A new perspective to cognitive neuroscience: understanding the information processing strategies of human brain

Yaocong Duan

Submitted in fulfilment of the requirements for the

Degree of Doctor of Philosophy

School of Engineering

College of Science and Engineering

University of Glasgow, UK

July 2023

Abstract

Visual categorization is one of the most fundamental and crucial cognitive functions of the human brain. An unsolved mystery about human visual categorization ability is its remarkable efficiency and generalization. These abilities rely on the brain's sophisticated system that actively extracts and processes the task-relevant features, through an integration of both bottom-up processing of visual input and top-down information of task context. This thesis argues that the complexity of features represented in the brain has not been adequately considered. Reconstructing how the brain actively transforms complex visual input into task-relevant features could shed light on the underpinnings of visual efficiency and generalization.

In this thesis, I employed a novel approach to tackle this problem. By precisely controlling the visibility of individual pixels in images and utilizing high spatial and temporal resolution neuroimaging data, combined with information-theoretic analysis framework, I reconstructed the specific pixels that are collectively represented by neural activity, thereby dynamically revealing the visual content (i.e. features) within images that are represented by brain and which brain regions, at what time, transform the complex visual input into task-relevant features.

Moreover, I also introduced a new information-theoretic measure, termed Samplewise Mutual Information (SMI), which quantifies the single-trial relationships between two variables. By applying SMI to characterize the single-trial relationship between participant behavior and the image samples, and employing Non-negative Matrix Factorization (NMF) clustering algorithm to learn the local parts of pixels that collectively influence the participant's behavior, I identified the finer components of the features that can better predict the participant's behavior. These feature components could serve as the minimal processing units by the brain.

These works advance our understanding of information processing strategies of the human brain for visual categorization tasks and open up a new analysis framework for future research.

Acknowledgements

I express my deepest gratitude to my supervisors, Prof. Philippe Schyns and Dr. Robin Ince, for their invaluable guidance, support, help, and mentorship throughout my PhD journey. The knowledge and wisdom that I have learned from them are immeasurable.

I would like to thank all the colleagues in Prof. Philippe Schyns' lab and Prof. Racheal Jack's lab. It’s so nice to meet them all in Glasgow. Particularly, thank Jiayu, Tian, Kasia, Nicola, Chaona, Racheal and Joachim for their kind help to me.

I am also very grateful to all staff at School of Psychology, School of Engineering and College of Science and Engineering for their kind administrative support.

Thank all my friends in Glasgow for their companionship and kind help to me.

Last but not least, I express sincere gratitude to the China Scholarship Council for providing me the scholarship.

Contents

List of Tables

List of Figures

1 General introduction

1.1 Overview

Visual categorization is one of the most important cognitive functions of the brain. By rapidly classifying visual inputs into meaningful categories, this ability allows us to efficiently interpret and interact with our surrounding environment. In the animal kingdom, fast and accurate visual categorization is vital for survival. For example, distinguishing between potential threats and non-threatening objects, or identifying safe versus dangerous food sources, can significantly impact an animal's chances of survival.

The brain's visual categorization abilities are not only highly efficient but also exhibit exceptional generalization capabilities, as demonstrated by the ability to apply learned categories to new, unfamiliar stimuli under varying conditions of lighting, angle, or cluttered scenes, and identify objects even when they are partially obscured (Biederman, 1987; DiCarlo et al., 2012; DiCarlo & Cox, 2007; Logothetis & Sheinberg, 1996). Despite significant advancements in artificial intelligence (AI), the human brain still outperforms AI in classifying visual inputs with greater ease and accuracy, as well as in generalizing across different contexts. Such efficiency and flexibility of the brain in visual categorization are attributed to its sophisticated architecture where bottom-up (feedforward) visual information is processed in a hierarchical manner (DiCarlo et al., 2012; Serre et al., 2007; VanRullen & Thorpe, 2001). This bottom-up processing is complemented by a top-down attentional mechanism (Desimone & Duncan, 1995, 1995; Evans et al., 2011; Harel et al., 2014; Kanwisher & Wojciulik, 2000; Moore & Zirnsak, 2017; Moran & Desimone, 1985), which allows the brain to selectively extract and process diagnostic features that are relevant to the task at hand, rather than processing all visual information. Similar attention mechanisms have also significantly enhanced AI capabilities (Vaswani et al., 2017).

Given the crucial role that selective feature extraction and processing play in understanding the efficiency and generalization of visual categorization, it has been a longstanding and important question in the field of visual cognitive neuroscience and still remains unclear.

One of the most influential model in cognitive neuroscience posits the brain as an information processing system (Marr & Ullman, 2010). Within this framework, a central goal of cognitive neuroscience is to understand what specific information the brain processes, and where, when, and how the brain processes this information to achieve visual

categorization behavior. To this end, an extensive body of work in cognitive neuroscience has been done to understand what specific visual features human participants used to achieve behavioral decisions (Bonnar et al., 2002; Gosselin & Schyns, 2001; Schyns et al., 2002; van Rijsbergen et al., 2014), where, when and how the brain represents these visual features into their neural activities (Huth et al., 2012; Schyns et al., 2007, 2009; M. L. Smith et al., 2004; Teichmann et al., 2023; Zhan, Ince, et al., 2019), the transition of visual features in brain networks (R. A. A. Ince et al., 2015; R. A. A. Ince, Jaworska, Gross, Panzeri, Van Rijsbergen, et al., 2016) and how top-down information of task contexts, attention and predictions can change the representation of visual features (Harel et al., 2014; Hebart et al., 2018; K. Kay et al., 2023; Schyns & Oliva, 1999; Yan et al., 2023; Zhan, Ince, et al., 2019).

1.1.1 An overlooked complexity in feature representation

Although these approaches have substantially advanced the understanding of visual information processing underlying categorization behaviors, there remains an overlooked complexity on feature representation that impedes a comprehensive elucidation of the visual cognitive system. Visual categorization relies on the mental/internal representation of features (Baker et al., 2022; Brinkman et al., 2017; DiCarlo & Cox, 2007; Murray, 2011). Given the modulation of spatial, object and feature attention (Brignani et al., 2010; Carrasco & Barbot, 2019; Evans et al., 2011; Moore & Zirnsak, 2017), the brain does not represent the entirety of visual input. Instead, the mental/internal representation of visual input only encompasses a subspace of their features. However, a single input image, and even a single object within this image, typically affords multiple different categorization behaviors. Observers can use, therefore represent, distinct features derived from the same images or objects to perform different categorizations. A study using ambiguous images has demonstrated that even without explicit categorization tasks, observer can perceive totally different content from a static image due to the perception of different features within the image (Bonnar et al., 2002; Zhan, Ince, et al., 2019). Therefore, the internal representation of the same input is not unique but can vary depending on task contexts, leading to distinct brain activity dynamics and behaviors.

Current perspectives posit that the neural representation of an input stimulus represents not only the properties of stimulus features, but also the internal states of the observer, including top-down information of task contexts, attention and prior knowledge (Çukur et al., 2013; Harel et al., 2014; Hebart et al., 2018). This perspective aligns with the idea that the features of a stimulus are not fixed or inherent but are actively and flexibly extracted by the brain

depending on the classification task at hand. Rather than passively receiving pre-existing features and selecting and enhancing the relevant ones, the brain engages in an active process to construct the necessary features from the visual input (Schyns et al., 1998; Schyns & Rodet, 1997). It highlights a dynamic transformation process of represented content during visual information processing.

This complexity of stimulus feature representations is often neglected in vision study because either the effect of the task is neglected (K. Kay et al., 2023) or only stimulating with full images, as is typical (Russakovsky et al., 2015; Rust & Movshon, 2005; Schrimpf et al., 2018). While studies using reverse correlation techniques uncover the task-relevant (i.e., diagnostic) features that participants use for visual categorization task (Gosselin & Schyns, 2001; Schyns et al., 2007; M. L. Smith et al., 2004; Zhan, Ince, et al., 2019), they do not unravel the specific task-dependent transformations of represented contents in the brain under a multiple-task design. Besides, Reverse correlation techniques describe the diagnostic features as the mental/internal representation of the stimulus because it is assumed that brain must process the features underlying behavior. However, it is not necessarily the case that the diagnostic features must be the minimum feature unit that the brain processes. If the mental representation provided by reverse correlation encompasses multiple processing units within the brain, rather than isolating each unit individually, the brain's processing of each independent unit might not be accurately captured. This poses a significant obstacle to elucidating the cognitive system since a crucial point in cognitive neuroscience is determining the specific visual information (the “what” question) that the brain processes to achieve the behavior (Schyns, 2018; Schyns et al., 2022).

1.1.2 The Aim of this thesis

The aim of this thesis is to develop a method combining reverse correlation (Gosselin & Schyns, 2001; Murray, 2011), neuroimaging (Baillet, 2017) and information theory (Cover & Thomas, 2012; R. A. A. Ince et al., 2017) to reconstruct the transformations of visual contents represented by neural representations throughout the process of neural information processing depending on the task at hand. Reverse correlation serves to manipulate and sample the visual contents within image stimuli. Neuroimaging serves to record brain activities. Information theory serves as an analysis framework to build the statistical relationship between stimuli and brain activity, quantifying the neural representation of stimuli, transformations of represented visual content and higher-order interactions about stimulus representations.

This method helps to investigate how the brain selects and extracts diagnostic features from complex visual inputs depending on the tasks. Reverse correlation methods can approximate the mental representations of stimulus features by human subjects based on their behavior. By reconstructing the transformations in the visual content represented by neural source activities under high spatial and temporal resolution (Baillet, 2017; Hillebrand & Barnes, 2005), we can observe when and in which brain regions these neural representations begin to resemble the mental representations. Since the neural representation of visual content depends on the specific visual information that the cognitive system is utilizing from images, analyzing how this represented content transforms can help us understand the brain's strategies for processing information.

1.1.3 Framing the question and the solution

The cognitive system receives the sensory inputs of external stimuli. The incoming sensory information is then handled by a series of cognitive processes implemented by neurons, to ultimately give rise to a behavioral decision. To formally frame this procedure, the cognitive processes can be studied in a system comprising four levels of concepts (see Figure 1-1):

1. external stimuli;

2. physical properties of the brain;

3. emerging mental properties of the mind;

4. behavioral decisions.

In this system, we can manipulate the external stimuli and ask participants to perform specific behavioral tasks while recording the (parts of) neural activity of their brain. Thus, stimuli, behavior and neural activity (physical brain properties) can be thought of as three measurable levels, while emerging mental properties of the mind are immeasurable. In this system, the aim of cognitive neuroscience can be framed as understanding the immeasurable emerging mind from the data of the other three measurable levels.

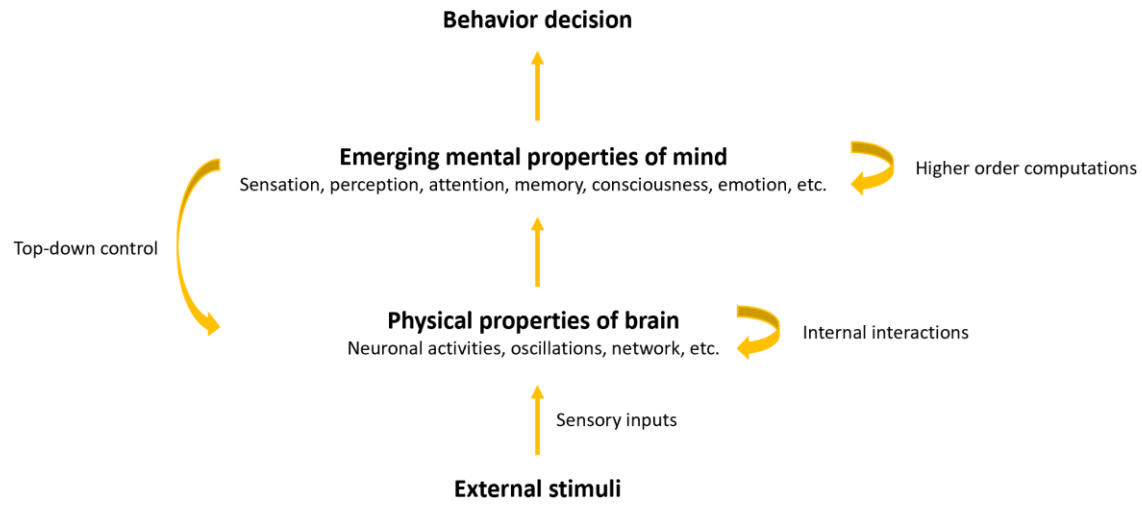

Figure 1-1 The architecture of cognitive neuroscience.

To understand the immeasurable level of the emerging mind, one approach is to analyze mental representations. There are two primary methods to characterize mental representations. The first method involves using reverse correlation techniques, where visual content is randomly sampled to determine which features of the stimuli influence the subject's behavior (Brinkman et al., 2017; Gosselin & Schyns, 2001; Murray, 2011; Schyns et al., 2002). Because the visual features influencing the behavior are thought to be represented. This approach provides an approximation of the mental representation based on behavioral responses. The second method involves measuring how visual content is represented in neural activity. Here, neural representations are considered as approximations of mental representations (Baker et al., 2022; Poldrack, 2021). However, it is important to note that the content represented by neural activity is dynamic and continuously evolving, rather than fixed. The transformation of neural representations depends on how the cognitive system utilizes information from visual inputs. Therefore, by reconstructing how the content of neural representations transforms over time, we can gain insights into the information processing strategies employed by the cognitive system, offering a way to explore the higher-level processes associated with the emerging mind.

It is important to bear in mind that none of them are the true mental representation. The features influencing behaviors are a subset of the mental representation as it can’t measure those features represented in the brain but not directly related to the behavior. For example, the salient distractor that are task-irrelevant can be dominant in the brain response (Lin et al., 2024).

1.1.4 Practical principles and the challenges of analysis

To deal with the above system, this thesis focuses on combining reverse correlation, neuroimaging techniques and information theory as a practical framework to understand cognitive systems.

I would like to highlight the importance of using high dimensional stimuli and reverse correlation technique. In a normal behavioral experiment, researchers must have a prior assumption of the stimulus properties that influence participant's behavior and randomly sample those properties during the experiment to observe their elicited responses. In reverse correlation experiment, these assumptions are not needed. Any property influencing participant's behavior will be automatically revealed if they are covered by the space of stimuli to which random noise is added. On the other hand, classical analysis methods in cognitive neuroscience focus on investigating spatiotemporal dynamics of neural activity that represent the stimulus, but they do not unravel how stimulus representations are transformed during neural processing. Characterizing the patterns of pixels that are collectively represented in neural activity and how the patterns change at different stages of neural processing provides a means to track such transformation of stimulus representations by neural processing.

Neuroscience makes the elusive aim of understanding the cognitive system feasible. However, it also introduces a very large and complex system that is difficult to deal with. How we study the complex relationships within this large system becomes a crucial problem in above analysis framework. Particularly when experiments involve a high dimensional stimulus (e.g., realistic scene images where each pixel is sampled) because the relationships between high-dimensional stimuli and high-dimensional brain activity need to be resolved. For example, in a 6mm source reconstructed magnetoencephalography (MEG) data there will be 12,773 voxels (sources) with also a high temporal resolution (e.g., usually 500-1000Hz sampling rates which means each voxel will have 500-1000 samples per second recording). Such a huge amount of data makes it difficult to both compute and visualize the relationships in the data.

Information theory, as a tool for measuring the variability and the relationships among variables, provides a solution for uncovering the complex interactions within the system efficiently. Normally, different mechanisms of relationships could produce the same effect size in terms of information theoretic measures. That means that one information theoretical

quantity will not correspond to one specific relationship but a group of relationships. A practical principle for information theoretic analysis is to use its quantities as an index to search the effects of interest from high dimensional data space and then disentangle the exact relationship by visualizing the structure of the relationships.

1.1.5 Organization of the thesis

To address these challenges, I used an analysis framework combining neuroimaging techniques, reverse correlation and information theory method. In the rest of this chapter, I will review reverse correlation, neuroimaging, the method of source reconstruction that I used in this thesis, the concept of visual representation and information theory. For information theory, I focus on understanding the concepts of information, using information theory as a tool for measuring relationships, entropy (as the core of information theory), Mutual Information (MI), Point-wise Mutual Information (PMI) and Co-Information (CoI), since these concepts are used in this thesis for measuring and decomposing relationships. I also briefly review the common-used information theory quantities.

The following chapters are organized as follows.

Chapter 2

Addressing these challenges requires a high-dimensional, fine-granularity control of the stimuli, multiple tasks experiment design and an efficient analysis framework. To this end, I designed an experiment comprising four 2-Alternative-Forced-Choice (AFC) categorization tasks applied to the same realistic, complex city street scene images randomly sampled with Bubbles procedure (Gosselin & Schyns, 2001). Bubbles approach randomly samples the pixels (i.e. fine granularity) of a stimulus image with Gaussian apertures. It ensures that the participant can only correctly categorize the images when the randomly sampled pixels show the features needed for categorization (Gosselin & Schyns, 2001; Schyns et al., 2002), enabling reverse correlating each pixel of image to the participant's behavior and brain activity. Considering each pixel as a feature dimension (i.e. high-dimensional), features that are represented into brain activity would be a subset of pixels representing a geometric subspace of the image. To enable the fast computation of the relationships between high-dimensionally sampled stimulus images and high-dimensional brain activity data as well as tracking task modulations on them, I employed information theory analysis using Gaussian Copula Mutual Information (GCMI) (R. A. A. Ince et al.,

2017). The results reveal three internal transformation stages of the image representation guided by pre-frontal cortex. From high-dimensional representation at stage 1 (50-120ms) where occipital sources represent more image features than the task requires, to stage 2 (121-150ms) where feature representations reduce to lower-dimensional manifolds, which then transform into the task-relevant features underlying categorization behavior over Stage 3 (161-350ms).

Chapter 3

While chapter 2 reverse engineered how tasks change the internal transformation of visual content represented in the brain, each task in chapter 2 only relied on a simple unitary feature of the images. In more complex scenarios, e.g., the perception of Dali's ambiguous painting (Bonnar et al., 2002; Zhan, Ince, et al., 2019), participants' perceptual decisions relied on multiple pieces of features in the image. It triggered the next critical question: What is the minimum representational unit that is represented into brain activity? Addressing this question requires a more comprehensive characterization of the relationship structures between stimulus samples and brain activity. To this end, I developed a new quantity under information theory framework—Sample-wise Mutual Information (SMI), inspired by Point-wise Mutual Information (PMI) in information theory (Cover & Thomas, 2012; R. A. A. Ince et al., 2017), to measure the specific contribution of each individual sample to the overall relationships between two variables. PMI measures the specific contribution of any combination of values of two variables. In this chapter, I introduce PMI and SMI and demonstrate their application on various datasets.

In next chapter, I will apply these two quantities to decompose the task-relevant features in a Dali's ambiguous painting perception experiment (Bonnar et al., 2002; Zhan, Ince, et al., 2019).

Chapter 4

To address the question: What is the minimum representational unit of feature manifolds that are represented into brain activity? I specifically test if the decomposition of task-relevant features can predict the participants' behavior better. The idea is that SMI between stimulus image samples and participants' behavior response can attribute the credit to each pixel that influence participants' behavior on individual trials. Therefore, it can reveal single-trial task-relevant feature images (i.e. single-trial classification images). Then we could cluster the pixels into their best local parts based on their trial-by-trial credit patterns.

I apply SMI on task-relevant features that are obtained from MI between participants' behavior responses and stimulus image samples to decompose single-trial relationship between them. I further apply the Non-negative Matrix Factorization (NMF) algorithm (Lee & Seung, 1999) on matrix of single-trial task-relevant features to learn the parts that collectively support participant's behavior. Results show that NMF is able to decompose task relevant features into parts of eyes, nose, mouth, similar to stimulus features represented in participants' brain activity.

Chapter 5

In this chapter, I will give a general discussion on the results of this thesis and the future research plan on this topic.

1.2 Psychophysical reverse correlation techniques: measuring the mind through behaviors

1.2.1 Psychophysical reverse correlation

Psychophysics is a branch of psychology that deals with the relationship between physical stimuli and the perceptions they elicit, which provides crucial theoretical and methodological foundations for understanding human perceptual systems. The history of psychophysics can be traced back to the 19th century with the work of German psychologist Gustav Fechner, who proposed that the relationship between physical stimuli and perception (e.g., the threshold of detection, discrimination and identification of visual stimuli) could be quantified using mathematical equations (Gescheider, 2013).

Within the framework of psychophysics, the reverse correlation technique (also known as the classification images technique for visual study) emerged as a novel research paradigm to find an approximation of mental representation of stimuli by inferring critical stimulus features influencing participants' responses (Murray, 2011). This method is originally derived from auditory research of auditory tone detection in noise (Ahumada & Lovell, 1971) that explores the characteristics of the auditory stimulus features that best predict the observer's decision variable.

The name "reverse correlation" comes from the fact that the method is reversed from the traditional correlation method. In traditional correlation experiments, researchers present a set of predefined stimuli and observe what responses are elicited by each stimulus. In reverse correlation experiments, participants are presented with noisy stimuli (base stimuli added with random noise that varies from trail to trial) and are asked to perform a specific behavioral task over hundreds to thousands of trials of repeated experiments. The rationale behind reverse correlation is that if random noise distorts the task relevant features in stimuli, participants will be unable to correctly recognize the stimuli. Therefore, only when the relevant stimulus features are maintained in the randomly varying noise can participants make accurate behavioral responses. Reverse correlation then seeks to reveal the pattern (e.g., subareas on the images) in varied stimuli that is most likely to elicit a particular response, by relating the random noise to participants' behavior responses. For example, to discover which facial features underlie the recognition of gender, participants are presented with face images altered by random noise and are asked to judge the gender of the faces. Every pixel in the image changes from trial to trial due to noise, but only the variations in the pixels that are relevant to gender recognition will affect the accuracy of participants' judgment, while variations in other pixels will not. Therefore, the patterns of pixels that are collectively associated with accuracy can reveal the face regions that support gender judgment (Brinkman et al., 2017; Gosselin & Schyns, 2001; Schyns et al., 2002). The outcome of reverse correlation is called classification image (CI) which shows stimulus features that participants use in the task and these features are also called task-relevant features or diagnostic features (Gosselin & Schyns, 2001; Schyns et al., 2002; Zhan, Ince, et al., 2019).

There are two main ways to compute CI in reverse correlation experiments. One way is to average the random noise for a particular response and the resulting CI is thought to be the typical value of stimuli that elicit this response. The other way is to calculate the correlation between random noise and responses. This way reveals which stimulus features influences responses and the magnitude of correlation quantifies the weight of each stimulus feature to determine the response.

