American Psychological Association

Emotion

© 2019 American Psychological Association
ISSN: 1528-3542

2021, Vol. 21, No. 1, 96–107
http://dx.doi.org/10.1037/emo0000685

Leveraging Facial Expressions and Contextual Information to Investigate Opaque Representations of Emotions

Stefano Anzellotti
Boston College

Sean Dae Houlihan
Massachusetts Institute of Technology

Samuel Liburd Jr.
University of the Virgin Islands

Rebecca Saxe
Massachusetts Institute of Technology

Observers attribute emotions to others relying on multiple cues, including facial expressions and information about the situation. Recent research has used Bayesian models to study how these cues are integrated. Existing studies have used a variety of tasks to probe emotion inferences, but limited attention has been devoted to the possibility that different decision processes might be involved depending on the task. If this is the case, understanding emotion representations might require understanding the decision processes through which they give rise to judgments. This article 1) shows that the different tasks that have been used in the literature yield very different results, 2) proposes an account of the decision processes involved that explain the differences, and 3) tests novel predictions of this account. The results offer new insights into how emotions are represented, and more broadly demonstrate the importance of taking decision processes into account in Bayesian models of cognition.

Keywords: emotions, Bayesian models, facial expressions, context, cue integration

Supplemental materials: http://dx.doi.org/10.1037/emo0000685.supp

Understanding other people's emotional experience is critical to guide social interactions; it helps us to predict what they will do and how they will respond to our actions. Facial expressions have been traditionally considered an important source of information about other people's emotions (Ekman & Oster, 1979). Observers are relatively accurate at recognizing the emotions of posed facial expressions (Calder et al., 2003). According to the theory of Basic Emotions (Adolphs, 2010; Darwin & Prodger, 1998; Ekman, 1992), emotional experience is subdivided into a small number of emotion categories, associated with the production of specific behaviors and expressions that evolved as adaptive responses.

However, Carroll and Russell (1996) found that the same facial expressions could be interpreted as reflecting different emotions depending on the context. Aviezer, Trope, and Todorov (2012) have shown that participants were at chance at judging the valence of expressions displaying extreme disappointment or extreme happiness; when facial expressions were cropped and paired with a positive or negative valenced context, the context was found to dominate the emotion inferences performed by participants (Aviezer et al., 2012). A growing body of evidence points to the importance of context for emotion attribution (Barrett, Mesquita, & Gendron, 2011; Hassin, Aviezer, & Bentin, 2013). Therefore, studying how information about facial expressions and information about context are integrated is key to understand emotion attribution.

Recent research has proposed to investigate emotion attribution within a Bayesian framework (Ong, Zaki, & Goodman, 2015; Saxe

This article was published Online First October 3, 2019.

Stefano Anzellotti, Department of Psychology, Boston College; Sean Dae Houlihan, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology; Samuel Liburd Jr., Department of Biology, University of the Virgin Islands; Rebecca Saxe, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology.

We thank Alex Todorov and Hillel Aviezer for sharing their stimuli; Jonathan Phillips for our conversation on opaque representations, epistemic transparency, competence, and performance; and James Russell, Alex Todorov, and Kathryn C. O'Nell for invaluable comments on a draft of the article. Stefano Anzellotti was supported by Boston College and a Pilot Award from the Simons Foundation Autism Research Initiative (SFARI). This material is based upon work supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216. Author contributions: Stefano Anzellotti developed the study concept. All authors contributed to the study design. Testing, data collection, and data analysis were performed by Stefano Anzellotti, Samuel Liburd Jr., and Sean Dae Houlihan. Stefano Anzellotti and Rebecca Saxe interpreted the results. Stefano Anzellotti drafted the manuscript, and all authors provided critical revisions. All authors approved the final version of the manuscript for submission. Open practices statement: The experiments reported in this article were not formally preregistered. Requests for the data or materials can be sent via email to the corresponding author at stefano.anzellotti@bc.edu.

Correspondence concerning this article should be addressed to Stefano Anzellotti, Department of Psychology, Boston College, McGuinn Hall Room 334, 275 Beacon Street, Chestnut Hill, MA 02467. E-mail: stefano.anzellotti@bc.edu

& Houlihan, 2017; Wu & Schulz, 2018). In this perspective, human inferences about the emotions of others are modeled as a process that maps the observable inputs (the causal context **c** and the observable facial expression **x**) onto a probability distribution over different emotions **e** following the rules of probabilistic Bayesian inference: P(**e**|**c**,**x**). A Bayesian approach to emotion attribution offers several advantages: It leads to formal models that generate quantitative predictions, and it models not only the participants' judgments, but also their uncertainty.

Testing Bayesian models of emotion attribution relies on measuring behaviorally the probabilistic relationships between cues and inferred emotions. Several different tasks have been used to this end in the previous literature. For example, Ong et al. (2015) asked how a character felt given a specific outcome (Experiment 1) and how likely was an outcome given the emotion that ensued (Experiment 2); Wu and Schulz (2018) asked, given an expected outcome, how likely were different facial expressions.

If participants' responses in all of these tasks reflect the probabilities they attribute to different emotions, using different tasks should yield the same responses, and the different tasks could be used interchangeably (we will say in this case that emotion representations are "transparent"). By contrast, participants' responses in the different tasks might be the outcome of distinct decision processes operating on the same emotion representations, and thus different tasks might yield different responses (we will say in this case that emotion representations are "opaque"; Figure 1). In this case, only some, and possibly none, of these distributions might correspond to the probabilistic representations that participants use for inference. Therefore, understanding emotion attribution requires determining whether emotion representations are transparent or opaque, and if they are opaque, it requires modeling jointly the decision processes and the emotion representations to which they are applied.

This article is organized in two parts. In the first part, we investigated whether probabilistic representations of emotions are transparent or opaque. To do so, we asked participants to perform emotion inferences given facial expressions, contextual information, or both. We compared the distributions computed from participants' ratings using three types of tasks that have been used in the emotion attribution literature. Given a set of cues (i.e., a facial expression or a context), the first type of task asked participants about the valence of the emotion experienced by a character; the second provided a fixed emotion valence, and asked how likely it was; the third asked participants to produce a probability distribution over emotional valences.

Since we found that representations of emotions are opaque, in the second part of the article we attempted to recover the participants' internal representations of emotions. This is a challenging inverse problem. We introduced two key criteria that must be satisfied by internal representations of emotions, and we tested whether one of the three tasks commonly used in the field produces judgments that satisfy the criteria.

The first criterion is based on cue integration. If Bayesian cue integration well approximates the participants' inferences, and behavioral ratings for a type of task yield distribution close to the probabilistic representations used for inference, then Bayesian cue integration should well approximate the relationship between the behavioral ratings as well. By contrast, if behavioral ratings are the output of a decision process that transforms the internally represented probabilities, that decision process could have led to information loss, and Bayesian integration might no longer account for the relationship between the behavioral ratings for context, emotions, and expressions.

The second criterion is based on "linking functions." If the behavioral ratings obtained in one task T_0 are close to the internally represented probabilities, there should exist for each other task T_i "linking functions" such that the judgments in the tasks T_i can be obtained as a function $f_i(T_0)$ of the judgments in task T_0. These "linking functions" are none other than the functions implemented by the decision processes J_i for each of the other tasks ($f_i \approx J_i$). We started from the task that best satisfied the first criterion, and formulated hypotheses on the decision processes used to generate the ratings in the other tasks. Next, we assessed whether these hypotheses could account for the observed data, and we tested new predictions they generated with novel experiments and simulations. Ultimately, the two strategies converged to indicate that distributional probability judgments are a good candidate to approximate participants' internal probabilistic representations of emotions.

