How Expression, Context and Perspective Determine Judgments of Emotion

Bin Han, *Student Member, IEEE* Jessie Hoegen, *Student Member, IEEE,* Su Lei, *Member, IEEE* Gale Lucas, *Member, IEEE* Danielle Shore, *Member, IEEE* Brian Parkinson, *Member, IEEE* Jonathan Gratch, *Senior Member, IEEE*

Abstract—Within the field of Affective Computing, facial expressions were historically assigned emotion labels by annotators without knowledge of the context in which the expressions were produced. Often these 3rd-person impressions were used instead of 1st-person judgments about the target's feelings. The field now appreciates that 1st- and 3rd-person judgments can differ dramatically and that context plays a central role in explaining these differences. More recently, there is growing appreciation of the importance of a 2nd-person perspective. When people are engaged in a social task, impressions of their partner's emotions differ from the impressions of detached bystanders. In this paper, we explore how facial expressions, contexts and perspective (1st, 2nd-, and 3rd-person) interact to determine judgments of emotion. We explore this using automatic facial expression analysis of natural expressions produced in the context of a social task (the prisoner's dilemma). Our findings suggest that expression and context combine to determine judgments, but the way they combine differs depending on perspective. We discuss the implication of these findings for automatic emotion recognition methods and the inferences such methods can support.

Index Terms—context, facial expression, perspective, valence, prisoner's dilemma

1 Introduction

Affective computing approaches emotion recognition from multiple perspectives. Some algorithms adopt a 1st-person perspective by training on data where people self-label their emotions. By examining facial, vocal, or postural cues, these algorithms aim to predict subjective feelings such as pain [1], depression [2], or product-related happiness [3]. Other approaches are trained from a 3rd-person perspective. For example, annotators are asked to label what they believe someone is feeling from prerecorded expressions. Predicting these 3rd-person labels could be useful for understanding stereotypes [4] or training people to produce easily recognizable expressions [5]. Finally, some work distinguishes 2nd- from 3rd-person judgments [6], [7], [8], emphasizing that perceptions formed during social tasks may differ from those of detached observers [9]. Thus, predicting 2nd-person labels may be particularly useful for understanding how interactions unfold [10], [11].

There is growing appreciation that the same expression may be labeled quite differently depending on the perspective of the annotator, which creates problems if labels formed from one perspective are used to predict a different perspective [12], [13]. This is especially problematic when 3rd-person annotators have limited access to the context around how the expression arose [14], but even when context is provided, strong differences remain, particularly when expressions arise from a social task. 1st-person annotators assess their feelings directly through introspection, but observers must infer them from expressions and context [15], [16], [17]. Thus, context and expression might interact in complex ways to determine the observer's label (e.g., perhaps a smile conveys pleasure in a positive situation but displeasure in a negative one [18]). Less research has compared 2nd- and 3rd-person labels, but psychological findings suggest they differ in important ways. For example, someone else's anger feels more intense and evokes greater physiological activation when it is directed at you, rather than being observed from a detached perspective [19], [20]. Given their direct involvement in the social situation, 2nd-person annotators may also be more motivated to attend to facial and contextual cues of their partner than a detached bystander (in that correct inferences could benefit future interaction). Indeed, 3rd-party observers often apply simple normative rules ("that action is unfair") whereas actors in the situation put more emphasis on the presumed intentions behind the action [21]. This suggests that 3rd-person labels may be more influenced by the observable context, whereas 2nd-person may exhibit more nuanced judgments. Table 1 summarizes these different processes and the suitable uses for algorithms trained on different perspectives.

We examine how perspective and context shape annotation labels in a social task. Although the impact of context and perspective has been examined in psychological and affective computing research, several limitations motivate the present study. First, psychological research on context effects has relied on stylized stimuli such as artificially generated facial expressions [15], [16], [30]. Here, we examine spontaneous facial expressions generated during a social task. Second, while affective computing approaches increasingly acknowledge differences between 1st- and 3rd-person labels, rarely have all three perspectives been considered in a single analysis, and multi-perspective corpora

- *Institute for Creative Technologies, University of Southern California, Los Angeles, CA 90094 USA. E-mail: bhan, jhoegen, slei, lucas, gratch @ict.usc.edu*
- *Department of Experimental Psychology, University of Oxford, Oxford, UK. E-mail: danielle.shore, brian.parkinson@psy.ox.ac.uk*

Manuscript received May 14, 2024;

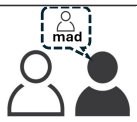
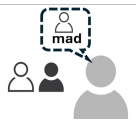
Perspective		Labeling processes	Example applications
1st-person		Labeled through introspection Expression directed away from self High involvement in situation	Predict sender reported feelings [22] Predict why feeling occurs [23] Predict sender actions [3], [24]
2nd-person		Labeled via inference from cues and context; Qualified by presumed intentions Expression directed as self (more intense?) High involvement in situation	Predict impact on receiver feelings [25] Predict receiver inferences [26] Predict receiver actions [27]
3rd-person		Labeled via inference from cues and context; Less regard to intentions Expression directed at 3rd party No involvement in situation	Predict bystander bias [4], [28] Predict bystander actions [29]

TABLE 1: Three different perspectives on emotion recognition

often involve few participants [31]. Third, prior work on context-based recognition has emphasized how facial expression perception is shaped by body language [32] or the background scene [33], while we complement this work by focusing on how social actions shape emotion perception. Specifically, we study emotion perceptions in the iterated prisoner's dilemma (IPD) and how player actions in the dilemma create context that shapes perception of emotional expressions. IPD is commonly used to investigate the impact of emotion in social contexts [34]. Prior IPD studies find participants are highly expressive, their expressions contain information about what just occurred [35], partner perceptions change according to the context [11], and people are poor at predicting their partner's emotions [36]. Here we extend our previous work on differences between 1st- and 2nd-person emotion perceptions [37] by contrasting further with 3rd-person judgments.

First, we consider two competing hypotheses for how facial expressions and social context shape emotion labels.

H1: Expression and context act as independent predictors of emotion. Support for H1 would be indicated by main effects only—for example, smiles predicting positive valence across all contexts, with context simply shifting ratings upward or downward.

H2: Expression and context interact, such that the same expression is interpreted differently depending on the context. Support for H2 would be indicated by significant expression × context interactions (e.g., a smile conveying pleasure in mutual cooperation but being ignored when the expresser is exploited).

If H1 holds, this simplifies the development of automated emotion recognition methods. It would mean that if smiles predict positive emotion, this relationship holds across contexts, and context simply adds or subtracts from this main effect. Existing "context-free" algorithms could easily be adjusted via late fusion. If H2 holds, it might require early fusion of expressive and contextual cues, (though see [16] for an alternative view).

Second, we consider whether these processes vary across perspectives.

H3: Expression–context interactions are more likely for observer judgments (2nd- and 3rd-person) than for self-reports (1st-person). Support for H3 would be indicated by stronger interactions in the observer conditions.

We test these hypotheses across three studies, examining how three factors – facial expressions (coded as six facial factors), context (in terms of IPD outcomes), and perspective – influence emotion valence ratings. We evaluate H1–H3 using lens model analyses (described below). Note that Study 1 originally appeared in a conference [37].

2 Related Work

Recognition Perspectives. Emotion recognition methods have long adopted a range of perspectives in assigning labels but the difference between these perspectives has sometimes been overlooked, leading to applications that have attempted to recognize 1st-person emotion using 3rd-person labels. This confusion was perhaps encouraged by Ekman's promotion of Basic Emotion theory which argued that expressions were universal signifiers of underlying emotional state [38]. Recent work has been more careful to distinguish 1st- and 3rd-person labels (the latter often referred to as "perceived emotion recognition"). Common approaches are to document discrepancies between 1st- and 3rd-person labels [13], to minimize these discrepancies [39], or at minimum, to clearly acknowledge the perspective used when labeling.

