# Dynamic representations of behaviourally relevant features support rapid face processing in the human ventral visual stream

Diana C. Dima*[1] and Krish D. Singh[1]

[1] *Cardiff University Brain Research Imaging Centre (CUBRIC), School of Psychology, Cardiff University, Cardiff CF24 4HQ, United Kingdom*

*Corresponding author: `dimadc@cardiff.ac.uk` (D.C. Dima)

**Abstract**

Humans are remarkably good at extracting information from faces even in challenging viewing conditions, yet the neural representations supporting this ability are still not well understood. Here, we varied the presentation duration of backward-masked facial expressions to investigate their representational dynamics using magnetoencephalography (MEG). The results highlighted a two-stage process supporting face perception, with the ventral stream encoding facial features at an early stage and facial configuration at a later stage. Reducing presentation time modulated this process: early responses were transient, while featural and configural representations emerged later. These patterns overlapped with representations of behaviour in ventral stream areas, pointing to their importance in extracting task-relevant information. Although both face presence and expression were decodable from MEG data when stimuli were presented as briefly as 30 ms, only face presence could be decoded outside of subjective awareness. Together, our results describe the adaptable system linking facial configuration, brain and behaviour.

# 1 Introduction

Faces are rapidly and preferentially processed in a distributed network of brain areas which extract and relay information, such as emotional cues, with remarkable efficiency (Grill-Spector et al., 2017; Ishai, 2008; Pessoa and Adolphs, 2010). This advantage is thought to extend to unconscious processing, although the extent and mechanisms of face perception outside of awareness, as well as the underlying representations, are still not well understood. In the current study, we addressed this question by interrogating the neural representations of rapidly presented emotional faces using multivariate analysis of magnetoencephalography (MEG) data.

Using different methods of rendering faces "invisible", such as binocular suppression or backward masking, many experiments have shown some degree of unconscious face processing, demonstrated at the behavioural or neural levels (see Axelrod et al., 2015 for a review). However, electrophysiological investigations paint a complex picture of the underlying mechanisms: while many studies using binocular suppression have detected evoked responses to invisible faces (Sterzer et al., 2009; Jiang et al., 2009), other studies report no such effect, particularly when using backward masking (Reiss and Hoffman, 2007; Fisch et al., 2010; Navajas et al., 2013; Rodriguez et al., 2013), which is thought to disrupt re-entrant processing through conflicting input from feedforward connections (Lamme et al., 2002).

Facial expression has been shown to modulate the early stages of visual perception (Aguado et al., 2012; Dima et al., 2018a) and to elicit non-conscious responses in numerous studies (Tamietto and De Gelder, 2010). Evidence of "blindsight" (non-conscious perception despite visual cortex lesions; e.g. Pegna et al., 2005) has led to considerable debate about the automaticity of emotion perception and the role of a subcortical route in facilitating it (Pessoa et al., 2005; Pessoa, 2005), with most of the evidence showing a processing advantage for invisible fearful faces (e.g. Bertini et al., 2017; Jiang and He, 2006; Williams et al., 2004). However, some studies show evidence against the non-conscious processing of expression (Schlossmacher et al., 2017; Hedger et al., 2016). Furthermore, it is unclear whether the advantage found in many experiments for fearful faces generalizes to threat-related expressions, or is linked to characteristic low-level properties (Hedger et al., 2015). This idea is reinforced by inconsistent effects found for angry faces: while some experiments show evidence of non-conscious perception of angry faces (e.g. Almeida et al., 2013; Adams et al., 2010), other studies show no effect or even a disadvantage in the competition for awareness (Hedger et al., 2015; Gray et al., 2013; Hedger et al., 2016).

In addition, although evidence of rapid face processing points to highly efficient feature extraction, the mechanisms supporting this are still the subject of debate. It is widely believed that faces are perceived holistically, unlike other stimuli (Farah et al., 1998; Richler and Gauthier, 2014); however, some behavioural goals, such as identity recognition, are thought to rely on facial features and not on holistic perception (Visconti Di Oleggio Castello et al., 2017). Classic models support a configural model of face perception (Calder et al., 2000; Namdar et al., 2015), from the detection of a first-order configuration (face features) to the perception of a second-order configuration determined by relationships between features (Piepers and Robbins, 2012; Maurer et al., 2002). Although classic paradigms like face inversion or the composite face have shown how the highly specialized mechanisms for face perception can break down in the presence of configural disruption (Behrmann et al., 2014), the spatiotemporal dynamics of these processes remain less well understood.

In the current study, we collected MEG data while participants viewed briefly presented and backward-masked happy, angry, neutral, and scrambled faces. By varying stimulus duration, we were able to interrogate the neural representations underpinning rapid face and expression perception, and we tracked how they change in the presence of limited visual input. We then used multivariate methods to assess the presence of neural responses to faces presented outside of subjective awareness. We reliably detected a

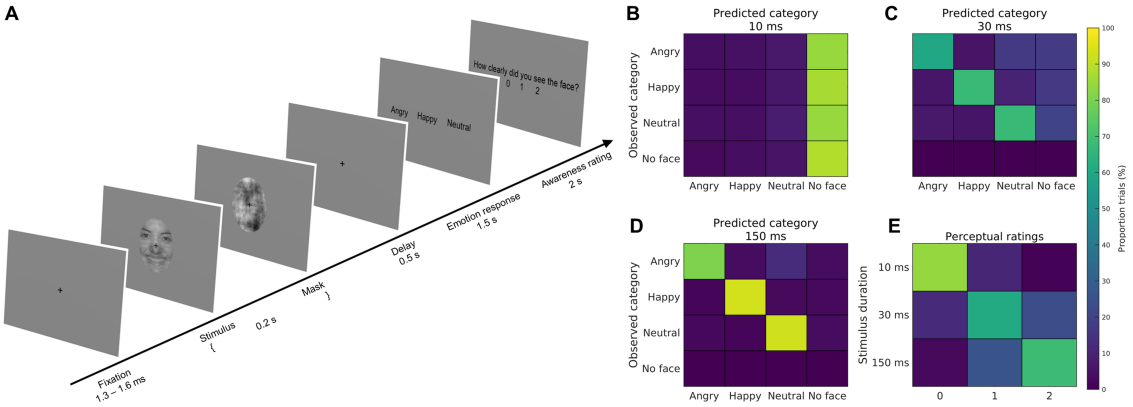



Figure 1: Overview of the experimental paradigm and behavioural results. **A**. Stimuli were presented on screen for 150 ms, 30 ms, or 10 ms, and were followed by a 50 ms, 170 ms, or 190 ms scrambled mask. **B-D**. Confusion matrices mapping the average proportion of trials receiving each of the possible responses (X-axis) out of the trials belonging to each category (Y-axis). "No response" trials were excluded for statistical analysis, but are shown here as representing a "no face" response. **E**. Perceptual ratings for each stimulus duration summarized as average proportion of trials.

response to subliminal faces, but no expression modulation outside of awareness, adding to recent evidence against the automaticity of expression perception. Finally, we used representational similarity analysis (RSA) to tease apart the contributions of first-order and second-order face configuration and to explore the link between behaviour and ventral stream responses to faces.

## 2 Results

In order to investigate rapid face and expression processing, we recorded MEG data while participants ($N$=25) viewed emotional faces presented for 150, 30 and 10 ms, followed by scrambled masks presented for 50, 170, and 190 ms respectively (Figure 1A). We collected 80 trials for each emotional expression in each stimulus duration condition, as well as 40 control trials in which scrambled masks were presented for 10 ms instead of faces. Stimuli were matched in terms of low-level properties (including Fourier amplitude spectra), were aligned using automated eye detection, and cropped to an oval mask. Participants performed an emotion discrimination task using button responses counterbalanced across sessions, and, in addition, rated their perceptual awareness on a 3-point scale ranging from 0 (no face perceived) to 2 (clear perception).

## 2.1 Perception and behaviour

In order to assess the effects of stimulus duration and face expression on behaviour, we calculated confusion matrices mapping the expression discrimination responses to each stimulus category (Figure 1). Performance was quantified as proportion correct trials after excluding trials with no response, and a rationalized arcsine transformation was applied prior to statistical analysis (Studebaker, 1985). We then performed a $3 \times 3$ repeated-measures ANOVA with factors *Duration* (levels: 10 ms, 30 ms, 150 ms) and *Expression* (levels: angry, happy, neutral).

As expected, stimulus duration had a strong effect on expression discrimination performance, with average performance not exceeding chance level at 10 ms ($33.45\% \pm 2.99$) and rising well above chance at 30 and 150 ms ($78.62\% \pm 2.11$ and $91.83\% \pm 1$ respectively). This was reflected in a significant main effect of duration in the ANOVA ($P < 0.0001$, $F(1.21, 29.06) = 221.05$, $\eta^2 = 0.9$). Face expression had a weak effect, with angry faces categorized less accurately than both happy and neutral faces ($P = 0.046, F(1.95, 46.71) = 3.33, \eta^2 = 0.12$), with no significant interaction effect ($P = 0.23$, $F(1.74, 41.83) = 1.53$, $\eta^2 = 0.06$).

Participants found the task challenging, as reflected in the perceptual awareness ratings: 84.5% of the 10 ms trials were rated as not containing a face (Figure 1E). This suggests that participants were complying with the task with respect to both expression discrimination and perceptual rating. Importantly, for faces presented for 10 ms, there was no difference in accuracy between expressions ($P = 0.43, F(1.65, 39.5) = 0.8$) or between any pair of cells in the confusion matrix ($P = 0.6, F(3.42, 82.07) = 0.64$), suggesting that faces presented at this duration were equally likely to be categorized as any expression. Note that the expression discrimination task here was not a forced-choice task (participants could refrain from responding) and these tests were performed on the small subset of 10 ms trials that received a response; references to awareness in this paper thus refer exclusively to subjective awareness, as indicated by perceptual ratings.

## 2.2 Spatiotemporal dynamics of face perception

To investigate face processing as a function of stimulus duration, we performed within-subject decoding of responses to faces vs. scrambled stimuli. The analysis included three components: sensor-level time-resolved classification to evaluate the progression of condition-related information; sensor-level temporal generalization to assess the temporal structure of this information; and source-space decoding to obtain spatial information

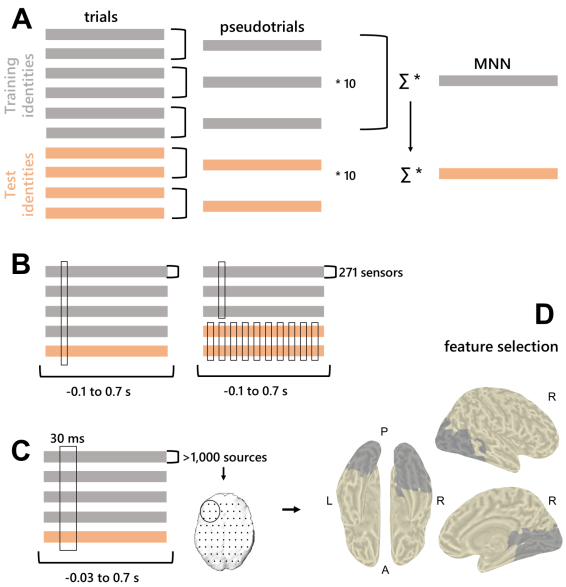



Figure 2: Overview of the MVPA analysis pipeline.**A**. Trial averaging and multivariate noise normalization (MNN) procedure. Σ is the error covariance matrix. **B**. Sensor-space time-resolved decoding (left) and temporal generalization (right). **C**. Source-space searchlight decoding procedure. **D**. Sources included in the representational similarity analysis based on face vs. scrambled classification results. P: posterior; A: anterior; L: left; R: right.

about subliminal responses to faces (Figure 2).

We used cross-exemplar cross-validation to ensure that stimulus repetitions and face identity effects did not bias our results, and we assessed the significance of decoding performance using randomization testing (Nichols and Holmes, 2001; Singh et al., 2003; see 4.9).

We first performed time-resolved classification on responses to neutral faces and scrambled stimuli using data from all MEG sensors, separately for each stimulus duration. In the case of faces presented for 10 ms, we excluded any trials reported as containing a face, to ensure that we assessed responses outside of subjective awareness. Scrambled stimuli could be discriminated from faces presented for 150 and 30 ms as early as 100 ms, as reflected by above-chance decoding performance on the MEG sensor set (Figure 3A). After the initial peak in performance, decoding accuracy decreased, but remained well above chance for the remainder of the decoding time window. For faces presented for 10 ms and reported as not perceived, there was only a weak increase in decoding performance, which reached significance at 147 ms and dropped back to chance level after ~350 ms (Supplementary Table 1).

To assess how well face representations generalized across stimulus durations, we repeated this analysis by training and testing on stimulus exemplars presented for different

amounts of time (Figure 3B). Decoding accuracy was high when cross-decoding between 30 ms and 150 ms faces; interestingly, after an initial peak (100-200 ms), performance decreased, and started increasing again after 300 ms, suggesting that representations become more similar over time. On the other hand, representations only generalized to 10 ms faces for a limited time window, with a peak at M170 latencies.

Using temporal generalization decoding (King and Dehaene, 2014), we investigated the temporal structure underpinning face decoding, and we found that this changed with stimulus duration. For faces presented for 150 ms, successful temporal generalization started at ~93 ms in a diagonal pattern suggestive of transient representations, with more sustained representations (square patterns) arising at M170 latencies and after 300 ms (Figure 3D-E). For 30 ms stimuli, a diagonal generalization pattern started at ~110 ms after stimulus onset and sustained representations only arose later (~400 ms). Early processing thus appears to be heavily biased by stimulus presentation duration, with 30 ms faces failing to elicit a stable representation at M170 latencies. For faces presented for 10 ms, only few transient clusters survived correction for multiple comparisons, with the largest one occurring after 200 ms.

