that there were nine alternatives to choose from. In the following discussion, we use only the corrected percentages that can be directly compared to the theoretical chance level.

All except one of the AU combinations that was predicted to produce a specific appraisal inference exceeded chance level, in many cases by very substantial margins. The exception was AUs 4 + 10 + 22 + 27, which was not identified as confrontation, but rather as not wanted (which is the usual precursor for confrontation). The important confusions, exceeding double the chance level (22.22), are mostly attributable to semantic similarity, for example, within valence: "wanted" with "pleasant" on the one hand and "not wanted" with "unpleasant" on the other. An interesting case is the confusion between "scrutiny" on the one hand and "not wanted" and "confrontation" on the other. This effect is most likely due to frowning (AU 4) or a piercing look (AU 7), which are both frequent facial responses in cases where things do not work out as expected. On the whole, the results strongly support the underlying theoretical assumption that observers can reliably deduce highly specific appraisal results from appropriate facial action configurations.

To examine individual differences in this ability, we computed the percentage of hits, as defined above, for each participant. The average proportion across all 134 participants was .38. There were no gender or age differences in accuracy, but a strong correlation, $r = .63$, $p < .001$, with emotion recognition ability, as measured by the GERT. This finding suggests that the ability to correctly infer appraisals from the face might be a precondition for the ability to correctly deduce the nature of the corresponding emotion episode.

Task 2: Emotion judgment. The profile correlations of the separate response distribution matrices for the two avatar identities (male and female) reached $r = .89$, and thus Table 6 shows the overall accuracy matrix for the emotion judgments for both avatar genders combined. In this case, theoretical chance level amounts to 6.66%, given that there were 15 alternatives to choose from. As for Task 1, we corrected the raw percentages by the category-specific response bias (see Table S10 in SOM) and compare the corrected percentage to the theoretical chance level. Hit rates exceeding chance level by approximately a factor of 2 or more are found for surprise, happiness, enjoyment, interest, sadness, disgust, disappointment, and anger (alternative 2 only). Except for anger and surprise, the accuracy percentages are relatively low, but virtually all confusions concern highly related emotions. Thus, as one might expect from similar results for the recognition of actor-portrayed emotions (see Bänziger & Scherer, 2010, Table 6.1.5; Bänziger et al., 2012, Table 6 and Table C2 in their supplemental materials), there are clusters of confusion between (a) happiness, enjoyment, and contentment; (b) anger, disgust, and contempt; and (c) fear, sadness, and despair. Interestingly, in the current study, disappointment is a frequent inference applied to fear, sadness, disgust, and contempt target stimuli. Thus, a central appraisal signal of obstructiveness ("not expected, not wanted") seems to be a major guideline in determining the nature of an emotion, which supports the CPM prediction that novelty, pleasantness, and goal conduciveness are the first appraisals in the predicted sequence, orienting the emotion to a major direction of negative valence. Overall, these results encourage further attempts to understand the mechanisms underlying emotion recognition as an application of deduction rules from appraisal inferences that are generated by specific appraisal configurations.

We computed the proportion of responses reflecting the choice of the primary target for each participant, which yielded an average hit rate proportion (in the sense of highly specific target prediction) of .37 across all 134 participants. There were no gender differences, but we found a low positive correlation of hit rate with age, $r = .19$, $p < .05$. As in the case of appraisal inference, emotion recognition from avatar videos correlated strongly with GERT-S

Table 5
Overall Response Percentages and Chance Levels for Task 1: Appraisal Judgments

Appraisal checks	Action unit configurations	Sudden	Scrutiny	Pleasant	Unpleasant	Wanted	Not wanted	Resignation	Confrontation	None of these
Novelty										
Sudden	AUs 1, 2, 5, 26	**50.6**	4.3	11.7	2.1	6.8	2.7	3.3	2.2	4.9
Scrutiny	AUs 4, 7	5.2	**31.9**	.8	9.0	1.8	11.3	3.3	5.9	13.6
Intrinsic pleasantness										
Pleasant	AUs 6, 13, 14, 23, 43, 53	3.4	1.1	**44.5**	.2	40.0	.4	2.4	9.1	5.8
Unpleasant	AUs 9, 10	.9	4.9	2.4	**37.8**	2.1	6.9	6.1	7.0	8.7
Goal conduciveness										
Conducive/Wanted	AUs 5, 6, 12, 25, 27	8.6	2.3	32.4	.5	**38.2**	1.6	1.4	11.8	8.7
Obstruction/Not wanted (1)	AUs 4, 11, 14, 17, 24	4.0	19.5	2.0	11.4	2.1	**16.3**	5.7	9.7	7.8
(2)	AUs 4, 11, 15, 20, 23	3.1	2.6	.8	19.3	.7	**18.6**	20.8	3.2	9.7
Coping/Power										
Low/Resignation	AUs 1, 4, 16, 25, 26, 43, 54	11.7	9.2	1.2	6.0	2.9	14.3	**32.5**	3.2	7.8
High/Confrontation (1)	AUs 7, 17, 53	2.8	18.4	2.0	4.2	3.9	6.0	17.0	**33.9**	28.2
(2)	AUs 4, 10, 22, 27	9.8	5.7	2.0	9.5	1.4	21.8	7.5	**14.0**	4.9