The 2nd-person perspective has recently gained traction within social neuroscience though it remains understudied in affective computing (perhaps due to the comparative difficulty of obtaining 2nd-person labels at a scale that is comparable to 3rd-person labels). Schilbach et al. [40] examined the neural distinctions between interacting with another person and merely observing them. They found that active engagement significantly alters neural activation patterns, emphasizing the unique influence of the 2nd-person perspective. Building on this foundational research, the same authors' more recent study [7] proposed that this direct engagement perspective involves unique emotional dynamics and reciprocity, thereby influencing the neural underpinnings of social cognition in ways not captured by third-person observations.

Most affective computing approaches and datasets tend to focus on a single perspective, though that is beginning to change. For example, the K-EmoCon dataset includes multimodal signals and labels from all three perspectives, though the number of participants is quite small [31]. Our

article advances this line of research by exploring these three emotion perspectives within the same social task.

Context in Emotion Recognition. Recent studies highlight the significant impact of context on emotion perception. Visual background [41], vocal expressions [42], body language [32], other faces [43], and linguistic cues [42], [44] strongly shape how emotions are interpreted from facial expressions. These findings challenge the assumption that emotional states are clearly recognizable from facial expressions, suggesting a complex interplay with contextual information [12]. Our article contributes to this body of work on the impact of context in interpreting facial cues.

There have been several attempts to investigate how contextual information modulates the perception of facial emotions using various notions of context. Carroll and Russell [45] paired facial expressions prototypical of basic emotions with emotional narratives and found that the tone of the narrative interacted with the face to shape perceived emotion. Righart and Gelder [46] reported a similar effect using background imagery rather than a text narrative to manipulate context. More recently, Ong et al. [16] showed these context effects extend to task-oriented behavior. For example, when interpreting the emotions of a player in a game, third-party observers attend to both game outcomes and facial expression to infer the player's feelings.

These effects are consistent with the *"Kuleshov Effect"* wherein movie viewers' perceptions of a character's emotional expression change based on the contextual scene that precedes it. Motivated by the Kuleshov effect, Mobbs et al. [47] explored how even neutral faces can be perceived as emotional depending on the context and began to unpack the neurological basis of the effect providing insights into the brain mechanisms underlying the integration of facial and contextual information.

Context-Aware Emotion Recognition Methods. Recent research accounts for context effects using a range of definitions of context. Recognition accuracy improves by incorporating body posture or gestures [41] or detecting objects in the background of the image [48], [49]. For example, Zhang et al. [50] employed a Contrastive Language-Image Pretraining(CLIP)-based architecture using activity descriptions as contextual information. These studies highlight the critical role of context in emotion recognition and suggest that the sources and methods for deriving context may vary across different research fields and tasks.

Social Events as Context. While most of the above-mentioned studies involve 3rd-party impressions of a single individual paired with some context, 2nd-person emotions involve expressions that arise within an interdependent social task. Here, aspects of the task itself might naturally be seen as determining the context. For example, Robert Frank explored how expressions in social games like the prisoner's dilemma might shape behavior [34], In such a game, each player's decision forms a context for interpreting their expressions. In a similar vein, Hazarika et al. [51] explored emotion recognition within task-oriented dialog, utilizing the historical context of previous speech acts and their emotional content to predict the emotional state of a given utterance. Even subtle features such as the orientation of gaze might shape interpretation in a social task. For example, Mumenthaler et al. [43] showed how head/gaze movements of one avatar influence third-person perceptions of emotions displayed by another.

3 Prisoner's Dilemma

We explore the role of perspective and context in the iterated prisoner's dilemma (IPD), a canonical task for studying facial expressions and their influence on social decision-making [34]. Like many social tasks, the prisoner's dilemma involves a mixture of cooperation and competition and players have been shown to attend to each other's emotional expressions, form inferences from these expressions about their partner's goals or character, and use these inferences to inform subsequent intentions to cooperate or compete [30], [52]. Both studies reported below use the same procedures, task structure, and software framework for administering the game and collecting facial responses. We describe this framework in some detail before proceeding to the studies.

The iterated prisoner's dilemma was administered in a laboratory setting. All participants provided informed consent and the study was approved by the university's research ethics committee. Participants received a base fee to perform the task, but to help provoke emotional responses, participants played the game for lottery tickets that were entered into a prize draw which could yield a substantial additional bonus. Participants who earned a higher score (for example, by successfully collaborating or by stealing more tickets from their partner) had a higher chance of winning this additional bonus.

Participants were randomly paired, situated in separate rooms, and participated via the software framework. Participants first received instructions on the task and potential for bonus, played the game, then completed post surveys. The game interface was displayed on the screen, alongside a live video of their partner recorded by a webcam. The participants could see each other while playing the game but were unable to communicate by speech, as audio was not transmitted by the framework. To help make the rules intuitive to participants, it was based on the British TV show *Golden Balls*. In each round of the 10-round game, players were given ten lottery tickets and had to decide whether to *split* (corresponding to the cooperate decision in the prisoner's dilemma) or *steal* (corresponding to the defect decision). The consequences of splitting or stealing depended on the other player's decision. Figure 1 shows the game interface and Table 2 shows the payoff matrix for the game. After both participants made a decision, the joint decision would be displayed to both participants (e.g., one participant might have chosen to split, the other to steal) following a short animation. The joint decisions were logged in a database for further analysis. This sequence of events repeated until 10 rounds were completed.

		Participant B	
		cooperate (C)	*defect (D)*
Participant A	*cooperate (C)*	A = 5, B = 5	A = 0, B = 10
	defect (D)	A = 10, B = 0	A = 1, B = 1

TABLE 2: Payoff matrix used in the game

Given the structure of the experiment, the primary way players interact is through their choice of actions over 10

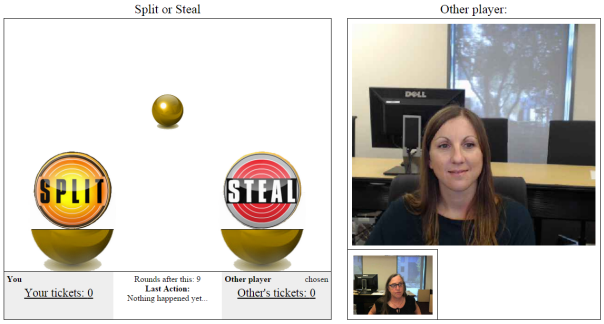


Fig. 1: The game interface. Participants make choices by selecting either "split" or "steal" in each round. They see the number of tickets won by themselves and by the other player as well as the other player's webcam video (top right) and their own webcam video (bottom right).

rounds and their emotional expressions. Prior work shows that these emotional expressions are typically associated with player actions. For example, Lei and Gratch found that players were most expressive immediately after the joint decision of a round was revealed [35]. Thus, it is reasonable to assume that this joint outcome provides a context for interpreting the associated facial expressions (see also [11], [16], [52]). Therefore, we operationalize *context* as the joint decision that co-occurs with each facial expression being rated. As each participant has the choice to either cooperate or defect during the prisoner's dilemma, there are a total of 4 possible contextual situations associated with facial expressions: The participants both cooperate (designated as 'CC'), both defect ('DD'), a participant is exploited by their partner ('CD') or a participant exploits their partner ('DC').

4 Study 1: 1st and 2nd Person Perspective

To contrast 1st and 2nd person judgments of emotion, we adopt a lens model analysis [53]. Brunswik's lens model has been widely used to study how people infer psychological states like pain [54] or personality [55] from observable behavior, and how these inferences differ by perspective (e.g., the behaviors that predict self-reported pain might differ from the features observers use to infer pain). The model assumes that perceivers rely on available cues in a situation (e.g., facial tension or body posture) to infer psychological states. Accuracy is assessed by the degree of correspondence between perceivers' judgments of targets' states and the targets' own reports of those states. The relation between observable cues and the self-reported state is termed cue validity, whereas the degree to which perceivers' judgments are shaped by those cues is termed cue utilization. The familiar lens-shaped diagram provides a concise way to visualize these correspondences, but the model also includes a mathematical decomposition of such judgments, making it a powerful tool for analyzing perception, decision-making, and bias.

4.1 Emotion Perceptions

We utilize data from the Oxford Split-Steal corpus, which includes 100 participants (61 females and 29 males, mean age = 26.42, SD = 7.44) playing a 10-round IPD and who could earn money based on how many points they eared in the game [36]. Due to some recording issues, 907 videos were available for analysis.