Finally, we spatially localized the subliminal response to faces in source space. All participants with one exception acquired a structural MRI, which was used to source localize the MEG data using a Linearly Constrained Minimum Variance (LCMV) beamformer (Van Veen et al., 1997). We performed whole-brain searchlight classification of 10 ms faces vs. scrambled stimuli ($N$=24), using source clusters with a radius of 10 mm and time windows of 30 ms. Faces were successfully decoded in a right occipital area at M170 latencies (Figure 3C), with a later stage associated with ventral patterns. Given the disruption of recurrent processing through backward masking in this paradigm, the early occipital sources likely reflect the feedforward nature of this response.

## 2.3 Temporal dynamics of expression perception

We performed sensor-level time-resolved decoding of all pairs of emotional expressions separately for each stimulus duration. The analysis was performed similarly to the time-resolved face decoding analysis described above.

The highest decoding performance was achieved on late responses to expressions presented for 150 ms (Figure 4A). Expressions presented for 30 ms also achieved above-chance decoding, although these effects were more transient. We also performed this analysis on pooled datasets (faces presented for 30 and 150 ms), as the face cross-decoding analysis

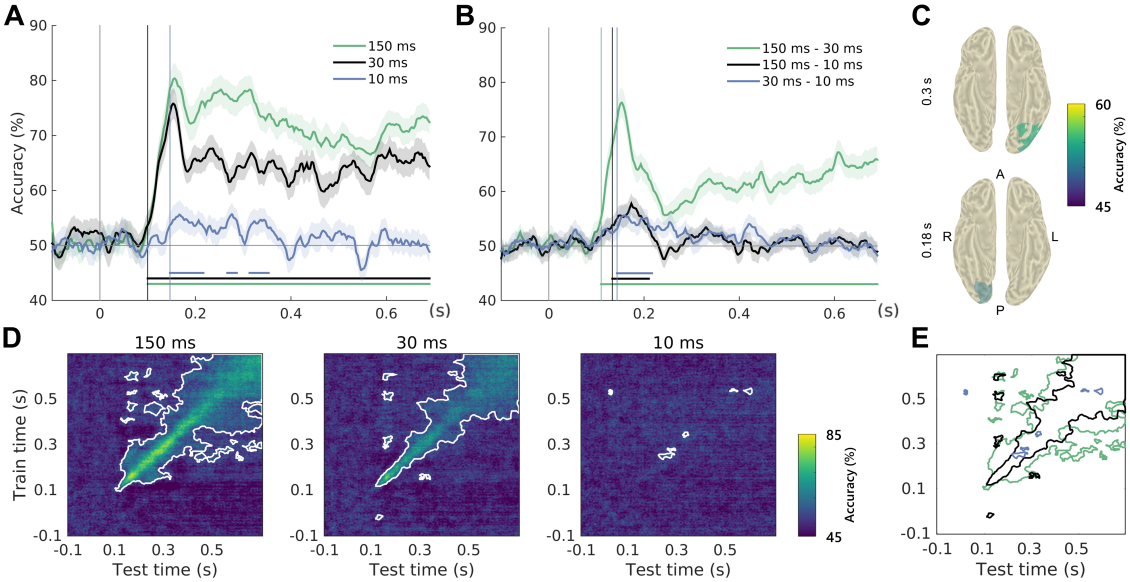



Figure 3: Face vs. scrambled decoding results. **A**. Sensor-space time-resolved decoding accuracy for all stimulus durations. Vertical bars mark above-chance decoding onset and horizontal lines show significant time windows ($P$<0.05, corrected). **B**. Sensor-space time-resolved cross-decoding for all pairs of stimulus durations. Cross-validation was performed across exemplars and accuracies were averaged over the two training/test directions. **C**. Sources achieving above-chance decoding of 10 ms faces outside awareness at M170 latencies in source space ($P$<0.005, corrected). **D**. Sensor-space temporal generalization accuracy and significant clusters (white contours; $P$<0.05, corrected) for all stimulus durations. **E**. Significant temporal generalization clusters for all three stimulus durations, showing more sustained representations of faces presented for 150 ms (legend as in A).

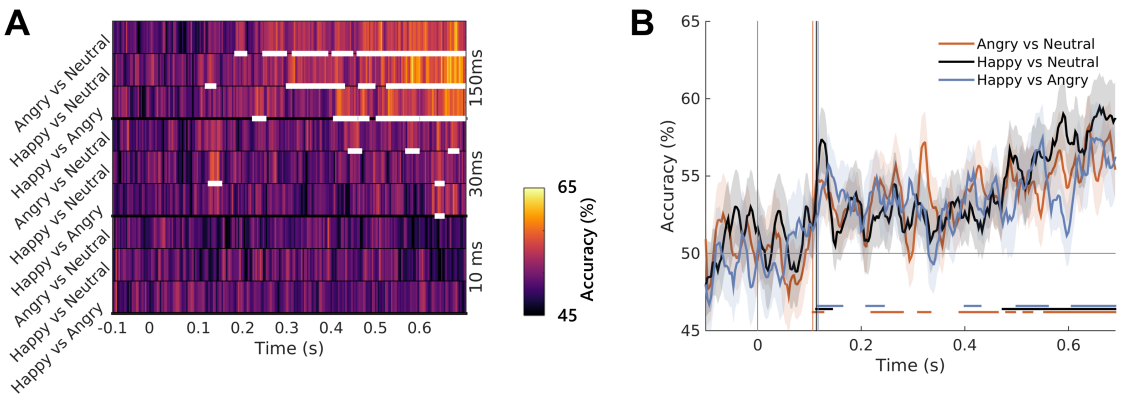



Figure 4: Expression decoding results. **A**. Time-resolved decoding accuracy for the three expression decoding problems and the three stimulus durations. White horizontal lines show significant time windows ($P$<0.05, corrected). **B**. Time-resolved accuracy for the three expression decoding problems using the pooled datasets (30 + 150 ms).

showed that responses generalized between these two categories (Figure 3B). Complementary results were obtained using the pooled datasets (faces presented for 30 and 150 ms), which revealed a multi-stage progression for all expressions, with transient early decoding at M100 latencies and an increasing accuracy in late time windows (Figure 4B). We found no above-chance performance when decoding 10 ms expressions. This finding adds to mounting evidence against the automatic processing of expression outside awareness (Koster et al., 2007; Pessoa et al., 2006; Hedger et al., 2016), and we explore potential reasons for this result below (3).

### 2.4 Face representations in occipitotemporal cortex

To interrogate the content of neural representations in space and time, we performed representational similarity analysis (RSA) using a searchlight approach at the source level (Su et al., 2012). A face-responsive area of interest was selected by performing whole-brain searchlight classification of faces (150 ms) vs. scrambled stimuli and identifying the sources achieving high decoding accuracy ($99.5^{th}$ percentile). Time-resolved neural representational dissimilarity matrices (RDMs) were constructed using a spatiotemporal searchlight approach in this ROI (4.8). To create the RDMs, we computed cross-validated Mahalanobis distances between cross-subject responses to all pairs of face stimuli. RSA analyses were separately conducted for the three stimulus duration conditions.

We investigated the temporal dynamics of face perception by assessing the similarity between MEG patterns and models quantifying behaviour, expression, identity and visual properties. We evaluated 9 different model RDMs at each stimulus duration: a behavioural

RDM; a face identity model; three high-level expression RDMs investigating different models of expression perception; a model based on automatically detected Action Units (movements of individual facial muscles; Ekman and Friesen, 1977); two face configuration models capturing first-order and second-order characteristics; and a spatial envelope model (Figure 5).

We quantified the similarity between MEG patterns and each of the models at each time window and searchlight using Spearman's rank correlation (Nili et al., 2014), and we performed a further partial correlation analysis to measure the unique contribution of the significantly represented visual models.

### 2.4.1 Occipitotemporal cortex encodes behavioural responses

To assess the link between behaviour and neural patterns, we calculated cross-validated model RDMs based on expression discrimination patterns across participants. Among the other model RDMs tested, behavioural RDMs correlated most with the high-level expression models (particularly the angry-vs-others model at 30 ms and 150 ms, Spearman's $\rho = 0.29$ and $\rho = 0.34$). At 150 ms, the behavioural RDM also correlated with the configural face models ($\rho = 0.22$ and $\rho = 0.18$). As expected based on performance, behavioural RDMs at 10 ms did not correlate with the other two ($\rho = -0.05$ and $\rho = -0.09$ respectively), while behavioural RDMs at 30 and 150 ms were positively correlated ($\rho = 0.38$; Figure 5B).

Based on these links, face configuration, together with facial expression, appears to partially explain behavioural responses. To more directly test this, we performed a variance partitioning analysis, using hierarchical multiple regression to quantify the unique and shared variance explained by facial configuration and high-level expression models in behavioural responses (4.10.4). In the 10 ms condition, the neutral-vs-others model and the two configural models explained 25.1% of the variance; in the 30 ms and 150 ms conditions, the angry-vs-others model and the configural models explained up to 45.7% of the variance in behaviour. Furthermore, while the expression model contributed most of the variance, over 75% of this variance was shared with the configural models. The unique contribution of configural models increased with stimulus duration (from ~2% at 10 ms, to ~20% at 150 ms). Together, these results point to the role of face configuration in driving high-level representations and behaviour. Note that for the 10 ms condition, we were unable to decode expression from the MEG data; however, expression and configuration explained a portion of the variance in behaviour, suggesting that they may contribute to

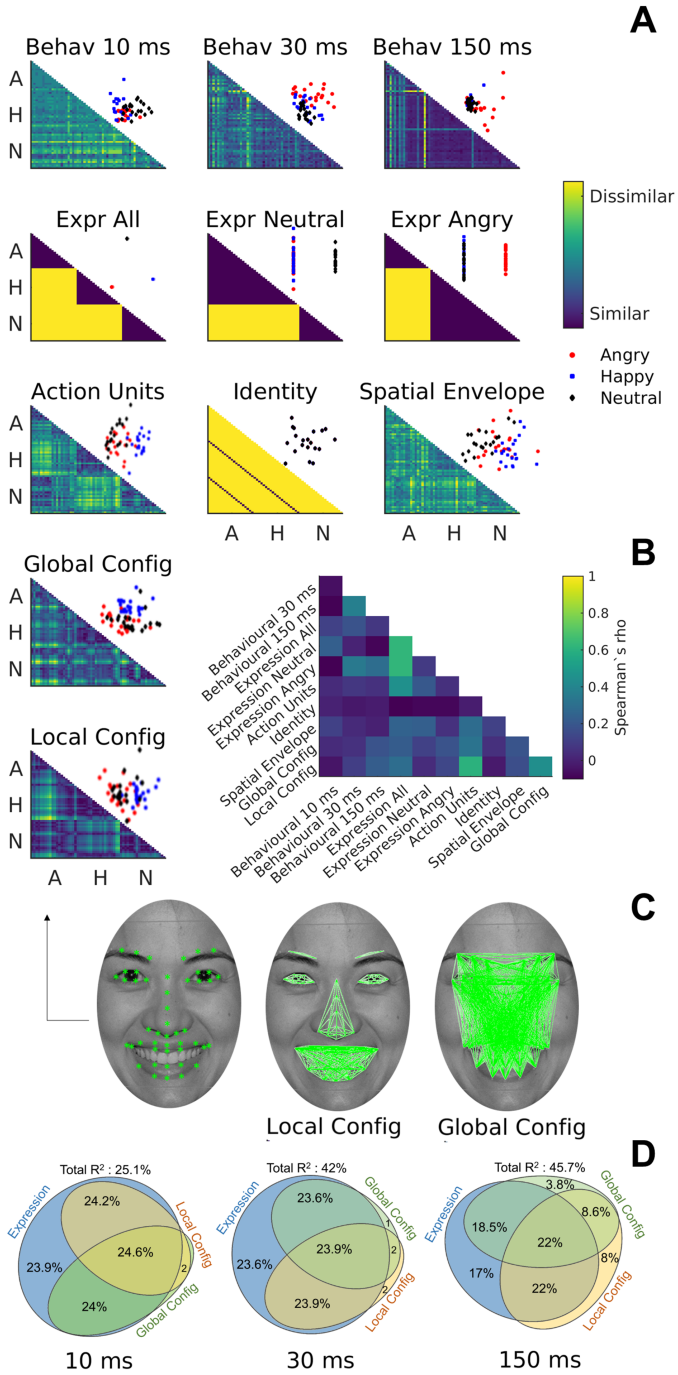



Figure 5: Models used in RSA analysis. **A**. Model RDMs showing predicted distances between all pairs of stimuli (lower triangles). A:Angry; H: Happy; N: neutral. Stimuli are sorted according to face identity. Upper triangles show 2D multidimensional scaling (MDS) plots for each model, which help visualize the distances between stimuli according to each model. **B**. Model inter-correlations (Spearman's $\rho$). **C**. Metrics used to derive the local and global face configuration models. The left-hand panel shows automatically detected facial landmarks for an example stimulus, while the other two panels depict the pairwise Euclidean distances used to calculate the two model RDMs. Behav: behavioural models; Expr: high-level expression models (all-vs-all, neutral-vs-others, and angry-vs-others); Config: face configuration models. **D**. Variance partitioning results, showing the contributions of expression and face configuration models to behavioural responses at each stimulus duration. Values represent % of the total $R^2$.

the subliminal response to faces.