Note. For each action unit (AU) configuration, rater bias corrected percentages are provided (the raw percentages are listed in Table S10 in SOM). Values in boldface represent the predictions listed in Table 4. At the bottom of the table, the chance levels corrected for rater bias in the direction of specific categories are shown.

Table 6
Overall Response Percentages and Chance Levels for Task 2: Emotion Judgment

Target emotions and predicted appraisal configurations	Emotion rating labels														
	Pride	Happiness	Enjoyment	Contentment	Disappointment	Interest	Relief	Surprise	Fear	Despair	Sadness	Disgust	Contempt	Anger	None of these
Surprise/sud	1.3	1.3	4.2	3.4	.3	9.3	6.5	**51.2**	1.1	1.6	.5	.3	.7	1.1	3.5
Enjoyment/ple	25.6	19.3	**15.8**	27.9	.3	3.4	3.8	.8	.0	.0	.5	.5	2.0	.3	4.5
Contentment/con	8.8	29.6	24.1	**5.8**	.0	5.3	6.0	1.6	1.1	.4	.0	.8	1.2	.5	4.5
Disappointment/obs (1)	2.0	1.3	1.5	3.8	**12.9**	8.0	.6	1.1	.0	6.2	4.7	16.1	15.0	10.2	6.0
Disappointment/obs (2)	1.3	.8	1.2	2.1	**12.2**	3.4	.6	.0	5.1	8.5	22.2	18.0	9.1	4.0	8.9
Interest/upr + ple	16.8	10.5	**10.1**	21.2	1.0	**13.9**	29.2	1.9	.0	.0	.5	.5	2.7	.5	8.0
Happiness(Pride) ple + con	**12.8**	**28.8**	23.9	5.8	.4	6.3	8.1	1.9	.0	.0	.9	.3	.4	.0	2.0
Sadness(Despair)/obs + lop (1)	2.0	.8	2.4	2.5	**14.1**	6.3	10.3	5.6	18.0	**19.0**	**15.1**	3.0	4.3	.5	8.0
Sadness(Despair)/obs + lop (2)	2.0	1.3	1.5	5.5	**11.1**	5.3	6.0	4.5	20.9	**19.0**	**18.9**	3.6	5.2	.0	11.4
Fear (Despair)/sud + obs + lop (1)	2.0	1.5	3.3	2.5	10.8	7.0	8.7	**13.7**	**16.9**	**13.9**	13.7	3.0	3.6	.3	7.5
Fear (Despair)/sud + obs + lop (2)	2.7	1.5	3.3	3.8	11.8	7.6	7.1	**9.5**	**22.0**	**15.1**	11.4	3.3	5.2	.3	8.0
Anger (Contempt)/upr + obs + hip (1)	12.1	1.5	2.4	7.5	10.8	12.1	6.5	.8	2.9	6.2	6.2	7.6	**19.3**	**3.0**	8.4
Anger (Contempt)/upr + obs + hip (2)	2.0	1.0	1.5	1.2	2.9	.6	.6	4.5	7.0	3.8	.9	10.1	**4.3**	**39.0**	4.9
Disgust(Contempt)/unp + obs + hip (1)	7.4	.5	3.3	6.3	7.9	9.8	4.4	.5	.0	4.7	3.8	**17.2**	**20.9**	3.2	10.4
Disgust(Contempt)/unp + obs + hip (2)	1.3	.3	1.5	.8	3.6	1.7	1.6	2.5	5.1	1.6	.5	**15.6**	**5.9**	37.1	4.0