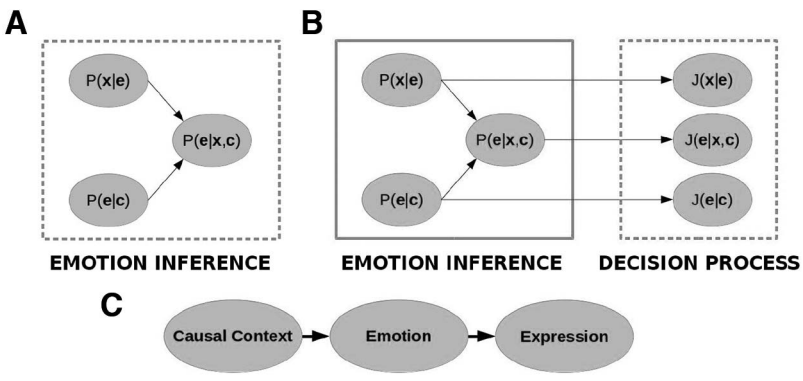

Figure 1. (A) Transparent representations: judgments in different tasks directly reflect the emotion probabilities; (B) Opaque representations: judgments are the outcome of decision processes applied to latent (non-directly-observable) emotion probabilities; (C) directed acyclic graph depicting the dependencies between causal context, emotion, and expression.

Method

Testing the Opacity of Emotion Representations

In the first part of the article, we tested whether representations of emotions are transparent or opaque. To do so, we performed three experiments asking participants to infer emotions from facial expressions, contextual information, or both. Since the paradigms used in the three experiments are similar, to avoid repetitions we describe them jointly as follows.

Stimuli

All experiments including face images used eight face images chosen from the stimuli of Aviezer et al. (2012): two of winning female tennis players, two of losing female tennis players, two of winning male tennis players, and two of losing male tennis players.

For experiments including context, participants were told that the target character was a tennis player who had just won (in half of the trials; or lost, in the remaining trials) an important point. For experiments including partial context, participants were only told that the target character was a tennis player at the end of a rally, but were not told whether the player had won or lost.

Participants

A total of 1,755 participants were recruited through Mechanical Turk Marketplace. Thanks to Mechanical Turk, we could collect large samples of participants, leading to highly significant results in our statistical tests. For the valence ratings and the pointwise probability ratings, participants completed a very brief online task (approximately 3 min) and received $0.15 regardless of their performance on the task. For the pointwise distributional ratings, participants completed a brief online task (approximately 6 min) and received $0.80 regardless of their performance on the task. Data were collected with the approval of the MIT Institutional Review Board.

Before and after the ratings of interest, participants had to complete simple control tasks designed to ensure that they were reading and following the instructions. The data from participants who failed the control tasks were discarded prior to the analysis.

Experiment Design

In each Mechanical Turk assignment, a participant completed a single trial of interest for the analysis, to remove effects of prior judgments on later judgments. Participants were shown a cue and were asked to perform a judgment about the emotion of the target character given that cue.

We used five types of cues, and three types of tasks, for a total of 15 conditions (Figure 2B). Different participants took part to each different condition. The cue was either an image of a target character's facial expression, a sentence providing contextual information about her/his situation, or both; the sentences could provide either positive contextual information ("the tennis player won a point") or negative contextual information ("the tennis player lost a point").

The task could request to provide a judgment of the valence of emotions, a pointwise probability judgment, or a distributional probability judgment (tasks are described in more detail below).

For judgments of the valence of emotions, participants were asked to report how negative or positive was the character's emotion by adjusting a slider from 0 (*very negative*) to 48 (*very positive*). The choice of a scale with 49 steps was motivated by our aim to subdivide it into seven equal bins of seven steps.

For pointwise judgments of the probability of different emotion valences, participants were shown a slider fixed on a specific value from 0 (*very negative*) to 48 (*very positive*). Participants were provided with a second slider to judge how likely the character was to experience the fixed emotional valence given the available cue, going from very unlikely to very likely.

For distributional judgments of the probability of different emotion valences, participants were asked how likely the target character was to experience emotions going from negative (1) to neutral (4) to positive (7), using an adjustable bar chart with seven bars, with heights that automatically maintained a total sum of 100. Prior to the trial of interest, participants were familiarized with the adjustable bar graph using two example trials, in which they were asked to express a given distribution over emotion valences. In the first practice trial, they were asked to raise the bar for "neutral" to 100. In the second practice trial, they were asked to raise the bar for "very happy" to 60, for "happy" to 30, and for "somewhat happy" to 10. Participants could proceed from the practice trial to the trial of interest only if they completed the practice trials correctly within a margin of error of 10 on each bar.

Data Analysis

Judgments of emotion valence were binned into seven bins of seven values each, and the proportion of participants whose judgments fell within each of the bins was plotted.

Pointwise probability judgments were grouped based on the fixed emotional valence that was presented. For each of the seven emotional valence bins, the corresponding probability judgments were averaged. The resulting distribution was then normalized to sum to 1.

Distributional judgments were averaged yielding seven mean probabilities, one for each of the valence levels from 1 to 7. The significance of the difference between the distributional probability judgments and the other distributions was tested with a Hotelling T^2 test.

Testing Different Types of Behavioral Judgments as Approximations of Internal Representations

In the second part of the article, we aimed to tackle the challenges posed by opaque representations, testing whether one of the three tasks typically used in the field is a good candidate to reflect internal representations of emotions. In general, inverting the decision functions mapping from internal representations to behavioral judgments may not be possible. Such functions might be noninjective (as in the case of maximum likelihood estimator; MLE). However, we defined two criteria that behavioral judgments must satisfy if they do match closely internal representations of emotions.

Criterion I: Operations on Representations

The first criterion is that if behavioral judgments closely match internal representations, then it should be possible to find opera-