Behavioural RDMs showed the strongest and most sustained correlations with MEG patterns in ventral stream areas, including sources corresponding to the location of the fusiform face area (FFA) and OFA (Figure 6). Behavioural representations evolved differently in time for the three stimulus durations. For 10 ms faces, behaviour explained the data starting at 120 ms until the end of the analysis time window. Representations emerged similarly early for 150 ms faces and reached the noise ceiling before falling back to low $\rho$ values at 400 ms. For 30 ms faces, correlations were significant starting at 210 ms in a relatively focal right temporal area. Patterns were more posterior for 10 ms faces and more extensive, including sources corresponding to the OFA and FFA, for 150 ms faces.

The correlation time-courses suggest interesting differences in processing as a function of the information available: for clearly perceived faces, features relevant in behaviour are extracted between 120-400 ms, while behavioural responses for briefly presented faces appear to require sustained processing, as reflected by behaviour-related correlations not dropping back to zero. These results are in line with previous evidence of behavioural representations in ventral stream areas in scene and object perception (e.g. Walther et al., 2009), and suggest that visual feature processing, even at early stages, is closely linked to behavioural goals.

### 2.4.2 Configural face processing from featural to relational

The two face configuration models were also represented in the MEG patterns. In the correlation analysis, the local and global configuration models explained representations in partially overlapping areas of the ventral stream (corresponding to the right FFA location), with local configuration representations arising earlier (at 120 ms for 150 ms faces, and 300 ms for 30 ms faces). Our RSA method (4.10.3) favoured sustained correlations over transient peaks; note that the global configuration model correlation approached the noise ceiling during a transient time window at M170 latencies for both 150 ms and 30 ms faces, suggesting a contribution of second-order characteristics, although this occurred later than first-order feature representations (Supplementary Figure 4). The partial correlation analysis revealed further differences between conditions: for 150 ms faces, the local and global models made unique contributions in explaining the data; conversely, for 30 ms faces we detected no unique contributions, suggesting that the extraction of configural information from faces occurs differently in the absence of sufficient information. None of the models significantly correlated with MEG patterns elicited by 10 ms faces.

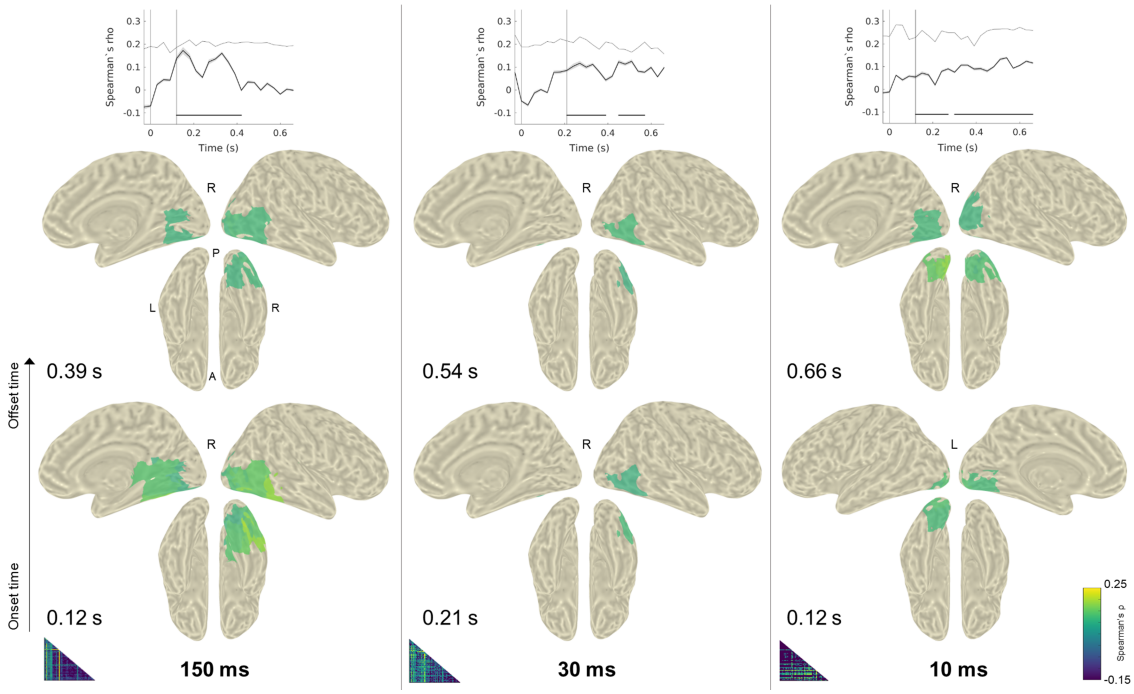



Figure 6: Correlations between MEG patterns and behavioural model RDMs for each stimulus condition duration (vertical columns). The top panels show correlation time-courses averaged across all significant searchlights; the noise ceiling is shown as a dotted horizontal line and is only approached in the 150 ms condition. The cortical maps show significant correlation coefficients for the first and last significant time windows (onset and offset times) on the inflated template MNI brain. The hemisphere shown is indicated with the letter R/L. Model RDMs are shown in the lower left corner of each column. See SourceMovies1 for movies showing the evolution of behavioural representations in time.

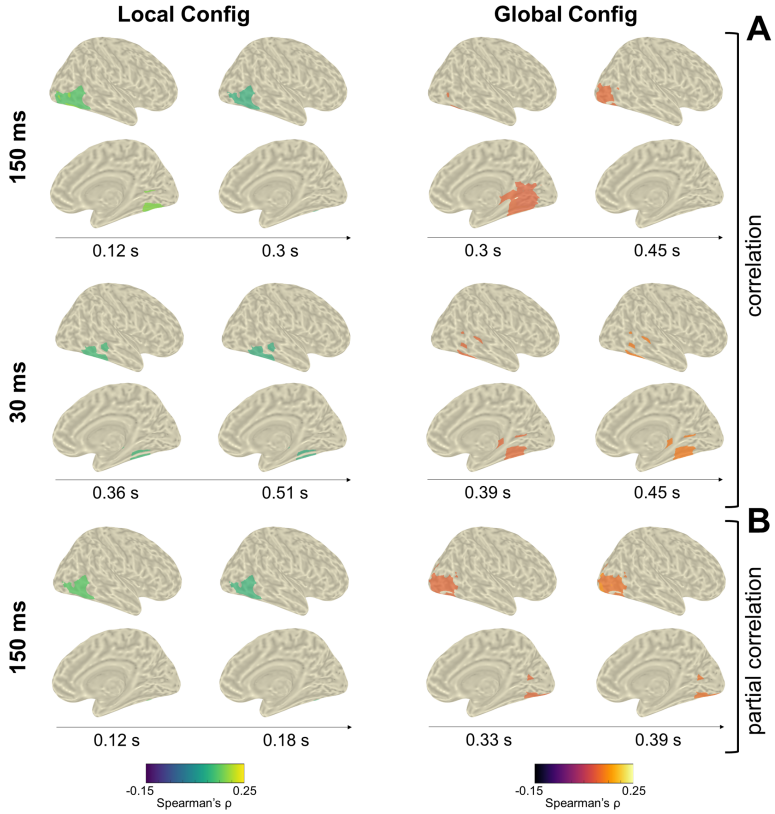



Figure 7: Significant correlations between MEG patterns and configural model RDMs. **A**: Correlation analysis results are significant for the 150 ms and 30 ms conditions. **B**: Partial correlation results are significant for the 150 ms condition. Only right hemisphere searchlights correlate with the configural models. Maps are shown for the onset and offset times of significant correlation. See SourceMovies2 for movies showing the evolution of behavioural representations in time.

Note that although both internal (eyes, nose, mouth) and external (face shape, hair) face features have been shown to contribute to neural responses to faces (Axelrod, 2010), we focus here on internal features; for the purposes of this paper, external features were excluded from the stimuli and we refer to the second-order configuration of distances between internal features as "global configuration". Internal features are relevant to the context of expression discrimination and have been shown to be more reliable even in facial recognition contexts (e.g. Kemp et al., 2016; Longmore et al., 2015).

### 2.4.3 Transient representations of visual and high-level models

Two other models elicited brief representations in the MEG data. For 150 ms faces, the spatial envelope model explained left hemisphere occipital representations starting at ~400 ms, suggesting sustained processing of visual features, potentially based on feedback mechanisms.

For 30 ms faces, a high-level expression model (neutral-vs.-others) was represented in

the MEG data starting at 300 ms (Figure 8). This can be speculatively explained by the formation of task-related representations in the absence of sufficient information. Note that when faces are clearly presented, only specific facial feature models are represented, while categorical models show no contribution to occipitotemporal representations. On the contrary, when faces are briefly presented, the configural models do not contribute unique information, and only the high-level expression model is significant in the partial correlation analysis.

Although correlation coefficients between the models and neural data are generally low (Supplementary Table 3), the noise ceiling shows that the maximal correlation possible with our data is also low (mean $\rho$=0.21); this is not surprising, considering the low $\rho$-values usually found in MEG RSA studies, and the fact that our paradigm involved complex, high-level visual stimuli and a demanding task. In this case, the noise ceiling serves as a useful benchmark for the explanatory power of our models. For example, the behavioural RDM reaches the noise ceiling in the 150 ms condition, but not for briefer stimuli, suggesting that behavioural representations fully explain the data when stimuli are clearly perceived. The local configuration model also shows good explanatory power at its earliest stage, and the same is true for the global model for a brief time window. With time, both models fall away from the noise ceiling, while other significant models also fail to fully explain the data (Supplementary Figure 4).

Given the complex face processing and task-related activity reflected by the MEG patterns, it is not surprising that most models do not approach the noise ceiling. In fact, the explanatory power of the configural models at early stages (100-200 ms) is striking, as is the strength of behavioural representations in ventral stream within 400 ms. Furthermore, the initial peak in performance of the behavioural model overlaps with the peak of the local configuration model. Together with the shared variance between configuration, expression and behaviour shown in the variance partitioning analysis (Figure 5D), this points to the role played by facial configuration in the extraction of emotional cues essential in the expression discrimination task.

# 3 Discussion

In the current study, we investigated how face representations in MEG sensor-level and source-space patterns vary with expression and with stimulus presentation duration. Using MVPA, we found a response to faces presented for 10 ms occurring at M170 latencies outside of subjective awareness, but no such response to expression. Furthermore, neural

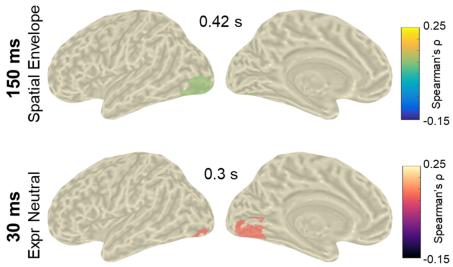



Figure 8: Significant correlations between: (1) MEG patterns for the 150 ms condition and the spatial envelope model RDM (**top**); (2) MEG patterns for the 30 ms condition and the high-level neutral-vs-others model (**bottom**). Only left hemisphere searchlights correlate with the two models. Maps are shown for the onset time of significant correlation, as clusters are sustained until offset (top: 0.54 s, bottom: 0.36 s).

responses became more transient when presentation time was reduced. Finally, we showed that behaviour and face configuration drive representations in face-responsive occipitotemporal cortex, with temporal dynamics varying as a function of stimulus duration.

### 3.1 Face and expression processing with limited visual input

When decoding faces and scrambled stimuli, we found early effects for 150 ms and 30 ms faces (~100 ms), as well as above-chance decoding of 10 ms faces shown outside of subjective awareness (147 - 350 ms), in line with previous studies showing evidence of face perception outside of awareness (Axelrod et al., 2015). Furthermore, temporal representations underpinning classification performance varied with stimulus duration: for 150 ms faces, a sustained representation emerged at M170 latencies which was absent for 30 ms faces. This suggests that clearly presented faces are perceived through a multi-stage process, while disrupted recurrent processing leads to delayed stable representations.

Conscious perception may be supported by temporally stable representations, while processing of stimuli outside subjective awareness may require a sequence of transient stages (Dehaene, 2016). Since above-chance decoding of 10 ms faces is transient in the current study, temporal generalization reveals only few transient clusters along the diagonal. On the other hand, the patterns differentiating 30 ms and 150 ms faces suggest that longer stimulus durations elicit an earlier stable representation, reflective of conscious perception and likely to be supported by recurrent processes. It has previously been suggested that faster stimulus presentation leads to more transient representations (Mohsenzadeh et al., 2018); however, since the backward masking procedure used here disrupts the formation of a stable representation by entering the visual stream, it is unclear whether different methods of preventing awareness would lead to the same results.

Alternative explanations are possible when interpreting temporal generalization patterns. First, SNR decreases as a function of stimulus duration, and this could lead to lower accuracies and less sustained representations. However, we find that the most striking difference in temporal generalization patterns occurs at M170 latencies, which is the time window exhibiting comparable decoding accuracies between 150 ms and 30 ms faces. Thus, the transient patterns characterizing M170 responses for rapidly presented faces are more likely to reflect a change in temporal dynamics. Second, it has been suggested that the transience of neural states can be overestimated in temporal generalization decoding due to trial-to-trial variability in effect onsets (Vidaurre et al., 2018); but given that the conditions we are comparing differ only in stimulus duration, the progression from sustained to transient observed here is unlikely to be explained by differences in onset variability.

