The Neurodynamic Basis of Real World Face Perception

Arish Alreja

April 2024

Joint Ph.D. Program in Neural Computation & Machine Learning
Neuroscience Institute & Machine Learning Department
Carnegie Mellon University, Pittsburgh, PA

Dissertation Committee

Avniel S. Ghuman	University of Pittsburgh (Chair)
Robert E. Kass	Carnegie Mellon University (Chair)
Leila Wehbe	Carnegie Mellon University
Charles E. Schroeder	Columbia University

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy.

This work was partially supported by the National Science Foundation (1734907), the National Institutes of Health (R01MH132225, R01MH107797), the Richard King Mellon Foundation, and the Beckwith Foundation.

Keywords: Computer Vision, Mobile Eye Tracking, Face AI models, State Space models, Sparse Canonical Correlational Analysis, Neural Decoding, Neural Reconstruction, Mixture Models, Representational Similarity Analysis, Axis based Code, Norm based Code, Tuning Curves, Neural Population Tuning, Weber's law, intracranial EEG, Real World Neuroscience, Face Perception, Fusiform Face Area, Facial Expressions, Facial Motion, Neurodynamics.

Not all those who wander are lost.

ABSTRACT

Understanding how our brains process information while we interact with the real world is a central objective for neuroscience. However, most important neuroscientific discoveries have come from studying brain activity that was recorded while people performed tightly controlled laboratory experiments, which leaves us with open questions about how those findings relate to the brain in the real world. Recent advances in technology have made it possible to record the natural environment, behavior, and brain activity simultaneously and at scale, making it possible to study the brain in the real world. However, realizing the potential of these advances for scientific discovery requires confronting two intertwined questions: Can we even model the uncontrolled variability that arises in the real world? And if we can, then can we learn anything about the brain by doing so? This thesis attempts to answer these questions in the context of face perception during natural social interactions. It introduces methods that address the engineering and analytical challenges necessary to harness large datasets and transform the uncontrolled variability in real world behavior from a challenge into an asset that enables scientific discovery.

ACKNOWLEDGEMENTS

I would like to thank all the surgical epilepsy patients whose participation in research made this thesis possible. Drug resistant epilepsy is a challenging condition to endure and treatment i.e., seizure localization, requires patients undergo neurosurgery to implant electrodes that monitor their brain activity, and spend 1-2 weeks under observation in a hospital's Epilepsy Monitoring Unit (EMU). This environment also provides a rare opportunity to study the human brain, and I am deeply grateful to our patients for their willingness to participate in research and for their generous enthusiasm for science. The opportunity to work with them has been a defining part of my graduate training.

I am also thankful to Taylor Abel, Mark Richardson, James Castellano, and the EMU teams at UPMC Presbyterian and Children's Hospital of Pittsburgh for creating and nurturing a space for research in their EMUs. Working with patients under your care has been a privilege.

I am grateful to the members of my dissertation committee for their time, wisdom, and guidance. Rob Kass, for making sure I square up to important questions, practical and conceptual, about the work at hand as well as the future. Charlie Schroeder, for leaving me with nuggets of wisdom in every one of our conversations, no matter how brief. Leila Wehbe, for her enthusiasm for studying the brain in the real world, within and beyond the scope of this thesis.

I am especially grateful to my advisor Avniel Ghuman, for a great many lessons, which if

listed, could become a dissertation by themselves. Of them, three have been particularly important in shaping my growth from an engineer into a scientist. The first is the art of simplification, to render complex questions tractable. The second is the openness to welcome new, unfamiliar, and even uncomfortable ideas, and to find the courage to question prevailing wisdom. The third and arguably most important is to find joy and adventure in the process of doing science.

I am thankful to my lab mates Yuanning Li, Matthew Boring, Shahir Molawei, Brett Bankson, Maxwell Wang, David Geng, Jhair Colan, Mary Kate Richey, Irisin Yu, and Witold Lipski for making our lab a wonderful place to work. I am also thankful to Kyle Rupp, Jasmine Hect, Emily Harford, and Sreekrishna Ramakrishnapillai in Taylor Abel's lab, for warmly welcoming me into their midst and for the opportunity to work with their patients at the Children's Hospital of Pittsburgh. I am also grateful to Qianli Ma, Nicole Silverling, and Taylor Gautreaux in L-P. Morency's group at Carnegie Mellon for their collaboration. I am also grateful to my former co-advisor Max G'Sell for his guidance through my early years in graduate school, the fruits of which are embedded in Chapter 2. Lastly, I owe a debt of gratitude to our former lab manager and my dear friend Michael Ward, who taught me both the skills and the ethos of working with patients in the EMU.

I owe thanks to Melissa Stupka and Diane Stidle. You have made the Neuroscience Institute and the Machine Learning Department wonderful homes that have enriched my time at Carnegie Mellon; and your guidance and support has helped me navigate various aspects of the Joint Program in Neural Computation and Machine Learning.

My experiences outside the academy have been a significant influence on my journey into research. Ken Thompson and Cynthia Correa constantly encouraged me to aim higher while I interned at startups during my undergraduate years at Georgia Tech, at a time when circum-

stances inhibited my ability to dream. Microsoft was a particularly important training ground in this regard, where my managers Neil Deason and William Looney taught me how to identify, approach, and solve broad multi-dimensional problems, and an ecosystem of mentors including Tony Bell, Anand Lakshminarayanan, Abhi Abhishek, Srikanth Shoroff, and Vishal Thakkar took me under their wing to shaped my worldview about leadership, innovation, technology, and business in diverse ways. I am grateful to all of them for planting seeds and nurturing ideas, each in their own way, all of which have led me to where I am today.

My lack of neuroscientific background made leaving industry to pursue graduate school in neuroscience a highly uncertain path. I would not have been able to navigate it without the mentoring and support I received at my alma mater, Georgia Tech. I am deeply grateful to Professor Robert Butera for planting the seeds of neuroscience research in my mind when I was an undergraduate in 2005, and for helping me find my way back to Georgia Tech when they finally sprouted a decade later. Rob helped get me oriented to computational neuroscience by directing me to classes taught by Dieter Jaeger, Astrid Prinz, Christopher Rozell, and Garrett Stanley, and he suggested Chris as a potential mentor. I'm also grateful to Chris for taking me under his wing, introducing me to neurotheory research, and to Ilya Nemenman whose mentorship grew my confidence. I am also grateful to Garrett Stanley for the opportunity to participate in optogenetics experiments in his lab, and to Michael Bolus and Adam Willats for watchful mentorship that ensured my effort amounted to some benefit for their projects. These opportunities and experiences were instrumental in preparing me for graduate training in computational neuroscience.

Finally, I am thankful to my family including my father and mother in law Rajesh and Sandhya Bhat for their unwavering encouragement and support over the last several years. Above all, I'm grateful to my wife Meera; I would not have embarked upon this adventure without your encouragement and I would not have seen it through without your immovable faith.

Contents

CHAPTER 1

INTRODUCTION

Understanding how the brain processes information as we interact with the real world is a central objective for neuroscience. Observing the brain in the real world is an intuitive though challenging way to approach this goal. Some of the earliest neuroscientific studies made such observations by mapping brain lesions to deficits in function (1, 2, 3, 4). Despite their inherent limitations, lesion based studies helped describe some of the earliest functional maps of the human brain and remain an important tool for neuroscientific inquiry (5, 6).

Psychophysics took a different approach to study human cognition, seeking a more granular understanding with highly controlled laboratory experiments designed to test specific hypotheses, one at a time. Measuring behavior in these experiments enabled researchers to make inferences about neural representations i.e., what we can and cannot do reveals the nature of information representation *somewhere* in our brain (7, 8, 9, 10). Subsequent advances in technology enabled recording brain activity during experiments (11), connecting behavior to the underlying neural substrate and giving rise to the modern neuroscience experiment (11, 12, 13, 14).

Much of our knowledge about human cognition comes from laboratory experiments. However, the ecological validity of that knowledge is an open question because of the stark differences between laboratory and the real world. Growing evidence shows that ecological approaches evoke different responses from the brain compared to controlled experiments (15, 16, 17, 18, 19)

underscoring the importance of studying the brain in the real world to gain new insights that might not be obtained from controlled experiments. Even among skeptics, studying the brain in the real world is important, at least to validate experimental findings and determine how they generalize. Despite a growing number of calls to action (18, 20, 21, 22, 23, 24, 25) emphasizing the importance of real world neuroscience, the number of actual research efforts studying the human brain in the real world has remained fairly small (26, 27, 28, 29, 30). A big reason for this is the three major challenges that must be addressed to realize the promise of real world neuroscience.

The first major challenge is to record the natural environment, behavior, and brain activity simultaneously and with high fidelity. Obtaining rich large scale multi modal recordings requires multiple devices that must work together in synchrony with each other. This can be a significant practical challenge due to heterogeneity of the data streams being recorded and varying tolerances of hardware devices. The exact specifications of the engineering problems that arise also varies based on the aspect of cognition being investigated. For instance, investigating the neural correlates of visual cognition requires recording where participants look and what they see using mobile eye-tracking, but participant mobility is not essential. In contrast, participant mobility and movement tracking is critical for studying navigation in the real world (31), but tracking eye-movements may be unnecessary. The complexity of this challenge is also visible in literature, where separate methods papers (30, 31) tend to precede actual scientific results (26).

The second major challenge is to develop analysis approaches that can model the uncontrolled variability of natural behavior in the real world effectively. This challenge arises because two features of controlled experiments around which analysis is organized are absent during natural behavior. The first is the absence of control over the timing, presentation, and the nature of stimuli - which is replaced by the uncontrolled variability of natural environments. The second

is the absence of control over what participants do and when, in the form of task instructions - which is replaced by unscripted natural behavior. A practical advantage of studying natural behavior is the smaller burden on participants compared to controlled experiments, which makes it easier to collect a larger volume of data. However, realizing the benefit of large scale recordings requires analytical frameworks that can model the uncontrolled variability of natural behavior.

The third major challenge is for analytical approaches to provide an interpretable understanding of neural representations underlying cognition. Addressing this challenge is important both for real world neuroscience and for experimental studies. Early neuroscientific experiments found interpretability by profiling neural tuning in simple geometric spaces where stimuli were parameterized and hypotheses about neural tuning could be tested (32). Advances in technology enriched stimuli at the cost of interpretable parameterization, and analysis frameworks like Representational Similarity Analysis (RSA)(33) advanced distance matrices as an approach to parameterizing stimuli and brain activity to fill the gap. Deep neural networks have emerged as an attractive and scalable alternative to parameterize and relate stimuli to the brain (34) because they can represent different computational hypotheses based on their architecture, optimization objectives, and diet of training data (35, 36). These approaches enable analysis of large neuroscientific datasets, but the insights they provide are often limited to a "score" which is hard to compare/compete between different models on a scoreboard (37). A big reason for this is that the geometry of neural network's parameter spaces is inaccessible or hard to interpret. One approach to fill this gap is with methods that can identify or learn shared tuning spaces in which aspects of brain activity and stimuli are strongly related, and whose underlying assumptions ensure the geometry of tuning spaces is interpretable. This idea is represented in both traditional Statistics and contemporary Machine Learning/Artificial Intelligence, but instantiating it in neuroscience requires careful consideration of variables relevant for each cognitive domain, particularly when modeling the uncontrolled variability of natural behavior in the real world.

This thesis engages with these challenges by studying the brain during natural behavior in the real world and in controlled experiments, with a focus on face perception. Chapter 2 investigates the relationship between face viewpoint and identity using intracranial brain recordings from humans doing controlled experiments. It introduces an interpretable mixture model approach to learning representational spaces from neural data, which is then used to compare and compete data driven and literature based hypotheses about face viewpoint representations. Chapter 3 shifts the focus to studying the brain during unscripted social interactions in the real world. Specifically, it addresses engineering and technical challenges to establish a paradigm for studying the neural basis of social behavior in the real world in humans using intracranial brain recordings collected in an inpatient environment where participants interact with friends, family, clinicians, and researchers. Chapter 4 uses this paradigm to investigate face perception. Specifically, it establishes an analytical framework (and general principles that underpin it) for modeling the uncontrolled variability of the real word, demonstrates the robustness of this approach for face processing, and demonstrates that interpretable tuning spaces can be learned from data. This thesis concludes by using this framework to test hypotheses about the neural representation for facial expressions observed during unscripted social interactions in the real world.

LIST OF PUBLICATIONS

- Chapter 2 - "Temporal Dynamics of Face Viewpoint and Identity Representations in Human Ventral Temporal Cortex", In Prep.
- Chapter 3 - "A New Paradigm for Investigating Real-World Social Behavior and its Neural Underpinnings", Behavior Research Methods, 2022.
- Chapter 4 - "Reconstructing the neural code for real world face perception", In Prep.
- Chapter 4 - "Reconstructing the neural code for face perception in the real world", U.S. Provisional Patent 63/565,173 , filed March 14, 2024.

CHAPTER 2

THE NEURAL CODE FOR FACE VIEWPOINT AND IDENTITY

2.1 INTRODUCTION

An influential cognitive model of face processing (38) suggests face viewpoint centric representations arise from a structural encoding scheme and precede the rise of identity representations. Experimental studies in human (39) and non-human primates (40) suggest that these representations arise alongside each other in face areas in ventral temporal cortex (VTC), and how they relate changes as visual information advances from posterior to anterior regions. Specifically, identity representations are thought to be dependent upon face viewpoint in posterior VTC, evolving to partial (mirror) invariance, before they completely disentangle into a viewpoint invariant identity representation in anterior VTC. Non-human primate studies (40) posit that feedforward propagation is sufficient to account for the temporal dynamics of these representations, but the temporal dynamics of these representations in humans remain unclear because of the limited temporal resolution of imaging studies (39, 41). Human studies with high resolution intracranial recordings have illuminated temporal dynamics underlying different facets of face perception (42, 43), suggesting they may do the same for face viewpoint and identity representations in humans.

This chapter investigates representational dynamics of face viewpoint and their relationship with identity using intracranial EEG recordings from face processing areas in the ventral temporal cortex. Intracranial EEG recordings were collected from 75 face selective electrodes in 18 subjects, located in the face processing network in human ventral temporal cortex, while they viewed faces at different viewpoints in a gender discrimination task. Multivariate Temporal Pattern Analysis (MTPA) was performed on data from these electrodes to relate neural activity with respect to face viewpoint and identity. A novel mixture model approach for representational analysis is developed, revealing new characteristics in the neural representation for face viewpoint and capturing qualitative observations from existing literature. Representational Similarity Analysis (RSA) against a biologically plausible deep learning model of face processing concurs with representational analysis using the new method. The results show previously unreported characteristics in the face viewpoint representations. Identity decoding in a subset of 7 subjects reveals that the representational hierarchy associated with the identity code (viewpoint dependence $\rightarrow$ mirror invariance $\rightarrow$ viewpoint invariant). The relationship between identity and face viewpoint representations is examined and reveals the mirror symmetric face viewpoint representation (with weak mirror confusion) as a correlate of the identity code. Notably, we find the idea of purely feedforward propagation of visual information from posterior to anterior face areas does not account for the observed dynamics of face viewpoint and identity representations in human VTC, where different representation may rise and dissipate over time in the same cortical location.

2.2 Materials and Methods

2.2.1 Data

Intracranial EEG recordings were collected from 18 human subjects (11 males, 7 females). Each subject participated in 2 experiments as part of this study. Experiment 1 was a functional

localizer experiment (a one back task) with images of faces (50% males), bodies (50% males), words, hammers, houses, and phase scrambled faces were used as visual stimuli. Experiment 2 was a face perception experiment (gender discrimination). Face stimuli with 5 distinct viewpoints (either Left Away, Right Tilt, Straight or Right Away, Left Tilt, Straight) with 50% male and 50% female faces, were taken from the Karolinska Directed Emotional Faces (KDEF) stimulus database (44). Three variants of the stimulus set existed for Experiment 2. Variant 1 included 40 individuals (50% male) each with 5 facial expressions and 3 distinct face viewpoints (either Left Away, Right Tilt, Straight or Right Away, Left Tilt, Straight). The 600 unique images were each shown once for a total of 600 trials. Variant 2 included 8 individuals (50% male) each with all 5 facial expressions and 5 face viewpoints. The 200 unique images were shown 3 times each for a total of 600 trials. Variant 3 included 4 individuals (50% male) with all 5 face viewpoints and a neutral expression.

2.2.2 Multivariate Temporal Pattern Analysis

Multivariate methods were used instead of traditional univariate statistics because of their superior sensitivity (42, 45, 46, 47). In this study, Multivariate Temporal Pattern Analysis (MTPA) decoders were used to estimate the coding of different stimulus conditions in recorded neural activity from individual electrodes. MTPA estimates decoding accuracy at a given timepoint with classifiers that use recorded neural activity as input features, within a time window (100 ms wide in this study) which follows the timepoint. The time course of decoding accuracy for a trial is estimated by sliding the time window over the duration of neural activity for the trial. We also utilized more granular performance measures in addition to decoding accuracy, to gain insight into the neural representation where relevant. The first of these were confusion matrices, which were estimated for each MTPA time step. The second granular metric was d', derived from the confusion matrices and calculated as Z(true positive rate) – Z(false positive rate), where Z is the inverse of the Gaussian cumulative distribution function. d' was used because it is an unbiased

Figure 2.1: Pairwise Identity Classification problem examples for a single identity pair. The analysis iterates over all identity pairs as well as iterates over face viewpoints so that all face viewpoints serve as a test set. **(a)** Viewpoint Dependent Pairwise Identity Classification at the Away Right Viewpoint. **(b)** Mirror Invariant Identity Classification Tested on the Away Left Viewpoint. **(c)** Viewpoint Invariant Identity Classification trained on Away Right Viewpoint and tested on all viewpoints except Away Left (sidestepping the mirror invariance code).

measure of effect size and one that takes into account both the true positive and false positive rates. It also has the advantage that it is an effect size measure that has similar interpretation as Cohen's d (48, 49) while also being applicable to multivariate classification. Previous studies have demonstrated that both the low-frequency and the high frequency neural activity, i.e., Evoked Response Potentials (ERP) and Evoked Response Broadband (ERPP), contribute to the coding of facial information (42, 47, 50), therefore, both ERP and ERBB signals in the time window are combined as input features for the MTPA classifier. For each electrode, permutation tests with FDR corrections were used to assess statistical significance and control for multiple comparisons inherent in MTPA. Mixed effects analysis to account for subject, electrode level variability in estimating the standard error for population averaged classification accuracy over time was implemented using a hierarchical bootstrap procedure (51) that is a non-parametric approach capable of capturing linear and non-linear effects and providing a more conservative view of variability compared to linear (or non-linear) mixed effects analysis models.

MTPA is utilized in two contexts in this chapter. The first is 5 way decoding of face viewpoint. The second is pairwise identity decoding for the different identity classification problems enumerated in Fig. 2.1.

CONFUSION MATRIX MIXTURE MODEL

A mixture model approach is developed for data driven representational analysis. The derivation, algorithm, bootstrap procedures to estimate variability in model parameters, construction of a perfectly 'saturated' model to assess quality of fits in terms of variance explained and neuroscientifically meaningful interpretation of model parameters are detailed as follows.

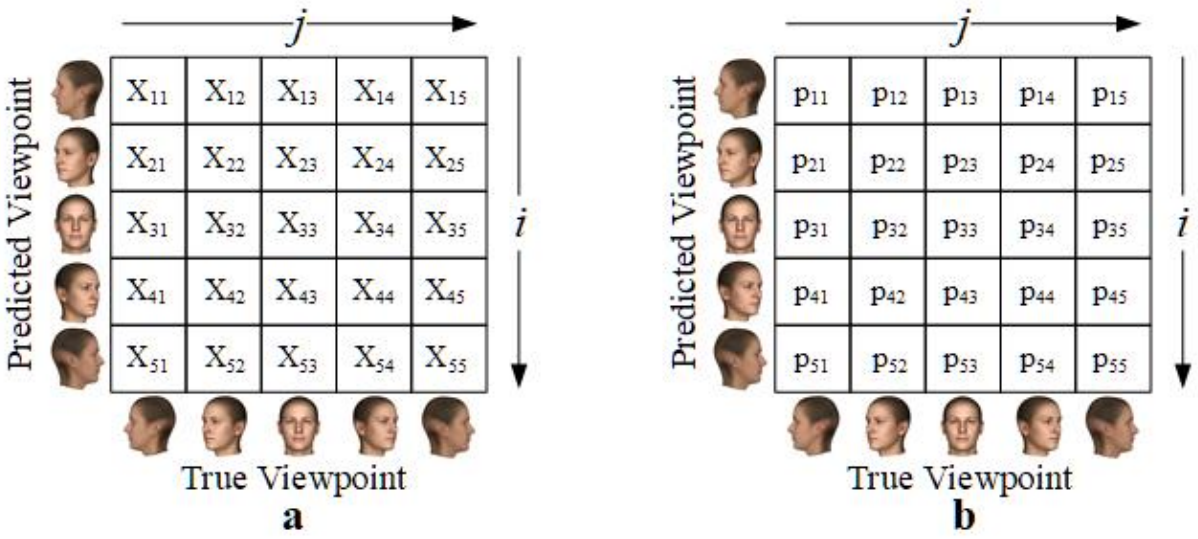

Figure 2.2: **(a) Data** : Each sample ($x \in \mathbb{Z}_{\geq}^{D \times D}$) is a matrix where each entry is a whole number. These samples represent un-normalized confusion matrices generated from a D category classification problem, with $D = 5$ categories representing 5 different face viewpoints in this case. $\sum_{i=1}^{D} x_{ij}$ is the number of trials for the j^{th} true category of face viewpoint and $\sum_{i=1}^{D} \sum_{j=1}^{D} x_{ij}$ is the total number of trials in the confusion matrix sample x. **(b) Parameters** : A normalized confusion matrix where each entry p_{ij} represents the probability of a trial from the j^{th} category being classified as the ith category. We observe that $p_{ij} \in [0, 1] \; \forall \; i, j$ and use $p_j = [p_{1j}, p_{2j}, ..., p_{Dj}]$ as shorthand for the multinomial random variable p_j that represents each column.

Consider a data set $\boldsymbol{X} : [\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^N]$ of N of un-normalized confusion matrices (Fig. 2.2.a) generated as part of the 5–way face viewpoint classification problem. A single sample from this dataset, $\boldsymbol{x^n} \in \mathbb{Z}_{\geq}^{D \times D}, \; \forall \; n \in [1, N]$ is visualized in Fig.2.2.a. Such data set can be fit to templates which are parameterized as normalized confusion matrices, with each true category modeled as a D outcome multinomial random variable and considered a column of the

template as shown in Fig. 2.2.b. This implies that $\boldsymbol{p_j} = \sum_{i=1}^{D} p_{ij} = 1 \; \forall \; j \in [1, D]$, where $\boldsymbol{p_{ij}} \in [0, 1] \; \forall \; i, j \in [1, D]$. The multinomial random variables that form each column are considered independent as a result of the experimental conditions under consideration (i.e. modeling confusion matrices). Under this parameterization, the likelihood $\boldsymbol{f(x^n, p_j)}$ of a multinomial random variable representing a single column, with a single sample (x^n) can be written as

$$f(x^n, p_j) = \frac{(\sum_{i=1}^{D} x_{ij}^n)!}{\prod_{i=1}^{D}(x_{ij}^n!)} \prod_{i=1}^{D} p_{ij}^{x_{ij}^n} \tag{2.1}$$

Next, recalling that multinomial random variables for each column ('True Category') are independent we can write down the likelihood $\boldsymbol{f(x^n, p)}$ for a sample (x^n) against a parameterized confusion matrix as

$$f(x^n, \boldsymbol{p}) = \prod_{j=1}^{D} f(x^n, p_j) = \prod_{j=1}^{D} \left[\frac{(\sum_{i=1}^{D} x_{ij}^n)!}{\prod_{i=1}^{D}(x_{ij}^n!)} \prod_{i=1}^{D} p_{ij}^{x_{ij}^n} \right] \tag{2.2}$$

and extend to estimate the likelihood for the entire dataset $\boldsymbol{f(X, p)}$

$$f(\boldsymbol{X}, \boldsymbol{p}) = \prod_{n=1}^{N} f(x^n, \boldsymbol{p}) = \prod_{n=1}^{N} \prod_{j=1}^{D} f(x^n, p_j) = \prod_{n=1}^{N} \prod_{j=1}^{D} \left[\frac{(\sum_{i=1}^{D} x_{ij}^n)!}{\prod_{i=1}^{D}(x_{ij}^n!)} \prod_{i=1}^{D} p_{ij}^{x_{ij}^n} \right] \tag{2.3}$$

Maximizing the likelihood in Eq. 2.3 (and/or its log) would give us the optimal parameters to fit the data against a single template. With a joint distribution and a likelihood function that can be maximized to obtain parameters that best fit the data, it is natural to consider the scenario, where the population of confusion matrices in a data set corresponds to multiple distinct latent factors/confusion matrix templates, present in the data in different proportions. Using Eq 2.3 as a building block, we define a K component mixture model with a prior $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_K]$ and

$P(z = k) = \pi_k$, such that $0 \leq \pi_k \leq 1$; $\sum_{k=1}^{K} \pi_k = 1$. We extend the notation such that $\boldsymbol{p}$ for the k^{th} component is denoted as $\boldsymbol{p}^k$ and p_{ij} for the k^{th} component will be denoted as p_{ijk}. Finally, using $\boldsymbol{\Theta} = \{[\boldsymbol{p^k}, \pi_k] \; \forall \; k \in [1, K]\}$ as shorthand for the parameter set, we can write down the likelihood (conditioned on Θ) as

$$P(\boldsymbol{X}|\Theta) = \sum_{k=1}^{K} f(\boldsymbol{X}, \boldsymbol{p^k})\pi_k = \sum_{k=1}^{K} \pi_k \prod_{n=1}^{N} \prod_{j=1}^{D} f(x^n, p_j^k) = \sum_{k=1}^{K} \pi_k \prod_{n=1}^{N} \prod_{j=1}^{D} \left[\frac{(\sum_{i=1}^{D} x_{ij}^n)!}{\prod_{i=1}^{D}(x_{ij}^n!)} \prod_{i=1}^{D} p_{ijk}^{x_{ij}^n} \right] \tag{2.4}$$

ALGORITHM: An Expectation–Maximization (EM) procedure (Algorithm 1) is used to estimate model parameters that maximize a mixture model's overall likelihood (Eq. 2.4) for a given data set. Since the EM procedure is not guaranteed to reach a global minima, we estimate parameters for 25 different initializations when learning a model and pick the model which converges to the highest likelihood among them.

Algorithm 1 Expectation Maximization algorithm to estimate mixture model parameters

Initialize θ randomly, rel_tol = 1, last_likelihood=0
while rel_tol$> 1e-9$ **do**
    # Expectation Step (hold parameters fixed)
    for each sample 'n' $\in [1, N]$ **do**
        for each component 'k' $\in [1, K]$ **do**
            $w_{nk} \leftarrow \frac{f(x^n, \boldsymbol{p}^k)\pi_k}{\sum_{l=1}^{K} \pi_l f(x^n, p^l))}$
    # Maximization Step (hold posteriors w_{nk} fixed) and estimate model parameters
    for each component 'k' $\in [1, K]$ **do**
        $\pi_k \leftarrow \frac{\sum_{n=1}^{N} w_{nk}}{\sum_{l=1}^{K} \sum_{n=1}^{N} w_{nl}}$
        for i $\in [1, D]$ **do**
            for j $\in [1, D]$ **do**
                $p_{ijk} \leftarrow \frac{\sum_{n=1}^{N} w_{nk} x_{ij}^n}{\sum_{n=1}^{N} \sum_{i=1}^{D} w_{nk} x_{ij}^n}$
    rel_tol $\leftarrow$ last_likelihood - Eq 2.4
    last_likelihood $\leftarrow$ Eq 2.4

MODEL SELECTION: Selecting the appropriate number of components for a mixture model is an empirical problem. In this study, we used 5 fold cross validation to evaluate log likelihoods

(Eq. 2.4) for models with up to 10 components, in a similar way as MTPA with each fold serving as the test set once. The average test log likelihood across the 5 folds was examined to choose the optimal number of model components based on the 1 standard error rule (52), which posits that the smallest/simplest model with an average test score within 1 standard error of the model with the optimal test score should be chosen. This rule favors smaller models i.e. fewer components.

SATURATED MIXTURE MODEL: In order to assess a mixture model's fit to the data, what portion of variance in the data the models capture. To address this issue, model loglikelihoods are normalized between a 1 component (average confusion matrix) model (serving as the floor) and a 'saturated' model (serving as the ceiling). The 'saturated' model is constructed to grant a parameter for each dimension of each sample confusion matrix. The model is handcrafted with 7575 components corresponding to 7575 confusion matrices in the data, and each components' parameters are normalized versions of the sample confusion matrix i.e., the model is handcrafted to fit the data perfectly.

INTERPRETING CORTICO-TEMPORAL DYNAMICS FROM POSTERIOR PROBABILITIES: The constrained multinomial mixture model does not confer it with any notion of time or awareness of which electrode a data sample corresponds to. As a result, the representations it learns are not anchored in any way to these variables. However, since the knowledge of these variables exists outside the model, the model output (posterior probability predictions) can be rearranged as a #of Electrodes $\times$ # of MTPA time points $\times$ # of components tensor. Visualizing these time series for each component as population averages (weighted or otherwise) over all electrodes reveals how the face viewpoint representation evolves over time across the population of Ventral Temporal Electrodes used in this study.

We also visualize a cortical probability map of the learned face viewpoint representation using weighted Kernel Density Estimation, where posterior probabilities from the mixture model,

for each electrode (e) at each MTPA timepoint (t) serve as weights. A gaussian kernel (Euclidean distance) with a bandwidth of 5 mm was used under the assumption that volume conduction was the appropriate underlying model for signal propagation. The density function is estimated for each model component (k) at each MTPA timepoint (t) as follows, and normalizing it across components allows us to visualize how the face viewpoint representation changes over cortical space and time for all model components.

$$\hat{f_{kt}}(x) = \frac{1}{Eh}\sum_{e=1}^{E} w_{kt}^{e}\Phi\left(\frac{x - x^{e}}{h}\right), \text{ where } h = 5 \text{ mm}, \Phi(x) = \frac{1}{\sqrt{2\pi}}e^{\frac{-x^2}{2}}, \text{ where } x \text{ is an MNI location} \tag{2.5}$$

ESTIMATION OF REPRESENTATIONAL VARIABILITY: We estimate the variability, specifically the standard error of model parameters using confusion matrices from the MTPA bootstrap described earlier (see *Estimation of variability in classification results*). We repeat the same learning procedure described in Algorithm 1 for each of the re-sampled data sets, which corresponds to running a non-parametric bootstrap in the context of mixture models.

Bootstrapping mixture models is accompanied by the 'label switching problem', where due to random initialization of mixture components the learned components may be in a different order for each bootstrap run. Correcting this and remapping components to ensure that they are in the same order is necessary to estimate the variability of learned parameters for each component (and its prior probability) correctly. To address the label switching problem, we align the components learned from each bootstrap run's data against the baseline model using Hellinger Distance (Eq. 2.6), which is an f-divergence measure that quantifies the distance between two

discrete distributions and is defined as follows for mixture components shown in Fig. 2.2.b

$$d_{Hellinger}(k_1, k_2) = \frac{1}{\sqrt{2}} \times \sqrt{\sum_{i=1}^{D}\sum_{j=1}^{D}\left(\sqrt{p_{ijk_1}} - \sqrt{p_{ijk_2}}\right)^2}, \quad \text{where } k_1, k_2 \in [1, K] \tag{2.6}$$

For each bootstrap model, a $K \times K$ distance matrix $d_{Hellinger}$ is calculated between the baseline mixture model (indexed by k_1) and the bootstrap mixture model (indexed by k_2). The mapping procedure shown in Algorithm 2 is used to assign each component of the baseline mixture model to the 'closest' component of the bootstrap mixture model, while ensuring a 1 to 1 mapping between baseline and bootstrap mixture model components.

Algorithm 2 Mapping procedure to map baseline and bootstrap mixture model components

```
Initialize minmaps as None
for k1 ∈ [1, K] do
    min_distances = d_Hellinger(k1, :)
    Initialize match = False, index = 0
    while match is False do
        if min_distances[index] in minmaps then
            index = index + 1
        else
            minmaps[k1] = min_distances[index]
            match=True
```

Once all the bootstrap mixture models are aligned to the baseline mixture model, we estimate the standard deviation for each model parameter across the bootstrap models, which corresponds to the standard error for those parameters in the baseline model.

STRUCTURED COMPONENT TEMPLATES IN MIXTURE MODELS: In addition to a purely data driven approach to learning representational structure, we defined templates with a reduced number of free parameters in a manner which constrained their structure. The structural constraints were motivated by results about the representational structure of face viewpoint from existing studies in both primates and humans. These structured templates included a linear angle code and a mirror symmetric code. There were uniform/strict and relaxed versions of each. A

fixed template with no free parameters (i.e. noise) was also defined. The EM procedure defined in Algorithm 1 applies with minor modifications to steps 10,11,12. For the Noise template (Fig. 2.7.a), Steps 10,11 and 12 from Algorithm 1 are simply skipped and for the remaining templates in Fig. 2.7.b,c,d,e, they are substituted with the steps in Algorithm's 3,5,6,4 respectively.