1st- and 2nd-person emotion impressions were gathered using video-cued recall. After the 10 IPD rounds, participants saw 5-second clips of webcam footage from the game they had just played. These clips corresponded to the 5 seconds immediately following the moment their joint decision was revealed. Prior research suggests people are mostly facially expressive during this window [35]. Participants viewed clips of both themselves and their partner from each round in chronological order.

4.2 Measures

4.2.1 Emotion Perceptions

For each 5-second clip presented during the video-cued recall procedure, participants provided 1st-person and 2nd-person ratings of emotional valence. Participants were first presented with the 10 clips showing their own facial reactions and asked to attend to their own reactions and self-report their emotions on a continuous valence scale running from -50 (negative valence) to +50 (positive valence) for each clip. They were next presented with the 10 clips showing their partner's facial reactions and asked to rate perceived emotion using the same -50 to +50 scale ("How positive or negative was your (or your partner's) emotion?") on a continuous scale from –50 (very negative) to +50 (very positive). Participants were also asked to rate the extent to which they regulated their expressions and their perceptions of their partner's regulation for each clip, but responses to these questions are not used in the present analysis or discussed further in this paper – see [36] for details.

4.2.2 Automatic Facial Expression Annotations

Using the timestamps stored with each joint decision, the 5-second clips of joint outcome events used during the video-cued recall procedure were stored on a server. We obtained facial expression ratings for each clip automatically using commercial software originally based on CERT [56]. This system extracts Action Units (AU) as defined by the Facial Action Coding System (FACS) from each frame of the participant's video. The AU values are reported as 'evidence' values, representing the likelihood of the AU being active in a particular frame.

AUs afford a rich representation of facial movements, allowing for a nuanced analysis of expression, but the large number of units leads to a sparse representation. Further, many AUs tend to co-occur in common patterns (e.g., smiles often involve AU12 and AU6). Therefore, it is common to reduce the full list of AUs into a smaller number of orthogonal dimensions. This approach has been applied to analyze emotional styles [57], expressions used to deliver good or bad news [58], and the prisoner's dilemma [59].

Following Stratou and colleagues [59], we represent facial expressions in terms of six orthogonal facial factors that combine commonly co-occurring AUs (illustrated in Figure 2). Factor one (F1) corresponds to a 'smile' consisting of AU6, AU7 and AU12, F2 corresponds to 'eyebrows-up'

Fig. 2: Facial factors. Enjoyment smile (F1), eyebrows up (F2), open mouth (F3), mouth tightening (F4), eye tightening (F5), and mouth frown (F6).

consisting of AU1 and AU2, F3 corresponds to an 'open-mouth' consisting of AU20, AU25 and AU26, F4 corresponds to 'mouth-tightening' consisting of AU14, AU17 and AU23, F5 corresponds to 'eye-tightening' consisting of AU4, AU7 and AU9 and F6 corresponds to 'mouth-frown' and consists of AU10, AU15 and AU17.

4.3 Study 1 Analysis

To assess H1 (i.e., expression and context independently shape labels), we correlate facial factors and valence scores, then examine if this correlation changes as a function of context (e.g., do smiles predict valence differently for mutual cooperation compared with mutual defection). Since we had both 1st-person and 2nd-person valence ratings, we examine this relationship separately for (1st-person and 2nd-person ratings.

We use moderated multiple linear regression to investigate our second hypothesis H2 (i.e., expression and context interact). We created regression models to predict the expressed and perceived ratings of valence, in order to find the relation between expression and context. Expressions were represented by the average value for a reveal event of the six previously defined factors that were automatically extracted from the videos following the reveal event.

We used dummy coding to encode context using the variables 'decision self' and 'decision partner.' These decisions were coded differently based on the joint outcome: For joint cooperation (CC) 'decision self' and 'decision partner' were coded 0 for cooperating and 1 for defecting. For joint defect (DD) they were given the reverse coding, so that 0 represented defecting and 1 represented cooperating. For exploited by other (CD) 'decision self' was coded as 0 for cooperating and 1 for defecting, while 'decision partner' was coded as 0 for defecting and 1 for cooperating. Finally, for exploiting other (DC) 'decision self' was coded as 0 for defecting and 1 for cooperating and 'decision partner' was coded as 0 for cooperating and 1 for defecting.

In order to model the relation between expressions and each player's choice in the round, the decision variables (i.e., 'decision self' and 'decision partner') were used as two moderating variables in the regression model predicting the dependent variable of either expressed or perceived valence. Thus, the regression equation was:

$$Y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4(x_1 \times x_2) + b_5(x_1 \times x_3) + b_6(x_2 \times x_3) + b_7(x_1 \times x_2 \times x_3) + \varepsilon \quad (1)$$

where x variables are the values of the independent variables, x_1 is the facial factor, x_2 is 'decision self' and x_3 is 'decision other'. Y is the value of the dependent variable (self-reported or perceived valence) and ε is the error term.

4.4 Study 1 Results

Overall, results suggest that facial cues and context interact to determine 2nd-person perceptions of emotion (supporting H2 and rejecting H1). By contrast, face and context were independent sources of information for 1st-person perceptions (supporting H1 and failing to support H2). Together, these two findings provide support for H3 (that face and context interact to determine observer inferences)

The results are summarized with the lens model in Figure 3. The left side of the diagram (referred to as *cue validity*) shows the relationship between some latent variable (in our case, emotion) and observable perceptual cues. The right side of the diagram (referred to as *cue utilization* illustrates how a perceiver uses cues to reconstruct the latent variable (in our case, the observer's perception of their partner's emotion). Each link in the diagram shows the correlation with the latent construct (i.e., self-reported emotion on the left and perceived emotion on the right). Symmetry between the two sides of the "lens" implies that the perceiver correctly utilizes the valid cues. Asymmetry indicates perceptual errors and helps identify misconceptions.

In our case, we are particularly interested in how context (i.e., joint outcomes of the game) interacts with facial cues. In Figure 3, solid lines indicate that context is an independent predictor. Dashed lines indicate that the cue and context interact. For example, the dashed line between smile and perceived emotion indicates that the way the perceiver uses the smile-cue changes based on the joint outcome of the game. The top connecting line shows the correlation between self-reported and perceived judgments (r(907) = 0.397, p $<$ .001). We first discuss the significant main effects of face and context separately before discussing the interaction between these two factors.

4.4.1 Impact of the face alone

We examined facial cue validity (i.e., 1st-person perceptions) while ignoring context by calculating correlations between each facial factor and self-reported valence. We found significant correlations for the smile (r(907) = .147, p $<$ .001) and for the open mouth (r(907) = .148, p $<$ .001) factors. This indicates that smiles and open-mouth are valid cues of self-reported emotion, and each cue shows a positive relationship (i.e., ignoring context, more smiles mean more positive self-reported valence).

We examined 2nd-person perceptions, ignoring context, by correlating each facial factor with 2nd-person perceptions that the partner felt positive or negative (i.e., perceived valence). We found significant correlations for the smile (r(907) = 0.208, p $<$ .001), the open-mouth (r(907) = .145, p $<$ .001), the mouth-tightening (r(907) = .072, p = .30) and the frown (r(907) = .069, p = .038) factors. This shows that observers utilize the same two valid cues as the players that they are observing (smiles and open mouth), but also attend

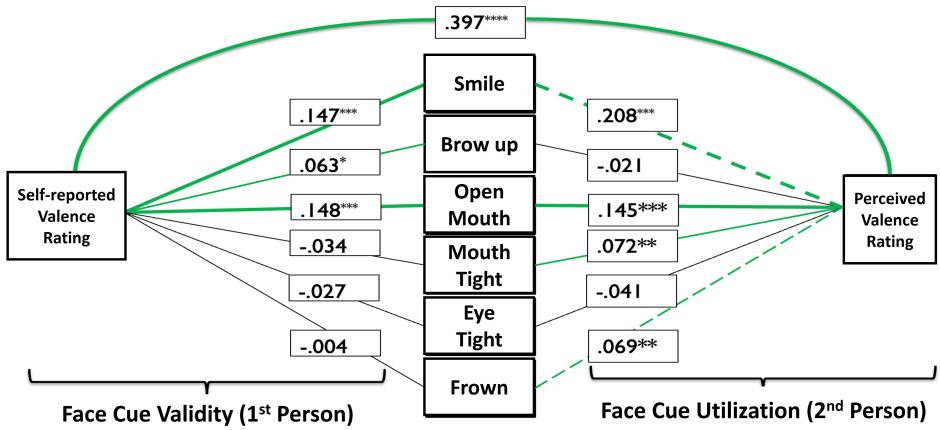


Fig. 3: Brunswik's Lens model of correlations between facial factors and participant valence ratings. Green lines show significant correlations. Dashed lines show factors with significant expression–context interactions.