Information supporting face decoding outside of subjective awareness was localized mainly to occipital cortex in our searchlight source-space decoding analysis (Figure 3C). Given the suppression of sustained neural activity in backward masking, the early stages of this response can be attributed to either purely feedforward activity, or to feedback connections, which have been shown to target V1 at early stages of recurrent processing (Wyatte et al., 2014; Mohsenzadeh et al., 2018). If backward masking truly disrupts recurrent processing when associated with a lack of visual awareness (Lamme et al., 2002; Boehler et al., 2008), a feedforward pattern (or one based on local recurrent circuits) is the most likely explanation. Furthermore, the fact that we detect a response to faces, and not to expression, suggests that two different stages of identification and categorization may be supported by qualitatively different mechanisms. It is still the subject of debate whether feedforward processing can support categorization (DiCarlo et al., 2012; Howe, 2017), and our results support the idea that some degree of recurrent processing is necessary (Maguire and Howe, 2016; Lamme and Roelfsema, 2000).

Note that the spatial resolution of MEG prevents us from drawing strong conclusions on the origin of this response to faces. Furthermore, recent observations have been made about concerns of information spreading in source-space MVPA analyses of MEG data, potentially overestimating the spatial extent of effects (Sato et al., 2018). In the current study, we restricted our source-space decoding analysis to localizing effects identified at the sensor level, and we applied randomization testing with an omnibus threshold in order to avoid spurious effects (4.9) and to alleviate the trade-off between maximizing information and reducing false positives.

All expressions presented for at least 30 ms were decodable from MEG data. In a previous MEG study involving passive viewing of emotional faces (Dima et al., 2018a), we found early (M100) decoding of angry expressions compared to happy and neutral faces. Here, we show early decoding of both face and expression (~100 ms), with only a slight advantage for angry expressions (107 ms; Supplementary Table 2), suggesting a contribution of task-related effects to early visual processing. Furthermore, it is important to note that all analyses described here were performed across facial identity and that stimuli were controlled in terms of low-level properties. The current results thus support the idea that expression categorization begins at the early stages of visual perception with preferential processing of emotional cues.

### 3.2 Expression and awareness: no subliminal threat advantage

In this experiment, we measured subjective visual awareness using a perceptual awareness scale. Subjective and objective measures of awareness both have their strengths and limitations; although subjective measures pose a criterion problem (Szczepanowski and Pessoa, 2007), objective measures (such as performance on a forced-choice task) may reflect unconscious processing (Lau, 2008; Song and Yao, 2016; Wierzchoń et al., 2014). We restricted our experiment to subjective awareness, shown to be effectively captured by perceptual awareness scales, particularly when employed after discrimination tasks (Wierzchoń et al., 2014; Sandberg et al., 2010). Here, the discrimination task was used to verify subjects' compliance and assess the presence of potential expression biases in responses given to subliminal faces (2.1).

It is not surprising that we detected a subliminal response to faces outside of subjective awareness, considering the wealth of evidence on non-conscious face processing (Axelrod et al., 2015). However, given our definition of visual awareness, the absence of a subliminal expression effect is striking. Despite existing evidence of non-conscious expression processing, some studies cast doubt on this effect (Schlossmacher et al., 2017; Hedger et al., 2016; Koster et al., 2007; Pessoa et al., 2006; see also inconsistent results for angry faces discussed in 1). In our study, three main aspects may have contributed to this result: (1) stimuli were normalized in terms of low-level properties, minimizing the detection of visual differences at early stages of perception; (2) we used a cross-identity classification approach, ensuring that we investigate categorical differences; (3) we used a very short stimulus presentation time, reducing the amount of information available to the visual system and limiting the possibility of residual awareness.

Furthermore, although absence of evidence cannot be taken as evidence of absence, we were able to detect a subliminal response to faces despite a lower number of scrambled trials, as well as expression effects to faces presented for longer than 10 ms (using similarly sized datasets). As the MVPA framework and the analysis pipeline were chosen to maximize signal and statistical power, it is likely that this result reflects a true absence of an effect in our data. Although MEG patterns reflect non-consciously extracted facial features, the extraction of expression-specific cues from our controlled stimulus set appears to require some conscious processing. If the face-specific response is supported by feedforward processing, the absence of expression-related information would indicate that the extraction of features and categories may require sustained activity.

### 3.3 Ventral stream representations of behaviour and face configuration

To understand the representations underlying our decoding results, we investigated the similarity between MEG patterns and models based on behavioural performance, as well as facial expression, identity, configuration, and spatial envelope.

We found that ventral stream areas encoded sustained and extensive behavioural representations as early as 120 ms after stimulus onset (Figure 6). This suggests that the extraction of features essential in behavioural decision-making is a rapid process accomplished in face-responsive cortex, similarly to evidence found in higher-level object and scene perception (Walther et al., 2009; Cohen et al., 2017; Bankson et al., 2018; Groen et al., 2018) and in line with previous studies showing that the perceptual similarity of faces is represented in neural patterns (Said et al., 2018; Furl et al., 2017). For faces presented for 150 ms, behavioural correlations emerge within 400 ms after stimulus presentation, suggesting that representations leading to behavioural response are rapidly formed in the ventral stream.

Moreover, we found representations of face configuration in ventral stream areas, with first-order features being represented earlier and followed by second-order features. Facial configuration has long been thought to play an important part in identity and expression perception (Calder et al., 2000), and in our RSA analysis the configural models show some of the strongest contributions among the nine models tested. In fact, we show that with the exception of a brief time window, no "categorical" representations, as quantified by the high-level models, are formed in occipitotemporal cortex; instead, configural representations appear to overlap with representations of behaviour, suggesting that it is face configuration that drives expression-selective responses in ventral stream areas and guides

behaviour.

The contribution of local features prior to the global configuration model adds to evidence suggesting that emotional face perception is supported by the processing of diagnostic features, such as the eyes and mouth (Fox and Damjanovic, 2006; Wegrzyn et al., 2017). Recent studies have shown that the recognition of familiar faces may not rely on holistic face processing, but on specific features (Visconti Di Oleggio Castello et al., 2017; Mohr et al., 2018), and it has been suggested that responses in face-selective areas such as the OFA may represent faces in terms of topological maps or feature-based models (Henriksson et al., 2015). Particularly for expression perception, feature-based processing provides an efficient mechanism for the rapid extraction of visual cues essential in human interaction, as reflected by the ability of the Action Unit coding system to quantify facial expressions (Ekman and Friesen, 1977; Srinivasan et al., 2016). However, we note that the Action Unit model RDM assessed here did not significantly correlate with the MEG patterns, probably due to the static and brief nature of our stimuli.

Previous studies have shown differential modulation of ERP components by first-order and second-order face configuration. Some studies have shown early components (P1, N170) to encode the former only (e.g. Mercure et al., 2008; Zion-Golumbic and Bentin, 2007), while others have also shown effects of second-order configuration at N170 latencies (Eimer et al., 2011). Furthermore, fMRI studies have reported a division of labour in the face-selective network, with the FFA thought to play a special role in representing both types of configural information (Liu and Ioannides, 2010; Golarai et al., 2015). Recently, it has been suggested that featural and configural processing of even non-face objects elicit face-like responses in the OFA and FFA (Zachariou et al., 2018). Here, we combined the strengths of source-localized MEG data and the RSA framework to tease apart the two models using a single stimulus set. The searchlight RSA analysis revealed that the two models overlap spatially in a right ventral stream area corresponding to the FFA, but are dissociated temporally: for 150 ms faces, representations switch from first-order to second-order at ~300 ms after stimulus onset, bringing together previous fMRI and electrophysiological findings.

Furthermore, this two-stage process appears to depend on the amount of information available to the visual system. For 150 ms faces, local and global configuration models make unique, temporally distinct contributions to explaining the data, as shown in the partial correlation analysis. For 30 ms faces, no unique variance is explained by the two models; furthermore, representations are temporally overlapping in the correlation

analysis and occur after 300 ms (Figure 7). This complements our sensor-level temporal generalization findings: 30 ms faces are processed through a series of transient coding steps at early stages and a stable representation is formed after 300 ms, when both first-order and second-order features are represented. On the other hand, for 150 ms faces, a two-stage process takes place, with an initial stable representation emerging at M170 latencies and supported mainly by first-order features, and a later representation after 300 ms encoding second-order configuration. Feature representations thus appear to be linked to the late emergence of stable representatons, thought to be reflective of recurrent processing and categorization (Mohsenzadeh et al., 2018; Tang and Kreiman, 2017). Importantly, this idea is supported by spatially and temporally overlapping behavioural representations in ventral stream areas.

Together, the findings we present here constitute a stepping stone towards a better understanding of high-level representations in face perception. While binary categorical models can estimate high-level representations and task-related processing, the code supporting visual perception is likely to be better understood in terms of behavioural goals and the visual features supporting them. We show that face-responsive cortex dynamically encodes facial configuration starting with first-order features, and that this supports behavioural representations when participants are performing an expression discrimination task. Furthermore, we show that the cascade of processing stages changes with stimulus duration, pointing to the adaptability of the face processing system in achieving goals when visual input is limited. Finally, although we find evidence of a subliminal response to faces, we do not detect a subliminal response to expression. Our results bridge findings from previous fMRI and electrophysiological research, revealing the spatiotemporal structure of face representations in human occipitotemporal cortex.

# 4 Materials and Methods

## 4.1 Participants

The participants were 25 healthy volunteers (16 female, age range 19-42, mean age 25.6 $\pm$ 5.39). All volunteers gave written consent to participate in the study in accordance with The Code of Ethics of the World Medical Association (Declaration of Helsinki). All procedures were approved by the ethics committee of the School of Psychology, Cardiff University.

## 4.2 Stimuli

Stimuli were 20 faces with angry, neutral and happy expressions (10 female faces; model numbers: 2, 6, 7, 8, 9, 11, 14, 16, 17, 18, 22, 23, 25, 31, 34, 35, 36, 38, 39, 40) from the NIMSTIM database (Tottenham et al., 2009). The eyes were aligned using automated eye detection as implemented in the Matlab Computer Vision System toolbox (Mathworks, Inc., Natick, Massachusetts). An oval mask was used to crop the faces to a size of $378 \times 252$ pixels subtending $2.6 \times 3.9$ degrees of visual angle. All images were converted to grayscale. Their spatial frequency was matched by specifying the rotational average of the Fourier amplitude spectra as implemented in the SHINE toolbox (Willenbockel et al., 2010), and Fourier amplitude spectra for all faces were set to the average across the face set.

Masks and control stimuli were created by scrambling the phase of all face images in the Fourier domain. This was achieved by replacing the phase information in each of the images with phase information from a white noise image of equal size (Perry and Singh, 2014). To ensure matched low-level properties between face and control stimuli, pixel intensities were normalized between each image and its scrambled counterpart, using the minimum and maximum pixel intensity of the scrambled image.

## 4.3 Experimental design

At the start of each trial, a white fixation cross was centrally presented on an isoluminant gray background. Its duration was pseudorandomly chosen from a uniform distribution between 1.3 and 1.6 s. A face stimulus was then centrally presented with a duration of either 10 ms, 30 ms or 150 ms; the stimulus was followed by a phase-scrambled mask with a duration of 190 ms, 170 ms or 50 ms respectively (for a constant total stimulus duration of 200 ms). In each block, 10 trials contained no face; instead, a phase-scrambled control stimulus was flashed for 10 ms and followed by another mask.

After a 500 ms delay intended to dissociate face perception from response preparation, participants had to correctly select the expression they had perceived out of three alternatives presented on screen (Figure 1A). They had 1.5 seconds to make a button press; if they were sure that no face had been presented, they could refrain from responding. The mapping of the response buttons to emotional expressions changed halfway through the experiment so as to ensure that emotional expression processing would not be confounded by specific motor preparation effects.

Next, participants had to rate how clearly they had seen the face using a 3-point scale starting from 0. They were instructed to only select 0 if no face had been perceived, 1

if they had perceived a face but not clearly, and 2 if they had clearly perceived the face. They had 2 seconds to make this response.

In each of four blocks, each face was presented once with each of the three possible stimulus durations. We thus collected 80 trials per condition, except for the control condition (containing no face) which only had 40 trials.

## 4.4 Data acquisition

All participants with one exception acquired a whole-head structural MRI using a 1 mm isotropic Fast Spoiled Gradient-Recalled-Echo pulse sequence.

Whole-head MEG recordings were made using a 275-channel CTF radial gradiometer system (CTF, Vancouver, Canada) at a sampling rate of 1200 Hz. Four of the sensors were turned off due to excessive sensor noise. An additional 29 reference channels were recorded for noise rejection purposes and the primary sensors were analyzed as synthetic third-order gradiometers (Vrba and Robinson, 2001).

Stimuli were presented using a ProPixx projector system (VPixx Technologies, Saint-Bruno, Canada) with a refresh rate set to 100 Hz. Images were projected to a screen with a resolution of 1920 x 1080 pixels situated at a distance of 1.2 m from the participant. Recordings were made in four blocks of approximately 15 minutes each, separated by short breaks. The data were collected in 2.5 s epochs beginning 1 s prior to stimulus onset.

Participants were seated upright while viewing the stimuli and electromagnetic coils were attached to the nasion and pre-auricular points on the scalp in order to continuously monitor head position relative to a fixed coordinate system on the dewar. To help co-register the MEG data with the participants' structural MRI scans, we defined the head shape of each subject using an ANT Xensor digitizer (ANT Neuro, Enschede, Netherlands). An Eyelink 1000 eye-tracker system (SR Research, Ottawa, Canada) with a sampling rate of 1000 Hz was used to track the subjects' right pupil and corneal reflex.