Algorithm 3 Linear Angle Parameter Estimate

$\theta \leftarrow \frac{\sum_{n=1}^{N}\sum_{i=1}^{D} w_{nk}x_{ii}^{n}}{\sum_{n=1}^{N}\sum_{i=1}^{D}\sum_{j=1}^{D} w_{nk}x_{ij}^{n}}$
$\theta' \leftarrow \frac{1-\theta}{4}$

Algorithm 4 Mirror Symmetric Relaxed Parameter Estimate

$\theta_1 \leftarrow \frac{\sum_{n=1}^{N} w_{nk}(x_{11}^{n}+x_{15}^{n}+x_{51}^{n}+x_{55}^{n})}{\sum_{n=1}^{N}\sum_{j=1}^{D}\sum_{i=1}^{D} w_{nk}x_{ij}^{n}}$
$\theta_2 \leftarrow \frac{\sum_{n=1}^{N} w_{nk}(x_{22}^{n}+x_{24}^{n}+x_{42}^{n}+x_{44}^{n})}{\sum_{n=1}^{N}\sum_{j=1}^{D}\sum_{i=1}^{D} w_{nk}x_{ij}^{n}}$
$\theta_3 \leftarrow \frac{\sum_{n=1}^{N} w_{nk}x_{33}^{n}}{\sum_{n=1}^{N}\sum_{j=1}^{D}\sum_{i=1}^{D} w_{nk}x_{ij}^{n}}$
$\theta'_1 \leftarrow \frac{1-2\theta_1}{3}$
$\theta'_2 \leftarrow \frac{1-2\theta_2}{3}$
$\theta'_3 \leftarrow \frac{1-2\theta_3}{4}$

Algorithm 5 Linear Angle Relaxed Parameter Estimate

for j $\in [1, D]$ **do**
$\theta_j \leftarrow \frac{\sum_{n=1}^{N} w_{nk}x_{jj}^{n}}{\sum_{n=1}^{N}\sum_{i=1}^{D} w_{nk}x_{ij}^{n}}$
$\theta'_j \leftarrow \frac{1-\theta_j}{4}$

Algorithm 6 Mirror Symmetric Parameter Estimate

$\theta \leftarrow \frac{\sum_{n=1}^{N}\sum_{i=1}^{D} w_{nk}(x_{ii}^{n}+x_{D-ii+1}^{n})}{\sum_{n=1}^{N}\sum_{i=1}^{D}\sum_{j=1}^{D} w_{nk}x_{ij}^{n}}$
$\theta' \leftarrow \frac{1-2\theta}{3}$;
$\theta'' \leftarrow \frac{1-2\theta}{4}$

2.3 RESULTS

2.3.1 FACE SELECTIVE ELECTRODES

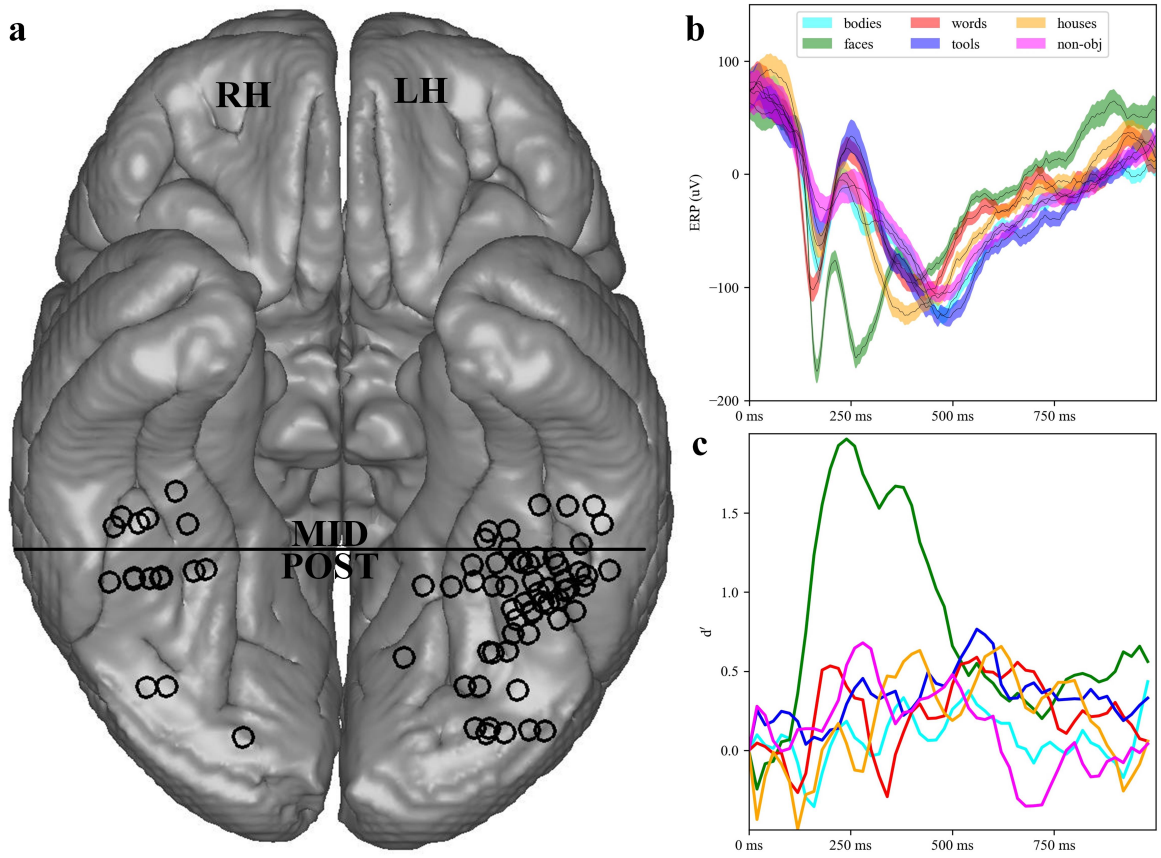

Figure 2.3: **Face Selectivity:** 75 electrodes characterized as face selective (57/18 in Left/Right hemisphere of which 41/12 were posterior to MNI Y coordinate = -45) based on **(a)** Anatomical constraints that restrict them to the mid-fusiform sulcus (MFS), or the fusiform, obitotemporal and parahippocampal gyri, or in the sulci adjacent to the fusiform gyrus. **(b)** Greater signal strength (ERP) for electrodes in response to faces as compared to other object categories from the localizer experiment (representative example). **(c)** Greater statistical specificity/sensitivity (d′) for faces as compared to other object categories for electrodes from Multivariate Temporal Pattern Analysis (MTPA) (see *Methods* for details) of neural activity from the localizer experiment (representative example).

A three fold criterion was used to identify face selective electrodes using data from a localizer experiment. The first part was anatomical constraints that limited analysis to electrodes in ventral

temporal cortex (Fig. 2.3.a). The second was the relative strength of response to face vs. non-face stimuli (Fig. 2.3.b). The third was statistically significant face sensitivity (Fig. 2.3.c). These criteria identified 75 face selective electrodes from 19 subjects for analysis, with 57/18 in Left/Right Hemispheres of which 41/12 lay posterior to MNI Y coordinate = -45. The imbalance in hemispheric sampling reflects an imbalance in implanted electrodes in the subject population.

2.3.2 The Neural Code for Face Viewpoint

Face Viewpoint Classification

Neural decoders with Linear Discriminant Analysis (LDA) classifiers were trained to predict face viewpoint, using combining the ERP and ERBB activity over a 100 ms sliding time window as input and classification accuracy (and d′) versus time trends curves were estimated for each face selective electrode using 5 fold cross validation and Multivariate Temporal Pattern Analysis (MTPA). 70 out of the 75 electrodes exhibited statistically significant classification accuracy (Fig.2.4.a), and ≈86% of all electrodes were significant 170 - 260 ms after stimulus onset (Fig.2.4.b). Population averaged classification accuracy (Fig.2.4.c) revealed a single peak at 220 ms.

Population averaged d′ for each face viewpoint offered a finer grained view of discriminability underlying classification accuracy (Fig. 2.4.d). Peak discriminability and discriminability versus time differed by face viewpoint, but in a mirror symmetric manner (e.g., similar peaks/time courses partial side profiles (Tilt/45°). Front facing profiles exhibited the strongest discriminability curves, followed by side (Away/90°) profiles. Notably, partial side profiles (Tilt/45°) exhibited the weakest peak discriminability. We refer to this effect as ‘anchoring’, where extreme face viewpoints are strongly encoded in neural activity, with relatively weaker encoding for intermediate viewpoints. A hemispheric bias effect (53) was also observed for discriminability of face viewpoint in the population averages for each hemisphere.

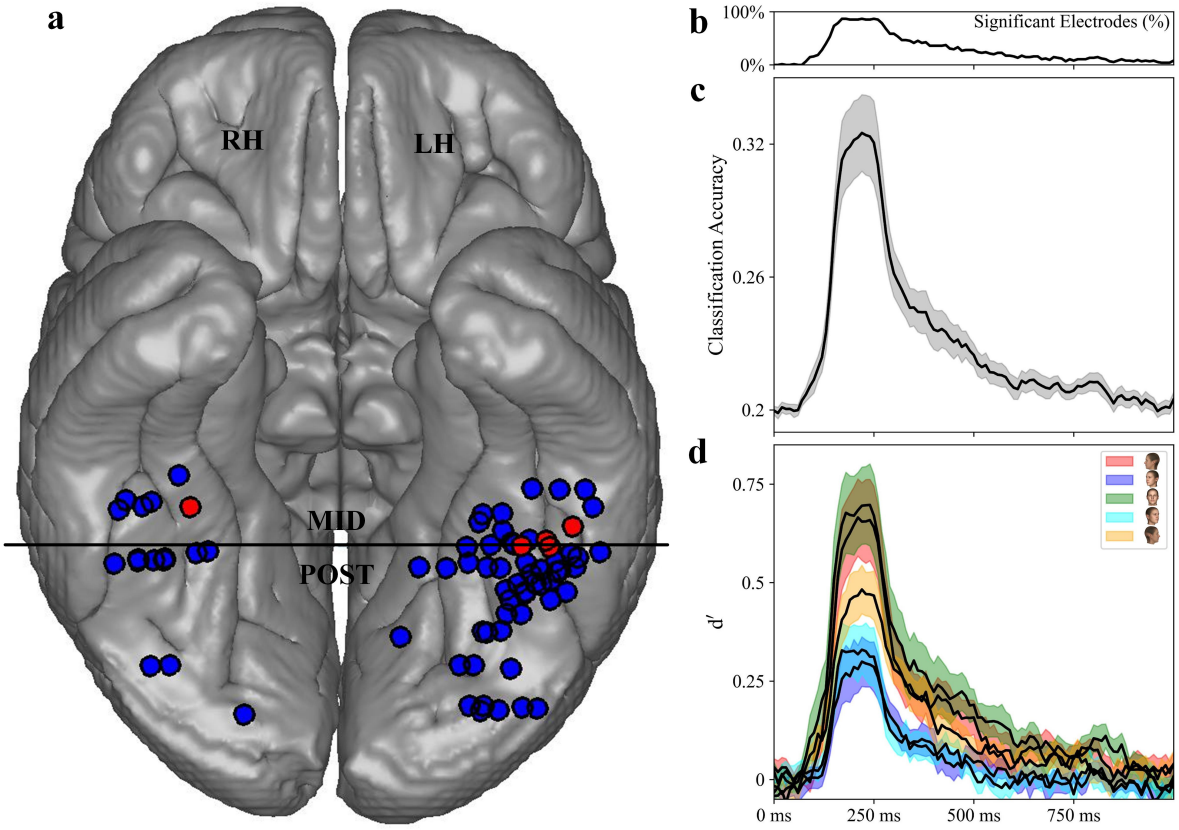

Figure 2.4: **Face Viewpoint Classification:** **(a)** Cortical distribution of statistically significant ($p<0.05$ in blue) electrodes (70 of 75) for face for face viewpoint classification. **(b)** Percentage of electrodes that are statistically significant over time. **(c)** Population averaged classification accuracy (chance = 0.2) $\pm$ standard error (shaded band) vs. time. **(d)** Population averaged $d' \pm$ standard error (shaded band) vs. time for each face viewpoint.

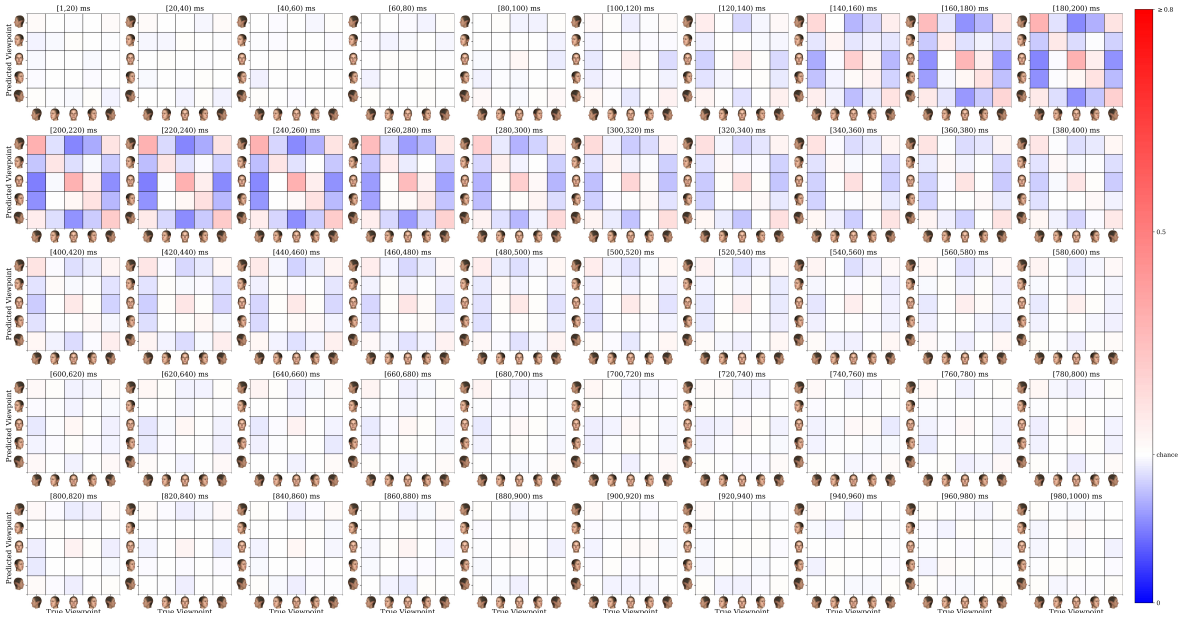

Figure 2.5: Population and time window (20 ms) averaged confusion matrices from face viewpoint classification starting from stimulus onset (top row left most confusion matrix) to the end of trial (bottom row right most confusion matrix) 1000 ms later, with each row depicting a 200 ms interval. Cells closer to chance (0.2) are depicted in paler/whitish hues. Cells with predictions significantly above and chance are depicted with redder and bluer hues respectively.

Confusion matrices estimated during face viewpoint classification reflect the neural representation of face viewpoint as seen by classifiers. A confusion matrix time series is estimated for each electrode during MTPA. A population averaged time series of confusion matrices (Fig. 2.5) suggests a mirror symmetric representation is part of the face viewpoint code, but visualizing averages is a coarse approach, limited in its ability to reveal representational diversity that may underlie the average.

In order to reveal the underlying representational structure for face viewpoint in the neural code, a novel mixture model that uses confusion matrices as mixture components was developed (See *Appendix A* for details) and we performed representational analysis using three different

approaches. The first approach used a purely data driven mixture model. The second approach used structured templates motivated by prevailing hypotheses about face viewpoint representations, constraining the degrees of freedom in the mixture model. The third approach involved representational similarity analysis (33), comparing neural representations of face viewpoint with a recent biologically plausible computational model of face processing implemented as a deep learning network (54).

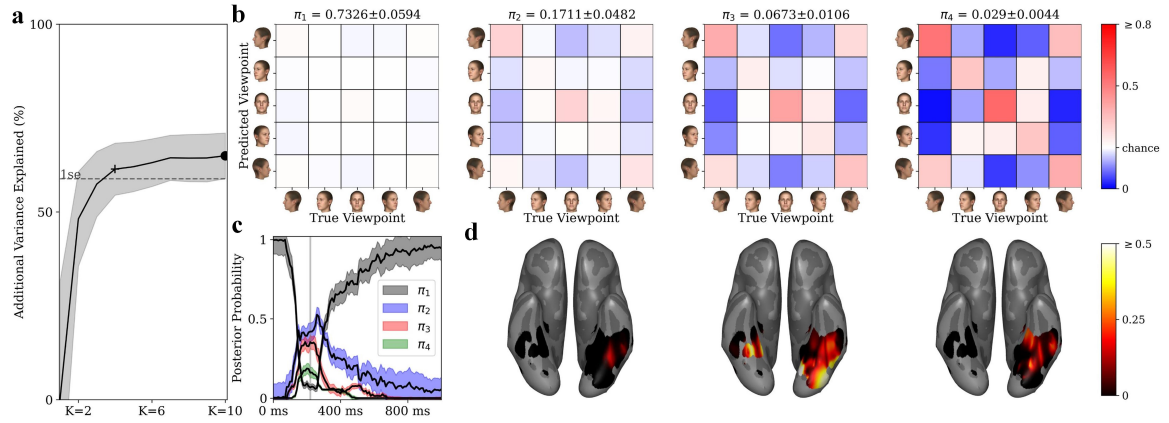

Figure 2.6: **Data driven neural representation of face viewpoint: (a)** mean (black line) $\pm$ standard error (gray band) depicting percentage of variance explained vs. # of components (K $\in [1, 10]$) by mixture models in cross validation based model selection. The average confusion matrix (K=1) serves as the baseline (0%) and a perfect 'saturated' model (See *Appendix A* for details) constructed to assess goodness of fit serves as the ceiling (100%). The optimal model size (K=4; marked by 'x') is chosen by applying the 1 SE rule (52) to the best fit model (K=10; marked by 'o') and it captures 61% of the variance on this scale. **(b)** Mixture model components are displayed in descending order of prior probabilities $\pm$ standard error (on top for each component) for the optimal (K=4) model. The standard error for each component's parameters (i.e., individual cells in each 5 x 5 component) is < 0.02 across all parameters and components. **(c)** The time course ($\pm$ standard error) of population averaged posterior probabilities for each model component with grey vertical line at 220 ms denoting time of peak classification accuracy (per Fig. 2.4.c). **(d)** Cortical heat maps with the probability of each model component (except π_1, the additive inverse of the visible heatmaps) 220 ms after stimulus onset. See *Supplementary Video* visualizing temporal dynamics of cortical heat maps.

DATA DRIVEN FACE VIEWPOINT REPRESENTATIONS Model selection was performed by training mixture models varying in size ($K \in [1, 10]$ components) using 5 fold cross validation, paired with the 1 SE model selection rule (52) that favors parsimonious models. Model fits were

evaluated by normalizing test loglikelihoods on a scale with the $K = 1$ component model as the baseline (0%) and a perfect 'oracle' model constructed to assess goodness of fit serves as the ceiling (100%) (See *Appendix A* for details). The optimal ($K = 4$ components) model captures ≈61% additional variance (relative to the baseline $K = 1$ component model/average confusion matrix) on this scale (Fig. 2.6.a).

The 4 component data driven model revealed a vivid representational structure. Notably, a null component reflecting (π_1 in Fig.2.6.b) the absence of face viewpoint information for all viewpoints was the most dominant component in the mixture ($\pi_1 = 0.7326 \pm 0.0594$). 'Anchoring' i.e., strong coding at front facing and side (Away/90°) profiles and weak coding for partial side (Tilt/45°) profiles, was a common feature across the remaining components (π_2–π_4). Unambiguous characterization of these components (π_2–π_4) as linear angle or mirror symmetric was challenging from visual inspection. The second most prevalent component's ($\pi_2 = 0.1711 \pm 0.0482$) structure exhibited qualitative similarities to a linear angle code with 'anchoring' (π_2 in Fig.2.6.b), although the anti-diagonal elements for the side profiles (Away/90°) prevent ruling out a mirror symmetric code. The structure of the remaining two components (π_3, π_4 in Fig.2.6.b) exhibited qualitative similarities to a mirror symmetric viewpoint (i.e., confusion between left and right profiles) with 'anchoring'. In these mirror symmetric components, mirror confusion at the anti-diagonal appeared weaker than correct classification along the diagonal. π_3 and π_4 were contrasted from one another by the strength of coding, and appeared to be scaled versions of one another, with the relatively more abundant mirror symmetric component ($\pi_3 = 0.0673 \pm 0.0106$) exhibiting relatively weaker coding (diagonal) and mirror symmetry (anti-diagonal) in comparison to the least abundant component ($\pi_4 = 0.029 \pm 0.0044$).

Posterior probabilities for confusion matrices from each electrode, organized by time and averaged across the electrode population visualized temporal dynamics, and the relative strength of

different representational components in the electrode population over time (Fig. 2.6.c). Spatio-temporal heatmaps (Fig.2.6.d and supplementary video) for π_2–π_4 incorporating posterior probabilities of different components for each electrode revealed different face viewpoint representations can rise and fall over time in sampled regions of ventral temporal cortex.

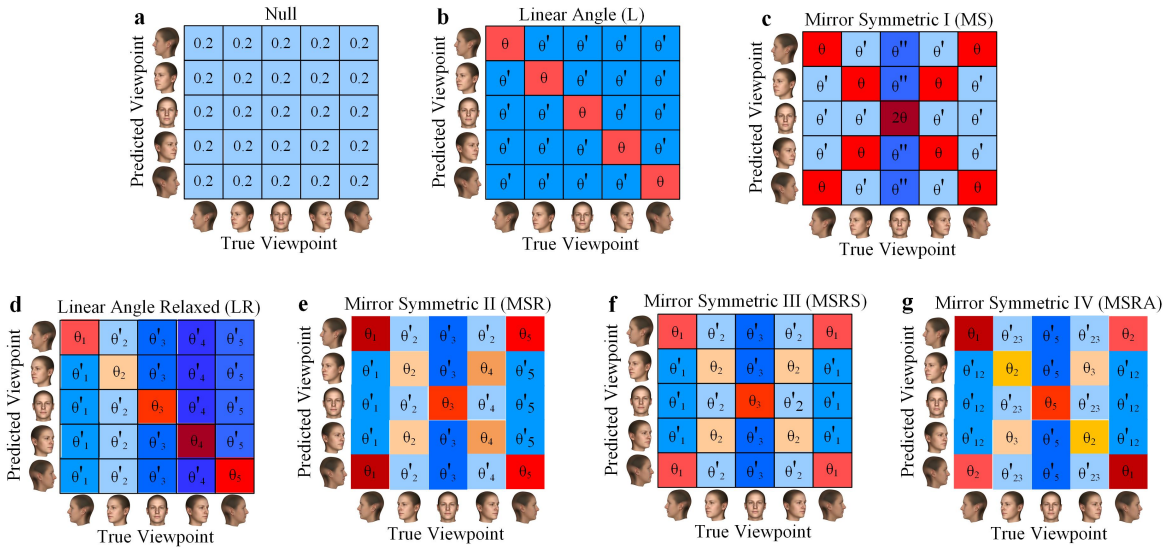

Figure 2.7: **Structurally constrained, minimally parameterized representational templates:** Confusion matrix templates corresponding to specific hypotheses about the face viewpoint code including **(a)** Null **(b)** A linear angle code **(c)** A relaxed linear angle code (referred to as 'LR'), where the specificity of individual face viewpoints varies. **(d)** A mirror symmetric face viewpoint code (referred to as 'MS') **(e)** A relaxed mirror symmetric face viewpoint code (referred to as 'MSR'), where the mirror symmetric specificity of different face viewpoint profiles (away, tilt, straight facing) varies. **(f)** A symmetrically relaxed mirror symmetric viewpoint code (referred to as 'MSRS'), where the mirror symmetric specificity of away and tilt profiles is identical, independent of direction (left/right). **(g)** A symmetrically relaxed mirror symmetric viewpoint code allowing unequal mirror confusion confusion (referred to as 'MSRA'), where the mirror symmetric specificity of away and tilt profiles is identical, independent of direction (left/right) and the diagonal/anti-diagonal parameters can differ.

HYPOTHESIS DRIVEN FACE VIEWPOINT REPRESENTATIONS Structurally constrained and minimally parameterized templates were developed to capture canonical linear angle, mirror symmetric face viewpoint and null representations. Additional relaxed versions of canonical representations were developed based on phenomena (e.g., 'anchoring', hemispheric bias, and

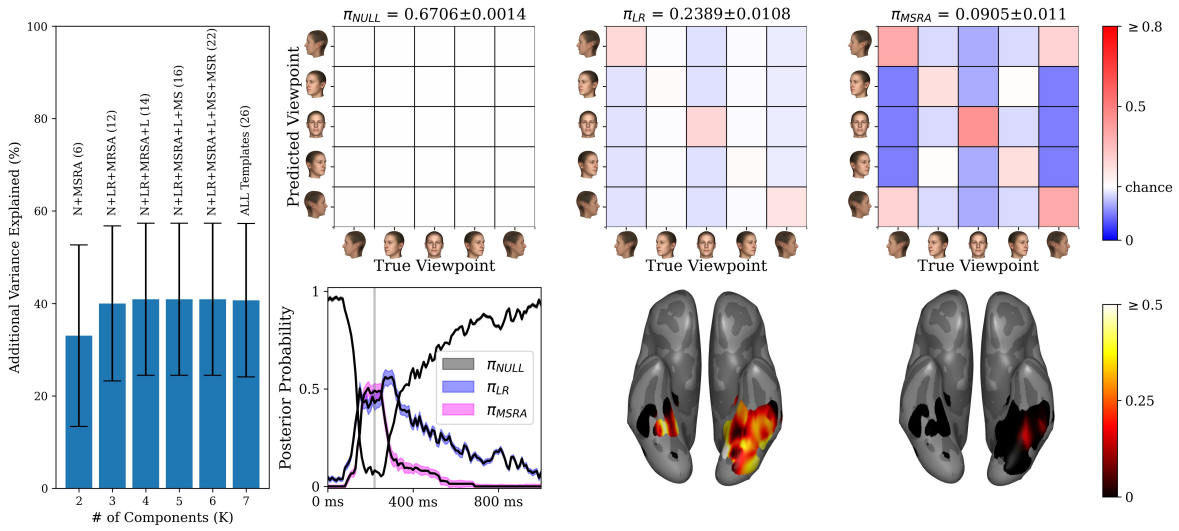

Figure 2.8: **Hypothesis driven face viewpoint representation:** **(a)** mean (bars) ± standard error (black line) depicting percentage of variance explained for the models that contain 2 - 7 components, with unique structured templates shown in Fig. 2.7 serving as components (and the 'Null' component mandated). The average confusion matrix serves as the baseline (0%) and a perfect 'saturated' model (See *Appendix A* for details) constructed to assess goodness of fit serves as the ceiling (100%). The text on top of each bar enumerates the structured templates present in the best model and (i.e., combination of 2,3,4,5,6 and 7 templates that provides the best fit for each model size) and the number of free parameters they add up to. A clear knee in additional variance explained is visible for the best 3 component model which contains the Linear Relaxed ('LR') and Mirror Symmetric IV ('MSRA'), that emerge as the most essential out of the 7 structured templates and account for 40% of additional explained variance in the data **(b)** Mixture model components in descending order of prior probabilities (± standard error; shown on top for each component) for the optimal (denoted by * in **(a)**) 3 component mixture model which features a 'Null', 'Linear Relaxed' and 'Mirror Symmetric IV' component (which allows for weaker mirror confusion, but forces symmetry across left and right viewpoints). The standard error for each component's parameters (i.e., individual cells in each 5 x 5 component) is < 0.02 across all parameters and components. **(c)** The time course (± standard error) of population averaged posterior probabilities for each model component. **(d)** Cortical heat maps showing the probability of each model component over cortical regions sampled as part of this study at a single MTPA timepoint 120 ms after stimulus onset (moment of peak classification accuracy per Fig. 2.4.b) (see *Appendix A* for details). See *Supplementary Video* visualizing the temporal dynamics of these heat maps during stimulus presentation.

weaker mirror confusion) observed in the population average time series and data driven models (Fig. 2.5,2.6). As a starting point for model selection, a 7 component mixture model with one template of each type (a total of 7) was evaluated using 5 fold cross validation. The template which contributed the least was identified and removed (with the exception of the null component) to arrive at the best 6 component model, and subsequently, the best 5, 4, 3 and 2 component models. The best 3 component model was the optimal choice (Fig. 2.8.a), explaining ≈40% additional variance, a noticeable improvement over ≈33% in the best 2 component model, whereas models with additional components offered little improvement (<41%) over the best 3 component model.

The best 3 component model featured relaxed versions of a linear angle and mirror symmetric representation. 'Anchoring' emerged as an essential property of the linear angle representation, and weaker mirror confusion as an essential property of the mirror symmetric representation (Fig. 2.8.b). Within the model, the 'linear relaxed' component was dominant ($\pi_{LR} = 0.2389 \pm 0.0108$) compared to the 'mirror symmetric' template with weaker mirror confusion ($\pi_{MSRA} = 0.0905 \pm 0.011$).

Posterior probabilities for confusion matrices from each electrode, organized by time and averaged across the electrode population visualized temporal dynamics, and the relative strength of different representational components in the electrode population over time (Fig. 2.8.c). Correspondence between the time courses of posterior probabilities was also observed between the best 3 component model and the data driven model. π_{MSRA} mirrored the combined time courses of π_3 and π_4 (mirror symmetric components with weaker mirror confusion in the data driven model), whereas π_{LR} mirrored π_1, supporting an 'anchored' linear angle code as the latter's essential representational structure. Spatio-temporal heatmaps (Fig.2.8.d and supplementary video) for π_{LR} and π_{MSRA} incorporating posterior probabilities of different components for each electrode also

revealed qualitative similarities in the spatio-temporal dynamics of both models.

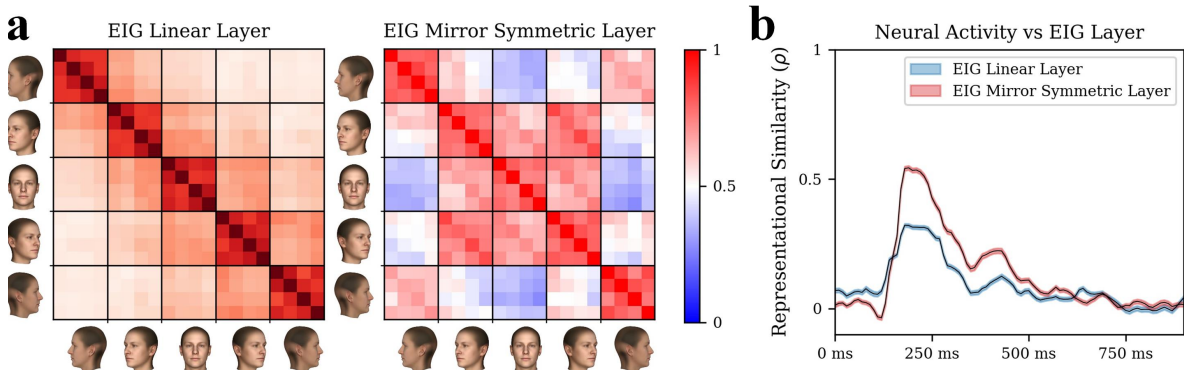

Figure 2.9: **(a)** Representational Similarity Matrices for EIG's putative Linear and Mirror Symmetric Layers, for stimulus images from Variant 3 of the face perception experiment using 'correlation' distance (represented by the color bar) **(b)** Population averaged Correlation $\pm$ standard error (shaded bands) between EIG and Neural representational similarity matrices vs. time.

REPRESENTATIONAL SIMILARITY ANALYSIS BETWEEN A COMPUTATIONAL MODEL FOR FACE VIEWPOINT AND NEURAL ACTIVITY Representational Similarity Analysis (RSA) was carried out to compare neural activity and a computational model called the Efficient Inverse Graphics (EIG) Network (54). The EIG network is a deep learning model whose architecture is constrained to match the face patch system of non-human primates, and trained to reconstruct faces from images by predicting the coefficients of a linear face model (55). The EIG's network activations has shown strong correspondence with face viewpoint and identity representations in single unit recordings from non-human primates (40). Here, representational similarity matrices were computed using correlation distance for time windowed neural activity for each electrode, and the Efficient Inverse Graphics (EIG) Network's putative linear angle and mirror symmetric layers using stimulus images from Variant 3 of the face perception experiment. The representational similarity matrix corresponding to the EIG network's putative linear angle layer lacked the 'anchoring' effect observed in neural data reflected in preceding representational analyses. The representational similarity matrix for the EIG network's putative mirror symmetric layer revealed an 'anchored' mirror symmetric representation with weaker mirror confusion, qualitatively sim-

ilar to the mirror symmetric components (π_3, π_4 in Fig. 2.6.b, π_{MSRA} in Fig. 2.8.b) in the data driven and template based mixture models.

Population averages of correlation versus time trends between neural and EIG representational similarity matrices for face viewpoint estimated using MTPA for each face selective electrode revealed early peaks (Fig. 2.9.b) for both EIG layers. Notably, the correlation between the 'anchored' mirror symmetric layer and neural data achieved a higher peak relative to the putative linear angle layer. This represents a reversal of the relative dominance of the linear angle code (over the mirror symmetric code) in its 'anchored' forms in preceding representational analysis using mixture models.

2.3.3 The Neural Code for Identity

Neural decoders with Linear Discriminant Analysis (LDA) classifiers were trained to predict identity for each within-gender identity pair, combining the ERP and ERBB activity over a 100 ms sliding time window as input and classification accuracy versus time trends curves averaged across identity pairs and face viewpoints were estimated for each face selective electrode using 5 fold cross validation and Multivariate Temporal Pattern Analysis (MTPA). For each electrode, permutation tests with FDR corrections were used to assess statistical significance and control for multiple comparisons inherent in MTPA. Identity classification analysis was undertaken for 47 electrodes from subjects who performed Variant 3 of Experiment #2 only, to ensure that for each identity pair, there were sufficient trials train classifiers. 12 of 47 eligible electrodes ($\approx$ 25%) were statistically significant for viewpoint dependent identity classification (Fig. 2.10.a), 5 of 47 electrodes ($\approx$ 10%) for mirror invariant identity classification (Fig. 2.10.b) and 14 of 47 electrodes ($\approx$ 30%) for viewpoint invariant identity classification (Fig. 2.10.c).

Classification accuracy versus time curves (Fig. 2.10.d) averaged across the statistically sig-

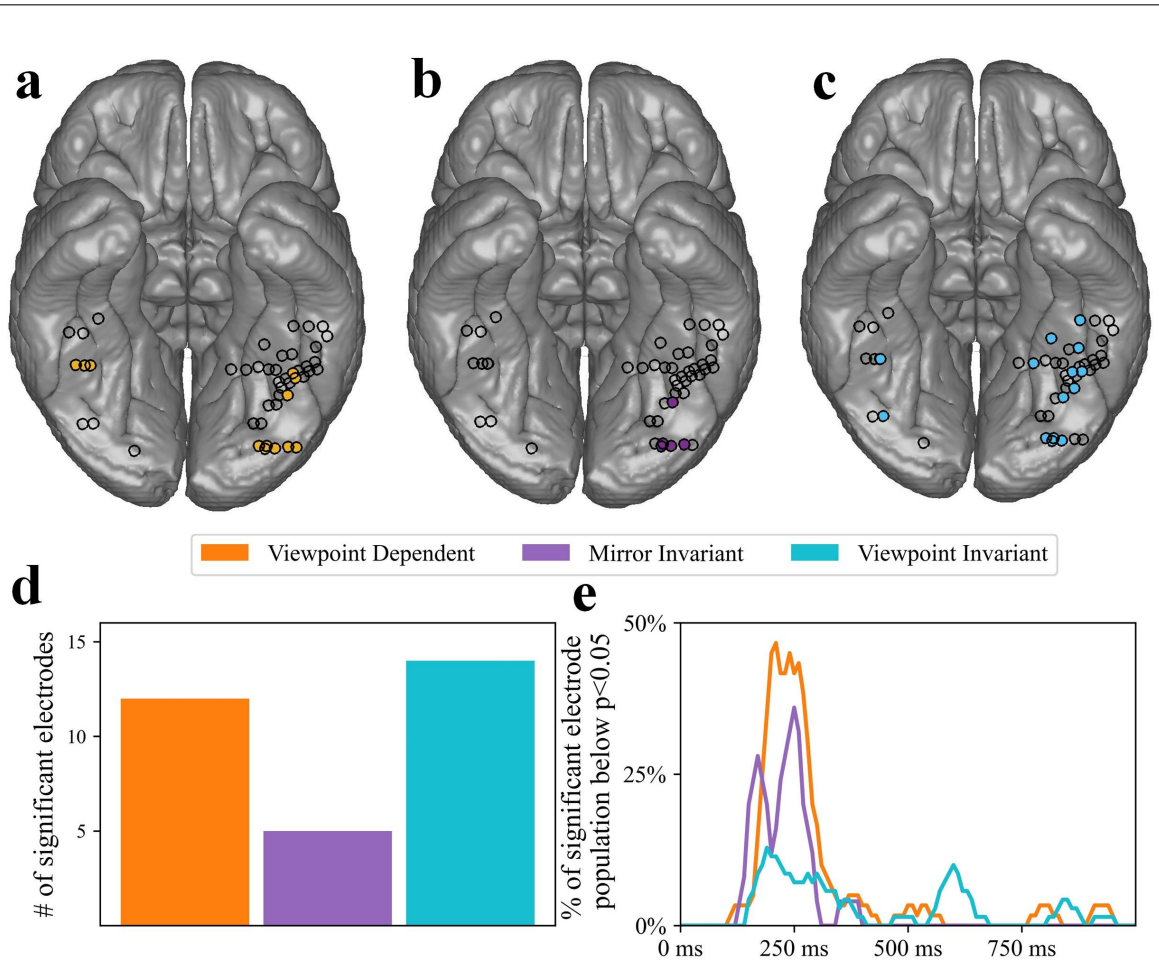

Figure 2.10: Cortical locations of electrodes that show statistically significant ($p<0.05$) identity decoding in a **(a)** Viewpoint Dependent context **(b)** Mirror Invariant context **(c)** Viewpoint Invariant context. **(d)** The number of electrodes which achieve statistically significant identity decoding (at any point in time) Viewpoint Dependent, Mirror Invariant, and Viewpoint Invariant identity decoding. **(e)** Percentage of statistically significant electrode population (quantified in **d)** exceeding statistical significance ($p<0.05$, FDR corrected) over time (smoothed over 50 ms) for Viewpoint Dependent, Mirror Invariant, and Viewpoint Invariant identity decoding.

nificant electrode population suggests (supported by Fig. 2.10.e) that statistically significant electrodes for viewpoint dependent and mirror invariant identity classification may be statistically significant at similar time frames to each other, whereas electrodes that achieve statistical significance for viewpoint invariant identity classification do so at different time frames.

