nature human behaviour

ARTICLES

https://doi.org/10.1038/s41562-022-01390-y

Check for updates

Amplification in the evaluation of multiple emotional expressions over time

Amit Goldenberg[1,8] ✉, Jonas Schöne[2,8], Zi Huang[1], Timothy D. Sweeny[3], Desmond C. Ong[4,5], Timothy F. Brady[6], Maria M. Robinson[6], David Levari[1], Jamil Zaki[7] and James J. Gross[7]

Social interactions are dynamic and unfold over time. To make sense of social interactions, people must aggregate sequential information into summary, global evaluations. But how do people do this? Here, to address this question, we conducted nine studies ($N = 1{,}583$) using a diverse set of stimuli. Our focus was a central aspect of social interaction—namely, the evaluation of others' emotional responses. The results suggest that when aggregating sequences of images and videos expressing varying degrees of emotion, perceivers overestimate the sequence's average emotional intensity. This tendency for overestimation is driven by stronger memory of more emotional expressions. A computational model supports this account and shows that amplification cannot be explained only by nonlinear perception of individual exemplars. Our results demonstrate an amplification effect in the perception of sequential emotional information, which may have implications for the many types of social interactions that involve repeated emotion estimation.

Imagine a friend telling you a story about a situation that made them very angry, or a colleague describing an exciting new idea. As their narrative unfolds, they express emotions of different intensities over time. These expressions are then summarized into a more general understanding of the person's degree of anger or excitement, a process referred to as ensemble coding[1–3]. Ensemble coding—particularly the extraction of averages—occurs without intention and plays an important role in social judgements[4,5]. However, even though exposure to sequentially unfolding emotional expressions is ubiquitous and consequential to social interactions, the vast majority of emotion research has focused on people's ability to evaluate single emotional occurrences rather than sequences of expressions, leaving the question of how people aggregate multiple sequential emotional representations under-researched.

This gap in our understanding is problematic because the aggregation of sequential expressions of differing emotional intensities may favour stronger emotional responses. This is because people attend and respond to faces expressing emotion faster than those not expressing emotion[6,7], and they also seem to find it more difficult to detach their attention from more emotional faces[8]. Increased attention to emotional facial expressions is also associated with stronger visual working memory[9–12] and long-term memory[13,14]. Differences in the quality of memory are seen as a function of whether the emotion expressed was positive or negative[10,11,13,15,16], with what seems to be stronger visual working memory in response to negative stimuli[12].

How might the priority given to more versus less emotional expressions affect the evaluation of expressions appearing in sequence? One clue is that people often remember only a subset of items within a sequence[17,18]. Given the primacy that emotional faces have in memory, salient expressions may be more likely to be preferentially remembered in long sequences, which should impact people's overall evaluation of the sequence's average intensity[19]. Furthermore, as longer sequences are statistically more likely to contain stronger emotional responses than shorter sequences, there are more opportunities to remember more emotional expressions, which should displace less salient expressions in working memory. Therefore, simply as a matter of sampling, amplification seems likely to be stronger for longer sequences. Finally, given that visual working memory is enhanced for negative faces, it is possible that when people are asked to average a sequence of emotions immediately after they occur, amplification is larger for negative emotions than for positive ones.

The handful of studies that have examined ensemble coding of emotions appearing in sequence have mostly focused on the tendency to overweight recent faces[1,20] or to accurately perceive certain frames within a sequence[21], but they have not explored the tendency for amplification. However, indirect support for the possibility of amplification in the evaluation of emotional sequences comes from ensemble coding research that is not specifically focused on emotions. For example, in two studies that examined people's ability to evaluate the average size of circles appearing in sequence, people overweighted larger circles—which are more salient in perception—in estimating the mean size[20,22]. Furthermore, recent findings suggest that people may overestimate the emotional intensity of crowds due to increased attention to emotional faces[8,23], and may evaluate dynamic emotional expressions as more intense than static ones[24–26].

Perhaps the strongest indirect support for the idea that people may be biased in estimating sequences of emotion comes from the peak–end rule, originally introduced by Kahneman and colleagues[27–30]. According to the peak–end rule, people evaluate subjective affective experiences by averaging the peak and the end of the experience[27,31]. While the peak–end rule is broadly consistent with amplification in emotional sequences, there are also clear differences. First, the peak–end rule has been examined only by evaluating people's subjective experience, rather than social

[1]Harvard Business School, Harvard University, Boston, MA, USA. [2]Department of Experimental Psychology, University of Oxford, Oxford, UK. [3]Department of Psychology, University of Denver, Denver, CO, USA. [4]Department of Information Systems and Analytics, National University of Singapore, Singapore, Singapore. [5]Institute of High Performance Computing, Agency for Science, Technology and Research, Singapore, Singapore. [6]University of California, San Diego, CA, USA. [7]Department of Psychology, Stanford University, Stanford, CA, USA. [8]These authors contributed equally: Amit Goldenberg, Jonas Schöne. ✉e-mail: agoldenberg@hbs.edu

perceptions. Second, the peak–end rule focuses only on the peak and the end of a sequence, while we wish to argue that memory promotes more salient expressions in general, not only the peak expressions, which should lead to different predictions in the degree of amplification (see the Supplementary Information for direct comparisons). Finally, Kahneman and colleagues argue that the length of the sequence should not affect its affective evaluation (duration neglect)[28], while we argue that length matters.

We conducted a set of nine studies with the goal of examining the occurrence of amplification in the evaluation of emotional sequences, identifying its driving mechanisms and providing evidence for its occurrence in more natural settings. The first set of studies (Studies 1–4) was designed to detect the existence of amplification in the evaluation of emotional sequences, addressing potential challenges to the main finding in each study. These studies had three main pre-registered hypotheses (https://osf.io/ag8nv). Our first hypothesis was that the participants would estimate the average emotional intensity expressed in a sequence of emotions as more intense than it actually is (amplification effect) (H1). Our second hypothesis was that amplification would be stronger for longer sequences (H2). Our third hypothesis was that we would see (slightly) increased amplification for sequences of negative emotions[8] (H3). The second set of studies (Studies 5–8) was designed to test whether enhanced memory for emotional faces was driving the sequential amplification effect. Finally, in Study 9, we addressed limitations in external validity and examined the occurrence of amplification in the evaluation of natural videos depicting emotional stories using the Stanford Emotional Narratives Dataset (SEND)[32].

Results

Analysis for all studies was conducted in R using mixed models for repeated measures. Assumptions of normality (Kolmogorov–Smirnov test) and equal variance (Levene test) were checked for all main analyses (see the Supplementary Information for the full report). In the cases in which these assumptions were violated, we conducted a robust estimation of mixed effects using the package robustlmm[33] and found similar results in all cases (Supplementary Information).