## 4.5 Behavioural analysis

The effect of stimulus duration and emotional expression on participants' expression discrimination accuracy (percentage correct responses) was analyzed after applying a rationalized arcsine transformation (Studebaker, 1985) using a 3x3 repeated-measures ANOVA with factors Duration (levels: 10 ms, 30 ms, and 150 ms) and Expression (levels: angry, happy, and neutral).

## 4.6 MEG multivariate pattern analysis (MVPA)

To test for differences between conditions present in multivariate patterns, we used a linear Support Vector Machine (SVM) classifier with L2 regularization and a box constraint $c = 1$. The classifier was implemented in Matlab using LibLinear (Fan et al., 2008) and the Statistics and Machine Learning Toolbox (Mathworks, Inc.). We performed binary classification on (1) responses to neutral faces versus scrambled stimuli (face decoding); (2) all three pairs of emotional expressions (expression decoding).

For face decoding, time-resolved classification was performed separately for each stimulus duration. To assess the presence of subjectively non-conscious responses, the classification of faces presented for 10 ms was performed after excluding any trials reported as containing a face. To ensure that decoding results were not biased by stimulus repetitions or recognition of face identities across the training and test sets, cross-exemplar five-fold cross-validation was used to assess classification performance: the classifier was trained on 16 of the 20 face identities and 8 of the 10 scrambled images, and tested on the remaining 4 faces and 2 scrambled exemplars.

To assess similarities between responses across stimulus duration conditions, face cross-decoding was also performed, whereby a decoder was trained on 150 ms faces and tested on 30 ms faces and viceversa. The analysis was repeated for all pairs of conditions, using cross-exemplar cross-validation to ensure true generalization of responses; the resulting accuracies were averaged across the two training/testing directions, which led to similar results.

The temporal structure of face-related was assessed through temporal generalization decoding (King and Dehaene, 2014). Classifier models were trained on each sampled time point between -0.1 and 0.7 s and tested on all time points in order to evaluate the generalizability of neural patterns over time at each stimulus duration. For this analysis, a cross-exemplar hold-out procedure was used to speed up computation (the training and test sets each consisted of 10 face identities/5 scrambled exemplars).

For expression decoding, classification was separately applied to all pairs of emotional expression conditions for each stimulus duration. As low trial numbers were a limitation of the study design, we increased the power of our analysis by also pooling together trials containing faces shown for 30 ms and 150 ms (which were shown to share representations in the cross-decoding analysis). Performance was evaluated using five-fold cross-exemplar cross-validation. Note that splitting the datasets according to perceptual rating led to largely similar results (Supplementary Figure 3).

To achieve equal class sizes in face decoding, face trials were randomly subsampled (after cross-exemplar partitioning) to match the number of scrambled trials. For expression classification, trial numbers did not significantly differ between conditions after artefact rejection ($F(1.92, 46.18) = 0.15, P = 0.85, \eta^2 = 0.0062$).

## 4.7 MEG sensor-level analyses

MEG data were analyzed using Matlab (Mathworks, Inc.) and the Fieldtrip toolbox (Oostenveld et al., 2011). Prior to analysis, trials containing excessive eye or muscle artefacts were excluded based on visual inspection, as were trials exceeding 5 mm in head motion (quantified as the displacement of any head coil between two sampled time points). Using eyetracker information, we also excluded trials containing saccades and fixations away from stimulus or blinks during stimulus presentation. A mean of 8.71% $\pm$9.4% of trials were excluded based on this procedure.

For all analyses, MEG data were downsampled to 300 Hz and baseline corrected using the 500 ms before stimulus onset. A low-pass filter was applied at 100 Hz and a 50 Hz comb filter was used to remove the mains noise and its harmonics.

To improve SNR (Grootswagers et al., 2017), each dataset was divided into 20 equal partitions and pseudo-trials were created by averaging the trials in each partition. This procedure was repeated 10 times with random assignment of trials to pseudo-trials and was performed separately for the training and test sets.

To improve data quality, we performed multivariate noise normalization (MNN; Guggenmos et al., 2018). The time-resolved error covariance between sensors was calculated based on the covariance matrix ($\Sigma$) of the training set ($X$) and used to normalize both the training and test sets, in order to downweight MEG channels with higher noise levels (Equation 1).

$$X^* = \Sigma^{-\frac{1}{2}} X \tag{1}$$

In sensor-level MVPA analyses, all 271 MEG sensors were included as features and decoding was performed for each sampled time point between -0.1 and 0.7 s around stimulus onset.

## 4.8 MEG source-space analyses

For source analyses, participants' MRI was coregistered to the MEG data by marking the fiducial coil locations on the MRI and aligning the digitized head shape to the MRI with Fieldtrip. Note that the participant who had not acquired an MRI was excluded

from source-space analyses. MEG data were projected into source space using a vectorial Linearly Constrained Minimum Variance (LCMV) beamformer (Van Veen et al., 1997). To reconstruct activity at locations equivalent across participants, a template grid with a 10 mm isotropic resolution was defined using the MNI template brain and was warped to each participant's anatomical MRI. The covariance matrix was calculated based on the average of all trials across conditions bandpass-filtered between 0.1 and 100 Hz; this was then combined with a single-shell forward model to create an adaptive spatial filter, reconstructing each source as a weighted sum of all MEG sensor signals (Hillebrand et al., 2005). To alleviate the depth bias in MEG source reconstruction, beamformer weights were normalized by their vector norm (Hillebrand et al., 2012).

To improve data quality, MNN was included in the source localization procedure. As beamforming constructs a common filter based on pooled data (thus introducing no condition-related bias), the error covariance was in this case also calculated based on the pooled data. We then multiplied the normalized beamformer filters by the error covariance matrix, ensuring that the filters downweighted sensors with higher noise levels. The time-courses of virtual sensors were then reconstructed at all locations in the brain by multiplying the sensor-level data by the corresponding weighted filters. This resulted in three time-courses for each source, containing each of the three dipole orientations, which were concatenated for use in the MVPA analysis in order to maximize classification performance (Gohel et al., 2018). Preprocessing (baseline correction and downsampling) was performed as for sensor-level analyses.

A searchlight approach was used in source-space classification, whereby clusters with a 10 mm radius were entered separately into the decoding analysis. To exclude sources outside the brain and in regions such as the cerebellum, we restricted our searchlight analysis to 1256 sources included in the 90-region Automated Anatomical Clustering (AAL) atlas (Tzourio-Mazoyer et al., 2002). Given the 10 mm resolution of our sourcemodel, this amounted to a maximum of 27 neighbouring sources being included as features (mean 26.9, median 27, SD 0.31). Decoding of subliminal faces vs. scrambled stimuli was performed on 30 ms time windows with a 3 ms overlap using the time windows identified in sensor-space decoding in order to reduce computational cost.

We also performed supraliminal face decoding (150 ms faces vs. scrambled stimuli) in order to identify a face-responsive ROI for use in the RSA analysis. This was accomplished by identifying searchlights achieving a cross-subject accuracy above the 99.5th percentile ($P<0.005$, 66 searchlights; Figure 2).

## 4.9 Significance testing

We evaluated decoding performance using the averaged accuracy across subjects (proportion correctly classified trials) and assessed its significance through randomization testing (Nichols and Holmes, 2001; Jamalabadi et al., 2016; Noirhomme et al., 2014).

For sensor-level decoding, we repeated the cross-exemplar decoding procedure with 1,000 label shuffling iterations across the training and test sets. To speed up computation, the null distribution was estimated based on the time point achieving maximum overall accuracy in the MVPA analysis (Dima et al., 2018b). Observed time-resolved accuracies were then compared to the group maps to calculate P-values. For all analyses, omnibus correction was applied across tests, time points and sources where applicable (Nichols and Holmes, 2001; Singh et al., 2003), with a supplementary false discovery rate correction applied for tests where the null distribution was not separately estimated.

For whole-head sensor-space decoding, p-values were calculated using the maximal null distribution across tests and corrected with a false discovery rate of 0.05, and a threshold of at least 5 consecutive significant time points was imposed. For temporal generalization decoding, the maximal distribution was created across tests and time points, and contiguous clusters of at least $5^2$ time points were considered significant.

To detect above-chance decoding of subliminal faces in source space, we performed 100 randomization iterations for each source cluster and subject in order to minimize computational cost. We then randomly combined the individual randomized accuracies into $10^3$ whole-brain group maps (Stelzer et al., 2013). The analysis was restricted to time windows identified at sensor level (140-350 ms). A minimal extent of three consecutive time windows with a FDR-corrected $P < 0.005$ was applied.

## 4.10 Representational Similarity Analysis (RSA)

### 4.10.1 Neural patterns and analysis framework

To interrogate the content of neural representations in space and time, we performed representational similarity analysis (RSA). For this analysis, MEG data were source reconstructed as described above and trials were sorted according to expression and face identity. RSA was performed separately for each stimulus duration and only trials containing faces were included in the analysis.

To offset computational cost, a searchlight analysis was performed using occipitotemporal sources identified in face decoding, with a temporal resolution of 30 ms, as in the source-space decoding analysis (4.8). All three dipole orientations were concatenated for

each source. The exclusion of responses to scrambled stimuli from the RSA ensured that feature selection was based on an orthogonal contrast (Figure 2).

To create MEG representational dissimilarity matrices (RDMs), we calculated the squared cross-validated Euclidean distance between all pairs of face stimuli (Guggenmos et al., 2018). Note that as the data were multivariately noise-normalized, this is equivalent to the squared cross-validated Mahalanobis distance (Walther et al., 2016). For each participant, the data were split into a training set (the first 2 sessions) and a test set (the last 2 sessions). The two stimulus repetitions contained in each set were averaged, and these were averaged across subjects to create training and test sets. To compute the cross-validated Euclidean distance between two stimulus patterns ($X^*$, $Y^*$), we calculated the dot products of pattern differences based on the training set and the test set (Equation 2). This procedure has the advantage of increasing the reliability of distance estimates in the presence of noise.

$$d^2(X^*, Y^*) = \sum_{i=1}^{n} (X_i^* - Y_i^*)_{train} (X_i^* - Y_i^*)_{test} \tag{2}$$

The spatiotemporally resolved MEG RDMs were then correlated with several model RDMs to assess the contribution of different features to neural representations. In an initial analysis, we calculated Spearman's rank correlation coefficients between each model RDM and the MEG RDM (Nili et al., 2014). To further investigate the unique contribution of each model, we entered the significantly correlated models based on visual features of the images into a partial correlation analysis, where each model's correlation to the MEG data was recalculated after partialling out the contribution of the other models.

Note that a model based on behaviour, which was also represented in the MEG data for all stimulus duration conditions, was not included in the partial correlation analysis; the rationale is that we were interested in the contribution of each visual property independently of the others, but we did not expect a unique contribution of behaviour in the absence of expression-related visual properties, and partialling out the behavioural model from the visual models would not be easily interpretable. Instead, we preferred to independently describe the correlations between behaviour and visual models, brain and behaviour, and brain and visual models, as the three main factors of interest in our analysis.

### 4.10.2 Model RDMs

We investigated the temporal dynamics of face perception by assessing the similarity between MEG patterns and 9 models quantifying behaviour and facial/visual properties (Figure 5).

To create behavioural model RDMs, we calculated the number of error responses made by each participant to each stimulus and summed these up to create a cross-subject behavioural RDM. For each stimulus duration, we created separate behavioural RDMs by calculating pairwise cross-validated Euclidean distances between error response patterns, using a cross-session training/test split as described above.

To create a high-level identity model, we assigned distances of 0 to pairs of face identities repeated across emotional expression conditions, and distances of 1 to pairs of different face identities. We used a similar strategy to create high-level emotional expression models. An all-versus-all model was created by assigning distances of 0 to all faces belonging to the same emotional expression condition, and distances of 1 to pairs of faces differing in emotion. We also tested a neutral-versus-others model by assigning distances of 0 to all emotional faces (happy + angry), and an angry-versus-others model by assigning distances of 0 to all benign faces (happy + neutral).

To account for variability in expression that is not captured by such high-level binary representations, we also tested a model based on Action Units. Action Units quantify changes in expression by categorizing facial movements (Ekman and Friesen, 1977). We used OpenFace (Baltrusaitis et al., 2016) to automatically extract the intensity of 12 Action Units in our image set (Supplementary Table 4), and we calculated pairwise Euclidean distances between these intensities for all pairs of faces in our stimulus set to obtain an Action Unit RDM.

To create face configuration RDMs, we first used OpenFace (Baltrusaitis et al., 2016) to automatically detect and label face landmarks. The software created 68 2D landmarks for each face. We removed landmarks corresponding to the face outline and the 2 outermost eyebrow landmarks, to account for cases in which these landmarks were cropped out by the oval mask used in the MEG stimulus set. The final landmark set consisted of 47 coordinates for 6 facial features (eyes, eyebrows, nose, and mouth), which were visually inspected to ensure that they were correctly marked. To capture feature-based (local) facial configuration, we calculated within-feature pairwise Euclidean distances between landmarks (Figure 5C). To quantify global face configuration, we calculated between-feature Euclidean distances (the distances between each landmark and all landmarks belonging

to different facial features). Distances were then concatenated to create feature vectors describing each face in terms of its local/global configuration, and Euclidean distances between them gave the final configural model RDMs. The local/global configurations correspond to the first-order (isolated) and second-order (relational) features in classic configural models of face perception (Diamond and Carey, 1986; Piepers and Robbins, 2012).

Finally, a spatial envelope model was created in order to capture image characteristics using the GIST descriptor (Oliva and Torralba, 2001). This procedure extracts 512 values per image by applying a series of Gabor filters at different orientations and positions, and thus quantifies the average orientation energy at each spatial frequency. To obtain the spatial envelope RDM, we calculated pairwise Euclidean distances between all images using the GIST values.