CI is also often regarded as an approximation of mental representation of the stimuli. Mental representations refer to the theorical internal cognitive constructs that mirror the properties of external stimuli, such as visual images. Directly measuring mental representations poses significant challenges due to their subjective and intangible nature. Reverse correlation offers a novel approach to visualize these mental representations. In reverse correlation, it is

assumed that people do visual categorization tasks by matching visual stimuli to a mental template/representation (Brinkman et al., 2017). For example, people identify a face as male face because the visual input of the face matches their mental template of male more than that of female. Different people may use different templates to do the same task thus subjective difference exists. Reverse correlation provides an approach to visualize the mental template/representation by inferring which facial features influence observers' judgments of a face's gender.

1.2.2 Bubbles

The reverse correlation method used in this thesis is called Bubbles (Gosselin & Schyns, 2001), a data-driven method that allows researchers to study which area of images (spatial visual information) the participants uses to perform visual categorization tasks. Specifically, Bubbles technique adds randomly positioned Gaussian apertures to the images that support the categorization task to vary the visibility of every pixel on the image. Bubbles sampling ensures the participant can only correctly categorize the stimulus when the random samples reveal the features that the participant uses in the task. After hundreds to thousands of trials, classification images can be generated by measuring the relationships between the bubbles sampling and participant’s behavior responses.

1.2.3 Classification images vs. fixation of eye movements

Fixation of eye movements and classification images are two different methods used to study visual perception and attention (Murray, 2011; Rayner, 1998). Fixation of eye movements is a method that involves tracking the movement of the eyes as a participant views a visual scene. By measuring where and how long a participant fixates their gaze for, researchers can gain insight into the visual features that capture attention, and how explicit attention is directed to different areas of a scene. This method is widely used in cognitive science and psychology to study visual attention, visual search, and visual perception.

One advantage of classification images over fixation of eye movements is that it allows researchers to study the internal representation (mental representation) of the visual stimulus rather than just the position of gaze. By analyzing the patterns of responses across a set of noise images, researchers can infer the visual features that support the task at hand and understand how the visual system uses this information to perform visual tasks.

1.2.4 Limitations of Reverse Correlation

While reverse correlation has proven to be a powerful tool in psychological study, it also has its limitations. One significant limitation of reverse correlation is that it requires features within images to be consistently positioned. For example, facial stimuli are well-suited for reverse correlation because key facial features like eyes, nose, and mouth are generally in the same relative locations across different faces. This consistency allows reverse correlation to effectively identify how variations in these features can influence perception, such as recognizing expressions. However, if features are not in a consistent position, reverse correlation becomes less effective, as the technique relies on stable spatial relationships to infer mental representations.

Another limitation is that reverse correlation typically requires a large number of trials to achieve reliable results, which can be time-consuming and potentially fatiguing for participants. Therefore, it is not suitable for populations of kids and older people.

1.3 Tracking neural representations with neuroimaging

The human brain is made up of approximately 85-100 billion neurons (Herculano-Houzel, 2009), which are the primary functional units of the nervous system. These nerve cells are connected to each other to form complex neural networks, whose activity encodes and transmits information, providing foundations for higher-level cognitive functions, such as vision, sensation, emotions, memory, and decision making (Ju & Bassett, 2020).

1.3.1 Neural representation

Neural representation refers to the process by which the brain encodes external stimuli into patterns or states of neural activity, essentially creating an internal "copy" or representation of the external world. This internal representation allows the brain to process and respond to stimuli in a meaningful way. At its core, neural representation involves a population of neurons using their specific activity states to encode external stimuli (Baker et al., 2022; K. N. Kay, 2011; Kriegeskorte & Kievit, 2013; Poldrack, 2021).

In cognitive psychology, a core concept is mental representation, which is a theoretical construct believed to underlie higher cognitive functions. However, since it is theoretical,

mental representations cannot be directly observed. Even though psychophysical methods, such as reverse correlation techniques, can approximate mental representations, these approximations do not directly capture the mental representations themselves. Neural representation offers an alternative approach from neuroscience perspective. While we cannot directly observe mental representations, we can observe neural representations. These neural representations may serve as the physical basis for mental representations or, at the very least, provide another means of approximating them. This connection between neural and mental representations opens up a new avenue for understanding the physical underpinnings of cognitive processes.

In cognitive neuroscience research, the term representation often refers to the systematic relationship between neural activity and the variations in external stimuli (Baker et al., 2022; K. N. Kay, 2011; Poldrack, 2021). This relationship is based on the premise that if a group of neurons represents an external stimulus through its specific activity states, then their activity must be correlated with variations in the stimulus. This approach does not directly explore the specific state-based mechanisms by which neurons encode a stimulus (Kriegeskorte & Kievit, 2013). Instead, the strength of this representation can be quantitatively assessed, allowing researchers to infer which brain regions, and at what times, are involved in encoding and processing information about external stimuli. This method helps us understand the underlying cognitive system.

1.3.2 Neuroimaging: measuring physical activities of the brain

To explore neural representations in the brain and to investigate how the brain processes information, it is essential to have methods that allow for the direct observation of neural activity. Neuroimaging techniques provide us with such a means, enabling researchers to measure neural activity in the brain (the physical properties of the brain). Various techniques in neuroimaging are used to measure and visualize the structure and function of the brain. The most commonly used noninvasive neuroimaging techniques include electroencephalography (EEG), magnetoencephalography (MEG), and functional magnetic resonance imaging (fMRI). Each of these techniques has its own pros and cons.

EEG is a non-invasive method to record an electrogram of the electrical activity of the brain using electrodes placed on the scalp. EEG has a high temporal resolution and is relatively inexpensive and portable, making it a popular technique in cognitive neuroscience research. However, EEG signals are sensitive to noise and artifacts, and the spatial resolution is

relatively low. MEG is another non-invasive technique that measures the magnetic fields generated by the brain electrical activity. Like EEG, it provides high temporal resolution recording of neural activity and has a better spatial resolution than EEG. Besides. MEG signals are less sensitive to noise and artifacts than EEG. Particularly, MEG signals are less distorted than EEG signals by the skull and scalp, which results in an advantage of source reconstruction for MEG (Hari et al., 2010). EEG and MEG are also considered to be two complementary recordings because MEG is more sensitive in detecting currents that are tangential to the surface of the scalp while EEG is more sensitive to tangential and radial neuronal activities. Thus, EEG can detect neuronal activities both in the sulci and at the top of the cortical gyri, whereas MEG is more sensitive to neuronal activities in sulci (Hari et al., 2010).

The neuroimaging technique used in this thesis is MEG considering its advantages in reconstructing the activity sources.

1.3.3 ERP

Event-related potentials (ERP) is an electrical brain response that is time-locked to specific events. ERPs are measured by EEG, normally calculated by averaging EEG signals that are time-locked to specific events over multiple trails. The rationale is that averaging can cancel out random noise in EEG signals (the average value of noise is approximately 0), while event-related EEG activity accumulates to form specific waveforms (Luck & Kappenman, 2012). ERPs provide a fundamental understanding of how brain processes information through temporal perspective.

Although this thesis is based on the MEG technique, I also aligned the results (i.e. the reconstructed internal transformations of representations) into different ERP time window. These cover:

C1

The ERP component C1 can be a negative-going component or a positive-going component with its peak normally observed in the 65–90 ms post-stimulus. Its thought to be linked with occipital hemifield responses (Slotnick, 2018).

P100

The ERP component P100 is a positive deflection in the waveform that peaks approximately 100 milliseconds after the presentation of a visual stimulus. The P100 is typically observed over the occipital scalp regions and is primarily associated with early visual processing, early attention/stimulus representation.

N170

The ERP component N170 is a negative deflection in the waveform that peaks approximately 170 milliseconds after the presentation of a visual stimulus. The N170 is typically observed over the occipito-temporal scalp regions and is primarily associated with the processing of faces, familiar objects or words (Bentin et al., 1996, 2007; R. A. A. Ince, Jaworska, Gross, Panzeri, Rijsbergen, et al., 2016; Kanwisher et al., 1997; Rossion et al., 2003; Rousselet et al., 2004; Schyns et al., 2007).

N250

The ERP component N250 is a negative deflection in the waveform that peaks approximately 250 milliseconds after the presentation of a visual stimulus. The N250 is typically observed over central and parietal scalp regions and is primarily associated with the cognitive processes related to attention, decision-making, and semantic processing N250 (Kaufmann et al., 2008).

P300

The ERP component P300 is a positive deflection in the waveform that peaks approximately 300-400 milliseconds after the presentation of a visual stimulus. The P300 is typically observed over parietal and central scalp regions and is primarily associated with attention, memory, categorization, and decision-making. P300 is considered to be an endogenous potential, as its occurrence links not to the physical attributes of a stimulus, but to a person's reaction to it (Nieuwenhuis et al., 2005; Ratcliff et al., 2009).

1.3.4 Source reconstruction

The EEG/MEG signals recorded from the scalp do not directly reflect the activated neuronal sources in the brain but actually are a mixing of those source signals. Source reconstruction is a method used to estimate the location and strength of the brain's neuronal sources activities from recorded scalp signals (Hillebrand & Barnes, 2005).

The problem that source's locations of activity have to be estimated from scalp recording data is called inverse problem. A main challenge of inverse problem is that theoretically it may not have a unique solution, just like inferring a three-dimensional scene from a two-dimensional image does not have a unique solution. That means multiple sources distribution can give rise to the same scalp recordings. Source reconstruction methods seek to find the best solution for the inverse problem using models involving prior knowledge of brain activity (Hillebrand & Barnes, 2005).

A main technique for source reconstruction is beamforming, where a theoretical model of the magnetic field produced by a given current dipole is used as a prior, along with second-order statistics of the data in the form of a covariance matrix, to calculate a linear weighting of the sensor array (the beamformer) via the Backus-Gilbert inverse. This is also known as a linearly constrained minimum variance (LCMV) beamformer. (Hillebrand & Barnes, 2005; Oostenveld et al., 2011). There are two main advantages of beamforming technique. One is that induced changes in cortical oscillatory power that do not result in a strong average-evoked response, known as event-related synchronization (ERS) and event-related desynchronization (ERD), can be identified and localized. The other one is that beamforming is practically simple and user-friendly. Beamforming does not need a predefined number of active sources as a priori and the only parameters that a user needs are the size of grid for the reconstruction, the time-frequency window over which to run the analysis, and optionally the amount of noise regularization (Hillebrand & Barnes, 2005).

1.4 Information theory

With methods to record neural activity (i.e. neuroimaging) and techniques to manipulate visual stimuli (i.e. psychophysics), a critical step in understanding neural representation and how the brain processes information is to measure the systematic relationships between these variables. This includes examining the relationship between two variables, such as the stimulus and the neuronal activity, as well as exploring more complex interactions, such as how neural activity represents multiple stimulus features or how multiple neurons collectively represent a stimulus. These relationships can be quantitatively assessed using information-theoretic approaches, which provide a framework for measuring the informational content, their interactions and transmissions within neural activities (R. A. A. Ince et al., 2017).

Information theory is a branch of mathematics that deals with the quantification, storage, and communication of information. It was developed by mathematician Claude Shannon in his 1948 seminal paper "A Mathematical Theory of Communication" as a way to understand the fundamental limits of communication and information processing (Shannon, 1948). Information theory provides the mathematical foundation for many modern technologies such as data compression, data encryption, and wireless communication. It has been widely used in fields like communication, machine learning, statistics, neuroscience, biology and physics.

1.4.1 The concept of information

Intuitively, information can be understood as the thing that reduces one's uncertainty about the outcome of an event. That is, if you gain the information about an event, you will go from being more uncertain ("unknown") to being more certain ("known") about that event and you can better predict the outcome. Thus, the amount of information is often quantified by the magnitude of reduction in uncertainty. In information theory, this is the basic idea behind the concept of entropy, which measures the *uncertainty* (interchangeably used with *variability* and *randomness*) of a random variable or more generally a system (e.g., brain). When new information is received, it can be used to reduce the entropy of a system, which is against the second law of thermodynamics (Friston, 2010a). The total entropy of a variable or system is considered as the amount of information carried by that variable or system.

It is noteworthy that unlike the information we casually refer to, information in information theory does not have any specific semantic meaning, but only relates to the number of elementary events or messages. Information theory was originally designed to study communication. As Shannon describes in his information theory paper: the fundamental problem of communication is that of reproducing at one point ("receiver") either exactly or approximately a message selected at another point ("sender"). Often, the messages have meaning, but these semantic aspects of communication are irrelevant to the formal problem of their engineering. The significant aspect is that the actual message is that selected from a set of possibilities. So, the system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design (Shannon, 1948).

1.4.2 Information theory as a tool for measuring the relationships

In this thesis, I will use information theory as a framework for studying the relationship between variables (R. A. A. Ince et al., 2017). Using information theory to measure the relationship between variables is a natural endeavor. The goal of communication is to reproduce at the receiver end the signal of the sender. If reproduction is perfect, the received signal contains all the information of the sent signal, and the sent and received signals are completely correlated. If the signal varies during transmission, thereby differing with the original signal, the received signal has only part of the information of the original signal, and the two signals are partially correlated. Therefore, it is intuitive that the degree to which two signals share each other's information can be used as a measure of their relationships.

To illustrate usage of information theory to study a system (e.g., our goal: cognitive system), consider the example of the eyes (Figure 1-2). The eye is a complex system that captures external light and converts it into electrical signals, which are then processed by the brain to form visual perception. Our two eyes capture a 2D projection of the outside world meaning that a copy of information about the outside world is represented on the retina and its receptors, which can be measured with information theoretical quantity Mutual Information (MI). The 2D information (e.g., color information) is largely redundant between the two eyes, meaning that when one eye is closed, this information is not lost. This redundancy provides the visual system with robustness, as the loss or mutation of information in one eye does not lead to an error in the system, because the counterpart redundant information in the other eye is still available. However, 3D depth information from stereopsis, which is not seen by either eye alone (unlike 3D depth from shading), can only be perceived by the two eyes working together (i.e., the information from the two eyes is integrated). This kind of information is called synergy, which refers to the additional information that two variables convey only when they work together, while not conveyed by either variable individually (Luppi et al., 2022).

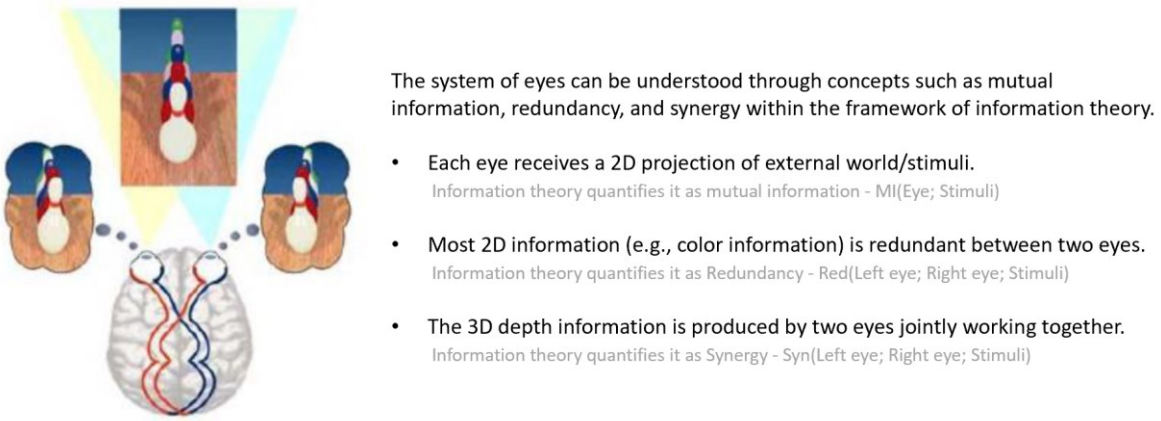

Figure 1-2 An illustration of using information theory to understand the eyes system.

In the next section, I review information theoretic quantities, including entropy, mutual information, redundancy, synergy and explain how to quantify these abstract concepts mathematically.

1.4.3 Review of information theory quantities

Information theoretic quantities are measures of the amount of information that is contained in a variable or variables in a system. They are used to understand and quantify the fundamental limits of communication and information processing.

1.4.3.1 Entropy

Entropy is a fundamental quantity in information theory that quantifies the *uncertainty*, interchangeably used with *variability* and *randomness*, of a random variable (Shannon, 1948). It is fundamental because all other information theory quantities can be derived from entropy (like the building blocks) in a simple and intuitive way (only involving addition and subtraction). Thus, understanding entropy is central for understanding information theory.

For a discrete random variable X with probability mass function $p(x)$ for each outcome value x, a low probability outcome means that outcome is less likely to occur, and so would be more surprising to an observer if it did occur. A high probability outcome would be less surprising. Information theory says that a more surprising outcome conveys more information, which is called the *surprisal* or *self-information* of that outcome. Mathematically, this notion can be expressed as

$$h(x) = -\log p(x)$$

The information value has units (e.g., bits or nats) depending on the base of logarithm. Conventionally the base of logarithm is taken to 2, corresponding to units of bits. Entropy of variable X is then defined as the expected (i.e., average) information (i.e., surprisal) conveyed by observing the outcome value x, that is

$$\begin{aligned} H(X) &= \sum_x p(x)h(x) \\ &= -\sum_x p(x)\log p(x). \end{aligned}$$

A higher entropy variable has a higher degree of uncertainty or variability, which means it is harder to predict the outcome. A lower entropy variable is easier to predict. Shannon proved that his entropy formula is the unique function that satisfies the axioms of being continuous, non-decreasing and additive for independent events.

In addition to information entropy, variance is also a representation of information. Variance is a statistical measure of the spread of a dataset by measuring the average of the squared deviation of each data point from the mean. In dimensionality reduction method Principal Component Analysis (PCA), variance is used to measure information that is retained in the transformed dataset (a matrix with multiple variables). The goal of PCA is to find the directions in the data with the most variance, and these directions are considered to be the directions that contain the most information. By projecting the data onto these directions, the dimensionality of the data can be reduced while retaining the most important information. Unlike information entropy, variance is defined on a set of data (i.e., a set of samples or trials), while information entropy is defined on the probability distribution. Therefore, for empirical data, calculating entropy usually needs to estimate the probability distribution from samples. Besides, a major advantage of information entropy for neuroimaging applications is that it provides a common meaningful effect size (i.e., bits) across many different statistical tests (with discrete, continuous, and multidimensional variables).

1.4.3.2 Mutual information

Though entropy per se has a wide range of applications, researcher often are not interested in the entropy of a certain variable, but instead, they want to know whether the variabilities of two different variables are mutually independent or correlated. Mutual information (MI)

is a measure of the amount of shared information between two variables (Shannon, 1948). It is defined as the amount of reduction in the uncertainty[1] of one variable that can be achieved by obtaining knowledge about the other variable. MI between variables X and Y is formulated as:

$$\begin{aligned} MI(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \end{aligned}$$

Another way to understand MI is by considering the joint entropy $H(X,Y) = -\sum_x \sum_y p(x,y) \log p(x,y)$ based on the joint distribution $P(X,Y)$. In the Venn diagram in Figure 1-3, $H(X)$ and $H(Y)$ can be thought of as the area of the left and right circles respectively, and $H(X,Y)$ is represented by the total area occupied by both circles. Then $MI(X,Y)$, which is represented as overlapping part of two circles, can be calculated as

$$MI(X;Y) = H(X) + H(Y) - H(X,Y)$$

These three formulas are completely equivalent, and they can be written into the same form by substituting the entropy formula as

$$MI(X;Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

This formula reveals the essence of measuring relationships through mutual information by comparing the joint probability of two variables and the product of the individual probabilities. Probability theory shows that when two variables are independent, the product of their probabilities equals the joint probability. In this case, the above formula takes a value of 0. The more correlated the two variables are, the larger the value of the above formula will be.

[1] As mentioned in section of “concept of information”, information reduces uncertainty.

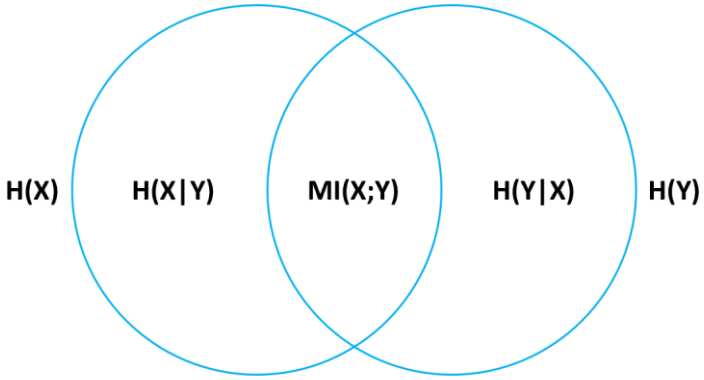

Figure 1-3 Venn diagram of mutual information.

1.4.3.3 Pointwise mutual information

Pointwise Mutual Information (PMI) is a concept similar to MI for quantifying the association between two variables. The difference is that PMI measures the relationship between each individual outcome of the two variables, whereas MI measures the overall association between the two variables. PMI of a pair of outcomes x and y from discrete random variables X and Y is defined as:

$$\begin{aligned} PMI(x;y) &= h(x) - h(x|y) \\ &= \mathrm{h(y)} - \mathrm{h(y|x)} \\ &= \mathrm{h(x)} + \mathrm{h(y)} - \mathrm{h(x,y)} \\ &= \log\frac{p(x,y)}{p(x)p(y)} \end{aligned}$$

The formula for PMI is similar in form to MI, but PMI is calculated based on surprisal instead; MI is expected (i.e., average) PMI:

$$MI(X,Y) = \sum_{x}\sum_{y} p(x,y)\, PMI(x,y)$$

Unlike MI, the value of PMI can be either positive or negative. The positive PMI value means that the outcome x and y are more likely to occur together than they would if they are independent. Negative PMI value means outcome x and y are less likely to occur together.

PMI is widely used in multiple fields, such as natural language processing, information retrieval, and machine learning. In natural language processing, it is used to evaluate the association between words in a corpus of text, and is often used to identify collocations (i.e., words that tend to occur together). In information retrieval, PMI is used to estimate the relevance of a document to a query and to rank search results. In machine learning, it is used to identify features that are highly correlated with a target variable and to enhance the performance of classifiers and other algorithms.

However, PMI also has some limitations. One of them is that PMI is sensitive to the rarity of outcomes. It can give a large value for the association between two rare outcomes, which may not be meaningful. In this case, weighted PMI (wPMI) is often used.

$$\begin{aligned} wPMI(x,y) &= p(x,y)\ PMI(x,y) \\ &= p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \end{aligned}$$

1.4.3.4 Co-information (interaction information)

Co-information (CoI), also known as interaction information with a flip sign, is a generalization of mutual information for three variables, which measures the relationships in terms of information among them. For random variables X, Y and Z, the positive co-information is called redundancy, which measures the amount of information shared by three variables. The negative co-information (or positive interaction information) is call synergy, which measures the amount of additional information about Z that can be obtained by knowing X and Y together compared to knowing X and Y separately. In other words, redundancy measures how well three variables covary. While synergy measures how one variable changes (modulates) the relationship between the other two. It is a measure of the degree to which the three variables interact with each other, rather than just co-varying.