Studies 1–4. The goal of Studies 1–4 was to test the three hypotheses described above (https://osf.io/ag8nv). The basic structure of the task was similar in all of these studies. The participants were exposed to a sequence of 1–12 faces of the same identity expressing different intensities of either anger or happiness (but not both). The participants were then asked to evaluate the average emotional intensity expressed in the sequence by morphing a face bearing the same identity (Fig. 1a). We modified the location and starting point of the scale in Studies 1–4 (see the detailed description in the Methods). To measure amplification (H1), we conducted a mixed model analysis of repeated measures, comparing the actual mean emotion expressed in each set with the participants' estimated mean emotion. We added by-participant and by-face-identity random intercepts. Supporting the first hypothesis, the estimated mean emotion was higher than the actual mean emotion in all of the studies (Table 1; Study 1: $b=0.75$; $t(9{,}704)=4.35$; $P<0.001$; $R^2=0.05$; 95% confidence interval, (0.41, 1.09); Study 2: $b=1.40$; $t(8{,}964)=8.50$; $P<0.001$; $R^2=0.05$; 95% confidence interval, (1.07, 1.72); Study 3: $b=1.31$; $t(10{,}012)=7.71$; $P<0.001$; $R^2=0.04$; 95% confidence interval, (0.98, 1.65); Study 4: $b=3.95$; $t(9{,}395)=22.22$; $P<0.001$; $R^2=0.07$; 95% confidence interval, (3.60, 4.30)). The second and third hypotheses were tested with a single model. We created a difference score between the participants' estimation of the sequence average and the actual average, and we used both the sequence length (H2) and the valence (H3) as predictors. We used the same random intercepts as in the previous model. Our results suggest that an increase in the number of expressions in the sequence led to an increase in amplification in all of the studies (Table 1; Study 1: $b=0.33$; $t(4{,}820)=8.91$; $P<0.001$; $R^2=0.14$; 95% confidence interval, (0.24, 0.40); Study 2: $b=0.30$; $t(4{,}428)=8.45$; $P<0.001$; $R^2=0.13$; 95% confidence interval, (0.22, 0.36); Study 3: $b=0.36$; $t(4{,}988)=10.36$; $P<0.001$; $R^2=0.12$; 95% confidence interval, (0.29, 0.43); Study 4: $b=0.19$; $t(4{,}965)=5.36$; $P<0.001$; $R^2=0.12$; 95% confidence interval, (0.12, 0.26)). Further analysis conducted on sequence length (reported in detail in the Supplementary Information) indicated that amplification was evident only in sequences larger than four to six expressions, depending on the study, which is congruent with research on working visual memory. The participants' accuracy in evaluating shorter sequences provides additional support that the amplification effect occurs due to the aggregation of long sequences. Finally, we found an overall increased amplification for negative sequences compared with positive sequences in two of these four studies, and as expected, even in the studies in which this effect was found, it was relatively weak (Table 1; Study 1: $b=0.56$; $t(4{,}823)=2.20$; $P=0.02$; $R^2=0.13$; 95% confidence interval, (0.23, 1.07); Study 2: $b=0.91$; $t(4{,}420)=3.74$; $P<0.001$; $R^2=0.14$; 95% confidence interval, (0.43, 1.39); Study 3: $b=0.40$; $t(4{,}987)=1.68$; $P=0.09$; $R^2=0.12$; 95% confidence interval, (−0.06, 0.88); Study 4: $b=-0.77$; $t(4{,}967)=-3.03$; $P<0.001$; $R^2=0.075$; 95% confidence interval, (−1.28, −0.27)).

Study 5. The goal of Study 5 was to validate two previously reported findings related to the importance of memory to sequence evaluation (pre-registration: https://osf.io/j4kqz/). Study 5 included two blocks. In the first block (30 trials), the participants were asked to evaluate the average emotional intensity expressed in sequences of eight expressions similar to Study 1. In the second block (20 trials), the participants saw sequences of eight expressions and then performed a memory test by choosing between a true target expression that appeared in the sequence and a false target expression that did not appear in the sequence ($N=150$; men, 51; women, 98; other, 1; age: mean = 38.35, s.d. = 12.37). Looking first at our first block, we were able to replicate the findings of Studies 1–4 (Supplementary Information). We then turned to evaluating the participants' tendency to correctly choose the true target face in the memory test. We first confirmed that the participants were more successful in the

Fig. 1 | Structure and results of Studies 1–4. a, The structure of the amplification task used in Studies 1–4. The participants saw a sequence of 1–12 facial expressions, expressing different degrees of either anger or happiness, that appeared on the screen for one second ((i) represents one expression in the sequence). Between each expression, the participants saw a fixation cross for 400–600 ms (ii). The participants were then asked to move the mouse to the left of the line to begin the evaluation stage (iii). They were then asked to evaluate the average emotion expressed by these expressions by adjusting the intensity of a single morphed face (1–50, (iv)). **b**, Two samples (left, NimStim; right, Radboud) of three facial expressions from the neutral-to-angry scale (top) and from the neutral-to-happy scale (bottom) that were used in the studies. Values of 25 and 50 correspond to 50% and 100% intensities in our morph range, respectively. **c**, A summary of the results of the comparison between estimated and actual average ratings of the sequence in Studies 1–4 ($N=377$). The analysis was done using mixed models (*t*-test). The *x* axis represents the number of facial expressions in the sequence. The *y* axis represents the difference between the participants' estimation of the average sequence and the actual average. The data are presented as mean values ± confidence intervals. The green dotted line represents the average amplification across studies. The red dotted line represents the actual facial expression mean. Male and female images in **b** reproduced with permission from refs. [40,41], respectively.

memory task when the true target face corresponded to a face that occurred later in the sequence (Supplementary Information). We then examined whether the emotional intensity of the target facial expression predicted the probability of remembering the expression. We constructed a generalized linear mixed model in which we used the emotional intensity of the true target expression as the predictor and whether the participants chose this expression correctly or not as the dependent variable. We added a covariate to the model of the distance between the false and the true target, as this distance is likely to affect the participants' ability to remember. We also