#### 4.10.3 Significance testing

To assess the significance of spatiotemporally resolved correlation maps, we used a randomization approach (4.9). Model RDMs were shuffled 1,000 times and correlations were recomputed for each of the 66 searchlights using the time window achieving the maximal correlation coefficient across models for each of the stimulus duration conditions. Since negative correlations were not expected and would not be easily interpretable, P-values were calculated using a one-sided test (Furl et al., 2017). To correct for multiple comparisons, P-values were omnibus-corrected by creating a maximal distribution of randomized correlation coefficients across searchlights, models and conditions, and FDR and cluster-corrected across timepoints ($\alpha = 0.05$, thresholded at 3 consecutive time windows).

#### 4.10.4 Variance partitioning

To gain more insight into the relationship between behavioural responses, expression categories and face configuration models, we used a variance partitioning approach (Greene et al., 2016; Groen et al., 2018). For each stimulus duration condition, the corresponding behavioural RDM was entered into a hierarchical multiple linear regression analysis, with three model RDMs as predictors: the two facial configuration models and the most correlated high-level expression model (10 ms: neutral-vs-others; 30 and 150 ms: angry-vs-others). These models were selected to reduce the predictor space before performing variance partitioning. To quantify the unique and shared variance contributed by each model, we calculated the $R^2$ value for every combination of predictors (i.e. all three mod-

els together, each pair of models separately, and each model separately). The EulerAPE software was used for visualization (Micallef and Rodgers, 2014; Figure 5).

## Acknowledgements

The authors would like to thank Lorenzo Magazzini and Gavin Perry for advice on the study design, and Dimitrios Pantazis for helpful comments on the manuscript. The study was supported by the UK MEG Partnership Grant (MRC/EPSRC, MR/K005464/1), CUBRIC and the School of Psychology at Cardiff University.

## Conflict of interest

The authors declare no competing interests.

## References

W. J. Adams, K. L. H. Gray, M. Garner, and E. W. Graf. High-Level Face Adaptation Without Awareness. *Psychological Science*, 21(2):205–210, 2010. doi:10.1177/0956797609359508.

L. Aguado, B. Valdés-Conroy, S. Rodríguez, F. J. Román, T. Diéguez-Risco, and M. Fernández-Cahill. Modulation of early perceptual processing by emotional expression and acquired valence of faces: An ERP study. *Journal of Psychophysiology*, 26(1): 29–41, 2012. doi:10.1027/0269-8803/a000065.

J. Almeida, P. E. Pajtas, B. Z. Mahon, K. Nakayama, and A. Caramazza. Affect of the unconscious : Visually suppressed angry faces modulate our decisions. *Cognitive, Affective, & Behavioral Neuroscience*, 13:94–101, 2013. doi:10.3758/s13415-012-0133-7.

V. Axelrod. The Fusiform Face Area: In Quest of Holistic Face Processing. *Journal of Neuroscience*, 30(26):8699–8701, 2010. doi:10.1523/JNEUROSCI.1921-10.2010.

V. Axelrod, M. Bar, and G. Rees. Exploring the unconscious using faces. *Trends in Cognitive Sciences*, 19(1):35–45, 2015. doi:10.1016/j.tics.2014.11.003.

T. Baltrusaitis, P. Robinson, and L.-P. Morency. OpenFace: an open source facial behaviour analysis toolkit. *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10, 2016.

B. Bankson, M. Hebart, I. Groen, and C. Baker. The temporal evolution of conceptual object representations revealed through models of behavior, semantics and deep neural networks. *NeuroImage*, 178:172–182, 9 2018. doi:10.1016/J.NEUROIMAGE.2018.05.037.

M. Behrmann, J. J. Richler, G. Avidan, and R. Kimchi. Holistic Face Perception. *Oxford Handbook of Perceptual Organization*, pages 1–21, 2014. doi:10.1093/oxfordhb/9780199686858.013.010.

C. Bertini, R. Cecere, and E. Làdavas. Unseen fearful faces facilitate visual discrimination in the intact field. *Neuropsychologia*, (June):0–1, 2017. doi:10.1016/j.neuropsychologia.2017.07.029.

C. N. Boehler, M. A. Schoenfeld, H.-J. Heinze, and J.-M. Hopf. Rapid recurrent processing gates awareness in primary visual cortex. *Proceedings of the National Academy of Sciences*, 105(25):8742–8747, 2008.

A. J. Calder, A. W. Young, J. Keane, and M. Dean. Configural information in facial expression perception. *Journal of Experimental Psychology: Human Perception and Performance*, 26(2):527–551, 2000. doi:10.1037/0096-1523.26.2.527.

M. A. Cohen, G. A. Alvarez, K. Nakayama, and T. Konkle. Visual search for object categories is predicted by the representational architecture of high-level visual cortex. *Journal of neurophysiology*, 117(1):388–402, 1 2017. doi:10.1152/jn.00569.2016.

S. Dehaene. Decoding the Dynamics of Conscious Perception : The Temporal Generalization Method. In B. G and C. Y, editors, *Micro-, Meso- and Macro-Dynamics of the Brain*, pages 85–97. Springer, 2016. doi:10.1007/978-3-319-28802-4.

R. Diamond and S. Carey. Why faces are and are not special: an effect of expertise. *Journal of experimental psychology*, 115(2):107–117, 1986.

J. J. DiCarlo, D. Zoccolan, and N. C. Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–34, 2 2012. doi:10.1016/j.neuron.2012.01.010.

D. C. Dima, G. Perry, E. Messaritaki, J. Zhang, and K. D. Singh. Spatiotemporal dynamics in human visual cortex rapidly encode the emotional content of faces. *Human brain mapping*, pages 1–14, 2018a. doi:10.1002/hbm.24226.

D. C. Dima, G. Perry, and K. D. Singh. Spatial frequency supports the emergence of categorical representations in visual cortex during natural scene perception. *NeuroImage*, 179:102–116, 2018b. doi:10.1016/j.neuroimage.2018.06.033.

M. Eimer, A. Gosling, S. Nicholas, and M. Kiss. The N170 component and its links to configural face processing: A rapid neural adaptation study. *Brain Research*, 1376: 76–87, 2 2011. doi:10.1016/J.BRAINRES.2010.12.046.

P. Ekman and W. Friesen. *Facial action coding system: a technique for the measurement of facial movement*. Consulting Psychologists Press, Palo Alto, CA, 1977.

R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*, 9(2008):1871–1874, 2008. doi:10.1038/oby.2011.351.

M. J. Farah, K. D. Wilson, and J. N. Tanaka. What Is " Special " About Face Perception ? *Psychological review*, 105(3):482–498, 1998. doi:10.1037//0033-295X.105.3.482.

L. Fisch, E. Privman, M. Ramot, M. Harel, Y. Nir, F. Andelman, M. Y. Neufeld, U. Kramer, I. Fried, and R. Malach. Neural "Ignition": Enhanced Activation Linked to Perceptual Awareness in Human Ventral Stream Visual Cortex. *Neuron*, 64(4):562–574, 2010. doi:10.1016/j.neuron.2009.11.001.Neural.

E. Fox and L. Damjanovic. The eyes are sufficient to produce a threat superiority effect. *Emotion*, 6(3):534–9, 2006. doi:10.1037/1528-3542.6.3.534.

N. Furl, M. Lohse, and F. Pizzorni-Ferrarese. Low-frequency oscillations employ a general coding of the spatio-temporal similarity of dynamic faces. *NeuroImage*, 157:486–499, 2017. doi:10.1016/j.neuroimage.2017.06.023.

B. Gohel, S. Lim, M.-Y. Kim, H. Kwon, and K. Kim. Dynamic pattern decoding of source-reconstructed MEG or EEG data: Perspective of multivariate pattern analysis and signal leakage. *Computers in Biology and Medicine*, 93(October 2017):106–116, 2018. doi:10.1016/j.compbiomed.2017.12.020.

G. Golarai, D. G. Ghahremani, J. L. Eberhardt, and J. D. E. Gabrieli. Distinct representations of configural and part information across multiple face-selective regions of the human brain. *Frontiers in psychology*, 6:1710, 2015. doi:10.3389/fpsyg.2015.01710.

K. L. H. Gray, W. J. Adams, N. Hedger, K. E. Newton, and M. Garner. Faces and awareness: low-level, not emotional factors determine perceptual dominance. *Emotion*, 13(3):537–44, 2013. doi:10.1037/a0031403.

M. R. Greene, C. Baldassano, A. Esteva, D. M. Beck, and L. Fei-fei. Visual Scenes are Categorized by Function. *Journal of Experimental Psychology: General*, 145(1):82–94, 2016. doi:10.1037/xge0000129.Visual.

K. Grill-Spector, K. S. Weiner, K. Kay, and J. Gomez. The Functional Neuroanatomy of Human Face Perception. *Annual Review of Vision Science*, 3:167–196, 2017. doi:10.1146/annurev-vision-102016.

I. I. Groen, M. R. Greene, C. Baldassano, L. Fei-Fei, D. M. Beck, and C. I. Baker. Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. *eLife*, 7:e32962, 2018. doi:10.7554/eLife.32962.

T. Grootswagers, S. G. Wardle, and T. A. Carlson. Decoding dynamic brain patterns from evoked responses: A tutorial on multivariate pattern analysis applied to time-series neuroimaging data. *Journal of cognitive neuroscience*, 29(4):677–697, 2017.

M. Guggenmos, P. Sterzer, and R. M. Cichy. Multivariate pattern analysis for MEG: A comparison of dissimilarity measures. *NeuroImage*, 173:434–447, 6 2018. doi:10.1016/J.NEUROIMAGE.2018.02.044.

N. Hedger, W. J. Adams, and M. Garner. Fearful faces have a sensory advantage in the competition for awareness. *Journal of Experimental Psychology: Human Perception and Performance*, 41(6):1748–1757, 2015. doi:10.1037/xhp0000127.

N. Hedger, K. L. H. Gray, M. Garner, and W. J. Adams. Are visual threats prioritised without awareness? A critical review and meta analysis involving 3 behavioural paradigms and 2696 observers. *Psychological Bulletin*, 142(9):934–968, 2016.

L. Henriksson, M. Mur, and N. Kriegeskorte. Faciotopy-A face-feature map with face-like topology in the human occipital face area. *Cortex*, 72:156–167, 2015. doi:10.1016/j.cortex.2015.06.030.

A. Hillebrand, K. D. Singh, I. E. Holliday, P. L. Furlong, and G. R. Barnes. A new approach to neuroimaging with magnetoencephalography. *Human brain mapping*, 25 (2):199–211, 6 2005. doi:10.1002/hbm.20102.

A. Hillebrand, G. R. Barnes, J. L. Bosboom, H. W. Berendse, and C. J. Stam. Frequency-dependent functional connectivity within resting-state networks

: An atlas-based MEG beamformer solution. *NeuroImage*, 59(4):3909–3921, 2012. doi:10.1016/j.neuroimage.2011.11.005.

P. D. L. Howe. Natural scenes can be identified as rapidly as individual features. *Atten*, 79:1674–1681, 2017. doi:10.3758/s13414-017-1349-y.

A. Ishai. Let's face it: It's a cortical network. *NeuroImage*, 40(2):415–419, 2008. doi:10.1016/j.neuroimage.2007.10.040.

H. Jamalabadi, S. Alizadeh, M. Sch, C. Leibold, and S. Gais. Classification Based Hypothesis Testing in Neuroscience : Below-Chance Level Classification Rates and Overlooked Statistical Properties of Linear Parametric Classifiers. *Human Brain Mapping*, 37(January):1842–1855, 2016. doi:10.1002/hbm.23140.

Y. Jiang and S. He. Cortical Responses to Invisible Faces: Dissociating Subsystems for Facial-Information Processing. *Current Biology*, 16(20):2023–2029, 2006. doi:10.1016/j.cub.2006.08.084.

Y. Jiang, R. W. Shannon, N. Vizueta, E. M. Bernat, and C. J. Patrick. Dynamics of processing invisible faces in the brain: Automatic neural encoding of facial expression information. *NeuroImage*, 44(3):1171–1177, 2009. doi:10.1016/j.neuroimage.2008.09.038.Dynamics.

R. I. Kemp, A. Caon, M. Howard, and K. R. Brooks. Improving Unfamiliar Face Matching by Masking the External Facial Features. *Applied Cognitive Psychology*, 30(4):622–627, 7 2016. doi:10.1002/acp.3239.

J.-R. King and S. Dehaene. Characterizing the dynamics of mental representations : the temporal generalization method. *Trends in Cognitive Sciences*, 18(4):203–210, 2014. doi:10.1016/j.tics.2014.01.002.

E. H. W. Koster, B. Verschuere, B. Burssens, R. Custers, and G. Crombez. Attention for Emotional Faces Under Restricted Awareness Revisited : Do Emotional Faces Automatically Attract Attention ? *Emotion*, 7(2):285–295, 2007. doi:10.1037/1528-3542.7.2.285.

V. A. F. Lamme and P. R. Roelfsema. The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences*, 23(11):571–579, 2000. doi:10.1016/S0166-2236(00)01657-X.

V. A. F. Lamme, K. Zipser, and H. Spekreijse. Masking Interrupts Figure-Ground Signals in V1. *Journal of Cognitive Neuroscience*, 14(7):1044–1053, 2002.

H. C. Lau. Are we studying consciousness yet? In *Frontiers of Consciousness*, pages 245–258. Oxford University Press, 10 2008. doi:10.1093/acprof:oso/9780199233151.003.0008.