The formula of co-information is:

$$CoI(X;Y;Z) = MI(X;Z) + MI(Y;Z) - MI([X,Y];Z)$$

Since co-information is symmetric, it can be also written as

$$\begin{aligned} CoI(X;Y;Z) &= MI(X;Y) + MI(Z;Y) - MI([X,Z];Y) \\ &= MI(Y;X) + MI(Z;X) - MI([Y,Z];X) \end{aligned}$$

1.4.3.5 Partial information decomposition

Partial information decomposition (PID) is a framework to analyze the information conveyed by multiple variables $\boldsymbol{R} = \{R_1, R_2 \dots R_n\}$ about a target variable S (Williams & Beer, 2010; R. Ince, 2017). PID decomposes the total information conveyed by $\boldsymbol{R}$ into the sum of unique, redundant and synergistic information subsets. For 3-variable interaction (which is the state-of-the-art), PID decompose $MI([R_1\ R_2];\ S)$ into $Unq(R_1) + Unq(R_2) + Red + Syn$, where Unq denotes unique information conveyed by the variable about S, and Red and Syn denote redundant and synergistic information respectively. Unlike CoI that can only be either redundancy or synergy, PID says redundancy and synergy can exist at once and CoI = redundancy – synergy.

The main aim of PID is to deal with the problem of CoI that redundancy and synergy are confounded. When any two among three variables are independent, redundancy doesn't exist, and therefore there is no confusion between redundancy and synergy. In this case, the classic CoI is safe to use. Otherwise PID can provide a better understanding of information in system.

1.4.3.6 Conditional mutual information

Conditional mutual information (CMI) is simply the MI that considers the effect of other variables in the system. The CMI between random variables X and Y conditioned on variable Z is defined as:

$$CMI(X;Y|Z) = MI(X;Y) - CoI(X;Y;Z)$$

It is noteworthy that the explanation of CMI is different depending on CoI since CoI can be either positive or negative. When CoI is positive, $CMI(X;Y|Z)$ measures shared information between X and Y removing the part that also shared with Z. When CoI is negative, $CMI(X;Y|Z)$ measures shared information between X and Y adding the synergy with Z.

1.4.3.7 Transfer entropy

Transfer entropy (TE), also known as directed information (DI), is a measure of the time lagged relationship between two variables (Schreiber, 2000). It is defined as conditional mutual information between variables X in earlier time and Y in later time conditioning out the Y in earlier time. That is, the information about Y provided by earlier X that is not

included in earlier Y. Transfer entropy can be thought as the information communicated from earlier X to later Y.

1.4.3.8 Directed feature information

Directed feature information (DFI) is an extension of DI. It is defined as

$$DFI = DI - DI|F$$

DFI measures the causal transfer of information specific to a stimulus feature (i.e., variable), which can be used to determine the content of communications (R. A. A. Ince et al., 2015).

2 Brain networks compute low-dimensional categorization-relevant feature manifolds that support behavior

2.1 Summary

To interpret our surroundings, the brain uses a visual categorization process. Current theories and models suggest that this process comprises a hierarchy of different computations that transforms complex, high-dimensional inputs into lower-dimensional representations (i.e. manifolds) in support of multiple categorization behaviors. In this chapter, I tested this hypothesis by analyzing these transformations reflected in dynamic MEG source activity while individual participants actively categorized the same stimuli according to different tasks: face expression, face gender, pedestrian gender, vehicle type. Results reveal three transformation stages. At Stage 1 (high-dimensional, 50-120ms), occipital sources represent both task-relevant and task-irrelevant stimulus features; task-relevant features advance into higher ventral/dorsal regions whereas task-irrelevant features halt at the occipital-temporal junction. At Stage 2 (121-150ms), stimulus feature representations reduce to lower-dimensional manifolds, which then transform into the task-relevant features underlying categorization behavior over Stage 3 (161-350ms). The findings detailed in this chapter shed light on how the brain's network mechanisms transform high-dimensional inputs into the specific feature manifolds that support multiple categorization behaviors.

2.2 Introduction

Despite the intricate and detailed nature of the visual input, our ability to categorize relies on extracting the essential elements of this information—i.e. the features that are crucial for the task at hand. For example, whereas categorizing the scene in Figure 2-1A as a "happy face" requires processing the mouth of the central face, categorizing this same picture as a "SUV" requires processing the shape of the right-flanked vehicle, or the left "female pedestrian" with the bodily features that disclose its gender, and so forth. The key point is that a single input image, and even a single object within this image, typically affords multiple different categorization behaviors (e.g. "happy," or "female" for the same central face). When our brain categorizes these input images or objects, it doesn't just passively treat them as unitary wholes. Instead, current theories and models suggest that brain networks actively transform their representations of the complex input images into task-specific image subspaces characterized by distinct geometric structures—i.e. diagnostic

feature manifolds (Cichy & Kaiser, 2019; De Melo et al., 2022; K. Kay et al., 2023; Naitzat et al., n.d.; Schyns et al., 2002, 2022; M. L. Smith et al., 2012; Zhan, Ince, et al., 2019). Hence, the actual visual information processed by the brain for categorization does not encompass the entirety of the stimulus images or objects presented in the visual field, but rather, it only comprises a geometrical subset of them, and critically, this subset varies depending on the task at hand. Here, I am testing this fundamental hypothesis, by reverse engineering, at a system level, the dynamic brain networks that actively transform identical input scene images for distinct categorization behaviors.

Significant progress in understanding visual categorization resulted from accurately mapping the brain regions that respond to various categories of images, e.g., those of faces, bodies, objects and scenes (Bracci & Op De Beeck, 2023; DiCarlo & Cox, 2007; Grill-Spector & Weiner, 2014). These regions comprise primarily the occipito-ventral/dorsal pathways that respond to different image categories, from their early split projection in left and right occipital cortex to their later categorical/semantic representations in the right fusiform gyrus, including how feedback reverses this flow to predict the input stimulus (Friston, 2010b; Lawrence et al., 2019; Yuille & Kersten, 2006). Though this approach proved invaluable to investigate where and when different brain regions are involved with processing full images, it overlooked how the task itself changes the actual feature manifolds that are processed by the brain. That is, how does categorizing the scene as “happy face,” or “SUV” or “female pedestrian” differently transform this fixed input into the specific feature manifolds that support task behavior?

Research into eye movements (Henderson & Hayes, 2017; Malcolm et al., 2014), attention (Brignani et al., 2010; Carrasco & Barbot, 2019), reverse correlation (Gosselin & Schyns, 2001; M. L. Smith et al., 2012; Zhan, Ince, et al., 2019) and neural network modelling (Schyns et al., 2022) suggests that categorization mechanisms in capacity-limited systems actively and flexibly transform the representation of high-dimensional input images into the low-dimensional feature manifolds that specifically support different categorization behavior (e.g. the smiling mouth feature manifold for responding “happy”, or the car shape manifold for responding “car” from the same image in Figure 2-1A), guided by frontal-parietal network mechanisms (Shashidhara et al., 2019). Critically, what emerges is an active process whereby brain networks process feature manifolds that are not inherently given, but instead dynamically extracted from the image depending on the participant’s categorization task and their individual strategy.

To track such transformations of feature manifolds into neural responses requires a broad, systems-level approach with fine-granularity control of the stimuli. Stimulating with categories of uncontrolled faces, cars and pedestrians full images, as is typical (Russakovsky et al., 2015; Rust & Movshon, 2005; Schrimpf et al., 2018), makes it practically unfeasible to precisely track where, when and particularly how the individual brain transforms the representation of these images (and objects within them) into the specific feature manifolds needed for categorization behavior–i.e. in the finite time of a neuroimaging experiment. Instead, these features hide behind the notorious "wall of [image] pixels (De Melo et al., 2022; Schyns et al., 2022)." To break through this wall, and reveal the feature manifolds used for categorization, I applied the Bubbles procedure (Gosselin & Schyns, 2001; Schyns et al., 2002). Bubbles randomly samples, with Gaussian apertures, the pixels of a stimulus image that each participant then categorized in four different ways–i.e. as *face expression*, *face gender*, *pedestrian gender* and *vehicle type* (see Figure 2-1A). Bubbles ensures that the participant can only correctly categorize the images when the randomly sampled pixels show the features needed for categorization (Gosselin & Schyns, 2001; Schyns et al., 2002). With such control, we could reverse engineer (1) the feature manifolds that each participant processes for categorization behavior in each task (Jack & Schyns, 2017) and critically, (2) where (i.e. which networks of brain regions), when (i.e. during which time windows) and how (i.e. with what transformations) the activity of 5,107 cortical MEG sources (every 1.67 ms between 0 to 450 ms post stimulus) transformed the representation of the same images into distinct task-specific feature manifolds that support behavior.

To preview the findings, the categorization task modulates internal representations of the visual input and their transformations over three systems-level Stages. At Stage 1 (high-dimensional, 50-120ms), occipital sources represent more stimulus features than the task requires–i.e. including opponent source-level representations (Buchsbaum et al., 1983; Graham & Wolfson, 2004; Popivanov et al., 2016; G. Rhodes et al., 2013) of a feature when it is task-relevant vs. irrelevant. While task-relevant features advance into ventral/dorsal pathways (Cichy et al., 2014; K. N. Kay et al., 2015; Kietzmann et al., 2019; Kriegeskorte et al., 2008; Margalit et al., 2020), irrelevant ones are halted at the occipital-temporal junction. At Stage 2 (high-to-low dimensional, 121-150ms), occipital sources reduce most irrelevant features, while ventral-dorsal pathways represent manifolds that keep transforming over Stage 3 (low-dimensional 161-350ms) into the task-relevant features (Frangou et al., 2019; Hanks & Summerfield, 2017; Jaworska et al., 2022; Ratcliff et al., 2009; Shashidhara et al., 2019) underlying categorization behavior (e.g. smiling mouth in "happy" vs. car features in "car"). Furthermore, I show that Pre-Frontal Cortex (PFC)

interacts with ventral/dorsal pathways early on (during Stages 1-2, from 71-95 ms post-stimulus), to guide stimulus feature representations based on their task-relevance.

2.3 Results

2.3.1 Experiment

The experiment comprised four 2-Alternative-Forced-Choice (AFC) categorization tasks applied to the same 64 base images of a realistic, complex city street scene (see Figure 2-1A). These images comprised varying embedded targets–i.e. 8 (4 male + 4 female) different face identities (Dailey et al., 2001) x 2 expressions x 2 vehicles x 2 pedestrian). Each participant (N = 10, within-participant statistics) performed each 2-AFC task in different blocks of 1,536 trials (i.e. precision neuroscience, with dense sampling (Poldrack, 2017)). Figure 2-1A illustrates, with two examples of the base images, the combinatorics of stimulus and 2-AFC task-response differences (i.e. face expression, face gender, pedestrian gender and vehicle type).

Each trial started with a fixation cross presented in the middle of screen for a random time interval between 500-1000 ms, followed by one base image for 150 ms, whose pixels were randomly sampled with the Bubbles procedure (Gosselin & Schyns, 2001; Schyns et al., 2002), see Figure 2-1 and *Methods, Stimuli*. As explained, Bubbles sampling ensures that the participant can only correctly categorize the stimulus when the random samples reveal by chance the pixels of the features that the participant requires to resolve the task. For example, the randomly sampled pixels of trial *n* in Figure 2-1 would enable categorization of "happy" in the *face expression* block, but not categorization of "SUV" in the *vehicle type* block, and vice versa with the samples of trial *m* (see *Methods, Task Procedure*). Critically, to eliminate typical low-level effects when different stimuli are associated with different categorization tasks (e.g. of different contrast energy profiles in different categories), here the set of randomly sampled stimuli was identical in each participant and blocked task (with stimuli presented in a random order in each block). Furthermore, to control the active engagement of cognitive processes from stimulus onset to categorization response, I kept the retinal locations of the image components constant across the experiment. This enabled processes from predictions of the spatial locations of task-relevant features, to discrimination of their attended pixel contents for decisions–e.g. the vehicle was always presented at a right image location, with "car" vs. "SUV" pixel contents to categorize. On each trial, the

participant's dynamic MEG activity (localized with a beamformer to 5,107 cortical sources, see *Methods, MEG*) and categorization responses are concurrently recorded.

2.3.2 Behavior: Task-relevant feature manifolds

To reconstruct the categorization-feature manifolds that support task performance (see Table 2-3 and Table 2-4), in each participant, I quantified the cross-trial relationship between pixel presence vs. absence due to random sampling (Figure 2-1C) and corresponding behavioral correct vs. incorrect in each categorization task—computed with Mutual Information (MI) (Cover & Thomas, 2012; R. A. A. Ince et al., 2017) as MI(pixel visibility; correct vs. incorrect categorization), maximum statistics (T. E. Nichols & Holmes, 2002; T. Nichols & Hayasaka, 2003), controlling the Family Wise Error Rate (FWER) over 101,920 pixels at $p < 0.05$, see *Methods, Analyses, Participant features*.

Figure 2-1C shows that participants use different features per task from an identical set of sampled images—e.g. mouth features to categorize *face expression*; left and/or right eye features for *face gender*; body parts for *pedestrian gender*; different features for *vehicle type*. Importantly, different participants often use different features to classify the same object with the same labels—e.g. Figure 2-1C illustrates that participant 1 uses the windshield and a large portion of the front fender and bonnet to classify *vehicle type* as "car" or "SUV" whereas participant 10 uses the shape of the alloy wheel and the car badge on the hood to produce the same category labels (Figure 2-7 develops all these results per participant). This demonstrates that a similar stimulus-response relationship across participants (or participant and models) does not warrant internal processing of the same stimulus features.

However, with these low-dimensional manifolds of categorization features now identified in each participant, I can uniquely examine how their brain transforms the same high-dimensional stimuli into the specific low-dimensional feature manifolds to enact behavior in each task.

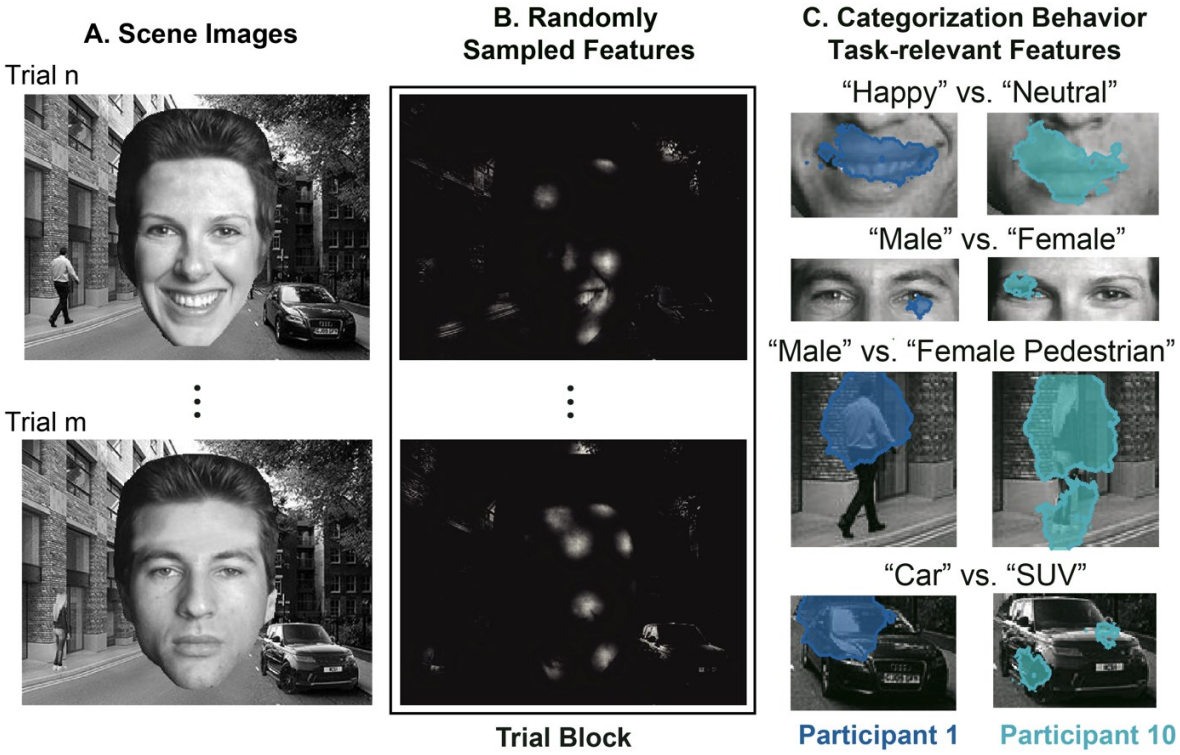

Figure 2-1 Categorization design and task-relevant features. *A. Scene Images.* I used 64 original images of a street scene that comprises a central face, to its left, a pedestrian on a sidewalk, to its right a parked vehicle. *B. Randomly Sampled Features.* On each trial, Bubbles randomly sampled the pixel-composite features of one of the 64 original images to synthesize a sampled stimulus. I used the same set of sampled stimuli (Gosselin & Schyns, 2001) presented in a random order in each categorization task, so that each participant (N = 10) saw each sampled stimulus image 8 times (twice per task). *C. Categorization Behavior; Task-Relevant Features.* The stimulus set afforded two different categorization responses in four different two-alternative forced choice categorization tasks: *face expression*, "happy vs. neutral" responses; *face gender*, "male vs. female"; *pedestrian gender*, "male vs. female"; *vehicle type*, "car vs. SUV." *Task-relevant features (color-coded for Participants 1 and 10, see* Figure 2-7 for all participants*).* For each participant, I computed MI(pixel-visibility; correct vs. incorrect categorization) (R. A. A. Ince et al., 2017) for each image pixel to reveal the pixels that significantly (FWER corrected $p < 0.05$) modulate categorization accuracy–i.e. color-coded in example participants 1 and 10 in each categorization task to illustrate the key point that participants often use different (sometimes even mutually exclusive) features to produce the same responses (e.g. for "male" vs "female pedestrian", upper body in participant 1 and full body in participant 10).

2.3.3 Brain: Systems-level time-courses of task-dependent stimulus transformations

To identify the stages that convert the representation of high-dimensional input images to the low-dimensional feature manifolds underlying categorization behavior, I started with a data-driven analysis. This analysis computed the representation of each varying scene pixel across trials (due to random Bubbles sampling) into the corresponding variations of MEG source amplitude responses post-stimulus–i.e. computing MI(pixel-visibility, MEG_t), for each pairing of 61 x 47 image pixels and 5,107 cortical sources, across 271 time points. For

visualization, I summarized the results into short consecutive periods of source response: [50-70], [71-80], [81-90], [91-120], [121-140], [141-150], [161-200], [221-280], [291-350] ms post-stimulus. These periods cover five dynamic neural events involved with visual categorizations–i.e. C1 (Slotnick, 2018), occipital hemifield responses; P100, early attention/stimulus representation and ensuing N170 (Bentin et al., 1996, 2007; R. A. A. Ince, Jaworska, Gross, Panzeri, Rijsbergen, et al., 2016; Rossion et al., 2003; Rousselet et al., 2004; Schyns et al., 2007), faces/familiar object representations; N250 (Kaufmann et al., 2008) and P300 (Ratcliff et al., 2009), attention/decision mechanisms. This analysis visualizes the image pixels that dynamic brain activity represents within each period and transforms across periods.

Figure 2-2A summarizes the results, showing different dynamic transformations of the same stimuli in each categorization task (rows), where orange-to-yellow colors indicate number of participants whose sources represent this image pixel in each period; maximum = 9/10 participant, maximum a posteriori probability (MAP) [95% highest posterior density interval (HPDI)] estimate of the population prevalence (R. A. Ince et al., 2021) of the effect of 9/10 participant replications = 0.9 [0.61 - 0.99], see *Methods, Analyses, Global representation of image pixels in brain networks*; *Methods, Bayesian population prevalence.*

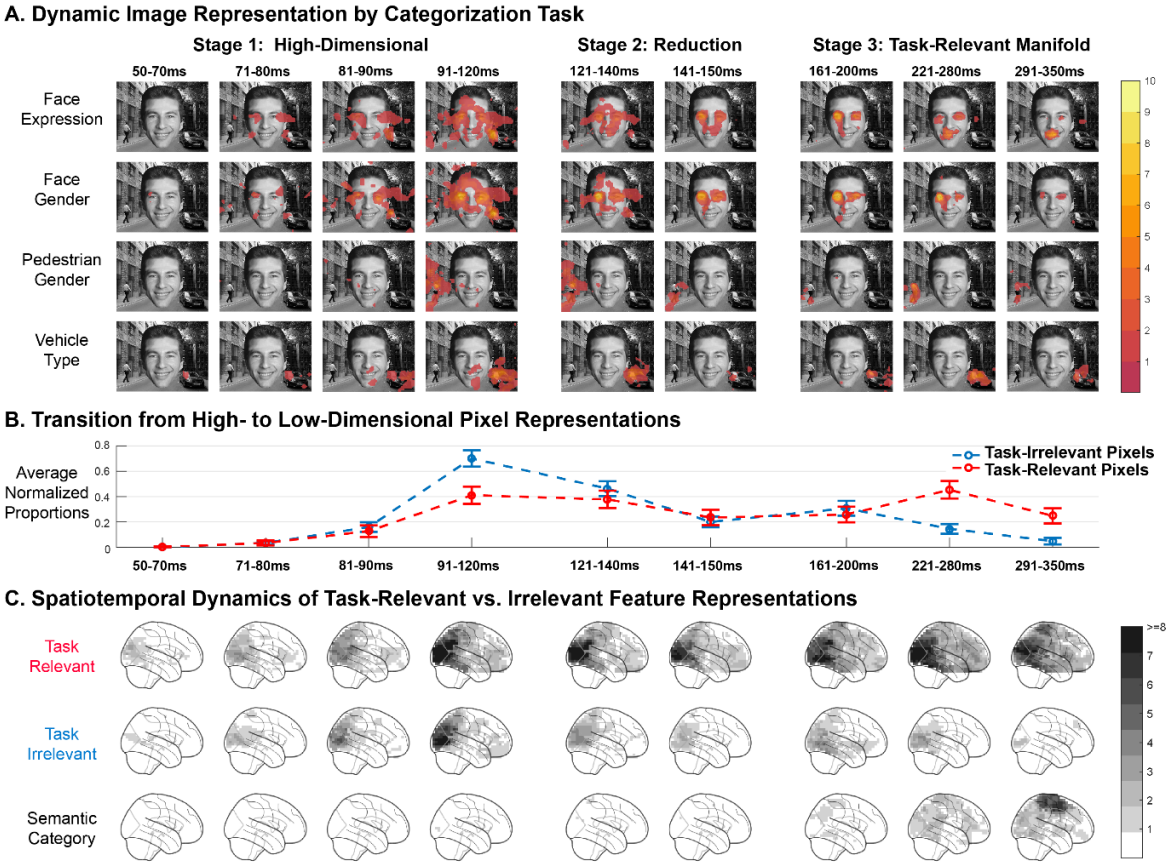

Figure 2-2 Systems-level transformations of images into categorization feature manifolds. *A. Dynamic stimulus representation by categorization task.* In each participant and MEG source, I computed the cross-trial relationship between each pixel's visibility and the MEG source amplitudes at time *t* post-stimulus–i.e. MI(pixel-visibility, MEG_t). I segmented the post-stimulus time course into 9 consecutive periods. In each period, and for each categorization task, I pooled pixels significantly represented on at least one MEG source, False Discovery Rate (FDR) test with q = 0.001. To visualize the transformations of stimulus representations across participants, I summarize their per-period results, by revealing the image pixels that their MEG sources represent in each task–i.e. where orange-to-yellow colors indicate the number of participants whose MEG sources represent this pixel, maximum = 9/10 participant, Maximum A Posteriori (MAP) [95% Highest Posterior Density Interval (HPDI)] prevalence (R. A. Ince et al., 2021) = 0.90 [0.61 - 0.99]). *B. Transition from high- to low-dimensional pixel representations.* In each period, I computed across participants and tasks the average number of task-relevant (vs. task-irrelevant) pixels–i.e. normalized per participant and task to the maximum task-relevant (vs. irrelevant) pixel numbers, with standard error bars. The resulting curves cross-over between Stages 1-2 and Stage 3 showing the transition from higher-dimensional inputs (with both task-relevant and irrelevant pixels) to lower-dimensional, task-relevant feature manifolds (with mainly task-relevant pixels). *C. Spatio-temporal dynamics of task-relevant vs. irrelevant feature representations.* For each participant, task and behavioral categorization feature (Figure 2-1C and Figure 2-7), I quantified how each source represents this feature (F) in its amplitude at each post-stimulus time point *t*–i.e. MI(F; MEG_t) (R. A. A. Ince et al., 2017), FWER corrected in each participant over sources and time points, $p < 0.01$. Greyscale sources reveal the number of participants that represented at least one feature in each period, when the feature is task-relevant vs. irrelevant, maximum = 8/10 participant, MAP [95% HPDI] prevalence (R. A. Ince et al., 2021) = 0.80 [0.49 - 0.96]. Category information provides a ground-truth reference of the MEG source representation of category information across participants—computed e.g. in *vehicle type* as MI(car vs. SUV stimulus; MEG_t), FWER corrected over sources and time points, $p < 0.05$, maximum = 8/10 participant, MAP [95% HPDI] prevalence (R. A. Ince et al., 2021) = 0.80 [0.49 - 0.96].