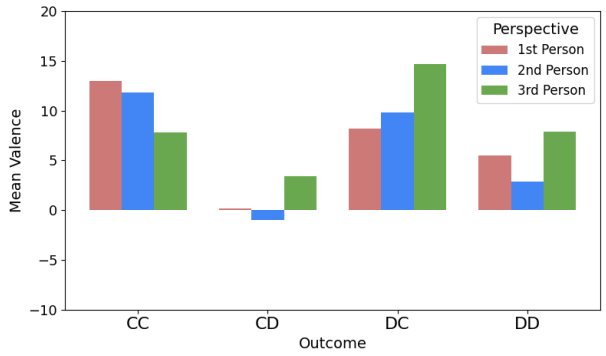


Fig. 4: The effect of context alone on valence ratings. Here we also include results for Study 2 for comparison. Vertical bars are 95% confidence interval of the mean.

to two invalid cues (mouth tightening and frown). Thus, ignoring the role of context, observers pay closest attention to the valid cues of smiling and open mouth, but also attend to irrelevant cues and thus could misinterpret their partner's self-reported feelings.

4.4.2 *Impact of the context alone*

We examined the impact of context while ignoring facial cues by examining the average valence rating for each of the four game outcomes (see Figure 4). This shows context has a similar overall effect on both self-reported and perceived emotion. Regardless of whether participants are reporting their own feelings or estimating their partner's feelings, valence is reported as highest following mutual cooperation (CC). As expected from the structure of the payoff matrix, valence is reported as negative when a player has been exploited (DC) and positive when a player successfully exploits their partner. Interestingly, valence is reported as somewhat positive following mutual defection, but this might be explained as a response to finding out that the other player had a similar intention to you. One might expect the strongest positive emotions when exploiting one's partner (as this has the greatest payout), but this was not the case. This may reflect the fact that satisfaction was tinged with feelings of guilt about exploiting one's partner [60].

4.4.3 *How do face and context combine?*

In order to contrast H1 (the face and context are independent predictors) with H2 (the face and context interact), we used moderated multiple linear regressions as described in Section 4.3. We used moderated regressions for each of the factors based on both display and rated valence, looking at the interaction between expression and context, i.e., a total of 6 models for self-reported emotion and 6 models for perceived emotion.

For 1st-person emotion, we failed to find support for Hypothesis 2. None of the facial factors showed a significant interaction with context. Thus, for self-reported emotion, context and facial cues seem to offer independent sources of information for determining emotion (H1).

In contrast, we find clear support for Hypothesis 2 for 2nd-person perceptions of emotion. Smiles and frowns showed clear evidence of an interaction. Smiles (F1) signal happiness when the outcome benefited the player that was smiling (CC or DC), but smiles were largely ignored when the perceiver knew that the outcome caused the other player harm (indeed, the more someone smiled when exploited had a negative association with perceived valence). Specifically, the moderated regression showed how one's own decision (C or D) and partner's decision (C or D; together representing CC CD DC DD) moderated the association between smile (F1) and perceived valence.

The interaction term between own decision, partner's decision, and Smile (F1) was significant, B = 0.308, t = 2.103, p = .036. To break down the interaction, we conducted simple slope analysis. This revealed a significant effect of factor 1 on the perceived expression following mutual cooperation (CC: BB = 0.395, t = 9.550, $p < .001$), and when the player exploited their partner (DC: BB = 0.229, t = 2.932, p = .003), but not for other outcomes (DD: BB = 0.119, t = 1.430, p = .153, CD: BB = -0.023, t = -0.277, p = .781).

Frowns (F6) had little information value *except* when the person frowning was exploited. In this case, contrary to folk-wisdom, frowns were interpreted as a positive signal. Specifically, we found a significant interaction in the

moderated regression examining how own decision (C or D) and partner's decision (C or D; together representing CC CD DC DD) moderated the association between frown (F6) and perceived valence. The interaction term between own decision, partner's decision and frown reached significance, B = 0.344, t = 2.281, p = .023. To break down the interaction, we conducted a simple slope analysis. This showed that there was a significant effect of frown on the perceived expression when the player was exploited (CD: BB = 0.250, t = 2.872, p = .004), but not for the other outcomes (CC: BB = 0.026, t = 0.628, p = .530; DD: BB = -0.026, t = -0.317, p = .751; DC: BB = 0.095, t = 1.157, p = .248).

Figure 5 shows the regression coefficients (B) for the moderated regressions on smile and frown. In summary, expressions and context interact to determine second-person perceptions of emotion. Smiles indicate pleasure unless the context suggest pain. Frowns are largely ignored unless the context suggests the expressor was exploited.

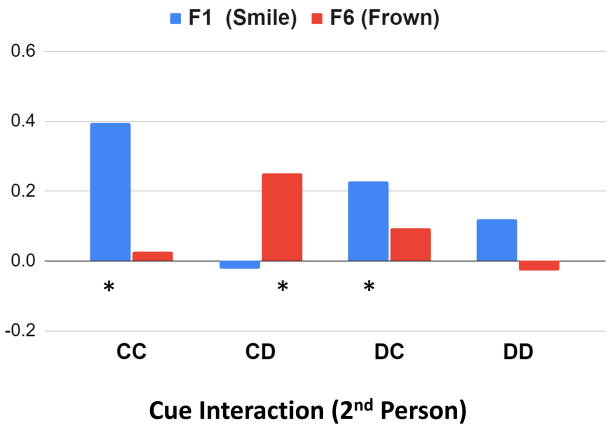


Fig. 5: The regression coefficients for smile and frown based on game state for 2nd-person perceived valence.

5 Study 2: 1st 2nd 3rd Person Perspective

Study 1 provides support that observer judgments involve complex interactions between facial and situational cues in IPD but only compared 1st- and 2nd-person perspectives. To allow a direct comparison of all three perspectives, we collected 3rd-person emotion annotations using the Prolific platform.

To mimic Study 1, annotators received an explanation of the Split or Steal game and used the same measures as 2nd-person annotators. They were shown outcomes and facial expressions for 10 trials in sequence. After each video, they answered the question: "How positive or negative was the emotion that Player A was expressing?" using a valence scale ranging from -50 (very negative) to +50 (very positive). Each annotation task took approximately 15 minutes to complete, and compensation was provided at a range of $12 USD according to Prolific's guidelines. To ensure data quality, we included an attention check during the task. We collected responses from 10 annotators per video, and used the average valence score across annotators as the final label for each video (following common practice in emotion recognition corpora). A subset of 40 games (400 videos) was annotated. We first examine how perspective influences valence scores and how these are shaped by context. We next replicate the same interaction analysis from Study 1 to assess if facial cues and context interact.

5.1 Study 2 Results

5.1.1 Impact of perspective and context on valence

To test our hypotheses, we examined expression and context effects and their interaction, with H3 predicting stronger effects for observers. We performed a 3 (perspective: 1st, 2nd, 3rd) × 4 (context: CC, CD, DC, DD) ANOVA to assess valence ratings. Results showed a main effect of context, $F(3,1167)=26.93$, $p < .001$, no main effect of perspective, $F(2,1167)=0.85$, $p=.428$, and a significant perspective × context interaction, $F(9,1062)=5.58$, $p< .001$, indicating that context was used differently depending on perspective.