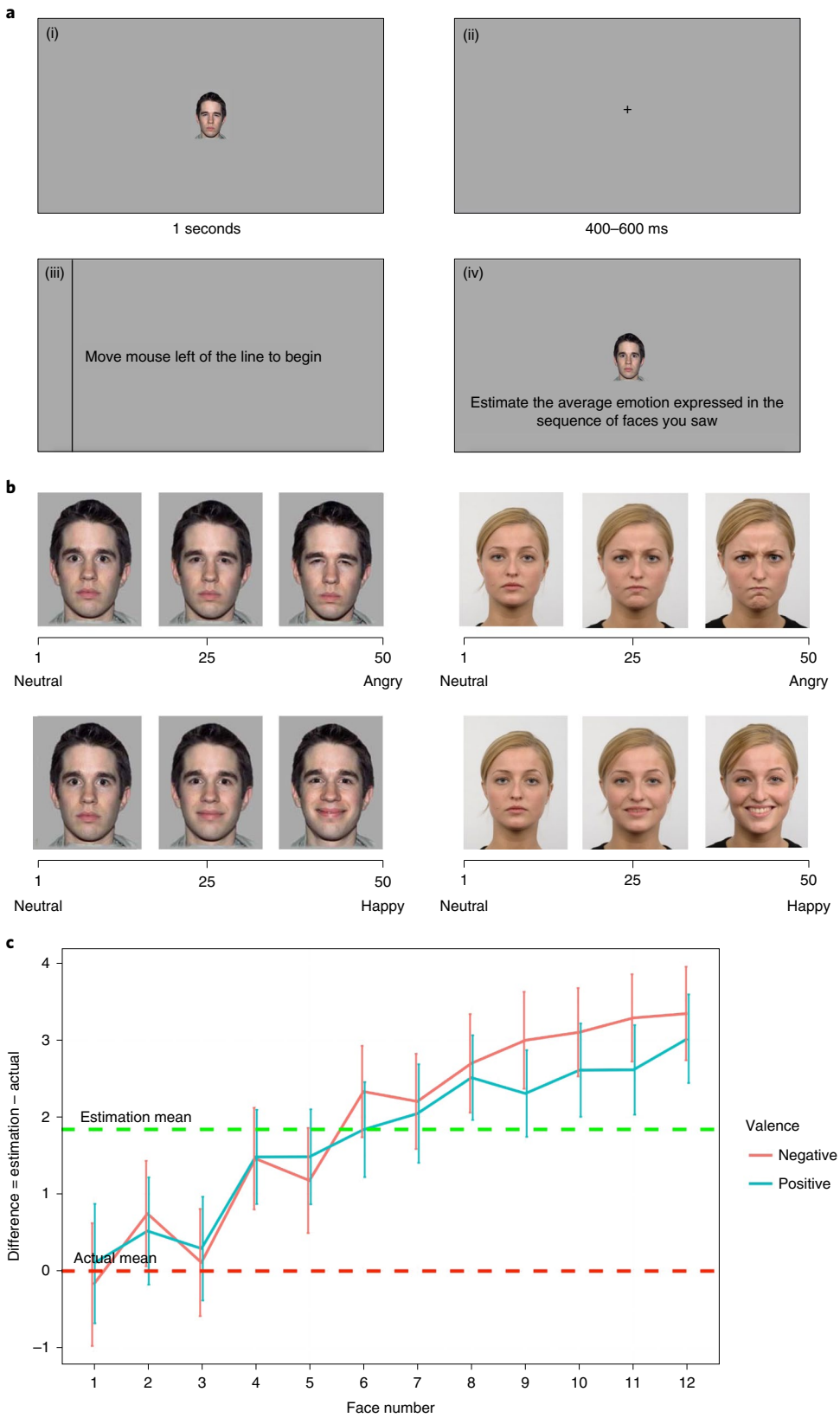

Table 1 | Summary of the results of Studies 1–4 divided by the three hypotheses

Hypothesis	Study	*b* (confidence interval), (s.e.)	*t* (d.f.)	*P*	R^2
H1: general amplification (positive numbers indicate amplification)	1: establishing effect	0.75 (0.41, 1.09), (0.17)	4.35 (9,704)	<0.001***	0.05
	2: replication with a new morph set	1.40 (1.07, 1.72), (0.16)	8.50 (8,964)	<0.001***	0.05
	3: scale starts on the right side	1.31 (0.98, 1.65), (0.17)	7.71 (10,012)	<0.001***	0.04
	4: scale starts with strong intensity	3.95 (3.60, 4.30), (0.17)	22.22 (9,395)	<0.001***	0.07
H2: amplification as a function of sequence length (positive numbers indicate increased amplification with sequence length)	1: establishing effect	0.33 (0.25, 0.40), (0.03)	8.91 (4,820)	<0.001***	0.14
	2: replication with a new morph	0.30 (0.22, 0.36), (0.03)	8.45 (4,428)	<0.001***	0.13
	3: scale starts on the right side	0.36 (0.29, 0.43), (0.03)	10.36 (4,988)	<0.001***	0.12
	4: scale starts with strong intensity	0.19 (0.12, 0.26), (0.03)	5.36 (4,965)	<0.001***	0.073
H3: amplification as a function of sequence valence (positive numbers indicate that negative sequences were stronger than positive sequences)	1: establishing effect	0.56 (0.23, 1.07), (0.26)	2.20 (4,823)	0.02*	0.13
	2: replication with a new morph	0.91 (0.43, 1.39), (0.24)	3.74 (4,420)	<0.001***	0.14
	3: scale starts on the right side	0.40 (−0.06, 0.88), (0.24)	1.68 (4,987)	0.09	0.12
	4: scale starts with strong intensity	−0.77 (−1.28, −0.27), (0.25)	−3.03 (4,967)	<0.001***	0.075

The first hypothesis was that the participants would tend to evaluate the sequence mean as more intense than it actually was. The second hypothesis was that amplification in the evaluation of the sequences would increase with sequence length. The third hypothesis was that amplification would be stronger in negative sequences. The asterisks denote levels of significance.

added a by-individual random intercept and a random intercept of the face identity. The results suggest that the intensity of the facial expression predicted the probability of memory ($b=0.04$; $z=15.23$; $P<0.001$; $R^2=0.10$; 95% confidence interval, (0.37, 0.47)). We conducted additional analysis to make sure that this effect was not driven solely by participants merely choosing the more emotional expression in each trial (Supplementary Information).

Studies 6 and 7. The goal of Studies 6 and 7 (reported only in the Supplementary Information) was to manipulate the tendency to remember certain faces and test its effect on the tendency for amplification. In Study 6 ($N=96$; men, 62; women, 33; other, 1; age: mean = 25.49, s.d. = 7.35), we manipulated salience in memory by taking advantage of the recency bias found in Study 5 and manipulating the intensity of emotions that the participants viewed either at the beginning or at the end of the sequence (pre-registration: https://osf.io/sgbzy/). Our procedure was the same as for Study 1 with two differences. First, the sequence lengths were 2–12 and included only even sequence numbers. Second, each of the 50 trials that the participants completed was divided into a high-intensity end and a low-intensity end. In the high-intensity end trials, the expressions in the first half of the sequence were randomly drawn from only the low-intensity emotions (1–25 on our scale), and the second half were randomly drawn from only the high-intensity emotions (26–50 on our scale). The low-intensity end trials were structured in the opposite manner. We designed the task so that the low-end and high-end trials would mirror each other completely. In addition to providing support for the three hypotheses (Supplementary Information), we examined the difference between the participants' estimation and the actual sequence average as the dependent variable, and the order of high- and low-intensity expressions as the independent variable, including by-participant and by-face-identity random intercepts. The results suggest that in trials in which the high-intensity expressions were presented at the end, the participants' estimations were amplified compared with the actual mean ($b=2.87$; $t(13.64)=6.33$; $P<0.001$; $R^2=0.18$; 95% confidence interval, (1.95, 3.80)). The results also pointed to a significant de-amplification in the low-intensity end condition ($b=-1.31$; $t(13.61)=-2.90$; $P=0.012$; $R^2=0.18$; 95% confidence interval, (−2.24, −0.38)). Importantly, however, despite the fact that recency seemed to have a tremendous effect on evaluation, the amplification in the high-intensity end condition was larger than the de-amplification in the low-intensity end condition (see the Supplementary Information for further analysis). To provide further support for the memory mechanism, in Study 7 we manipulated memory by changing the salience of high- or low-intensity emotions by adding a red square around some of the faces (as reported fully in the Supplementary Information). Taken together, these studies provide strong evidence for memory of more salient expressions as driver of amplification.

Study 8. Differential memory for more emotional (intense or salient) stimuli is one possible mechanism. However, perceptual characteristics of the stimulus space, rather than changes in memory, could also be driving the observed amplification effects. To address this possibility, we used computational modelling to separately quantify the psychophysical similarity between expressions, and we used these similarity data to estimate what biases in memory for ensembles would be expected on the basis of similarity alone. In Study 8, we first empirically tested how people perceived distances between emotional intensities at different points of our emotional scale showing some nonlinearity such that expressions in the middle of the scale were more differentiated than those at the edges of the scale (Fig. 2 and Supplementary Information). We then built on an existing computational model[34] that was designed to simulate ensemble memory with specific attention to nonlinearity in similarity, by comparing three models: (1) a baseline model that only incorporated nonlinearity in similarity (found in our pilot), (2) a recency model that was based on the baseline model but also assumed stronger weight in memory for more recent items and (3) an amplification model that was based on the recency model but added an assumption of increased weight to more emotional expressions. We used the results of Study 6 to compare these three models' fit (see the full description in the Supplementary Information). The results suggest that the amplification model yielded the best fit, providing additional support for amplification over and above recency and nonlinearity.

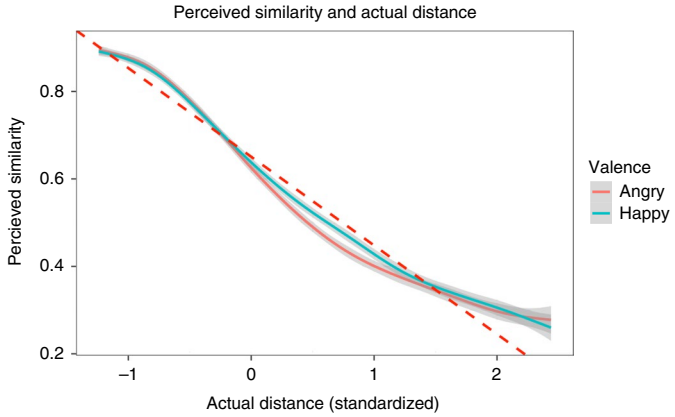

Fig. 2 | Results from the similarity analysis in Study 8 ($N = 100$). The x axis represents the difference between item 1 and item 2. The data are presented as mean values ± standard errors. The analysis was done using mixed models (t-test). As indicated by the changes in the degree of slope, both shorter distances and longer distances were perceived as more similar than distances between these two extremes.