Note. For each target emotion and judgment category rater bias corrected percentages (in italics) are provided. Values in bold represent the predictions (both primary and secondary, expected because of confusions) listed in the text and in Table S8 in SOM. At the bottom of the table, the chance levels corrected for judge bias for choosing a certain category are listed (raw response percentages and the theoretically expected chance level are provided in Table S10 in SOM). sud = Sudden; upr = Unpredictable + Pleasant; ple = Pleasant; unp = Unpleasant; con = Conducive; obs = Obstructive; lop = Low Power; hip = High Power.

scores, $r = .61$, $p < .001$. An important additional finding was a correlation of .53 ($p < .001$) between correct appraisal inference (Task 1) and emotion recognition (Task 2) from the avatar videos. In other words, participants who were very good at recognizing the targeted individual appraisals were also very good at identifying the targeted emotions based on sequences of appraisal markers. It should be noted that these correlations of hit rate are based on the very stringent definition of hit in terms of the primary target only. We would expect still higher correlations if secondary target choices were also considered. This pattern confirms the assumption that the ability to correctly infer appraisals from the face might be a precondition for the ability to correctly deduce the nature of the corresponding emotion episode.

These results demonstrate that the design decisions taken based on the results of Study 2 paid off. First, rather than using the classic AU pairs suggested in the literature for facial emotion expression, it is necessary to carefully construct specific AU combinations for the different appraisal checks, up to six or seven AUs, to obtain clear discrimination. Second, given the important halo effects, especially a strong valence superfactor, it is advisable to choose a forced-choice paradigm rather than open dimension ratings to demonstrate the discrimination ability of human judges for appraisal differences. Third, it was useful to reduce the number of additional cues (gender-typed facial architecture, background) to focus judgments on the synthesized facial movements (reducing the effects of avatar identity).

General Discussion and Conclusion

In this series of three studies, we used highly innovative technology to experimentally manipulate the central variable to investigate the recognition of emotion from facial expression and the precise nature of the configuration of facial movements. Given the lack of established models for the facial synthesis of emotions and appraisals, as well as the complex judgment procedures required by the theoretical design, we attempted to consecutively maximize power across consecutive studies by applying what has been learned from the earlier results. For example, experimental testing of the optimal intensity of AU expressions allowed us to find the right balance between strength of impact and credibility/authenticity. Similarly, we continuously refined our judgment/rating procedures from the occurrence of halo effects in the earlier studies.

We paid particular attention to the issue of statistical power. As the design does not allow for strict hypothesis testing on the basis of individual observations, we could not use the classic methods of estimating sample size. Rather, given that the data essentially consist of differences between group means of proportions (e.g., average hit rates for different emotions), our concern was to ensure the stability of group means. Means tend to become stable after about 15 cases and most studies in this field work with groups of about 20–30 judges. In Studies 1 and 2, we therefore targeted and obtained an N of about 30. Given the complexity of Study 3, and the interest in individual differences, we targeted a much larger sample of over 130 judges.

The latter decision was also motivated by the concern for sufficient diversity of our participant sample to ensure that the mechanism predicted to underlie emotion recognition from the face would hold across gender, age, and social background differences. For Study 1 (which had to be conducted in the laboratory to fine-tune the new technology) we used a gender-balanced student sample and a relatively large age range. For studies 2 and 3, we used professional survey samples to obtain large groups of participants from a wide variety of socioeconomic and cultural backgrounds, equalized for gender and systematically selected to cover three major age groups. Using this sampling procedure allowed us not only to demonstrate that the existence of the predicted mechanism is largely independent of special group effects, but also to discover massive correlations between the performance in the experimental judgment tasks and a validated test of emotion recognition competence.

All in all, the significance and stability of the results obtained, despite the unusual stimuli and the decision tasks with a large and heterogeneous sample of adults, suggest that the findings can be generalized and thus contribute to the cumulative theoretical knowledge in psychology and serve as a solid basis for future work in this area. In particular, the use of dynamic facial expression synthesis with naturalistic avatars (see also de Melo, Carnevale, Read, & Gratch, 2014; Jack, Garrod, & Schyns, 2014; Joyal, Jacob, Cigna, Guay, & Renaud, 2014) has made it possible, for the first time, to perform tightly controlled experiments on facial emotion recognition, rather than relying on judgment studies with either real-life recordings (which are beset by a multitude of uncontrollable factors) or actor portrayals of emotion (which may suffer from effects of differential expression abilities of the actors). Most important, in studies on emotion recognition based on these two types of stimuli, researchers had to work with the inference of full-blown emotions, generally only prototypical renderings of a few basic emotions.