We analyzed this interaction with four repeated-measures one-way ANOVAs, one per IPD outcome. The effect was driven by the 3rd-person annotators, specifically their labeling of mutual-cooperation (CC) and exploitation (DC) (Fig. 4). Perspective impacted CC ($p< .001$), with t-tests showing differences between 2nd- and 3rd-person ratings ($t = 4.47$, $p< .001$) and 1st- and 3rd-person ratings ($t = 5.40$, $p < .001$), but not 1st- and 2nd-person ($t = 1.25$, $p = 0.21$). A similar pattern emerged for exploiting one's partner (DC). The ANOVA was significant ($p = .008$), and post hoc paired-samples t-tests showed significant differences between 2nd- and 3rd-person ratings ($t = -2.52$, $p = .014$) and between 1st- and 3rd-person ratings ($t = -3.44$, $p < .001$), but not between 1st- and 2nd-person ($t = -0.50$, $p = 0.62$). Perspective did not have an effect in the CD ($p = .12$) or DD ($p = .25$) conditions.

Overall, 3rd-person annotators were more impacted by context. They rated participants as happier when they stole from their partner and less happy when they split the money (compared with 1st- and 2nd-person raters), perhaps equating happiness with the amount of money earned. This is consistent with prior research that 3rd-person annotators are more driven by abstract norms [21].

5.1.2 How do face and context combine

Following Study 1, we contrasted H1 and H2 on the 3rd-person ratings using moderated multiple linear regression. Context and facial expressions did not interact to determine 3rd-person annotations (supporting H1 and failing to support H2, and thus failing to find support for H3 with regard to our 3rd-person annotators.

Fig. 6 shows the correlations between facial factors and valence ratings. Expressions involving open-mouthed smiles were rated as more positive. In comparison with Study 1, smiles were much stronger predictors of valence ratings ($r=.509$) than for the 1st- ($r=.147$) and 2nd-person ($r=.208$) perspectives.

6 Study 3: 3rd Person Replication

To further explore the perceptions of 3rd-person observers, we examine archival data from the University of Southern California's Split-Steal dataset, which involved the identical task and data collection framework as used in Study 1 [35]. Unlike Study 1, participants were recruited from the Los

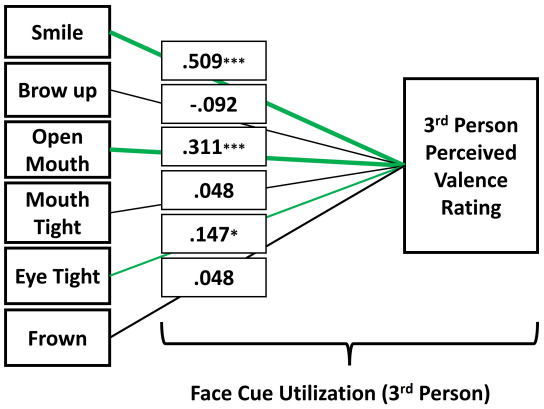


Fig. 6: Correlation between facial expression and 3rd-person perceived valence in Study 2.

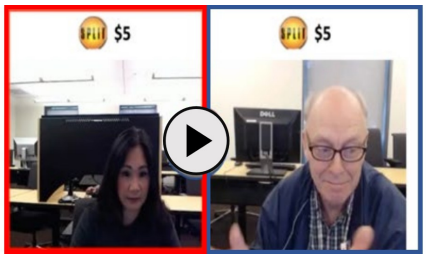


Fig. 7: Facial reactions were annotated with the joint game outcome (here, mutual cooperation).

Angeles area and did not engage in the video-cued recall process. Players were incentivized by playing for lottery tickets for several $100 USD lotteries. The experiment was otherwise identical. This dataset includes 3rd-person annotations for a subset of videos which we used for analysis.

6.1 Emotion Perceptions

3rd-person judgments of emotion were obtained via Amazon Mechanical Turk. To maximize understanding, annotators were first explained the game rules, then annotated a set of videos along with an explanation of the game outcome. A subset of highly expressive videos was selected to remove the influence of neutral expressions, and the number of videos was balanced across outcomes (whereas Study 1 consisted of a much greater proportion of mutual cooperation videos). Videos consisted of 25 of the most expressive for each possible outcome in the game, totaling 100 videos. Expressivity [35] was determined using the automatic labels provided with each video.

Annotators were told to focus on the emotional reactions of one player (highlighted by a red box) and asked to guess how they might feel. Figure 7 illustrates what the annotators saw, which highlights the game outcome (e.g., CC) to provide context. Annotators assessed Valence using a 5-point Likert scale. Each annotator evaluated ten videos selected at random, and we gathered 20 ratings per video and used the average of these 20 ratings to represent perceived emotion (compared with 10 ratings per video in Study 2).

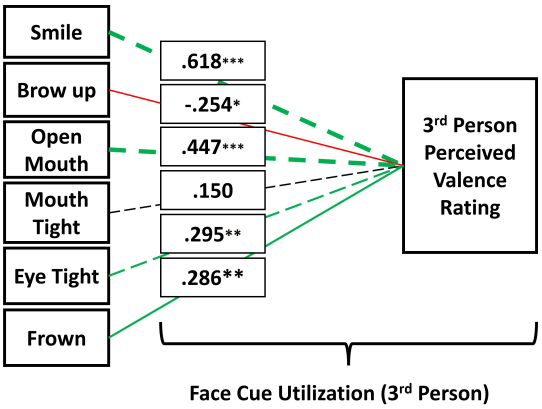


Fig. 8: Correlation between facial expression and 3rd-person perceived valence. Green lines show significant correlations. Dashed lines show factors with significant expression–context interactions.

6.2 STUDY 3: Results

6.2.1 *Impact of the face alone*

To test our hypotheses, we examined whether expressions predicted valence (H1), interacted with context (H2), and whether such interactions were stronger for observers (H3). We first examined how bystanders utilize facial cues, ignoring the role of context. Specifically, we calculated the correlation between the partner's facial expressions and 3rd-person perceived valence. The result is shown in Figure 8 We found significant correlations for the smile ($r(100) = 0.618$, $p < .001$), the open-mouth $r(98)) = -0.254$, $p < .011$), the mouth-tightening ($r(98) = 0.447$,$p < .001$) and the eye tight ($r(98) = 0.295$, $p = .003$) factors.

Compared with Study 1, 3rd-person judgments show a remarkably similar pattern to the second-person judgments: Smiles and open mouth serve as the best predictors of positive valence. Interestingly, frowns again are associated with positive valence (contrary to expectations from the decontextualized interpretation of frowns). The associations are much stronger than in Study 1, but this could reflect the averaging of multiple annotations, which removes individual variance that was present in the first study. However, as Study 2 used a different participant pool, there are potential issues with comparing the two datasets directly.

6.2.2 *Impact of the context alone*

We next examined the influence of context on 3rd-party judgments while disregarding the impact of facial cues. Specifically, we analyzed the average valence scores for each context category, as presented in Figure 9.

The analysis indicates that context plays a key role in how 3rd-person viewers perceive emotions, in line with the trends seen in self and 2nd-person perceptions outlined in Section 4.4.2. In the 3rd-person perspective, all views agree on the positive nature of mutual cooperation (CC). Statistical analysis using t-tests revealed significant differences between CC and CD, as well as between CC and DD ($p<.001$). There are also significant differences between CD and DC, and between DC and DD ($p<.001$)

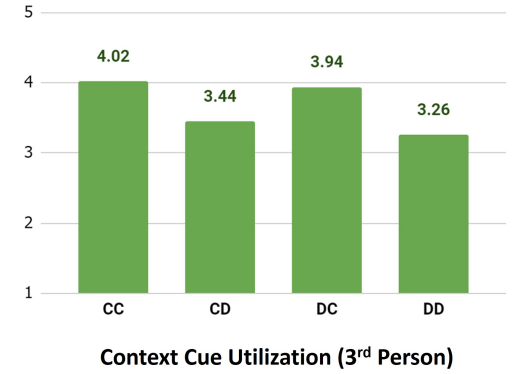


Fig. 9: Effect of context on valence ratings (in Study 3, valence rated on a 5-point Likert scale.