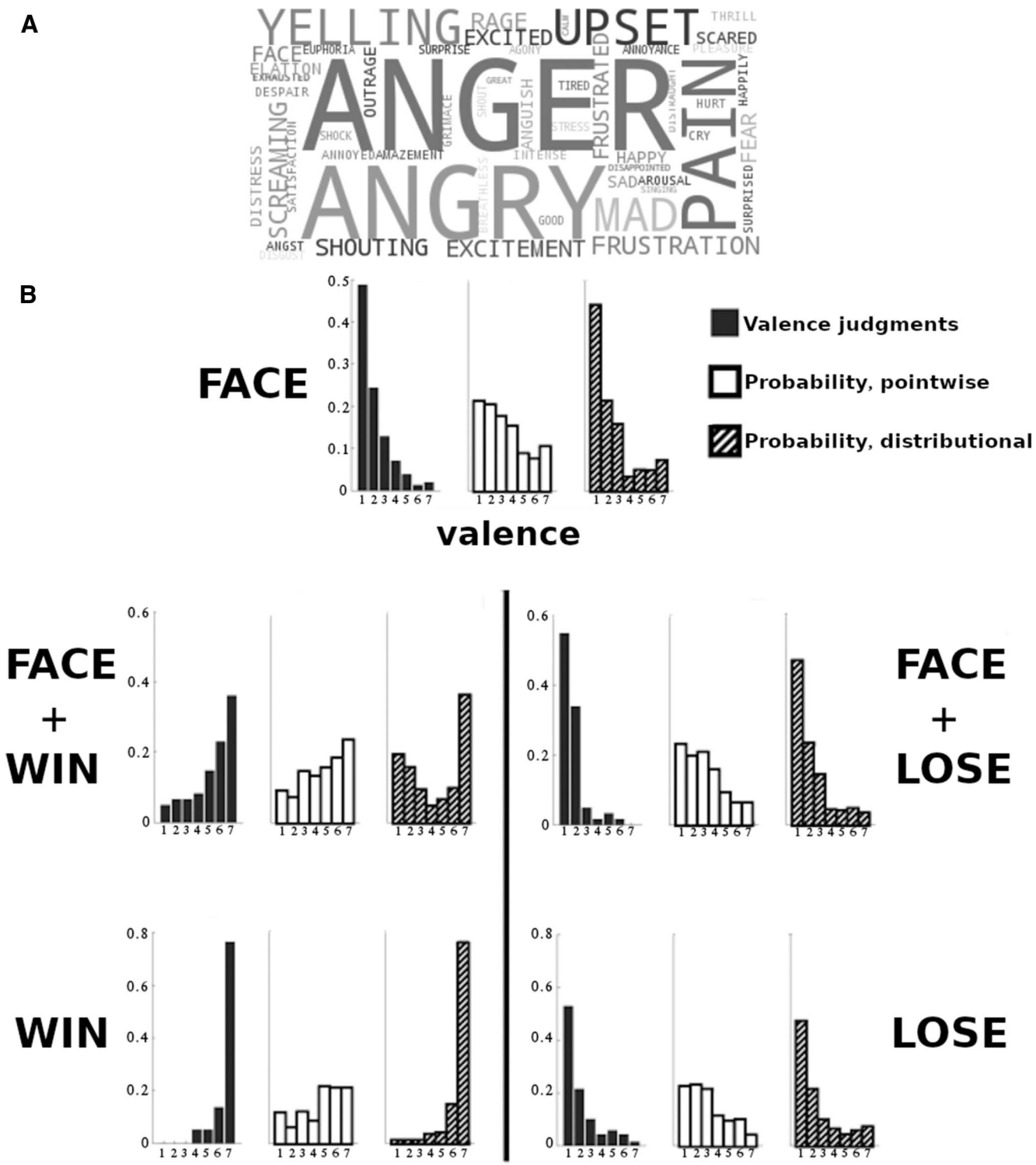

Figure 2. (A) Word cloud depicting free labeling of the emotions inferred by participants given the facial expressions; (B) Participants' ratings of the valence of the emotion (left), of the probability of the emotion given a fixed valence ("pointwise" probability ratings, center), and of the probability distribution over possible valences ("distributional" probability ratings, right) given facial expressions (top), facial expressions and context (middle), and context only (bottom).

tions on the behavioral judgments that correspond to the operations on representations. Let's consider representations α, β, γ, and an operation on representations f such that $\gamma = f(\alpha, \beta)$. Let's also consider a decision function J that maps the representations to behavioral judgments $J(\alpha)$, $J(\beta)$, $J(\gamma)$. We can try to find a function g such that $J(\gamma) = g(J(\alpha), J(\beta))$.

However, a function g that satisfies this requirement may not exist. The operation f occurs before the decision process; therefore, it can use all the information in α, β, and γ. But since representations of emotions are opaque, J could transform the internal representations in ways that lead to information loss. Then, it might be no longer possible to find a g such that $J(\gamma) = g(J(\alpha), J(\beta))$.

In the special case of a task such that the behavioral judgments match closely internal representations (that is, $J(\alpha) \approx \alpha$, $J(\beta) \approx \beta$, $J(\gamma) \approx \gamma$), it is guaranteed that we can also find a function g that satisfies the requirement. In fact, if we choose $g = f$, this choice of g will satisfy the requirement. Summing up, if behavioral judgments measured with a particular task closely match internal representations, for each operation f on internal representations we should be able to find a corresponding operation g on the behavioral judgments.

Application of Criterion I: Cue Integration

To apply criterion I, we need an operation over representations. Cue integration is an operation over representations that has been studied in previous work: It has been proposed that cue integration is Bayesian (Ong et al., 2015). Furthermore, earlier research has taken advantage of the simplifying assumption that there is a uniform prior on emotions ($P(\mathbf{e}) \approx constant$) to formulate the following cue integration equation (the proof is reported in the online supplemental materials):

$$P(\mathbf{e}\,|\,\mathbf{c}, \mathbf{x}) \propto P(\mathbf{e}\,|\,\mathbf{x})P(\mathbf{e}\,|\,\mathbf{c}). \quad (1)$$

This assumption is helpful but needs to be used with caution, as it is quite possible that the true distribution $P(\mathbf{e})$ is not uniform. We can observe that the equation gives us an operation that we can use to test criterion I in the case of emotions. In fact, we can pose the following:

$$\alpha = P(\mathbf{e}\,|\,\mathbf{x}) \quad (2)$$

$$\beta = P(\mathbf{e}\,|\,\mathbf{c}) \quad (3)$$

$$\gamma = P(\mathbf{e}\,|\,\mathbf{c}, \mathbf{x}) \quad (4)$$

and

$$\gamma \propto f(\alpha, \beta) = \alpha\beta. \quad (5)$$

If behavioral ratings obtained with one task closely match the underlying representations of emotion probabilities used for inference, we can expect the behavioral ratings for emotions given context and expression to be related to the behavioral ratings for emotions given context alone and to the behavioral ratings for emotions given expression alone following Equation 1. That is, if a task T yields behavioral judgments J_T such that $J_T(P(\mathbf{e}\,|\,\mathbf{x})) \approx P(\mathbf{e}\,|\,\mathbf{x})$ (and so on), the behavioral judgments need to satisfy the condition

$$J_T(\mathbf{e}\,|\,\mathbf{c}, \mathbf{x}) \propto J_T(\mathbf{e}\,|\,\mathbf{x})J_T(\mathbf{e}\,|\,\mathbf{c}). \quad (6)$$

We tested to which extent different behavioral tasks produce ratings that satisfy this condition (see Figure 3).

Criterion II: Linking Functions

One limitation of criterion I is that, while behavioral judgments that match internal representations need to satisfy it, it is not entirely guaranteed that if a set of behavioral judgments satisfy the criterion, they must be identical to the internal representations. For this reason, we aimed to strengthen the conclusions obtained by applying criterion I using a second criterion, based on linking functions.

Let's consider a representation α, a task T_0 that yield judgments close to the internal representations (such that $J_{T_0}(\alpha) \approx \alpha$), and other n tasks $T_1, \ldots, T_n$. Then, for each $i = 1, \ldots, n$ there must exist "linking functions" $f_1, \ldots, f_n$ such that $J_{T_i}(\alpha) \approx f_i(J_{T_0}(\alpha))$. In fact, if we choose $f_i = J_{T_i}$, since $J_{T_0}(\alpha) \approx \alpha$, we have that $f_i(J_{T_0}(\alpha)) \approx f_i(\alpha) = J_{T_i}(\alpha)$. In other word, criterion II states that if a task T_0 yields behavioral judgments that match closely internal representations, the behavioral judgments in response to all other tasks must be computable as functions of the behavioral judgments in task T_0. We tested whether we could find such linking functions between the task that best satisfied criterion I (distributional probability judgments) and the other two tasks.

Linking Distributional Probability Judgments and Judgments of Emotion Valence

In addition to testing which behavioral ratings are best approximated by Bayesian cue integration, we can test hypotheses about the decision processes that map probabilistic representations onto observed judgments. When participants are given a context or are shown a character's facial expression, and they are asked to produce a judgment about the character's emotion ("is Sue somewhat happy, or very happy?") as opposed to a judgment about probability ("how likely is it that Sue is very happy?"), they may be faced with the task of converting the representation of a probability distribution ($P(\mathbf{e}|\mathbf{c})$ or $P(\mathbf{e}|\mathbf{x})$) into a single emotion value ($J(\mathbf{e}|\mathbf{c})$ or $J(\mathbf{e}|\mathbf{x})$), i.e.

$$J(\mathbf{e}\,|\,\mathbf{x}) = f(P(\mathbf{e}\,|\,\mathbf{x})). \quad (7)$$

In particular cases, the assumption that the probability distribution $P(\mathbf{e}|\mathbf{x})$ is equal to the distribution across multiple trials of the judgments of valence $J(\mathbf{e}|\mathbf{x})$ might hold. For example, this is the case if the judgments are generated sampling an emotion from the set of emotions with probability given by the distribution $P(\mathbf{e}|\mathbf{x})$. However, in general the distribution of judgments of valence will be different from the distribution of probabilities. For example, a likely mechanism to generate judgments of valence $J(\mathbf{e}|\mathbf{x})$ can be selecting the emotional valence $\mathbf{e}$ with the highest probability (henceforth "maximum likelihood estimator" or "MLE"):

$$J(\mathbf{e}\,|\,\mathbf{x}) = \arg\max_{\mathbf{e}}(P(\mathbf{e}\,|\,\mathbf{x})). \quad (8)$$