Much of this earlier research has been atheoretical in nature, focusing mostly on differential recognition ability for different emotions. In contrast, in this study, we are concerned with the underlying mechanisms of emotion recognition—and the degree to which this is linked with the mechanism underlying emotion expression. On the basis of the CPM, we postulate that it is the cumulative sequence of individual appraisals that shapes dynamic changes in facial movements and constitutes emotion expression. This mechanism allows an enormous range and variety of different expression patterns that largely surpass the few prototypical expressions that have been proposed for basic emotions. The virtually limitless possibilities of AU combinations in complex dynamic time series affords the expression of a large variety of affective episodes, many of which are much more subtle than the classic list of major emotions. We expect this expression mechanism to be mirrored by similarly structured recognition mechanisms that operate in reverse, in which specific appraisals are recognized from their signature AUs and the most likely emotion category inferred from the appraisal results.

Although the results obtained in the study by Sergi et al. (2016) and in the current research program provide no direct evidence for the existence of such a mechanism, they are highly compatible with it. Our predictions on the appraisal meaning of the synthesized facial actions were based on theoretical assumptions about the functions of specific facial actions, as first proposed by Darwin (1872/1998; see also Lee et al., 2013), and on extrapolation from empirical evidence for the occurrence of specific AUs in real-life or portrayed emotions. We have shown that observers strongly agree on the meaning of these facial action configurations and accurately infer the nature of the respective appraisal.

Critics of this model could argue that observers might first recognize an emotion and only then infer the appraisal that is most likely to be associated with it. However, this alternative explana-

tion is not supported by our results, given that Tables 5 (Task 1 of Study 3) and 6 (Task 2 of Study 3) show that both the agreement and the hit rate is lower for emotions than for appraisals (even though more cues are provided in the case of the former), suggesting a primary, more powerful signal structure for appraisal. Furthermore, accurate appraisal inference correlates as highly with test scores for general emotion recognition ability as it does for emotion recognition from avatar faces. This would be unlikely if emotions were inferred first and appraisals deduced only subsequently, as the deduction process should introduce additional error.

Critics may also argue that the results reported here are mainly due to guessing and elimination effects. Obviously, such artifacts can never be completely ruled out in judgment studies. Unfortunately, so far no appropriate alternative methods have been developed. We have done our best to avoid such artifacts by providing a large number of credible choices and by correcting scores and chance level for response bias, making the guessing option rather unlikely.

Other critics will probably argue that our results have been obtained in the absence of context cues and that different contexts will make these results disappear. Of course, real life perception and inference rarely occur outside of a social context providing a myriad of more or less relevant cues. As social perception researchers have pointed out very early (e.g., Bruner & Tagiuri, 1954), human observers are strongly influenced by social context cues. However, this does not invalidate research on reliable facial markers of cognitive appraisal and behavior preparation, especially as many of these can be shown to have (or at least have had in an evolutionary perspective) concrete functions. Outside of the laboratory, social context cues are likely to be often concordant with the facial expressions shown by the protagonists, reinforcing the information provided by faces. As Wallbott (1988) has shown, situation information is more important to judges if it is discrepant rather than consonant with facial cues. As to sender characteristics, our results consistently show the absence of interaction effects between appraisal or emotion judgments on the one hand and gender or identity of the avatars on the others. In any case, as outlined in the introduction, the purpose of this study was to examine the *nature of inference patterns* based on theoretically predicted patterns of facial expression, not the relative importance of face cues versus context cues in impression formation in social settings.

Finally, it can be argued that our results have been obtained with Western participants and may well not hold up in very different cultures. This is possible, of course, but it is an empirical question. Given the evolutionary stable, functional underpinnings of many of the facial patterns studied here, it is not impossible to find at least some degree of communality. In fact, we hypothesize that appraisal judgments should be more stable across cultures than emotion judgments, given that they are more basic building blocks and that their definition is less affected by differences in the meaning of verbal emotion labels.

Further research is needed to provide replication and to examine the hypothesized mechanisms in greater detail and across different cultures. The facial synthesis method proposed here provides almost limitless possibilities for further experimental studies with a high level of control of the relevant variables. Further improvement can be made for the number and complexity of AU combinations to be used, the precise control of duration parameters of single AUs within the combination, and the relative strength of the expression of different AUs, to name but a few of the variables that are likely to enhance our understanding of the underlying processes. For example, it may be possible to explain the common patterns of confusion in recognition from the relative power of certain AU combinations.