2.3.4 Face Viewpoint and Identity

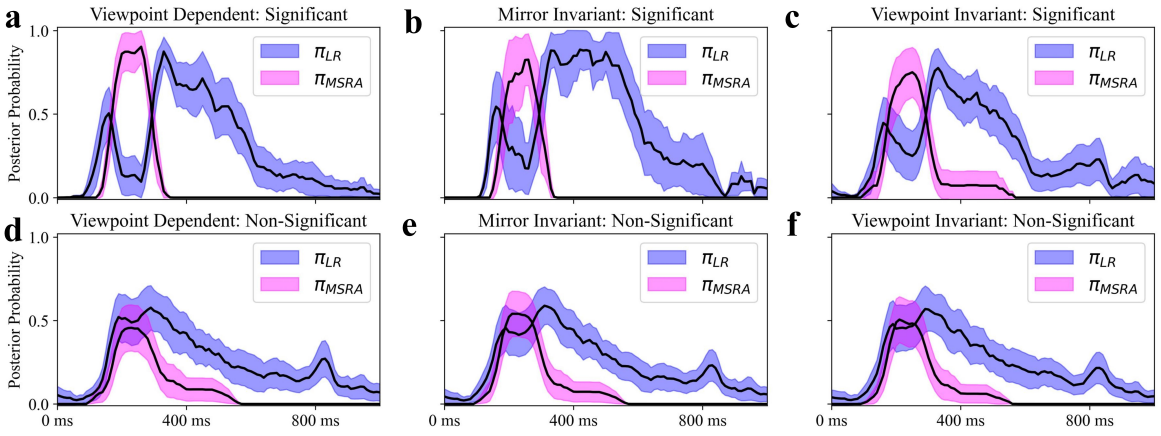

Figure 2.11: Posterior probabilities of Face Viewpoint Representations for non-null components (π_{LR} and π_{MSRA}) of the best 3 component structured mixture model in Fig. 2.8.b, averaged over Significant and Non Significant Identity Classifying Electrode Populations for **(a,d)** Viewpoint Dependent Identity Classification. **(b,e)** Mirror Invariant Identity Classification. **(c,f)** Viewpoint Invariant Classification.

Time courses of face viewpoint representations from the data driven model were averaged for statistically significant and non-significant electrode sub-populations for Viewpoint Dependent, Mirror Invariant and Viewpoint Invariant Identity Classification (Fig. 2.11). For each type of identity classification, the statistically significant electrodes exhibited a distinct temporal signature in the face viewpoint space, relative to the non-significant electrodes. In statistically significant electrodes subpopulations, the 'anchored' linear angle representation dips early, making room for the 'anchored' mirror symmetric components peak. The 'anchored' linear angle representation exhibits its peak later (≈300-500 ms after stimulus onset) as the mirror symmet-

ric representation decays into 'anchored' linear angle representation (see supplementary video). This is a sharp contrast to the non-significant electrode population where those components exhibit temporal signatures similar to each other and aligned with the population average (Fig. 2.8.c). This observation that a substantial weak mirror confusion component in the face viewpoint representation is distinctly correlated with the neural code for face identity.

2.4 Discussion

This study combines a novel mixture model framework for latent representational analysis with high resolution (iEEG) recordings of brain activity from 19 subjects to study face viewpoint and identity representations in the human face processing network. The results reveal previously unreported aspects of the face viewpoint and identity representations, accompanied by a fine grained view of their neurodynamics, and elucidate how face viewpoint and identity representations relate. These insights establish the novel method's efficacy in identifying a latent representational basis from a population of neural similarity matrices under different (purely data driven, hypothesis constrained and purely hypothesis driven) analytical settings, which underscores its promise as a tool for latent representational analysis beyond the neuroscientific questions explored here.

2.4.1 Face Viewpoint and Identity: Representation and Neurodynamics

The latent representational basis for face viewpoint estimated from neural data has several characteristics that validate the mixture model framework's efficacy. One such observation is the emergence of a dominant 'null' representational component which reflects the absence of face viewpoint information in portions of brain activity evoked by faces. Reorganizing model posterior probabilities by unknown external covariates such as time and electrode reveals temporal

dynamics for face viewpoint representations that are consistent with classification time courses. Lastly, the hemispheric bias effect (53) in the structure of estimated representational components reflects the imbalanced hemispheric sampling of the electrode population.

Representational analysis for face viewpoint spanning data driven, hypothesis constrained and hypothesis driven settings all converge to similar conclusions and add novel insights to the broad consensus about face viewpoint representations in literature. Specifically, the model reveals weaker mirror confusion (i.e., there is a statistically significant difference between viewpoint coding and mirror confusion) in partially (mirror) symmetric representations that has not been reported previously. Similarly, previously unreported 'anchoring' effects emerge for the viewpoint dependent representation, showing consistency with partially (mirror) invariant representations for which 'anchoring' has been reported (40, 56). Notably, the population level representational findings (i.e., 'anchoring', weaker mirror confusion) find some support at the single-unit level (32) and correlate with face viewpoint aftereffects reported by psychophysical studies (57, 58, 59, 60), which support the idea of a multichannel coding scheme for face viewpoint.

The neurodynamics of face viewpoint representations across the face processing network are visualized at high spatial and temporal resolution by fusing model predictions with external covariates (time, electrode, and cortical location). They reveal an 'anchored' viewpoint dependent representation as an early baseline response, which gives way to an 'anchored' mirror symmetric representation over most of the sampled areas of the face processing network, that decays back to the early baseline response eventually dissipating to a null representation. These findings present a contrast to existing results from human imaging (39, 61) that suggest similar compartmentalization of face viewpoint as observed in non-human primates (40) and are supported by spatially coarser results from human scalp EEG studies (56). These findings also offer potential resolution existing debates about viewpoint dependence (41, 62) or mirror symmetry (39, 61) being the de

facto representation in the fusiform, which has roots in methodological considerations (62) that do not arise in iEEG analysis using MTPA. By offering a potential resolution to the ”either or” ambiguity about the fusiform and showing a lack of purely spatial compartmentalization for face viewpoint representations, the neurodynamics for face viewpoint representations suggest the that the information processing hierarchy (for face viewpoint) in the face processing network spans cortical space *and* time.

Identity decoding reveals statistically significant decoding for viewpoint dependent, mirror invariant and viewpoint invariant identity representations across the face processing network, with multiple identity representations arising in the same electrode in several instances. These findings present a similar lack of spatial compartmentalization along a posterior-anterior gradient as observed for face viewpoint, contrasting with existing results (39, 40, 61). Temporal dynamics offer a contrasting picture for different identity representations, with electrodes significant for viewpoint dependent and mirror invariant identity classification achieving significance early ($<$300 ms), whereas statistical significance for viewpoint invariant electrodes is distributed over time in an early ($<$300 ms) and late ($>$500 ms) window. These observations suggest that feedforward propagation alone does not explain information processing for identity in the face processing network.

Lastly, independent of the type of identity representation (viewpoint dependent, mirror invariant or viewpoint invariant), the face viewpoint representations for electrodes that exhibit statistically significant identity decoding reveals a distinctly stronger mirror symmetric face viewpoint representation (which dissipates by 500 ms) relative to the non-significant electrode population.

2.4.2 Latent Representational Analysis using Confusion Matrix Mixture Models

The mixture model framework expands the possibilities for latent representational analysis within the RSA framework. It's use confers several conceptual and functional advantages but requires care toward operational considerations.

Conceptual Advantages

Conceptual advantages emerge from the use of confusion matrices as a representational primitive and the inherent linear structure of mixture models. Confusion matrices capture neural representations for decoded stimulus parameters, as seen by the underlying classifier through the lens of a distance measure. Depending upon the classifier, confusion matrices can subsume traditional distance measures used in RSA (e.g., correlation, cosine, or mahalanobis), but they can also capture representations extracted from patterns of neural activity by non-parametric classifiers (e.g., k Nearest Neighbors or Decision Trees) that have nebulous relationships with established distance measures. This broad, information focused reach of confusion matrices makes them a compelling representational primitive to support the RSA framework's own broad scope. Indeed, the RSA framework recognizes this promise by admitting confusion matrices as representational primitives (33). However, prior efforts have not realized this promise fully, possibly due to a lack of statistical tools for second latent representational analysis (63, 64). Next, modelling confusion matrices as a collection of multinomial random variables incorporates a natural noise model for the distribution of errors in discrete outcomes (corresponding to distinct stimuli), relative to the gaussian noise assumptions inherent in existing regression based approaches. Lastly, mixture probabilities (for models) and posterior probabilities (for samples) in a linear model provide a normalized and intuitive interpretation about the importance of different components in the latent representational basis at different scales (i.e., samples, population).

FUNCTIONAL ADVANTAGES

Functional advantages arise from the framework's versatility in supporting multiple analytical settings, and fusion of model predictions and external covariates (e.g., time, electrode and cortical location in this study) to obtain neuroscientifically meaningful interpretation (neurodynamics in the present study) beyond the structure of the model itself.

Purely data driven estimation of a latent representational basis lets the data speak for itself without being constrained by existing (model derived or researcher defined) representational hypotheses. Such data driven exploration can enable development of new representational hypotheses where none exist, or support the refinement and validation of existing hypotheses (as shown here). The development of new hypothesis constrained templates (Fig. 2.7.d,e,f,g) for face viewpoint in this study capture previously unreported representational characteristics revealed by data driven analysis serves as an example of fruitful interplay between data driven and hypothesis constrained analyses supported by the framework. Computational models, such as encoding models are a compelling way to explore information processing in the neural substrate, particularly when experiments are challenging or impossible to conduct. Given the large search space of possible computational models, validating model behavior and representations against neural ground truth becomes an important prerequisite to gaining insights from them, and may also guide their development. The hypothesis driven analytical setting of the statistical framework supports validating model representations against neural data with import of a fixed (i.e., non-parameterized) latent representational basis that may be obtained from computational models. Combining hypothesis driven analysis with representational results from data driven and hypothesis constrained analyses, as shown here, reveals insights about agreement (EIG_{MS}) or divergence between models and neural representations (lack of 'anchoring' in the $EIG_{LIN.}$). Usage of the framework for face viewpoint analysis demonstrates convergence toward a set of complementary representational results that combine the framework's capabilities (analytical flexibility, fusion with external co-

variates) to converge to robust insights.

OPERATIONAL CONSIDERATIONS

MODEL SIZE AND MODEL SELECTION Data driven modeling induces a natural bias toward larger models for better fits. Model selection criteria such as cross validated model selection (combined with the 1 standard error rule) (52) balance the tradeoff between model size and quality of fit, in favor of parsimony. The use of hypothesis constrained templates whose parameters are estimated in a data driven manner introduces a distinct challenge in the model selection procedures, which can be addressed by adding a simple combinatorial step into the model selection process (i.e., which combination of templates is the best in a 2 component model?). The combinatorial aspect of the problem can be quite challenging, particularly if the number of hypothesis templates available is non-trivial, and our approach (see *Methods* for exact details) constrains the combinatorial complexity by constraining the search space in a systematic manner (i.e., one a hypothesis template is removed from a large model because it fares poorly in terms of explaining the data compared to other options, it does not get considered again). Lastly, we add flexibility and introduce a search space using the dilution approach for externally imported hypotheses, since they may struggle to fit the data in cases of variable SNR in the neural activity. All three approaches for all three analytical settings are aligned with cross-validated model selection, but with slight variations for customization to unique considerations that arise in each analytical setting. The model selection approach is also flexible in terms of the measures used and cross-validated fits can easily be substituted with the Akaike Information Criterion (AIC) (65) or the Bayesian Information Criterion (BIC) (66), if desired.

MIXED EFFECTS ANALYSIS FOR REPRESENTATIONAL RESULTS The assessment of variability for results, within individuals, a group of individuals or across groups is an important step toward generalizing neuroscientific findings emerging from any analysis, including representational analysis. Mixed effects analysis, using linear or non-linear models is a frequently used

approach. However, non-parametric procedures such as the simple bootstrap or a hierarchical bootstrap (67, 68, 69) enable model free assessment of variability with the advantage of accounting for linear and non-linear effects, conservative estimates of variability as compared to model based methods for mixed effects analysis (51). The results presented here demonstrate the use of a hierarchical bootstrap approach for representational analysis to account for variability arising from subject, electrode and trial level effects on representational results.

SAMPLE COMPLEXITY The number of parameters estimated for a K component model with D categories in a purely data driven model scales as $O(KD^2)$. The number of samples required to estimate model parameters reliably increases with the number of parameters to be estimated. Sample complexity manifests as two considerations in the context of representational analysis in the context of the mixture model. The first is to have a sufficient number of confusion matrices in the data set. The second is to have a sufficient number of instances (trials) in the cells of confusion matrices being used for modeling. Sample complexity is subject to the true nature of the distribution from which samples come. For example, if the underlying rank of the high dimensional data is low and then fewer samples may be sufficient to estimate mixture parameters. However, in practice it may not be possible to make or verify such assumptions, and sample complexity remains an important consideration for the use of these models. The use of structurally constrained and minimally parameterized templates has the potential to mitigate sample complexity challenges for representational analysis of small datasets. With well framed representational hypotheses, models with fewer parameters may offer meaningful and reliable representational insights.

CHAPTER 3

A NEW PARADIGM FOR INVESTIGATING REAL WORLD SOCIAL BEHAVIOR AND ITS NEURAL UNDERPINNINGS

Eye tracking and other behavioral measurements collected from patient-participants in their hospital rooms afford a unique opportunity to study immersive natural behavior for basic and clinical translational research. We describe an immersive social and behavioral paradigm implemented in patients undergoing evaluation for surgical treatment of epilepsy, with electrodes implanted in the brain to determine the source of their seizures. Our studies entail collecting eye tracking with other behavioral and psychophysiological measurements from patient-participants during unscripted behavior, including social interactions with clinical staff, friends and family, in the hospital room. This approach affords a unique opportunity to study the neurobiology of natural social behavior, though it requires carefully addressing distinct logistical, technical, and ethical challenges. Collecting neurophysiological data synchronized to behavioral and psychophysiological measures helps us to study the relationship between behavior and physiology. Combining across these rich data sources while participants eat, read, converse with friends and family, etc., enables clinical-translational research aimed at understanding the participants' disorders and clinician-patient interactions, as well as basic research into natural, real-world behavior.

We discuss data acquisition, quality control, annotation, and analysis pipelines that are required for our studies. We also discuss the clinical, logistical, and ethical and privacy considerations critical to working in the hospital setting.

3.1 INTRODUCTION

Real-world behaviors such as social interactions are traditionally studied using simplified laboratory conditions in order to control for inherent natural complexities. Real-world environments offer the opportunity to study behavior, and its physiological correlates, in ecologically valid settings. Technological advances in recent decades have enabled us to capture and analyze critical behavioral and physiological variables in real time, over long periods of time, with greater fidelity than ever before (31, 70, 71, 72) to enable modeling real-world variability and complexity through large datasets using modern computational methodology. Doing so in real-world environments allows us to convert real-world complexities from problems to assets, which can prove transformative for understanding natural behavior and its relationship to physiology (15, 16, 18, 19).

The inpatient hospital environment is a distinctive real-world setting for investigating the relationship between behavior and physiology. It features monitoring of physiological data (electrocardiograms, electromyograms, heart rate, blood pressure, neural recordings, etc.) as part of standard care that can be augmented with behavioral monitoring (eye–tracking, egocentric video and audio recording, etc.). It also offers the opportunity to observe the relationship between behavior, perception, and physiology before, during, and after clinical events relevant to the patients' pathology. ***From a clinical perspective***, a deeper grasp of the relationship between behavior and physiology accompanying clinical events has broad implications for diagnostics and our understanding of physiological-behavioral relationships in clinical disorders (73, 74, 75). In addition, the hospital setting provides the opportunity to capture key caregiver–patient interactions,

whose salience for patients in such an environment cannot be overstated (76, 77). Modeling these interactions has deep implications in terms of understanding joint clinical decision-making, clinical information transfer, patient outcomes, patient satisfaction and the informed consent process in ways that cannot be replicated in controlled lab environments (78, 79, 80, 81, 82). ***From a basic science perspective***, the inpatient hospital environment also offers a compelling immersive environment to advance basic knowledge by studying natural behavior, such as interactions with friends and family, clinicians, eating, reading etc., in patients that have simultaneous behavioral, physiological and psychophysiological monitoring (83).

Real-world behavior encompasses a multitude of physiological and behavioral processes unfolding at different timescales, which are affected by 'change events' in the environment itself (84). This makes them challenging to study. Successfully studying the relationship between behavior and physiology in such settings requires extracting meaningful insights from data that are rich, complex and heterogeneous in nature and varied in time. Inpatient hospital settings are subject to these considerations, as well as the additional complexity of hospital environments where unpredictable and potentially adverse events may unfold for patients. In addition, they give rise to ethical considerations that include patient privacy and well-being (and potentially the privacy of others), and the confidentiality of clinical information and doctor–patient interactions (85).

This paper presents methodology for collecting behavioral and physiological data in epilepsy patients who undergo extra–operative invasive monitoring for seizure localization. Patients are implanted with intracranial electrodes (superficial, depth or a combination of both) and then are admitted to the Epilepsy Monitoring Unit (EMU) for 1–2 weeks for clinical identification of the epileptogenic zone and for functional mapping. This clinical setting presents a unique opportunity to capture behavioral data (eye–tracking using eye–tracking glasses, audio, and video

recordings) synchronized with neural activity recorded by intracranial electrodes implanted in the patient's brain, during real-world social interactions with friends, family, clinicians and researchers. We discuss the privacy and ethical considerations that arise in this paradigm and how they can be addressed, as well as logistical challenges such as fitting seizure prone patients, who have significant head bandaging protecting their implantation sites, with eye–tracking glasses to collect data in a safe and robust manner. Finally, we describe data preprocessing and data fusion pipelines that can be used to construct a high-quality multimodal data set that blends real-world social behavior and neural activity, allowing us to study the neural correlates of real-world social and affective perception in the human brain.

3.2 Materials & Methods

3.2.1 Participants

A total of 6 patients (4 men, 2 women) underwent surgical placement of subdural electrocorticographic electrodes (ECoG) or stereoelectroencephalography (SEEG) depth electrodes as standard of care for epileptogenic zone localization. Together ECoG and SEEG are referred to here as iEEG. The ages of the participants ranged from 22 to 64 years old (mean = 37 years, SD = 13.47 years). No ictal events were observed during experimental sessions.

3.2.2 Informed Consent

All participants provided written informed consent in accordance with the University of Pittsburgh Institutional Review Board. The informed consent protocols were developed in consultation with a bioethicist (Dr. Lisa Parker) and approved by the Institutional Review Board of the University of Pittsburgh. Audio and video of personal interactions were recorded during experimental sessions. Our protocol incorporated several measures to ensure privacy considerations and concerns could be addressed based on the preferences of individual participants. First, the

timing of recording sessions was chosen based on clinical condition and participant preference, to ensure that they were comfortable with recording of their interactions with the visitors present (and/or expected to be present). Second, all visitors present in the room were notified about the nature of the experiment at the beginning of each recording session and given the opportunity to avoid participation. Third, a notification was posted at the entrance of the patient room informing any entrants that an experiment was being conducted where they might be recorded so that they could avoid entering if they chose to. It is notable that there are no reasonable expectations of privacy other than for the patient, and this work was considered to meet the criteria for waiver of informed consent for everyone other than the participants themselves. Finally, at the end of each experimental recording, participants were polled to confirm their consent with the recording being used for research purposes, and offered the option to have specific portions (e.g., a personal conversation) or the entire recording deleted if they wished. Thus, explicit "ongoing consent" was acquired through written informed consent at the beginning and end of each session; providing participants the opportunity both affirm their willingness to participate and to consider the content of the recordings before giving final consent. None of our participants thus far have asked to have recordings partially or fully deleted after the recording session was complete.

3.2.3 ELECTRODE LOCALIZATION

Coregistration of grid electrodes and electrode strips was adapted from the method of (86). Electrode contacts were segmented from high-resolution postoperative CT scans of participants coregistered with anatomical MRI scans before neurosurgery and electrode implantation. The Hermes method accounts for shifts in electrode location due to the deformation of the cortex by utilizing reconstructions of the cortical surface with FreeSurfer^{TM} software and co-registering these reconstructions with a high-resolution postoperative CT scan. All electrodes were localized with Brainstorm software (87) using postoperative MRI coregistered with preoperative MRI images.

3.2.4 DATA ACQUISITION

Multimodal behavioral data (audio, egocentric video, and eye–tracking) as well as neural activity from up to 256 iEEG contacts can be recorded simultaneously during unscripted free viewing sessions in which participants wore eye–tracking glasses while they interacted with friends and family visiting them, clinicians and hospital staff responsible for their care, and members of the research team. In addition, participants also engaged in other activities like eating meals, reading, and watching television. The type and duration of activities varied across different recording sessions. The timing and duration of recording sessions were determined based on clinical condition, participant preference and to coincide with the presence of visitors in the hospital room, where possible.

Behavioral data were captured by fitting each participant with SensoMotoric Instrument's (SMI) ETG 2 Eye Tracking Glasses (Fig. 3.1.a,c). An outward facing egocentric camera recorded video of the scene viewed by participants at a resolution of 1280 x 960 pixels at 24 frames per second (Fig. 3.1.b). Two inward facing eye–tracking cameras recorded eye position at 60 Hz (Fig. 3.1.c,d). Audio was recorded at 16 KHz (256 Kbps) using a microphone embedded in the glasses. SMI's iView ETG server application, running on a laptop received and stored streaming data for all three modalities from the eye–tracking glasses by way of a USB2.0 wired connection. The iView ETG software also served as an interface for researchers to calibrate the eye-tracking glasses to each participant with a 3 point calibration procedure that enabled the accurate mapping of eye–tracking data to specific 'gaze' locations on video frames, and to initiate and stop the recording of behavioral data.

Electrophysiological activity (Field Potentials) can be recorded from up to 256 iEEG electrodes at a sampling rate of 1 KHz using a Ripple Neuro's Grapevine Neural Interface Processor (NIP) (Fig. 3.2). Common reference and ground electrodes were placed subdurally at a location

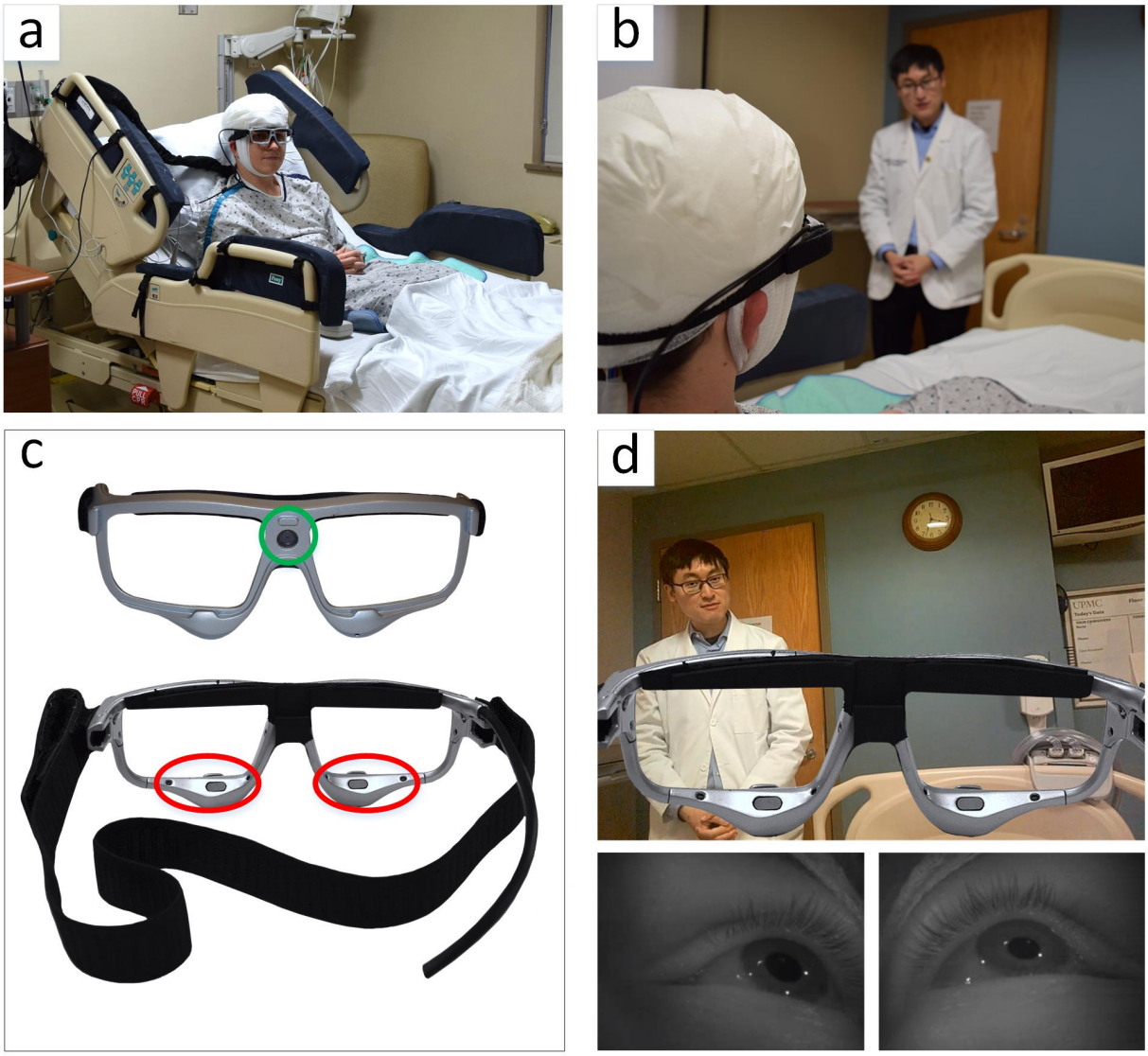

Figure 3.1: **a)** A participant in the UPMC Epilepsy Monitoring Unit implanted with iEEG electrodes, secured with bandaging, and fitted with SensoMotoric Instrument's (SMI) ETG 2 Eye Tracking Glasses that have been modified with an ergonomic Velcro strap. **b)** An over the shoulder view of the participant and the visual scene during an interaction with a researcher. **c)** Front (top) and Back (bottom) view of the SMI ETG 2 Eye Tracking Glasses with the egocentric video camera (green circle) and inward facing eye–tracking cameras (red ellipses). **d)** A snapshot of the participant's view (top) through the SMI ETG 2 Eye Tracking Glasses corresponding to panel b), and their eye movement (bottom) captured by the inward facing eye–tracking cameras.

distant from any recording electrodes, with contacts oriented toward the dura.

A MATLAB™ script, running on the same laptop as the SMI iView ETG Server software, broadcasts numbered triggers every 10 seconds, injecting them simultaneously into the neural data stream via a Measurement Computing USB-204 data acquisition (DAQ) device connected to the NIP's digital port and into eye–tracking event stream via SMI's iView ETG server application via a sub millisecond latency local loop back network connection using UDP packets (Fig. 3.2). These triggers were used to align and fuse the heterogeneously sampled data streams after the experiment, during the *Data Fusion* stage (see below for details).

BEST PRACTICES FOR BEHAVIORAL RECORDING

In each recording session, neural activity recording was initiated followed by simultaneous initiation of recording of eye–tracking, egocentric video, and audio recording via the SMI ETG 2 Eye Tracking Glasses using the SMI iView ETG Software Server. Once the recording of all modalities was underway, the MATLAB™ script was initiated to generate and transmit triggers. At the end of each recording session, the tear down sequence followed the reverse order: 1) the MATLAB™ script was terminated, marking the end of the recording, 2) the SMI iView ETG Software Server recording was halted, 3) the neural data recording stream was stopped on the NIP. Excess data from prior to the first numbered trigger and after the last numbered trigger were discarded for all modalities.

Shift in the placement of the eye–tracking glasses is possible if the participant inadvertently touches or moves them during a recording session. Such disruption can introduce systematic error(s) in eye gaze data captured after the disruption(s), although errors can be mitigated with gaze correction (see *Data Preprocessing* for details). The potential for such an event increases with the duration of a recording session. To minimize the risk of such error(s), we first instruct

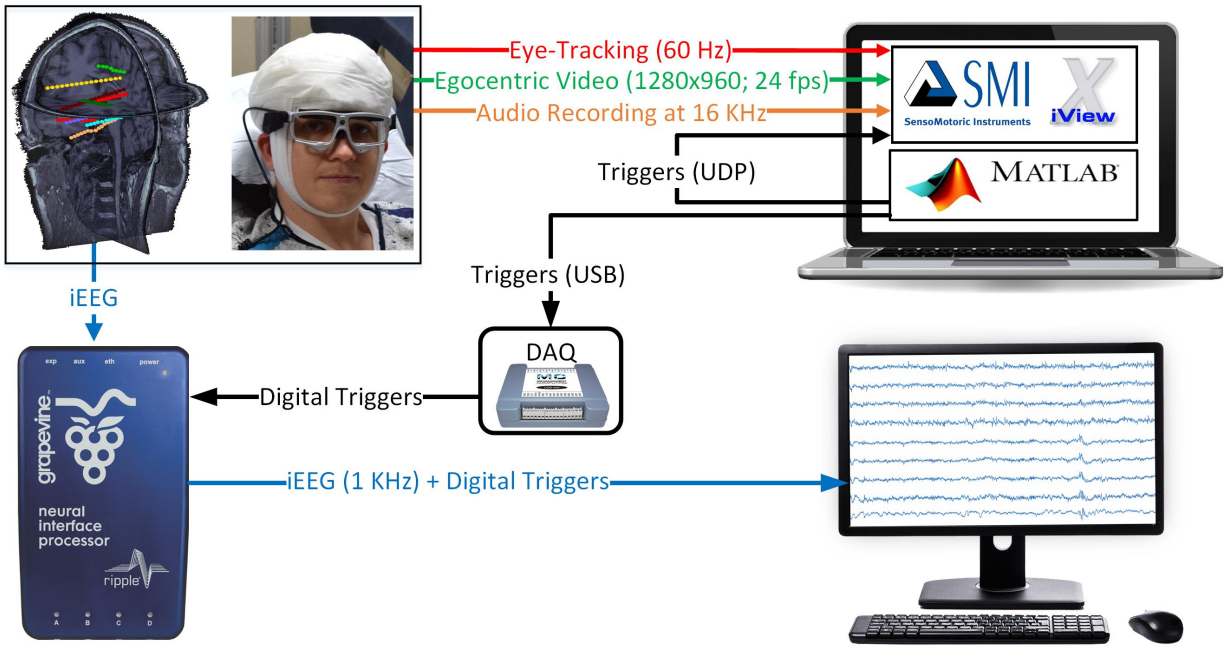

Figure 3.2: A system diagram of the experimental setup for the collection of synchronized behavioral (egocentric video, eye–tracking and audio) and physiological (iEEG recordings) from participants during real world social interactions. The green, red and blue lines represent egocentric video (1280x960 pixels; 24 fps), eye–tracking (60 Hz), and audio (16 KHz). Digital Triggers, represented by black lines, are inserted in the eye–tracking and iEEG recordings via a sub millisecond local loopback UDP connection and a DAQ respectively. iEEG recordings from up to 256 electrodes (visualized in MRI) are digitized at 1 KHz and combined with digital triggers using Ripple Neuro's Grapevine Neural Interface Processor (NIP) are transmitted and stored on a computer.

participants to avoid touching or nudging the glasses during a recording session to avoid disrupting the eye–tracking calibration completed at the beginning of the recording session. Second, we strive to reduce such errors by limiting an individual recording session to one hour and including a short break for participants. During this interlude, the recording is terminated, and participants are offered the opportunity to remove the eye tracking glasses before initiation of the next session. The interlude serves two purposes: 1. it gives the participant a break from wearing the eye–tracking glasses, helping to alleviate fatigue and discomfort; 2. initiating a new recording allows the research team to re-secure and re-calibrate the eye–tracking glasses, renewing the accurate mapping of gaze to the egocentric video. Although we prefer $\approx$1 hour recordings as a best practice, maintaining this practice depends upon participants' preference and the number visitors. In some cases, recording sessions may be longer.

3.2.5 Ergonomic Modifications to Eye Tracking Glasses

Standard clinical care following iEEG implantation involves the application of a bulky gauze head dressing. This bandaging is applied around the head to protect the operative sites where the iEEG electrodes are secured with bolts. The dressing also includes a chin wrap to provide further support in preventing dislodgement of the iEEG electrodes by securing the connector wires that carry electrical activity to clinical and/or research recording systems like the Ripple Neuro Grapevine NIP. In our studies, the bandaging typically covered the participants' ears, rendering the temples on the eye-tracking glasses unusable. To overcome this challenge, we modified the structure of the eye-tracking glasses, removing the temples and substituting them with an adjustable elastic band. We attached the elastic band to the frame of the eye-tracking glasses using Velcro patches sown at each end. The modification permitted secure placement of the glasses on the face of a participant, with the elastic band carefully stretched over the head dressing to avoid disturbing the operative sites (Fig 3.1.c). To reduce any pressure the eye-tracking glasses placed on the participants' faces as a result of the elastic band alteration, we further modified the glasses

by adding strips of adhesive backed craft foam to the nose bridge and upper rims of the frame. These ergonomic solutions enabled correct, robust, and comfortable placement of eye-tracking glasses for each participant with flexibility to adjust to individual bandaging and electrode placement configurations. As an added measure to minimize the possibility of movement for eye-tracking glasses during recording sessions, the USB cable connecting the eye-tracking glasses to the laptop was secured to the participants' hospital gowns near the shoulder with a large safety pin to prevent the weight of the remaining length of cable from pulling on and displacing the glasses during a recording session. Sufficient slack was left in the cable segment between the glasses and the fixation point on the participants' gowns to allow for free head movement while preventing the secured cable segment from pulling on and potentially displacing the eye tracking glasses.