Considering each pixel as a stimulus dimension, each task shows MEG sources transitioning from an initially high-dimensional stimulus representation of large parts of the scene (Stage 1, 50-120 ms, periods 1 to 4) to a more focussed representation of only the task-relevant pixels–i.e. the lower-dimensional manifolds that develop between Stage 3, periods 7 to 9, 161-350 ms, compare with Figure 2-7. Stage 2 (periods 5 to 6) therefore marks the critical transition from higher-dimensional Stage 1 to task-relevant feature manifolds Stage 3.

To better formalize these transitions, I grouped image pixels as either task-relevant or irrelevant based on participant behavior–cf. Figure 2-1C and Figure 2-7 and *Methods, Analyses, Feature mask and visibility*. In Figure 2-2B, the red curve shows across different periods the number of task-relevant pixels while the blue curve shows task-irrelevant ones–i.e. computed as the cross-participant-and-tasks average of significant pixels in each period, reported with error bars. The cross-over of these curves between Stages 1-2 and Stage 3 identifies the transition from high-dimensional representations to task-specific feature manifolds.

2.3.4 Brain: Systems-level localizations of task-dependent stimulus transformations

To examine how the localized MEG sources represent and transform images based on task demands, I compared the source representation of an identical feature when it is task-relevant, or not–e.g. Participant 1's blue mouth in Figure 2-1C in *face expression* vs. in all other tasks. For each behavioral categorization feature (in Figure 2-7), I therefore determined the per-trial visibility score F, by intersecting this feature's pixels (e.g. Participant 1's blue mouth) with the pixels randomly sampled by Bubbles (cf. Figure 2-1, and *Methods, Analyses, Feature visibility*). I then quantified how variations of F across trials (Cover & Thomas, 2012; R. A. A. Ince et al., 2017) are represented in corresponding variations of MEG source amplitude responses–i.e. as $MI(F; MEG_t)$ (R. A. A. Ince et al., 2017), FWER corrected in each participant over sources and time points, $p < 0.01$ see *Methods, Feature representation on MEG sources.* Thus, the resulting MI curves (which I will develop in Figure 2-3, see Equation 1 in *Methods, Analyses*) are **not** curves of brain activity. Rather, they are curves of feature F representation into source magnetic field amplitude (MEG_t).

First, I show summary results in Figure 2-2C that reveal where (which sources) and when (which Stage/time period) features are transformed when they are task-relevant and used for behavior (Figure 2-2C, task-relevant) vs. task-irrelevant (Figure 2-2C, task-irrelevant). Each

glass brain displays the number of participants (gray levels) whose MEG sources represent at least one such feature as task-relevant or irrelevant–i.e. maximum = 8/10 participant, MAP [95% HPDI] prevalence (R. A. Ince et al., 2021) = 0.80 [0.49 - 0.96]. For reference, I also show how these MEG sources represent category information for at least one task, computed for instance for *vehicle type* as MI(car vs. SUV stimulus; MEG_t), FWER corrected over sources and time points, $p < 0.05$, plotted again as number of participants, maximum = 8/10 participant, MAP [95% HPDI] prevalence (R. A. Ince et al., 2021) = 0.80 [0.49 - 0.96].

Figure 2-2C reveals that occipital cortex sources represent task-relevant and task-irrelevant features during Stage 1, accounting for its higher dimensionality. However, there is already an effect of task because early representations distribute around the locations of the task-relevant features in the image but their effect sizes (i.e. MI) are weaker for surrounding pixels (see Figure 2-2A). Task-relevant features move along the ventral/dorsal pathways, but if they are irrelevant, they halt at the occipital-temporal junction. Occipital sources reduce most task-irrelevant features while ventral/dorsal sources form a lower-dimensional feature manifold during Stage 2. During Stage 3, the ventral/dorsal pathways keep transforming the low-dimensional feature manifolds into task-relevant features (at Stage 3, period 9). And category information is increasingly represented from 161-280 ms (at Stage 3, period 7 and 8), peaking at the parietal-frontal juncture post ~291ms (at Stage 3, period 9).

2.3.5 Brain: Systems-level expansion of task x feature transformations

Figure 2-3 expands the summary of Figure 2-2C, by displaying the representational dynamics in a grid with feature on the rows and task on the columns. Each panel shows the cross-participant average curves of significant feature representation–i.e. MI(F; MEG_t)–every 2 ms, for each color-coded MEG source, progressing from cyan (occipital) to yellow (frontal) as indicated in the reference glass brain.

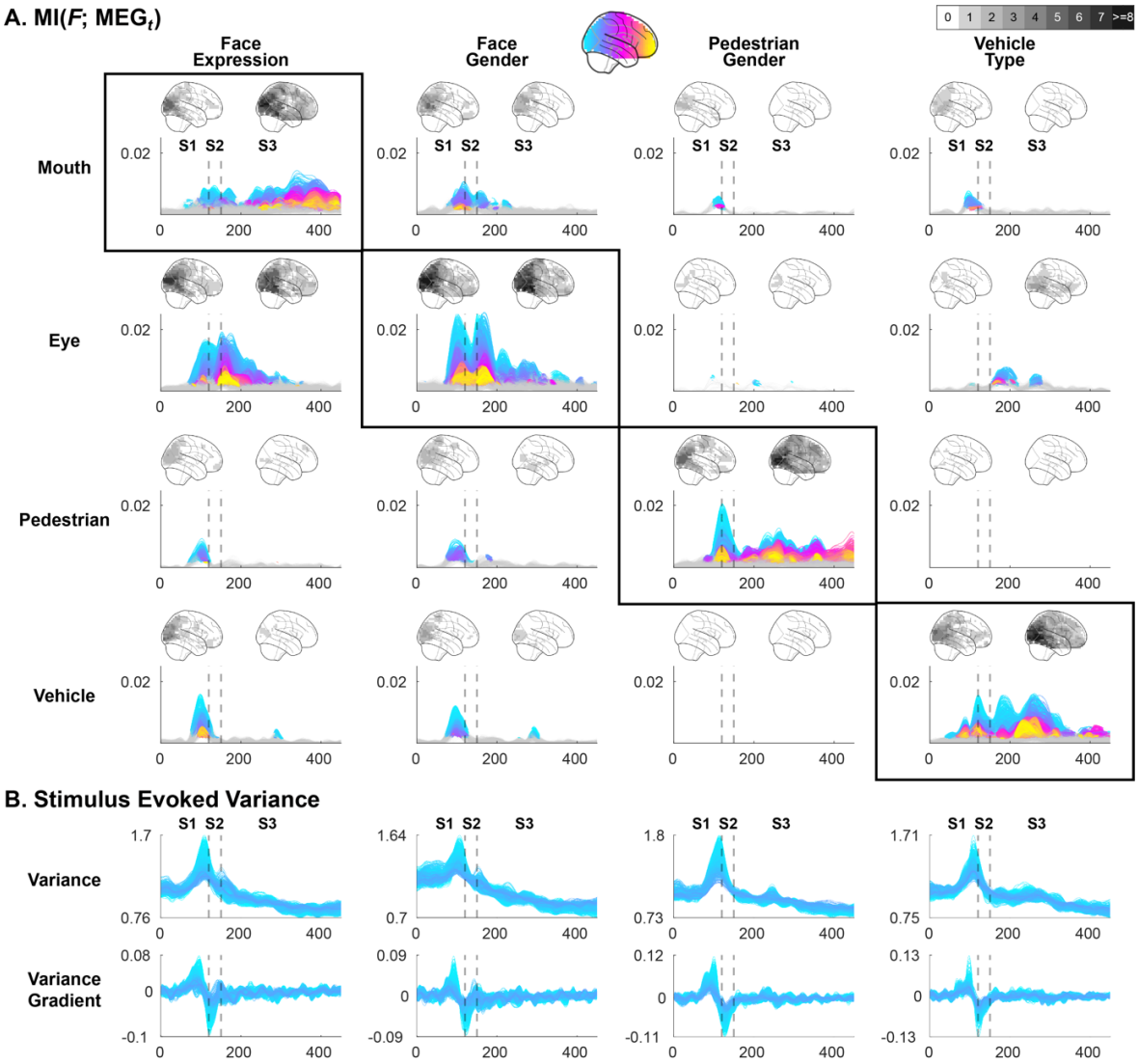

Figure 2-3 Dynamic representations of stimulus features across categorization tasks. *Dynamic representation of stimulus features (rows) in categorization tasks (columns).* Curves in each cell show the average (n = 10 participants) time-course of significant feature representation—of each participant's feature shown in Figure 2-1C and Figure 2-7, computed as MI(*F*; MEG$_t$) (R. A. A. Ince et al., 2017)—on MEG sources, each color-coded by its location on a posterior-to-anterior axis (cyan-yellow)—FWER $p < 0.01$, permutation maximum statistics per participant. Dashed lines (at 120 and 150 ms) delineate stages S1 to S3 reported from Figure 2-2. Small brains flanking the dashed lines(Bentin et al., 1996, 2007; R. A. A. Ince, Jaworska, Gross, Panzeri, van Rijsbergen, et al., 2016; Rousselet et al., 2004) show the participant prevalence of feature representations 50-150 ms (left brain) and 150-450 ms (right brain) post-stimulus, Bayesian maximum a posteriori (MAP) population prevalence(R. A. Ince et al., 2021), n = 10 participants, p = 0.01. Each row reveals qualitatively different representation dynamics of the same stimulus feature when it is task-relevant (matrix diagonal, highlighted with a box) vs. task-irrelevant (off diagonal). *B. Stimulus evoked variance and gradient* of MEG occipital source signal (cyan colored) in stages S1 to S3, averaged per source across participants.

The diagonal of Figure 2-3A shows the transformation of task-relevant features (*Fs*) through Stage 1 to 3 (black box highlights, e.g. vehicle feature in *vehicle type*; dashed lines delineate stages S1 to S3), with representations progressing from occipital to higher-level regions,

seen in the cyan-to-yellow occipital-to-frontal time-courses. Off diagonal plots, on the other hand, display short-lived representations of task-irrelevant features confined to occipital cortex (evident in cyan curves, cf. vehicle feature, the first three columns of the fourth row). Small brains localize the within-participant inference of these divergent representations across participants, during Stage 1-2 (left) and Stage 3 (right), FWER $p < 0.01$, permutation test, maximum statistic, see *Methods, Feature representation on MEG sources*. An exception is the central face's eyes, which remain represented in the two face tasks, consistent with previous studies (Schyns et al., 2003, 2007; M. L. Smith et al., 2004). This observation will be revisited in the Discussion.

The comparison in Figure 2-3A between features when they are task-relevant (diagonal) vs. irrelevant (each row, off-diagonal) is noteworthy. It shows similar initial representations in the occipital sources (cyan) for the same feature. However, by Stage 2, these representations diverge, with the same feature reduced in occipital cortex vs. passed into ventral/dorsal pathways. Next, I develop the representation mechanisms behind this divergence at the level of individual occipital sources.

2.3.6 Brain: Source-level representation of task-relevant vs. irrelevant F

First, I draw attention to the higher-dimensional feature representations on cyan occipital sources (Figure 2-3A) which align at Stage 1 with the peak cross-trial variance of their evoked MEG responses (Figure 2-3B). During 120-150ms Stage 2, this variance drops (negative gradient in 3B), marking the time when occipital sources reduce task-irrelevant features while relevant features progress into ventral/dorsal pathways. Figure 2-9 further shows that occipital representations of a given feature peaks slightly sooner when it is task-irrelevant than when it is task-relevant (Wilcoxon rank sum test, $p < 0.001$, MI averaged at each time point across participants, tasks and sources). Now, I delve into how the variance of an occipital source marks the identical feature variations *F* as task-relevant (and passed for further processing) or task-irrelevant (and reduced in occipital cortex).

Figure 2-4A, illustrates the variations *F* of the vehicle feature visibility from participant 8 as disks with varying radii (representing relative feature visibility). Figure 2-4B shows the MI representation curve of *F* on an example cyan occipital source at Stage 1, when the feature is relevant (in *vehicle type,* Figure 2-4B, solid cyan MI representation curve for this source)

vs. irrelevant (in *all other tasks*, Figure 2-4B, dashed cyan MI curve). Figure 2-4C shows the corresponding ERF and variance underpinning these representations of *F*.

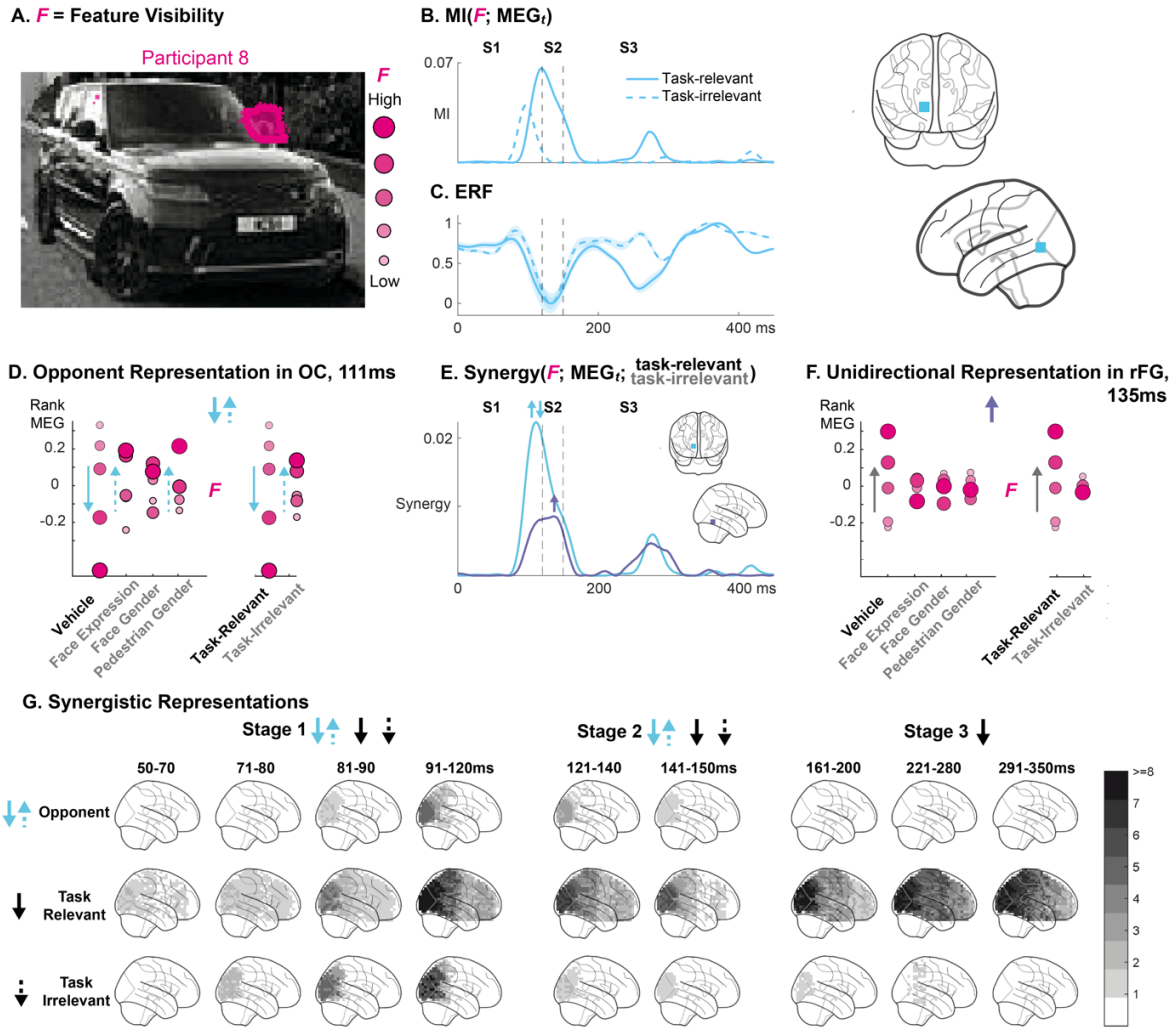

Figure 2-4 Task-modulations of feature representations. *A. Feature visibility, F.* Random stimulus sampling across trials varies visibility of vehicle feature *F*, represented as 5 varying radii (example from participant 8, Figure 2-7). *B. MI(F; MEG)* quantifies the dynamic representation of *F* in one example occipital source (located in small brains) when *F* is task-relevant (plain cyan curve) vs. task-irrelevant (dashed cyan curve). C. *Event-Related Field (ERF)* underlying the MI curves (panel B) for this source whose MEG amplitudes variations (shaded area, variance) represent *F*. *D. Opponent occipital source representations of F.* At 111 ms, the MEG amplitude variations of the same source (y axis) differently represent identical variations of feature *F* (circle radii, panel A) when it is task-relevant vs. irrelevant. Cyan arrows indicate these opposite representational directions when *F* is task-relevant (plain arrow, in *vehicle type*) and passed later into rFG vs. irrelevant (dashed arrow, other tasks) and reduced in occipital cortex. *E. The cyan synergy* curve quantifies the time course of these opponent representational interactions (Cover & Thomas, 2012; R. A. A. Ince et al., 2017) (that panel D illustrates at peak 111 ms, indicated with opponent arrows). The dark blue synergy curve illustrates another representational interaction in the rFG source shown in panel F (located in adjacent small brain). *F. Unidirectional Representations.* Dark blue rFG source represents *F* at 135 ms peak synergy, but here only when the feature is relevant in *vehicle type*. *G. Synergistic Representations*. Synergy(*F*; MEGt; Task-relevance) quantifies how brain sources differently represents identical *F* over Stages 1 to 3 depending on task-relevance vs. task-irrelevance, covering

three types of source-level representations. *Opponent synergy* indicates number of participants with significant opponent representations of the same *F* when task-relevant vs. irrelevant (cf. panel D); *Task-relevant (or irrelevant) synergy* indicates unidirectional representations of *F* when either task-relevant (cf. panel F) or irrelevant.

Figure 2-4D shows how *F* is differently represented based on its task-relevance. At Stage 1 (111 ms post stimulus), identical vehicle feature variations *F* (varying disk radii) exhibit MEG amplitude responses in opposite directions on the occipital source. Importantly, this depends on whether *F* is task-relevant (solid arrow in Figure 2-4D), and subsequently passed into the ventral/dorsal pathways vs. task-irrelevant (dashed arrow), and subsequently reduced in occipital cortex (i.e *vehicle type* vs. all other tasks). Figure 2-4E quantifies such opponent representations with information theoretic synergy(*F*; MEG_t; task-relevance vs. task-irrelevance), a double interaction that Figure 2-4D illustrates at its 111 ms peak (indicated with opponent cyan arrows in Figure 2-4E cyan curve). Task-synergy quantifies how the same feature is differently represented on the same source[2] depending on task, leading to different fates (i.e. passed vs. reduced).

Figure 2-4G shows that such opponent sources (cf. opposite cyan arrows) are mainly found in occipital cortex during Stage 1 and are consistent across participants (FWER $p < 0.01$ corrected over MI-significant sources * 271 time points, see *Methods, Analyses, Opponent feature representations*). And as previewed, the direction of these amplitude responses in Stage 1 could determine whether the feature will be reduced at Stage 2 or prominently represented in Stage 3 for behavioral responses.

In contrast, task-synergy can also indicate a feature that is unidirectionally represented either when it is task-relevant or irrelevant. Figure 2-4E displays the synergy curve of an example right Fusiform Gyrus (rFG) source, marked in purple, with a 135 ms peak (single arrow on the curve) during transition Stage 2. In Figure 2-4F, this rFG source unidirectionally represents the same vehicle feature *F*, but here only when it is task-relevant. Figure 2-4G extends this observation, illustrating across sources and time the count of participants who have at least one such exclusive task-relevant (or task-irrelevant) feature representation

[2] Although the sign of MEG responses is arbitrary(Gross et al., 2013), opponent signs reliably indicate the task-relevance vs. irrelevance of the same feature.

(FWER $p < 0.01$ corrected over MI-significant sources * 271 time points, see *Methods, Analyses, Task-relevant feature selection*).

2.3.7 Brain: Network interactions with prefrontal cortex modulate early source representations by task

Figure 2-4 shows that amplitude variations in occipital sources can represent the same feature differently by task relevance: either in opposite directions or unidirectionally. Here, I test the hypothesis that network interactions during Stages 1 and 2, specifically between Pre-Frontal Cortex (PFC) and the occipital-ventral/dorsal pathways, top-down modulate these early feature representations, to determine whether the same physical feature is passed into the ventral pathway for further processing, or instead reduced early within occipital cortex.