Notably, the appraisal-based mechanism advocated here does not contradict other theories of facial expression and recognition such as the basic emotion (Ekman, 2004), dimensional-contextual (Russell, 1997), or action tendency (Frijda & Tcherkassof, 1997) perspectives. Advocates of these theories accept the important role of appraisal processes in emotion elicitation, and the mechanisms described here are compatible with and complementary to their proposals. In consequence, adoption of the approach described and empirically investigated in this article may encourage efforts to further our theoretical and empirical understanding of the facial emotion expression and the recognition process via hypothesis-based research paradigms.

References

Adolphs, R. (2002). Recognizing emotion from facial expressions: Psychological and neurological mechanisms. *Behavioral and Cognitive Neuroscience Reviews, 1,* 21–62. http://dx.doi.org/10.1177/1534582302001001003

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,* 347–357. http://dx.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,* 158–164. http://dx.doi.org/10.1016/j.biopsycho.2008.04.004

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology, 70,* 614–636. http://dx.doi.org/10.1037/0022-3514.70.3.614

Bänziger, T., Mortillaro, M., & Scherer, K. R. (2012). Introducing the Geneva Multimodal expression corpus for experimental research on emotion perception. *Emotion, 12,* 1161–1179. http://dx.doi.org/10.1037/a0025827

Bänziger, T., & Scherer, K. R. (2010). Introducing the Geneva Multimodal Emotion Portrayal (GEMEP) corpus. In K. R. Scherer, T. Bänziger, & E. B. Roesch (Eds.), *Blueprint for affective computing: A sourcebook* (pp. 271–294). Oxford, England: Oxford University Press.

Bruner, J. S., & Tagiuri, R. (1954). The perception of people. In G. Lindzey (Ed.), *Handbook of Social Psycholgy* (Vol. 2, pp. 634–654). Cambridge, MA: Addison Wesley.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. Berkeley, CA: University of California Press.

Bühler, K. (1934). *Sprachtheorie* [Theory of language]. Jena, Germany: Fischer.

Campos, B., Shiota, M. N., Keltner, D., Gonzaga, G. C., & Goetz, J. L. (2013). What is shared, what is different? Core relational themes and expressive displays of eight positive emotions. *Cognition and Emotion, 27,* 37–52. http://dx.doi.org/10.1080/02699931.2012.683852

Carroll, J. M., & Russell, J. A. (1997). Facial expressions in Hollywood's portrayal of emotion. *Journal of Personality and Social Psychology, 72,* 164–176. http://dx.doi.org/10.1037/0022-3514.72.1.164

Darwin, C. (1998). *The expression of the emotions in man and animals* (3rd ed., P. Ekman, Ed.). London, England: HarperCollins. (Original work published 1872)

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,* 316–328. http://dx.doi.org/10.1037/a0015369

de Melo, C. M., Carnevale, P. J., Read, S. J., & Gratch, J. (2014). Reading people's minds from emotion expressions in interdependent decision making. *Journal of Personality and Social Psychology, 106,* 73–88. http://dx.doi.org/10.1037/a0034251

DiGirolamo, M. A., & Russell, J. A. (2017). The emotion seen in a face can be a methodological artifact: The process of elimination hypothesis. *Emotion, 17,* 538–546. http://dx.doi.org/10.1037/emo0000247

Du, S., Tao, Y., & Martinez, A. M. (2014). Compound facial expressions of emotion. *PNAS Proceedings of the National Academy of Sciences of the United States of America, 111,* E1454–E1462. http://dx.doi.org/10.1073/pnas.1322355111

Ekman, P. (2004). What we become emotional about. In A. S. R. Manstead, N. H. Frijda, & A. H. Fischer (Eds.), *Feelings and emotions: The Amsterdam Symposium* (pp. 119–135). Cambridge, United Kingdom: Cambridge University Press.

Ekman, P., & Friesen, W. V. (1978). *Facial action coding system: A technique for the measurement of facial movement*. Palo Alto, CA: Consulting Psychologists Press.

Ekman, P., Friesen, W. V., & Hager, J. C. (2002). *The facial action coding system* (2nd ed.). Salt Lake City, UT: Research Nexus eBook.

Ellsworth, P. C., & Scherer, K. R. (2003). Appraisal processes in emotion. In R. J. Davidson, K. R. Scherer, & H. Goldsmith (Eds.), *Handbook of the affective sciences* (pp. 572–595). New York, NY: Oxford University Press.