It can be shown that choosing the emotion with the highest probability is the optimal decision mechanism predicted by Decision Theory (Berger, 2013; Körding & Wolpert, 2006), if the utility function is $u(J(\mathbf{e}|\mathbf{x})) = 1$ if $J(\mathbf{e}\,|\,\mathbf{x}) = \tilde{\mathbf{e}}$ and 0 otherwise. In

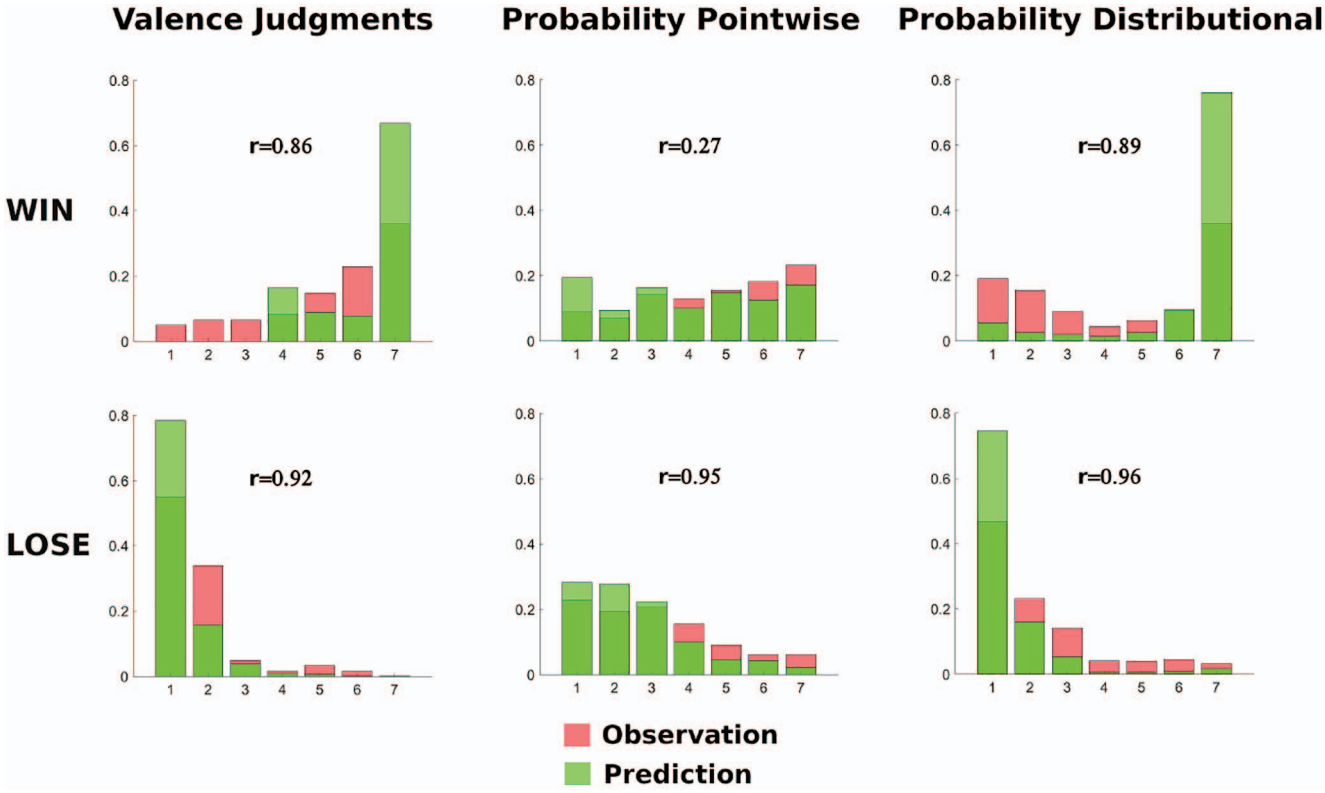

Figure 3. Comparison between the observed integration between expression and context information, and the integration predicted by a Bayesian model relying on the participant's emotion inferences given expression alone, and context information alone. The distributional ratings for the integration of expression and context show the highest correlations to the integration predicted by the model for both win and lose contexts. See the online article for the color version of this figure.

this case, the expected utility function is equal to the probability distribution: $\mathbb{E}(u(J(\mathbf{e} \mid \mathbf{x}))) = P(\mathbf{e} \mid \mathbf{x})$, and therefore the judgment is $\arg\max_{\mathbf{e}}[\mathbb{E}(u(J(\mathbf{e} \mid \mathbf{x}))] = \arg\max_{\mathbf{e}}[P(\mathbf{e} \mid \mathbf{x})]$.

Judgments of emotion valence generated applying Maximum Likelihood (ML) to a probability distribution $P(\mathbf{e}|\mathbf{x})$ will have reduced tails as compared to the original probability distribution, and if the probability distribution is bimodal, the nondominant peak should be reduced in the distribution of judgments of valence $J(\mathbf{e}|\mathbf{x})$.

Simulations. We generated simulated valence judgments applying ML to the distributions obtained from the distributional probability task. For each bin (1 to 7), a gamma distribution was used to model the distribution of values across participants for that bin (the values are non-negative). We then used the gamma distributions to simulate the distributional probability judgments of 100 subjects. For each simulated subject, we computed the valence judgment that subject would produce applying ML to its distributional probability judgments. The resulting judgments were analyzed as the valence judgments in the experiments with real data, yielding a distribution of simulated valence judgments. This procedure was iterated 100 times for each set of cues. In addition, we used the same procedure to generate simulations using softmax (instead of ML). The softmax decision function is widely used in the literature (Daw & Doya, 2006), and might be explained by resource-limited sampling (Vul, Goodman, Griffiths, & Tenenbaum, 2014).

Comparison between simulated and observed valence judgments. We used the Hotelling T^2 test to compare the observed valence judgments and the simulated valence judgments. In these simulations, each participant yields only one value, instead of seven values; therefore, in the Hotelling T^2 we used the variance instead of a 7×7 covariance matrix. When we compared the patterns of results obtained with different tasks, we tested whether the valence judgments averaged by bin across participants were significantly different from the distribution of the distributional probability judgments. In keeping with this approach, here we tested whether the mean of the observed valence judgments was significantly different from the distribution of the simulated valence judgments.

Linking Distributional Probability Judgments and Pointwise Probability Judgments

Even when participants are asked about the likelihood of an emotion (or of an event), there are differences in the ways a judgment of probability can be elicited. Two strategies that have been used in the prior literature consist in 1) specifying one given emotion or event, and asking participants how likely that emotion

or event is in isolation ("pointwise"); or 2) specifying the full set of possible emotions or events, and asking participants to produce a probability distribution over the set ("distributional").

In the case of the "pointwise" judgments, one possible concern is that inquiring about the likelihood of one specific emotion would lead participants to imagine contexts that can lead to that emotion, overestimating how likely those contexts are. This phenomenon can be modeled within a probabilistic framework by suggesting that participants marginalize over the possible contexts:

$$P(\mathbf{e}\,|\,\mathbf{x}) = \int_{\mathbf{c}} P(\mathbf{e}\,|\,\mathbf{c}, \mathbf{x})P(\mathbf{c})d\mathbf{c} \quad (9)$$

and that the estimate of $P(\mathbf{c})$ is susceptible to contexts that are highly available to the participants. This would be the case, for instance, if participant used kernel density estimation to estimate $P(\mathbf{c})$. Imagining contexts that can lead to the emotion given would make them more available, leading to an increase in the estimated probability $P(\mathbf{c})$ for those contexts.