3.2.6 DATA PREPROCESSING

The behavioral (eye-tracking, video, audio) and physiological (neural) data streams captured during a real-world vision recording were preprocessed as follows before *Data Fusion* was initiated.

EYE-TRACKING

The eye-tracking data stream is composed of time series data sampled at 60 Hz, where each sample (referred to as an eye-tracking trace) contains a recording timestamp, an eye gaze location (X,Y coordinates in the space of egocentric video) and is labeled by the SMI iView ETG platform as belonging to a fixation, a saccade or a blink. Consecutive eye-tracking traces with the same label (fixation, saccade, or blink) are interpreted as belonging to a single eye-tracking 'event' of that type, whose duration is the difference in recording timestamps of the last and first eye-tracking traces in the block of consecutive traces with the same label (fixation, saccade or blink).

As an example, a set of 36 eye-tracking traces (amounting to 0.6 second of recorded activity), where the first 18 are labeled as fixation, the next 3 labeled as saccade, followed by the final 15 labeled as fixation, would be interpreted as a fixation event ≈300 ms long (18 samples at 60 Hz), followed by a saccade event ≈50 ms long (3 samples at 60 Hz) followed by a fixation event ≈250 ms (15 samples at 60 Hz).

We developed custom Python scripts that parse eye-tracking traces and construct logs of eye-tracking events for each recording session. In addition to the duration of each eye-tracking event, the median gaze location (median is used for robustness to outliers) was logged for each fixation event and the start/end gaze locations were captured for each saccade event. Blink traces are denoted by a loss of eye-tracking (i.e. absence of gaze location) and as a result only the duration of blink events was tracked in the consolidated eye-tracking event logs.

Preprocessing of eye-tracking data also incorporates the detection and correction of systematic errors in gaze angle estimation that can be induced by the movement of eye-tracking glasses during recording sessions (e.g., if a participant inadvertently touches and moves the glasses due to fatigue), which disrupts the calibration of eye–tracking glasses (see *Data Acquisition* for details). Such issues were detected by manually viewing all experimental recordings using SMI's BeGaze application, which renders eye-gaze, audio and egocentric video together. The disruption of calibration for eye gaze tracking is visually detectable when viewing egocentric video overlaid with eye-tracking and audio because visual behavior is altered such that the gaze data fails to make sense consistently after loss of eye-gaze calibration (e.g., the subject is scrolling through a phone or reading a book or watching tv or talking to someone, but the gaze location is visibly shifted away from the obvious target). These issues were corrected using the SMI BeGaze application, which allows researchers to apply a manual correction (i.e., an offset) to eye

gaze at any time point in a recording, which applies to all eye gaze data following the corrected time point. The corrections were verified by reviewing the video that followed the correction, to ensure that corrected eye gaze data made sense consistently. Corrections to eye-tracking data preceded preprocessing in such cases.

VIDEO

Recordings of egocentric (head-centered) videos offer a broad range of visual stimuli, including objects, people and faces. Since the video recordings come from a camera mounted on the same glasses as the eye tracker they provide an egocentric view, i.e. the recorded videos capture the scene corresponding to where the participant is facing and the perspective moves as the participant's head moves. As a broad research goal, we wanted to know what objects were present in the recorded scenes. Our primary object's of interest were visitors' faces and bodies, given the objective of examining social interactions. We processed videos to identify the location of faces and body parts of people in the video recordings. As a secondary objective, we were also interested in identifying other non-face and non-body objects. Finally, for all face locations, we extracted several higher-level measures about human visual behaviors, including head pose (including orientation and position of the head), eye gaze (e.g., toward vs away from the observer) and facial expressions.

To automatically identify faces, people, and other objects, we used a computer vision algorithm - YOLO v3 (88) for object detection on each video frame. The algorithm identified bounding boxes and labels for each object present in a video frame, including faces and people. A total of 1,449,098 video frames were processed this way. While there has been great progress in computer vision for automated object detection in the last decade, it is not perfect. For example, algorithms such as YOLO v3 are trained on image data sets which contain a predetermined list of object categories, that may not include many objects that are present in a clinical

setting. In addition, objects belonging to the predetermined list of object categories may also be mis-detected (false positives or false negatives). Since the annotations were supposed to serve as ground truth for analysis of neural data, their accuracy was essential and we implemented a second stage of annotation based on human judgement, to confirm the quality of automated object detection and correct mis-detection. To avoid the time intensive prospect of manually annotating all video frames in the second stage, we annotated the first video frame corresponding to each fixation, because fixations are typically brief (a few hundred milliseconds), defined by the lack of significant eye movement, and thus it is reasonable to assume that participants look at the same location/object in a relatively unchanging scene during a fixation. We identified the video frame corresponding to the beginning of each fixation using video timestamps present in eye-tracking traces. Human annotators provided coordinates of bounding boxes for each face, or person present in video frames for a total of 125,996 frames as part of the second stage of annotation.

Finally, we used the OpenFace software (89), a facial behavior analysis toolkit using computer vision technology, to extract additional high-level information for face regions. For each face region, OpenFace provides information about (1) the position of 64 facial landmarks including eye, nose and mouth positions, (2) head orientation and position, (3) eye gaze direction and (4) facial expression information encoded following the Facial Action Coding System standard (90).

AUDIO

Audio recordings from a microphone embedded in the eye–tracking glasses capture sound from the participant's perspective. The clarity of recorded audio is influenced by the loudness of sounds and the distance of the source from the participant. Since our objective involves examining social interactions, speech detection and speaker identification are "events" of interest.

To detect time segments with speech in the audio recording and to diarize the audio (i.e. to determine who spoke when) we use state of the art deep learning speech detection (91) and speaker identification (92) pipelines available as part of an open source toolbox (93). Even these state-of-the-art models have unacceptably high error rates (particularly for diarization) for them to provide useful annotations as labels in analysis of behavior-physiology relationships. In order to overcome this hurdle, we configured these models to be highly sensitive (leading to higher false positives, but very few false negatives) and then manually reviewed model predicted time segments for speech and speaker identification, to identify and correct false positives. Outside of parameters that control the sensitivity of the deep learning models, the efficacy of speech detection and diarization is influenced by the loudness of the speakers themselves, as well as their distance from the participant (i.e., the microphone). This means that the participant's speech is the most reliably detected, while the quality of speech detection (and therefore speaker identification) for other speakers may vary. As a result, we chose to collapse audio diarization into two categories during manual review, the participant and speakers other than the participant. Segments with concurrent speech from the participant and other speakers were labeled as participant speech.

INTRACRANIAL RECORDINGS

Response potentials and broadband high frequency activity (BHA) were extracted from the raw iEEG recordings for statistical analysis using MATLAB™. Response potentials were extracted using a fourth order Butterworth bandpass ([0.2 Hz, 115 Hz]) filter to remove slow linear drift and high-frequency noise, followed by line noise removal using a fourth order Butterworth bandstop ([55 Hz, 65 Hz]) filter.

BHA extraction involved two steps. First, the raw signal was filtered using a fourth order

Butterworth bandpass ([1 Hz, 200 Hz]) filter followed by line noise removal using notch filters at 60, 120 and 180 Hz to obtain local field potentials. Next, Power spectrum density (PSD) between 70–150 Hz was calculated for the local field potentials with a bin size of 2 Hz and a time-step size of 10 ms using Hann tapers. For each electrode, the average PSD across the entire recording was used to estimate a baseline mean and variance of the PSD for each frequency bin. The PSD was then z-scored using these baseline measurements for each frequency bin at each electrode. Finally, BHA is estimated by averaging the z-scored PSD across all frequency bins (excluding the line noise frequency bin at 120 Hz).

iEEG recordings were subjected to several criteria for inclusion in the study. Any recordings with ictal (seizure) events were not included in the study. Artifact rejection heuristics were implemented to avoid potential distortion of statistical analyses due to active interictal (between seizure) or outliers. Specifically, we evaluated the filtered iEEG data against three criteria that are applied to each sample i.e., each time point in iEEG recordings, which corresponds to 1ms of neural activity. These criteria were applied to the filtered iEEG signal for each electrode, as well as the averaged (across all electrodes) iEEG signal. The first criterion labels a sample as 'bad' if it exceeds 350 μV in amplitude. The second criterion labels a sample as bad if the maximum amplitude exceeds 5 standard deviations above/below the mean. The third criterion labels a sample as bad if consecutive samples (1 ms apart at a 1000 Hz sampling rate) change by 25 μV or more. For the averaged iEEG signal, any sample satisfying any of these three rejection criteria is labeled as bad. Further, if more than 10 electrode contacts (out of a typical 128) satisfy the bad sample criterion for a particular sample, it is labeled as a bad sample. Less than 10% of the samples in experimental recordings were labeled as bad samples. All data types were dropped from analysis for fixations that contained bad samples.

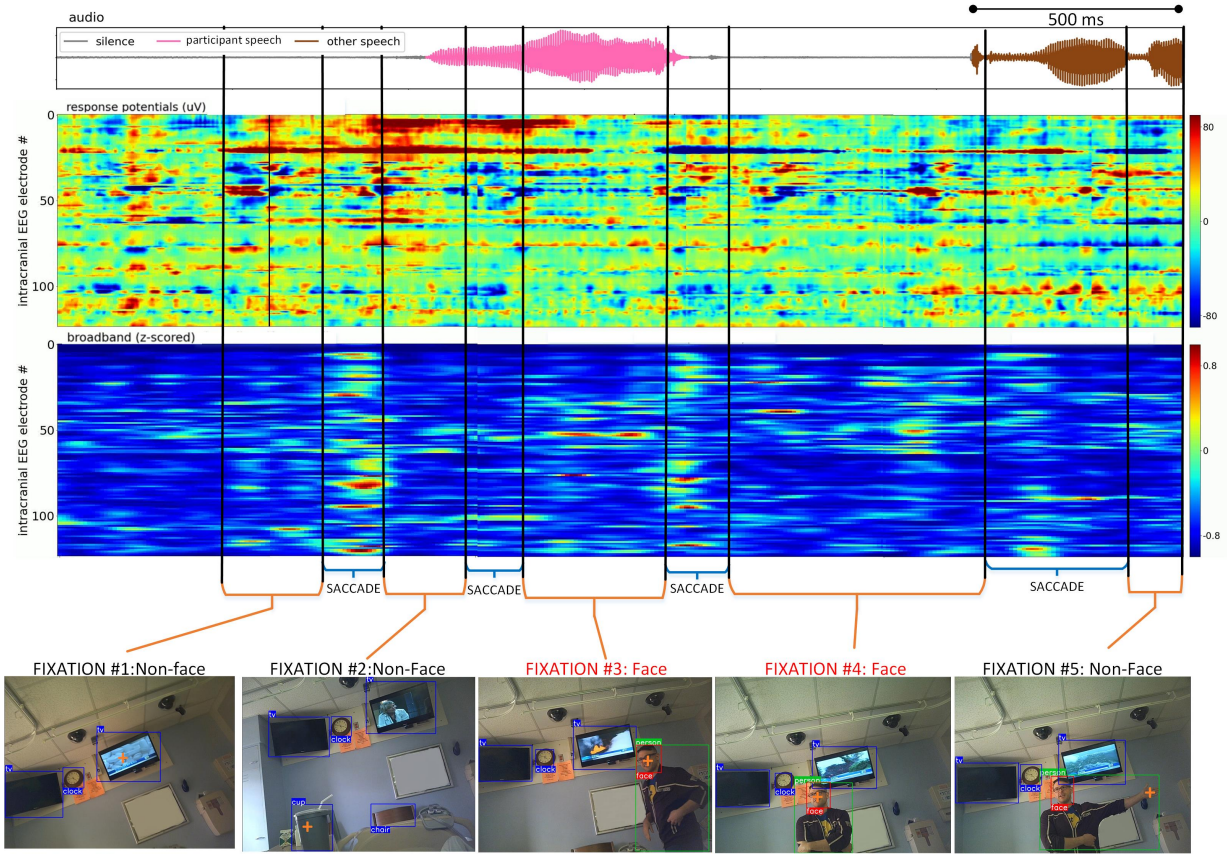

Figure 3.3: **Fused multimodal data set from a real world vision recording:** The audio waveform is shown on top, with gray, pink and brown segments denoting silence, participant speech and speech from other speakers. Response potentials and broadband high frequency activity heat maps from a 124 iEEG electrode montage are shown below the annotated audio. Vertical black lines demarcate fixations and saccades, which are marked underneath the audio and neural time series with orange and blue braces respectively. The bottom row shows video frames corresponding to each fixation event, with an orange '+' denoting eye gaze location and bounding boxes identifying the location of different objects, including persons and faces.

3.2.7 DATA FUSION

Precise fusion of heterogeneous behavioral (eye-tracking, egocentric video and audio) and physiological (neural) data streams is essential for constructing a multimodal data set to answer our questions about the neural correlates of real-world vision. In our approach, eye-tracking provides the reference modality against which video/audio, psychophysiological, and neurophysiological (neural activity) data streams are aligned in time (Fig 3.3). Each eye-tracking event is mapped to a corresponding egocentric video frame. For fixation events, we combine eye gaze location with bounding box locations/sizes from annotations for the egocentric video frame to determine what object (face or non-face) the participant is fixating upon. Each eye-tracking event is mapped to an auditory time segment and labeled as belonging to a speech or silence segment, with additional labeling for speaker identity in the case of a speech segment. Finally, neural recordings are also aligned in time to eye-tracking events based on the temporal offset of eye-tracking events and neural data, from trigger events which are injected in both data streams at 10 second intervals during recording sessions.

The quality of multimodal data sets assembled by the data fusion process described above is reliant on the quality of the heterogeneously sampled behavioral, psychophysiological, and physiological data streams fed into the data fusion process. Acquisitional variability, if present and left undetected, can severely degrade the quality of fused data sets by introducing alignment issues, and dropped video frames and/or recording offsets are common. Our methodology includes cross-verification procedures that guard against such issues with careful examination of the raw data streams for each modality. These procedures assume that the raw data captured for any modality contains accurate and sufficient timing information to diagnose and correct such issues. As long as hardware/software systems in use meet this basic assumption about raw data capture, the cross-verification approach we describe should scale. Below, we detail two specific issues that arose in our recordings using SMI ETG 2 Eye Tracking Glasses and illustrate how we

addressed them to ensure data quality in the fused data set.

SAMPLING RATE VARIABILITY

Variability in sampling rates is observed in engineered systems and can arise due to a variety of reasons ranging from temperature dependent variation in the frequencies of crystal oscillators that drive digital clock signals to propagation delays in circuit boards and circuitry running at different clock rates. If a fixed sampling rate is assumed, then these variations can accumulate as sampling drift over time and potentially lead to significant timing offsets over long periods of operation. These phenomena are addressed in engineered systems in various ways including using clocks far faster than the sampling rates desired and frequent resetting/calibration to minimize drift accumulation.

Here, we describe our approach to detect and remove such issues from the final multimodal data set that results from our data fusion procedures. We evaluated variability in the sampling rate of eye–tracking traces based on their timestamps. Since audio, video and neural data are anchored to eye–tracking events, minor sampling variability for eye–tracking does not introduce any error as long as other data streams can be aligned to eye–tracking correctly. We evaluated the timing and mapping of all other modalities (audio, egocentric video and neural data) against eye–tracking. Specifically, we found the need to address sampling rate variability that arose in the egocentric video stream, so it could be reliably mapped to eye–tracking data.

The inter-frame interval for the video stream can vary systematically by small amounts from the rated 41.66 ms (24 fps) for a recording session. These deviations can be a source of error in the mapping of eye–tracking traces to video frames unless they are addressed during data fusion. A critical marker of this problem is an inconsistency between the number of frames present in the video and the number of video frames estimated from eye–tracking traces using Eq 3.1. It

is important to note that this variability is not always accounted for in manufacturer software or documentation. The solution to this issue is relatively simple because the eye–tracking traces include a 'Video Time' column which has millisecond resolution. Instead of assuming a fixed frame rate as Eq 3.1 does

$$\underbrace{\text{Video Frame Number}}_{\text{in .avi file}} = \underbrace{\text{Video Time in seconds}}_{\text{from eye-tracking traces}} \times 24 \text{ frames per second} \tag{3.1}$$

We estimated video frame numbers corresponding to each eye-tracking trace using the 'Video Time' in them as follows

Algorithm 7 Sampling rate variability resistant mapping of video frames to eye-tracking traces

```
Initialize trace_counter = 0, frame_num = 0, video_time = 0, et_traces (from file)
while trace_counter<N do
    frame_num += round[(et_traces[trace_counter].video_time - video_time) × 24]
    video_time = et_traces[trace_counter].video_time
    et_traces[trace_counter].video_frame = frame_num
    trace_counter = trace_counter + 1
```

ADDRESSING DATA GAPS OR CORRUPTION FROM BEHAVIORAL MODALITIES

Loss or corruption of data during media recordings on embedded devices is demonstrable, and is a potential source of error for a fused multimodal data set that relies on precise alignment of multiple heterogeneously sampled modalities. As a result, our data fusion process pays close attention to identifying and characterizing such issues and addressing them to ensure data integrity. Here, we qualitatively describe different classes of issues observed in our data and how we address them to ensure data quality.

We observed missing frames in the egocentric video stream. Specifically, after correcting for sampling rate variability, we observed residual discrepancies between the number of frames that were expected per the video timestamps in the eye-tracking logs and the number of actual frames present in the video files from recordings. By evaluating timestamps for each frame in the '.avi'

files using OpenCV (94), we found that the lost frames were at the beginning of the video stream (i.e., the first K frames of an N frame video are missing) frames. We confirm this diagnosis with an additional form of verification, which used low level audio and video processing tools to manually blend audio and video streams with and without a correction for missing frames and visually verifying the absence of lip-audio synchronization issues in the resulting video. Finally, we obtained an additional point of manual verification by visualizing the ostensibly lost frames (decoders discard frames they deem corrupt when parsing a file, but they are present in the files) from the video file on a frame by frame basis, confirming that they are corrupted/garbled. The specific pattern of error (first K frames missing) observed with our experimental equipment (SMI ETG 2 Eye Tracking Glasses) may not replicate with other hardware, though given engineering constraints, other errors may arise instead. As an example, other eye-tracking glasses may have frame loss/corruption intermittently during a recording instead of at the beginning. However, our observations suggest that such issues may exist with other eye-tracking devices and data fusion pipelines should incorporate verification stages that can identify and correct such issues, with a preference for multiple modes of verification that are consistent with each other.

Blinks are a natural part of visual behavior and the eye-tracking records denote them as such. Since eye–tracking is lost during blinks, there is usually no information about gaze, pupil dilation etc. available for blink events. We see blinks interspersed among fixations and saccades, and they are typically a few hundred milliseconds long. However, we observed longer periods lasting several seconds in multiple recordings. To understand this phenomenon better, we viewed the videos for periods where this happened, with gaze information overlaid using SMI's BeGaze software. We found these anomalous blinks to be masking a real phenomenon, where the participant may be looking out the corner of their eye, which takes their eye-gaze outside of the field of vision of the egocentric camera or upon occasion, potentially taking their pupils outside of the field of vision of the eye–tracking camera. Since the system cannot accurately capture visual

behavior as it relates to the video in these conditions, it labels those periods as blinks. These scenarios are easy to spot during manual inspection because the eye-gaze location before and after the blink tends to be near the periphery of the video frame. These conditions are not a significant challenge for data quality, because they can be easily dropped from analysis. However, awareness of their existence is meaningful for data fusion pipelines.

3.3 RESULTS

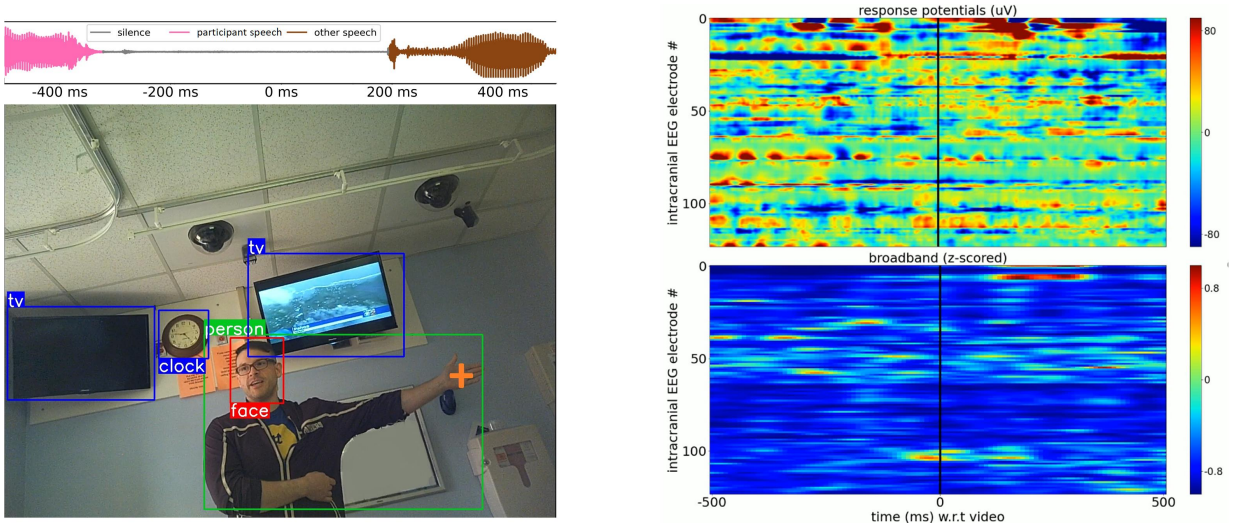

Figure 3.4: **A snapshot of fused multimodal (audio, egocentric video, eye-tracking and iEEG):** On the left, an annotated audio snippet (top) and video frame (bottom) visualizes the world through the participant's eyes and ears as they interact with friends and family visiting them during a recording session. Speech/silence and speaker diarization labels color the audio signal on top. The annotated video frame below depicts the participant's eye gaze location with orange '+' marker with colored bounding boxes identifying the location and sizes of different objects detected by computer vision models and verified by human annotators. The panel on the right visualizes 1 second of neural activity across 124 iEEG electrodes, corresponding to the video frame/audio on the left, with response potentials on top and broadband activity at the bottom (see *Supplemental Video* for a dynamic version of this figure).

We collected iEEG recordings from patients in the Epilepsy Monitoring Unit (EMU) who wore SMI ETG 2 Eye Tracking Glasses as they went about their day interacting with friends and family visiting them as well as members of the clinical team. We used computer vision

models to identify objects, faces and persons (bodies) in videos of the visual scenes in front of the participants during these sessions. Similarly, we used speech processing models to identify speech intervals and diarize the audio recorded from the internal microphone in the SMI ETG 2 Eye Tracking Glasses. All annotations from computer vision and speech processing models were validated and corrected, if necessary, by human annotators to ensure data quality. Here, we show that fused multimodal datasets (see Fig 3.4 for a snapshot; see *Supplemental Video* for a dynamic version) which include annotated audio, eye-tracking, annotated video, and iEEG, can be developed using this process. Such datasets can help advance visual neuroscience research beyond traditional experimental paradigms and explore the neural correlates of real-world social vision.

3.3.1 Behavioral Data

We collected data from 6 participants across 11 different free viewing recording sessions which ranged from 41 - 143 minutes long and added up to a total of 16 hours and 48 minutes. Social contexts differed across recording sessions and sometimes within a recording session, in terms of the number of individuals present, the number of interactions they had with the participant and the nature of those interactions.

Visual Behavior

SMI Eye Tracking glasses captured visual behavior, categorizing each moment's sample as belonging to a saccade, fixation, or blink. Visual behavior varied depending upon the social context during recording sessions. Saccades usually accounted for 10 - 15% of the recording duration (Fig 3.5.a), even though they account for nearly half the events (after accounting for blinks and occasional loss of eye-tracking) (Fig 3.5.b) as a result of the saccade–fixation–saccade structure of the active sensing cycle, a contrast highlighted by the skew in the distribution of saccade durations and fixation durations (Fig 3.5.c). Saccades and fixations are not perfectly balanced

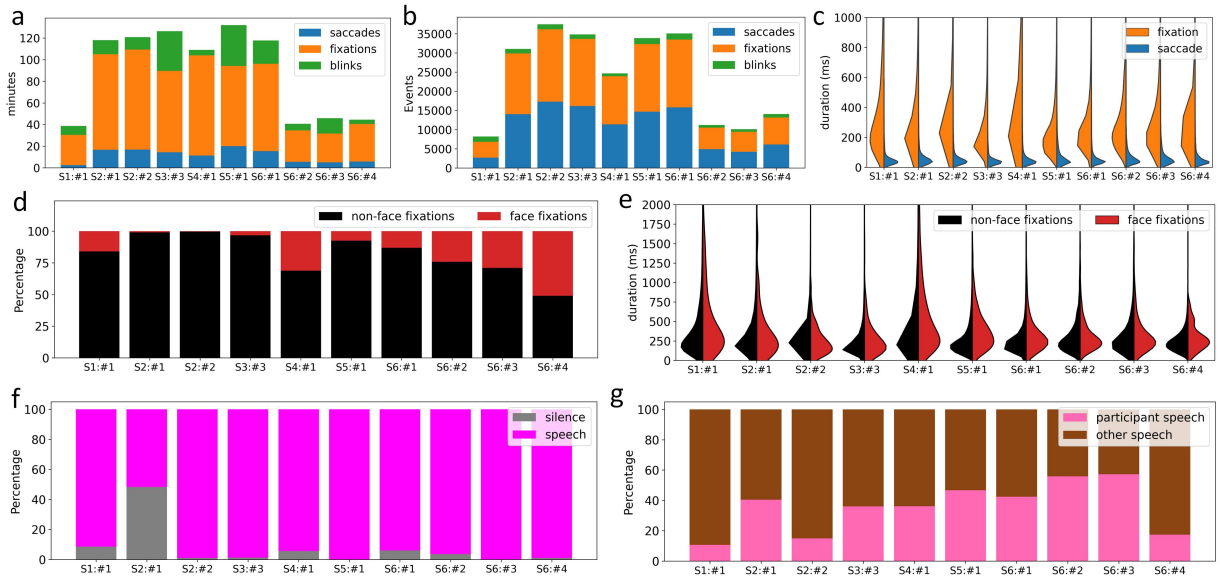

Figure 3.5: **Summary of dataset spanning visual behavior and the auditory environment: a)** The duration of each recording session (with multiple sessions for each participant) broken down by time spent in different visual behaviors (saccades, fixations and blinks). **b)** Similar to a), but counting distinct events for each visual behavior instead of time. **c)** Saccade and fixation duration distributions for each recording session. **d)** The fraction of time fixations were on faces and non-face objects for each recording session. **e)** Fixation duration distributions for face and non-face targets for each recording session. **f)** The fraction of each recording session broken down by time spent in silence and speech. **g)** The fraction of each recording session broken down by speech from the participant and other speakers.

due to the loss of eye-tracking from blinks and other reasons (e.g., noisy conditions, participants closing their eyes for brief periods or looking out of the corner of their eye during the recording sessions).

We identify fixation targets by combining gaze location from eye-tracking with bounding boxes from the video frame corresponding to each fixation. We categorize fixations as face and non-face fixations, reflecting our focus on the social aspects of real-world vision. The social context during a recording session has a natural influence on the distribution of fixation targets. We found that participants fixated on faces less than 30 - 40% of the total time spent fixating during a recording session (Fig 3.5.d), even in the most social situations (e.g., EMU room full of multiple family and friends, with active conversations). The distribution of fixation durations for the two fixation categories showed that face fixations tend to be a little bit longer (Fig 3.5.e), indicating that even during the most social situations with familiar people, we look at the faces of people around us infrequently but when we do look at them we tend to hold them in our gaze a little longer.

AUDITORY CONTEXT

The SMI ETG 2 Eye Tracking glasses also recorded audio using an in-built microphone. We used deep learning models (93) to do auditory scene analysis, augmenting it with manual annotation to ensure high reliability. Once again, depending upon the social context during each recording session, we observed varying levels of verbal discourse (Fig 3.5.f). We observed that speech could be detected from both the participant and others in the room, but the participant was reliably comprehensible due to their proximity to the microphone, whereas the volume and comprehensibility of the voices of other speakers would vary based on how close they were to the participant, making source identification more challenging even for manual annotation. To avoid potential confusion during manual annotation, we restricted speech diarization during supple-

mental manual annotation/verification to classifying speech as originating from the participant or other persons in the room. We found that the participant's own verbal behavior varied across recording sessions, with comparable speech in the room, across recording sessions, even for the same participant (Fig 3.5.g).

BEHAVIORAL ANNOTATION: RELIABILITY AND ITS COST

Egocentric Video

Automated software driven annotation of video frames is straightforward and fast, but accompanied by a trade-off between speed and accuracy. The speed of automated annotation depends upon the algorithms used for object detection. YOLO v3 (88) is a popular algorithm for object annotation (detection), performing at a rate of 45 fps on a NVIDIA K40 Graphics Processing Unit (GPU), or 5 fps on a standard CPU. This means annotating an hour of video takes 32 minutes with a GPU, or close to 5 hours with a CPU.

The accuracy of annotation algorithms is not high. We measured the quality of the automated annotations by comparing the automated annotations from software with human annotations for all sessions. We found that software-driven annotation only achieved an average of 69.5% Intersection over Union score (a measurement for evaluating object detection algorithms, higher the better, with a threshold of 100%). This means that the overlap ratio between the software's bounding boxes and the human annotators' bounding boxes was only 69.5%, suggesting the accuracy of automated software-driven annotation may be limited.

Although human annotators produce higher-quality annotations, the process is time and labor-intensive. Human annotators annotated 125,996 frames out of 16 hours and 48 minutes of videos. The total time spent on annotating 125,996 frames was 104 hours, with an average of 6.19 hours for an experienced human annotator to annotate an hour of video.

For quality control, 3% of frames from two sessions (sessions from S5:#1 and S6:#4) were randomly sampled and verified by a second annotator. The overlap ratio between the first annotator's bounding boxes and the second annotators' bounding boxes was 97.3% and 97.4% respectively for the two sampled sessions. This process underscored the significantly higher quality of human annotation over the automated software-driven annotation.

Speech Detection and Diarization

Automated speech detection and speaker identification were computationally efficient, with an hour's audio being processed within 1 - 2 minutes. Manual verification and correction of mis-detection was done with manual annotator's listening to the audio, and correcting false negatives/missed speech and false positives/speech labeled as silent and required an hour of manual effort for each hour of audio recording. Comparisons of automated speech detection with manually verified/corrected speech intervals for the first 10 minutes of each recording session revealed mis-detection (speech classified as silence or vice versa) for $\approx$ 4% of the annotated audio. Manual annotation for speaker identification involved collapsing the automated speaker diarization labels into two categories, 'participant speech' and 'other speech'. Speech segments where the participant and other individuals were speaking concurrently were labeled as 'participant speech'. Manual annotators listened to the full length of the recording assigning new labels to each speech segment manually, which took 75 minutes for each hour of speech.

DATA FUSION ISSUES: DETECTED AND CORRECTED

Next, we show some results which motivate careful evaluation of the raw data for each modality before data fusion of heterogeneously sampled data streams from an experimental recording is attempted. Specifically, we describe and quantify alignment issues between eye-tracking and video data collected using SMI ETG 2 Eye Tracking glasses, that were identified and corrected

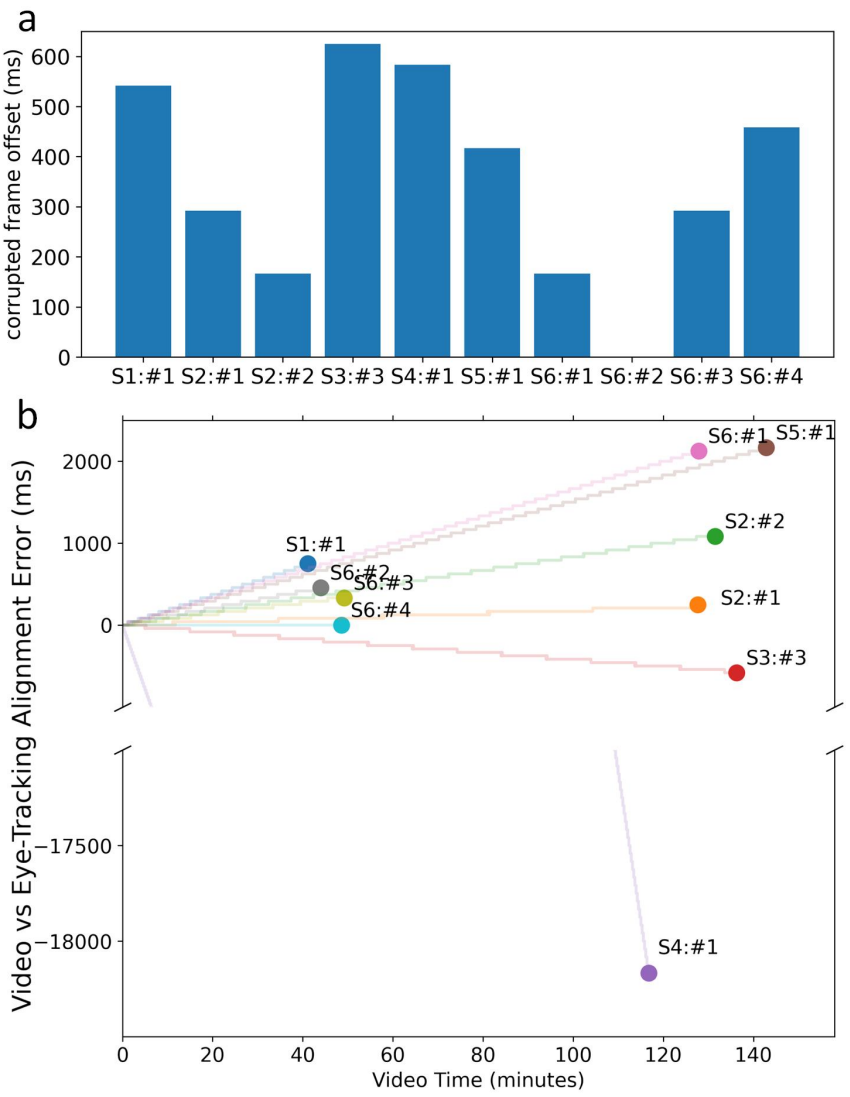

Figure 3.6: **The potential effects of video frame corruption and video frame rate variability on the accuracy of data fusion:** Visualization of timing error for each recording session introduced in the alignment of eye–tracking events and corresponding egocentric video frames in the case **a)** Corrupted frames at the beginning of each video file are not detected and corrected in the eye–tracking to video frame alignment procedure. This is a fixed error that affects all eye-tracking events in a session. **b)** The procedure to map frames to eye–tracking traces does not address small variations in frame rate (i.e., Eq. 3.1 instead of Algorithm 1). This is a time varying error which accumulates over the duration of a recording (scatter points indicate the final accumulated error at the end of each recording session) and its rate of accumulation (slope of shaded lines in the background) depends upon the magnitude of the deviation in video frame rate is from 24 fps. We observe deviations as small as 1 frame (41.67 ms) over a 49 minute recording for S6:#4 and as large as 432 frames (18 seconds) over a 2 hour recording for S4:#1.