Fernández-Dols, J. M., & Russell, J. A. (Eds.). (2017). *The science of facial expression*. Oxford, UK: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780190613501.003.0024

Frijda, N. H., & Philipszoon, E. (1963). Dimensions of recognition of expression. *The Journal of Abnormal and Social Psychology, 66,* 45–51. http://dx.doi.org/10.1037/h0042578

Frijda, N. H., & Tcherkassof, A. (1997). Facial expressions as modes of action readiness. In J. A. Russell & J. Fernández-Dols (Eds.), *The psychology of facial expression* (pp. 78–102). Cambridge, UK: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511659911.006

Galati, D., Scherer, K. R., & Ricci-Bitti, P. E. (1997). Voluntary facial expression of emotion: Comparing congenitally blind with normally sighted encoders. *Journal of Personality and Social Psychology, 73,* 1363–1379. http://dx.doi.org/10.1037/0022-3514.73.6.1363

Gentsch, K., Grandjean, D., & Scherer, K. R. (2015). Appraisals generate specific configurations of facial muscle movements in a gambling task: Evidence for the component process model of emotion. *PLoS ONE, 10,* e0135837. http://dx.doi.org/10.1371/journal.pone.0135837

Gosselin, P., Kirouac, G., & Doré, F. Y. (1995). Components and recognition of facial expression in the communication of emotion by actors. *Journal of Personality and Social Psychology, 68,* 83–96. http://dx.doi.org/10.1037/0022-3514.68.1.83

Hareli, S., & Hess, U. (2012). The social signal value of emotions. *Cognition and Emotion, 26,* 385–389. http://dx.doi.org/10.1080/02699931.2012.665029

Jack, R. E., Garrod, O. G. B., & Schyns, P. G. (2014). Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. *Current Biology, 24,* 187–192. http://dx.doi.org/10.1016/j.cub.2013.11.064

Joyal, C. C., Jacob, L., Cigna, M.-H., Guay, J.-P., & Renaud, P. (2014). Virtual faces expressing emotions: An initial concomitant and construct validity study. *Frontiers in Human Neuroscience, 8,* 787. Advance online publication. http://dx.doi.org/10.3389/fnhum.2014.00787

Kaiser, S., & Wehrle, T. (2001). Facial expressions as indicators of appraisal processes. In K. R. Scherer, A. Schorr, & T. Johnstone (Eds.), *Appraisal processes in emotions: Theory, methods, research* (pp. 285–300). New York, NY: Oxford University Press.

Kaiser, S., Wehrle, T., & Schmidt, S. (1998). Emotional episodes, facial expression, and reported feelings in human-computer interactions. In A. H. Fischer (Ed.), *Proceedings of the Xth conference of the International Society for Research on Emotions* (pp. 82–86). Würzburg, Germany: ISRE.

Krumhuber, E. G., Kappas, A., & Manstead, A. S. (2013). Effects of dynamic aspects of facial expressions: A review. *Emotion Review, 5,* 41–46. http://dx.doi.org/10.1177/1754073912451349

Krumhuber, E. G., & Scherer, K. R. (2011). Affect bursts: Dynamic patterns of facial expression. *Emotion, 11,* 825–841. http://dx.doi.org/10.1037/a0023856

Krumhuber, E. G., Tamarit, L., Roesch, E. B., & Scherer, K. R. (2012). FACSGen 2.0 animation software: Generating three-dimensional FACS-valid facial expressions for emotion research. *Emotion, 12,* 351–363. http://dx.doi.org/10.1037/a0026632

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

Lee, D. H., Susskind, J. M., & Anderson, A. K. (2013). Social transmission of the sensory benefits of eye widening in fear expressions. *Psychological Science, 24,* 957–965. http://dx.doi.org/10.1177/0956797612464500

Meaux, E., & Vuilleumier, P. (2016). Facing mixed emotions: Analytic and holistic perception of facial emotion expressions engages separate brain networks. *NeuroImage, 141,* 154–173. http://dx.doi.org/10.1016/j.neuroimage.2016.07.004

Mehu, M., Mortillaro, M., Bänziger, T., & Scherer, K. R. (2012). Reliable facial muscle activation enhances recognizability and credibility of emotional expression. *Emotion, 12,* 701–715. http://dx.doi.org/10.1037/a0026717

Mehu, M., & Scherer, K. R. (2012). A psycho-ethological approach to social signal processing. *Cognitive Processing, 13,* 397–414. http://dx.doi.org/10.1007/s10339-012-0435-2

Mehu, M., & Scherer, K. R. (2015). Emotion categories and dimensions in the facial communication of affect: An integrated approach. *Emotion, 15,* 798–811. http://dx.doi.org/10.1037/a0039416