Importantly, for contexts that are already very likely, the increase in probability would be smaller (they were already likely to begin with), while for contexts that are less likely, the increase in probability would be greater. As a consequence, distributions measured by asking participants to produce pointwise judgments could be higher entropy and closer to uniform. We tested this prediction as well as two other predictions of this hypothesis.

First, if participants compute $P(\mathbf{e}|\mathbf{x})$ marginalizing over possible contexts, then when we change the set of possible contexts (even without specifying one specific context), we would expect their estimates of $P(\mathbf{e}|\mathbf{x})$ to change accordingly. To test this prediction, we showed participants the subset of stimuli from (Aviezer et al. 2012) and told participants that the facial expressions were produced in a tennis game at the end of a rally (without disclosing whether the player had won or lost). We predicted that while in naturalistic settings high-intensity emotions are usually negative, in the context of a tennis game they are similarly likely to be negative or positive. If participants estimate $P(\mathbf{e}|\mathbf{x})$ marginalizing over context, we predicted we would see a shift from a more unimodal distribution toward a more bimodal distribution when participants are told that the context is that of a tennis game.

Second, if the distribution for $P(\mathbf{e}|\mathbf{x})$ obtained with the pointwise judgments is more uniform because of inflated estimates of $P(\mathbf{c})$, if we analyzed participants' pointwise judgments of $P(\mathbf{e}|\mathbf{x})$ without normalizing the distribution to 1, we should observe that the probabilities are at least as high as the ones observed in the distributional judgments of $P(\mathbf{e}|\mathbf{x})$, and they should sum to more than 1. We tested this possibility reanalyzing the pointwise judgment data without normalizing the responses to sum to 1.

Results

Pilot

In a pilot study, we asked participants ($n = 190$) to label the face images with up to two words describing the emotion experienced by the character. Participants overwhelmingly rated the images as reflecting a negative emotion, regardless of whether the stimulus was that of a player winning or losing a point (Figure 2A). This finding is in line with the observations in the article by Aviezer et al. (2012).

Experiment 1

In Experiment 1, we measured emotion judgments made on the basis of facial expressions. We recruited three groups of participants ($n = 160, 135, 200$, respectively). Data from participants who failed the control trials were discarded prior to the analysis, leaving data from $m = 156, 119$, and 136 participants respectively.

The first group of participants was asked to perform judgments of emotion valence, the second group was asked to perform pointwise judgments of the probability of emotion valence, and the third group was asked to perform distributional judgments of the probability of emotion valence (see Method). The judgments of individual participants were used to compute a distribution over emotion valence for each of the three ways to elicit judgments.

If these different ways of probing the distribution $P(\mathbf{e}|\mathbf{x})$ are interchangeable, we should obtain the same results in all three experiments. In stark contrast with this prediction, we observed different distributions in the first, second, and third group of participants (Figure 2B, top).

We used Hotelling's T-squared test to assess the significance of the differences between the distributional probability ratings, and the means of the valence ratings and the pointwise probability ratings, respectively. The distributional probability ratings were significantly different from the mean of the valence ratings (Hotelling $T^2(7, 129) = 38.31$, $p < .0001$), and from the mean of the pointwise probability ratings (Hotelling $T^2(7, 129) = 112.10$, $p < .0001$). With the same stimuli, different types of tasks yielded different distributions.

The differences between the valence judgments and the distributional probability judgments were smaller than the differences between these two tasks and the pointwise probability judgments. Despite this, they were still significant—the Hotelling T2 test is a sensitive test, and we had good statistical power thanks to a relatively large number of participants. Whether or not valence judgments and distributional probability judgments may be used interchangeably depends on the size of the difference between their distributions, and on whether it is sufficiently small to lead to equivalent conclusions. For these reasons, interchangeable use of the two tasks is sound only when distributions for both tasks are known, and the size of the differences are known to be inconsequential for the conclusions reached.

Experiment 2

In Experiment 2, we measured emotion judgments made on the basis of the context that caused the emotion using the following sentence: "A professional tennis player just won a very important point." ("won" was replaced with "lost" with 50% probability). We recruited three groups of participants ($n = 150, 150, 150$, respectively). Data from participants who failed the control trials were discarded prior to the analysis, leaving data from $m = 129$, 113, and 114 participants respectively.

As in Experiment 1, the participants in the three groups performed different types of tasks, producing valence ratings (first group), pointwise probability ratings (second group), and distributional probability ratings (third group; Figure 2, bottom). Data were analyzed separately depending on whether the context specified that the player won or lost the point. We used Hotelling's T-squared test to assess the significance of the differences between

the distributional probability ratings, and the means of the valence ratings and the pointwise probability ratings, respectively.

In the case in which the player won the point (Figure 2B, bottom left), estimation of Hotelling T^2 generated a near-singular matrix and yielded negative T^2 values, which are not interpretable. In the case in which the player lost the point (Figure 2B, bottom right), distributional probability ratings were significantly different from the mean of the valence ratings (Hotelling $T^2(7, 51) = 15.55$, $p = .0296$), and from the mean of the pointwise probability ratings (Hotelling $T^2(7, 51) = 107.24$, $p < .0001$). Also in this experiment, the distributions obtained with the three tasks were not the same.

Experiment 3

In Experiment 3, we measured emotion judgments made on the basis of both facial expressions and the context that caused the emotion. We recruited three groups of participants ($n = 150, 320, 150$, respectively). Data from participants who failed the control trials were discarded prior to the analysis, leaving data from $m = 123, 272$, and 109 participants respectively.

As in Experiment 1, the participants in the three groups performed different types of tasks, producing valence ratings (first group), pointwise probability ratings (second group), and distributional probability ratings (third group; Figure 2, middle). Data were analyzed separately depending on whether the context specified that the player won or lost the point. We used Hotelling's T-squared test to assess the significance of the differences between the distributional probability ratings, and the means of the valence ratings and the pointwise probability ratings, respectively.

In the case in which the player won the point (Figure 2B, middle left), distributional probability ratings were also significantly different from the mean of the valence ratings (Hotelling $T^2(7, 50) = 68.94$, $p < .0001$), and from the mean of the pointwise probability ratings (Hotelling $T^2(7, 50) = 30.14$, $p = .0001$). Across all three experiments, the distributions obtained with the three types of tasks were not the same. In the case in which the player lost the point (Figure 2B, middle right), distributional probability ratings were significantly different from the mean of the valence ratings (Hotelling $T^2(7, 45) = 33.79$, $p < .0001$), and from the mean of the pointwise probability ratings (Hotelling $T^2(7, 45) = 77.53$, $p < .0001$).

Cue Integration

To test the extent to which different behavioral tasks generated judgments consistent with a Bayesian cue integration model (see Method), we used the emotion ratings given facial expressions only and the emotion ratings given context information only to compute predicted emotion ratings given both cues (see Figure 3). The accuracy of the predictions was evaluated computing their Pearson correlation with the empirically observed emotion ratings generated by participants given both expression and context information.

Ratings obtained requesting participants to report a full probability distribution best aligned with the predictions generated by the Bayesian cue integration model (Figure 3, right column). Distributional probability ratings obtained with individual cues generated predictions that correlated highly with the distributional ratings obtained with cues both for the "win" context ($r = .89$) and for the "lose" context ($r = .96$). Judgments of emotion valence fared slightly worse, with a correlation of $r = .86$ for the "win" context and $r = .92$ for the lose context. Finally, the pointwise probability ratings aligned well with the Bayesian cue integration model for the "lose" context ($r = .95$), but poorly for the "win" context ($r = .27$), failing to account for the reduction in the probability assigned to negative emotions. Across all tasks, the Bayesian cue integration model had more difficulty predicting cue integration for the "win" context.