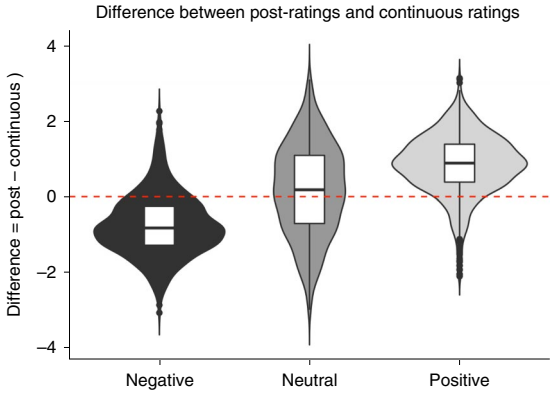

Fig. 3 | The difference between post-ratings and continuous ratings for the three types of videos in Study 9 ($N = 565$): neutral, negative and positive. The data are presented as median values. The boxes represent the interquartile range (IQR). The lines represent the first quartile − 1.5 IQR and the third quartile + 1.5 IQR. The dots represent all participants ouside that range. Comparison between the conditions was done using mixed models (t-test). A positive number indicates that the post-rating was more positive than the average continuous rating, and a negative number indicates that the post-rating was more negative than the continuous rating.

Study 9. In Study 9, we sought to generalize the findings and examine them in more naturalistic interactions. We used data from the SEND[32], in which observers watched and provided emotional ratings of videos of a diverse set of targets telling personal emotional stories. The participants were asked to provide two types of ratings in response to each video: a continuous, real-time evaluation of the degree of negativity and positivity of each video as they unfolded over time, and a global evaluation of the target emotionality after watching the whole video. We then compared the average of the participants' real-time emotional evaluations with their post-video global evaluations. To account for differences that may have been caused by different scales, we treated the difference between the post-rating and the continuous rating of the neutral videos as our baseline comparison. We then conducted a mixed model analysis using the difference between the global evaluation and the average of the real-time continuous evaluation as our dependent variable and the valence of the video as the dependent variable, including by-participant and by-video random intercepts. The results for the neutral videos were not different from zero ($b = 0.16$; $t(191) = 1.39$; $P = 0.17$; $R^2 = 0.52$; 95% confidence interval, (−0.06, 0.39)). For the negative videos, the difference between the post-ratings and continuous ratings was significantly more negative than in the neutral condition ($b = -0.92$; $t(189) = -5.66$; $P < 0.001$; $R^2 = 0.52$; 95% confidence interval, (−1.24, −0.60)). In contrast, and also congruent with the tendency for amplification, the difference between the post-ratings and continuous ratings in the positive videos was significantly more positive than for the neutral videos ($b = 0.70$; $t(189) = 4.87$; $P < 0.001$; $R^2 = 0.52$; 95% confidence interval, (0.42 0.98); Fig. 3). To summarize, the analysis of the SEND videos pointed to an amplification effect that was similar to the one evident using sequences of static emotional expressions in the previous studies.

Discussion

Our evaluations of others' emotions are central to almost any social interaction, and because emotions unfold over time, such evaluations hinge on how we aggregate and evaluate sequential emotional information. The aggregation of others' emotions also impacts crucial decisions. Examples include job interviews, in which candidates are evaluated partly on the basis of their passion for the job, and medical decisions, which are based in part on doctors' perceptions of the degree to which their patients are in pain. In nine studies, we have demonstrated that these sorts of assessments may involve systematic amplification, which is caused by increased memory for more emotional expressions in the sequence.

In some circumstances, amplification might impair optimal decision-making. In many others, however, amplification in the evaluation of emotional sequences might be helpful. Expressing emotions, particularly negative ones, is against the norm in many cultures; this often leads people to try to conceal their emotional expressions[35]. Detecting even minor emotional expressions can be extremely informative regarding others' thoughts, goals and future behaviour. The importance of emotional expressions may be one of the reasons why more emotional expressions are more likely to be remembered. Furthermore, when a face changes back and forth from emotional states to more neutral states, it makes sense to ignore the less emotional expressions and to interpret the more emotional ones as reflecting the true emotion of that person. These tendencies may play a role even when people are asked to average the emotional intensity expressed in a sequence, such as in the present studies. It is therefore important to note that we do not see this amplification as unnatural or necessarily unhelpful, but rather as an important feature of social cognition.

We suspect that our findings may not be unique to emotions but instead might be generalizable to sequential perception of any type of stimulus that is asymmetrical in terms of salience, such that some features are more salient than others. Because much of the world unfolds over time, a large portion of our perception and cognition involves aggregating sequential information into single representations. This is done not only for lower-level features such as the size and orientation of a moving object but also for higher-level cognitions such as determining an individual's competence on the basis of their completion of multiple tasks. Future work should therefore examine amplification in variety of stimuli occurring sequentially, especially for stimuli that may have such asymmetric features. These might include evaluating sequence averages of money (the biggest sums are more salient), diversity (diverse members are more salient[36]) and even morality (immoral behaviour is more salient). Future work should attempt to broaden the insight of amplification to these domains.

Limitations and future directions. Our studies have several noteworthy limitations and leave open questions regarding how

amplification manifests itself in natural social interactions. One main limitation relates to the idea that asking participants to evaluate the emotional average of a sequence is dissimilar to the way such evaluations are done naturally. Research suggests that people both naturally segment continuous stimuli[37] and extract average features from a set even when they are not asked to do so[5,38]. Furthermore, in Study 9, the participants were not asked to evaluate the sequence on average but rather to provide a global evaluation of emotionality, leading to similar results. However, despite these findings, it is possible that sequences are naturally aggregated in a way that differ from the mean.

A second limitation relates to the stimuli and the way they were presented to the participants. In real life, when people estimate emotional expressions unfolding over time, emotions tend to cluster together. One question is how the distribution of emotions in a sequence impacts the evaluation of a person's emotionality. Furthermore, in many of our studies (except Study 9), the participants evaluated sequences of clear facial expressions by white men, while the participants themselves were not experiencing any emotion. People's ability to aggregate multiple emotions is likely to depend on target attributes such as gender, ethnicity, age and culture, as well as the perceiver's emotional state. Further studies should manipulate these aspects and examine their contributions to amplification.

This project reveals an important aspect of social cognition that is central to many social interactions. However, there is no reason to assume that amplification in the evaluation of sequences applies only to emotions. Further understanding of how people integrate sequential information of various kinds is therefore likely to reveal other biases in the way we understand our social world.

Methods

This research was approved by the Human Subject Committee at Harvard University (IRB20-0091). All participants provided informed consent and were compensated for their time.

Stimuli for Studies 1–8. To create the stimulus set, we created facial expression morphs from two face sets. The first set was developed for a recent investigation of ensemble face perception[39] and was based on four exemplar faces of men from the NimStim face set[40] (four men). The second set was developed by us to increase the gender diversity of the faces, using eight exemplar faces (four women and four men) based on the Radboud face sample[41]. Morphing was done using the software Fantamorph. We used the neutral and emotional faces of the set to linearly interpolate 48 morphs (that is, 'morph units') between each actor's neutral expression and that same actor's angry and happy expressions. The morphed sets for each identity were on a scale of 0% (completely neutral) to 100% (completely emotional) in increments of 2% (1–50 scale; Fig. 1b). We conducted a pre-test to confirm that people could track differences in the created morphs, finding that the participants were accurate in identifying the intensity of a single facial expression evaluation (Supplementary Information). Note that our 50 face morphs do not imply that every emotional unit corresponds to a categorically distinct emotion. Additionally, the morphs, while mathematically linearly related, were not necessarily psychophysically linear. We directly addressed issues that relate to potential effects of nonlinearity in Study 8.

Amplification task (used in Studies 1–7). In each trial, the participants first saw a sequence of 1–12 facial expressions with the same identity expressing different intensities of emotion from either a neutral-to-angry (anger condition) or a neutral-to-happy (happiness condition) continuum. Each facial expression in the sequence was presented in the middle of the screen for 1,000 ms on a grey background and was randomly taken from a 1–50 morph (Fig. 1b). Between each expression, the participants saw a fixation cross for a duration between 400 and 600 ms (randomly determined on each trial). The sequence's average expressive intensity was therefore normally distributed around 25.5, which was the middle of the scale, with varied degrees of variance. The valence and the number of expressions presented in each sequence were also chosen randomly in each trial. We did not mix the happy and angry expressions because doing so could undermine our ability to detect an amplification effect: if the participants fixated on one extremely negative and one extremely positive expression, then on average, their estimate of the sequence could seem to be relatively accurate despite the fact that they were biased by emotional intensity in their sampling of expression. After the sequence of facial expressions, the participants were asked to evaluate the average emotion expressed in the sequence. To start the measurement phase, the participants were asked to move their cursor beyond a vertical line, which was located on the left (Studies 1–2 and 5–7) or the right side of the screen (Studies 3–4; Fig. 1a(iii)). Once the pointer crossed the vertical line, a single face bearing a neutral expression was presented on the screen (except for Study 4; see below). The identity of the face in the scale matched that of the faces in the previous sequence. The participants were then asked to move the pointer away from the starting point to modify the facial expression from neutral to emotional. The participants had as much time as they needed to estimate the mean intensity of the emotional sequence. After completing the main task, the participants filled out a short survey that was designed to examine potential moderators for the effect that could be manipulated in future studies (see the Supplementary Information for the full description).