Mendolia, M. (2007). Explicit use of categorical and dimensional strategies to decode facial expressions of emotion. *Journal of Nonverbal Behavior, 31,* 57–75. http://dx.doi.org/10.1007/s10919-006-0020-4

Mortillaro, M., Mehu, M., & Scherer, K. R. (2011). Subtly different positive emotions can be distinguished by their facial expressions. *Social Psychological and Personality Science, 2,* 262–271. http://dx.doi.org/10.1177/1948550610389080

Mortillaro, M., Mehu, M., & Scherer, K. R. (2013). The evolutionary origin of multimodal synchronization and emotional expression. In E. Altenmüller, S. Schmidt, & E. Zimmermann (Eds.), *Evolution of emotional communication: From sounds in nonhuman mammals to speech and music in man* (pp. 3–25). New York, NY: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780199583560.003.0001

Mortillaro, M., Meuleman, B., & Scherer, K. R. (2012). Advocating a componential appraisal model to guide emotion recognition. *International Journal of Synthetic Emotions, 3,* 18–32. http://dx.doi.org/10.4018/jse.2012010102

Pope, L. K., & Smith, C. A. (1994). On the distinct meanings of smiles and frowns. *Cognition and Emotion, 8,* 65–72. http://dx.doi.org/10.1080/02699939408408929

Psychology Software Tools, Inc. (2012). *E-prime 2.0* [software]. Retrieved from http://www.pstnet.com

Roesch, E., Tamarit, L., Reveret, L., Grandjean, D., Sander, D., & Scherer, K. R. (2011). FACSGen: A tool to synthesize emotional facial expressions through systematic manipulation of facial action units. *Journal of Nonverbal Behavior, 35,* 1–16. http://dx.doi.org/10.1007/s10919-010-0095-9

Russell, J. A. (1997). Reading emotion from and into faces: Resurrecting a dimensional-contextual perspective. In J. A. Russell & J. Fernández-Dols (Eds.), *The psychology of facial expression* (pp. 295–320). Cambridge, UK: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511659911.015

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. (1986). Vocal affect expression: A review and a model for future research. *Psychological Bulletin, 99,* 143–165. http://dx.doi.org/10.1037/0033-2909.99.2.143

Scherer, K. R. (1992). What does facial expression express? In K. Strongman (Ed.), *International review of studies on emotion* (Vol. 2, pp. 139–165). Chichester, England: Wiley.

Scherer, K. R. (2001). Appraisal considered as a process of multi-level sequential checking. In K. R. Scherer, A. Schorr, & T. Johnstone (Eds.), *Appraisal processes in emotion: Theory, methods, research* (pp. 92–120). New York, NY: Oxford University Press.

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

Scherer, K. R. (2013a). Emotion in action, interaction, music, and speech. In M. A. Arbib (Ed.), *Language, music, and the brain: A mysterious relationship* (pp. 107–140). Cambridge, MA: MIT Press. http://dx.doi.org/10.7551/mitpress/9780262018104.003.0005

Scherer, K. R. (2013b). The nature and dynamics of relevance and valence appraisals: Theoretical advances and recent evidence. *Emotion Review, 5,* 150–162. http://dx.doi.org/10.1177/1754073912468166

Scherer, K. R., & Bänziger, T. (2010). On the use of actor portrayals in research on emotional expression. In K. R. Scherer, T. Bänziger, & E. B. Roesch (Eds.), *Blueprint for affective computing: A sourcebook* (pp. 166–178). Oxford, UK: Oxford University Press.

Scherer, K. R., & Ceschi, G. (1997). Lost luggage emotion: A field study of emotion-antecedent appraisal. *Motivation and Emotion, 21,* 211–235. http://dx.doi.org/10.1023/A:1024498629430

Scherer, K. R., Clark-Polner, E., & Mortillaro, M. (2011). In the eye of the beholder? Universality and cultural specificity in the expression and perception of emotion. *International Journal of Psychology, 46,* 401–435. http://dx.doi.org/10.1080/00207594.2011.626049

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

Scherer, K. R., & Grandjean, D. (2008). Facial expressions allow inference of both emotions and their components. *Cognition and Emotion, 22,* 789–801. http://dx.doi.org/10.1080/02699930701516791

Scherer, K. R., & Meuleman, B. (2013). Human emotion experiences can be predicted on theoretical grounds: Evidence from verbal labeling. *PLoS ONE, 8,* e58166. http://dx.doi.org/10.1371/journal.pone.0058166

Scherer, K. R., Mortillaro, M., & Mehu, M. (2013). Understanding the mechanisms underlying the production of facial expression of emotion: A componential perspective. *Emotion Review, 5,* 47–53. http://dx.doi.org/10.1177/1754073912451504

Scherer, K. R., Mortillaro, M., & Mehu, M. (2017). Facial expression is driven by appraisal and generates appraisal inference. In J.-M. Fernández-Dols & J. A. Russell (Eds.), *The science of facial expression* (pp. 353–373). New York, NY: Oxford University Press.