L. Liu and A. A. Ioannides. Emotion separation is completed early and it depends on visual field presentation. *PLoS ONE*, 5(3), 2010. doi:10.1371/journal.pone.0009790.

C. A. Longmore, C. H. Liu, and A. W. Young. The importance of internal facial features in learning new faces. *Quarterly Journal of Experimental Psychology*, 68(2):249–260, 2 2015. doi:10.1080/17470218.2014.939666.

J. F. Maguire and P. D. L. Howe. Failure to detect meaning in RSVP at 27 ms per picture. *Attention, Perception, & Psychophysics*, 78(5):1405–1413, 7 2016. doi:10.3758/s13414-016-1096-5.

D. Maurer, R. L. Grand, and C. J. Mondloch. The many faces of configural processing. *Trends in Cognitive Sciences*, 6(6):255–260, 6 2002. doi:10.1016/S1364-6613(02)01903-4.

E. Mercure, F. Dick, and M. H. Johnson. Featural and configural face processing differentially modulate ERP components. *Brain Research*, 1239:162–170, 11 2008. doi:10.1016/J.BRAINRES.2008.07.098.

L. Micallef and P. Rodgers. eulerAPE: Drawing Area-Proportional 3-Venn Diagrams Using Ellipses. *PLoS ONE*, 9(7):e101717, 7 2014. doi:10.1371/journal.pone.0101717.

S. Mohr, A. Wang, A. D. Engell, and S. M. Hall. Early identity recognition of familiar faces is not dependent on holistic processing. *bioRxiv preprint*, (March 2018):1–27, 2018.

Y. Mohsenzadeh, S. Qin, R. M. Cichy, and D. Pantazis. Ultra-Rapid serial visual presentation reveals dynamics of feedforward and feedback processes in the ventral visual pathway. *eLife*, 7(e36329), 2018.

G. Namdar, G. Avidan, and T. Ganel. Effects of configural processing on the perceptual spatial resolution for face features. *Cortex*, 72:115–123, 2015. doi:10.1016/j.cortex.2015.04.007.

J. Navajas, M. Ahmadi, and R. Quian Quiroga. Uncovering the Mechanisms of Conscious Face Perception: A Single-Trial Study of the N170 Responses. *Journal of Neuroscience*, 33(4):1337–1343, 2013. doi:10.1523/JNEUROSCI.1226-12.2013.

T. E. Nichols and A. P. Holmes. Nonparametric Permutation Tests For Functional Neuroimaging : A Primer with Examples. *Human brain mapping*, 25(15):1–25, 2001.

H. Nili, C. Wingfield, A. Walther, L. Su, W. Marslen-Wilson, and N. Kriegeskorte. A Toolbox for Representational Similarity Analysis. *PLoS Computational Biology*, 10(4), 2014. doi:10.1371/journal.pcbi.1003553.

Q. Noirhomme, D. Lesenfants, F. Gomez, A. Soddu, J. Schrouff, G. Garraux, A. Luxen, C. Phillips, and S. Laureys. Biased binomial assessment of cross-validated estimation of classification accuracies illustrated in diagnosis predictions. *NeuroImage: Clinical*, 4: 687–694, 2014. doi:10.1016/j.nicl.2014.04.004.

A. Oliva and A. Torralba. Modeling the Shape of the Scene : A Holistic Representation of the Spatial Envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

R. Oostenveld, P. Fries, E. Maris, and J. M. Schoffelen. FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, 2011, 2011. doi:10.1155/2011/156869.

A. J. Pegna, A. Khateb, F. Lazeyras, and M. L. Seghier. Discriminating emotional faces without primary visual cortices involves the right amygdala. *Nature Neuroscience*, 8(1): 24–25, 2005. doi:10.1038/nn1364.

G. Perry and K. D. Singh. Localizing evoked and induced responses to faces using magnetoencephalography. *European Journal of Neuroscience*, 39(9):1517–1527, 2014. doi:10.1111/ejn.12520.

L. Pessoa. To what extent are emotional visual stimuli processed without attention and awareness? *Current Opinion in Neurobiology*, 15(2):188–196, 2005. doi:10.1016/j.conb.2005.03.002.

L. Pessoa and R. Adolphs. Emotion processing and the amygdala: from a 'low road' to 'many roads' of evaluating biological significance. *Nature Reviews Neuroscience*, 11(11): 773–783, 2010. doi:10.1038/nrn2920.

L. Pessoa, S. Japee, and L. G. Ungerleider. Visual Awareness and the Detection of Fearful Faces. *Emotion*, 5(2):243–247, 2005. doi:10.1037/1528-3542.5.2.243.

L. Pessoa, S. Japee, and D. Sturman. Target Visibility and Visual Awareness Modulate Amygdala Responses to Fearful Faces. *Cerebral Cortex*, 16(March):366–375, 2006. doi:10.1093/cercor/bhi115.

D. W. Piepers and R. A. Robbins. A review and clarification of the terms "holistic," "configural," and "relational" in the face perception literature. *Frontiers in Psychology*, 3:1–11, 2012. doi:10.3389/fpsyg.2012.00559.

J. E. Reiss and J. E. Hoffman. Disruption of early face recognition processes by object substitution masking. *Visual Cognition*, 15(7):789–798, 2007. doi:10.1080/13506280701307035.

J. J. Richler and I. Gauthier. A meta-analysis and review of holistic face processing. *Psychological Bulletin*, 140(5):1281–1302, 2014. doi:10.1037/a0037004.

V. Rodriguez, M. Brett, and M. Valdés-Sosa. Absence of Face-specific Cortical Activity in the Complete Absence of Awareness: Converging Evidence from Functional Magnetic Resonance Imaging and Event- related Potentials. *Journal of cognitive neuroscience*, 24 (2):396–415, 2013. doi:10.1162/jocn.

C. P. Said, C. D. Moore, A. D. Engell, and J. V. Haxby. Distributed representations of dynamic facial expressions in the superior temporal sulcus. *Journal of Vision*, 10(2010): 1–12, 2018. doi:10.1167/10.5.11.Introduction.

K. Sandberg, B. Timmermans, M. Overgaard, and A. Cleeremans. Measuring consciousness: Is one measure better than the other? *Consciousness and Cognition*, 19(4): 1069–1078, 2010. doi:10.1016/j.concog.2009.12.013.

M. Sato, O. Yamashita, M.-a. Sato, and Y. Miyawaki. Information spreading by a combination of MEG source estimation and multivariate pattern classification. *PLOS ONE*, 13(6):e0198806, 6 2018. doi:10.1371/journal.pone.0198806.

I. Schlossmacher, M. Junghöfer, T. Straube, and M. Bruchmann. No differential effects to facial expressions under continuous flash suppression: An event-related potentials study. *NeuroImage*, 163(August):276–285, 2017. doi:10.1016/j.neuroimage.2017.09.034.

K. D. Singh, G. R. Barnes, and A. Hillebrand. Group imaging of task-related changes in cortical synchronisation using nonparametric permutation testing. *NeuroImage*, 19: 1589–1601, 2003. doi:10.1016/S1053-8119(03)00249-0.

C. Song and H. Yao. Unconscious processing of invisible visual stimuli. *Scientific Reports*, 6(August):15–20, 2016. doi:10.1038/srep38917.

R. Srinivasan, J. D. Golomb, and A. M. Martinez. A Neural Basis of Facial Action Recognition in Humans. *Journal of Neuroscience*, 36(16):4434–4442, 2016. doi:10.1523/JNEUROSCI.1704-15.2016.

J. Stelzer, Y. Chen, and R. Turner. Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): Random permutations and cluster size control. *NeuroImage*, 65:69–82, 2013. doi:10.1016/j.neuroimage.2012.09.063.

P. Sterzer, L. Jalkanen, and G. Rees. Electromagnetic responses to invisible face stimuli during binocular suppression. *NeuroImage*, 46(3):803–808, 2009. doi:10.1016/j.neuroimage.2009.02.046.

G. Studebaker. A "rationalized" arcsine transform. *Journal of speech and hearing research*, 28:455–462, 1985. doi:10.1044/jshr.2803.455.

L. Su, E. Fonteneau, W. Marslen-wilson, and N. Kriegeskorte. Spatiotemporal Searchlight Representational Similarity Analysis in EMEG Source Space. In *Second International Workshop on Pattern Recognition in NeuroImaging Spatiotemporal*, 2012. doi:10.1109/PRNI.2012.26.

R. Szczepanowski and L. Pessoa. Fear perception: Can objective and subjective awareness measures be dissociated? *Journal of Vision*, 7(4):10–10, 2007. doi:10.1167/7.4.10.

M. Tamietto and B. De Gelder. Neural bases of the non-conscious perception of emotional signals. *Nature Reviews Neuroscience*, 11(10):697–709, 2010. doi:10.1038/nrn2889.

H. Tang and G. Kreiman. Recognition of Occluded Objects. In Q. Zhao, editor, *Computational and Cognitive Neuroscience of Vision*, number October. Springer, Singapore, 2017. doi:10.1007/978-981-10-0213-7.

N. Tottenham, J. W. Tanaka, A. C. Leon, T. McCarry, M. Nurse, T. a. Hare, D. J. Marcus, A. Westerlund, B. J. Casey, and C. Nelson. The NimStim set of facial expressions: judgments from untrained research participants. *Psychiatry research*, 168(3):242–9, 8 2009. doi:10.1016/j.psychres.2008.05.006.

N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot. Automated Anatomical Labeling of Activations in SPM Using a Macroscopic Anatomical Parcellation of the MNI MRI Single-Subject Brain. *NeuroImage*, 15:273–289, 2002. doi:10.1006/nimg.2001.0978.

B. Van Veen, W. van Drongelen, M. Yuchtman, and A. Suzuki. Localization of brain electrical activity via linearly constrained minimum variance spatial filtering. *IEEE Transactions on Biomedical engineering*, 44(9):867–880, 1997. doi:10.1109/10.623056.

D. Vidaurre, N. Myers, M. Stokes, A. C. Nobre, and M. W. Woolrich. Temporally unconstrained decoding reveals consistent but time-varying stages of stimulus processing. *bioRxiv*, page 260943, 2018. doi:10.1101/260943.

M. Visconti Di Oleggio Castello, K. G. Wheeler, C. Cipolli, and I. Gobbini. Familiarity facilitates feature-based face processing. *PLoS One*, 12(6):1–14, 2017.

J. Vrba and S. E. Robinson. Signal processing in magnetoencephalography. *Methods*, 25 (2):249–71, 10 2001. doi:10.1006/meth.2001.1238.

A. Walther, H. Nili, N. Ejaz, A. Alink, N. Kriegeskorte, and J. Diedrichsen. Reliability of dissimilarity measures for multi-voxel pattern analysis. *NeuroImage*, 137:188–200, 2016. doi:10.1016/j.neuroimage.2015.12.012.

D. B. Walther, E. Caddigan, L. Fei-Fei, and D. M. Beck. Natural Scene Categories Revealed in Distributed Patterns of Activity in the Human Brain. *Journal of Neuroscience*, 29(34):10573–10581, 2009. doi:10.1523/JNEUROSCI.0559-09.2009.

M. Wegrzyn, M. Vogt, B. Kireclioglu, J. Schneider, and J. Kissler. Mapping the emotional face . How individual face parts contribute to successful emotion recognition. *PLoS ONE*, 12(5):1–15, 2017.

M. Wierzchoń, B. Paulewicz, D. Asanowicz, B. Timmermans, and A. Cleeremans. Different subjective awareness measures demonstrate the influence of visual identification on perceptual awareness ratings. *Consciousness and Cognition*, 27(1):109–120, 2014. doi:10.1016/j.concog.2014.04.009.

V. Willenbockel, J. Sadr, D. Fiset, G. O. Horne, F. Gosselin, and J. W. Tanaka. Controlling low-level image properties: The SHINE toolbox. *Behavior Research Methods*, 42(3):671–684, 2010. doi:10.3758/BRM.42.3.671.

L. M. Williams, B. J. Liddell, J. Rathjen, K. J. Brown, J. Gray, M. Phillips, A. Young, and E. Gordon. Mapping the Time Course of Nonconscious and Conscious Perception of Fear: An Integration of Central and Peripheral Measures. *Human Brain Mapping*, 21(2):64–74, 2004. doi:10.1002/hbm.10154.

D. Wyatte, D. J. Jilk, and R. C. O'Reilly. Early recurrent feedback facilitates visual object recognition under challenging conditions. *Frontiers in Psychology*, 5:674, 7 2014. doi:10.3389/fpsyg.2014.00674.

V. Zachariou, Z. N. Safiullah, and L. G. Ungerleider. The Fusiform and Occipital Face Areas Can Process a Nonface Category Equivalently to Faces. *Journal of Cognitive Neuroscience*, pages 1–18, 2018. doi:10.1162/jocn_a_01288.

E. Zion-Golumbic and S. Bentin. Dissociated Neural Mechanisms for Face Detection and Configural Encoding: Evidence from N170 and Induced Gamma-Band Oscillation Effects. *Cerebral Cortex*, 17(8):1741–1749, 8 2007. doi:10.1093/cercor/bhl100.