Studies 1–4: establishing amplification. The goal of Studies 1–4 was to test the three hypotheses described above (https://osf.io/ag8nv). Given that recent online studies examining the evaluation of crowds' emotions used 100 participants completing 50 trials[8], we decided on a sample size of 100 participants per study completing 50 trials of the task. In all of the studies in this set, of the 100 participants that completed the task, we removed participants whose average estimation was below 10 or higher than 40, which could occur only if ratings were conducted to finish the task quickly without any regard to the averages. Our final samples were as follows: for study 1, $N=93$ (men, 52; women, 41; age: mean = 28.23, s.d. = 9.54); for Study 2, $N=94$ (men, 35; women, 55; other/did not specify, 4; age: mean = 33.60, s.d. = 11.43); for Study 3, $N=98$ (men, 35; women, 62; other/did not specify, 1; age: mean = 25.78, s.d. = 9.08); and for Study 4, $N=92$ (men, 31; women, 66; other/did not specify, 1; age: mean = 24.67, s.d. = 6.42) (see the breakdown of the participants' ethnicities in the Supplementary Information). All of our participants were recruited through Prolific and received US$2.30 for their participation. Studies 1–4 all had the same basic structure described above, but we modified the location and starting point of the scale to make sure that amplification was not caused by the way emotions were measured. In Studies 1 and 2, the initial expression in the scale was anchored on a neutral expression starting on the left side of the screen: 1 on the scale from 1 to 50. This was done because previous research indicated that the initial location of the scale in ensemble coding tasks led to an anchoring effect, such that estimations were closer on average to that of the initial location[4]. In Study 3, the scale also started from neutral, but the direction of the scale was reversed, from right to left. Finally, in Study 4, the scale was initiated on the left side of the screen, but the starting point of the scale was the most emotional expression. This was done to eliminate the possibility that starting from neutral led the participants to 'overshoot' in their estimation of the average.

Study 5: testing memory based on facial expression intensity. The goal of Study 5 was to validate two previously reported findings related to the importance of memory to sequence evaluation. Our starting sample was 150 participants, but we were left with 136 participants after removing participants on the basis of our pre-registered criteria (see the Supplementary Information for the power analysis). Given that 137 participants would put us under a power of 80%, we then collected an additional 13 participants prior to examining or analysing these data, bringing us to $N=150$ (men, 51; women, 98; other, 1; age: mean = 38.35, s.d. = 12.37). All of our participants were recruited through Prolific and received US$2.30 for their participation. The task included two separate blocks that were always presented to the participants in the same order. The first block was the amplification test block, which was 30 trials long. The participants completed an amplification task similar to that of Studies 1–4 with one difference: the length of the sequence was always eight facial expressions. We kept the sequence length equal for two reasons. First, we wanted our amplification trials to have the same sequence length as the memory trials. Second, we wanted to be able to compare the contribution of the order of each expression in the sequence to the participants' estimation. We hypothesized that we would see amplification in this block (H1) as well as stronger amplification for negative emotions (H2; pre-registration: https://osf.io/j4kqz/). The second block, which was 20 trials long, was designed to examine the participants' ability to remember certain emotional expressions in the sequence. Similar to the first block, each trial started with a sequence of eight expressions, either neutral-to-negative or neutral-to-positive. Following each sequence, the participants saw two target expressions: a true target that appeared in the sequence and a false target that did not appear in the sequence. The false target expressions were chosen in each trial by finding the two expressions in the sequence that had the biggest difference between them and taking the midpoint of that difference. For example, if 5 and 13 had the biggest difference in the sequence, the false target expression would be 9. The two target facial expressions appeared on the screen right after the final fixation cross of the sequence (400–600 ms). The participants had as much time as they needed to make their choice. We hypothesized that people would be more likely to succeed in the memory test in the trials in which the true target expression expressed stronger emotion (H3). Following the task, the participants completed a survey similar to the previous studies (see the Supplementary Information for the full analysis).

Study 6: manipulating recency of strong-intensity emotions. The goal of Study 6 was to examine the effect of memory on the participants' tendency for

amplification by manipulating the intensity of emotions that the participants viewed at either the beginning or the end of the sequence. Our pre-registered hypotheses (https://osf.io/sgbzy/) were that we would see amplification (H1), that amplification would increase with sequence length (H2) and that amplification would be stronger for negative emotions (H3). Finally, we hypothesized that amplification would be stronger for trials in which stronger emotions were presented at the end of the sequence, compared with the beginning (H4). We recruited participants from Prolific in exchange for US$2.30. Our sample was $N = 100$, similar to those of Studies 1–4. Congruent with our pre-registered criteria, we removed four participants for providing average ratings of below 10 or above 40. Our final sample was therefore $N = 96$ (men, 62; women, 33; other, 1; age: mean = 25.49, s.d. = 7.35). Our procedure was identical to that of Study 1 with two differences. First, the sequence lengths were 2–12 and included only even sequence numbers. The was because we wanted to divide each sequence into two halves and manipulate the intensity of each half. The second difference from the tasks in Studies 1–4 was that each of the 50 trials that the participants completed was divided into two conditions: high-intensity end and low-intensity end. In the high-intensity end trials, the expressions in the first half of the sequence were randomly drawn from only the low-intensity emotions (1–25 on our scale), and the second half were randomly drawn from only the high-intensity emotions (26–50 on our scale). The low-intensity end trials were structured in the opposite manner, such that the first half included only high-intensity facial expressions, and the second half only low intensity. We designed the task so that the low-end and high-end trials would mirror each other completely. The order of the high-end and low-end trials was random. Following the task, the participants completed a survey similar to the previous studies (see the Supplementary Information for the full analysis).

Study 7: manipulating salience. The goal of Study 7 was to examine the effect of memory on the participants' tendency for amplification by manipulating the salience of either high- or low-intensity emotions. Salience was manipulated by adding a red square around either the high-intensity or low-intensity expressions in a sequence. This study is reported in full in the Supplementary Information.

Study 8: testing nonlinearity in emotion perception. Thus far, our analysis of underlying mechanisms has focused on differential memory for more emotional (intense or salient) stimuli. However, another explanation of the apparent bias we observe is that low-level perceptual rather than emotional characteristics of emotional faces give rise to nonlinear integration of emotional faces into ensembles[42]. Given this concern, we used a computational modelling approach to separately quantify the psychophysical similarity between expressions, and we used these similarity data to estimate what biases in memory for ensembles would be expected on the basis of similarity alone. To achieve this goal, in the first phase we empirically tested how people perceived distances between emotional intensities at different points of our emotional scale. We then built on an existing computational model that was designed to simulate ensemble memory with specific attention to nonlinearity in similarity[34], by comparing three hypothetical models of ensemble coding: a baseline model that only incorporated nonlinearity in similarity, a recency model that was based on the baseline model but also assumed stronger weight in memory for more recent items, and an amplification model that was based on the recency model but added an assumption of increased weight to more emotional expressions. In the second phase, we used the results of Study 6 to compare these three models' fit. For the first phase, we recruited participants from Prolific in exchange for US$2.70. We aimed for a similar number of participants as in our other studies. No participants were excluded from the study. Our final sample was $N = 100$ (men, 37; women, 62; other, 1; age: mean = 35.99, s.d. = 12.69). To evaluate similarity in emotional perception, we modified the similarity task that was used by Schurgin and colleagues[42]. In each trial, the participants saw two expressions on the screen and were asked to evaluate to what degree these two expressions were similar to each other on a 1–7 scale (1, not similar at all; 7, very similar). The participants had as much time as they needed to make their selection. The similarity between two expressions was measured using a seven-point Likert scale, where $S_{min} = 1$ and $S_{max} = 7$. To generate the psychophysical similarity function, we simply normalized these data to range from 0 to 1, giving a psychophysical similarity metric, such that $f(x) = ((S_x - S_{min})/(S_{max} - S_{min}))$. To cover the whole 1–50 scale, one of every five expressions was selected and compared with all other expressions in increments of 5. For example, an expression of emotional intensity 1 was compared with 1, 5, 10, 15, 20, 25, 30, 35, 40, 45 and 50. Completing all comparisons within a certain scale required conducting 66 comparisons. In each study, the participants completed 264 (66 × 4) comparisons, which meant that each participant completed all possible comparisons in four of the eight expression–emotion continua: 4 identities × 2 valences (neutral-to-happy and neutral-to-angry). The four identities were chosen randomly for each participant. For the analysis of the results of this study, see the Supplementary Information.