(see *Methods for details)* during data fusion. We found two issues in the video stream, which would lead to misalignment between eye-tracking traces and the video frame they correspond to.

The first issue was related to corrupted and unrecoverable egocentric video frames at the beginning of each recording (see *Methods* for details). The duration of egocentric video lost as a result of this issue varied by recording, and ranged from the first 0 ms - 625 ms (Fig 3.6.a). In a video with the first N frames corrupted, this issue would lead to incorrect mapping of eye-tracking traces to a video frame $N+1$ frames later than the egocentric video frame they corresponded to, which could lead to errors in annotation of fixations (e.g., as face or non-face fixation) across the entire video. After correction, the only impact of this issue is that eye-tracking traces/neural data for the first few frames that are corrupted and discarded cannot be used for analysis, which is a very minor loss.

The second issue was related to variability in the average frame rate for egocentric video recorded from each session. We observed that for different sessions, the average frame rate of the recorded video was slightly above or below 24 frames per second. Eye-tracking traces are mapped to video frames using a 'Video Time' variable embedded in them. Estimating the video frame number corresponding to an eye-tracking trace using Eq. 3.1 which assumed a frame rate of 24 fps that was slightly higher or lower than the real frame rate of the video. The discrepancy led to an error between the estimated frame and the real frame corresponding to eye-tracking traces, which accumulated as the video recording progressed (Fig 3.6.b) and became visible with the eye-tracking traces mapping to far fewer/greater frames than were present in the video at the end of the recording. This problem was avoided by using the procedure defined in Algorithm 7, which is robust to these small variations in frame rate (see *Methods* for details). Both these problems co-occurred and addressing them as described in the *Methods* section gave us perfect consistency between the number of frames estimated in the eye-tracking traces and the number

of frames present in the egocentric video. Lastly, we also evaluated audio and neural activity for similar alignment inconsistencies with the eye–tracking logs and found no issues with alignment.

3.3.2 Neural Correlates of Real-World Social Vision

The number and cortical locations of intracranial EEG electrodes from which neural data were recorded varied by participant with a total of 686 cortical locations distributed across the temporal, parietal, occipital, frontal and cingulate areas of participants (Fig 3.7.a, b).

Finally, we aligned neural activity recorded from intracranial EEG electrodes to the composite behavioral (eye-tracking + visual behavior + auditory context) log using digital triggers embedded in the neural and the eye-tracking data streams. This final step allows identification and extraction of neural activity corresponding to individual eye-tracking events (saccades, fixations, and blinks).

Our analysis of real-world vision is anchored to fixations, and Fig 3.7.c visualizes average Fixation Response Potentials (FRPs) and Fixation Related Broadband High Frequency Activity (FRBHA) for face and non-face fixations from several of the 686 intracranial EEG electrodes for which real-world vision data were collected. Typical aspects of the FRP (e.g. enhanced N170 for faces, particularly in ventral temporal cortex locations) and FRBHA (42, 43, 98, 99, 100) are well represented for electrodes from multiple lobes suggesting the alignment of neural activity and eye–tracking events is robust and provides a key ”proof-of-principle” for this real-world paradigm, similar to that provided by recent studies in macaque monkeys engaged in free viewing of natural scenes (101).

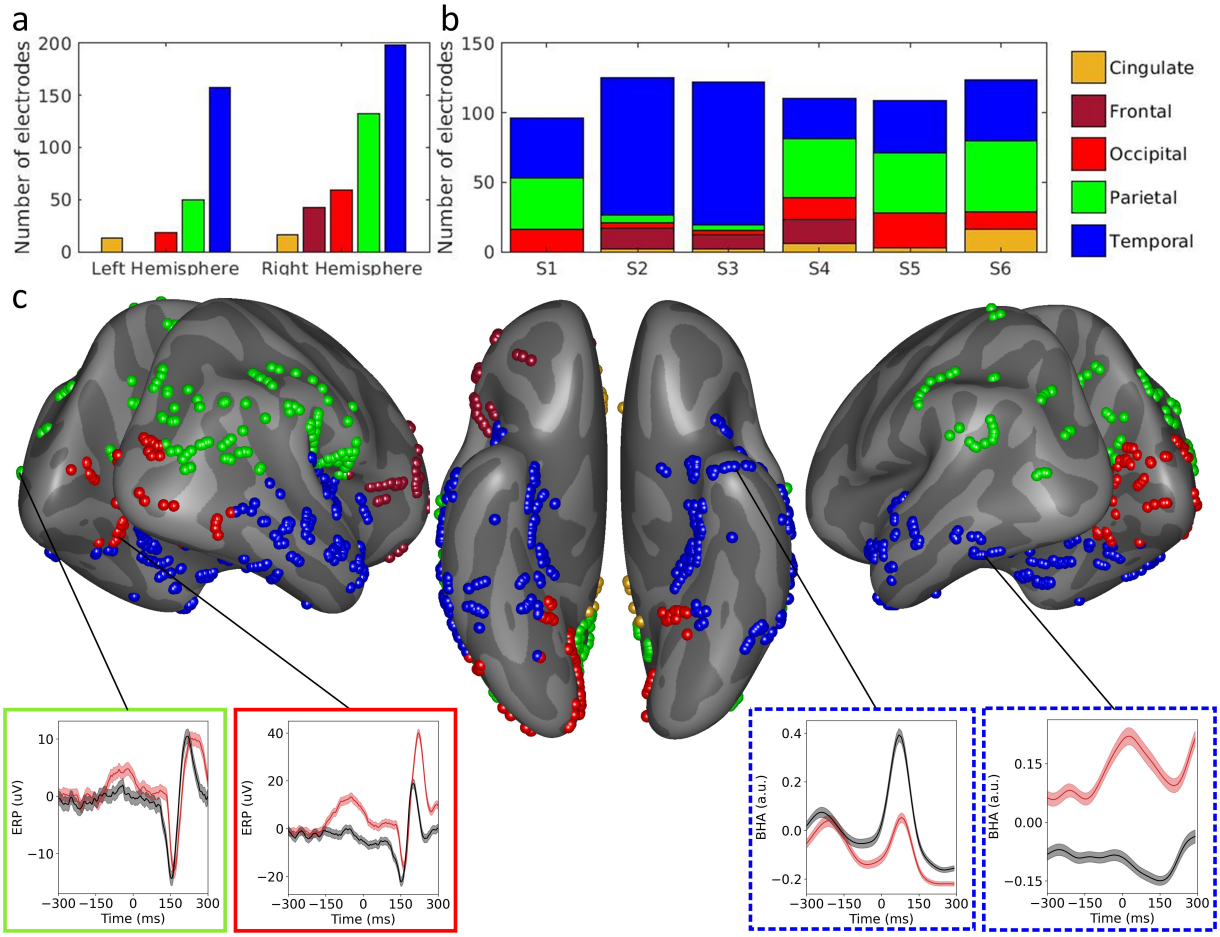

Figure 3.7: **a)** Cortical distribution of the 686 intracranial EEG electrodes from 6 participants over different lobes across the Left and Right Hemispheres per the Desikan Killiany atlas (95). **b)** Per participant electrode distribution across different cortical regions **c)** Visualization of locations of electrodes from all participants on an inflated cortical surface with ventral, lateral (left and right), posterior and anterior views. Average fixation locked neural activity from electrodes sampled across all participants and recording sessions. The colors of the boxes correspond to the lobe of the cortical location being sampled and the outlines denote the neural signal that is visualized (solid lines denote Fixation Response Potentials (uV) and dashed lines denote Fixation Response Broadband Activity (a.u.)). The average fixation locked response to ≈1000 fixations of face (red) and non-face objects (black) each is shown for each cortical location. One notable result is that differences in the neural response between face and non-face fixations appear prior to fixation onset, suggesting predictive activity/"pre-saccadic preview"(96, 97)

3.4 DISCUSSION

We investigated the feasibility of combining neural recordings from iEEG electrodes with eye-tracking, video and audio recordings collected using eye-tracking glasses and annotated using computer vision and speech models to generate robustly fused multi-modal data sets from unscripted recording sessions in an inpatient hospital environment. Fusion of visual behavior with neurophysiological recordings enables investigation of the neural correlates of real-world social vision and affective perception. Summary views of the data highlight the heterogeneity that emerges in uncontrolled behavior in ecologically valid settings, and underscore the need for care when trying to assess generalizability of observed effects across individuals. A natural approach to address these challenges is to define summary variables or learn then using data driven approaches like multiset canonical correlation analysis (102). The efficacy of our methodology is validated in the context of real-world social vision by fixation locked neural activity (FRPs and FRBHA) for face and non-face fixations from ventral temporal electrodes, which show category selective neural signatures that are also observed in traditional visual neuroscience experiments. Our initial findings also point to several potential opportunities for the enrichment of behavioral and physiological data collection as well questions of significant interest for clinical and translational research.

3.4.1 ENRICHING BEHAVIORAL MONITORING

HIGHER FIDELITY CAPTURE OF VISUAL BEHAVIOR

From analyzing the data sets presented here, three natural opportunities to improve the capture of visual behavior are apparent. The first entails higher fidelity data acquisition for behavioral data streams that we already capture. The eye-tracking glasses used in this study feature a single head-centered perspective (egocentric) video camera operating at 24 frames per second with a resolution of 1280 x 960 pixels capturing a 60° (horizontal) by 46° (vertical) region of the

field of vision, with 2 eye-tracking cameras operating at 60 Hz. Increasing the spatial resolution of the video camera in pixels, improving the temporal resolution of both eye-tracking and video and capturing a larger fraction of the field of vision can aid in better tracking of visual behavior over a more complete portion of the field of the vision. The second opportunity requires adding a new data modality (head position) using an Inertial Measurement Unit (IMU), that can provide tracking for the physical frame of reference corresponding to each video frame. The third opportunity involves considering the addition of depth perception information for eye-gaze, which may potentially be supported by the addition of a second egocentric camera or LIDAR (103). A review of available research grade hardware (104) provides an account of the capabilities of several research grade devices, which can be evaluated for their suitability with respect to each of these possibilities.

AURAL SCENE CAPTURE

Analysis and annotation of the auditory scene recorded using the in-built microphone embedded in the eye-tracking glasses reveals the potential advantages of capturing the aural scene as well as the limitations of having a single microphone physically attached to the patient. The potential addition of high definition microphone arrays in the room can enable a complete recording the auditory scene, including the capture and source localization of all sound, including speech. In the context of social behavior, such an enriched capture offers the opportunity to go beyond speech and speaker detection and into speech recognition, and its conversion to text (105, 106, 107) thereby allowing the use of language models that could add an additional behavior modality for semantic and sentiment analysis (108).

FROM MONITORING VISUAL BEHAVIOR TO VISUAL MONITORING OF BEHAVIOR

Heavily monitored inpatient hospital environments like an EMU are typically equipped with cameras that allow clinical care teams to monitor patient behavior. The same video streams

also capture the physical behavior of other individuals (e.g., doctors, nurses, family) who are present. These video streams hold the potential to add two additional behavioral modalities to the multi-modal data set we have described. The first modality is affective behavior, for the patient and other individuals present, extracted using facial analysis tools like OpenFace (89). The second modality is physical behavior using tools like OpenPose (109) and DeepLabCut (110, 111, 112, 113, 114), which may enable us to explore the relationship between physiology and behavioral phenomena like interpersonal synchrony (115).

3.4.2 ENRICHING PHYSIOLOGICAL MONITORING

As part of standard care, inpatient hospital environments feature the monitoring of a wide variety of physiological data like EKGs, EMGs, heart rate, pupillometry, blood pressure, neural recordings, pulse oximeter readings, saliva samples, urine samples as well as clinical events. A richer physiological data set than the one presented here – one that contains a greater number of the physiological modalities – can combine powerfully with behavioral markers to allow pursuit of highly relevant clinical and translational research questions.

As an example, attention and arousal are thought to be modulated by the locus coeruleus-noradrenergic (LC-NE) system. Pupil size (116, 117, 118) in absence of lighting change and heart rate (119) are both considered proxies for locus coeruleus (LC) activity. A data set that fuses EKG and pupillometry with human intracranial EEG along with visual behavior recorded during real–world social interactions, such as those between patient-participants and clinicians, can enable investigation of the neural correlates of arousal and attention in ecologically valid and clinically salient settings.

3.4.3 ETHICAL CONSIDERATIONS

Ethical considerations presented by research involving video and audio recording of real-world behavior in a clinical environment include issues of privacy protection, data sharing and publication of findings, and challenges of obtaining informed consent (120, 121). Studies involving such recording affect the privacy of not only participants, but also the visitors, clinicians, and researchers with whom they interact. We believed, and the institutional review board concurred, that with regard to those interacting with participants, this study met the criteria for waiver of informed consent, because obtaining consent was impracticable and the study presented only minimal risks to visitors and others interacting with participants. Instead, a notice was placed on the door of patient rooms to alert anyone entering the room that video and audio recordings would be acquired. Visitors could opt-out by not visiting, or by requesting that their visit not be one of the interactions recorded (perhaps by rescheduling the visit). Clinicians were not able to opt-out of entering and being recorded, as they were required to provide standard care; however, they were informed in advance that the study was being conducted and could raise concerns about their presence and interactions being recorded. These concerns are addressed on a case-by-case basis. (One can imagine, for example, that for reasons of personal safety a clinician might not want her employment location to be made public through future publication/presentation of study findings.) Moreover, the faces of those interacting with participants are to be obscured in all tapes/photos that are either shared or published.

The risks to participant privacy were more substantial, and were simultaneously compounded and mitigated by the clinical environment. In comparison to home environments, inpatient settings afford a lower expectation of privacy, with hospital staff coming and going, rooms often left open to the hallway, and, in some cases, rooms being under video and audio monitoring for reasons of clinical care. Patients generally trade-off their privacy for the prospect of clinical benefit. Nevertheless, the study involved greater reduction in privacy and for reasons that afforded

no direct benefit to the participants themselves.

Participants were asked to give informed consent to study participation, including the video and audio recording, collection of physiological data, data sharing, and publication of study findings. Study procedures — putting on, calibrating and wearing the eye-tracking glasses — served to remind participants that their behavior was being recorded. At the end of each recording session, patient-participants were asked to consider the events that happened and explicitly consent to the recording being used for research purposes. In addition, separate consent/release was acquired for use of the video and audio recordings in figures for publications or in presentations. This is especially important because the study took place in a particular clinical setting, and thus for participants who are identifiable in the recordings, publication/presentation of findings would reveal health-related information about them—namely that they were in an Epilepsy Monitoring Unit.

The question of data sharing for recordings that are inherently not de-identifiable is an additional issue to consider. Processed data (annotations with identifiable information removed, for example audio diarization and generic aspects of the computer vision annotations) could likely be shared openly as long as substantial care was taken to assure de-identification. Sharing raw data is a bigger challenge and would require additional layers of consent such as consent procedures used when creating public behavioral databases, though even with this level of protection care must be taken given the potential sensitive nature of the recordings in a clinical environment. Thus, at most, well curated snippets of raw data may be publicly shareable, and sharing of raw data would likely have to be done under IRB approval at both institutions with a data use agreement.

In this study, we sought to study natural real-world social interactions and thus avoided

recording doctor-patient interactions or clinical events. For studies that seek to understand doctor-patient interactions or clinical events, these protections and privacy concerns become even more acute and participants should be reminded when acquiring both pre- and post- session consent that the video/audio recordings will include sensitive clinical information.

3.4.4 Implications for Clinical and Translational Research

Real-world social interactions in an inpatient hospital setting include caregiver–patient interactions (81, 82, 122), which include interactions with neurosurgeons and epileptologists in the case of patients in the EMU. Capturing physiological and behavioral data corresponding to these interactions offers a unique opportunity to understand how clinical decision making in these dyadic interactions is affected by different circumstances based on factors like the severity of clinical issues involved, the presence of family, the patient's mental health. A deeper understanding of the relationship between patient physiology and behavior that accompanies clinically important interactions has profound implications for clinical practice, patient outcomes and patient satisfaction (123). Lastly, the described workflow can be applied to better understand seizure semiology, which is the keystone for seizure localization and directly related to optimal post-operative results in curative epilepsy surgery.

3.4.5 Neural Basis of Real-World Behavior

Ecological validity is essential to the investigation of social behavior in the real world. The experimental paradigm we describe here is part of an emerging effort to address this challenge (15, 31, 124, 125, 126). Laboratory psychology and neuroscience allows for tightly controlled experiments that are crucial for the advancement of knowledge and many aspects of what is discovered in these tightly controlled experiments have external validity (127). However, an ecological approach often yields results that differ from those of laboratory experiments (15, 16, 17, 18, 19). For example, recent studies have shown that eye gaze patterns for static

faces or even movie faces are very different from those observed during actual face-to-face interactions (128, 129, 130, 131, 132, 133) and real world settings have been shown to activate broader brain networks than do artificial conditions (126, 134, 135). Moreover, the "naturalistic intensity" (134) of an interaction with one's loved ones or a doctor or a threatening stranger is a key element of real-world experience that cannot be fully captured in a laboratory. Basic aspects of the organization of the "social brain" (136) are unlikely to change in real-world environments, for example regions of the brain that show face selectivity in the lab (98, 137, 138) remain face selective in natural conditions (Fig. 3.7), as expected given that disruption to these regions cause real-world face processing abnormalities (139, 140, 141). However, important aspects of how these regions code and process social information are likely to reflect real-world processes that differ from the laboratory environment. At a minimum, it is important to validate laboratory findings in real world settings to determine the generalizability of models derived from controlled experiments (127).

The complexity of studies in the real-world is that there is enormous uncontrolled variability in natural environments. However, modern computational studies, such as those in artificial intelligence and computer vision, show that real-world variability can be well-modeled with sufficient data. Our paradigm is designed to enable real-world neuroscience by facilitating the collection and processing of large datasets combining behavior, physiology, and neural recordings that can be analyzed using modern computational techniques to test hypotheses about social behavior and its neural bases in natural environments.

The movement towards studying the neural basis of real-world behavior has also been seen in recent studies with non-human subjects, enabled by the potential of telemetric recordings that allow for neural activity to be recorded during natural behavior (142, 143, 144). Parallel studies of natural neuroscience in non-human primates has the potential to allow for a deeper under-

standing of the cellular-to-systems mechanisms for basic pan-specific aspects of social behavior and cognition. Advances in computer vision provide the opportunity to annotate nonhuman animal behavior and in relation to details of a natural environment (113) just as they do in human studies. Recent work has also demonstrated that restraint free, real-world eye tracking is also possible in non-human primates (145, 146). Thus, the approach described in this work could be adapted to parallel studies in non-human primates, leveraging the higher resolution methods that are possible to use in nonhuman primates, to allow a cellular-to-systems understanding of the neural basis of real-world cognition and perception.

3.5 CONCLUSION

We view the approach outlined above as part of an ongoing paradigm shift in approach towards studying real-world behavior and cognition and their neural underpinnings. Real-world "naturalistic intensity" and ecological validity is particularly important for studying social interactions and their neural correlates. Our current methodology augments eye-tracking and behavioral monitoring in experimental recording sessions in the EMU with neurophysiological monitoring. Extending behavioral monitoring to unscripted and more real-world contexts can enable the collection of multi-modal data sets that are large enough for cutting edge machine learning techniques like deep learning to be pressed into service to learn relationships between behavior and physiology. Combined behavioral and physiological data can be used both for studying basic cognitive phenomenon and can also be used to find markers that are predictive for clinically significant events like seizures, cardiac events, respiratory events, and others.

CHAPTER 4

RECONSTRUCTING THE NEURAL CODE FOR REAL WORLD FACE PERCEPTION

A central goal for neuroscience is to understand how our brains process information in real life, such as faces during natural social interactions. We harnessed multi-electrode intracranial recordings from hours of unscripted interactions participants had with friends, family, etc. Videos of faces being viewed could be reconstructed from brain activity alone and vice versa, which emphasized the importance of the social-vision pathway to natural face perception. Sharper neural tuning was revealed for the type of facial expression over its intensity. There was greater sensitivity for subtle differences from a person's resting expression than from strong expressions – a Weber's law for facial expressions. These results suggest that oval-shaped neural tuning for the kind and intensity of facial expressions reflects the neural code for real-world face processing.

4.1 INTRODUCTION

How does your brain code your daughter's facial expressions and movements while playing snakes and ladders together? This question illustrates a central goal of neuroscience – we seek to understand how the brain processes information during natural behavior in the real world. We study face perception to understand how our brains process the identity, expressions, and facial movements on people's faces during natural social interactions in the real world. Important

discoveries, such as the existence of an extended face processing network and aspects of how it codes for faces, have come from laboratory paradigms that monitor brain activity while participants view faces on a screen under tight experimental constraints (147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157). However, the fundamental question of how our brains process the expressions and movements of real faces during natural, unscripted interactions with people remains unanswered. Addressing this central neuroscientific goal requires the answer to two intertwined questions: Can we model the unconstrained variability of faces during free, natural social interactions in the real world? And, if so, how can we understand the underlying neural representation by testing hypotheses about the neural code for facial expressions and movement during real-world interactions?

Here, we harnessed simultaneous mobile eye-tracking and intracranial recordings from five human participants undergoing clinical monitoring for seizure localization for 1-2 weeks. This environment afforded a rare opportunity to study their brain during hours of natural interactions with friends, family, clinicians, and researchers. Patients who chose to participate in this research wore eye-tracking glasses (Fig. 4.1A) that recorded both their field of view (Fig. 4.1B) with an outward facing world camera and where they looked in this scene with inward facing eye-tracking cameras. Computer vision was used to detect faces in the video from the world camera and combined with eye-tracking to determine when patients looked at faces (Movie 1). Brain activity recorded from intracranial electrodes (Fig. 4.1C) was aligned with eye-movements (fixation-locked brain activity) (Fig. 4.1D) to probe the neural signatures of faces. The pose, eye-gaze, shape, texture, expressions, and movement of each face were parameterized using face AI models that create an interpretable, linear space representing faces. (Fig. 4.1E).

A defining aspect of the modeling approach is a jointly learned neuro-perceptual space (Fig. 4.2A) where each axis of the space corresponds to aspects of brain activity and sets of dynamic

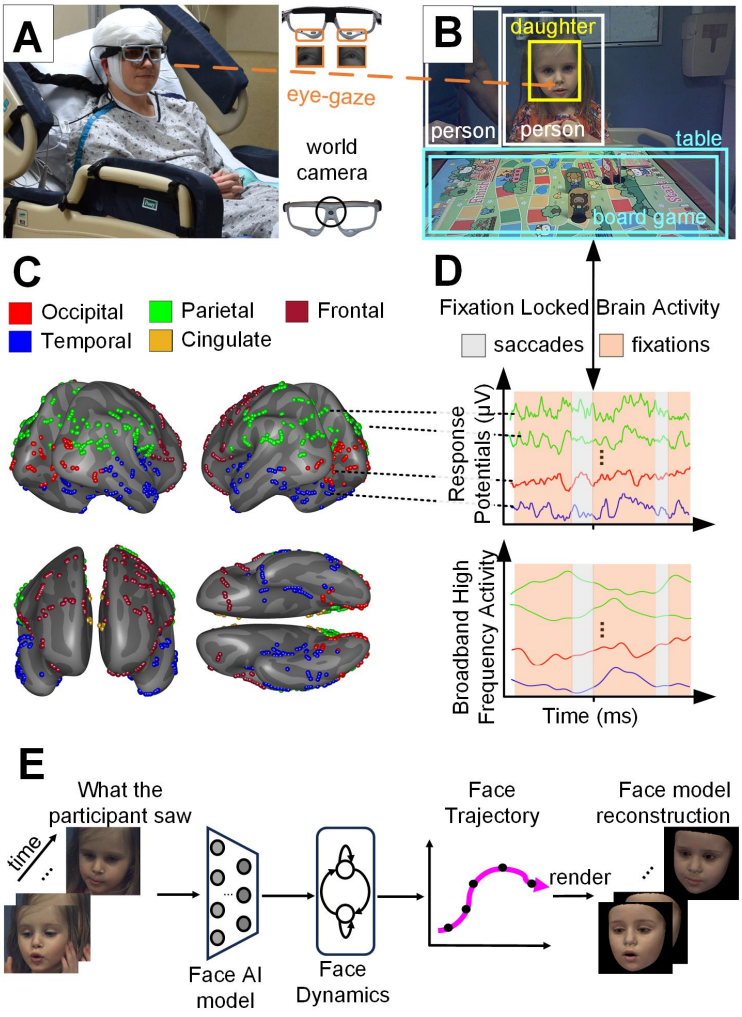

Figure 4.1: **Simultaneously recorded brain activity and unscripted natural interactions (A)** Participants implanted with intracranial electrodes wore eye-tracking glasses that captured their field of view with a world camera (black circle) and where they looked with inward facing eye-tracking cameras (orange rectangles) that tracked their eye-gaze (orange dashed line). **(B)** Eye-tracking was combined with computer vision annotations of world video to determine when participants looked at faces and who they looked at. In this example frame, the participant was looking (orange cross) at their daughter's face (see Movie 1 for a real time view). **(C)** Intracranial electrodes implanted for clinical treatment in 5 participants (681 electrodes total) captured brain activity from Temporal (195 electrodes), Parietal (244 electrodes), Occipital (90 electrodes), Cingulate (41 electrodes), and Frontal (207 electrodes) areas. **(D)** Intracranial EEG recordings were aligned in time (black arrow) with eye movement (saccades and fixations) and world video annotations in (B) and preprocessed to obtain Response Potentials and Broadband High Frequency Activity (BHA) for all 681 electrodes. **(E)** Real faces recorded in world video were parameterized so they could be reconstructed with high fidelity, using face AI models that estimated the pose, eye gaze, identity, expression, and texture of each face in a linear face model, and a linear dynamical system that tracked facial motion.

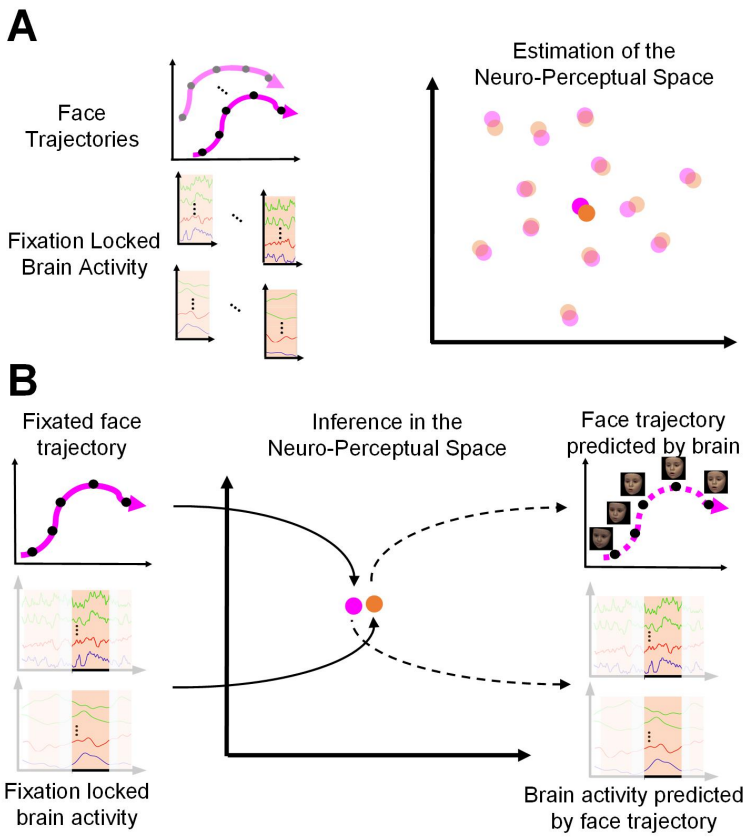

Figure 4.2: **Learning and Inference in a Neuro-Perceptual space: (A)** A computational model was trained to identify aspects of brain activity (orange dots) and sets of facial features (magenta dots) that were highly correlated. **(B)** Held out Fixation Locked brain activity was projected into the neuro-perceptual space (orange dot) and projected out to the face space to predict a video of fixated face. Reversing this process predicted brain activity from a face trajectory (magenta dot) in the neuro-perceptual space.

facial features that are strongly correlated with each other. Moving in this neuro-perceptual space corresponds to both a parametric change in brain activity and complementary parametric changes in the perceptual features that corresponded to that brain activity. Neural tuning curves are defined as the relationship between parametric changes in the percept and corresponding parametric changes in brain activity (158, 159); thus, the neuro-perceptual space is a tuning space. Learning a linear neuro-perceptual space provides straightforward interpretability of the geometry of the data manifold – if the data occupy a linear subspace, there is a simple linear relationship between parametric differences in the face and parametric difference in brain activity; if the data occupy a non-linear manifold, the geometry of this manifold can be probed to unravel aspects of faces that the brain is more or less sensitive to.

4.2 RESULTS

We first tested the robustness of this approach to model face perception during natural social interactions by reconstructing faces, including their motion and expression, from brain activity alone and vice versa. Specifically, to reconstruct a face participants viewed, its corresponding fixation locked brain activity was projected into the model's neuro-perceptual space. This neuro-perceptual representation was then projected out to a face space and the predicted face was visualized (Fig. 4.2B), then this process was reversed to create a movie of the predicted brain activity based on face information alone. We then examined the tuning geometry to unravel the neural representation of real-world facial expressions and face motion.

Across all participants, qualitatively accurate movies of the faces that participants viewed could be made using brain activity alone. For example, Fig. 4.3A and Movie 2 show the face of Participant #1's daughter while they played snakes and ladders. Fig. 4.3C and Movie 3 show additional reconstructed faces from other participants' interactions with friends and family as well as clinicians and researchers. To quantify reconstruction quality, we assessed pairwise classifica-

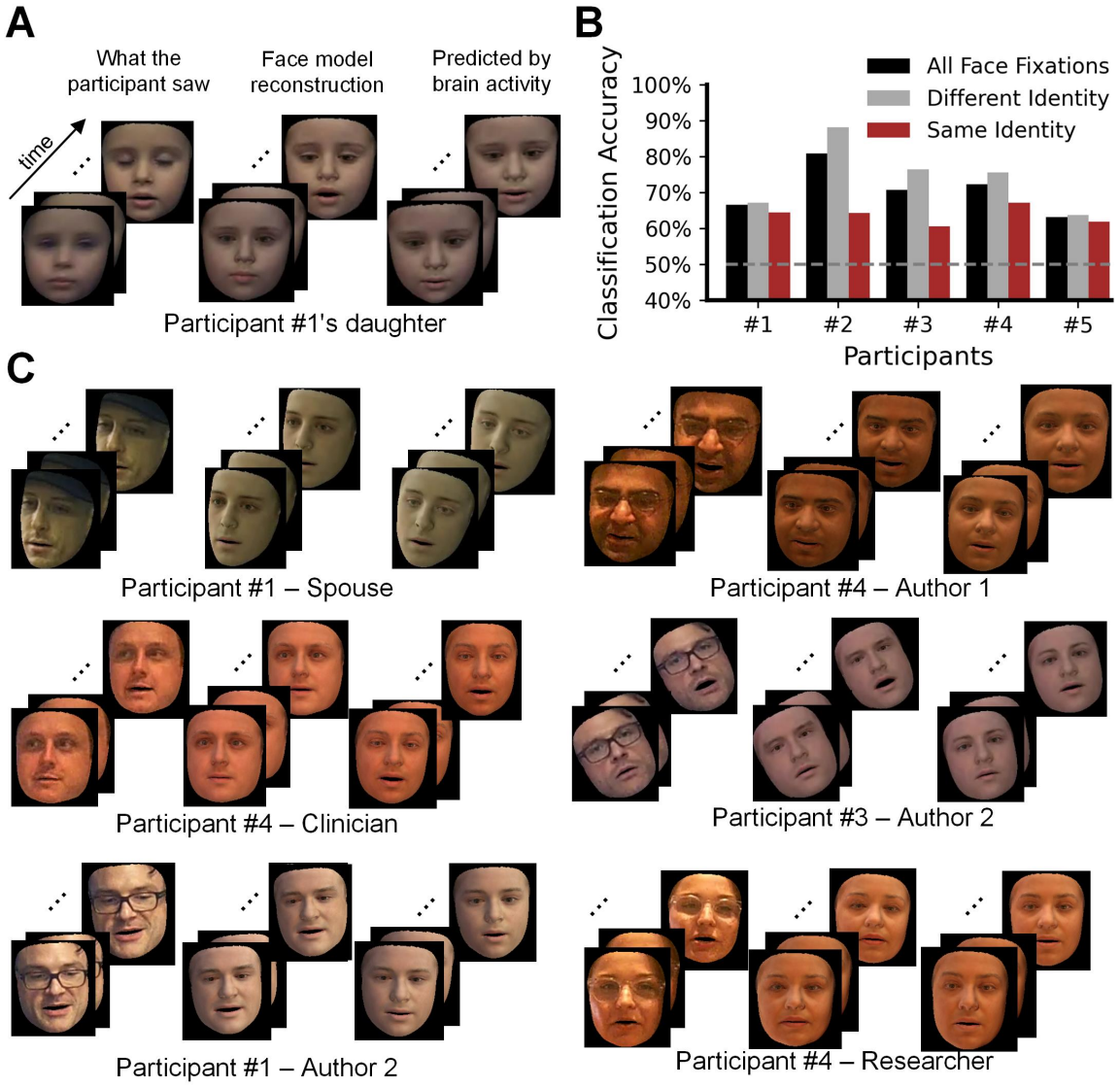

Figure 4.3: **(A) Face reconstruction:** An example face during a fixation (left), its face model representation (middle), and the face reconstructed using brain activity alone (right) for Participant #1's daughter while playing snakes and ladders (see Movie 2). **(B) Top level Statistics:** Pairwise classification accuracy (see Methods for details) was significantly ($p<0.05$ with permutation tests) above chance (50%) for all patients (black), not only between identities (gray), but also between instances of an individual's face (brown). **(C) Additional face reconstructions:** Faces of multiple individuals were reconstructed for each participant and faces of researchers and clinicians present in recording sessions for different participants could be reconstructed from each participant's brain activity. See Movie 3 for these and additional reconstructions from all participants. The original faces are not shown for some individuals who were either unreachable or declined consent to use their faces in print. In those cases, the face model representation for those individuals is rotated by a random matrix to obscure their identity.

tion accuracy by determining if the face reconstruction movie was more like the actual face from that fixation compared to other face fixations. Significant reconstruction accuracy (Fig. 4.3B) was observed in each participant; including for the case where the pairs of fixations being compared were restricted to different instances of the same individual's face (within identity reconstruction of dynamic facial expressions). The qualitative accuracy of individual reconstructions and quantitative statistics demonstrate that this paradigm and analytical framework is suitable for modeling the unconstrained variability of real world faces during free, natural social interactions. However, realizing the potential of this approach for neuroscientific discovery requires inverting this reconstruction, i.e., reconstructing brain activity from dynamic faces.