These results provide an initial suggestion that distributional probability ratings may yield data that are closer to the representations used by participants for inference. However, much stronger evidence is needed to support this interpretation. In particular, it is necessary to offer hypotheses for why the judgments in the other two tasks show greater deviation from Bayesian cue integration, and to test novel predictions generated by these hypotheses.

Simulations

A possible hypothesis for the difference between distributional judgments of emotion probabilities and judgments of emotional valence is that participants, when asked to produce a single valence judgment given the probability distribution they represent, select the most likely valence (maximum likelihood estimation or MLE). Using a decision theory framework, it can be shown that MLE judgments are optimal for many intuitive utility functions (see Method for details). In MLE, judgments are produced selecting the valence associated with the highest probability; therefore, if the "MLE hypothesis" is correct we would expect a reduction in the tails of the valence judgments distribution as compared to the distribution obtained with distributional probability ratings. We indeed found such a reduction in the tails of valence judgments (Figure 4A, black dots, higher values reflect greater tail reduction). To further test whether these observations could be well described by the MLE hypothesis, we applied MLE to the distributions measured with the distributional probability ratings to compute simulated valence judgments (Figure 4B). The simulations showed the expected tail reduction, and captured accurately the amount of reduction observed in the empirical data (Figure 4A, violin plots), with the exception of inferences about emotions given both expression and context in the "win" condition, where the empirical tail reduction was even greater than predicted by the MLE model.

In addition, we used the Hotelling T^2 test to assess the difference between the mean valence judgments and the distribution of valence judgments simulated by applying MLE to the distributions measured with the distributional probability ratings. Across all conditions, the mean of the observed valence judgments was not significantly different from the distribution of the simulated valence judgments (**e|x**: $T^2(1, 136) = 1.6412$, $p = .2002$; **e|c** win: $T^2(1, 56) = 1.6499$, $p = .1990$; **e|c** lose: $T^2(1, 58) = 0.1041$, $p = .7469$; **e|x, c** win: $T^2(1, 57) = 3.3935$, $p = .0655$; **e|x, c** lose: $T^2(1, 52) = 0.1741$, $p = .6765$).

An alternative model that could explain tail reduction is softmax. In fact, MLE can be seen as a particular case of softmax (in the limit for the base of the exponential going to infinity). We generated a second set of simulations applying softmax to the observed distributional probability judgments instead of MLE. Softmax is widely used in the literature on decision making (Daw

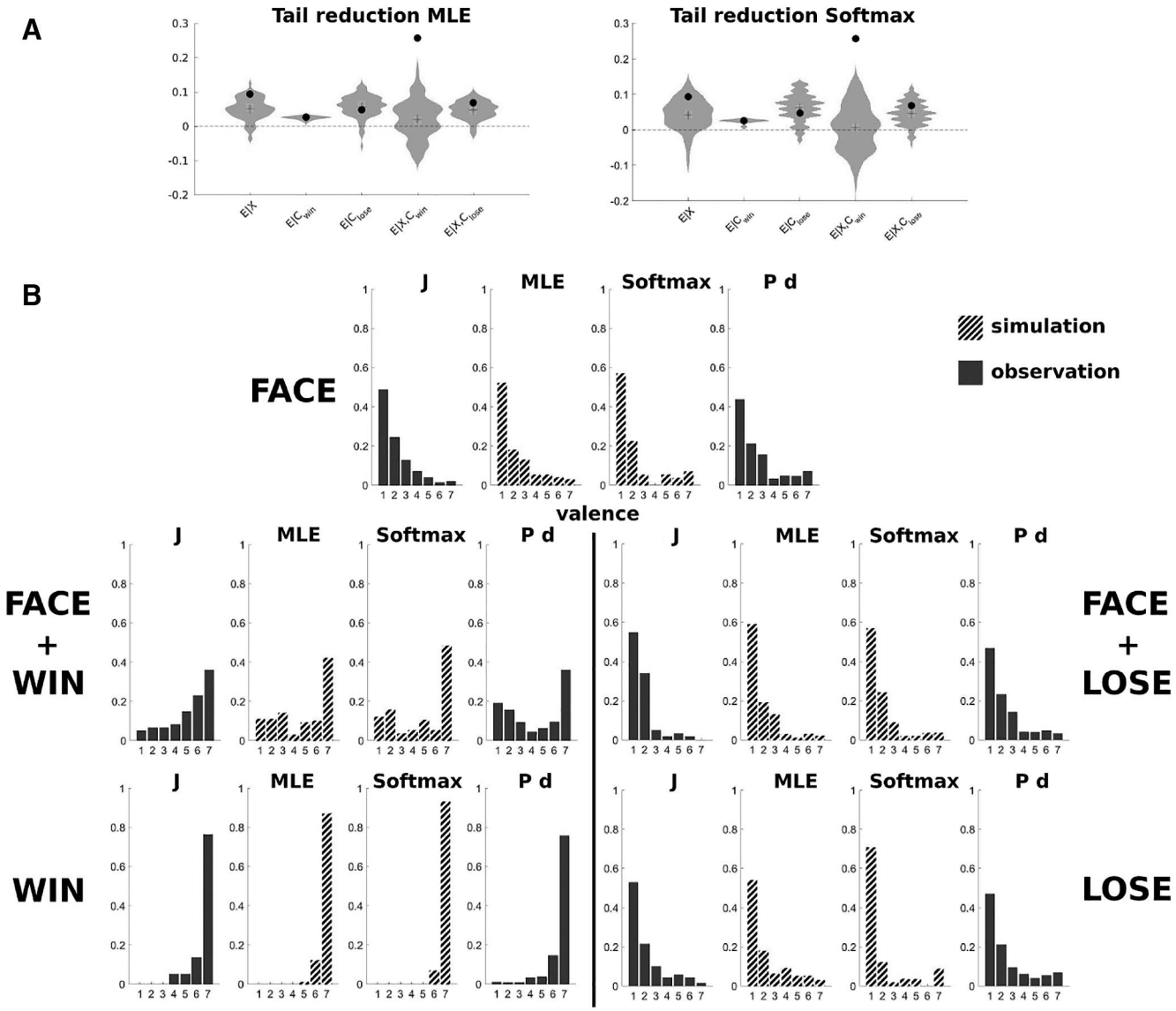

Figure 4. (A) Reduction in the frequency of ratings on the opposite end of the primary mode predicted by the application of ML inference (left) and softmax (right). Distributions obtained from 100 simulations with 100 subjects each is shown in the violin plots; the reduction in the second mode from the observed distributional probability ratings to the observed valence judgments is depicted with black dots. (B) Observed distribution of valence ratings (left), simulated distribution of valence ratings obtained applying ML to the distributional probability judgments (center-left), simulated distribution of valence ratings obtained applying softmax to the distributional probability judgments (center-right), and observed distributional probability judgments (right) given facial expressions (top), expressions and context (middle), and context only (bottom).

& Doya, 2006), and it has been shown that it could be the outcome of resource-limited sampling (Vul et al., 2014). Like MLE, softmax also yielded tail reduction (Figure 4A). We used the Hotelling T^2 test to assess the difference between the mean valence judgments and the distribution of valence judgments simulated by applying softmax to the distributions measured with the distributional probability ratings. The simulated and observed distributions were significantly different only in one condition (**e|c** win: $T^2(1, 57) = 33.4546$, $p < .001$); in all other conditions, they were not significantly different (**e|x**: $T^2(1, 136) = 2.3304$, $p = .1269$; **e|c** lose: $T^2(1, 58) = 1.0904$, $p = .2964$; **e|x, c** win: $T^2(1, 57) = 2.0965$, $p = .1476$; **e|x, c** lose: $T^2(1, 52) = 0.0114$, $p = .9149$). MLE and softmax were comparable in terms of their accuracy at generating valence judgments from distributional probability judgments.