Having established nonlinearity in similarity perceptions, we then took a computational modelling approach to validate that the amplification found in our studies did not stem from nonlinearity and perception of similarity. We adapted a recently developed model for ensemble memory[34], which is the first computational model to make high-precision predictions of performance in continuous-report memory ensemble tasks. In this work, we treat this memory for ensembles as a measurement model; that is, as explained, we use it to formally separate effects of psychophysical similarity from amplification memory biases. This model of memory for ensembles postulates that each stimulus evokes a distributed pattern of activation over feature values, and ensembles are computed by pooling over these patterns of activation at a relatively early perceptual stage of processing. Critically, within this modelling framework, the pattern of activation evoked by each stimulus depends on the psychophysical similarity of features to items held in memory, such that feature values that are more like items held in memory receive a higher boost in activation. This model directly links psychophysical similarity to memory processes by postulating that the patterns of activation elicited by each stimulus determine how familiar that feature and similar features will feel. For instance, if a task requires remembering a certain emotional intensity, the specific intensity will evoke a very strong familiarity signal, but so will similar emotional expressions. Finally, in line with mainstream signal detection models of memory[43], the model posits that ensemble memory representations are corrupted by noise and that the signal-to-noise ratio depends on factors that determine the top-down upweighting of features of individual items (for example, manipulation of memory load, delay or presentation format). Formally, the most straightforward version of this model for ensembles is given by the following equation:

$$R_{ENS} = \text{argmax}\left(\left(\sum_{i=1}^{N} f(x)_i\, d'\right) + \sigma_{Noise}\right) \quad (1)$$

where R_{ENS} is the reported feature on the ensemble task (that is, which expression is chosen), N is the total number of items in the ensemble memory array, $f(x)$ is the psychophysical similarity function of item i (that is, it captures how similar each of the 50 expressions is to item i; we describe the measurement of this below) and d' is a free parameter that determines the level of activation of each feature value for each item. Note that this version of the model postulates that on average, each item in the sequence generates the same familiarity signal, meaning that d' is the same value for each item in the sequence (that is, the model has only one free parameter d'). σ_{Noise} is a fixed amount of noise, which was set to one standard deviation of a Gaussian distribution, consistent with a signal detection model. Argmax denotes the decision rule that memory reports are based on the feature that generates the maximum familiarity signal. More precisely, the argmax argument is taken over a vector of random variables $(X_1, X_2, X_3, \ldots, X_{50})$, where each random variable is one of the 50 possible expressions on the self-report scale, each of which is distributed according to the model equation given in the parentheses. We refer to the above model as the baseline model because it assumes (1) that the familiarity of the ensemble is solely determined by its psychophysical similarity and (2) that, on average, there is equal weighting of each item in memory—that is, there are no sequential or amplification effects on memory (that is, no recency or exaggeration of the impact of negative expressions).

The second variant of the ensemble model we use is the recency model[34], which postulates that memory performance in the sequential paradigm is determined by psychophysical similarity as well as higher weighting of more recent items in memory (recency effects). In line with extant recency models of memory, the recency weights are quantified with a normalized exponential function (without base e) defined over the serial position of each stimulus in the sequence[17]. The recency model is given by the following equation:

$$R_{ENS} = \text{argmax}\left(\left(\sum_{i=1}^{N} f(x)_i\, d'\, w_i^{Recency}\right) + \sigma_{Noise}\right) \quad (2)$$

$$w_i^{Recency} = \frac{r^i}{\sum_{i=1}^{N} r^i} \quad (3)$$

where $w_i^{Recency}$ is the recency weight of the ith item in the sequence, and r is a free parameter that determines the rate of prioritization as a function of the serial position of a stimulus[17]. This version of the model therefore has two free parameters, d' and r. The critical point to note is that equations (1) and (2) are identical except that equation (2) can also capture higher weighting of more recent items. A comparison of these models thus provides insight into whether there is evidence for higher prioritization of more recent items in the sequence once psychophysical similarity is taken into account. Given prior evidence for recency effects in ensemble tasks, as well as in the studies reported above, we expected the recency model to outperform the baseline model.

The final model we refer to as the amplification recency model. This model of ensembles postulates that in addition to effects of psychophysical similarity and recency on memory, there is also amplification of emotional expressions. As noted, we use this model as a measurement model (meaning that we do not assume that it is the best descriptive model of amplification but rather use it to quantitatively separate amplification biases from psychophysical similarity and recency effects) to formally separate possible effects of amplification from psychophysical similarity and recency. Accordingly, in line with our behavioural results, we make the simplifying assumption that recency and amplification combine independently to bias memory, and that amplification increases exponentially as a function of an expression's emotional extremeness. The model equation is shown below:

$$R_{\mathrm{ENS}} = \mathrm{argmax}\left(\left(\sum_{i=1}^{N} f(x)_i\, d' w_i^{\mathrm{Recency}} w_i^{\mathrm{Amplification}}\right) + \sigma_{\mathrm{Noise}}\right) \quad (4)$$

$$w_i^{\mathrm{Amplification}} = \mathrm{e}^{A(j/50)} \quad (5)$$

As shown in the above equations, the $w_i^{\mathrm{Amplification}}$ weight is an exponential function of the item's emotionality, which is denoted by j (1–50) and a free parameter, A. Larger values of A indicate higher weighting of more emotional expressions, and we constrained A to be non-negative (zero inclusive) to capture the fact that there may be no amplification (when A equals zero). This model thus has three free parameters: d', r' and A. As before, the amplification recency model is like the baseline and recency models except that it posits that memory biases are jointly determined by psychophysical similarity, recency effects and amplification of more extreme expressions. A comparison of the amplification recency model with the baseline and recency models therefore provides direct insight into whether there are amplification memory biases once psychophysical similarity and recency effects are taken into account. See the Supplementary Information for the model fitting description.

Study 9: amplification in the evaluation of emotional videos. For Study 9, we used data from the SEND[32], in which observers watched and provided emotional ratings of videos of a diverse set of targets telling personal emotional stories. The participants were asked to provide two types of ratings in response to each video: a continuous evaluation of the degree of negativity and positivity of each video, and a global evaluation of the target emotionality after watching the whole video. These two measurements allowed us to compare the average of the participants' real-time emotional evaluations with their post-video global evaluations. We hypothesized that the participants' global evaluations would be stronger than the averages of their continuous evaluations (H1) and that amplification would be stronger for videos with negative than with positive narratives (H2). Given that the videos did not differ significantly in length, this study was not suitable to validate the association between length and amplification, but this association was tested nevertheless and showed no change (see the Supplementary Information for the full analysis). The target videos were collected as part of the SEND (see the full description of the data collection in the Supplementary Information)[32]. Of the videos that were produced by participants, 193 were selected, containing 49 unique targets (gender: men, 20; women, 27; other, 2; age: mean = 24.8, s.d. = 9.6; ethnicity: East Asian, 6; South Asian, 3; Black, 2; Hispanic, 4; Middle Eastern, 1; White, 16; mixed, 13; other, 4). The clips were also cropped for length, such that the final clips lasted on average 2 minutes 15 seconds. The videos were divided into four valence categories by the original authors, which we retained in this study. After we transformed the video ratings to be on a 0–100 scale (0, very negative; 50, neutral; 100, very positive), the videos were divided into four categories. Positive videos included videos that were rated by targets on average as higher than 60, with a minimum rating of 40 (N = 62). Negative videos had an average rating lower than 40 with a maximum rating of 60 (N = 33). Neutral videos were videos that had a maximum rating of 60 and a minimum rating of 40 (N = 30). All other videos were categorized as mixed (N = 68). See the original paper for the full description of the videos[32].