Scherer, K. R., Schorr, A., & Johnstone, T. (2001). *Appraisal processes in emotion: Theory, methods, research*. New York, NY: Oxford University Press.

Schlegel, K., & Scherer, K. R. (2016). Introducing a short version of the Geneva Emotion Recognition Test (GERT-S): Psychometric properties and construct validation. *Behavior Research Methods, 48,* 1383–1392. http://dx.doi.org/10.3758/s13428-015-0646-4

Sergi, I., Fiorentini, C., Trznadel, S., & Scherer, K. R. (2016). Appraisal inference from synthetic facial expressions. *International Journal of Synthetic Emotions, 7,* 45–63. http://dx.doi.org/10.4018/IJSE.2016070103

Shuman, V., Clark-Polner, E., Meuleman, B., Sander, D., & Scherer, K. R. (2017). Emotion perception from a componential perspective. *Cognition and Emotion, 31,* 47–56. http://dx.doi.org/10.1080/02699931.2015.1075964

Smith, C. A. (1989). Dimensions of appraisal and physiological response in emotion. *Journal of Personality and Social Psychology, 56,* 339–353. http://dx.doi.org/10.1037/0022-3514.56.3.339

Tanaka, J. W., Kaiser, M. D., Butler, S., & Le Grand, R. (2012). Mixed emotions: Holistic and analytic perception of facial expressions. *Cognition and Emotion, 26,* 961–977. http://dx.doi.org/10.1080/02699931.2011.630933

van Peer, J. M., Grandjean, D., & Scherer, K. R. (2014). Sequential unfolding of appraisals: EEG evidence for the interaction of novelty and pleasantness. *Emotion, 14,* 51–63. http://dx.doi.org/10.1037/a0034566

van Peer, J. M., Grandjean, D., & Scherer, K. R. (2016). Novelty and pleasantness appraisals interact: Evidence from electromyographic measurement of facial expressions. Manuscript in preparation.

Wagner, H. L. (1993). On measuring performance in category judgment studies on nonverbal behavior. *Journal of Nonverbal Behavior, 17,* 3–28. http://dx.doi.org/10.1007/BF00987006

Wallbott, H. G. (1988). In and out of context: Influences of facial expression and context information on emotion attributions. *British Journal of Social Psychology, 27,* 357–369. http://dx.doi.org/10.1111/j.2044-8309.1988.tb00837.x

Wehrle, T., Kaiser, S., Schmidt, S., & Scherer, K. R. (2000). Studying the dynamics of emotional expression using synthesized facial muscle movements. *Journal of Personality and Social Psychology, 78,* 105–119. http://dx.doi.org/10.1037/0022-3514.78.1.105

Wingenbach, T. S. H., Ashwin, C., & Brosnan, M. (2016). Validation of the Amsterdam Dynamic Facial Expression Set—Bath Intensity Variations (ADFES-BIV): A set of videos expressing low, intermediate, and high intensity emotions. *PLoS ONE, 11,* e0147112. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0147112

Appendix

Labels and Visual Illustrations of the Major AUs Investigated (Modified From Mortillaro et al., 2011)

AU	Action name	Illustration
AU1	Inner brow raiser	
AU2	Outer brow raiser	
AU4	Brow lowerer	
AU5	Upper lid raiser	
AU6	Cheek raiser	
AU7	Lid tightener	
AU9	Nose wrinkler	
AU10	Upper lip raiser	
AU12	Lip corner puller	
AU15	Lip corner depressor	
AU16	Lower lip depressor	
AU17	Chin raiser	
AU18	Lip puckerer	
AU20	Lip stretcher	
AU22	Lip funneler	
AU23	Lip tightener	
AU24	Lip pressor	
AU25	Lips part	
AU26	Jaw drops	
AU28	Lips suck	
AU43	Eye closure	

See the online article for the color version of this figure.

Received July 25, 2017
Revision received October 6, 2017
Accepted October 7, 2017 ■