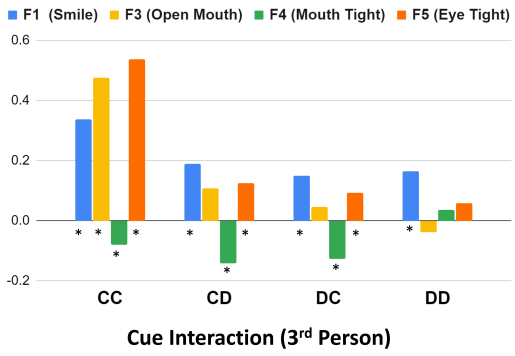


Fig. 10: Regression coefficients for facial factors based on game state for 3rd-person perceived valence.

6.2.3 How do face and context combine?

The 3rd-person perceived valence ratings were strongly impacted by context. We used the analysis from Study 1 to assess the effects of the facial factors [59] on 3rd-person valence ratings. Paralleling the analysis in Section 4.4.3, we characterized the game outcome in terms of whether it was good for the target player or if it was good for the player's partner. For 3rd-person emotion, we find clear support for Hypothesis 2 (the face and context interact). Smiles and open mouths showed clear evidence of an interaction. Smiles (F1) signal happiness when the outcome benefited the player who was smiling (CC or DC). Similarly, mouth tightening and eye tightening also demonstrate interactions. Notably, eye tightening (F5) signals happiness, especially when the player experiences mutual cooperation (CC).

The results show that F1, F3, F4, and F5 each interacted with game outcome in predicting 3rd-person valence ratings. The detailed analyses are shown below (see Figure. 10).

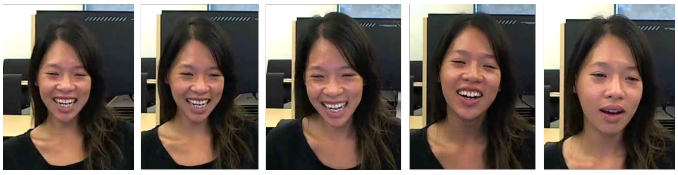


Fig. 11: Example video displaying high F1 (smile), F3 (open mouth), and F4 (eye tightening) scores in mutual cooperation (CC) outcome, indicative of a big smile.

The interaction term between own decision, partner's decision and Smile (F1) reached statistical significance, B = 0.162, t = 2.128, p = .036. A simple slope analysis revealed that F1 had a significant positive effect on perceived emotion across outcomes, but the relationship was much stronger for mutual cooperation (CC): CC (B = 0.338, t = 8.002, p< .001); CD (B = 0.190, t = 5.55, p < .001); DC (B = 0.150, t = 4.301, p< .001); DD (B = 0.165, t = 4.082, p < .001). Thus, smiles mainly contribute to perceived valence when the smile is known to have resulted from mutual cooperation.

The interaction between open mouth (F3) and game outcome was significant, B = 0.287, t = 3.108, p = .002. A simple slope analysis indicated that F3 had a strong significant positive effect on valence following mutual cooperation (CC) and a weak but significant effect following exploitation (CD), but there was no significant effect for the other outcomes: CC (B = 0.475, t = 7.483, p < .001); CD (B = 0.108, t = 4.055, p < .001); DC (B = 0.045, t = 0.884, p = .377); DD (B = -0.037, t = -1.040, p = .299). Thus, an open mouth mainly contributes to perceived valence when it is known to result from mutual cooperation.

The interaction between mouth tightening (F4) and game outcome was significant, B = 0.222, t = 2.54, p = .011. A simple slope analysis showed that F4 significantly impacted perceived emotion negatively, except for mutual defection (DD), where the effect was positive but not significant: CC (B = -0.08, t = -3.189, p = .001); CD (B = -0.142, t = -2.488, p = .013); DC (B = -0.128, t = -3.250, p = .001); DD (B = 0.035, t = 0.759, p = .448). Thus, mouth tightening predicts negative valence, except after mutual defection.

The interaction between eye tightening (F5) and game outcome was significant, B = 0.379, t = 3.24, p = .001. A simple slope analysis showed that F5 significantly impacted perceived emotion for all outcomes except mutual defection (DD), with the effect being much stronger following mutual cooperation (CC): CC (B = 0.536, t = 5.328, p < .001); CD (B = 0.124, t = 4.340, p < .001); DC (B = 0.092, t = 2.702, p = .007); DD (B = 0.058, t = 1.477, p = .140). Thus, eye tightening mainly contributes to perceived valence when occurring in the context of mutual cooperation.

Figure 10 overviews the regression coefficients (B) for the significant facial factors. Many facial cues exhibited a significant interaction with the context. Smiles (F1), open mouths (F3), and eye tightening (F4) were particularly strong indicators of positive valence in the context of mutual cooperation. For example, in the video snapshot from a mutual cooperation scenario (CC), as referenced in Figure 11, the participant displays a smile, an open mouth, and eye tightening, which collectively resemble a big smile. In contrast, smiles were the only cues associated with valence judgments following mutual defection (DD), and this correlation was relatively weak.

Overall, and in contrast with Study 2, these findings find support for H2 with regard to 3rd-person judgments and reinforce the findings of Study 1 by providing support for H3. Specifically, when observers make emotion judgments (in contrast to the first-person perspective) there is clear evidence of complex interactions between facial cues and context. Note that this stands in contrast with the results in

Study 2, thus offering mixed support for H3 overall.

7 General Discussion

Across three studies, we highlight the role of expression, context, and perspective in shaping emotion inferences. All studies find evidence that perspectives shape annotator labels and that the way expression and context combine to determine these labels also changes as a function of perspective. We find that context and expressions do not interact for 1st-person judgments but do for 2nd- and 3rd-person judgments. This supports hypotheses H3 (though the results for H3 were mixed for 3rd-person judgments). This is important because these interactions complicate approaches for context-based emotion recognition.

Study 1 examined the difference between 1st- and 2nd-person judgments. Open-mouth smiles proved an important feature that predicted positive valence from both perspectives, but there were important differences in how these cues were utilized. From the 1st-person perspective, open-mouth smiles were associated with positive emotions regardless of game outcome. By contrast, 2nd-person judgments were more nuanced: smiles only signaled positive emotion when the joint outcome benefited the smiler; otherwise, the smile was ignored. Frowns also interacted with context. Interestingly, frowns were seen as indicating positive emotion when the person frowning had just been exploited. As a result of these differences, partners were inaccurate in their judgments (i.e., weak correlation between self-reported and perceived valence).

Studies 2 and 3 considered the 3rd-person perspective. The pattern of 3rd-person judgments was similar to 2nd-person judgments (though smiles became a much stronger cue). These two studies showed inconsistent results about how facial cues and context interact, with only Study 3 finding strong evidence for an interaction. Overall, results reinforce prior research suggesting important differences between 2nd-and 3rd-person perspectives [7].

Some care must be taken in drawing firm conclusions across the two studies as there are important differences both in the participants and in how the judgments were obtained. Studies 1 and 2 showed raters each game outcome in succession. This allows contextual inferences, not only from the most recent action, but from the entire history of past actions. In contrast, Study 3 presented observers with a set of rounds drawn from different games, precluding their ability to reason about action sequences (e.g., that this was the second time in a row someone was exploited). Also, Study 1's participants were British, whereas Study 3 involved American participants. More broadly, annotators were forced to label expressions in terms of emotional meaning (e.g., valence), but more insight could be gained by allowing more open-ended questions or "think aloud" protocols to illuminate the meaning of these signals (see [18]).

8 Conclusion

Emotion recognition has been applied to a variety of perspectives. Some algorithms try to predict what someone feels (a 1st-person perspective). Others predict what observers think someone is feeling (a 3rd-person perspective). Finally, an algorithm might predict what someone engaged in social interaction thinks their partner is feeling (a 2nd-person perspective). While algorithms have tried to mix these perspectives (e.g., using labels provided by 3rd-person annotators to predict 1st-person feelings), emerging research in affective computing and psychology suggests this is problematic. In particular, observer judgments are strongly shaped by their understanding of the context that triggered these expressions.