We use the term 'observers' to describe participants who were recruited separately at a later date and were asked to provide their evaluation of the target's emotionality. Observers were recruited as part of the SEND database on Amazon Mechanical Turk to watch video clips and rate how the target in the video felt[32]. The observers saw each video along with a continuous sliding scale underneath that was designed for continuous emotional ratings. They were asked to dynamically adjust the scale as the video played to capture the emotional intensity of the target at each time point. The analogue scale was divided into 100 points (0, very negative; 50, neutral; 100, very positive) and sampled every 0.5 s. Seven hundred participants were recruited with the goal of getting at least 20 participants rating each video. Each participant watched eight videos. The final recruited sample was 695 participants, and 11 additional participants were removed for failing to correctly answer two comprehension checks. Of the remaining 684 participants, we divided the continuous data into windows of two seconds and removed any observer ratings for videos that included fewer than five ratings. This elimination standard was different from that of the original researchers, who only removed participants who provided zero continuous ratings. We believe that our criteria are a more conservative comparison for the analysis. However, using the original authors' criteria does not change the significance of the results. Our final sample therefore was N = 565 (age: mean = 37.23, s.d. = 11.23; gender: female, 279; male, 254; undefined, 32).

One concern that may be raised when comparing the continuous and post-rating measures is that the continuous rating included the beginnings of the videos, in which the participants did not change their ratings, meaning that their rating was de facto neutral. Keeping these ratings may artificially reduce the overall average of the continuous rating and further emphasize the amplification. To avoid this issue, we cut the continuous ratings to start only when the observers made their first change to the rating, thus removing sections in which the rating was neutral. We then averaged each continuous rating from the point at which the participants made their first rating to the end of the video. The observers provided two types of ratings in response to each video. The first rating was a continuous rating on a 0–100 scale, 0 indicating very negative, 50 indicating neutral and 100 indicating very positive. The ratings were sampled every 0.5 s. After watching the video, the participants were asked to rate the degree of the target positivity and negativity using two ratings on a 1–7 scale (1, neutral; 7, very emotional), one for positive emotions and one for negative emotions. Because the correlation between positive ratings and negative ratings was very strong (r = −0.79 (−0.75, −0.82)), and to compare the continuous ratings with the post-ratings, we averaged between the positive and negative post-ratings, creating one scale for post-ratings, 1 (very negative) to 7 (very positive). To compare the post-ratings with the continuous ratings, we converted the continuous ratings to a 1–7 scale by dividing them by 100, multiplying by 6 and adding 1. With this transformation, 100 on a continuous scale was equal to 7, and 0 was equal to 1.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data for Studies 1–8 are available at https://osf.io/krgcv/. The data for Study 9 are available at https://github.com/StanfordSocialNeuroscienceLab/SEND.

Code availability

The code for the analysis of Studies 1–8 is available at https://osf.io/krgcv/. The code for the tasks can be found at https://github.com/GoldenbergLab/task-sequential-faces-emotion-estimationMe.

Received: 23 June 2021; Accepted: 16 May 2022;
Published online: 27 June 2022

References

1. Haberman, J., Harp, T. & Whitney, D. Averaging facial expressions over time. *J. Vis.* **9**, 1 (2009).
2. Alvarez, G. A. Representing multiple objects as an ensemble enhances visual cognition. *Trends Cogn. Sci.* **15**, 122–131 (2011).
3. Whitney, D. & Yamanashi Leib, A. Ensemble perception. *Annu. Rev. Psychol.* **69**, 105–129 (2018).
4. Oriet, C. & Brand, J. Size averaging of irrelevant stimuli cannot be prevented. *Vis. Res.* **79**, 8–16 (2013).
5. Oriet, C. & Hozempa, K. Incidental statistical summary representation over time. *J. Vis.* **16**, 3 (2016).
6. Schirmer, A. & Adolphs, R. Emotion perception from face, voice, and touch: comparisons and convergence. *Trends Cogn. Sci.* **21**, 216–228 (2017).
7. Eimer, M. & Holmes, A. Event-related brain potential correlates of emotional face processing. *Neuropsychologia* **45**, 15–31 (2007).
8. Goldenberg, A., Weisz, E., Sweeny, T., Cikara, M. & Gross, J. J. The crowd emotion amplification effect. *Psychol. Sci.* **32**, 437–450 (2021).
9. Jackson, M. C., Wu, C. Y., Linden, D. E. J. & Raymond, J. E. Enhanced visual short-term memory for angry faces. *J. Exp. Psychol. Hum. Percept. Perform.* **35**, 363–374 (2009).
10. Sessa, P., Luria, R., Gotler, A., Jolicœur, P. & Dell'acqua, R. Interhemispheric ERP asymmetries over inferior parietal cortex reveal differential visual working memory maintenance for fearful versus neutral facial identities. *Psychophysiology* **48**, 187–197 (2011).
11. Jackson, M. C., Wolf, C., Johnston, S. J., Raymond, J. E. & Linden, D. E. J. Neural correlates of enhanced visual short-term memory for angry faces: an fMRI study. *PLoS ONE* **3**, e3536 (2008).
12. Lee, H. J. & Cho, Y. S. Memory facilitation for emotional faces: visual working memory trade-offs resulting from attentional preference for emotional facial expressions. *Mem. Cogn.* **47**, 1231–1243 (2019).
13. Kaufmann, J. M. & Schweinberger, S. R. Expression influences the recognition of familiar faces. *Perception* **33**, 399–408 (2004).
14. Gallegos, D. R. & Tranel, D. Positive facial affect facilitates the identification of famous faces. *Brain Lang.* **93**, 338–348 (2005).
15. Jackson, M. C., Linden, D. E. J. & Raymond, J. E. Angry expressions strengthen the encoding and maintenance of face identity representations in visual working memory. *Cogn. Emot.* **28**, 278–297 (2014).
16. D'Argembeau, A. & Van der Linden, M. Facial expressions of emotion influence memory for facial identity in an automatic way. *Emotion* **7**, 507–515 (2007).
17. Tong, K., Dubé, C. & Sekuler, R. What makes a prototype a prototype? Averaging visual features in a sequence. *Atten. Percept. Psychophys.* **81**, 1962–1978 (2019).
18. Wilken, P. & Ma, W. J. A detection theory account of change detection. *J. Vis.* **4**, 1120–1135 (2004).
19. Fiedler, K. Beware of samples! A cognitive–ecological sampling approach to judgment biases. *Psychol. Rev.* **107**, 659–676 (2000).
20. Hubert-Wallander, B. & Boynton, G. M. Not all summary statistics are made equal: evidence from extracting summaries across time. *J. Vis.* **15**, 5 (2015).