To investigate this, in each participant and task, I pinpointed two sources in the occipito-ventral/dorsal pathway during Stages 1 and 2: the source with strongest opponent representation of a given feature *F* (the synergistic "opponent seed" shown in Figure 2-5, color-coded by participant) and that with strongest unidirectional representation of *F* (the synergistic "unidirectional seed" also show in Figure 2-5). I then computed separately how opponent and unidirectional seeds interact will all PFC sources–i.e. by computing for each source pair the synergy(*F*; seed $source_t$; PFC $source_t$), separately for trials when *F* is task-relevant vs. *F* is task-irrelevant, FWER $p < 0.05$. Synergy emerges when two sources together predict more information about the feature than the sum of prediction by each source. Though PFC brain activity does not directly represent the feature, it does influence representation of the feature in the occipital-ventral/dorsal pathways. That is, when PFC activity changes, the relationship between feature visibility and activity in occipital-ventral/dorsal pathways changes. In this case, PFC sources and occipital-ventral/dorsal sources will generate synergy (the extra information about the feature that cannot be obtained from only occipital-ventral/dorsal sources without considering the PFC). Therefore, we need to explicitly consider PFC and occipital-ventral/dorsal activity together (synergistically) to understand the role of PFC on the occipital-ventral/dorsal representation of the feature.

This synergy analysis produced the four spatio-temporal maps per participant and task shown in Figure 2-5–i.e. opponent and unidirectional seeds x task-relevant and irrelevant feature conditions. It indicates where, when and how strongly each pair of PFC and occipito-

ventral/dorsal sources worked together as a network in representing feature *F*, separately for when *F* was task-relevant and irrelevant.

When *F* is task-relevant, Figure 2-5A shows that the orbitofrontal and ventromedial PFC (vmPFC) interact with both unidirectional and opponent seed sources during Stages 1 and 2 (96-150ms)–i.e. unidirectional seeds, maximum = 8/10 participant, (MAP) [95% HPDI] prevalence(R. A. Ince et al., 2021) = 0.80 [0.49 – 0.96]; opponent seeds, maximum = 7/10 participant, (MAP) [95% HPDI] prevalence (R. A. Ince et al., 2021) = 0.70 [0.38 – 0.90]). Critically, vmPFC interacts with occipital opponent sources primarily during Stage 2, when occipital cortex passes task-relevant features into the ventral pathway but reduces the features that are task-relevant. This suggests that vmPFC is involved with maintaining representations stimulus features when they are task-relevant across Stages 1 and 2, enabling their subsequent processing in the ventral pathway.

In contrast, when the same *F* is task-irrelevant, Figure 2-5B shows that the unidirectional occipital sources interact primarily with PFC orbitofrontal region (not with vmPFC), from early Stage 1 (71-95 ms); maximum = 9/10 participant, (MAP) [95% HPDI] prevalence(R. A. Ince et al., 2021) = 0.90 [0.61 – 0.99]. There is no clear PFC network interaction pattern for opponent seeds. These different network interactions for the representation of the same feature suggest that orbitofrontal PFC is primarily involved at Stage 1, with attentional or other mechanisms that mark the representations of task-irrelevant stimulus features for their subsequent reduction during Stage 2.

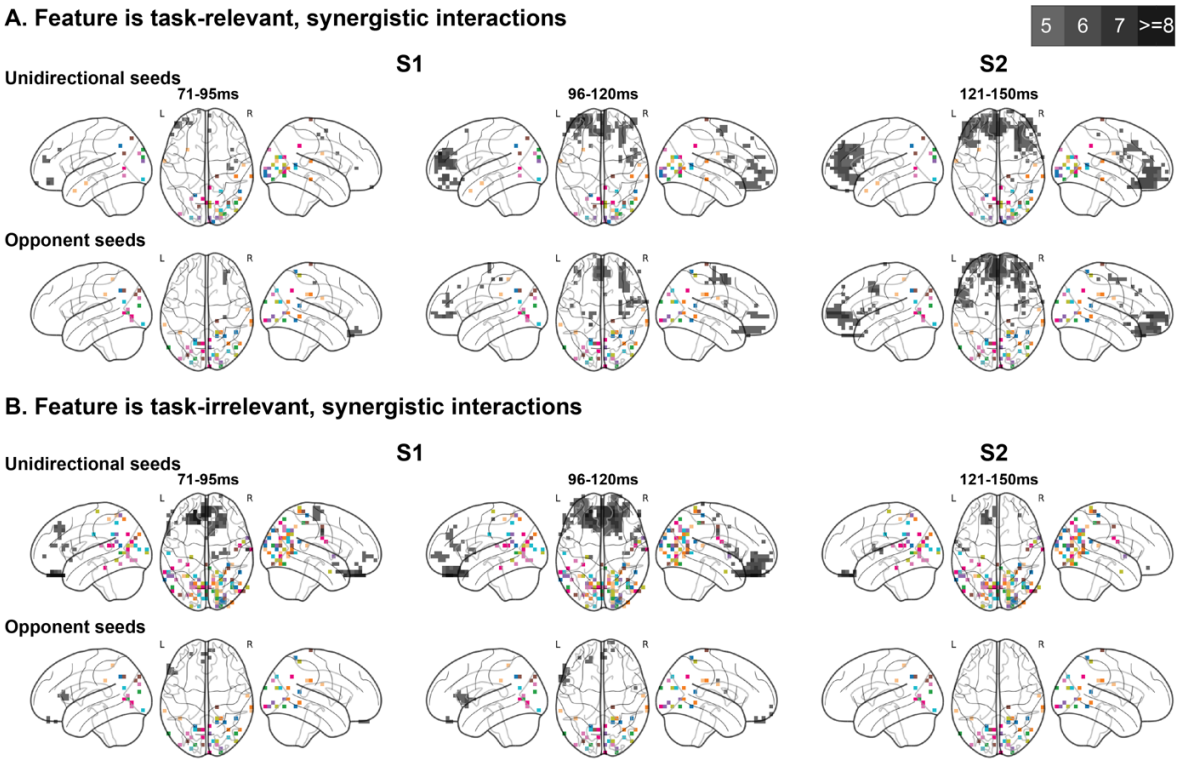

Figure 2-5 Early network interactions between PFC sources and occipito-ventral/dorsal sources. *A. Synergistic interactions when feature is task-relevant.* Unidirectional and opponent occipito-ventral/dorsal seed sources are color-coded by participant. Grey-levels indicate participant prevalence (>= 5) of synergistic interactions, computed as synergy(*F*; seed source$_t$; PFC source$_t$), revealing involvement of orbitofrontal and ventromedial PFC regions, from 96-120ms, permutation maximum statistics per participant, FWER $p < 0.05$. Unidirectional seeds, maximum = 8/10 participant, (MAP) [95% HPDI] prevalence(R. A. Ince et al., 2021) = 0.80 [0.49 – 0.96]; Opponent seeds, maximum = 7/10 participant, (MAP) [95% HPDI] prevalence(R. A. Ince et al., 2021) = 0.70 [0.38 – 0.90]. *B. Synergistic interactions when feature is task-irrelevant.* Unidirectional and opponent seeds synergistically interact mainly with orbitofrontal regions of PFC from 71-95ms, ending before the beginning of Stage 2 (121 ms). Unidirectional seeds, maximum = 9/10 participant, (MAP) [95% HPDI] prevalence(R. A. Ince et al., 2021) = 0.90 [0.61 – 0.99]; Opponent seeds, maximum = 6/10 participant, (MAP) [95% HPDI] prevalence(R. A. Ince et al., 2021) = 0.60 [0.28 – 0.83].

In sum, the network analyses show that different regions of PFC get involved with the early occipito-ventral/dorsal representations of stimulus features, depending on their task-relevance. Specifically, when a feature is task-relevant, orbitofrontal PFC and vmPFC guide its unidirectional and opponent representations during Stages 1 and 2 (~96-140ms), enabling the feature to progress from occipital into ventral/dorsal pathways for processing for behavior. In contrast, when the same physical feature is task-irrelevant, orbitofrontal PFC guides its unidirectional occipital representation at Stage 1 (~71-120ms), that occipital cortex then reduces from ~120ms. These distinct network interactions therefore suggest that PFC regulates how occipito-ventral/dorsal pathway transform representations of the same features based on their importance for the task at hand.

2.4 Discussion

At a systems-level, I studied where, when and how the brain networks of individual participants transform an identical set of high-dimensional input images into different low-dimensional manifolds of categorization features that support behavior in four different tasks—i.e. *face expression*, *face gender*, *pedestrian gender* and *vehicle type*. I revealed three stages that transform stimulus features into the task manifolds under the influence of pre-frontal cortex. During Stage 1 (50-120ms), higher-dimensional occipital sources represent more features than each task demands. If task-relevant, features advance into ventral and dorsal pathways; when irrelevant, their representations halt at the occipito-ventral junction. Occipital feature representations can be opponent or unidirectional. During Stage 2 (121-150ms), the representational focus narrows to low-dimensional task-relevant features manifolds because occipital cortex reduces most irrelevant features. During Stage 3 (161-350ms), the low-dimensional feature manifolds are further refined into the task-relevant features of decision behavior. Using precision neuroimaging and a dense-sampled design, I replicated these results in at least 8/10 participants, conferring a high Bayesian population replication probability (R. A. Ince et al., 2021) to these feature transformation mechanisms. Furthermore, top-down PFC influences modulate these task-dependent occipital representations from Stage 1. Specifically, orbito-frontal PFC is primarily involved with representations of task-irrelevant features at Stage 1 (~71-120ms) that occipital cortex reduces from ~100ms. When features are task-relevant, ventro-medial PFC is involved over Stages 1 and 2 (~96-140 ms) with unidirectional and opponent representations of task-relevant features that progress into ventral/dorsal pathways for behavior.

In both psychology and neuroscience, feature processing (Martínez et al., 1999; Noesselt et al., 2002) is foundational to numerous higher-level theories, spanning face, object and scene categorization and recognition (Humphreys, 2016; Nosofsky, 1986), as well as working (Baddeley, 2000; S. Rhodes & Cowan, 2018; van Moorselaar et al., 2018) and semantic memory (Grossman et al., 2002; Rubin, 2022) and even extending to conscious perception (Dehaene et al., 2014; Mashour et al., 2020). While feature processing encompasses feedforward and feedback communications across the occipito-ventral/dorsal and frontal-parietal-occipital networks, it is crucial to understand its role in interactive hierarchical models that disambiguate representations across layers (Friston, 2010b; Kietzmann et al., 2019; McClelland & Rumelhart, 1981; Yuille & Kersten, 2006). I align with this

understanding of a hierarchical interactive organization. For instance, a task like categorizing *vehicle type* should elicit specific information predictions (i.e. the participant's vehicle features), which would then flow downward through the hierarchy to the occipital cortex to interact with the incoming input. In this study, I demonstrated that distinct categorization tasks (e.g. *vehicle type* vs. *pedestrian gender*) invoked top-down PFC influences from Stage 1. These influences determined the relevance of the same physical feature in occipital cortex (and represented with opponent or unidirectional representations). It is crucial for categorization models, including Deep Neural Networks (DNNs), to replicate these three stages of dimensionality reduction, to yield similarly understandable feature manifolds for each categorization task (Schyns et al., 2022). Otherwise, while these models might predict category membership similarly to humans, the underlying features and transformations they compute might differ, see (Daube et al., 2021) for a solution to address this problem.

The critical first 150 ms

In the critical first 150 ms, the observations of the prevalent transformations from Stage 1 to Stage 2 during each categorization task largely align with classic early selection models of attention (Broadbent, 1957, 1958). There is an implementation whereby task-relevant features are channeled (or filtered in) through the ventral/dorsal pathways for further transformation and processing which eventually influences behavior. On the other hand, most task-irrelevant features, while initially represented, are subsequently reduced (or filtered out). The data indicate that interactions during Stage 1 particularly with the orbito-frontal and ventral-medial regions of the PFC, play a pivotal role in determining the trajectory of the same physical stimulus feature. This underscores potential constraints in how brain networks dynamically filter or allow features, especially in capacity-limited systems (Shiffrin & Gardner, 1972). Further research, especially those that delve into the finer granularities of neural responses (Huber et al., 2021; Lawrence et al., 2019; F. W. Smith & Muckli, 2010; Stephan et al., 2019), could shed more light on these mechanisms.

The two primary constraints underpinning the discussion are presented across Figure 2-2 to Figure 2-4. The first constraint emphasizes the sustained representation of the categorization feature manifolds in each task throughout the entire duration of information processing. A pressing question arises from this observation: How do gain functions and the recurrent/interactive activations in the cortical layers (specifically in V1-V4 and both the ventral and dorsal pathways) actively uphold these task-relevant feature representations from the moment of stimulus onset to the final behavior? The second constraint (cf. Figure 2-4)

notes that even when these same features are irrelevant in the task, the occipital cortex still briefly represents them. This brief representation leads us to consider the possibility of fusing individual MEG source amplitude data with fMRI cortical layer bold responses. Such a fusion could provide insights in how the inner and outer layers of the occipito-ventral (Huber et al., 2021; Margalit et al., 2020; Self et al., 2019) and dorsal cortex represent the same stimulus features, contingent on its task relevance. This approach may help elucidate how variations in layered cortical activity can result in opponent representations of the same feature, determining whether it is upheld or reduced. This distinction is pivotal to deepening the mechanistic understanding of the processes at play in Stages 1 and 2.

The 100-170 ms occipito-ventral/dorsal junction and subsequent visual categorizations

The task-dependent reduction and passing of stimulus features happen around the occipito-ventral and dorsal junction, before and after the timing and sources of the right occipito-ventral N/M170 Event Related Potentials (Bentin et al., 1996, 2007; R. A. A. Ince, Jaworska, Gross, Panzeri, van Rijsbergen, et al., 2016; Rousselet et al., 2004, 2014; Schyns et al., 2007) (ERPs). Past work showed that the N/M170 ERP reflects a network that communicates to the right fusiform gyrus the features contra-laterally represented in occipital cortices (R. A. A. Ince, Jaworska, Gross, Panzeri, van Rijsbergen, et al., 2016). The results suggest a reinterpretation of the N/M170. I showed that brain signal variance over the short (~50 ms) time window that precedes the N/M170 peak (cf. Figure 2-3B) reflects the junction during which brain networks transition from Stage 1 of high-dimensional stimulus representation to Stage 2 of lower dimensional processing of task-relevant feature manifolds. Transition to task-relevant categorization manifolds could explain why the N170 has been associated with multiple face, object and scene categorizations(Bentin et al., 1996, 2007; R. A. A. Ince, Jaworska, Gross, Panzeri, van Rijsbergen, et al., 2016; Rousselet et al., 2004). Developing further, Stage 2 transition is also when task-relevant features represented in left and right occipital cortices converge to the rFG (R. A. A. Ince, Jaworska, Gross, Panzeri, van Rijsbergen, et al., 2016; Zhan, Ince, et al., 2019), that seem to act as a buffer. Stage 3 processing could then integrate (Jaworska et al., 2022; Zhan, Ince, et al., 2019) these lateralized features into bi-lateral representations for multiple categorization behaviors. Results of decreasing lateralization of receptive fields (K. N. Kay et al., 2015) along the occipito-ventral pathway support such developments of bilateral representations.

Here again, further studies could fuse the precise temporal precision of MEG with the higher spatial resolution of fMRI (Finn et al., 2020; Huber et al., 2021; Lawrence et al., 2019; Self

et al., 2019), to better comprehend how the cortical layers of the ventral and dorsal pathways implement computations that integrate lateralized, buffered features into bilateral "stitched up," representations of the stimuli, pre- and post-170 ms, as shown with simpler stimuli and tasks (Jaworska et al., 2022; Schyns et al., 2009).

When these networks effectively categorize the stimulus is a fundamental question that relates to the visual information that is consciously perceived. Prevailing models(Dehaene et al., 2014; Mashour et al., 2020) suggest that stimulus features are "dispatched" to working memory for conscious perception. However, the data illustrates that categorization feature manifolds are maintained from occipital cortex to higher regions (R. A. A. Ince et al., 2015; R. A. A. Ince, Jaworska, Gross, Panzeri, van Rijsbergen, et al., 2016), potentially jointly acting as functional memory (Mashour et al., 2020) from ~100 ms post-stimulus until response. The features that constitute conscious perception(Schyns & Oliva, 1994, 1999) might align with the manifolds revealed at Stage 3, contrasting with the features that occipital cortex reduces at Stage 2 (see Figure 2-2 and Figure 2-3). This presents a tangible methodology to explore the complex landscape of conscious perception, including the influence of memory and prediction. The interplay between memory and categorization is evident as the feature manifolds of a categorization likely represent the predicted contents processed for categorization behavior when the stimulus appears. The findings also offer a robust framework to investigate the often-intangible contents of memory.

At this juncture, remember that I flagged that the eyes were processed in both face tasks, even when irrelevant in *face expression*. A similar result over the time-course of the N170 ERP is documented (Schyns et al., 2003, 2007), where the eyes are systematically represented, though not always necessary to judge the expression of a face. Others suggested that the first contact with a face is via the eyes (Niedenthal et al., 2010). The results of this chapter do indeed suggest that participants systematically represent the eyes and other face features in face categorizations tasks. These denser representations could explain why a deeper N170 ERP is often reported for faces (Bentin et al., 1996, 2007; Rousselet et al., 2014). Systematic representations of features spatially distributed across the face into the rFG could also explain its apparent "holistic representation" (Itier & Preston, 2018; Nemrodov et al., 2014; Richler & Gauthier, 2014).

The image is more broadly represented in the face tasks than in the other two tasks. This likely results from a combination of eye movements, inter-subject variability, cortical magnification, and attention. Specifically, the face is spatially broader in stimuli than the

pedestrian and vehicle. In the face tasks, participants likely attended to a larger region than in the pedestrian and vehicle tasks, which could also increase inter-subject variability. Besides, pedestrian and vehicle tasks are located in the two sides of the image, which leads to more eye movements. Together, these factors could have contributed to the broader image representation in the face tasks. However, these factors do not influence the conclusion because I aim to study the transformation of visual contents in natural vision, including the involvement of visual attention. Though the image representation is broader in face tasks, all four tasks demonstrated the same dimension-reduction transformations into task-relevant features over time.

Given the blocked task design, participants could use different decision-making strategies for the different blocks. However, because features obtained from reverse correlation cover different aspects of the image they are constrained to the same psychophysical scale between tasks. This is not problematic for rank-based information theoretic analysis.

I studied pervasive mechanisms that dynamically transform the same complex, high-dimensional input images for multiple visual categorizations. Within 150ms post-stimulus, the occipital cortex, under frontal guidance, either passes or reduces a feature based on its relevance in a categorization task, revealing opponent representational signatures at the MEG source-level. Following this, occipito-ventral and dorsal networks focus on the feature manifolds relevant to each categorization task. These feature transformations offer mechanistic insights into attention theories, face and object categorizations, and our understanding of conscious perception.

2.5 Methods

2.5.1 Participants

Ten participants (3 males and 7 females, age: M = 25.3, SD = 1.64, range = 23-28 years old) with normal or corrected to normal vision participated in all four tasks. All participants are right-handed. Gender and age were not considered in the study design and analysis. Participants were recruited via a database of participants at University of Münster. Informed consent was obtained from all participants. I designed and piloted the experiment. The formal experiment data was collected from University of Münster, Germany. The study was approved by the ethics committee of the University of Münster (2019-198-f-S) and conducted in accordance with the Declaration of Helsinki.

The number of participants and trials was determined based on statistical power estimation from the study employing within-participant statistics and population prevalence, which have demonstrated the ability to obtain robust and replicable effects (R. A. A. Ince et al., 2020, 2022).

2.5.2 Stimuli

I used 64 base greyscale images (8 face identities with 4 male and 4 female × 2 expressions × 2 pedestrians × 2 car) of a realistic city street scene comprising the combinations of varying embedded targets: a central face (which was male vs. female and happy vs. neutral), left flanked by a pedestrian (male vs. female), right flanked by a parked vehicle (car vs. SUV). The images were presented at 5.72° × 4.4° of visual angle, with 364 × 280 pixel size. I sampled information from each image, using the Bubbles procedure. Specifically, I multiplied the image with randomly positioned Gaussian apertures (sigma = 15 pixels) to vary the visibility of image features on each trial. I used 35 Gaussian apertures in all tasks, which was determined by a behavioral experiment pilot with 4 participants from Glasgow to trade-off the participant's performance among four tasks. I pre-generated 768 random bubble masks which were the same in all categorization tasks. On each session of trials, I applied the 768 masks to 12 repetitions of the original 64 images, for a total of 768 trials presented in a random order.

2.5.3 Task procedure

Each trial began with a fixation cross presented for a random time interval 500-1000 ms, followed by one of the original stimuli for 150 ms, whose features were randomly sampled with the Bubbles procedure. Participants were instructed to maintain fixation on each trial and respond as quickly and accurately as possible, by pressing one of two keys ascribed to each response choice—i.e. "happy" vs. "neutral" in *face expression*; "male" vs. "female" in *face gender* task; "male" vs. "female" in *pedestrian gender*; "car" vs. "SUV "in *vehicle type*. Each task comprised two sessions of trials, each comprising 768 trials (of 6 runs followed by a short break, each run comprising 128 trials = 8 identities × 2 expressions × 2 pedestrians × 2 cars × 2 repetitions).

2.5.4 MEG

Participants were seated upright in a magnetically shielded room while their MEG and behavior data were simultaneously recorded. Brain activity was recorded using a 275-

channel whole-head MEG system (OMEGA 275, VSM Medtech Ltd., Vancouver, Canada) at a sampling rate of 600 Hz. During MEG recordings, the head position was continuously tracked online by the CTF acquisition system. For MEG source localization, I obtained high-resolution structural magnetic resonance imaging (MRI) scans in a 3T Magnetom Prisma scanner (Siemens, Erlangen, Germany).

2.5.4.1 Pre-processing

I performed analyses with Fieldtrip (Oostenveld et al., 2010) and in-house MATLAB code, following recommended guidelines (Gross et al., 2013). I first visually identified noisy channels and trials with epoched data (-400 to 1500 ms around stimulus onset on each trial) high-pass filtered at 1 Hz (4th order two-pass Butterworth IIR filter). Next, I epoched the raw data into trial windows (-400 to 1500 ms around stimulus onset, 1-25 Hz band-pass, 4th order two-pass Butterworth IIR filter), filtered for line noise (notch filter in frequency space), applied fieldtrip build-in denoise function specific to the MEG system, and rejected noisy channels and trials identified in the first step. I then decomposed the data with ICA, and visually identified and removed the independent component corresponding to artifacts (eye blinks or movements, heartbeat).