Experiment 4

A possible hypothesis for the difference between distributional judgments of emotion probabilities and pointwise judgments of emotion probabilities is that participants automatically integrate

information about context in their inferences about emotions. In the pointwise judgments, fixing a single emotion would increase the availability of contexts that lead to that emotion, leading to a greater weighting of those contexts in the inference of how likely that emotion is. This hypothesis generates two novel predictions.

First, providing generic information about context might shift ratings of the probability of different emotions given a facial expression. In particular, we hypothesized that in naturalistic settings, facial expressions such as those used in this experiment tend to occur more frequently as a consequence of negative emotions than as a consequence of positive emotions (as suggested also by participants' ratings in the pilot study). However, in a tennis match, these types of facial expressions are far more likely to occur also as a consequence of positive emotions. If participants automatically integrate information about the context in their emotion inferences, we would expect that telling participants that a facial expression was made by a tennis player in a game should affect the participants' probability judgments, making positive emotions more likely. We tested this hypothesis in Experiment 4. Participants ($N = 200$, of whom $N = 153$ successfully completed the control trials and thus were analyzed) were shown a facial expression and were told that it had been produced by a tennis player in a game. Participants were asked to produce distributional probability judgments. As hypothesized, participants rated the positive emotions as more likely compared to when they were not given any contextual information (Figure 5A).

This evidence suggests that participants take into account automatically contextual knowledge when making emotion inferences. However, to further test the view that this mechanism accounts for the higher-entropy distributions observed in the pointwise distributional ratings, we sought to identify a signature of context-dependent effects within the data obtained from the pointwise distributional task.

If the valence fixed in the pointwise distributional led to overestimating the probability of contexts that can lead to that valence, we should observe an overall inflation of probabilities in the pointwise distributional ratings. To test this prediction, we reanalyzed the data obtained with pointwise distributional rating, now without normalizing the ratings to yield a probability distribution that sums to 1. In line with the prediction, the raw pointwise probability ratings were inflated across the board (Figure 5B).

Discussion

In this article, we have investigated whether representations of emotions are "transparent" or "opaque" by testing whether different behavioral tasks commonly used to probe emotion representations yield the same results. Behavioral ratings obtained with different tasks were very different (see Figure 2), demonstrating that emotion representations are opaque. The differences between the distributional probability judgments and the pointwise probability judgments were greater than the differences between the distributional probability judgments and the valence judgments.

We have introduced two criteria that can be used to test whether behavioral judgments obtained with a given emotion task might reflect the participants' internal representations of emotions. We have applied the criteria to judgments of emotions given facial expressions and context. Our results indicate that, among the type of judgments tested, distributional probability judgments are the most promising candidate to reflect internal representations of

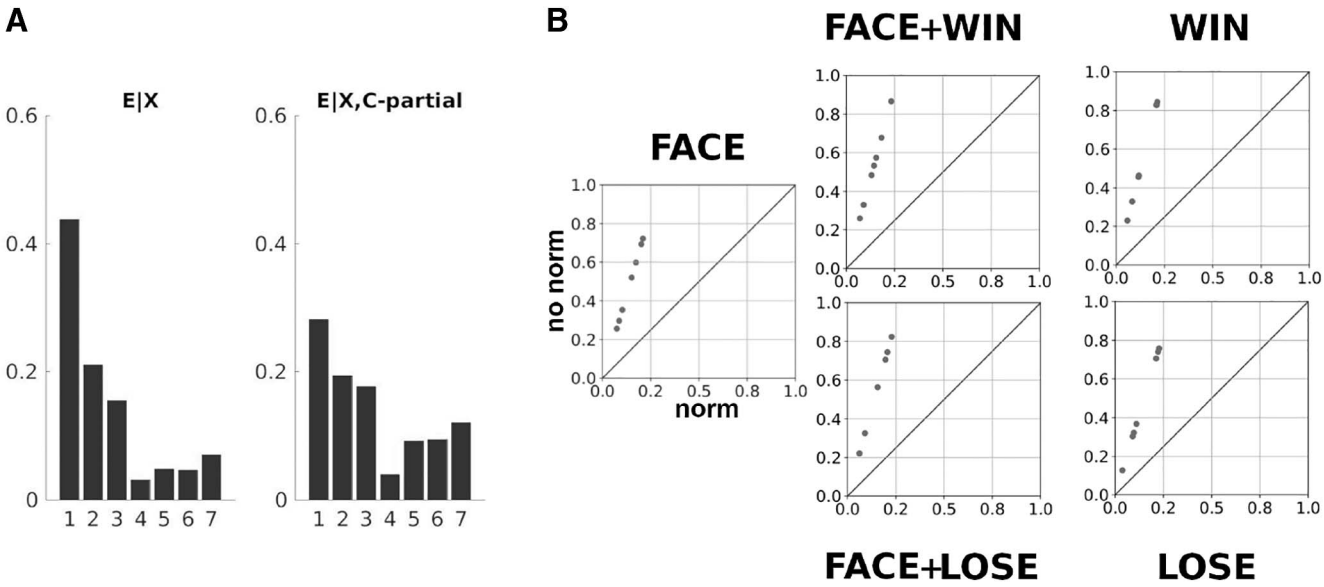

Figure 5. (A) Probability ratings (distributional) for emotion given expression ($E|X$) in the absence of context information, and for emotion given expression and context information specifying that the character experiencing the emotion is a tennis player at the end of a rally, but without specifying whether s/he won or lost the rally ($E|X$, C-partial). Even without specifying whether the player won or lost, knowledge that the context is a tennis game alters the distribution of emotions. (B) Scatter plot comparing pointwise probability ratings normalized to sum to 1 (x-axis) and non-normalized (y-axis). Dots above the $y = x$ diagonal indicate inflated probability judgments (for which the sum across all valences adds to more than 1).

emotions. More research is needed to replicate and strengthen this conclusion.

Ideally, other researchers who study emotion inferences will be able to use the criteria we have introduced to mitigate the challenges posed by opacity, testing whether the tasks they plan to use are likely to reflect internal representations of emotions. In cases in which this approach would be too complex and costly, we would provisionally favor tasks eliciting distributional probability judgments.

The distinction between opaque and transparent representations is related to but different from the notion of "epistemic transparency" (Boghossian, 1994). In fact, even if participants had access to their own representation of the probability distribution over different emotions, the "valence judgment" task still required them to generate as response a single valence value, leading to a discrepancy between the distribution of judgments and the represented distribution over emotions.

This distinction is also different from the distinction between "competence" and "performance" (see Hymes (1972)), as the discrepancy between performance and competence is usually thought to reflect "suboptimal" performance deriving from cognitive limitations, while behavioral judgments might deviate from the underlying representations due to task demands in ways that are "optimal" (e.g., we have discussed in the Method section the optimality of Maximum Likelihood given certain choices of utility functions).

The finding that different behavioral tasks commonly used in the field yield radically different emotion ratings indicates that great caution should be exerted in using interchangeably the ratings and distributions obtained from different tasks, not only in the case of emotion inferences but also in other cognitive domains.

Considering the present results obtained from different tasks and experimental manipulations, we propose that given limited information (such as a facial expression), participants infer emotions by generating plausible situations, and weighting each emotion by the probability of the situations that might cause it and are compatible with what is known (i.e., see Equation 9). This perspective accounts for the flexibility with which different emotions can be attributed to the same face. In this view, when participants were given a facial expression and an emotion, they might have oversampled situations that are compatible with both the expression and the emotion, overestimating the probability of that emotion. Finally, when participants were given partial context information (i.e., "the character is a tennis player just after a rally"), this information changed the space of plausible situations, altering the probability assigned to different emotions.

Distributional probability ratings were found to align well with predictions of a Bayesian cue integration model (see Figure 3), and applying a ML decision process to them yielded results closely matching the valence judgments (see Figure 4). These results converge to suggest that asking participants to produce distributional probability ratings might reflect more closely the uncertainties participants represent and use to reason about emotions.