In this paper, we carefully examined how perspective and context interact to shape the interpretation of facial expressions in a social task. We first highlighted psychological research showing why expressions are processed differently from different perspectives. We then used expressions produced in the iterated prisoner's dilemma to explore and characterize these differences. Our findings reinforce concerns about mixing perspectives. We find clear evidence that perspectives shape annotator labels and that the way expression and context interact to determine these labels changes as a function of perspective.

Taken together, our results reinforce previous findings that context and facial displays provide important cues about emotion. Past research has suggested that people are sometimes poor judges of emotion. Our study supports this, showing that people sometimes utilize invalid cues when estimating their partners' feelings. However, our more novel finding is that the way context and facial cues interact also differs between self-reported and perceived emotion. This may prove challenging for methods that aim to enhance the accuracy of emotion recognition by simply providing context to emotion annotators.

Despite the differences between the perspectives studied, our findings should not be interpreted as a basis for devaluing one perspective over another in practical applications such as emotional recognition or prediction tasks. Rather, it is important to recognize that each perspective offers valuable insights into emotional judgments that can be leveraged in complementary ways.

Acknowledgments

We thank Cleo Yao and Eli Fast for assisting Study 2. Work was supported by the European Office of Aerospace Research & Development (FA9550-18-1-0060) and the Army Research Office (Cooperative Agreement W911NF-20-2-0053). The views expressed are of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

[1] T. Hassan, D. Seuß, J. Wollenberg, K. Weitz, M. Kunz, S. Lautenbacher, J.-U. Garbas, and U. Schmid, "Automatic detection of pain from facial expressions: a survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 6, pp. 1815–1831, 2019.

[2] A. Jan, H. Meng, Y. F. A. Gaus, F. Zhang, and S. Turabzadeh, "Automatic depression scale prediction using facial expression dynamics and regression," *Avec '14*, p. 73–80, 2014.

[3] D. McDuff, R. E. Kaliouby, J. F. Cohn, and R. W. Picard, "Predicting ad liking and purchase intent: Large-scale analysis of facial responses to ads," *IEEE Transactions on Affective Computing*, vol. 6, no. 3, pp. 223–235, 2015.

[4] J. C. Peterson, S. Uddenberg, T. L. Griffiths, A. Todorov, and J. W. Suchow, "Deep models of superficial face judgments," *Proceedings of the National Academy of Sciences*, vol. 119, no. 17, p. e2115228119, 2022.

[5] I. Gordon, M. D. Pierce, M. S. Bartlett, and J. W. Tanaka, "Training facial expression production in children on the autism spectrum," *Journal of autism and developmental disorders*, vol. 44, pp. 2486–2498, 2014.

[6] E. Redcay and L. Schilbach, "Using second-person neuroscience to elucidate the mechanisms of social interaction," *Nature Reviews Neuroscience*, vol. 20, no. 8, pp. 495–505, 2019.

[7] L. Schilbach, B. Timmermans, V. Reddy, A. Costall, G. Bente, T. Schlicht, and K. Vogeley, "Toward a second-person neuroscience1," *Behavioral and brain sciences*, vol. 36, no. 4, pp. 393–414, 2013.

[8] B. Parkinson, "Heart to heart: A relation-alignment approach to emotion's social effects," *Emotion Review*, vol. 13, no. 2, pp. 78–89, 2021.

[9] C. Moore and J. Barresi, "The role of second-person information in the development of social understanding," *Frontiers in Psychology*, vol. 8, p. 277164, 2017.

[10] G. A. Van Kleef, "How emotions regulate social life: The emotions as social information (easi) model," *Current directions in psychological science*, vol. 18, no. 3, pp. 184–188, 2009.

[11] C. de Melo, P. J. Carnevale, S. J. Read, and J. Gratch, "Reading people's minds from emotion expressions in interdependent decision making," *Journal of Personality and Social Psychology*, vol. 106, no. 1, pp. 73–88, 2014.

[12] L. F. Barrett, B. Mesquita, and M. Gendron, "Context in emotion perception," *Current Directions in Psychological Science*, vol. 20, no. 5, pp. 286–290, 2011.

[13] C. Busso and S. S. Narayanan, "The expression and perception of emotions: comparing assessments of self versus others." in *Interspeech*, 2008, pp. 257–260.

[14] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak, "Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements," *Psychological Science in the Public Interest*, vol. 20, no. 1, pp. 1–68, 2019.

[15] H. Aviezer, R. Hassin, S. Bentin, and Y. Trope, "Putting facial expressions back in context," *First impressions*, pp. 255–286, 2008.

[16] D. C. Ong, J. Zaki, and N. D. Goodman, "Affective cognition: Exploring lay theories of emotion," *Cognition*, vol. 143, pp. 141–162, 2015.

[17] B. Han, C. Yau, S. Lei, and J. Gratch, " Knowledge-Based Emotion Recognition Using Large Language Models ," in *2024 12th International Conference on Affective Computing and Intelligent Interaction (ACII)*. Los Alamitos, CA, USA: IEEE Computer Society, Sep. 2024, pp. 1–9.

[18] M. Hladkỳ, R. Guerra, X. Cang, K. Maclean, P. Gebhard, and T. Schneeberger, "Modeling the 'kiss my ass'-smile: Appearance and functions of smiles in negative social situations," in *12th International Conference on Affective Computing and Intelligent Interaction (ACII). Glasgow, UK*, 2024.

[19] K. N'diaye, D. Sander, and P. Vuilleumier, "Self-relevance processing in the human amygdala: gaze direction, facial expression, and emotion intensity." *Emotion*, vol. 9, no. 6, p. 798, 2009.

[20] L. Rees and R. Friedman, "Not your garden (hose) variety emotion: An integrative review of the flows of anger and a path forward," *Academy of Management Annals*, vol. 19, no. 1, pp. 132–179, 2025.

[21] R. P. Cubitt, M. Drouvelis, S. Gächter, and R. Kabalin, "Moral judgments in social dilemmas: How bad is free riding?" *Journal of Public Economics*, vol. 95, no. 3-4, pp. 253–264, 2011.

[22] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "Avec 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proceedings of the 6th international workshop on audio/visual emotion challenge*, 2016, pp. 3–10.

[23] M. Wegge, E. Troiano, L. A. M. Oberlaender, and R. Klinger, "Experiencer-specific emotion and appraisal prediction," in *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, D. Bamman, D. Hovy, D. Jurgens, K. Keith, B. O'Connor, and S. Volkova, Eds. Abu Dhabi, UAE: Association for Computational Linguistics, Nov. 2022, pp. 25–32. [Online]. Available: https://aclanthology.org/2022.nlpcss-1.3

[24] T. Wakihira, M. Morimoto, S. Higuchi, and Y. Nagatomi, "Can facial expressions predict beer choices after tasting? a proof of concept study on implicit measurements for a better understanding of choice behavior among beer consumers," *Food Quality and Preference*, vol. 100, p. 104580, 2022.

[25] M. Angelika-Nikita, C. M. de Melo, K. Terada, G. Lucas, and J. Gratch, "The impact of partner expressions on felt emotion in the iterated prisoner's dilemma: An event-level analysis," *arXiv preprint arXiv:2207.00925*, 2022.

[26] D. Antos, C. De Melo, J. Gratch, and B. Grosz, "The influence of emotion expression on perceptions of trustworthiness in negotiation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 25, no. 1, 2011, pp. 772–778.

[27] R. Hoegen, G. Stratou, and J. Gratch, "Incorporating emotion perception into opponent modeling for social dilemmas," in *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, 2017, pp. 801–809.

[28] G. Bijlstra, R. W. Holland, and D. H. Wigboldus, "The social face of emotion recognition: Evaluations versus stereotypes," *Journal of Experimental Social Psychology*, vol. 46, no. 4, pp. 657–663, 2010.

[29] M. K. MacLin, C. Downs, O. H. MacLin, and H. M. Caspers, "The effect of defendant facial expression on mock juror decision-making: The power of remorse," *North American Journal of Psychology*, vol. 11, no. 2, pp. 323–332, 2009.