Fixation locked brain activity was reconstructed significantly across several cortical areas, including traditional face areas such as the fusiform in ventral temporal cortex. Notably, the most robust reconstructions of brain activity came from electrodes in areas around the temporal-parietal junction which correspond to the recently proposed third visual stream (156), posited to be a social-vision pathway. These observations were replicated in both hemispheres across different participants (Fig. 4.4A, Movie 4, Table in Fig. 4.6). The temporal dynamics of robustly reconstructed brain activity for electrodes in face areas included but were not limited to the N170 response. The reconstructed brain activity emphasized the putative social vision pathway as critical for face processing in the real world.

The above reconstructions (Fig. 4.3 and 4.4A) demonstrate both the robustness of the approach and the spatiotemporal patterns of brain activity that correspond to real world face perception. The neural code for dynamic facial expressions during real-world interactions can be further unraveled by examining aspects of the model itself. The reconstructions of faces from brain activity and vice versa were based on a neuro-perceptual space learned jointly from brain activity and faces. Moving along the axes of the neuro-perceptual space enables data-driven

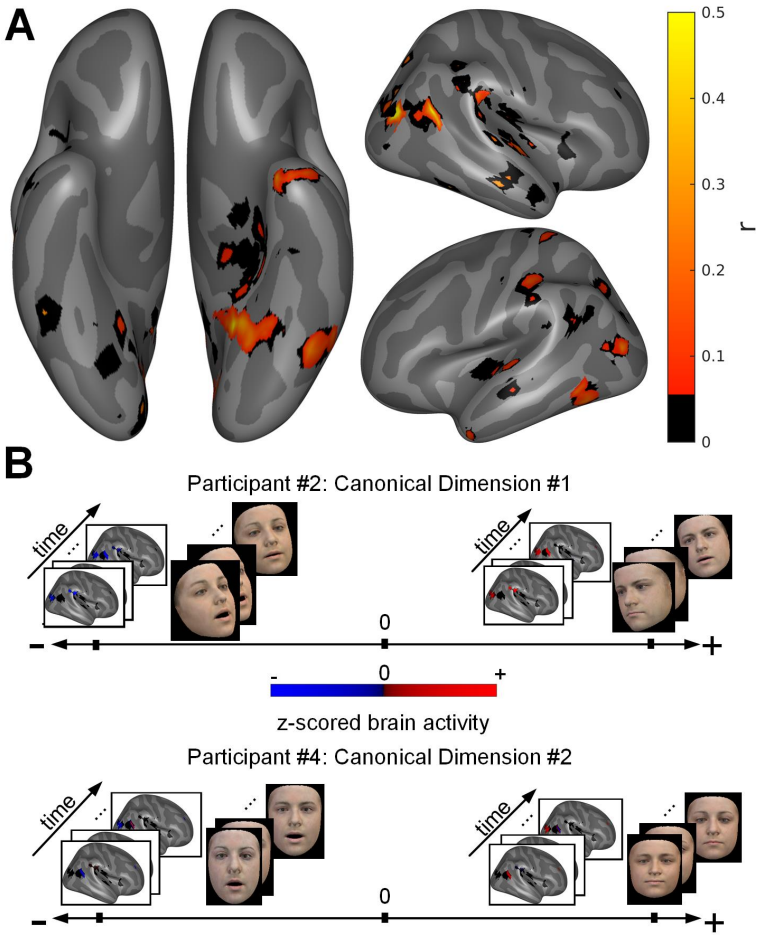

Figure 4.4: **(A) Reconstructed neurodynamics:** Significantly ($p<0.05$ with permutation tests) reconstructed Fixation Locked brain activity for all electrodes that model's make predictions for, visualized for all participants over the fixation duration (see Methods for details). Movie 4 depicts the underlying neurodynamics of this figure over all participants, and Movie 5 does the same separately for each participant. **(B) Neuro-Perceptual space:** Each axis specifies linear mappings between facial features and aspects of brain activity that are significantly correlated with each other. A step along any axis in this space changes both the predicted neurodynamics and the predicted face. The visualization depicts the predicted face trajectory and neurodynamics at the negative and positive ends of two dimensions for Participants #2 and #4. The canonical components visualized here are visibly sensitive to expression and motion (x-axis), as well as pose (y-axis). Movie 6 and Movie 7 visualize these dimensions.

discovery of how particular aspects of neural activity correspond to particular aspects of the perceptual input. For example, neuro-perceptual spaces across participants revealed dimensions sensitive to facial pose (Fig. 4.4B, Movie 6), expression and motion (Fig. 4.4B, Movie 7) and these example movies also show what aspects of the neural activity are different for parametric differences in the face. Mapping how parametric differences in stimuli such as faces correspond to parametric differences in neural activity is how we measure neural tuning and thus, neuro-perceptual spaces are tuning spaces, and the geometry of the data manifold embedded in these tuning spaces can be used to test hypotheses that reveal the neural code for faces.

We deployed these methods to investigate how our brains code for variations on someone's face during natural social interactions. We hypothesized that our brains code the facial expressions for a person as deviations from their resting face (which acts as a kind of "norm" expression for a person). To operationalize this, we first recentered the face space for each individual person participants saw by estimating the resting face and treating it as the origin of the space of a person's facial expressions. This recentering effectively removes face identity from the model, allowing us to examine how facial expressions are coded independent of whose face is making the expressions. Reconstruction accuracy remained significant even with identity removed. In two of our participants, there were sufficient fixations on multiple faces that we were able to show that we can train a model on one set of peoples' facial expressions and accurately predict the neural response for a different set of peoples' facial expressions and movements (cross-identity facial expression reconstruction; see *supplementary results*); i.e., given sufficient data, if we train a model only on "Mark's" facial expressions and movements, that model still accurately predicts "Sally's" facial expressions from brain activity and vice versa despite the fact that the model was not trained with any instances of seeing Sally's face. These results validated the robustness of recentering the face space to remove identity to test hypotheses about the coding of face information during real-world interactions.

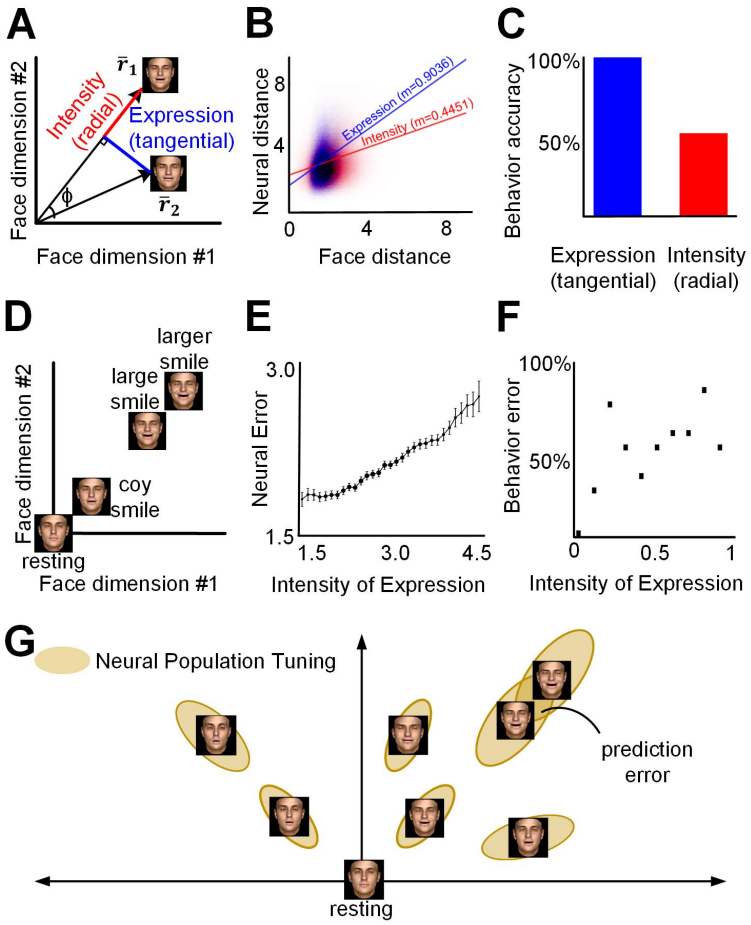

Figure 4.5: **Neural population tuning for facial expressions** **(A)** Hypothesis - our ability to tell facial expressions apart is greater for differences in expression (tangential distances) than differences in their intensity (radial distances). **(B)** Neural population tuning was sharper (steeper slope) for differences in expression relative to differences in the intensity of expressions (see Fig. S3 for all participants). **(C)** Behavior in a controlled face discrimination experiment was consistent with (B), showing greater accuracy for discrimination between expression type vs. intensity. **(D)** Hypothesis - our ability to distinguish expressions between faces close to each other is higher when the expression intensity (deviation from the resting face) is lower. **(E)** Neural sensitivity (error of neural prediction) for facial expression increased with the intensity of expressions (see Fig. S4 for all participants). **(F)** Behavioral responses were consistent with (E), showing decreasing sensitivity for facial expressions as expression intensity increased. **(G)** Neural population tuning for facial expressions of a person emerges as an oval shaped function (A - C), in which perceptual error i.e., tuning width grows based on distance from the resting facial expression of a person (D – F).

Next, we hypothesized that for the same sized visual changes, the brain would be more sensitive to differences in the kind of expression fixated upon than the intensity of the expression (Fig. 4.5A). In other words, we hypothesized that our brains need to be more sensitive to what kind of smile a person is making – a happy smile vs a sympathetic smile – compared to the intensity of the smile – a happy smile vs a slightly happier smile for example. The results supported this hypothesis for all subjects with sharper neural tuning for differences in expression (tangential distances) compared to differences in the intensity of expression (radial distances) (Fig. 4.5B). This finding was also validated behaviorally with a psychophysical experiment (Fig. 4.5C).

Finally, we hypothesized that the brain would be differentially sensitive to the same sized change in a face expression based on how close or far the expression was from the resting expression (Fig. 4.5D). In other words, we hypothesized that our brains should be more sensitive to the difference between a neutral expression and a coy smile compared to the difference between a large smile and a slightly larger smile, as an example. Indeed, the error in neural predictions increased (Fig. 4.5E) with the intensity of expressions (distance of fixated faces from the resting face) for all participants. This phenomenon was also validated behaviorally with a psychophysical experiment (Fig. 4.5F). The finding emerges as an analog (for facial expression) of Weber's law, which states that "the size of perceptible changes in stimulus intensity is proportional to the intensity of pre-existing stimulus" – another example of Weber's law is that we can easily tell the difference between a 1 and 2 pound weight but find it much harder to tell the difference between a 101 and 102 pound weight.

Taken together, these results demonstrate the neural code for facial expressions on a person's face during real-world interactions is defined by oval shaped tuning (Fig. 4.5G). The ovals are oriented toward the resting expression (norm) due to lesser sensitivity to differences along the

radial axis and greater sensitivity to differences tangential to this axis. The size of the ovals, i.e., the width of underlying tuning curves in the neural population, increase with distance from the resting expression.

4.3 DISCUSSION

In this study, we introduced an approach to model the uncontrolled variability of dynamic faces and brain activity by reconstructing them from each other during natural, unscripted interactions between participants and friends, family, researchers, etc. We used computer vision and a dynamical systems model to parameterize facial features, expressions, and motion, and a bidirectional model that reconstructed faces and brain activity from each other. We successfully reconstructed videos of faces being viewed during natural interactions based on neural activity alone, demonstrating the robustness of this approach. Reconstructed brain activity revealed the putative social-vision pathway (156) as important for face perception in real life, alongside traditional face areas in the ventral temporal cortex, and highlighted the importance of neural dynamics outside the N170 response. A central feature of the model was its jointly learned neuro-perceptual space which revealed the tuning of neural populations to facial features and enabled testing hypotheses about the neural code for dynamic facial expressions. We used this analytical framework to test hypotheses about a norm or resting expression centered code for facial expressions within a person's face. The results supported oval shaped tuning and revealed an analog of Weber's law for facial expressions in the process, which was subsequently tested and confirmed with a behavioral experiment that tested face perception.

The neural code for differences of facial expressions is relatively unknown compared to other aspects such as identity, where distinct coding schemes have been reported. Natural interactions like those recorded here offered a unique opportunity to address this gap in knowledge by probing natural variation in facial expressions. The results presented here supported the idea of a

norm-based neural code in which expressions on a person's face are coded as deviations from their resting facial expression (norm). Notably though, they do not rule out other possibilities such as a mixed axis-based and norm-based code because different hypotheses were not competed against each other. If anything, the results support a mixed code, because norm-based tuning was observed in a model whose basic structure was axis-based. An important aspect of neural population tuning that emerged was differences in neural sensitivity to changes in the type and intensity of expressions, and a decreasing sensitivity to facial expressions that were further away from the resting facial expression, which revealed an analog of the Weber's law for facial expressions. Taken together, these results portray neural tuning for people's facial expressions as oval shaped, where the ovals are pointed toward their resting facial expressions and get larger upon moving further away from the resting expression.

Real-world neuroscience has the potential to reveal novel observations that can be probed further in lab-based experiments and test how lab-based findings are implemented during natural behavior (160). Advances in technology enable recording natural behavior at high fidelity but using them to unravel brain-behavior relationships in the real world requires overcoming engineering and analytical challenges that are distinct for each aspect of cognition (26, 27, 28, 29). This study illustrates some of the key features for success in using uncontrolled real world recordings to understand the neural code are: appropriate behavioral events (fixations in this study), parameterization of stimuli or behavior being related to brain activity (projecting the faces into a parameterized face space), collection of large datasets that transform uncontrolled real-world variability from a challenge into an asset (hours of data), and statistical frameworks that robustly reveal the neural underpinnings of perception and behavior (the jointly learned neuro-perceptual space). These themes are also relevant for animal studies that are pushing the boundaries of brain recordings during natural behavior, driven by a rising interest in neuroethology (160, 161, 162, 163, 164, 165).

The jointly learned neuro-perceptual spaces in this study demonstrated robust reconstruction of dynamic facial expressions based on neural activity alone and vice versa (Fig. 3 and 4A). This jointly learned space facilitated both a data driven examination of neural tuning (Fig. 4B) as well as probing of the data manifold in these spaces, which allowed for the testing of specific hypotheses about the neural code (Fig. 5A, 5D). Examining the geometry of the data manifold generated novel hypotheses - the oval shaped tuning for facial expressions with larger ovals further away from the resting expression (Fig. 5G), that were tested using controlled experiments (Fig. 5C, 5F). Taken together, these neuro-perceptual spaces can reveal a picture of neural population tuning to natural stimuli in real-world settings, in much the same way studies in controlled settings have revealed tuning to aspects of vision (166), movement (167), navigation (168, 169, 170), etc. Understanding neural tuning for representations in real-world settings can not only provide ecologically valid substantiation of models developed in control experiments, but also generate new hypotheses that can be tested in controlled experiments. The modeling framework's capability to predict brain activity based on stimuli also has implications for brain-computer interfaces for vision restoration.

4.4 Materials and Methods

4.4.1 Participants

A total of five patients (three men, two women) underwent surgical placement of stereoelectroencephalography (intracranial EEG - iEEG) depth electrodes as standard of care for epileptogenic zone localization. The ages of the participants ranged from 22 to 64 years old (mean = 37 years, SD = 13.47 years). No ictal events were observed during experimental sessions.

4.4.2 INFORMED CONSENT

All participants provided written informed consent in accordance with the University of Pittsburgh Institutional Review Board. The informed consent protocols were developed in consultation with a bioethicist (Dr. Lisa Parker) and approved by the Institutional Review Board of the University of Pittsburgh. Audio and video of personal interactions were recorded during experimental sessions. Our protocol incorporated several measures to ensure privacy considerations and concerns could be addressed based on the preferences of individual participants. First, the timing of recording sessions was chosen based on clinical condition and participant preference, to ensure that they were comfortable with recording of their interactions with the visitors present (and/or expected to be present). Second, all visitors present in the room were notified about the nature of the experiment at the beginning of each recording session and given the opportunity to avoid participation. Third, a notification was posted at the entrance of the patient room informing any entrants that an experiment was being conducted where they might be recorded so that they could avoid entering if they chose to. Finally, at the end of each experimental recording, participants were polled to confirm their consent with the recording being used for research purposes and offered the option to have specific portions (e.g., a personal conversation) or the entire recording deleted if they wished. Thus, explicit "ongoing consent" was acquired through written informed consent at the beginning and end of each session; providing participants the opportunity both affirm their willingness to participate and to consider the content of the recordings before giving final consent. None of our participants thus far have asked to have recordings partially or fully deleted after the recording session was complete.

It is notable that there are no reasonable expectations of privacy other than for the patient, and this work was considered to meet the criteria for waiver of informed consent for everyone other than the participants themselves. Regardless, separate media releases were sought from individuals present in the video recordings to use their faces in publications. Some individuals

were either unreachable or declined consent to use their faces in print. In those cases, the original faces are not shown and the face model representation for those individuals is rotated by a random matrix to obscure their identity.

4.4.3 ELECTRODE LOCALIZATION

Electrodes were localized with Brainstorm software using high-resolution postoperative CT scans of participants that were co-registered with preoperative MRI images using FreeSurfer™.

4.4.4 DATA ACQUISITION

Multimodal behavioral data (egocentric video, and eye-tracking) as well as neural activity from 96-220 iEEG contacts were recorded simultaneously during unscripted free viewing sessions in which participants wore eye-tracking glasses while they interacted with friends and family visiting them, clinicians, and hospital staff responsible for their care, and members of the research team. The timing and duration of recording sessions were determined based on clinical condition, participant preference and to coincide with the presence of visitors in the hospital room, where possible.

Behavioral data were captured by fitting each participant with SensoMotoric Instrument's (SMI) ETG 2 Eye Tracking Glasses. An outward facing egocentric camera recorded video of the scene viewed by participants at a resolution of 1280 x 960 pixels at 24 frames per second. Two inward facing eye-tracking cameras recorded eye position at 60 Hz. SMI's iView ETG server application, running on a laptop received and stored streaming data for all modalities from the eye-tracking glasses by way of a USB2.0 wired connection. The iView ETG software also served as an interface for researchers to calibrate the eye-tracking glasses to each participant with a three-point calibration procedure that enabled the accurate mapping of eye-tracking data

to specific 'gaze' locations on video frames, and to initiate and stop the recording of behavioral data.

Electrophysiological activity (field potentials) was recorded from up to 220 iEEG electrodes at a sampling rate of 1 kHz using a Ripple Neuro's Grapevine Neural Interface Processor (NIP).

4.4.5 DATA SYNCHRONIZATION

A MATLAB® script, running on the same laptop as the SMI iView ETG Server software, broadcasted numbered triggers every 10 s, injecting them simultaneously into the neural data stream via a Measurement Computing USB-204 data acquisition (DAQ) device connected to the NIP's digital port and into the eye-tracking event stream via SMI's iView ETG server application via a sub-millisecond latency local loop back network connection using UDP packets. These triggers were used to align and fuse the heterogeneously sampled data streams after the experiment, during the Data Fusion stage.

In each recording session, neural activity recording was initiated, followed by simultaneous initiation of recording of eye-tracking and egocentric video via the SMI ETG 2 Eye Tracking Glasses using the SMI iView ETG Software Server. Once the recording of all modalities was underway, the MATLAB® script was initiated to generate and transmit triggers. At the end of each recording session, the tear down sequence followed the reverse order: 1) the MATLAB® script was terminated, marking the end of the recording, 2) the SMI iView ETG Software Server recording was halted, 3) the neural data recording stream was stopped on the NIP. Excess data from prior to the first numbered trigger and after the last numbered trigger were discarded for all modalities.

4.4.6 MINIMIZING EYE-TRACKING ERROR AND PARTICIPANT FATIGUE

Shift in the placement of the eye-tracking glasses is possible if the participant inadvertently touches or moves them during a recording session. Such disruption can introduce systematic error(s) in eye gaze data captured after the disruption(s), although errors can be mitigated with gaze correction (see Data Preprocessing for details). The potential for such an event increases with the duration of a recording session. To minimize the risk of such error(s), we first instructed participants to avoid touching or nudging the glasses during a recording session to avoid disrupting the eye-tracking calibration completed at the beginning of the recording session. Second, we strove to reduce such errors by limiting an individual recording session to 1 h and including a short break for participants. During this interlude, the recording was terminated, and participants are offered the opportunity to remove the eye tracking glasses before initiation of the next session. The interlude served two purposes: 1) it gave the participant a break from wearing the eye-tracking glasses, helping to alleviate fatigue and discomfort; 2) initiating a new recording allowed the research team to re-secure and re-calibrate the eye-tracking glasses, renewing the accurate mapping of gaze to the egocentric video. Although we prefer $\approx$ 1 h recordings as a best practice, maintaining this practice depended upon participants' preference and the number visitors. In some cases, recording sessions were longer.

4.4.7 ERGONOMIC MODIFICATIONS

Standard clinical care following iEEG implantation involves the application of a bulky gauze head dressing. This bandaging was applied around the head to protect the operative sites where the iEEG electrodes were secured with bolts. The dressing also included a chin wrap to provide further support in preventing dislodgement of the iEEG electrodes by securing the connector wires that carry electrical activity to clinical and/or research recording systems like the Ripple Neuro Grapevine NIP. The bandaging typically covered the participants' ears, rendering the temples on the eye-tracking glasses unusable. To overcome this challenge, we modified the structure

of the eye-tracking glasses, removing the temples and substituting them with an adjustable elastic band. We attached the elastic band to the frame of the eye-tracking glasses using Velcro patches sown at each end. The modification permitted secure placement of the glasses on the face of a participant, with the elastic band carefully stretched over the head dressing to avoid disturbing the operative sites. To reduce any pressure the eye-tracking glasses placed on the participants' faces as a result of the elastic band alteration, we further modified the glasses by adding strips of adhesive backed craft foam to the nose bridge and upper rims of the frame. These ergonomic solutions enabled correct, robust, and comfortable placement of eye-tracking glasses for each participant with flexibility to adjust to individual bandaging and electrode placement configurations. As an added measure to minimize the possibility of movement for eye-tracking glasses during recording sessions, the USB cable connecting the eye-tracking glasses to the laptop was secured to the participants' hospital gowns near the shoulder with a large safety pin to prevent the weight of the remaining length of cable from pulling on and displacing the glasses during a recording session. Sufficient slack was left in the cable segment between the glasses and the fixation point on the participants' gowns to allow for free head movement while preventing the secured cable segment from pulling on and potentially displacing the eye-tracking glasses.

4.4.8 BEHAVIORAL EXPERIMENT

A behavioral psychophysics experiment approved by the University of Pittsburgh's Institutional Review Board was conducted in a cohort of 8 participants (four men, four women) who were students and staff at the University of Pittsburgh. The ages of the participants ranged from 18 to 34 years old (mean = 26, SD = 5).

The behavioral paradigm required participants to determine whether two faces, shown one after the other, were the same or different. Each face was presented for 1 second. The interstimulus interval between faces was randomized, ranging from 500ms - 1100ms. Participants

were required to make a choice ("the faces were the same", "the faces were different") for each trial to advance through the experiment. A single run of the experiment comprised of 228 trials. The two faces presented in each trial featured one with a target expression at a specific intensity (the base face), and another face which was a radial or tangential perturbation to this base face. Radial perturbations could increase or decrease the intensity of the expression by a given step size in the face model space, but not change the expression itself. Tangential perturbations changed the expression (up or down an orthogonal direction) but did not affect the intensity of the original expression. The stimulus set consisted of 10 intensities for each expression i.e., there were 10 base faces for each expression which were presented alongside their 2 radial and 2 tangential perturbations. This added up to 40 trials for each expression i.e., 4 trials for each of the 10 expression intensity levels. Both radial and tangential perturbations were of the same step size in the face space. The stimulus set featured 6 different expressions of which 3 were canonical expressions (joy, fear, disgust), and 3 were randomly generated expressions. The maximum allowable expression intensity was the same across all expressions, and limited to ensure that presented faces were not aversive. All the faces presented in this paradigm were generated from a 3D morphable face model with a principal component space in which the canonical expressions (joy, disgust, fear, anger, surprise, sadness) were known. The paradigm featured sham trials in which the same face was shown twice, which accounted for 20% of the number of normal trials.

4.4.9 DATA PREPROCESSING

The physiological (neural) and behavioral (eye-tracking, video, audio) data streams captured during a real-world vision recording were preprocessed as follows before Data Fusion was initiated.

INTRACRANIAL RECORDINGS

Response potentials and broadband high frequency activity (BHA) were extracted from the raw iEEG recordings for statistical analysis using MATLAB. Response potentials were extracted using a fourth-order Butterworth bandpass ([0.2 Hz, 115 Hz]) filter to remove slow linear drift and high-frequency noise, followed by line noise removal using a fourth-order Butterworth band stop ([55 Hz, 65 Hz]) filter. BHA extraction involved two steps. First, the raw signal was filtered using a fourth-order Butterworth bandpass ([1 Hz, 200 Hz]) filter followed by line noise removal using notch filters at 60, 120, and 180 Hz to obtain local field potentials. Next, power spectrum density (PSD) between 70 and 150 Hz was calculated for the local field potentials with a bin size of 2 Hz and a time-step size of 10 ms using Hann tapers. For each electrode, the average PSD across the entire recording was used to estimate a baseline mean and variance of the PSD for each frequency bin. The PSD was then z-scored using these baseline measurements for each frequency bin at each electrode. Finally, BHA is estimated by averaging the z-scored PSD across all frequency bins (excluding the line noise frequency bin at 120 Hz). iEEG recordings were subjected to several criteria for inclusion in the study. Any recordings with ictal (seizure) events were not included in the study. Artifact rejection heuristics were implemented to avoid potential distortion of statistical analyses due to active interictal (between seizure) or outliers. Specifically, we evaluated the filtered iEEG data against three criteria that are applied to each sample i.e., each time point in iEEG recordings, which corresponds to 1 ms of neural activity. These criteria were applied to the filtered iEEG signal for each electrode, as well as the averaged (across all electrodes) iEEG signal. The first criterion labels a sample as 'bad' if it exceeds 350 μV in amplitude. The second criterion labels a sample as bad if the maximum amplitude exceeds 5 standard deviations above/below the mean. The third criterion labels a sample as bad if consecutive samples (1 ms apart at a 1000 Hz sampling rate) change by 25 μV or more. For the averaged iEEG signal, any sample satisfying any of these three rejection criteria is labeled as bad. Further, if more than ten electrode contacts (out of a typical 128) satisfy the bad sample

criterion for a particular sample, it is labeled as a bad sample. Less than 10% of the samples in experimental recordings were labeled as bad samples. All data types were dropped from analysis for fixations that contained bad samples.

EYE-TRACKING

The eye-tracking data stream is composed of time series data sampled at 60 Hz, where each sample (referred to as an eye tracking trace) contains a recording timestamp, an eye gaze location (X,Y coordinates in the space of egocentric video) and is labeled by the SMI iView ETG platform as belonging to a fixation, a saccade or a blink. Consecutive eye-tracking traces with the same label (fixation, saccade, or blink) are interpreted as belonging to a single eye-tracking 'event' of that type, whose duration is the difference in recording timestamps of the last and first eye-tracking traces in the block of consecutive traces with the same label (fixation, saccade, or blink). As an example, a set of 60 eye-tracking traces (amounting to 1 s of recorded activity), where the first 30 are labeled as fixation, the next 12 labeled as saccade, followed by the final 18 labeled as fixation, would be interpreted as a fixation event $\approx$ 500 ms long (30 samples at 60 Hz), followed by a saccade event $\approx$ 200 ms long (12 samples at 60 Hz) followed by a fixation event $\approx$ 300 ms (18 samples at 60 Hz). We developed custom Python scripts that parse eye-tracking traces and construct logs of eye-tracking events for each recording session. In addition to the duration of each eye-tracking event, the median gaze location (median is used for robustness to outliers) was logged for each fixation event and the start/ end gaze locations were captured for each saccade event. Blink traces are denoted by a loss of eye-tracking (i.e., absence of gaze location) and as a result only the duration of blink events was tracked in the consolidated eye-tracking event logs. Preprocessing of eye-tracking data also incorporates the detection and correction of systematic errors in gaze angle estimation that can be induced by the movement of eye-tracking glasses during recording sessions (e.g., if a participant inadvertently touches and moves the glasses due to fatigue), which disrupts the calibration of eye-tracking glasses (see

Data Acquisition for details). Such issues were detected by manually viewing all experimental recordings using SMI's BeGaze application, which renders eye-gaze, audio, and egocentric video together. The disruption of calibration for eye gaze tracking is visually detectable when viewing egocentric video overlaid with eye-tracking and audio because visual behavior is altered such that the gaze data fails to make sense consistently after loss of eye-gaze calibration (e.g., the subject is scrolling through a phone or reading a book or watching tv or talking to someone, but the gaze location is visibly shifted away from the obvious target). These issues were corrected using the SMI BeGaze application, which allows researchers to apply a manual correction (i.e., an offset) to eye gaze at any time point in a recording, which applies to all eye gaze data following the corrected time point. The corrections were verified by reviewing the video that followed the correction, to ensure that corrected eye gaze data made sense consistently. Corrections to eye-tracking data preceded preprocessing in such cases.

FACE DETECTION

Egocentric (head-centered) video recordings included a range of visual stimuli present in the room, including objects, people, and faces. We processed egocentric video recordings to detect faces, the primary object of interest in this study. Deep Learning based computer vision models (171, 172) developed for large-scale face detection and recognition applications were used to detect faces present in each video frame of egocentric video recordings. Manual review of egocentric videos with bounding boxes that annotated identifying detected faces showed model performance was robust to the variability of conditions present in the egocentric video recordings. Failure to detect a face was extremely rare, usually involving heavy occlusion, extremely poor lighting, or both. Cumulatively 761,510 faces were detected and labeled (see *Face Identification below*) across 1,136,208 frames of video corresponding to nearly 11 hours of recordings, which required $\approx$40 hours on a single NVIDIA 1080TI GPU. In addition to face detection, these models also generate a 512-dimensional embedding in a face space that can be used to train classifiers to

identify different individuals present in the video recordings.

FACE IDENTIFICATION

A neural network was trained to perform identity classification on faces detected across all video recordings for each patient. The network architecture featured two densely connected layers (128 ReLU units each) that were subjected to a 50% dropout rate to avoid overfitting. 512-dimensional embeddings generated by face detection models were used to predict identity. To prepare data for model training, identity labels were assigned manually to all faces present in a subset of video frames which corresponded to the beginning of each fixation. The annotation typically required 2 hours of effort to annotate a 1 hour video recording. Depending upon what participants did during a recording session, faces could also be detected on television screens, mobile phone screens, magazines, and even arise from false positives ($<$1%) none of which were in scope for this study. These extraneous faces arose in training data and were given a catch all identity label ("other") that separated them reliably from the faces of real people in the room. Models were trained using 5-fold cross-validation and high accuracy ($<$0.1% misclassification) and class balanced accuracy (since people were present in the recording for different amounts of time) were observed on held out data.

The trained identity classification networks for each patient were then used to label all the faces detected in each video frame of all their recording sessions. A final manual review of the fully annotated video was performed to ensure undesirable and unforeseen issues did not arise e.g., mislabeling sparsely present individuals as "other".

FACE PARAMETERIZATION

Each face detected in egocentric video recordings is represented in a linear face model that represents a 3D structure for them. Recent advances have enabled robust estimation and high-

fidelity reconstruction of 3D faces from monocular images that capture a 2D view of the original face. Face AI models that perform such reconstructions estimate the pose, shape, texture, expression of faces present in 2D images while accounting for extraneous factors such as camera position and lighting parameters, in a way that aligns important facial landmarks, facial appearance, and minimizes pixel level loss for the reconstructed face compared against the original 2D face image.

Here, we parameterized faces in each frame of egocentric video recordings to obtain the pose ($\theta \in \mathbb{R}^3$) as well as their shape ($s \in \mathbb{R}^{80}$), texture ($t \in \mathbb{R}^{80}$), and expression ($e \in \mathbb{R}^{64}$) in a generative 3D face model (173, 174) using Deep 3D Face (175). Separately, we obtained estimates for eye-gaze ($g \in \mathbb{R}^2$) for faces in each frame of egocentric video recordings using a state of the art neural network (176). Combining these results in a 229-dimensional representation for each face in each video frame of the egocentric video recordings.

FACE DYNAMICS MODEL

A state space model was trained to identify a low dimensional latent space where trajectories representing the dynamics of parameterized faces could be embedded and recovered to reconstruct the original 229-dimensional representation reliably. Pose and eye-gaze were low dimensional variables whose dynamics were tracked separately i.e., they are not embedded in the state space. Thus, the state space model represented $\mathbb{R}^{224}$ dimensional inputs spanning shape, texture, and expression in a latent space ($x \in \mathbb{R}^{30}$). The model structure described below follows smooth linear dynamics (A), linear coordinate transformations into the latent space (B), and a linear read out (C) of the latent variables back into to the original $\mathbb{R}^{224}$ face representation.

$$x_{t+1} = Ax_t + B \begin{bmatrix} s \\ t \\ e \end{bmatrix} + \delta_t$$

$$\begin{bmatrix} \hat{s} \\ \hat{t} \\ \hat{e} \end{bmatrix} = Cx \tag{4.1}$$

For each participant, these models were trained and validated on face trajectories from un-fixated faces that were not used in analysis. Validation also included a qualitative component, where researchers reviewed face videos reconstructed from latent representations for held out data, alongside the original faces which were not embedded in the model. The models were eventually used to generate trajectory embeddings for fixated faces that were used in analysis against brain activity.

RESTING FACE ESTIMATION

The resting facial expression for each individual present in egocentric video recordings was estimated using parameterized representations of un-fixated faces that were not used in analysis. The resting facial expressions were then subtracted from all fixated faces that were used in analysis against brain activity.

The first step in estimating resting facial expression was to regress out the effects of pose on face parameters. This was done by training a multiple regression model to predict the value of face shape, texture and expression parameters based on pose. Removing values predicted by the regression model provided a pose corrected parameterization for each fixated face. Subsequently, the average pose corrected shape, texture, and expression for each person in the recordings were

computed as the resting facial appearance and expression.

4.4.10 DATA ANALYSIS

Precise alignment of the heterogeneous behavioral (eye-tracking), environmental (egocentric video) and physiological (neural) data streams is essential for robust analysis, and this is achieved by using eye-tracking as a reference modality against which video and intracranial recordings are aligned in time as described in (30). All analysis in this study is anchored to behavioral events (fixations) and each fixation is mapped to corresponding egocentric video frames corresponding to it as well as brain activity (FLP and FLBHA).