2.5.4.2 Source reconstruction

I applied a Linearly Constrained Minimum Variance (LCMV) beamformer (Hillebrand & Barnes, 2005) to reconstruct the time series of 12,773 sources on a 6mm uniform grid warped to standardized MNI coordinate space. Using a Talaraich-Daemon atlas(Lancaster et al., 2000), I excluded all cerebellar and non-cortical sources, and performed statistical analyses on the remaining 5,107 cortical grid sources. I categorized cortical sources into four regions based on ROIs defined in the Talaraich-Daemon atlas (Lancaster et al., 2000).

Table 2-1 Cortical sources categorized into four regions of the Talaraich-Daemon atlas **(Lancaster et al., 2000)**.

Occipital region	Lingual gyrus (LG) Cuneus (CUN) Inferior Occipital Gyrus (IOG) Middle Occipital Gyrus (MOG) Superior Occipital Gyrus (SOG)
Temporal region	Fusiform Gyrus (FG) Inferior Temporal Gyrus (ITG) Middle Temporal Gyrus (MTG) Superior Temporal Gyrus (STG)
Parietal region	Superior Parietal Lobule (SPL)

	Inferior Parietal Lobule (IPL) Angular Gyrus (ANG) Supramarginal Gyrus (SMRG) Precuneus (PRECUN) Postcentral Gyrus (POSTCEN)
Frontal region	Anterior Cingulate (AC) Inferior Frontal Gyrus (IFG) Medial Frontal Gyrus (MeFG) Middle Frontal Gyrus (MiFG) Orbital Gyrus (OG) Paracentral Lobule (PL) Precentral Gyrus (PRECEN) Superior Frontal Gyrus (SFG)

2.5.5 Analyses

Feature representation

What is it?

Feature representation refers to a systematic relationship between a feature of the external world and neural activity (Baker et al., 2022; Poldrack, 2021). Our methodology quantifies the representation of a visual feature so that we can trace where, when and how the brain processes it.

How is a feature representation quantified?

In our data, the visibility of a feature in a stimulus varies in a continuous manner across trials–i.e. it is not a binary feature present vs. absent. To measure the representation of the feature into MEG activity, we use Mutual Information (specifically, the Gaussian Copula MI, GCMI) (R. A. A. Ince et al., 2017). GCMI quantifies across trials how strongly the variations of MEG amplitude represent the variations of feature visibility in the stimuli–i.e. as the information that MEG amplitude variations and feature visibility variations share, measured on the scale of bits.

For example, Figure 2-6A now plots the mean MEG amplitude response curves, where all trials are split into 5 equally occupied feature visibility bins–quintiles of the empirical CDF of feature visibility. Statistical difference between these mean curves is considered to reflect important processing differences across feature visibility conditions. Figure 2-6 clarifies that

the highest MI measure of feature representation corresponds to largest differences amongst mean MEG responses to the different bins of feature visibility.

The MEG amplitude curves evolve with peaks and troughs. These peaks and troughs can reflect representations of other features and/or cognitive variables. However, the feature representation curve underneath in Figure 2-6B does not mirror the MEG peaks because our information theoretic analysis specifically isolates, from raw MEG amplitude variations, the information that only pertains to the tested stimulus feature.

Feature manifolds

We used 'manifold' in its mathematical understanding, as a topological space that locally resembles Euclidean space. In neuroscience, 'neural manifold' is often used to refer to geometric structures in neural population activity–i.e. a subspace of neural state space.

We deliberately used 'feature manifold' to refer to the geometric structure of visual inputs (e.g. images) that are represented in neural activity. Object categorization relies on diagnostic features, which we show underline categorization behavior. However, a given object can be categorized in multiple different ways, each relying on distinct sets of diagnostic features. This implies that the brain must represent different stimulus feature manifolds for this object. This is often neglected in neuroimaging studies of visual categorization. We show that only a subspace of the 2D projection of the real-world (i.e. the image) is selected for categorization, in a task and participant-specific way.

Participant features

To reveal what image features each participant used to in each categorization task (i.e. the task-relevant features), I quantified the cross-trial statistical dependence between the visibility of each pixel due to bubbles sampling (Gosselin & Schyns, 2001) and the corresponding correct vs. incorrect categorization response of the participant in this task, computed as Mutual Information (Cover & Thomas, 2012; R. A. A. Ince et al., 2017), MI(pixel visibility; correct vs. incorrect categorization). I represented pixel visibility on each trial as a real number from 0 to 1 (low to high visibility), which I then binarized using a 0.2 threshold into 2 categories: 0 for low visibility and 1 for high visibility. To establish statistical significance, I ran a non-parametric permutation test with 1,000 shuffled repetitions, corrected over 101,920 (364 x 280) pixels using maximum statistics (FWER p

< 0.05). Significant pixels represent the participant's task-relevant features whose visibility influences their categorization behavior in each task (see Figure 2-1, Figure 2-2A and Figure 2-7).

Global representation of image pixels in brain networks

To visualize the global representational dynamics of the visual stimuli in each categorization task, I computed MI(pixel-visibility, MEG_t) for each one of the 364 x 280 stimulus pixels (downsampled to 61 x 47 for computational efficiency), 5,107 cortical MEG sources and 271 time points, producing a 3D matrix of MI values with dimensions of 2867 (61 x 47) pixels x 271 time points x 5,107 sources. MI quantifies the statistical dependence between two variables.

I then segmented the time dimension into nine intervals ([50-70], [71-80], [81-90], [91-120], [121-140], [141-150], [161-200], [221-280], [291-350] ms). To visualize the pixels that the MEG sources of each participant represent, I pooled all the pixels with statistically significant MI on at least one source on the considered interval. To compute this statistical significance in each participant, for each pixel, I took the maximum MI(pixel visibility, MEG_t) at each time point, resulting in a pixels x time matrix. I performed a FDR test on this matrix with a false discovery rate set at q = 0.001. For each image pixel, I color-coded the number of participants with such significant MI (maximum number = 9, Maximum A Posteriori (MAP) [95% Highest Posterior Density Interval (HPDI) prevalence = 0.90 [0.61 – 0.99]). Similarly, to visualize which MEG sources of the participant represent these pixels in each time interval, I pooled all sources with at least one significant MI(pixel visibility, MEG_t) in the interval. I reported the number of participants with a significant pixel To compute this statistical significance in each participant, for each source and time point, I took the maximum MI(pixel visibility, MEG_t) over all pixels, resulting in a sources x time matrix. I performed a FDR test on this matrix with a false discovery rate set at q = 0.1%.

Feature mask and visibility

As different participants can use different features in each task, to generalize analyses across participants, I transformed the data from levels of pixel visibility into levels of feature visibility (i.e. comprising the pixels making up the features of each participant). To this end, for each feature I selected the top 5% pixels with highest MI(pixel visibility; correct vs. incorrect categorization) to form feature masks. On each trial, I computed feature visibility

as the feature mask pixels shown by the bubbles sampling, weighted by the MI values of each pixel of the feature mask. I divided mouth (for face expression) and eyes (for face gender) features into their left and right components and considered them as a 2-dimensional feature variable in analyses. Figure 2-7 shows the feature masks of each participant and task.

$$Feature\ visibility = \sum_i MI(Pixel_i\ visibility; Behavior) \cdot Pixel_i\ visibility$$

Feature representation into MEG sources

To reconstruct where, when and how MEG sources represent each participant's features, I computed MI between the visibility of each feature and 5107 MEG source signals over 0 to 450 ms, in each task—i.e. when the feature is task-relevant, and also in the three other tasks when it is task-irrelevant, computed as MI(feature visibility; MEG_t) with GCMI as described above (R. A. A. Ince et al., 2017). To establish statistical significance, I ran a non-parametric permutation test with 1,000 shuffled repetitions, corrected (FWER $p < 0.01$) over 5107 sources x 271 timepoints with maximum statistics. This computation produces a 4 (tasks) x 4 (features) x 5,107 sources x 271 time points feature representation matrix for each participant.

Task modulation of feature representation on MEG sources

Synergy computes the difference between the overall representation strength of a feature (i.e. its visibility, F) in MEG activity, quantified across all trials ignoring the particular task (quantified with MI), and the average task-conditional representation strength (quantified with conditional MI):

$$Synergy(F; MEGt; Task) = MI(F; MEG_t \mid Task) - MI(F; MEG_t)$$

If the Task factor has no effect on representation, the above MI quantities will not differ resulting in zero synergy. Synergy results when the average representational strength of a feature is higher when controlling for task, meaning that we would better predict the stimulus from brain activity if we also knew the task performed.

To quantify the modulation effect of the four categorization tasks on the representation of the participant's features into MEG source activity, for each participant feature, I computed information theoretic synergy, as just defined, between 0 and 450 ms post-stimulus, where the categorization tasks variable has values of 1 to 4 to represent each task. To establish statistical significance, I use a nonparametric permutation test, with 1,000 repetitions, shuffling the task label of each trial, corrected over 5107 sources * 271 timepoints (FWER $p < 0.01$). This provides permutation samples from the null distribution where task does not affect feature representation.

Task-relevant vs. task-irrelevant

To quantify the specific modulation of task-relevance vs. irrelevance on the MEG source representation of each participant feature, I computed again synergy, this time as synergy(feature visibility; MEGt; task-relevance), where task-relevance could be 1 (for task-relevant) or 2 (for task-irrelevant). It was observed that synergy arose from two different representational mechanisms: Opponent feature representation and task-relevant feature selection. I define each below.

Opponent feature representations

Opponent feature representation on a given source means this: the same physical variations of feature visibility incur MEG amplitudes in opposite directions depending on whether this feature is task-relevant vs. task-irrelevant. Figure 2-6 illustrates this opposition in the shaded time window. We can see that the same changes in feature visibility give rise to MEG amplitude changes in opposite directions in the task-relevant and task-irrelevant binned MEG amplitude curves. Specifically, when the feature is task-relevant, the MEG amplitude response is more negative to higher feature visibility; in contrast, when the feature is task-irrelevant, the MEG amplitude response is more negative to lower feature visibility. It is important to note that this reversal refers to a difference in the sign of the correlation between feature and MEG–although we use MI, which is an unsigned measure. This reversal is not a statement about the sign of the evoked magnetic field. As shown in the example, there is a change in the sign of the correlation relationship, but the evoked MEG signal has negative sign in both cases. This implies that significant MI for F in multiple tasks, but their synergy reveals that this representational relationship depends on tasks.

I formalize this effect as the following logical conjunction:

Opponent feature representation:-

<significant task-relevant MI>
& <task-irrelevant MI>
& <significant synergy>
& <opponent signs for relevant vs irrelevant>

Task-relevant feature selection

Occurs when a given source represents a participant feature only when it is task-relevant. This synergy is logically defined as:

Unidirectional task-relevant feature representation:-

<significant task-relevant MI>
& <no significant task-irrelevant MI>
& <significant synergy>

Mutual Information (MI)

When testing whether a brain source is responsible for processing a stimulus feature, we conduct statistical tests to compare neural activities of the brain source (i.e. MEG amplitude in our case) under conditions where the feature is present versus absent. A significant difference in MEG amplitudes between two conditions indicates that the brain source is involved in processing that stimulus feature. In neuroscience, this systematic relationship between a particular feature of the external world and empirical data of neural activity is broadly called feature representation into neural activity. When feature visibility is not simply a binary state (present or absent), but rather exists on a continuous scale, we can quantify the relationship between feature visibility and MEG amplitudes using Mutual Information (MI) analysis. Figure 2-6 uses an example source to illustrate how the MI metric quantifies the relationship between feature visibility and MEG amplitude.

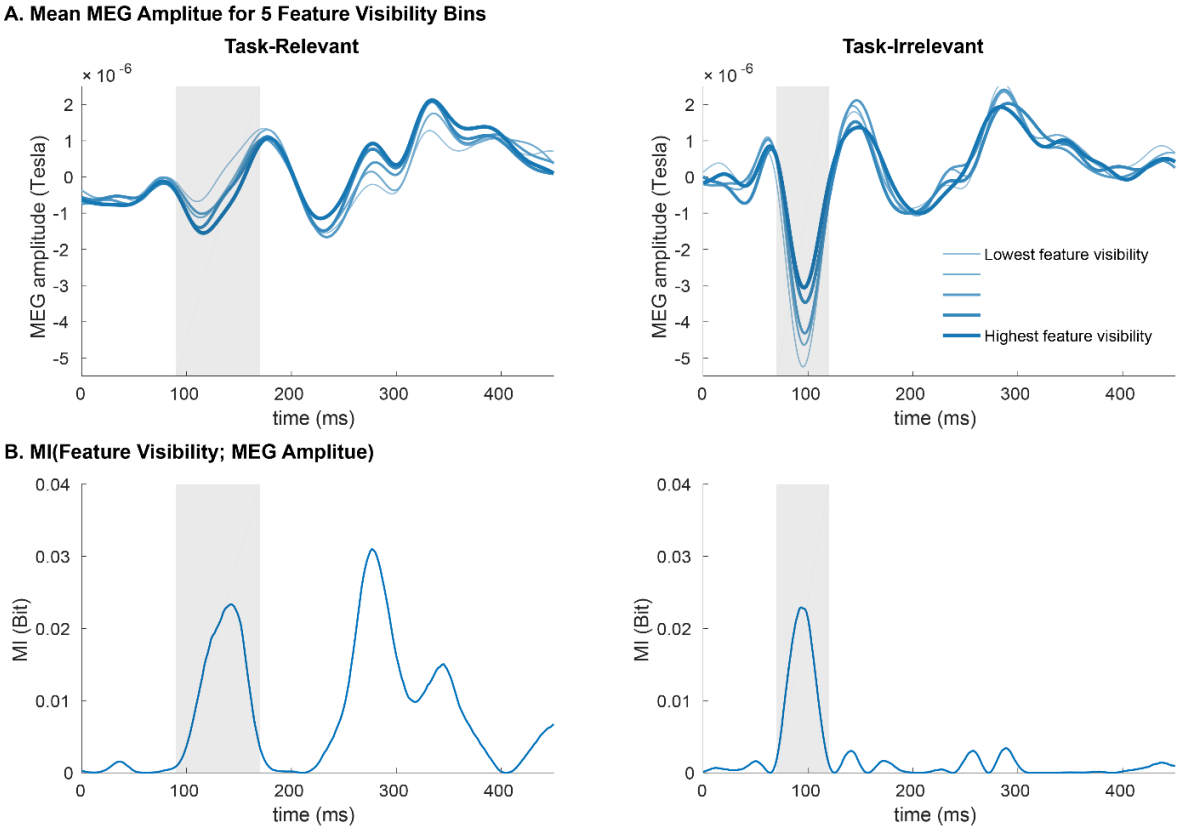

Figure 2-6 Illustration of Mutual Information (MI) and opponent representation. *A. Mean MEG amplitude for 5 feature visibility bins.* I split trials into 5 bins by feature visibility and plotted the mean MEG amplitude of an example occipital source for each visibility bins. In the period highlighted in grey, feature visibility was correlated with MEG amplitude. However, there was an opponent representation of the feature when it was task-relevant vs. task-irrelevant. Specifically, when the feature was task-relevant, higher feature visibility corresponded to more negative MEG amplitudes. In contrast, when the feature was task-irrelevant, lower feature visibility corresponded to more negative MEG amplitudes. *B. MI(Feature Visibility; MEG Amplitude).* MI therefore quantifies the relationship between feature visibility and MEG amplitude, which is broadly called feature representation in MEG amplitude in neuroscience.

Gaussian Copula Mutual Information (GCMI)

I calculate MI between the continuous valued pixel visibility (bubble mask value) and the continuous valued MEG amplitude at a given source and timepoint with Gaussian Copula Mutual Information (GCMI) (R. A. A. Ince et al., 2017). The empirical Cumulative Distribution Function (CDF) of the marginal distribution of each variable (pixel visibility and MEG) is estimated, and the data are transformed via the inverse CDF of a standard normal distribution. This results in a data set with perfect standard normal marginal distributions, but the same empirical copula as the original data. Then standard analytic expressions for bias-corrected Gaussian MI are used. As MI is invariant to marginal distributions, and Gaussian distribution has maximum entropy given constrained second

moments, this GCMI procedure provides a lower bound estimate of the true MI (R. A. A. Ince et al., 2017).

Bayesian population prevalence

Table 2-2 below provides a reference to transform the proportion of participants from the sample who have a significant effect into the Bayesian population prevalence(R. A. Ince et al., 2021). Population prevalence is a Bayesian estimate of the within-participant replication probability. Replicating a result in multiple participants offers a much higher standard of evidence than to declare statistical significance of a population mean effect. For example, $p = 0.05$ typically defines population mean statistical significance; $p < 0.001$ would be considered stronger evidence. In Figure 2-3A (diagonal of the matrix), I show 8/10 participants have significant MI task-relevant feature representations in occipital and ventral cortex (FWER $p < 0.01$). The frequentist p-value corresponding to this result under the global null that no one in the population shows this effect is $1.6x10^{-9}$. Under the global null the results are therefore 7 orders of magnitude more surprising than a typical mean demonstrating the experimental effect at the population level. Here, I report Bayesian estimates of the population parameter with their associated uncertainty. Given 8/10 participants significant at $p = 0.01$, we can be confident that the population replication probability is greater than 49%. I would expect the majority of the population to show this result if they were tested in the same experiment.

Table 2-2 Bayesian population prevalence: Maximum A Posteriori (MAP) [95% Highest Posterior Density Interval (HPDI)] for k significant participants out of 10.

	Within participant $\alpha = 0.05$	Within participant $\alpha = 0.01$
K=10	1 [0.75 - 1]	1 [0.75 – 1]
K=9	0.89 [0.61 – 0.99]	0.90 [0.61 – 0.99]
K=8	0.79 [0.49 – 0.96]	0.80 [0.49 – 0.96]
K=7	0.68 [0.38 – 0.90]	0.70 [0.38 – 0.90]
K=6	0.58 [0.28 – 0.83]	0.60 [0.28 – 0.83]
K=5	0.47 [0.19 – 0.75]	0.49 [0.19 – 0.75]
K=4	0.37 [0.11 – 0.66]	0.39 [0.11 – 0.66]
K=3	0.26 [0.05 – 0.56]	0.29 [0.05 – 0.56]
K=2	0.16 [0 – 0.44]	0.19 [0 – 0.44]
K=1	0.05 [0 – 0.34]	0.09 [0 – 0.34]

Table 2-3 reports the average categorization accuracy and reaction time performance (standard deviation in parentheses) in each task of the design (see Table 2-4 below for individual participants' data).

Table 2-3 Average accuracy and reaction times across participants in each categorization task.

	Face Expression	Face Gender	Pedestrian Gender	Vehicle
Accuracy	77.21% (1.45%)	75.12% (1.92%)	76.11% (1.56%)	65.25% (1.91%)
Reaction Time	753 ms (54.14)	738 ms (42.24)	779 ms (42.15)	978 ms (50.51)

Table 2-4 Per participant average accuracy and RT in each categorization task.

	Face Expression	Face Gender	Pedestrian Gender	Vehicle
Participant 1	70.31%	68.82%	75.20%	69.01%
Participant 2	78.06%	74.61%	73.57%	64.19%
Participant 3	83.59%	85.74%	80.99%	55.17%
Participant 4	76.63%	74.22%	67.84%	57.23%
Participant 5	76.30%	74.35%	71.74%	60.35%
Participant 6	76.17%	71.81%	74.35%	66.80%
Participant 7	84.44%	83.07%	83.07%	66.54%
Participant 8	70.64%	65.36%	79.62%	67.84%
Participant 9	78.26%	75.46%	73.31%	72.07%
Participant 10	77.67%	77.73%	81.38%	73.31%

	Face Expression	Face-Gender	Pedestrian Gender	Vehicle
Participant 1	614	583	651	775
Participant 2	1127	954	1058	1308
Participant 3	654	690	730	846
Participant 4	805	729	796	1037
Participant 5	865	903	824	890
Participant 6	684	816	745	1037
Participant 7	667	642	668	790
Participant 8	614	683	650	821
Participant 9	596	554	725	928

Participant 10	906	828	946	946

Supplementary Information

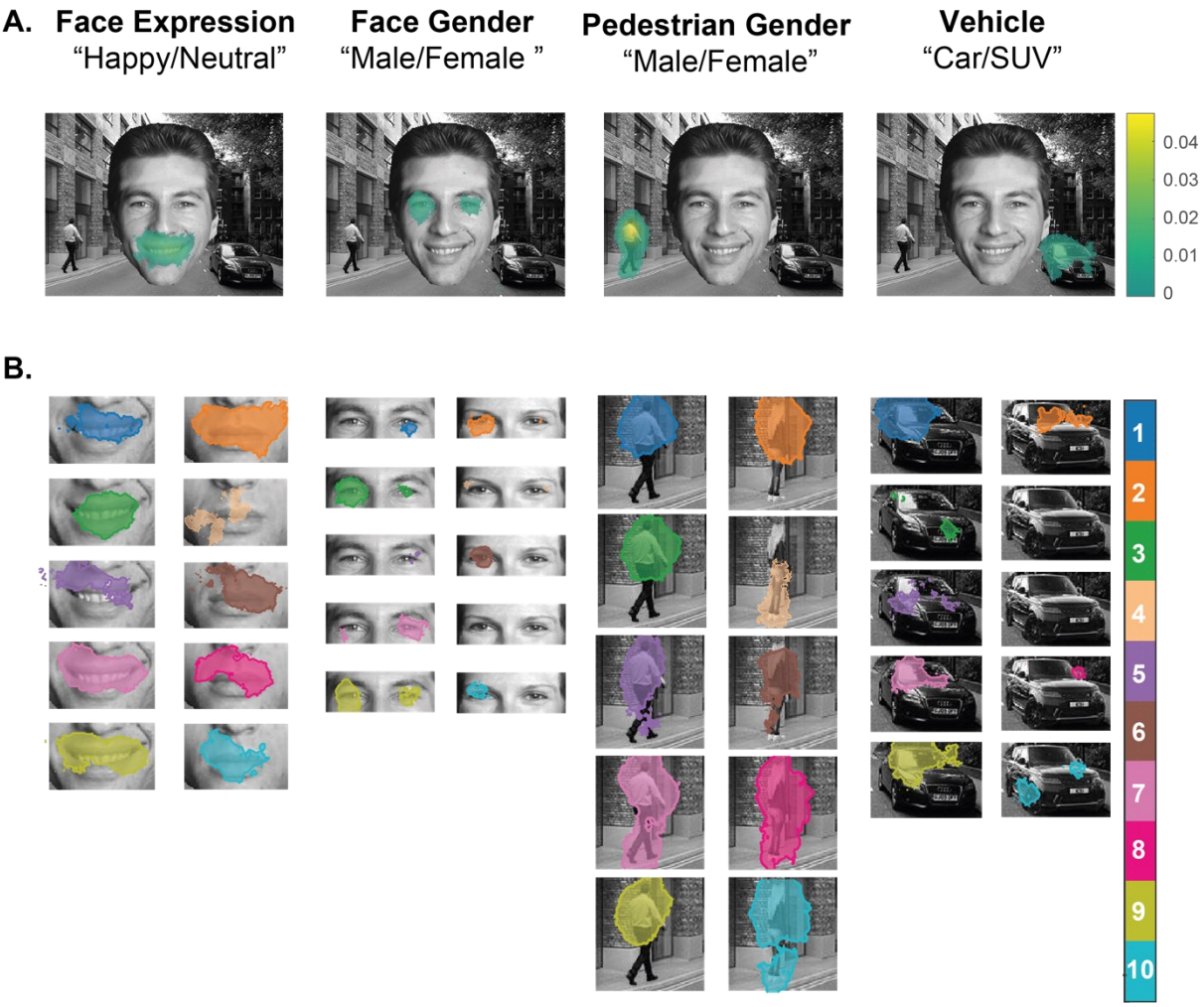

Figure 2-7 Task-relevant features. *A. Mean MI in each categorization task.* In each task (columns) and participant, for each image pixel I computed MI(<pixel visibility; correct vs. incorrect categorization>) (R. A. A. Ince et al., 2017), to reveal the significant (p < 0.05, FWER corrected) pixels that modulate categorization accuracy. For each pixel, i computed the mean MI across all ten participants. *B. Task-relevant features in each participant.* In each task (columns) the same olor-code represent the significant features for this participant. Note in each column (e.g. *pedestrian gender*) that different participants can use different (even mutually exclusive) features for the same categorization responses (e.g. for "male" vs "female pedestrian", upper body in participants 1, 2 and 3; lower body in participant 4). From the proportion of participants who significantly used each pixel, I estimated the population prevalence, expressed as a Bayesian maximum a posteriori (MAP) [95% Highest Posterior Density Interval (HPDI)] estimate. Face expressions: MAP [95% HPDI] = 1 [0.75 - 1]. Face gender: MAP [95% HPDI] = 0.58 [0.28 – 0.83]. Pedestrian: MAP [95% HPDI] = 1 [0.75 – 1]. Vehicle: [95% HPDI] = 0.47 [0.19 – 0.75].