# Supplementary Information

| | Sensor-space | | | Source-space |
|---|---|---|---|---|
| | 150 ms | 30 ms | 10 ms | 10 ms |
| Max % accuracy | 82.3 | 76.8 | 56.8 | 59.62 |
| SD (%) | 13.6 | 14.18 | 9.3 | 8.35 |
| Decoding onset (ms) | 100 | 100 | 147 | 120-150 |

Supplementary Table 1: Face decoding results.

| | Stimulus duration | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 150 ms | | | 30 ms | | | 10 ms | | | 30 + 150 ms | | |
| | A-N | H-N | A-H | A-N | H-N | A-H | A-N | H-N | A-H | A-N | H-N | A-H |
| Max % accuracy | 61.9 | 63.1 | 60.76 | 57.79 | 58.49 | 58.12 | 56.62 | 55.86 | 55.87 | 60.48 | 60.21 | 59.74 |
| SD (%) | 8.57 | 6.78 | 9.34 | 10.91 | 9.92 | 10.38 | 10.88 | 9.11 | 13.66 | 9.04 | 10.52 | 13.41 |
| Decoding onset (ms) | 180 | 113 | 220 | 437 | 120 | 633 | N/A | N/A | N/A | 107 | 113 | 117 |
| | Perceptual rating | | | | | | | | | | | |
| | 2 | | | 1 | | | 0 | | | 2 + 1 | | |
| | A-N | H-N | A-H | A-N | H-N | A-H | A-N | H-N | A-H | A-N | H-N | A-H |
| Max % accuracy | 59.55 | 62.54 | 64.03 | 56.56 | 56.88 | 56.63 | 57.64 | 55.32 | 56.01 | 60.43 | 62.25 | 60.24 |
| SD (%) | 12.24 | 11.6 | 10.82 | 12.1 | 13.63 | 13.21 | 14.46 | 10.24 | 12.47 | 11.95 | 12.07 | 12.25 |
| Decoding onset (ms) | 230 | 113 | 523 | 307 | 120 | 130 | N/A | N/A | N/A | 220 | 113 | 127 |

Supplementary Table 2: Expression decoding results.

| *Model* | Behavioural | Expression | | Spatial Envelope | | Global Config | | Local Config | |
|---|---|---|---|---|---|---|---|---|---|
| **150 ms** | $\rho$ | $\rho$ | $\rho_{part}$ | $\rho$ | $\rho_{part}$ | $\rho$ | $\rho_{part}$ | $\rho$ | $\rho_{part}$ |
| Max rho | 0.23 | 0.14 | 0.14 | 0.17 | 0.17 | 0.18 | 0.17 | 0.18 | 0.16 |
| SD | 0.12 | 0.03 | 0.03 | 0.06 | 0.06 | 0.09 | 0.08 | 0.07 | 0.06 |
| Onset (ms) | 120 | N/A | N/A | 420 | 390 | 300 | 330 | 120 | 120 |
| Offset (ms) | 390 | N/A | N/A | 540 | 540 | 450 | 390 | 300 | 180 |
| **30 ms** | $\rho$ | $\rho$ | $\rho_{part}$ | $\rho$ | $\rho_{part}$ | $\rho$ | $\rho_{part}$ | $\rho$ | $\rho_{part}$ |
| Max rho | 0.17 | 0.14 | 0.14 | 0.13 | 0.12 | 0.15 | 0.13 | 0.14 | 0.12 |
| SD | 0.07 | 0.03 | 0.03 | 0.05 | 0.04 | 0.07 | 0.03 | 0.05 | 0.04 |
| Onset (ms) | 210 | 300 | 300 | N/A | N/A | 390 | N/A | 360 | N/A |
| Offset (ms) | 540 | 360 | 360 | N/A | N/A | 450 | N/A | 510 | N/A |
| **10 ms** | $\rho$ | $\rho$ | $\rho_{part}$ | $\rho$ | $\rho_{part}$ | $\rho$ | $\rho_{part}$ | $\rho$ | $\rho_{part}$ |
| Max rho | 0.18 | 0.09 | 0.1 | 0.13 | 0.14 | 0.17 | 0.16 | 0.11 | 0.12 |
| SD | 0.04 | 0.03 | 0.03 | 0.04 | 0.05 | 0.04 | 0.04 | 0.03 | 0.04 |
| Onset (ms) | 120 | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 180 |
| Offset (ms) | 660 | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 240 |

Supplementary Table 3: RSA results for the 5 models achieving significant correlations.

| AU Code | Facial Action Coding System Name |
|---|---|
| AU01 | Inner brow raiser |
| AU02 | Outer brow raiser |
| AU04 | Brow lowerer |
| AU06 | Cheek raiser |
| AU09 | Nose wrinkler |
| AU10 | Upper lip raiser |
| AU12 | Lip corner puller |
| AU14 | Dimpler |
| AU15 | Lip corner depressor |
| AU17 | Chin raiser |
| AU20 | Lip stretcher |
| AU25 | Lips part |

Supplementary Table 4: Action Units (AU) used to create the Action Unit model RDM.

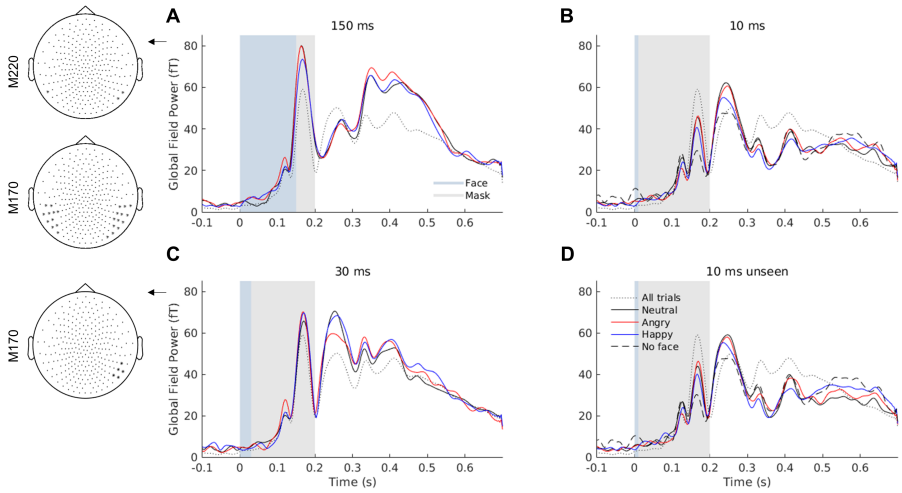
## Supplementary Analysis 1: Event-related field (ERF) analysis

We assessed the presence of difference between conditions in event-related fields (ERF). For the purposes of this analysis, MEG data were bandpass-filtered between 0.1 and 30 Hz and axial gradiometer event-related fields were averaged across subjects to calculate the global field power across all trials and conditions. This allowed us to determine three time windows of interest for evoked response component analysis: 63-137 ms (M100), 137-203 ms (M170), and 203 – 306 ms (M220).

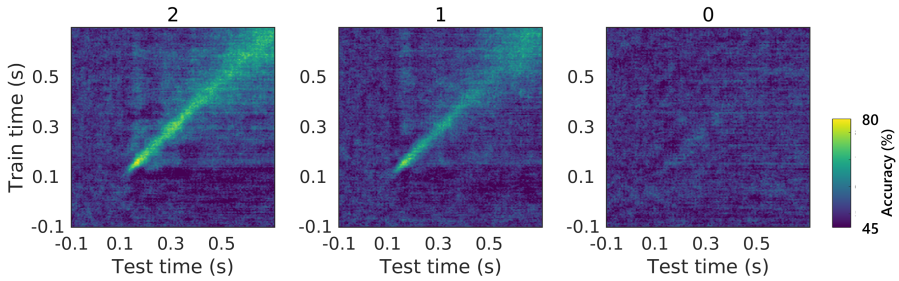
Next, we averaged evoked response fields for each condition and subject within the three time windows. We tested for differences between responses to faces and scrambled stimuli, and between responses to different emotional expressions, using paired t-tests and repeated-measures ANOVAs respectively at each sensor and time window. Significant sensors were determined using randomization testing (5000 iterations) and corrected for multiple comparisons using the maximal statistic distribution ($\alpha = 0.001$).

We assessed the presence of a response to faces by contrasting neutral faces with scrambled stimuli at each stimulus duration. For 150 ms faces, we found significant differences at M170 latencies and M220 latencies ($P < 0.0007, t(24) > 6.07$), but no significant effects at M100 latencies surviving our alpha of 0.001 (only one occipital sensor showed a non-significant effect with $P = 0.0059, t(24) = 4.89$). A significant, but smaller, cluster of right temporal sensors was also found for 30 ms faces at M170 latencies ($P < 0.0004, t(24) > 5.99$). No conclusive effects were found when contrasting faces presented for 10 ms with their scrambled counterparts, regardless of whether trials where a face was perceived were excluded or not ($P > 0.015, t(24) < 4.66$ across comparisons), and no effect of emotional expression was found at any of the stimulus durations

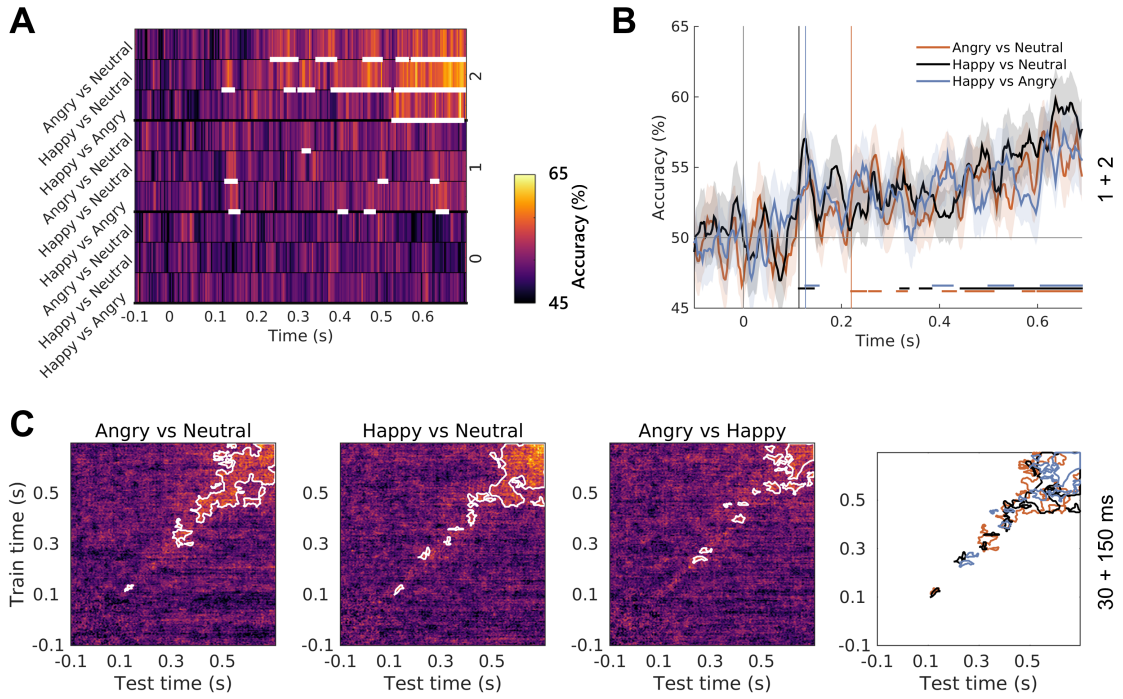
($P > 0.06, F(2, 48) < 8.59$). Several factors could explain the absence of emotional expression effects in our ERF data: (1) stimuli were highly controlled for low-level properties, minimizing visually-driven differences in early time windows; (2) our time windows of interest did not include late stages dominated by task-related processing of expression; (3) we performed a whole-brain analysis with a conservative correction for multiple comparisons.



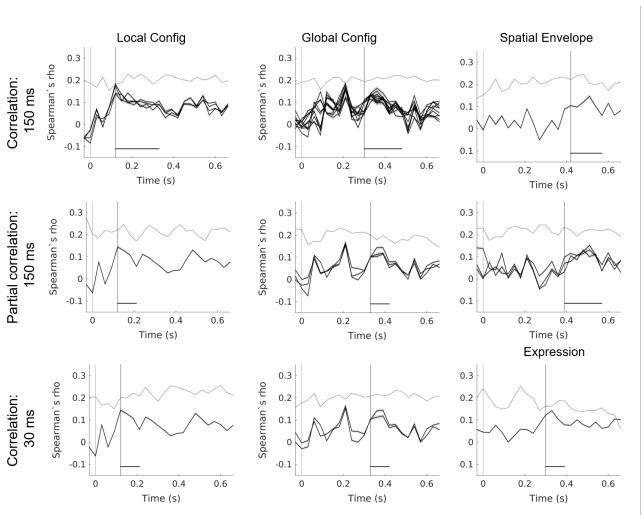

Supplementary Figure 1: ERF analysis results. **A-D**. Global field power averaged across participants and trials for each stimulus duration condition. Note decreasing M170 amplitudes with stimulus duration. **Left**. Significant sensors in the face vs scrambled contrast at M170 (137-203 ms) and M220 (203-306 ms) latencies ($P$<0.001 corrected).




Supplementary Figure 2: Face vs scrambled temporal generalization decoding for each perceptual rating category. The same progression from stable to transient representations is observed as when datasets were split according to stimulus duration.




Supplementary Figure 3: Expression decoding. **A**. Time-resolved decoding accuracy for each pair of expressions and perceptual rating, with above-chance time-windows highlighted in white ($P$<0.05 corrected). **B**. Accuracy time-courses obtained using pooled datasets (awareness ratings of 1 + 2). **C**. Temporal generalization accuracy and significant clusters (white contours; $P$<0.05, corrected) for the three decoding problems using the pooled datasets (duration of 30 + 150 ms). The last panel shows significant temporal generalization clusters for all three decoding problems. Angry vs neutral decoding leads to earlier stable representations.




Supplementary Figure 4: Correlation time-courses obtained in the RSA analysis. All significant searchlights are plotted separately against a noise ceiling averaged across significant searchlights.