A challenge for the investigation of opaque representations is that behavioral responses in a single task can be explained by multiple possible models of the underlying emotion representations. Changes in the model of the underlying representations can be compensated by appropriate changes in the decision function to produce the same outputs (Einhorn & Hogarth, 1981).

We have attempted to resolve this ambiguity taking advantage of the compositionality of representations, relying on the intuition that operations over representations are not invariant to the decision function. The ML decision function, for example, selects the emotion with highest probability, and thus induces a loss of information about the relative probability of the less likely emotions. In the presence of a bimodal representation, ML can lead to a reduction in the minor mode in the observed behavioral judgments (see Figure 4A)—which can lead to underestimating the contribution of the minor mode for cue integration (see Figure 3). Due to this information loss, there may be no operation on the judgments that can model cue integration. This strategy may not be able to resolve all ambiguity about the nature of representations, but it can restrict the space of possibilities. This approach is not limited to emotion inferences: It could be used in other domains in which judgments are generated thanks to the integration of multiple cues.

Contextual information is known to play a key role for emotion inferences (Aviezer et al., 2012; Barrett et al., 2011; Carroll & Russell, 1996; Hassin et al., 2013; Kayyal, Widen, & Russell, 2015). In this study, contextual information affected emotion inferences not only when it was clearly relevant to interpret facial expressions, informing the participant about whether the player had lost or won a point, but also when it was much more subtle (Figure 5A). Previous work (Barrett & Kensinger, 2010) has demonstrated that when participants are asked to label emotional faces, neutral contexts are automatically encoded in memory. These observations converge to suggest that humans spontaneously process contextual information and use it to update their expectations about the emotions that the people they encounter are likely to experience.

Across all behavioral tasks, Bayesian cue integration generated more accurate predictions for "lose" contexts than for "win" contexts. A possible explanation is that none of the behavioral tasks perfectly reflect the underlying emotion representations, and that the behavioral judgments are closer to the underlying representations for the "lose" context than for the "win" context. Alternatively, it is possible that Equation 1 does not perfectly reflect cue integration—for example, distinct mechanisms might be in play for the integration of facial expressions with incongruent contextual information.

In line with this possibility, ML applied to the distributional probability ratings led to a reduction in the minor mode of the distribution that matched closely the observed valence judgments in all cases, with the exception of the integration between facial expressions and a "win" context (Figure 4A), for which the empirical reduction in the minor mode exceeded that predicted by ML. In cases with conflicting information, integration might not be an appropriate framework for the decision process, and models of cue competition might generate more accurate predictions. An additional possibility is that including mental states such as beliefs and desires might be critical to account for the pattern of results observed in conditions in which information from different cues appears to be in conflict. Recent evidence (Wu, Baker, Tenenbaum, & Schulz, 2018) lends support to this possibility, indicating that richer models that include mental states, outcomes, and emotional expressions can capture the integration of congruent and incongruent cues with comparable accuracies. More generally, developing models that include multiple kinds of mental states is a key direction for future research.

Our results indicate that emotion representations might be more uncertain than it would appear just by analyzing judgments of emotion valence. A decision function like MLE or softmax can reduce the uncertainty of judgments as compared to the distribution that generated them. This occurs in particular in the presence of a minor mode in the distribution, as in the case of ambiguous facial expressions. Correctly recovering the presence of a minor mode is important to model inferential mechanisms such as cue integration. In fact, if an observer represents the likely emotions given a facial expression with distribution that has a minor mode, less novel contextual evidence would be needed to produce a shift in the observer's judgments.

One key component that will need to be included in future models is arousal (Barrett, 1998; Russell, 1980). Another limitation of the present study is that it is restricted to emotions—observers make a variety of inferences about other mental states including cognitive states (i.e., "confused") or states engaging theory of mind (i.e., "sympathy"). Future research can build on the conclusions reached in this article to test more complex models, choosing carefully which behavioral tasks to use.

Future work can attempt to estimate jointly the underlying representations and the decision function searching among a space of decision functions. Compositionality can be exploited following the strategy proposed in this article to search within a hypothesis space of decision functions. For example, the decision function can be modeled as generating judgments with a probability proportional to $p(E|cues)^b$, a family of decision functions that produces the underlying representations when $\beta = 1$, and converges to MLE when $b \to \infty$ (see Gershman, Pouncy, and Gweon (2017) for an example in a different context).

References

Adolphs, R. (2010). Emotion. *Current Biology, 20,* R549–R552.

Aviezer, H., Trope, Y., & Todorov, A. (2012). Body cues, not facial expressions, discriminate between intense positive and negative emotions. *Science, 338,* 1225–1229.

Barrett, L. F. (1998). Discrete emotions or dimensions? The role of valence focus and arousal focus. *Cognition & Emotion, 12,* 579–599.

Barrett, L. F., & Kensinger, E. A. (2010). Context is routinely encoded during emotion perception. *Psychological Science, 21,* 595–599.

Barrett, L. F., Mesquita, B., & Gendron, M. (2011). Context in emotion perception. *Current Directions in Psychological Science, 20,* 286–290.

Berger, J. O. (2013). *Statistical decision theory and Bayesian analysis*. New York, NY: Springer Science & Business Media.

Boghossian, P. A. (1994). The transparency of mental content. *Philosophical perspectives, 8,* 33–50.

Calder, A. J., Keane, J., Manly, T., Sprengelmeyer, R., Scott, S., Nimmo-Smith, I., & Young, A. W. (2003). Facial expression recognition across the adult life span. *Neuropsychologia, 41,* 195–202.

Carroll, J. M., & Russell, J. A. (1996). Do facial expressions signal specific emotions? Judging emotion from the face in context. *Journal of Personality and Social Psychology, 70,* 205–218.

Darwin, C., & Prodger, P. (1998). *The expression of the emotions in man and animals*. New York, NY: Oxford University Press.

Daw, N. D., & Doya, K. (2006). The computational neurobiology of learning and reward. *Current Opinion in Neurobiology, 16,* 199–204.

Einhorn, H. J., & Hogarth, R. M. (1981). Behavioral decision theory: Processes of judgement and choice. *Annual Review of Psychology, 32,* 53–88.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion, 6,* 169–200.

Ekman, P., & Oster, H. (1979). Facial expressions of emotion. *Annual Review of Psychology, 30,* 527–554.

Gershman, S. J., Pouncy, H. T., & Gweon, H. (2017). Learning the structure of social influence. *Cognitive Science, 41,* 545–575.

Hassin, R. R., Aviezer, H., & Bentin, S. (2013). Inherently ambiguous: Facial expressions of emotions, in context. *Emotion Review, 5,* 60–65.

Hymes, D. H. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics: Selected readings* (Part 2, pp. 269–293). Harmondsworth, England: Penguin.

Kayyal, M., Widen, S., & Russell, J. A. (2015). Context is more powerful than we think: Contextual cues override facial cues even for valence. *Emotion, 15,* 287–291.

Körding, K. P., & Wolpert, D. M. (2006). Bayesian decision theory in sensorimotor control. *Trends in Cognitive Sciences, 10,* 319–326.

Ong, D. C., Zaki, J., & Goodman, N. D. (2015). Affective cognition: Exploring lay theories of emotion. *Cognition, 143,* 141–162.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39,* 1161–1178.

Saxe, R., & Houlihan, S. D. (2017). Formalizing emotion concepts within a Bayesian model of theory of mind. *Current Opinion in Psychology, 17,* 15–21.

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive science, 38,* 599–637.

Wu, Y., Baker, C. L., Tenenbaum, J. B., & Schulz, L. E. (2018). Rational inference of beliefs and desires from emotional expressions. *Cognitive Science, 42,* 850–884.

Wu, Y., & Schulz, L. E. (2018). Inferring beliefs and desires from emotional reactions to anticipated and observed events. *Child Development, 89,* 649–662.

Received May 16, 2019
Revision received August 13, 2019
Accepted August 13, 2019 ■