21. Potter, M. Meaning in visual search. *Science* **187**, 965–966 (1975).
22. Albrecht, A. R. & Scholl, B. J. Perceptually averaging in a continuous visual world. *Psychol. Sci.* **21**, 560–567 (2010).
23. Goldenberg, A., Sweeny, T. D., Shpigel, E. & Gross, J. J. Is this my group or not? The role of ensemble coding of emotional expressions in group categorization. *J. Exp. Psychol. Gen*. https://doi.org/10.1037/xge0000651 (2019).
24. Krumhuber, E. G., Kappas, A. & Manstead, A. S. R. Effects of dynamic aspects of facial expressions: a review. *Emot. Rev.* **5**, 41–46 (2013).
25. Krumhuber, E. G. & Skora, L. in *Handbook of Human Motion* (eds Müller, B. & Wolf, S. I.) 1–15 (Springer International, 2016); https://doi.org/10.1007/978-3-319-30808-1
26. Yoshikawa, S. & Sato, W. Dynamic facial expressions of emotion induce representational momentum. *Cogn. Affect. Behav. Neurosci.* **8**, 25–31 (2008).
27. Kahneman, D., Fredrickson, B. L., Schreiber, C. A. & Redelmeier, D. A. When more pain is preferred to less: adding a better end. *Psychol. Sci.* **4**, 401–405 (1993).
28. Redelmeier, D. A. & Kahneman, D. Patients' memories of painful medical treatments: real-time and retrospective evaluations of two minimally invasive procedures. *Pain* **66**, 3–8 (1996).
29. Fredrickson, B. L. Extracting meaning from past affective experiences: the importance of peaks, ends, and specific emotions. *Cogn. Emot.* **14**, 577–606 (2000).
30. Rozin, P. & Rozin, A. Advancing understanding of the aesthetics of temporal sequences by combining some principles and practices in music and cuisine with psychology. *Perspect. Psychol. Sci.* **13**, 598–617 (2018).
31. Cojuharenco, I. & Ryvkin, D. Peak–end rule versus average utility: how utility aggregation affects evaluations of experiences. *J. Math. Psychol.* **52**, 326–335 (2008).
32. Ong, D. C. et al. Modeling emotion in complex stories: the Stanford Emotional Narratives Dataset. *IEEE Trans. Affect. Comput.* https://doi.org/10.1109/taffc.2019.2955949 (2019).
33. Koller, M. robustlmm: an R package for robust estimation of linear mixed-effects models. *J. Stat. Softw.* **75**, 1–24 (2016).
34. Robinson, M. & Brady, T. A quantitative model of ensemble perception as summed patterns of activation in feature space. Preprint at *PsyArXiv* https://doi.org/10.31234/osf.io/k3d26 (2022).
35. Eid, M. & Diener, E. Norms for experiencing emotions in different cultures: inter- and intranational differences. *J. Pers. Soc. Psychol.* **81**, 869–885 (2001).
36. Kardosh, R., Sklar, A. Y., Goldstein, A., Pertzov, Y. & Hassin, R. R. Minority salience and the overestimation of individuals from minority groups in perception and memory. *Proc. Natl Acad. Sci. USA* **119**, e2116884119 (2022).
37. Zacks, J. M. & Swallow, K. M. Event segmentation. *Curr. Dir. Psychol. Sci.* **16**, 80–84 (2007).
38. Dubé, C., Zhou, F., Kahana, M. J. & Sekuler, R. Similarity-based distortion of visual short-term memory is due to perceptual averaging. *Vis. Res.* **96**, 8–16 (2014).
39. Elias, E., Dyer, M. & Sweeny, T. D. Ensemble perception of dynamic emotional groups. *Psychol. Sci.* **28**, 193–203 (2017).
40. Tottenham, N. et al. The NimStim set of facial expressions: judgments from untrained research participants. *Psychiatry Res.* **168**, 242–249 (2009).
41. Langner, O. et al. Presentation and validation of the Radboud Faces Database. *Cogn. Emot.* **24**, 1377–1388 (2010).
42. Schurgin, M. W., Wixted, J. T. & Brady, T. F. Psychophysical scaling reveals a unified theory of visual memory strength. *Nat. Hum. Behav.* **4**, 1156–1172 (2020).
43. Wickens, T. D. *Elementary Signal Detection Theory* (Oxford Univ. Press, 2001).

Acknowledgements

J.S. is supported in part by the German Academic Scholarship Foundation (Promotionsförderung der Studienstiftung des deutschen Volkes) and in part by NIH grant no. 1R01MH112560-01. D.C.O. is supported in part by a Singapore Ministry of Education Academic Research Fund Tier 1 grant. T.F.B. is supported in part by NSF CAREER grant no. BCS-1653457. Finally, M.M.R. is supported by the National Research Service Award fellowship from the National Institute of Health no. 1F32MH127823-01. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

A.G., J.S. and Z.H. conceived and designed the experiments. A.G. and J.S. performed the experiments (Studies 1–8), analysed the data for these experiments and wrote the paper. D.C.O. and J.Z. designed and ran Study 9, and D.C.O. provided an initial analysis of the data. D.C.O. and J.Z. reviewed the paper. T.F.B. and M.M.R. analysed the data of Study 8, conducted the computational model of that study and reviewed the paper. D.L., T.D.S. and J.J.G. were involved in writing and reviewing the manuscript.

Competing interests

The authors declare no competing interests. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-022-01390-y.

Correspondence and requests for materials should be addressed to Amit Goldenberg.

Peer review information *Nature Human Behaviour* thanks Raoul Bell and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature portfolio

Corresponding author(s): Amit Goldenberg

Last updated by author(s): 2022-5-22

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data collection was done using a task that was built in jspsych.
Data analysis	Data analysis was done with R and Matlab

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Data for studies 1-8 is available here: https://osf.io/krgcv/
Data for Study 9 is available here: https://github.com/StanfordSocialNeuroscienceLab/SEND

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Studies are mostly repeated-measure within participant tasks evaluated quantitatively.
Research sample	All participants were recruited from prolific/mturk. Data is not representative in terms of population. Specific details: Study 1a: N=93; men: 52, women: 41; age: M= 28.23, SD = 9.54, Study 1b: N = 94; men: 35, women: 55 other/did not specify = 4; age: M= 33.60, SD = 11.43, Study 1c: N=98; men: 35, women: 62, other/did not specify= 1 ; age: M= 25.78, SD = 9.08, Study 1d: N=92; men: 31, women: 66, other/did not specify= 1; age: M= 24.67, SD = 6.42 Study 2: N= 150 men: 51, women: 98, 1 Other; age: M= 38.35, SD = 12.37 Study 3: N= 96; men: 62, women:33, other: 1, Age: M = 25.49, SD = 7.35 Study 4: N = 295; men: 119, women:175, other: 1, Age: M = 25.14, SD = 8.06 Study 5: N= 100; men: 37, women:62, other: 1, Age: M = 35.99, SD = 12.69 Study 6: N=565; age: M=37.23, SD =11.23, gender: female =279, male =254, undefined = 32
Sampling strategy	Sampling for all studies except for Study 6 was done in prolific. Sample sized were determined by power analyses that were conducted either as a result of previous findings (Study 1a), or based on results from the previous studies in the study sequences. All studies except for Study 6 were pre-registered and links are shared in the manuscript. Data from Study 6 is taken from a project by Ong, wu, Zhi-Xuan, Reddan, Kahhale, Mattek & Zaki, 2019 - Modeling emotion in complex stories: The stanford emotional narratives dataset.
Data collection	All of our data were collected online with a jspsych task that was designed by the research team. Participants were sent to the task and followed the instructions. Once they completed the task, they were sent to a Qualtrics survey and answered a few questions. In cases in which some variables were manipulated, for example when manipulating the order in which faces were presented, participants were not informed of such manipulations due to the fear that it would affect their performance in the task.
Timing	Start - Jan 2020 End - March 17 2021.
Data exclusions	Our exclusion criteria for studies 1-5 was participants who did not finish the study or whose average rating was lower than 10 or higher than 40. Given that average of all sequences was 25, the probability of getting to these averages is almost 0%. Study 1a: N=93; excluded - 7 Study 1b: N = 94; excluded - 6 Study 1c: N=98; excluded - 2 Study 1d: N=92; excluded - 8 Study 2: N= 150; excluded - 14 Study 3: N= 96; excluded - 4 Study 4: N = 295; excluded - 5 Study 5: N= 100; excluded - 0
Non-participation	*State how many participants dropped out/declined participation and the reason(s) given OR provide response rate OR state that no participants dropped out/declined participation.*
Randomization	In studies in which was done (Studies 3, 4) this was done randomly.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	See above
Recruitment	Participants were recruited through prolific.
Ethics oversight	Harvard University

Note that full information on the approval of the study protocol must also be provided in the manuscript.