Figure 2-8 Systems-level image transformations along the layers of the ventral pathway. A color-coded reference for the ventral pathway sources color-code their (cyan-to-yellow) depths (back-to-front). I sliced the ventral pathway into 7 layers of increasing depth and repeated for each layer (row), the analysis of Figure 2-2A to visualize the image transformations in the sources of each ventral layer (row) and time periods (columns in each task). Images show the number of participants (color-coded) that represent a given image pixel within a ventral layer and time window and task (FDR test with q=0.05). The result show task-specific transformations of broader image representations into lower-dimensional manifolds across the layers of the ventral pathway over ~91-150ms (when occipital cortex reduces task-irrelevant features).

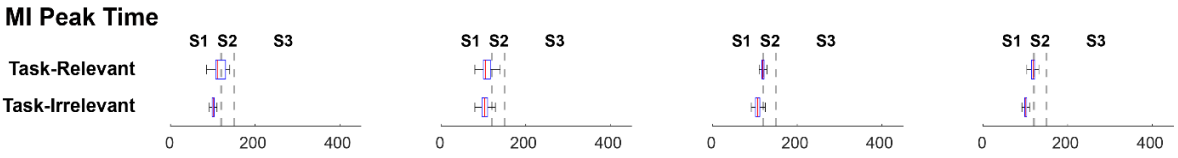

Figure 2-9 Distribution of peak time of feature representation. Across Stages 1 and 2, within 80-140 ms, for task-relevant vs. task-irrelevant features in each task.

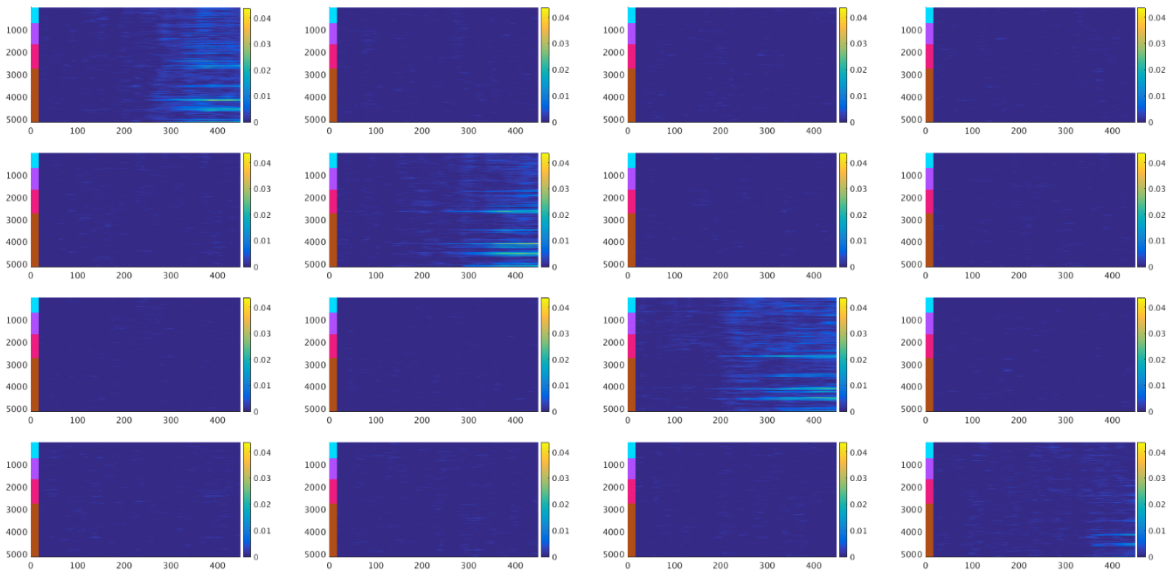

Figure 2-10 Categorical representation into MEG activity. Four rows are four tasks and four columns are four categorical information (i.e. categories of face expression, face gender, pedestrian gender and vehicle type). The plots show the mean MI(categorization response; MEG amplitude) over 10 participant. In each plot, x axis is time and y axis is sources (from top to bottom are occipital, temporal, parietal and frontal sources labeled with different respective). The plots on diagonal show that task-relevant categorical information is represented in brain activity after 200 ms post stimulus onset. While the plots off diagonal show that task-irrelevant categorical information is not represented in brain activity.

Figure 2-11 Categorization response representation into MEG activity with response-locked analysis. Four each task (each row), I plotted the prevalence of MI(categorization response; response-locked MEG activity) in grassbrains for 6 time windows across -200-400ms from participants' response. The results show a broad parietal-frontal network representing categorization responses, which peaks between 20-200ms post decision.

Note: Though the representation of categorization responses (behavior) per se is quite interesting, this study only focuses on the early representations of the feature manifolds and their transformations over time depending on the tasks.

3 Decomposing statistical dependence with pointwise and samplewise mutual information

3.1 Summary

In Chapter 2, although I reconstructed how the brain transforms the same complex scene images into task-relevant, low-dimensional features, the features used by participants in this experiment were relatively simple. However, many visual classification tasks rely on more complex features. This raises an important question: What is the minimum independent unit of representation in the brain for a visual categorization task? Addressing this question requires decomposing the reconstructed features, which necessitates a more precise characterization of the single trial relationship between stimuli and behavior, or between stimuli and the brain. With this refined single-trial relationship, clustering algorithms can learn about local clusters within these features. Therefore, in this chapter, I demonstrate how to use Pointwise Mutual Information (PMI) to decompose relationships and introduce a new information-theoretic measure inspired from PMI called Samplewise Mutual Information (SMI), which decomposes the relationship between two variables into the contribution of each trial/sample. I tested these measures using public datasets and will demonstrate how to use them and clustering algorithms to decompose diagnostic features in the next Chapter.

3.2 Introduction

Psychology depends heavily on statistical methods and modelling to gain insight from experimental data. Studies of decision making, categorization and information processing in psychology and neuroimaging rely on statistics to describe and quantify relationships in experimental data. Historically, the field has focused on the framework of Null Hypothesis Significance Testing (NHST), where an effect is statistically significant if it exceeds the chance level—defined as the level we would expect to observe 5% of the time under the null hypothesis that there is no effect. However, there is growing recognition of the insight that can be gained from quantifying statistical effects in more detail, as opposed to reducing them to a binary inferential result of significant vs. non-significant (McShane et al., 2019).

Although the focus is often on such binary inferences, most statistical methods provide an overall aggregate effect size – a single number that quantifies the strength of dependence between two measured variables in the observed data. For example, a correlation coefficient quantifies the strength of the linear relationship between two continuous variables. Mutual

information (MI) generalizes the measures (Cover & Thomas, 1991; R. A. A. Ince et al., 2017), to any form of dependence, linear or non-linear, and whether related to differences in means or in other higher order moments of a distribution. MI can be used as a statistical test, with either parametric inference based on the chi-squared null distribution or non-parametric permutation-based inference. However, a major advantage of MI for practical experimental statistics in psychology, neuroscience and neuroimaging is that it provides effect sizes on a common and meaningful scale (bits) across many different statistical tests with discrete, continuous, multidimensional and circular variables. These different families of variables are usually treated with different statistical tests whose effect sizes are not comparable.

Although MI provides meaningful and quantitatively comparable effect sizes across a range of different tests, it reduces the relationship between the samples of two considered variables to a single number. Two different data sets could produce statistical dependence of the same strength (same MI value), even with a very different relationship between the variables. Methods that describe the statistical relationships in more detail could then give greater insight, particularly when comparing between signals or responses as is common in cognitive neuroimaging (with high dimensional neural responses, stimuli and experimental parameters and behavioral variables).

Here, I present two extensions to MI that decompose the dependence in two different ways (Figure 3-1) to provide more detailed quantifications of the dependence relationship between two variables. The first, pointwise mutual information (PMI), quantifies the specific contribution of any particular combination of values of the variables. For example, consider an experiment where MI reveals a relationship between visual evidence (e.g. level of coherence of a pattern of random moving dots) and behavioral decision (e.g. perceived leftward vs rightward motion). To decompose this relationship, I would compute a PMI quantity for each specific combination of stimulus value (i.e. level of evidence and direction of motion) and response value (i.e. “left” vs. “right”). Figure 3-1 (third row) illustrates how PMI decomposes the relative contributions of each combination of the values of variables X (e.g. level of evidence for left and right motion) and Y (e.g. “left” vs. “right” decision) in both linear and nonlinear relationships. The second decomposition of MI, samplewise MI (SMI), quantifies the contribution of each individual sample (e.g. experimental trial) to the overall MI value. With SMI, researchers can quantify the degree to which individual trials follow or fail to follow (e.g. noisy trials) the overall pattern of dependence. Figure 3-1 (fourth row) shows the SMI value corresponding to each individual trial, where a high

positive (vs. negative) SMI (in red vs. blue) indicates that this trial contributes to (vs. contradicts) the relationship.

In sum, PMI and SMI decompose the aggregate MI measure of dependence between two variables. Whereas PMI quantifies the relative contribution of each possible combination of the *values* of the two variables, SMI accesses the contribution of each individual *sample* (e.g. experimental trial) to the aggregate effect. Both PMI and SMI provide further insight into the relationship that the overall MI effect size value does not reveal. For example, in Figure 3-1 the non-linear systems (see columns 4 and 5) have a similar strength of dependence measured with MI, but PMI and SMI reveal the different underlying structures of the relationships.

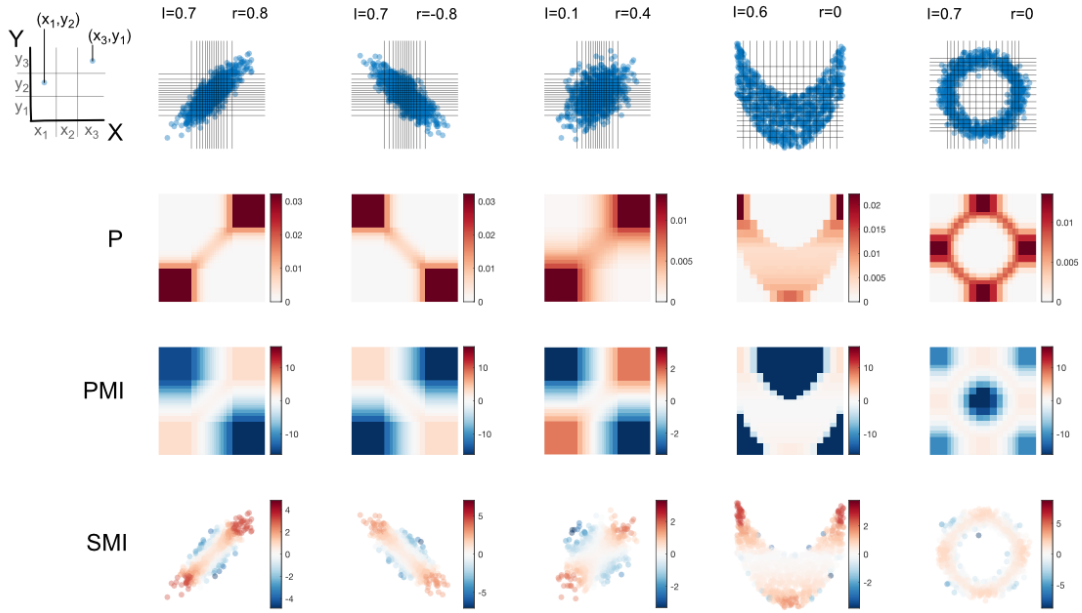

Figure 3-1 Examples of PMI and SMI. Each column shows an example of data simulated from a different system with varying dependence between two variables X and Y (x-axis and y-axis). From left to right, the first three columns show three systems where X and Y are jointly normally distributed with correlations 0.8, -0.8 and 0.4 respectively. In the last two examples, X and Y have a correlation of 0, but a clear non-linear relationship. Mutual Information (MI) is calculated by binning each variable into 16 equally occupied bins (top row; grey lines show bin edges) and sample the corresponding joint probability distribution as the normalized count of the number of samples in each joint bin (second row). Pointwise Mutual Information (PMI; third row): Red colors represent positive PMI, those values of X and Y are more likely to occur together than if the variables were independent. Blue colors represent negative PMI, those values of X and Y are less likely to occur together. The pattern of PMI over the input space reflects the structure of the underlying relationship (c.f. sign of correlation), both for linear (columns 1-3) and non-linear (columns 4-5) relationships. Samplewise Mutual Information (SMI; fourth row) shows the contribution of each sample (e.g. experimental trial), given by the PMI corresponding to the specific values of the variables that occurred in that sample. The mean of the SMI values over samples is equal to the overall MI. Positive SMI (red) indicates samples that contribute to the dependence; negative SMI (blue) indicates samples that are unlikely

given the overall dependence (i.e. noisy samples whose presence reduces the overall MI effect size value).

In this chapter, I introduce PMI and SMI and illustrate them on data from a simple 2-alternative forced choice (2-AFC) behavioral task. I show how PMI can quantify and classify the dependence between participants' response on one trial and the response for the previous trial (i.e. history bias). I then show how SMI between evidence and response quantifies an aspect of decision-making behavior that relates to reaction time. Next, I consider a more complicated reverse correlation task, with a natural stimulus image that is richly sampled across trials in a 3-alternative forced choice (3-AFC) design. Using PMI, I show the structure of serial dependence of behavioral responses in this experiment. Next, I will show how PMI and SMI can both be applied to the multi-class reverse correlation problem. Using PMI, I extract the image pixels that preferentially drive one or more responses (i.e. a unique vs. common effect). Using SMI I compute a classification image for each trial, and from these trial-based classification images I extract low-dimensional stimulus features. I validate these features by showing that they have a stronger relationship with participants' responses than other representations of the stimulus based on aggregate behavioral effects.

3.3 Results

Measuring serial dependence in a 2-AFC task with PMI

Serial dependence, or choice history bias, refers to the tendency of participants' responses on a trial, *t*, to be influenced by their response on the previous trial, *t-1*, even when this dependency is not optimal for decision making. Within the information theoretic framework, this serial dependence can be quantified with the mutual information between the responses on consecutive trials: $I(R_t; R_{t-1})$. MI expresses the strength of the serial dependence in bits (an unsigned positive value), without detailing the nature of the underlying statistical relationship. Here, I show how PMI can reveal this structure.

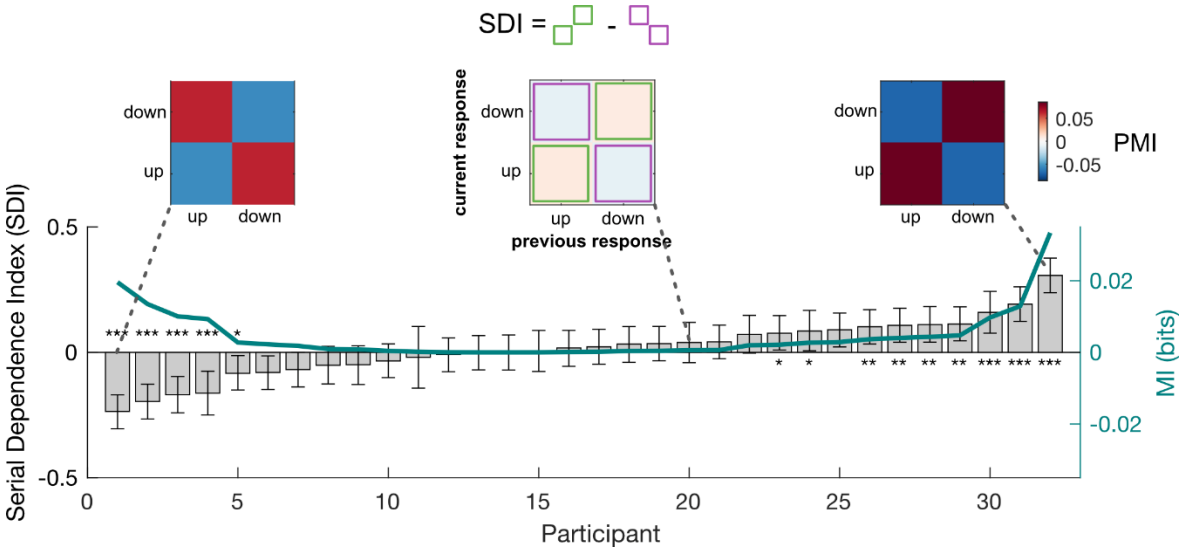

Figure 3-2 Quantifying serial dependence with PMI. In a 2-AFC experiment the MI between the current response choice (t) and the response choice on the previous trial (t-1) is calculated from four pointwise terms. I define a serial dependence index by considering the difference between the terms representing repetition of response choice (highlighted in green) and those representing alternation of response choice (highlighted in purple). Participants with similar values of MI (teal curve, right axis) can have different response patterns, which are revealed as different patterns of PMI (inset matrices, dashed lines indicate selected participant). SDI is shown for each participant (grey bars) together with 95% bootstrap confidence intervals. The statistical significance of SDI is determined from a two-sided permutation test; * = p<0.05, ** = p<0.01, *** = p<0.001.

Figure 3-2 (teal curve) shows MI, which quantifies the serial dependence between R_t and R_{t-1} for 32 participants in the visual fixed during 2-AFC discrimination task from (Urai et al., 2019). Inset matrices show PMI for three participants which illustrates the structure of the dependence relationship (see Methods). The left and right illustrated participant, have significant MI (permutation test, $p<0.001$), which shows that there is an above chance dependence between R_{t-1} and R_t. In contrast, there is no such serial dependence for the middle participant. The PMI values show markedly different patterns for relationships with similar strength (MI effect size). The participant shown on the right shows a tendency to repeat responses (e.g. more likely to respond "up" when response to the previous trial was also "up"). The participant shown on the left shows instead a tendency to alternate responses (e.g. more likely to respond "up" when the previous trial was "down"). Red colors indicate positive PMI, a response sequence more likely to occur than if each response was independent (optimal behavior given the task); blue colors indicate negative PMI, a response sequence less likely to occur given the observed serial dependence. Hence, this simple example illustrates how the aggregate MI effect size between two variables can be decomposed into the combinations of variable values that are more likely to occur (red, positive PMI) and those that are less likely to occur (blue, negative PMI), given the observed

dependence. The overall aggregate MI is the expectation of these PMI values over the observed joint distribution (see *Methods*).

To develop the analysis, I defined a Serial Dependence Index (SDI) as the difference between the repetition and alternation terms (see Methods, illustrated on the central participant in Figure 3-2). The bars in Figure 3-2 show the SDI for all 32 participants. 14 participants have a significant serial dependence (two-sided permutation test on SDI, $p<0.05$), 9 of them tend to repeat previous responses (i.e. positive SDI), whereas 5 participants alternated responses (i.e negative SDI). This analysis is applied to five different behavioral experiments from (Urai et al., 2019), a visual grating orientation decision task from (Benwell et al., 2019) and an unpublished auditory sweep direction decision task (see Methods). Based on the number of significant positive and negative SDI values, the population prevalence (R. A. A. Ince et al., 2020) proportion of individuals who would show such effects if tested is estimated. These results are shown in Figure 3-3. All experiments show reliable evidence for a non-zero prevalence of repeaters at the population level, with estimated values for the prevalence of repeating choice history behavior ranging from 15-50%. Only three experiments show reliable evidence for non-zero prevalence of alternators at the population level. For the auditory tasks there is evidence that the proportion of the population who show repeater behavior is greater than the proportion that show alternator behavior.

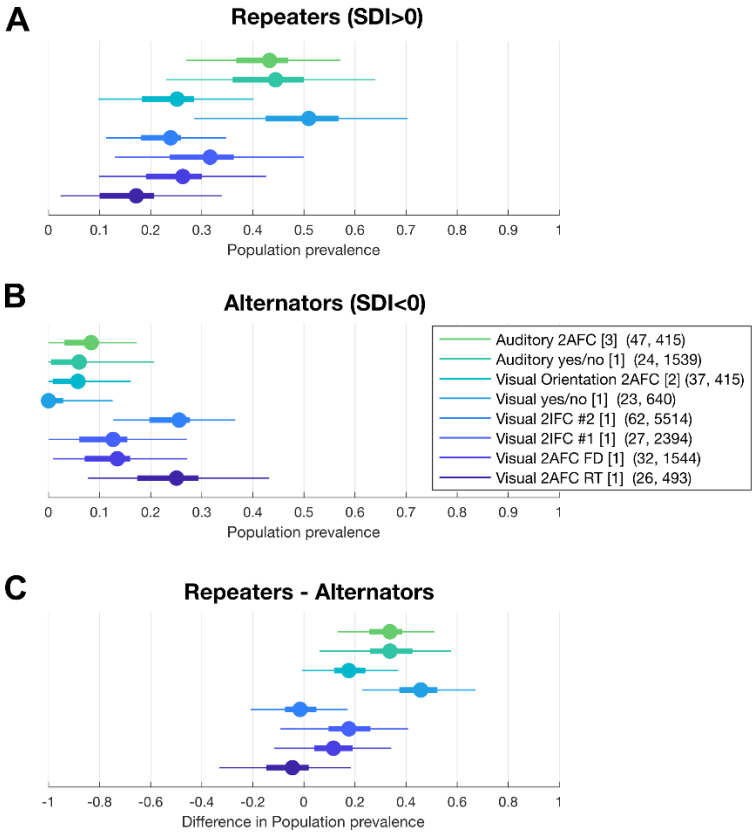

Figure 3-3 Population prevalence of repeating vs alternating serial dependence. The population prevalence proportion of participants who would show a true positive significant SDI value (R. A. A. Ince et al., 2020) is shown for a range of decision making tasks. For each participant a two-sided permutation test (p=0.05) is performed on the SDI value, the number of significant repeaters (SDI>0) and alternators (SDI<0). Circles show maximum a posteriori (MAP) estimate, thick lines show 50% highest posterior density interval (HPDI) and thin lines show 96% HPDI. Numbers in legend show number of participants and median number of trials for each experiment.