DATA PREPARATION

Fixations are determined to be on a face if the eye-gaze at the beginning of a fixation is on a person's face in the corresponding egocentric video frame. This is operationalized by determining if eye-gaze coordinates fall within a face bounding box identified by computer vision models (171, 172). Face fixations are filtered to ensure that they are 300ms or longer. Next, they are filtered to ensure that no fixation contains brain activity that has been characterized as having bad samples. Next, face parameters for the fixated person are retrieved for all the egocentric video frames corresponding to the face fixation, and fixations where this is not possible for any reason filtered out (e.g., because the face was not detected due to being occluded by obstacles like another person crossing them).

The fixations that satisfy these data quality criteria are then assembled into a dataset $(\mathbf{X}, \mathbf{Y})$, where $\mathbf{X} \in \mathbb{R}^p$ represents face trajectories in the face dynamics model and $\mathbf{Y} \in \mathbb{R}^q$ represents fixation locked brain activity (FRP and FRBHA) for all intracranial electrodes implanted in a participants brain. The dimensionality p of the face trajectories corresponds to collecting the pose ($\theta \in \mathbb{R}^3$), eye gaze ($g \in \mathbb{R}^2$), and latent face trajectories $x \in \mathbb{R}^{30}$ for 7 frames (corresponding

to ≈300 ms), which results in $p = 7 \times (3 + 2 + 30) = 245$ for basic face reconstruction and $p = 7 \times (2 + 30) = 224$ for the dataset to study facial expressions as deviations from the resting face. The dimensionality q of brain activity depends upon the number of electrodes (E) which differ for each patient-participant. For 300 ms, Fixation Response Potentials (FRPs) sampled at 1 KHz correspond to 300 dimensions for each electrode and Fixation Broadband High Frequency Activity (FRBHA) sampled at 100 Hz corresponds to 30 dimensions for each electrode, which results in $q = E \times (300 + 30)$. Participants in this study were implanted with anywhere between 96 to 220 electrodes which corresponds to $q \in \mathbb{R}^{31680-72600}$. The number of fixations ranges from $N \approx 10^2 - 10^3$ across different participants which means that $N << p, q$.

COMPUTATIONAL MODEL

Canonical correlation analysis (177) seeks to model the covariability between two multivariate datasets ($X \in \mathbb{R}^p$, $Y \in \mathbb{R}^q$) as a small number of strongly correlated latent variables (Canonical Components), to understand the relationship between them. It may also be described as a type of latent multi-view representational learning when viewed through a contemporary machine learning lens. In low dimensional data rich settings where $N > p, q$, CCA can be implemented using a Singular Value Decomposition (SVD) on $\Sigma_{YY}^{-\frac{1}{2}} \Sigma_{YX} \Sigma_{XX}^{-\frac{1}{2}}$ (177). However, this approach does not scale to high dimensions where $N << p, q$ due to challenges with inverting Σ_{XX}, Σ_{YY}. Different approaches have been proposed over the years (178, 179, 180, 181) primarily for applications in gene analysis, including those anchored around the idea of sparse canonical vectors and recent developments (181, 182) in this direction make fewer simplifying assumptions than earlier approaches (180).

Here, Sparse CCA is implemented by an iterative penalized least squares algorithm (182) which uses regularized regressions in an alternating manner to estimate canonical vectors for (X and Y), one canonical component at a time. Given a centered dataset $\mathbf{X} \in \mathbb{R}^{N \times p}, \mathbf{Y} \in \mathbb{R}^{N \times q}$

with sample covariances $\hat{\Sigma}_{XX} = \frac{1}{N}X^TX$, $\hat{\Sigma}_{YY} = \frac{1}{N}Y^TY$, $\hat{\Sigma}_{YX} = \frac{1}{N}Y^TX$, where the first $k-1$ pairs of canonical vectors $(\hat{w}^l_{brain}, \hat{w}^l_{face})\ \forall\ l \in (1, k-1)$ have been estimated, the k^{th} canonical vectors $(\hat{w}^k_{brain}, \hat{w}^k_{face})$ are estimated by solving

$$
\begin{aligned}
(\hat{w}^k_{brain}, \hat{w}^k_{face}) = \underset{w^k_{brain}, w^k_{face}}{\operatorname{argmin}} \ & \frac{1}{2N}\sum_{i=1}^{N}\left(Y_i^T w^k_{brain} - X_i^T w^k_{face}\right)^2 + \\
& w^k_{brain}\left(\sum_{l<k}\hat{\rho}_l \hat{\Sigma}_{YY}\hat{w}^l_{brain}\hat{w}^l_{face}\hat{\Sigma}_{XX}\right) + \\
& P_Y(w^k_{brain}) + P_X(w^k_{face}) \\
& s.t.\ w^k_{brain}\hat{\Sigma}_{YY}w^k_{brain} = 1,\ w^k_{face}\hat{\Sigma}_{XX}w^k_{face} = 1
\end{aligned} \tag{4.2}
$$

where $P_Y(w^k_{brain})$ and $P_X(w^k_{face})$ are regularization functions that may reflect the type of penalization in effect (e.g., group lasso, trend filtering). Here, we choose elastic penalties (Equation 4.3 that combine sparse feature selection with a smooth distribution of weights over the selected features. It is notable that although the optimization problem is nonlinear in nature, the model structure itself is linear.

$$
P(w, \lambda, \alpha) = \lambda\left(\alpha||w||_1 + \frac{(1-\alpha)}{2}||w||_2^2\right) \tag{4.3}
$$

The optimization problem in Equation 4.2 is solved using Algorithm 8 outlined in (182)

Algorithm 8 Iterative Penalized Least Squares

Initialize $(\hat{w}^k_{brain}, \hat{w}^k_{face})$ as first singular vectors of $\hat{\Sigma}_{YX} - \sum_{l=1}^{k-1} \hat{\rho}_l \hat{w}^l_{brain} \hat{w}^{l\,T}_{face}$

Initialize $R_k - 1 = diag(\hat{w}^1_{brain} \hat{\Sigma}_{YX} \hat{w}^1_{face}, ..., \hat{w}^{k-1}_{brain} \hat{\Sigma}_{YX} \hat{w}^{k-1}_{face})$

Initialize $\Omega_1 = I_n$, $\Omega_k = I_n - Y W^{k-1}_{brain} R_{k-1} W^{k-1}_{face} X^T / n$, where I_n is a $n \times n$ identity matrix.

while $(\hat{w}^k_{brain}, \hat{w}^k_{face})$ not converged **do**

  Set $\tilde{Y}_k = \Omega_k^T Y \hat{w}^k_{brain}$

  Compute $\breve{w}^k_{face} = \underset{w^k_{face}}{\operatorname{argmin}} \frac{1}{2n} ||\tilde{Y}_k - X w^k_{face}||_2^2 + P_X(w^k_{face}, \alpha_{face}, \lambda_{face})$

  $\hat{w}^k_{face} = \left[\breve{w}^{k\,T}_{face} \hat{\Sigma}_{XX} \breve{w}^k_{face} \right]^{-\frac{1}{2}} \times \breve{w}^k_{face}$

  Set $\tilde{X}_k = \Omega_k^T X \hat{w}^k_{face}$

  Compute $\breve{w}^k_{brain} = \underset{w^k_{brain}}{\operatorname{argmin}} \frac{1}{2n} ||\tilde{X}_k - Y w^k_{brain}||_2^2 + P_Y(w^k_{brain}, \alpha_{brain}, \lambda_{brain})$

  $\hat{w}^k_{brain} = \left[\breve{w}^{k\,T}_{brain} \hat{\Sigma}_{YY} \breve{w}^k_{brain} \right]^{-\frac{1}{2}} \times \breve{w}^k_{brain}$

Output $\hat{w}^k_{brain}$, $\hat{w}^k_{face}$ upon convergence.

TRAINING

Training data were demeaned and scaled to unit variance prior to model training. Models were trained with 5-fold cross-validation, which allowed each sample (fixation) to be in the test set once. The use of an elastic penalty function required choosing two parameters for each canonical component for each of the 5 models trained in this way. The first parameter was the regularization penalty (λ_{brain}, λ_{face}) and the second parameter was the elastic penalty (α_{brain},α_{face}). Both were identified during an additional 5-fold cross-validation procedure within the training data i.e., an additional inner cross-validation loop. An distinct relationship between α and λ (for brain activity and facial features) emerged during hyperparameter selection, where the amount of L1 penalty they collectively enforced ($\alpha \times \lambda$) remained identical i.e., increasing α led to a lower

optimal λ and decreasing α led to a higher optimal λ such that their product i.e., the L1 penalty, remained nearly constant. Such perturbations did not affect model performance in terms of statistics used to quantify model performance (see Inference section below for details) or in terms of the scientific conclusions drawn from examining model structure and the geometry of the data manifold in the neuro-perceptual space. These observations were therefore interpreted as a property of the data rather than the algorithm, and used to optimize the model training procedure by choosing a low value of $\alpha = 0.1$ (which was fixed) and λ was the only hyper parameter being optimized. The choice of α ensured the L2 penalty term, weighed as $\frac{1}{2}(1-\alpha)$ provided greater smoothing across the aspects of brain activity and facial features that were selected by the model.

In terms of model selection, canonical components which exhibited statistically significant correlation during the inner cross-validation loop were preserved and those that did not had their weights zeroed out. Models estimated up to 20 canonical components during training but the number of canonical vector pairs that survived cross-validation did not exceed 10 in any case. This approach of estimating a larger model ensured that no useful relationships were missed. The canonical space is also referred to as the neuro-perceptual space since it is jointly learned from brain activity and face trajectories.

INFERENCE

Inferences were drawn in three ways to assess model performance quantitatively and qualitatively. Brain activity (Y_i) and face trajectories (X_i) for held out fixations were first centered according to the mean and variances estimated from training data, and then projected into the neuro-perceptual space per Eq. 4.4.

$$\begin{aligned} Y_i^{CC} &= Y_i \times W_{brain}^K \\ X_i^{CC} &= X_i \times W_{face}^K \end{aligned} \tag{4.4}$$

Neuro-perceptual representations of brain activity were used to predict face trajectories, and vice versa as described by Eq. 4.5.

$$\hat{X}_i = Y_i^{CC} \times W_{face}^{\dagger\, K}$$
$$\hat{Y}_i = X_i^{CC} \times W_{brain}^{\dagger\, K} \tag{4.5}$$

where † represents the pseudoinverse of a sparse low rank projection matrix.

First, top level statistics were computed to assess model performance. Specifically, pairwise classification accuracy was computed for each held out fixation (i) by comparing the distance between the neuro-perceptual representation of its face trajectory (X_i^{CC}) and brain activity (Y_i^{CC}) against distances with all other held out fixations ($j \neq i$). If the former distance was smaller, the comparison was counted as accurate classification. The distances were weighted by the singular values (D) obtained during Step #1 of Algorithm 8. The average across all comparisons ($j \neq i$) was defined as the classification accuracy of each fixation. This metric was calculated separately for each held out fixation and then averaged across all fixations as described by Eq. 4.6.

$$\text{Accuracy} = \frac{1}{N}\sum_{i}^{N} \frac{1}{N-1} \sum_{\substack{j \\ j\neq i}}^{N} \mathbb{I}_{||D(X_i^{CC} - Y_i^{CC})||_2 < ||D(X_i^{CC} - Y_j^{CC})||_2} \tag{4.6}$$

The classification accuracy can be estimated in by comparing the neuro-perceptual face representation of a fixation (X_i^{CC}) with the neuro-perceptual brain activity representation of all other fixations (Y_j^{CC}) as described in Eq. 4.6 i.e., $||X_i^{CC} - Y_j^{CC}||_2$ or by comparing the neuro-perceptual brain representation (Y_i^{CC}) of a fixation with the neuro-perceptual face representation of all other fixations (X_j^{CC}), which changes Equation 4.6 to use $\mathbb{I}_{||D(X_i^{CC} - Y_i^{CC})||_2 < ||D(X_j^{CC} - Y_i^{CC})||_2}$. In practice, either of these variations resulted in similar top level statistics. Statistical significance thresholds for these statistics were estimated using permutation tests (see below for details).

Second, predicted brain activity was correlated with original neural activity (z-scored) for all electrodes across all time points for FRPs and FRBHA to determine how well the model predicted neurodynamics. Statistical significance thresholds for the correlations were estimated using permutation tests (see below for details).

Lastly, face trajectories predicted by brain activity were visualized as face videos alongside the original face and its face AI representation to qualitatively validate how neurally predicted faces looked compared to the original as shown in Fig. 4.3. Such face visualizations also helped visualize neural tuning by showing how movement along different dimensions of the neuro-perceptual space affected face appearance, dynamics, and predicted brain activity shown in Fig. 4.4B.

PERMUTATION TESTS

Permutation tests were implemented to estimate the statistical significance thresholds for top level statistics (pairwise classification accuracies) and reconstructed brain activity. Permutation tests with 1000 permutations were conducted separately for each participant. Three different flavors of permutation tests were implemented and results were consistent across them. In the first type of permutation test, the pairing between fixation locked brain activity and facial features was broken i.e., brain activity associated with a face fixation was permuted and assigned to a different face fixation. In the second type of permutation tests, fixation locked brain activity fixations that were not on faces at all was paired with facial features. In the third type of permutation test, brain activity was sampled randomly, breaking its anchoring to fixations, and facial features were randomly selected (from faces that were not fixated upon).

A Statistical significance threshold ($p<0.05$) for pairwise classification accuracies was calculated from a null distribution of those statistics estimated from 1000 permutations. Similarly, sta-

tistical significance thresholds ($p<0.05$) for correlations between actual and reconstructed brain activity were determined from a null distribution of those correlations estimated from 1000 permutations.

Finally, face predictions from the second flavor or permutation tests were also rendered to qualitatively assess what happens to the predicted faces when spurious brain activity is injected into the model. Visualization of faces predicted by permutation tests appeared close to the origin of the model with expressions and motions that resembled noise around the origin i.e., the mean face of the model. Geometrically, this suggested that non-face fixation locked brain activity disappeared into the null space of the model. Being a face model means its unavoidable that these models produce "a" face, but the predicted faces lack discriminable identity, expressions, and dynamics.

PROBING POPULATION TUNING IN THE NEURO-PERCEPTUAL LATENT SPACE

A step in the model's neuro-perceptual space changes the predicted face trajectory and the predicted pattern of brain activity. The step sizes of these changes are linearly dependent on the step size in the neuro-perceptual latent space. This coupling enables studying the tuning of neural populations by testing hypotheses and by exploring how the data manifold of brain activity and stimuli (face trajectories) relate.

A norm-based coding hypothesis for facial expressions was tested predicated on the assumption that facial expressions are coded as deviations from the resting face of each individual. If true, differences in neural tuning for the intensity of an expression (the radial aspect) compared to neural tuning for the type of expression (tangential aspect) would be observed. If the null hypothesis (no norm-based tuning) held, neural tuning would be the same for both. This hypothesis was tested in the neuro-perceptual latent space by computing radial and tangential distances be-

tween fixation pairs on the face trajectory manifold, validating they correlated significantly with pairwise distances on the neural manifold, and comparing fits (slopes) of radial and tangential distances to neural distances. Fixation pairs with $\omega > 90^\circ$ were ignored to ensure each fixation pair was only considered once. To be conservative, only fixation pairs on the same person and within a $\varphi < 45^\circ$ cone of each other were included in this analysis. The radial and tangential distances between fixation pairs on the face trajectory manifold were calculated as follows

$$\omega = \cos^{-1} \frac{(\bar{r}_1 - \bar{r}_2) \circ \bar{r}_1}{||\bar{r}_1 - \bar{r}_2||_2 \times ||\bar{r}_1||_2} \tag{4.7}$$

$$\text{radial} = ||\bar{r}_1 - \bar{r}_2|| \times \cos(\omega) \tag{4.8}$$

$$\text{tangential} = ||\bar{r}_1 - \bar{r}_2|| \times \sin(\omega) \tag{4.9}$$

Separately, the relationship between the data manifolds for brain activity and stimuli in the neuro-perceptual latent space. Specifically, the distance between the face and neural representation for each fixation was related to the intensity of expression (size of deviation away from the norm face) using Euclidean distances.

ANALYSIS OF BEHAVIORAL DATA

Participant responses were tallied against ground truth to determine response accuracy. Since there were no sham trials i.e., the faces were always different, this amounted to counting the number of trials where the participant response was "the faces were different" and dividing them by the total number of trials in consideration. Before being subject to this basic arithmetic, the trials were partitioned in two ways to compare the accuracy of behavioral responses. The first partition was between trials where the two faces were radially separated vs those that were tangentially separated. The second partition only featured trials where the faces were radially separated and tracked the accuracy of behavioral responses in those trials as a function of the expression intensity i.e., distance from the norm/origin.

4.5 SUPPLEMENTARY RESULTS

4.5.1 CROSS IDENTITY FACIAL EXPRESSION RECONSTRUCTION

Recordings from two participants featured sufficient fixations (>200) on an individual to attempt to decode their facial expressions using a model trained on the facial expressions and movements of other individuals. Data from the first participant's recordings featured enough fixations on two individuals, both of which exhibited significant top level statistics (56% with 463 samples and 58% with 687 samples respectively; $p<0.05$). Data from the second participant's recordings featured enough fixations on four individuals, three of which exhibited significant top level statistics (56% with 279 samples, 54% with 599 samples, 57.7% with 251 samples; $p<0.05$) while the fourth did not (51.9% with 855 samples).

4.5.2 CORTICAL DISTRIBUTION OF SIGNIFICANTLY RECONSTRUCTED ELECTRODES ACROSS PARTICIPANTS

Desikan Killiany Anatomical Area	Participant #1 # Significant Electrodes	Participant #1 # Total Electrodes	Participant #2 # Significant Electrodes	Participant #2 # Total Electrodes	Participant #3 # Significant Electrodes	Participant #3 # Total Electrodes	Participant #4 # Significant Electrodes	Participant #4 # Total Electrodes	Participant #5 # Significant Electrodes	Participant #5 # Total Electrodes	All Participants # Significant Electrodes	All Participants # Total Electrodes	# Participants with Significant Electrodes	# of Total Participants with Electrodes Present in Area
# SAMPLES (TOTAL)		1300		307		2283		618		207		N.A.		N.A.
Amygdala	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Bank of STS	1	1	0	0	0	0	0	0	0	0	1	1	1	1
Caudal Middle Frontal	0	0	0	0	0	0	14	35	0	0	14	35	1	1
Caudal Anterior Cingulate	1	3	0	0	0	0	2	3	0	0	3	6	2	2
Cuneus	3	3	7	7	2	2	0	0	3	3	15	15	4	4
Entorhinal	0	0	0	0	3	3	0	1	0	0	3	4	1	2
Fusiform	2	4	1	5	7	7	5	10	0	1	10	17	3	4
Hippocampus	0	0	0	0	0	0	0	0	0	0	5	10	1	1
Inferior Parietal	10	11	11	13	15	15	0	0	6	16	42	55	4	4
Inferior Temporal	0	0	4	6	5	6	0	1	1	3	10	16	3	4
Insula	2	9	0	1	4	4	4	5	2	5	12	24	4	5
Isthmus Cingulate	1	1	3	3	9	9	0	0	0	1	13	14	3	4
Lateral Occipital	3	4	13	13	0	0	0	0	5	7	21	24	3	3
Lateral Orbito Frontal	1	3	0	0	0	0	2	8	0	0	3	11	2	2
Lingual	2	4	2	2	6	6	0	0	5	6	15	18	4	4
Medial Orbito Frontal	0	0	0	0	0	0	1	3	0	0	1	3	1	1
Middle Temporal	2	3	5	7	2	5	8	13	1	5	18	33	5	5
Paracentral	0	0	1	1	5	5	0	0	1	1	7	7	3	3
Parahippocampal	0	0	1	2	11	12	0	0	0	0	12	14	2	2
Parsopecularis	1	1	0	0	0	0	5	16	0	0	6	17	2	2
Parsorbitalis	0	0	0	0	0	0	2	1	0	0	2	1	1	1
Parstriangularis	0	2	0	0	0	0	0	6	0	0	0	8	0	2
Pericalcerine	3	5	3	3	5	5	0	0	3	4	14	17	4	4
Post Central	4	5	0	0	4	4	0	0	7	15	15	24	3	3
Posterior Cingulate	0	0	0	0	7	7	0	4	1	5	8	16	2	3
Pre Central	1	1	0	0	0	0	14	38	6	7	21	46	3	3
Precuneus	5	6	9	12	10	10	0		4	11	28	39	4	4
Rostral Anterior Cingulate	1	1	0	0	0	0	2	3	0	0	3	4	2	2
Rostral Middle Frontal	5	7	0	0	0	0	4	23	0	0	9	30	2	2
Superior Frontal	2	3	0	0	0	0	19	46	4	5	25	54	3	3
Superior Parietal	0	0	5	6	7	7	0	0	9	11	21	24	3	3
Superior Temporal	9	10	5	8	2	2	0	2	0	0	16	22	3	4
Supramarginal	15	20	9	11	9	10	0	0	6	15	39	56	4	4
Temporal Pole	0	0	0	0	3	3	0	0	0	0	3	3	1	1
Transverse Temporal	2	2	3	3	2	2	0	0	1	5	8	12	4	4
Unknown	*1*	*1*	*2*	*6*	*0*	*0*	*0*	*0*	*0*	*0*	*3*	7	2	2
TOTAL	77	110	84	109	118	124	82	218	65	126	426	687	-	-

Figure 4.6: This table details the distribution of electrodes across cortical areas for each participant, and the number of electrodes in each area where some portion of brain activity was reconstructed significantly ($p<0.05$; permutation test) for each participant. The highlighted areas are those where brain activity is reconstructed consistently across multiple participant and they include areas corresponding to the social vision pathway, in addition to traditional areas in ventral temporal cortex involved in object recognition.

CHAPTER 5

CONCLUSION AND FUTURE DIRECTIONS

The history of neuroscientific progress is interwoven with that of technological progress; shaped by its limitations and accelerated by its advancement. Recent progress in miniaturizing sensing technology into wearable forms have enabled the simultaneous recording of natural environments, our behavior within them, and brain activity with high resolution. Advances in Machine Learning/AI promise that rich real world recordings can be characterized scalably and accurately for neuroscientists to model the relationship between our brains, behavior, and what happens in the real world. However, realizing these promises for neuroscientific discovery requires addressing engineering, analytical, bioethics and privacy challenge that arise in studying the brain during natural behavior in the real world.

Chapter 3 of this thesis explored how we can address the engineering, technical, bioethics and privacy challenges particular to studying social behavior in the real world. An important issue addressed in this work was the discovery and correction of subtle synchronization problems that can derail the fusion of simultaneous recordings of the environment, behavior, and brain which degrade data quality and prevent meaningful data analysis. Another important issue concerned effective use of Machine Learning/AI models to parameterize real world behavior with high fidelity to create accurate ground truth representations i.e., detecting faces, their expressions, speech, what was said. Their training on highly sanitized datasets resulted in brittle performance on noisy real world datasets like those collected here. The need for accurate ground

truth representations in these datasets made it untenable to use model predictions directly. To address these issues, Chapter 3 operationalized a human in the loop approach where annotators verified and corrected model annotation of real world recordings to obtain accurate ground truth representations for analysis. It also outlined best practices and scalability considerations (human and computational costs) for the proposed approach. Subsequent work in Chapter 4 improved the durability and scalability of the annotation approach for faces, reducing the human effort required by combining newer computer vision models with custom developed models to track face dynamics and identity. Lastly, Chapter 3 outlined best practices with respect to the bioethical and privacy considerations, operationalizing the idea of ongoing consent that empowers participants to determine how recordings of their natural behavior are used for scientific research.

Chapter 4 outlined an analytical framework to model the uncontrolled variability of natural behavior in the real world. Although the focus of the analysis was on faces, the framework's conceptual elements were generic and can be adapted to other cognitive domains. The main elements of the framework were 1) Identifying appropriate behavioral events to which analysis can be anchored (eye-gaze fixations in the case of vision), 2) Rich parameterization of stimuli and/or behavior being analyzed (projecting faces into a parameterized face space in this instance), 3) The collection of large datasets that transform uncontrolled real world variability from a challenge into an asset (hours of recordings of unscripted social interactions), and 4) The use of statistical models that robustly reveal the neural underpinnings of perception and behavior (a jointly learned neuro-perceptual space). This analytical framework has the potential to benefit real world neuroscience efforts in other domains of cognition, and potentially even studies of natural behavior in animal models.

Highly fidelity reconstruction of stimuli and brain activity from each other was an underlying theme for modeling faces in Chapter 4. Stimuli (faces) were parameterized richly so that highly

salient and subtle aspects of faces i.e., big smiles or small frowns, could be reconstructed with high fidelity. Models optimized to reconstruct this rich parameterization could therefore reveal which salient and subtle aspects of faces the brain cared about the most. The qualitative accuracy of reconstructed faces and quantitative strength of top level statistics validated the robustness of the analytical framework and robust reconstructions of brain activity revealed the neural substrate underlying face perception in the real world, highlighting the critical role and neurodynamics of the recently proposed social vision pathway (156). Taken together, these observations affirmed that it is indeed possible to model the uncontrolled variability of the real world robustly.

The robustness of modeling results was a necessary condition for scientific discovery, but not sufficient. Interpretability was a critical consideration that influenced modeling choices in Chapter 4. A central feature of the modeling approach was its ability to jointly learn a space in which aspects of brain activity and sets of facial features are highly correlated. Movement in this "neuro-perceptual" space represented parametric changes in both brain activity and facial features, and this coupling established it as a population tuning space learned from data (158, 159). Notably, the linearity of the model structure and other elements of the modeling pipeline (face model) lent interpretability to the data manifold in this space. Non-linearities that emerged in the data manifold reflected differences in the brain's sensitivity to different aspects of faces and testing hypotheses about its geometry could reveal neural tuning, providing a foundation for neuroscientific discovery.

How do our brains code for facial expressions during social interactions? This is a compelling neuroscientific question because the neural code for differences of facial expressions is relatively unknown compared to other aspects such as identity, where distinct coding schemes have been reported (148, 149). Experimental inquiry into facial expressions has also been far from ecological validity, relying on posed canonical expressions that caricaturize the diversity of facial

expressions in real life into a few categories. Moreover, there is mounting experimental evidence that introducing different elements of ecological validity for faces (e.g., static vs dynamic or real vs movie faces) changes both brain and behavior (126, 128, 129, 130, 131, 132, 133, 134, 135). In this landscape, the analytical framework described earlier combined with recordings of natural interactions provided a unique opportunity to probe how our brains code for facial expressions.

Chapter 4 operationalized this opportunity by removing face identity, centering each expression on an individual's face against their resting facial shape and expression. This transformation related robustly to brain activity, validating the idea of a norm-based code for facial expressions. Subsequent testing of hypotheses about the geometry of this code was undertaken by analyzing the data manifold in the model's neuro-perceptual space. This exercise revealed two important aspects of neural population tuning for facial expressions. The first was that neural sensitivity is different for changes in the type and intensity of expressions. The second was that neural sensitivity to facial expressions decreases as expressions get more intense, which is an analog of Weber's law for facial expressions. As a recap, the definition of Weber's law states that "the size of perceptible changes in stimulus intensity is proportional to the intensity of pre-existing stimulus" – an example of Weber's law is that we can easily tell the difference between a 1 and 2 pound weight but find it much harder to tell the difference between a 101 and 102 pound weight. Taken together, these results portrayed neural tuning for people's facial expressions as oval shaped, where the ovals are pointed toward their resting facial expressions and get larger upon moving further away from the resting expression. Subsequently, these validity of these findings was probed and asserted with a controlled experiment.

In addition to their relevance for face perception, an attractive aspect of these findings was how they came out, progressing from real world observations to experimental validation, resembling how discoveries in fields such as Physics typically progress. Such progressions are not

unknown in neuroscience, but they are atypical because technological obstacles have historically inhibited studying the brain in the real world in the past. As an instance of this virtuous progression, I hope this work will encourage the adoption of real world approaches in human studies across different cognitive domains, and in animal studies where there is a growing interest in computational approaches to studying ethology.

There are several avenues for future work which can break down into two main categories. The first involves using the analytical framework presented in this thesis to explore neuroscientific questions during natural behavior. The second involves advances methodology in terms of solving engineering problems and developing algorithms to scale real world neuroscience.

In the first category, an obvious extension is incorporate the reconstruction of bodies (pose, shape, dynamics) into the reconstruction paradigm. This effort is easy to operationalize because models that can parameterize bodies with high fidelity are readily available (183, 184, 185, 186). Expanding the scope of social interactions to include audition is another opportunity that can be explored with appropriate parameterization (187, 188, 189). Voices are the auditory analog of faces, and appropriate parameterization of voices during speech in social interactions can open the door to studying investigating the neural representations underlying voices during natural interactions. A natural question that arises from the findings of Chapter 4 is whether our brain's code for people's voices in a norm-based manner. Another potential expansion of scope would be to include the behavior of participants themselves such as their faces, gestures, and speech. Doing so requires extensions to the real world paradigm to record the participants themselves. The EMU environments already capture such recordings, but ensuring that is done with sufficient resolution and synchronizing them with existing behavioral recordings involves additional effort. The parameterization of such recordings can be achieved in the same way as it is for world video recordings, but identifying appropriate behavioral events and quanta to anchor analysis against

requires careful consideration because the scope expands beyond the participants visual behavior (eye-movements).

The Epilepsy Monitoring Unit environment also offers the opportunity to study the brain during other (not necessarily social) behaviors such as eating. Little is known about the neural representations of food and what we do know comes from experimental studies (190, 191). A 1-2 week stay in the EMU involves up to ≈ 40 meals being consumed by participants, presenting an opportunity to capture their brain activity and behavior as they eat. However, doing so means their behavior must be recorded during all meal times, which is a logistical challenge due to the burden of wearing mobile eye-tracking glasses at all times OR trying to set up an experiment at each meal time. Addressing this challenge effectively requires approaches to record the behavior and the environment without the burden of a wearable device.

The category of methodological advances breaks down into engineering problems to scale real world neuroscience and algorithm development to improve our ability to model brain behavior relationships in the real world.

Improvements in engineering methodologies may scale real world neuroscience in a variety of ways, each with its own benefits. For instance, scaling the recording of real world visual behavior and the environment to days instead of hours can allow collecting an order of magnitude more data, but requires substituting wearable mobile eye-tracking with a less burdensome non-wearable system. Alternatively, collecting human single unit brain activity during natural behavior can sharpen the resolution at which we understand neural tuning using the same approaches here, but operationalizing the real world vision paradigm at centers which collect such data presents a logistical and engineering challenge. Finally, a third approach is to deploy the approaches described in this thesis with non-invasive brain imaging e.g., scalp EEG. Doing so

requires addressing limitations in the resolution of these techniques, but has the potential to scale real world studies to a much higher number of participants and enable greater mobility.

Lastly, the continuing development of analytical approaches and algorithms to facilitate real world neuroscientific studies is essential and foundational work for future progress. One important opportunity is the incorporation of dynamics into mutually supervised models that learn interpretable neuro-cognitive spaces. A second opportunity involves incorporating non-linear coordinate transformations into models with care toward maintaining the interpretability of their geometry. Finally, a third is to consider the bidirectionality of models that are learned with large amounts of data as a potential avenue for BCI applications to drive neuromodulation, such as for a visual BCI.

REFERENCES

[1] Paul Broca. Remarques sur le siège de la faculté du langage articulé; suivies d'une observation d'aphémie (perte de la parole). *Bulletins de la Société Anatomique (Paris)*, 6: 330–357, 398–407, 1861.

[2] Carl Wernicke. Der aphasische symptomencomplex: Eine psychologische studie auf anatomischer basis. *Cohn & Weigert*, 1874.

[3] Korbinian Brodmann. *Vergleichende Lokalisationslehre der Grosshirnrinde in ihren Prinzipien dargestellt auf Grund des Zellenbaues*. Barth, 1909.

[4] Gordon Holmes and Wilder Penfield. The organization of the visual cortex in man. *Transactions of the Ophthalmological Societies of the United Kingdom*, 65:91–119, 1945.

[5] William B Scoville and Brenda Milner. Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery & Psychiatry*, 20:11–21, 1957.

[6] Antoine Bechara, Daniel Tranel, Hanna Damasio, and Antonio R Damasio. Double dissociation of conditioning and declarative knowledge relative to the amygdala and hippocampus in humans. *Science*, 293(5533):1389–1392, 2001.

[7] E. H. Weber. De tactu. *Handbuch der Physiologie des Menschen für Vorlesungen*, 2: 481–588, 1834.

[8] Gustav Theodor Fechner. *Elemente der Psychophysik*. Breitkopf & Härtel, 1860.

[9] Hermann Ebbinghaus. *Über das Gedächtnis: Untersuchungen zur experimentellen Psy-*

chologie. Duncker & Humblot, 1885.

[10] Hermann von Helmholtz. Die lehre von den tonempfindungen als physiologische grundlage für die theorie der musik. *Vieweg*, 1, 1852.

[11] Hans Berger. Über das elektrenkephalogramm des menschen. *Archiv für Psychiatrie und Nervenkrankheiten*, 87(1):527–570, 1929.

[12] Edgar D Adrian and W. T. Matthews. The berger rhythm: Potential changes from the occipital lobes in man. *Brain*, 57(4):355–385, 1934.

[13] H. H. Jasper and H. L. Andrews. Electro-encephalography: I. normal adults. *Journal of Experimental Psychology*, 23(3):213–244, 1938.

[14] W. G. Walter and G. O. Dove. Determination of relative energy changes in human brain during sleep by eeg patterns. *Journal of Neurophysiology*, 10(4):225–230, 1947.

[15] Pawel J. Matusz, Suzanne Dikker, Alexander G. Huth, and Catherine Perrodin. Are We Ready for Real-world Neuroscience? *Journal of Cognitive Neuroscience*, 31(3):327–338, 03 2019.

[16] Gijs A. Holleman, Ignace T. C. Hooge, Chantal Kemner, and Roy S. Hessels. The 'real-world approach' and its problems: A critique of the term ecological validity. *Frontiers in Psychology*, 11, 2020.

[17] James J. Gibson. *The Ecological Approach to Visual Perception: Classic Edition*. Houghton Mifflin, 1979.

[18] Jamil Zaki and Kevin Ochsner. The need for a cognitive neuroscience of naturalistic social cognition. *Annals of the New York Academy of Sciences*, 1167:16, 2009.

[19] Daniel L Powell and Gil G Rosenthal. What artifice can and cannot tell us about animal behavior. *Current zoology*, 63(1):21–26, 2017.

[20] Eleanor A. Maguire. Does memory research have a realistic future? *Trends in Cognitive Sciences*, 26(12):1043–1046, December 2022. ISSN 13646613.

[21] Gabriella Vigliocco, Laura Convertino, Sara De Felice, Lara Gregorians, Viktor Kewenig, Marie A. E. Mueller, Sebastijan Veselic, Mirco Musolesi, Andrew Hudson-Smith, Nick Tyler, Eirini Flouri, and Hugo Spiers. Ecological Brain: Reframing the Study of Human Behaviour and Cognition. preprint, PsyArXiv, April 2023.

[22] Sam Wass and Emily J.H. Jones. Editorial perspective: Leaving the baby in the bathwater in neurodevelopmental research. *Journal of Child Psychology and Psychiatry*, 64(8): 1256–1259, 2023.