[30] C. M. De Melo, P. J. Carnevale, S. J. Read, and J. Gratch, "Reading people's minds from emotion expressions in interdependent decision making." *Journal of personality and social psychology*, vol. 106, no. 1, p. 73, 2014.

[31] C. Y. Park, N. Cha, S. Kang, A. Kim, A. H. Khandoker, L. Hadjileontiadis, A. Oh, Y. Jeong, and U. Lee, "K-emocon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations," *Scientific Data*, vol. 7, no. 1, p. 293, 2020.

[32] S. K. D'mello and A. Graesser, "Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features," *User Modeling and User-Adapted Interaction*, vol. 20, pp. 147–187, 2010.

[33] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, "Context based emotion recognition using emotic dataset," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 11, pp. 2755–2766, 2020.

[34] R. H. Frank, *Passions within reason: The strategic role of the emotions*. WW Norton & Co, 1988.

[35] S. Lei and J. Gratch, "Emotional expressivity is a reliable signal of surprise," *IEEE Transactions on Affective Computing*, pp. 1–12, 2023.

[36] R. Hoegen, J. Gratch, B. Parkinson, and D. Shore, "Signals of emotion regulation in a social dilemma: Detection from face and context," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2019, pp. 1–7.

[37] J. Hoegen, G. Lucas, D. Shore, B. Parkinson, and J. Gratch, "How expression and context determine second-person judgments of emotion," in *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2023, pp. 1–7.

[38] P. Ekman *et al.*, "Basic emotions," *Handbook of cognition and emotion*, vol. 98, no. 45-60, p. 16, 1999.

[39] B. Zhang, G. Essl, and E. Mower Provost, "Automatic recognition of self-reported and perceived emotion: Does joint modeling help?" in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 2016, pp. 217–224.

[40] L. Schilbach, A. M. Wohlschlaeger, N. C. Kraemer, A. Newen, N. J. Shah, G. R. Fink, and K. Vogeley, "Being with virtual others: Neural correlates of social interaction," *Neuropsychologia*, vol. 44, no. 5, pp. 718–730, 2006.

[41] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, "Emotion recognition in context," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1667–1675.

[42] S. K. D'mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM computing surveys (CSUR)*, vol. 47, no. 3, pp. 1–36, 2015.

[43] C. Mumenthaler, D. Sander, and A. S. Manstead, "Emotion recognition in simulated social interactions," *IEEE Transactions on Affective computing*, vol. 11, no. 2, pp. 308–312, 2018.

[44] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using

deep neural networks," *IEEE Journal of selected topics in signal processing*, vol. 11, no. 8, pp. 1301–1309, 2017.
[45] J. M. Carroll and J. A. Russell, "Do facial expressions signal specific emotions? judging emotion from the face in context." *Journal of personality and social psychology*, vol. 70, no. 2, p. 205, 1996.
[46] R. Righart and B. De Gelder, "Rapid influence of emotional scenes on encoding of facial expressions: an erp study," *Social cognitive and affective neuroscience*, vol. 3, no. 3, pp. 270–278, 2008.
[47] D. Mobbs, N. Weiskopf, H. C. Lau, E. Featherstone, R. J. Dolan, and C. D. Frith, "The kuleshov effect: the influence of contextual framing on emotional attributions," *Social cognitive and affective neuroscience*, vol. 1, no. 2, pp. 95–106, 2006.
[48] T. Mittal, P. Guhan, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "Emoticon: Context-aware multimodal emotion recognition using frege's principle," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14 234–14 243.
[49] J. Lee, S. Kim, S. Kim, J. Park, and K. Sohn, "Context-aware emotion recognition networks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 10 143–10 152.
[50] X. Zhang, T. Wang, X. Li, H. Yang, and L. Yin, "Weakly-supervised text-driven contrastive learning for facial behavior understanding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 20 751–20 762.
[51] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, and R. Zimmermann, "Conversational memory network for emotion recognition in dyadic dialogue videos," in *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, vol. 2018. NIH Public Access, 2018, p. 2122.
[52] S. D. Houlihan, M. Kleiman-Weiner, L. B. Hewitt, J. B. Tenenbaum, and R. Saxe, "Emotion prediction as computation over a generative theory of mind," *Philosophical Transactions of the Royal Society A*, vol. 381, no. 2251, p. 20220047, 2023.
[53] E. Brunswik, *Perception and the representative design of psychological experiments*. Univ of California Press, 1956.
[54] M. A. Ruben and J. A. Hall, "A lens model approach to the communication of pain," *Health Communication*, vol. 31, no. 8, pp. 934–945, 2016.
[55] S. Hirschmüller, B. Egloff, S. Nestler, and M. D. Back, "The dual lens model: A comprehensive framework for understanding self–other agreement of personality judgments at zero acquaintance." *Journal of Personality and Social Psychology*, vol. 104, no. 2, p. 335, 2013.
[56] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett, "The computer expression recognition toolbox (cert)," in *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. IEEE, 2011, pp. 298–305.
[57] H. Ilgen, J. Israelashvili, and A. Fischer, "Personal nonverbal repertoires in facial displays and their relation to individual differences in social and emotional styles," *Cognition and Emotion*, vol. 35, no. 5, pp. 999–1008, 2021.
[58] D. M. Watson, B. B. Brown, and A. Johnston, "A data-driven characterisation of natural facial expressions when giving good and bad news," *PLOS Computational Biology*, vol. 16, no. 10, p. e1008335, 2020.
[59] G. Stratou, J. Van Der Schalk, R. Hoegen, and J. Gratch, "Refactoring facial expressions: An automatic analysis of natural occurring facial expressions in iterative social dilemma," in *2017 Seventh international conference on affective computing and intelligent interaction (ACII)*. IEEE, 2017, pp. 427–433.
[60] E. Fehr and K. M. Schmidt, "A theory of fairness, competition, and cooperation," *The quarterly journal of economics*, vol. 114, no. 3, pp. 817–868, 1999.

Bin Han is a PhD student in Computer Science at the University of Southern California. Her research interests include Human-Computer Interaction, Affective Computing, Virtual Humans, and Facial Emotion Recognition. She received her BS and MS degrees in Computer Science from Korea University in 2019 and 2021, respectively.

Jessie Hoegen Received her PhD in Computer Science from the University of Southern California (USC) in 2024. She received her MSc degree in Human Media Interaction at the University of Twente in 2015. Her research focuses on developing computational models of cognitive processes and human emotions, exploring how these interact and affect human behavior and decision-making.

Su Lei is an AI researcher in audEERING GmbH. She completed her Ph.D in Computer Science at the USC in 2023. Her research focuses on automatic understanding of human behavior that people interpret as emotional, including the role of context as well as facial expression synchrony.

Gale M. Lucas earned her Ph.D. in psychology from Northwestern University iin 2010. She is a Research Associate Professor at USC's Viterbi School of Engineering and works at the USC Institute for Creative Technologies. She works in the areas of human-computer interaction, affective computing, and trust-in-automation. Her research focuses on rapport, disclosure, trust, persuasion, and negotiation with virtual agents and social robots.

Danielle Shore is Deputy Research Director of Clinical Research and Training at the University of Oxford, a Research tutor at the Oxford Institute of Clinical Psychology Training and Research and a lecturer at Queen's College, Oxford. Her research aims to understand what shapes social interactions and decisions, focussing on the role of facial expressions. She also investigates how social interactions and decisions impact well-being.

Brian Parkinson is Professor of Social Psychology at the University of Oxford. He focuses on the social psychology of emotion, expression, and interpersonal regulation, particularly how emotions configure relationships and regulate orientations towards objects and events. His latest book, Heart to Heart: How Your Emotions Affect Other People, was selected as an Outstanding Academic Title of 2020 by Choice, the magazine of the American Library Association.

Jonathan Gratch is a Research Professor of Computer Science and Psychology at USC. He completed his Ph.D. at the University of Illinois in Urbana-Champaign in 1995. He examines models of human cognitive and social processes to advance theory and shaping human-machine interaction. He was the founding Editor-in-Chief of IEEE's Transactions on Affective Computing and Fellow of AAAC, AAAI and CogSci.