[23] Agustin Ibanez, Morten L. Kringelbach, and Gustavo Deco. A synergetic turn in cognitive neuroscience of brain diseases. *Trends in Cognitive Sciences*, page S1364661323003066, January 2024.

[24] Pawel J. Matusz, Suzanne Dikker, Alexander G. Huth, and Catherine Perrodin. Are We Ready for Real-world Neuroscience? *Journal of Cognitive Neuroscience*, 31(3):327–338, March 2019. ISSN 0898-929X, 1530-8898.

[25] Simone G. Shamay-Tsoory and Avi Mendelsohn. Real-Life Neuroscience: An Ecological Approach to Brain and Behavior Research. *Perspectives on Psychological Science*, 14(5): 841–859, September 2019. ISSN 1745-6916, 1745-6924.

[26] Matthias Stangl, Uros Topalovic, Cory S. Inman, Sonja Hiller, Diane Villaroman, Zahra M. Aghajan, Leonardo Christov-Moore, Nicholas R. Hasulak, Vikram R. Rao, Casey H. Halpern, Dawn Eliashiv, Itzhak Fried, and Nanthia Suthana. Boundary-anchored neural mechanisms of location-encoding for self and others. *Nature*, 589(7842):420–425, January 2021. Publisher: Nature Publishing Group.

[27] Matthias Stangl, Sabrina L. Maoz, and Nanthia Suthana. Mobile cognition: imaging the human brain in the real world. *Nature Reviews Neuroscience*, 24(6):347–362, June 2023. Publisher: Nature Publishing Group.

[28] Ariel Goldstein, Haocheng Wang, Leonard Niekerken, Zaid Zada, Bobbi Aubrey, Tom

Sheffer, Samuel A. Nastase, Harshvardhan Gazula, Mariano Schain, Aditi Singh, Aditi Rao, Gina Choe, Catherine Kim, Werner Doyle, Daniel Friedman, Sasha Devore, Patricia Dugan, Avinatan Hassidim, Michael Brenner, Yossi Matias, Orrin Devinsky, Adeen Flinker, and Uri Hasson. Deep speech-to-text models capture the neural basis of spontaneous speech in everyday conversations, June 2023. Pages: 2023.06.26.546557Section: New Results.

[29] Uri Hasson, Yuval Nir, Ifat Levy, Galit Fuhrmann, and Rafael Malach. Intersubject Synchronization of Cortical Activity During Natural Vision. *Science*, 303(5664):1634–1640, March 2004. Publisher: American Association for the Advancement of Science.

[30] Arish Alreja, Michael James Ward, Qianli Ma, Mark Richardson, Brian Russ, Stephan Bickel, Nelleke Van Wouwe, Jorge A González-Martínez, Lisa S Parker, Joseph Neimat, et al. A multimodal approach to investigate the neural mechanisms of real world social vision. *Behavior Research Methods*, 2022.

[31] Uros Topalovic, Zahra M Aghajan, Diane Villaroman, Sonja Hiller, Leonardo Christov-Moore, Tyler J Wishard, Matthias Stangl, Nicholas R Hasulak, Cory S Inman, Tony A Fields, Vikram R Rao, Dawn Eliashiv, Fried Itzhak, and Nanthia Suthana. Wireless programmable recording and stimulation of deep brain activity in freely moving humans. *Neuron*, 108(2):322–334, 2020.

[32] David I Perrett, PAJ Smith, DD Potter, AJ Mistlin, AS Head, Arthur David Milner, and MA Jeeves. Visual cells in the temporal cortex sensitive to face view and gaze direction. *Proceedings of the Royal society of London. Series B. Biological sciences*, 223(1232): 293–317, 1985.

[33] Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008.

[34] Charles F Cadieu, Ha Hong, Daniel LK Yamins, Nicolas Pinto, Diego Ardila, Ethan A Solomon, Najib J Majaj, and James J DiCarlo. Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLoS computational biology*, 10(12):e1003963, 2014.

[35] Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365, 2016.

[36] Chengxu Zhuang, Siming Yan, Aran Nayebi, Martin Schrimpf, Michael C. Frank, James J. DiCarlo, and Daniel L. K. Yamins. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3):e2014196118, 2021.

[37] Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? *bioRxiv*, 2020.

[38] Vicki Bruce and Andy Young. Understanding face recognition. *British journal of psychology*, 77(3):305–327, 1986.

[39] Vadim Axelrod and Galit Yovel. Hierarchical processing of face viewpoint in human visual cortex. *Journal of Neuroscience*, 32(7):2442–2452, 2012.

[40] Winrich A Freiwald and Doris Y Tsao. Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science*, 330(6005):845–851, 2010.

[41] Fernando M Ramírez, Radoslaw M Cichy, Carsten Allefeld, and John-Dylan Haynes. The neural code for face orientation in the human fusiform face area. *Journal of Neuroscience*, 34(36):12155–12167, 2014.

[42] Avniel Singh Ghuman, Nicolas M Brunet, Yuanning Li, Roma O Konecky, John A Pyles,

Shawn A Walls, Vincent Destefino, Wei Wang, and R Mark Richardson. Dynamic encoding of face information in the human fusiform gyrus. *Nature communications*, 5(1):1–10, 2014.

[43] Yuanning Li, R Mark Richardson, and Avniel Singh Ghuman. Posterior fusiform and midfusiform contribute to distinct stages of facial expression processing. *Cerebral Cortex*, 29(7):3209–3219, 2019.

[44] D Lundquist, A Flykt, and A Öhman. The Karolinska directed emotional faces. *Department of Neurosciences, Karolinska Hospital, Stockholm, Sweden*, 1998.

[45] James V Haxby, Andrew C Connolly, and J Swaroop Guntupalli. Decoding neural representational spaces using multivariate pattern analysis. *Annual review of neuroscience*, 37: 435–456, 2014.

[46] Elizabeth A Hirshorn, Yuanning Li, Michael J Ward, R Mark Richardson, Julie A Fiez, and Avniel Singh Ghuman. Decoding and disrupting left midfusiform gyrus activity during word reading. *Proceedings of the National Academy of Sciences*, 113(29):8162–8167, 2016.

[47] Kai J Miller, Gerwin Schalk, Dora Hermes, Jeffrey G Ojemann, and Rajesh PN Rao. Spontaneous decoding of the timing and content of human object perception from cortical surface recordings reveals complementary information in the event-related potential and broadband spectral change. *PLoS Computational Biology*, 12(1):e1004660, 2016.

[48] Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Academic press, 2013.

[49] Shlomo S Sawilowsky. New effect size rules of thumb. *Journal of Modern Applied Statistical Methods*, 8(2):26, 2009.

[50] Nicholas Furl, Nicola J Van Rijsbergen, Alessandro Treves, Karl J Friston, and Raymond J Dolan. Experience-dependent coding of facial expression in superior temporal sulcus.

Proceedings of the National Academy of Sciences, 104(33):13485–13489, 2007.

[51] Varun Saravanan, Gordon J Berman, and Samuel J Sober. Application of the hierarchical bootstrap to multi-level data in neuroscience. *Neurons, behavior, data analysis and theory*, 3(5), 2020.

[52] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Springer series in Statistics, New York, 2001.

[53] Annie WY Chan, Dwight J Kravitz, Sandra Truong, Joseph Arizpe, and Chris I Baker. Cortical representations of bodies and faces are strongest in commonly experienced configurations. *Nature neuroscience*, 13(4):417–418, 2010.

[54] Ilker Yildirim, Mario Belledonne, Winrich Freiwald, and Josh Tenenbaum. Efficient inverse graphics in biological face processing. *Science Advances*, 6(10), 2020.

[55] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *2009 sixth IEEE international conference on advanced video and signal based surveillance*, pages 296–301. Ieee, 2009.

[56] Tim C Kietzmann, Anna L Gert, Frank Tong, and Peter König. Representational dynamics of facial viewpoint encoding. *Journal of cognitive neuroscience*, 29(4):637–651, 2017.

[57] John E Hummel and Brian J Stankiewicz. Two roles for attention in shape perception: A structural description model of visual scrutiny. *Visual Cognition*, 5(1-2):49–79, 1998.

[58] Fang Fang and Sheng He. Viewer-centered object representation in the human visual system revealed by viewpoint aftereffects. *Neuron*, 45(5):793–800, 2005.

[59] Rebecca P. Lawson, Colin W. G. Clifford, and Andrew J. Calder. A real head turner: Horizontal and vertical head directions are multichannel coded. *Journal of Vision*, 11(9): 17–17, 08 2011.

[60] Juan Chen, Hua Yang, Aobing Wang, and Fang Fang. Perceptual consequences of face

viewpoint adaptation: Face viewpoint aftereffect, changes of differential sensitivity to face view, and their relationship. *Journal of Vision*, 10(3):12–12, 03 2010.

[61] J Swaroop Guntupalli, Kelsey G Wheeler, and M Ida Gobbini. Disentangling the representation of identity from head view along the human face processing pathway. *Cerebral Cortex*, 27(1):46–53, 2017.

[62] Fernando M Ramírez. Orientation encoding and viewpoint invariance in face recognition: inferring neural properties from large-scale signals. *The Neuroscientist*, 24(6):582–608, 2018.

[63] Jackson C. Liang, Anthony D. Wagner, and Alison R. Preston. Content Representation in the Human Medial Temporal Lobe. *Cerebral Cortex*, 23(1):80–96, 01 2012.

[64] Blair Kaneshiro, Marcos Perreau Guimaraes, Hyung-Suk Kim, Anthony M Norcia, and Patrick Suppes. A representational similarity analysis of the dynamics of object processing using single-trial eeg classification. *Plos one*, 10(8):e0135697, 2015.

[65] Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

[66] Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.

[67] B. Efron. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7 (1):1 – 26, 1979.

[68] Bradley Efron. *The jackknife, the bootstrap and other resampling plans*. SIAM, 1982.

[69] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[70] Mariana Jacob Rodrigues, Octavian Postolache, and Francisco Cercas. Physiological and behavior monitoring systems for smart healthcare environments: A review. *Sensors*, 20 (8):2186, 2020.

[71] P Johnson and DC Andrews. Remote continuous physiological monitoring in the home.

Journal of telemedicine and telecare, 2(2):107–113, 1996.

[72] Frank H Wilhelm, Monique C Pfaltz, and Paul Grossman. Continuous electronic data capture of physiology, behavior and experience in real life: towards ecological momentary assessment of emotion. *Interacting with Computers*, 18(2):171–186, 2006.

[73] Marta Vigier, Katherine R Thorson, Elisabeth Andritsch, Herbert Stoeger, Leonie Suerth, Clemens Farkas, and Andreas R Schwerdtfeger. Physiological linkage during interactions between doctors and cancer patients. *Social Science & Medicine*, 284:114220, 2021.

[74] Rosie Clark, James Blundell, Matt J Dunn, Jonathan T Erichsen, Mario E Giardini, Irene Gottlob, Chris Harris, Helena Lee, Lee Mcilreavy, Andrew Olson, Jay E. Self, Valldeflors Vinuela-Navarroo, Jonathan Waddington, J. Margaret Woodhouse, Iain D. Gilchrist, and Cathy Williams. The potential and value of objective eye tracking in the ophthalmology clinic. *Eye*, 33(8):1200–1202, 2019.

[75] Alexandra Wolf and Kazuo Ueda. Contribution of eye-tracking to study cognitive impairments among clinical populations. *Frontiers in Psychology*, 12:2080, 2021.

[76] Juhee Jhalani, Tanya Goyal, Lynn Clemow, Joseph E Schwartz, Thomas G Pickering, and William Gerin. Anxiety and outcome expectations predict the white-coat effect. *Blood pressure monitoring*, 10(6):317–319, 2005.

[77] Thomas G Pickering, William Gerin, and Amy R Schwartz. What is the white-coat effect and how should it be measured? *Blood pressure monitoring*, 7(6):293–300, 2002.

[78] Arnstein Finset and Trond A Mjaaland. The medical consultation viewed as a value chain: a neurobehavioral approach to emotion regulation in doctor–patient interaction. *Patient education and counseling*, 74(3):323–330, 2009.

[79] Donald J Kiesler and Stephen M Auerbach. Optimal matches of patient preferences for information, decision-making and interpersonal behavior: evidence, models and interventions. *Patient education and counseling*, 61(3):319–341, 2006.

[80] Sonja Weilenmann, Ulrich Schnyder, Brian Parkinson, Claudio Corda, Roland Von Kaenel, and Monique C Pfaltz. Emotion transfer, emotion regulation, and empathy-related processes in physician-patient interactions and their association with physician well-being: a theoretical model. *Frontiers in psychiatry*, 9:389, 2018.

[81] Jeffrey M. Girard, Alexandria K. Vail, Einat Liebenthal, Katrina Brown, Can Misel Kilciksiz, Luciana Pennant, Elizabeth Liebson, Dost Öngür, Louis-Philippe Morency, and Justin T. Baker. Computational analysis of spoken language in acute psychosis and mania. *Schizophrenia Research*, 2021.

[82] Michal Muszynski, Jamie Zelazny, Jeffrey M Girard, and Louis-Philippe Morency. Depression severity assessment for adolescents at high risk of mental disorders. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 70–78, 2020.

[83] Jasara N Hogan and Brian R Baucom. Behavioral, affective, and physiological monitoring. In *Computer-assisted and web-based innovations in psychology, special education, and health*, pages 3–31. Elsevier, 2016.

[84] Saul Shiffman, Arthur A Stone, and Michael R Hufford. Ecological momentary assessment. *Annu. Rev. Clin. Psychol.*, 4:1–32, 2008.

[85] Debra L Roter and Judith A Hall. Studies of doctor-patient interaction. *Annual review of public health*, 10(1):163–180, 1989.

[86] Dora Hermes, Kai J Miller, Herke Jan Noordmans, Mariska J Vansteensel, and Nick F Ramsey. Automated electrocorticographic electrode localization on individually rendered brain surfaces. *Journal of Neuroscience Methods*, 185(2):293–298, 2010.

[87] François Tadel, Sylvain Baillet, John C Mosher, Dimitrios Pantazis, and Richard M Leahy. Brainstorm: a user-friendly application for MEG/EEG analysis. *Computational Intelligence and Neuroscience*, 2011, 2011.

[88] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once:

Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[89] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 59–66. IEEE, 2018.

[90] E Friesen and Paul Ekman. Facial action coding system: a technique for the measurement of facial movement. *Palo Alto*, 3(2):5, 1978.

[91] Marvin Lavechin, Marie-Philippe Gill, Ruben Bousbib, Hervé Bredin, and Leibny Paola Garcia-Perera. End-to-end Domain-Adversarial Voice Activity Detection. In *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Barcelona, Spain, May 2020.

[92] Ruiqing Yin, Hervé Bredin, and Claude Barras. Neural Speech Turn Segmentation and Affinity Propagation for Speaker Diarization. In *19th Annual Conference of the International Speech Communication Association, Interspeech 2018*, Hyderabad, India, September 2018.

[93] Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. Pyannote.audio: Neural building blocks for speaker diarization. In *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Barcelona, Spain, May 2020.

[94] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[95] Rahul S Desikan, Florent Ségonne, Bruce Fischl, Brian T Quinn, Bradford C Dickerson, Deborah Blacker, Randy L Buckner, Anders M Dale, R Paul Maguire, Bradley T Hyman, Marilyn S. Albert, and Ronald J. Killiany. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *Neuroimage*,

31(3):968–980, 2006.

[96] Antimo Buonocore, Olaf Dimigen, and David Melcher. Post-saccadic face processing is modulated by pre-saccadic preview: Evidence from fixation-related potentials. *Journal of Neuroscience*, 40(11):2305–2313, 2020.

[97] Christoph Huber-Huber and David Melcher. The behavioural preview effect with faces is susceptible to statistical regularities: Evidence for predictive processing across the saccade. *Scientific reports*, 11(1):1–10, 2021.

[98] Matthew J. Boring, Edward H. Silson, Michael J. Ward, R. Mark Richardson, Julie A. Fiez, Chris I. Baker, and Avniel Singh Ghuman. Multiple adjoining word- and face-selective regions in ventral temporal cortex exhibit distinct dynamics. *Journal of Neuroscience*, 41(29):6314–6327, 2021.

[99] Corentin Jacques, Jacques Jonas, Louis Maillard, Sophie Colnat-Coulbois, Laurent Koessler, and Bruno Rossion. The inferior occipital gyrus is a major cortical source of the face-evoked n170: Evidence from simultaneous scalp and intracerebral human recordings. *Human brain mapping*, 40(5):1403–1418, 2019.

[100] Truett Allison, Aina Puce, Dennis D Spencer, and Gregory McCarthy. Electrophysiological studies of human face perception. i: Potentials generated in occipitotemporal cortex by face and non-face stimuli. *Cerebral cortex*, 9(5):415–430, 1999.

[101] Annamaria Barczak, Saskia Haegens, Deborah A Ross, Tammy McGinnis, Peter Lakatos, and Charles E Schroeder. Dynamic modulation of cortical excitability during visual active sensing. *Cell reports*, 27(12):3447–3459, 2019.

[102] Allan Aasbjerg Nielsen. Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data. *IEEE transactions on image processing*, 11(3):293–305, 2002.

[103] Jamie Roche, Varuna De-Silva, Joosep Hook, Mirco Moencks, and Ahmet Kondoz. A

multimodal data processing system for lidar-based human activity recognition. *IEEE Transactions on Cybernetics*, 2021.

[104] Matteo Cognolato, Manfredo Atzori, and Henning Müller. Head-mounted eye gaze tracking devices: An overview of modern devices and recent advances. *Journal of Rehabilitation and Assistive Technologies Engineering*, 5, 2018.

[105] Yi Ren, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Almost unsupervised text to speech and automatic speech recognition. In *International Conference on Machine Learning*, pages 5410–5419. PMLR, 2019.

[106] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Y. Ng. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014.

[107] Zixing Zhang, Jürgen Geiger, Jouni Pohjalainen, Amr El-Desoky Mousa, Wenyu Jin, and Björn Schuller. Deep learning for environmentally robust speech recognition: An overview of recent developments. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 9(5):1–28, 2018.

[108] Svetlana Kiritchenko, Xiaodan Zhu, and Saif M Mohammad. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, 50:723–762, 2014.

[109] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[110] Alexander Mathis, Thomas Biasi, Steffen Schneider, Mert Yuksekgonul, Byron Rogers, Matthias Bethge, and Mackenzie W. Mathis. Pretraining boosts out-of-domain robustness for pose estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1859–1868, January 2021.

[111] Jessy Lauer, Mu Zhou, Shaokai Ye, William Menegas, Tanmay Nath, Mohammed Mostafizur Rahman, V. Di Santo, Daniel Soberanes, Guoping Feng, V. Murthy, G. Lauder, C. Dulac, M. Mathis, and Alexander Mathis. Multi-animal pose estimation and tracking with deeplabcut. *bioRxiv*, 2021.

[112] Tanmay Nath*, Alexander Mathis*, An Chi Chen, Amir Patel, Matthias Bethge, and Mackenzie W Mathis. Using deeplabcut for 3d markerless pose estimation across species and behaviors. *Nature Protocols*, 2019.

[113] Alexander Mathis, Pranav Mamidanna, Kevin M. Cury, Taiga Abe, Venkatesh N. Murthy, Mackenzie W. Mathis, and Matthias Bethge. Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, 2018.

[114] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *European Conference on Computer Vision*, pages 34–50. Springer, 2016.

[115] Emilie Delaherche, Mohamed Chetouani, Ammar Mahdhaoui, Catherine Saint-Georges, Sylvie Viaux, and David Cohen. Interpersonal synchrony: A survey of evaluation methods across disciplines. *IEEE Transactions on Affective Computing*, 3(3):349–365, 2012.

[116] Peter R Murphy, Ian H Robertson, Joshua H Balsters, and Redmond G O'connell. Pupillometry and p3 index the locus coeruleus–noradrenergic arousal function in humans. *Psychophysiology*, 48(11):1532–1543, 2011.

[117] Peter R Murphy, Redmond G O'connell, Michael O'sullivan, Ian H Robertson, and Joshua H Balsters. Pupil diameter covaries with bold activity in human locus coeruleus. *Human brain mapping*, 35(8):4140–4154, 2014.

[118] Dag Alnæs, Markus Handal Sneve, Thomas Espeseth, Tor Endestad, Steven Harry Pieter van de Pavert, and Bruno Laeng. Pupil size signals mental effort deployed during multiple object tracking and predicts brain activity in the dorsal attention network and the locus

coeruleus. *Journal of vision*, 14(4):1–1, 2014.

[119] Ali Azarbarzin, Michele Ostrowski, Patrick Hanly, and Magdy Younes. Relationship between arousal intensity and heart rate response to arousal. *Sleep*, 37(4):645–653, 2014.

[120] Jessica W Berg, Paul S Appelbaum, Charles W Lidz, and Lisa S Parker. *Informed consent: legal theory and clinical practice*. Oxford University Press, 2001.

[121] Howard B Degenholtz, Lisa S Parker, and Charles F Reynolds III. Trial design and informed consent for a clinic-based study with a treatment as usual control arm. *Ethics & behavior*, 12(1):43–62, 2002.

[122] Anna L Gert, Benedikt V Ehinger, Silja Timm, Tim C Kietzmann, and Peter König. Wild lab: A naturalistic free viewing experiment reveals previously unknown eeg signatures of face processing. *bioRxiv*, 2021.

[123] Barbara M Korsch, Ethel K Gozzi, and Vida Francis. Gaps in doctor-patient communication: I. doctor-patient interaction and patient satisfaction. *Pediatrics*, 42(5):855–871, 1968.

[124] Fabio Babiloni and Laura Astolfi. Social neuroscience and hyperscanning techniques: past, present and future. *Neuroscience & Biobehavioral Reviews*, 44:76–93, 2014.

[125] Matthias Stangl, Uros Topalovic, Cory S Inman, Sonja Hiller, Diane Villaroman, Zahra M Aghajan, Leonardo Christov-Moore, Nicholas R Hasulak, Vikram R Rao, Casey H Halpern, Dawn Eliashiv, Fried Itzhak, and Nanthia Suthana. Boundary-anchored neural mechanisms of location-encoding for self and others. *Nature*, 589(7842):420–425, 2021.

[126] Uri Hasson and Christopher J Honey. Future trends in neuroimaging: Neural processes as expressed within real-life contexts. *NeuroImage*, 62(2):1272–1278, 2012.

[127] Craig A. Anderson, James J. Lindsay, and Brad J. Bushman. Research in the psychological laboratory: Truth or triviality? *Current Directions in Psychological Science*, 8(1):3–9,

1999.

[128] Gustav Kuhn, Robert Teszka, Natalia Tenaw, and Alan Kingstone. Don't be fooled! attentional responses to social cues in a face-to-face and video magic trick reveals greater top-down control for overt than covert attention. *Cognition*, 146:136–142, 2016.

[129] Ross G Macdonald and Benjamin W Tatler. Gaze in a real-world social interaction: A dual eye-tracking study. *Quarterly Journal of Experimental Psychology*, 71(10):2162–2173, 2018.

[130] Laura M Pönkänen, Annemari Alhoniemi, Jukka M Leppänen, and Jari K Hietanen. Does it make a difference if i have an eye contact with you or with your picture? an erp study. *Social cognitive and affective neuroscience*, 6(4):486–494, 2011.

[131] Evan F Risko and Alan Kingstone. Attention in the wild: Visual attention in complex, dynamic, and social environments. In Robert R. Hoffman, Peter A. Hancock, Mark W. Scerbo, Raja Parasuraman, and James L. Szalma, editors, *The Cambridge Handbook of Applied Perception Research*, chapter 23, page 466–487. Cambridge University Press, Cambridge, 2015.

[132] Evan F Risko, Kaitlin E Laidlaw, Megan Freeth, Tom Foulsham, and Alan Kingstone. Social attention with real versus reel stimuli: toward an empirical approach to concerns about ecological validity. *Frontiers in human neuroscience*, 6:143, 2012.

[133] Evan F Risko, Daniel C Richardson, and Alan Kingstone. Breaking the fourth wall of cognitive science: Real-world social attention and the dual function of gaze. *Current Directions in Psychological Science*, 25(1):70–74, 2016.

[134] Colin Camerer and Dean Mobbs. Differences in behavior and brain activity during hypothetical and real choices. *Trends in cognitive sciences*, 21(1):46–56, 2017.

[135] Uri Nili, Hagar Goldberg, Abraham Weizman, and Yadin Dudai. Fear thou not: activity of frontal and temporal circuits in moments of real-life courage. *Neuron*, 66(6):949–962,

2010.

[136] Yin Wang and Ingrid R Olson. The original social network: white matter and social cognition. *Trends in cognitive sciences*, 22(6):504–516, 2018.

[137] Doris Y Tsao and Margaret S Livingstone. Mechanisms of face perception. *Annu. Rev. Neurosci.*, 31:411–437, 2008.

[138] Nancy Kanwisher. Domain specificity in face perception. *Nature neuroscience*, 3(8): 759–763, 2000.

[139] Jason JS Barton, Daniel Z Press, Julian P Keenan, and Margaret O'Connor. Lesions of the fusiform face area impair perception of facial configuration in prosopagnosia. *Neurology*, 58(1):71–78, 2002.

[140] Josef Parvizi, Corentin Jacques, Brett L Foster, Nathan Withoft, Vinitha Rangarajan, Kevin S Weiner, and Kalanit Grill-Spector. Electrical stimulation of human fusiform face-selective regions distorts face perception. *Journal of Neuroscience*, 32(43):14915–14920, 2012.

[141] Jiedong Zhang, Jia Liu, and Yaoda Xu. Neural decoding reveals impaired face configural processing in the right fusiform face area of individuals with developmental prosopagnosia. *Journal of Neuroscience*, 35(4):1539–1548, 2015.

[142] Jaideep Mavoori, Andrew Jackson, Chris Diorio, and Eberhard Fetz. An autonomous implantable computer for neural recording and stimulation in unrestrained primates. *Journal of Neuroscience Methods*, 148(1):71–77, 2005.

[143] Sabyasachi Roy and Xiaoqin Wang. Wireless multi-channel single unit recording in freely moving and vocalizing primates. *Journal of Neuroscience Methods*, 203(1):28–40, 2012.

[144] Jose A Fernandez-Leon, Arun Parajuli, Robert Franklin, Michael Sorenson, Daniel J Felleman, Bryan J Hansen, Ming Hu, and Valentin Dragoi. A wireless transmission neural interface system for unconstrained non-human primates. *Journal of Neural Engineering*,

12(5):056005, aug 2015.

[145] Lydia M Hopper, Roberto A Gulli, Lauren H Howard, Fumihiro Kano, Christopher Krupenye, Amy M Ryan, and Annika Paukner. The application of noninvasive, restraint-free eye-tracking methods for use with nonhuman primates. *Behavior Research Methods*, pages 1–28, 2020.

[146] Amy M Ryan, Sara M Freeman, Takeshi Murai, Allison R Lau, Michelle C Palumbo, Casey E Hogrefe, Karen L Bales, and Melissa D Bauman. Non-invasive eye tracking methods for new world and old world monkeys. *Frontiers in behavioral neuroscience*, 13: 39, 2019.

[147] Nancy Kanwisher, Josh McDermott, and Marvin M. Chun. The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception. *The Journal of Neuroscience*, 17(11):4302–4311, June 1997.

[148] David A. Leopold, Igor V. Bondar, and Martin A. Giese. Norm-based face encoding by single neurons in the monkey inferotemporal cortex. *Nature*, 442(7102):572–575, August 2006.

[149] Le Chang and Doris Y. Tsao. The Code for Facial Identity in the Primate Brain. *Cell*, 169 (6):1013–1028.e14, June 2017.

[150] T. Allison, H. Ginter, G. McCarthy, A. C. Nobre, A. Puce, M. Luby, and D. D. Spencer. Face recognition in human extrastriate cortex. *Journal of Neurophysiology*, 71(2):821–825, February 1994. Publisher: American Physiological Society.

[151] Doris Y. Tsao, Sebastian Moeller, and Winrich A. Freiwald. Comparing face patch systems in macaques and humans. *Proceedings of the National Academy of Sciences*, 105 (49):19514–19519, December 2008. Publisher: Proceedings of the National Academy of Sciences.

[152] Shlomo Bentin, Truett Allison, Aina Puce, Erik Perez, and Gregory McCarthy. Electro-

physiological Studies of Face Perception in Humans. *Journal of Cognitive Neuroscience*, 8(6):551–565, November 1996.

[153] B. Rossion. A network of occipito-temporal face-sensitive areas besides the right middle fusiform gyrus is necessary for normal face processing. *Brain*, 126(11):2381–2395, November 2003.

[154] Winrich Freiwald, Bradley Duchaine, and Galit Yovel. Face Processing Systems: From Neurons to Real-World Social Perception. *Annual Review of Neuroscience*, 39(1):325–346, July 2016.

[155] Kalanit Grill-Spector, Kevin S. Weiner, Kendrick Kay, and Jesse Gomez. The Functional Neuroanatomy of Human Face Perception. *Annual Review of Vision Science*, 3(1):167–196, September 2017.

[156] David Pitcher and Leslie G. Ungerleider. Evidence for a Third Visual Pathway Specialized for Social Perception. *Trends in Cognitive Sciences*, 25(2):100–110, February 2021.

[157] Beatrice De Gelder and Marta Poyo Solanas. A computational neuroethology perspective on body and expression perception. *Trends in Cognitive Sciences*, 25(9):744–756, September 2021.

[158] E. D. Adrian. The impulses produced by sensory nerve endings: Part I. *The Journal of Physiology*, 61(1):49–72, March 1926.

[159] Daniel A Butts and Mark S Goldman. Tuning Curves, Neuronal Variability, and Sensory Coding. *PLoS Biology*, 4(4):e92, March 2006.

[160] Cory T. Miller, David Gire, Kim Hoke, Alexander C. Huk, Darcy Kelley, David A. Leopold, Matthew C. Smear, Frederic Theunissen, Michael Yartsev, and Cristopher M. Niell. Natural behavior is the language of the brain. *Current Biology*, 32(10):R482–R493, May 2022.

[161] M. Franch, S. Yellapantula, A. Parajuli, N. Kharas, A. Wright, B. Aazhang, and V. Dragoi.

Visuo-frontal interactions during social learning in freely moving macaques. *Nature*, 627 (8002):174–181, March 2024. Publisher: Nature Publishing Group.

[162] Neda Shahidi, Melissa Franch, Arun Parajuli, Paul Schrater, Anthony Wright, Xaq Pitkow, and Valentin Dragoi. Population coding of strategic variables during foraging in freely moving macaques. *Nature Neuroscience*, pages 1–10, March 2024. Publisher: Nature Publishing Group.

[163] Aniruddha Das, Sarah Holden, Julie Borovicka, Jacob Icardi, Abigail Oâ€™Niel, Ariel Chaklai, Davina Patel, Rushik Patel, Stefanie Kaech Petrie, Jacob Raber, and Hod Dana. Large-scale recording of neuronal activity in freely-moving mice at cellular resolution. *Nature Communications*, 14(1):6399, October 2023.

[164] Hristos S. Courellis, Samuel U. Nummela, Michael Metke, Geoffrey W. Diehl, Robert Bussell, Gert Cauwenberghs, and Cory T. Miller. Spatial encoding in primate hippocampus during free navigation. *PLOS Biology*, 17(12):e3000546, December 2019.

[165] Camille Testard, Sebastien Tremblay, Felipe Parodi, Ron W. DiTullio, Arianna Acevedo-Ithier, Kristin L. Gardiner, Konrad Kording, and Michael L. Platt. Neural signatures of natural behaviour in socializing macaques. *Nature*, March 2024.

[166] D H Hubel and T N Wiesel. Receptive Fields of Single Neurones in the Cat's Striate Cortex. *Journal of Physiology*, 3(148):574–591, 1959.

[167] Apostolos P. Georgopoulos, Andrew B. Schwartz, and Ronald E. Kettner. Neuronal Population Coding of Movement Direction. *Science*, 233(4771):1416–1419, September 1986. Publisher: American Association for the Advancement of Science.

[168] J. O'Keefe and J. Dostrovsky. The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Research*, 34(1):171–175, November 1971.

[169] John O'Keefe. Place units in the hippocampus of the freely moving rat. *Experimental*

Neurology, 51(1):78–109, January 1976.

[170] Torkel Hafting, Marianne Fyhn, Sturla Molden, May-Britt Moser, and Edvard I. Moser. Microstructure of a spatial map in the entorhinal cortex. *Nature*, 436(7052):801–806, August 2005. Publisher: Nature Publishing Group.

[171] Jia Guo, Jiankang Deng, Alexandros Lattas, and Stefanos Zafeiriou. Sample and computation redistribution for efficient face detection. *arXiv preprint arXiv:2105.04714*, 2021.

[172] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[173] *A 3D Face Model for Pose and Illumination Invariant Face Recognition*, Genova, Italy, 2009. IEEE.

[174] Yudong Guo, Juyong Zhang, Jianfei Cai, Boyi Jiang, and Jianmin Zheng. Cnn-based real-time dense face reconstruction with inverse-rendered photo-realistic face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(6):1294–1307, 2019.

[175] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019.

[176] Ahmed A Abdelrahman, Thorsten Hempel, Aly Khalifa, and Ayoub Al-Hamadi. L2cs-net: fine-grained gaze estimation in unconstrained environments. *arXiv preprint arXiv:2203.03339*, 2022.

[177] Harold Hoteling. Relations between two sets of variates*. *Biometrika*, 28(3-4):321–377, 12 1936.

[178] Elena Parkhomenko, David Tritchler, and Joseph Beyene. Genome-wide sparse canonical correlation of gene expression with genotypes. In *BMC Proceedings*, volume 1, page

S119. BioMed Central, 2007.

[179] Daniela M Witten and Robert J Tibshirani. Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical applications in genetics and molecular biology*, 8(1), 2009.

[180] Daniela M Witten, Robert Tibshirani, and Trevor Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534, 2009.

[181] Chao Gao, Zongming Ma, and Harrison H. Zhou. Sparse cca: Adaptive estimation and computational barriers. *The Annals of Statistics*, 45(5):2074–2101, 2017.

[182] Qing Mai and Xin Zhang. An iterative penalized least squares approach to sparse canonical correlation analysis. *Biometrics*, 75(3):734–744, 2019.

[183] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Collaborative regression of expressive bodies using moderation. In *International Conference on 3D Vision (3DV)*, 2021.

[184] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, October 2015.

[185] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[186] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), November 2017.

[187] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya

Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023.

[188] Paul Boersma. Praat, a system for doing phonetics by computer. *Glot. Int.*, 5(9):341–345, 2001.

[189] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O. Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. Deep voice 3: 2000-speaker neural text-to-speech. In *International Conference on Learning Representations*, 2018.

[190] Meenakshi Khosla, N Apurva Ratan Murty, and Nancy Kanwisher. A highly selective response to food in human visual cortex revealed by hypothesis-free voxel decomposition. *Current Biology*, 32(19):4159–4171, 2022.

[191] Nidhi Jain, Aria Wang, Margaret M Henderson, Ruogu Lin, Jacob S Prince, Michael J Tarr, and Leila Wehbe. Selectivity for food in human ventral visual cortex. *Communications Biology*, 6(1):175, 2023.