Indexing behavioral accuracy in a 2-AFC task with SMI between stimulus and response

I have shown how we can decompose the aggregate MI with PMI, to measure the contribution of each particular combination of variable values, and given an example where this pattern can be used to quantify meaningful features of the data. Now, I consider Samplewise Mutual Information (SMI), which decomposes an MI effect size by quantifying the specific contribution of each individual sample (here, experimental trial in the behavioral task).

To illustrate, I used the same 2-AFC motion perception task described and earlier (from Urai et al., 2019), where observers responded "left" vs. "right" to leftward vs. rightward moving random dot patterns with different levels of motion coherence. For each observer, I computed MI between the 10 different stimuli presented (5 levels of coherence each for left and right motion) and the observers' responses ("left" vs "right"). I then used PMI to decompose the aggregate MI relationship and reveal the behavioral pattern of each observer.

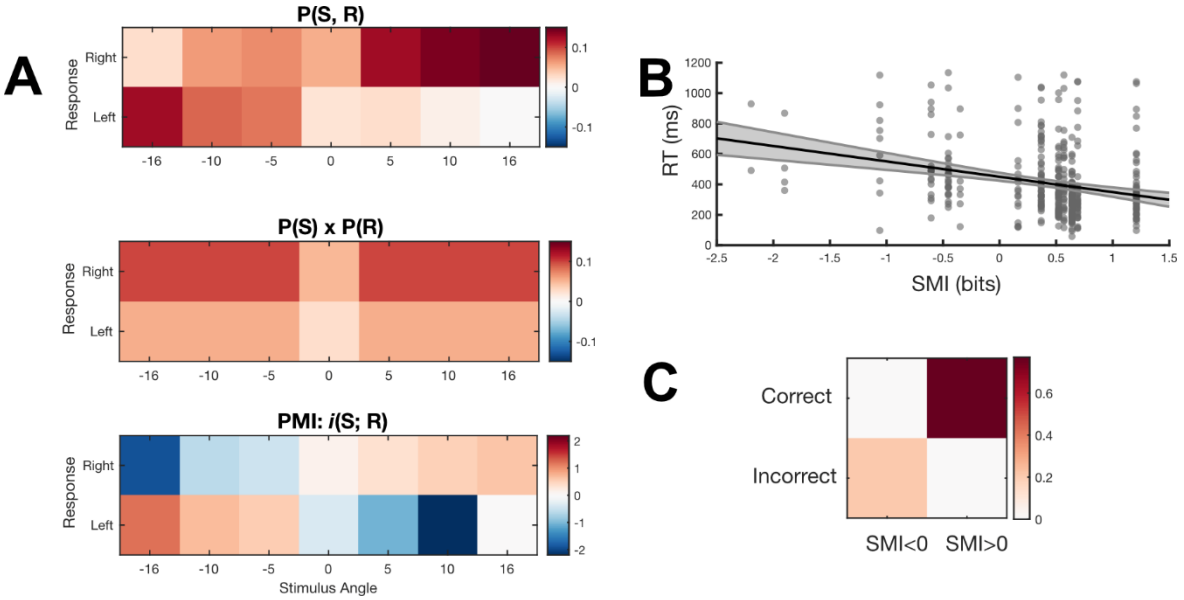

Figure 3-4 SMI between evidence and response in a 2-AFC task. A. *Top*: The joint probability distribution of stimulus evidence (here angle of a gabor patch, see Methods) and the participant response for an example participant. *Middle:* The joint probability distribution that would be expected if the participants response was statistically independent of the stimulus evidence. *Bottom*: The PMI for each combination of stimulus and response for this participant. B. The SMI value for each trial (obtained from the PMI for the particular stimulus and response combination of that trial) is negatively correlated with reaction time for this example participant. C: Combining trials across all participants (with SMI calculated based on each participant individual stimulus-response distribution), the sign of SMI is effectively equivalent to accuracy.

Whereas PMI decomposes MI into the specific values of the variables, SMI decomposes MI into the contribution of each specific trial. Each trial consists of one specific stimulus value that was shown on that trial, together with the particular response that the participant chose on that trial. The SMI value of each trial is obtained as the PMI value corresponding to the specific combination of stimulus and response which occurred on that trial. Hence, here each trial takes one of 20 possible PMI values from the joint space of 10 stimuli and 2 participant responses (Figure 3-4A). The mean SMI value over all trials is equal to the aggregate MI effect size (see Methods).

As for PMI, a positive (vs. negative) SMI value indicates that the combination of stimulus and response values on this trial are more (vs. less) likely to occur than if the variables were independent. In this experiment, all observers were accurate well above chance (mean 81%, range 70-87%, $p<0.001$ for each observer). For an observer performing above chance, positive SMI values generally indicate correct trials, as they indicate trials which follow the overall dependence pattern, which involves mostly correct responses due to the high accuracy. Negative SMI values generally indicate incorrect trials, as these are trials that don't follow the overall pattern of dependence in the participant's behavior. Note that SMI associated with individual trials can be higher, or lower than the aggregate MI, because MI is the average of the SMI values.

Measuring serial dependence in a 3-AFC task with PMI

The task consists of classifying spatial and spatial frequency sampled versions of a bistable Dali painting with three responses depending on which interpretation of the image was perceived – nuns (N, the two nuns in the slave market), voltaire (V, the disappearing bust of Voltaire) or don't know (DK). I show the PMI values for the 9 combinations of 3 responses over two consecutive trials (x-axis: previous response, y-axis: current response). Positive values of PMI mean that sequence of responses is more likely to occur than if the participants

were responding independently on each trial (red scale). Negative values of PMI mean that sequence of responses is less likely to occur (blue scale). Note that the serial dependence effect here is weak (0.0054 [0.0018-0.0188] bits mean [range] across participants) but significant, ($p<0.005$ for all subjects obtained from 100,000 permutations). This may be due to the presence of strong stimulus evidence on many trials (MI between dimensionality reduced stimulus and response is 0.36 [0.28-0.5] bits), which could override the bias from the previous response. PMI reveals different patterns of serial dependence: two participants (4 and 5, second row) show an alternating behavior between N and V: they are more likely to respond N when they responded V on the previous trial, and vice versa. Participants 1,2 and 3 on the other hand instead show repetition behavior, they are more inclined to repeat the same positive response (but not necessarily DK), and less likely to switch between N and V. In addition to these patterns there are some participant specific idiosyncrasies: participant 1 is more like to respond DK after V, and participant 3 is more likely to response N after DK. All participants are less likely to respond V after DK. Note that participant 2 is an outlier (10 x greater serial dependence MI than the other participants) and follows the repetition response pattern, suggesting there may have been periods where that participant paid less attention to the stimuli and instead responded repetitively.

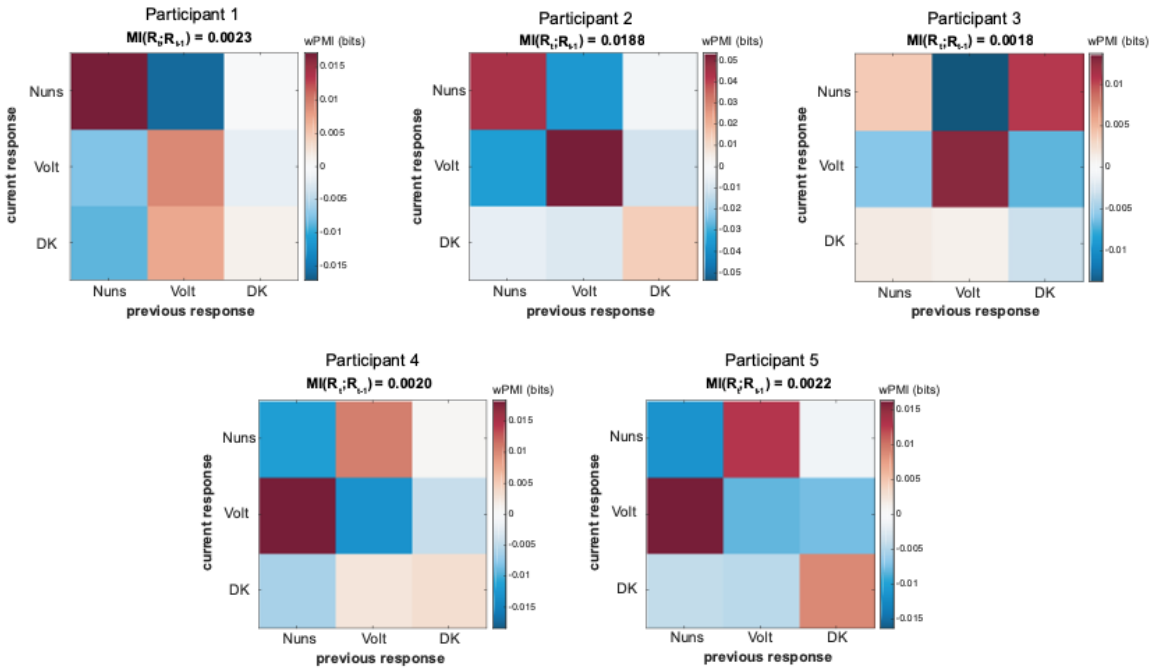

Figure 3-5 PMI for serial dependence in the 3-AFC Dali bubbles task

3.4 Methods

Pointwise Mutual Information (PMI)

Mutual information (MI) is the most general statistical test of dependence between two variables (Cover & Thomas, 1991; R. A. A. Ince et al., 2017). Here I consider the discrete formulation for categorical variables. Mutual information has a number of advantages as a statistical tool for practical data analysis. It makes minimal assumptions on the form of any dependence, is defined for any number of discrete classes and has a meaningful effect size in bits. The MI between two discrete variables *S* and *R* is usually defined as:

$$I(S;R) = \sum_{s\in S, r\in R} p(s,r)\log_2 \frac{p(s,r)}{p(s)p(r)}$$

Note that this is equivalent to expectation with respect to the joint distribution *P(S,R)* of a function (Bouma, 2009; Church & Hanks, 1990; Cover & Thomas, 1991; R. A. A. Ince, 2017; Lizier et al., 2014; Wibral et al., 2015):

$$i(s;r) = \log_2 \frac{p(s,r)}{p(s)p(r)} = \log_2 \frac{p(r|s)}{p(r)}$$

So that

$$I(S;R) = \langle i(s;r)\rangle_{P(S,R)} = \sum_{s\in S, r\in R} p(s,r)i(s;r)$$

The value of the function *i(s;r)* for specific values of *s*, *r* has been termed the *pointwise* (Bouma, 2009; Church & Hanks, 1990) or the *local* (Lizier et al., 2008) mutual information. While *I(S;R)* is always greater than or equal to zero, the pointwise terms *i(s;r)* can be either positive or negative. Note that $i(s;r)$ for a specific *r* and *s* is positive when $p(r|s) > p(r)$, and negative when $p(r|s) < p(r)$.

Applying a Bayesian perspective, *p(r)* can be thought of as the prior probability of *r*. *p(r|s)* can be interpreted as the posterior probability of *r*, after *s* has been observed. A Bayes optimal gambler who is able to observe values of *S* prior to betting on the outcome *R* (S-gambler) with knowledge of the overall participants performance *P(R,S)* makes bets based on *P(R|S)*. A gambler without access to S (blind gambler) would place bets based only on

the prior *P(R)*. Consider a simple experiment where the stimulus is an arrow pointing either left (s_L) or right (s_R). The participant is asked to respond indicating the observed direction (r_L, r_R). If the participant is performing the task successfully (i.e. above chance accuracy) then when s_L is presented our S-gambler will place a larger bet on the outcome r_L, as they know the participant is more likely to provide the correct response. If that response indeed occurs, then the S-gambler will have higher winnings than the blind gambler, as they bet more on the correct outcome r_L, which indeed occurred on this trial. This corresponds to a positive value of $i(r_L;s_L)$ because $p(r_L|s_L)>p(r_L)$, and means that, *when the particular values* s_L *and* r_L *observed*, our S-gambler won more than our blind gambler.

However, assuming imperfect performance of our participant, perhaps due to lapses in concentration, there will be some trials where they make an error and respond r_R even when the stimulus is s_L. In this case, our S-gambler will have observed s_L and bet accordingly on the outcome r_L which was more likely. But although r_L is more likely overall when stimulus s_L is presented, on this particular trial, it didn't occur – the participant was incorrect. While overall the participant is correct more than they are incorrect, they are still incorrect on some trials. So in this specific trial, our S-gambler would earn less than the blind gambler. This corresponds to a negative value of $i(r_R;s_L)$ because $p(r_R|s_L)<p(r_R)$. In general, negative values of pointwise information mean that, for that particular combination of values, the information provided by the stimulus was *misleading*, because an event it suggested was less likely to occur, nevertheless did occur. Hence negative pointwise information values have been termed *misinformation* (Wibral et al., 2015).

Of course, $I(S;R)>0$, and so, on average, the optimal posterior gambler will always do better in the long run. But pointwise mutual information can help us identify the contribution of specific combinations of values. Pointwise approaches have been applied in linguistics (Bouma, 2009; Church & Hanks, 1990; Recchia & Jones, 2009), in the study of complex systems such as cellular automata (Lizier et al., 2008, 2012, 2014) and recently in neuroscience (Martinez-Cancino et al., 2018; Wibral et al., 2014, 2015).

Throughout this paper when plotting PMI values I actually plot the summand of the mutual information expectation, which is the local information value weighted by the probability of those particular values: $p(s,r)i(s;r)$.

Samplewise Mutual Information (SMI)

Samplewise Mutual Information (SMI) is simply the PMI evaluated at the specific pair of values obtained in each sample. For example, if the experiment consists of a set of trials, where the stimulus presented on trial t is s_t and the response on the same trial is r_t, then the SMI for that trial is given by $SMI(t) = i(s_t, r_t)$. Then the overall MI is the mean of the SMI values over trials:

$$I(S;R) = \frac{1}{N_t}\sum_t SMI(t)$$

As for PMI, positive values occur on trials where the two observed values of *s* and *r* are more likely to occur together than they would if *S* and *R* were independent, and negative values occur on trials where the particular observed *s* and *r* are less likely to co-occur than if *S* and *R* were independent. In this sense, samples with positive SMI are those that follow the overall dependence relationship, while samples with negative SMI deviate from the overall relationship.

As well as considering positive vs negative values (i.e. $SMI(t) > 0$ or $SMI(t) < 0$) we can also consider samples with $SMI(t) > MI$ vs those with $SMI(t) < MI$. The former are trials which are driving the relationship to be stronger, while the latter are trials that are pulling down the overall dependence. Splitting on MI rather than 0 results in a more balanced binarization, especially when the MI is large.

In the field of complex systems both PMI and SMI have often been termed local mutual information.

Serial Dependence Index (SDI)

The serial dependence index for a 2-AFC task is defined in terms of the PMI as:

$$SDI = [i(r_{t-1} = 0, r_t = 0) + i(r_{t-1} = 1, r_t = 1)] \\ -[i(r_{t-1} = 0, r_t = 1) + i(r_{t-1} = 1, r_t = 0)]$$

That is the difference between the diagonal terms (representing choice repetition) and the off-diagonal terms (representing choice alternation). Positive values of SDI indicate repetition serial dependence, negative values indicate alternation.

SDT Experiment

I consider behavioral data from a two alternative forced choice (2AFC) fixed duration visual motion discrimination task described in (Urai et al., 2019). Data are available from https://doi.org/10.6084/m9.figshare.7268558. 32 observers performed a random dot motion discrimination (up vs down) task. After a fixation interval of 0.75—1.5s, random dot motion stimuli (0, 3, 9, 27 or 81% motion coherence) were displayed for 750ms. See (Urai et al., 2019) for full experimental details.

Dali Experiment

I consider behavioral data from a three alternative forced choice (3AFC) "bubbles" task (R. A. A. Ince et al., 2015; Zhan, Ince, et al., 2019). A ambiguous section of the Salvador Dali painting *"Slave Market with the Disappearing Bust of Voltaire"*, was sampled with bubbles (Gosselin & Schyns, 2001) across five spatial frequencies. The image was decomposed into 5 spatial frequency octaves, each of which were independently sampled by randomly positioned Gaussian apertures with standard deviation dependent on spatial frequency. These sampled spatial frequency slices were then recombined into a single presented stimulus image. Participants were asked to report whether they perceived the *nuns* in the slave market, the bust of *Voltaire*, or *don't know*, in case neither perception was clear.

3.5 Discussion

In event-related experimental designs there is increasing interest in what are often terms *single-trial* methods (Pernet et al., 2011), in which the trial-by-trial variability within participants is explicitly considered. For example, with events of two different classes, a classical approach would be to average the recorded neuroimaging signal to presentations of each class, and then look for a significant group level difference between those two mean responses across participants. But considering the variability at the trial level can give extra insight, particularly for reverse correlation experiments using high-dimensional sample data.

Therefore I developed a new information theory quantity SMI, inspired by PMI in information theory (Cover & Thomas, 2012; R. A. A. Ince et al., 2017), to quantify the trial-by-trial relationship between variables. Specifically, PMI gives details of the pattern of dependence between discrete variables with any number of categories, which is demonstrated here with examples of serial dependence and behavioral decisions. SMI gives the contribution of each individual trial sample to the overall dependence. This provides an avenue for higher order statistics, where the trial-by-trial relationship between two variables can be quantified, and then related to other experimental measures.

While I focus here on behavioral tasks, it is important to note that both PMI and SMI can be directly applied to neuroimaging data such as EEG, MEG or fMRI. I expect SMI in particular to have broad application.

In reverse correlation experiments, by establishing the relationship between noise stimuli and participant behavior, we can reconstruct the features that influence behavior, which are also referred to as mental representations (Gosselin & Schyns, 2001; Murray, 2011). These representations reflect how the stimulus is mentally perceived and processed under a given task. However, many visual tasks rely on complex features (Zhan, Ince, et al., 2019), which raises an important question: what is the minimum independent unit of representation that the brain processes? To address this, it is crucial to accurately characterize the relationship between stimuli and behavior, or between stimuli and brain activity, at the level of single trials. SMI characterizes the single-trial relationship between stimulus samples and participants' behavior, resulting in single-trial classification images. Clustering algorithms like Non-negative Matrix Factorization (NMF) can learn local pixel clusters (Lee & Seung, 1999) that share a common single-trial relationship structure as unitary features.

Moreover, contemporary cognitive neuroscience views the brain as a complex network (Bassett & Sporns, 2017), where cognitive functions are supported by connections between different regions of the brain. Given this framework, a critical question arises: how are features represented by connectivity between two units (e.g. neurons, voxels, channels) ranther than only within individual units? Since representation involves the relationship between stimuli and neural activity (Baker et al., 2022; Poldrack, 2021), this now involves examining the relationship between a feature and a relationship. However, measuring relationships typically requires a set of trials or samples, and the relationship itself is often a single statistical value summarized from a set of trials or samples, therefore without trial/sample dimension. To overcome this problem, we can decompose a relationship into

contributions from individual trials, providing a method to measure the relationship between a variable and a relationship, or even between two relationships. With these methods, we can explore scenarios such as whether a feature is represented in the connectivity between activities of two units or whether the activity of a unit can represent the feature representation state of another unit. In all such contexts, the ability to compute and decompose relationships becomes essential. This is where SMI proves valuable, offering a powerful tool for these complex analyses in future work.

4 Decomposing task-relevant features with trial-by-trial variations can better predict the behavior

4.1 Introduction

Reverse correlation techniques quantify the relationship between randomly sampled stimuli and behavioral responses, resulting in classification images that reveal which stimulus subspace are relevant to participants' behavior. (Gosselin & Schyns, 2001; Murray, 2011). The stimulus subspace revealed by classification images is termed task-relevant or diagnostic features, which are considered the mental representation of stimulus features (Brinkman et al., 2017). More recently, reverse correlation experiments have focused on measuring the neural representation of stimuli (Schyns et al., 2009; M. L. Smith et al., 2004; Zhan, Ince, et al., 2019). As shown in Chapter 2, this thesis also developed a systems-level analysis framework to reverse engineer the internal transformations of neural representational feature manifolds.

However, it is not necessarily the case that the task-relevant feature must be the minimum feature unit in mental representation. Thus, this thesis proposed a key challenge of decomposing the task-relevant features (i.e. diagnostic features) into their local parts that collectively influence participants' behavior trial by trial, serving as a proxy of the minimum feature unit of mental representation.

To address this challenge, Chapter 3 develops Pointwise Mutual Information (PMI) to quantify the specific contribution of any particular combination of values of the variables and Samplewise Mutual Information (SMI) to measure the contribution of each individual sample to the overall relationships between two variables within the information theory framework. This chapter further applies PMI and SMI on task-relevant features that are obtained from Mutual Information (MI) between participants' behavior and stimulus samples to decompose them into local parts respectively. PMI can be used to directly decompose the stimulus features into parts that are specifically contributing to each behavior response. While SMI does not directly decompose features into parts. Instead, SMI characterizes the single-trial relationship between stimulus samples and participants' behavior, resulting in single-trial classification images. I then apply clustering algorithms of Non-negative Matrix Factorization (NMF) to learn their local pixel clusters (Lee & Seung, 1999) that share a common single-trial relationship structure as unitary features. Finally, MI

between each type of decomposed features and participants behavior is calculated to compare whether the decomposition enhances the MI relationship effect.

This analysis is applied on the perception of Dali's ambiguous painting experiment (Bonnar et al., 2002; Zhan, Ince, et al., 2019). Participants can perceive either two nuns or a Voltaire bust from a static image due to processing of different features. In this experiment, participants utilized more complex features from the images for each perceptual decision than the experiment in chapter 2, therefore more suitable for the decomposition analysis. In this chapter, I used PMI to decompose task-relevant features into features that support each individual perception (i.e. Nuns features or Voltaire features) as well as features that can support both two perceptions. Besides, I used SMI and the NMF algorithm to decompose the task-relevant features into more granular local components. By calculating the relationship (i.e. MI) between these different features and behavior, the results showed that the more refined features derived using SMI and NMF algorithms could better predict participants' behavior.

4.2 Results

PMI decomposition of task relevant features

Pointwise Mutual Information (PMI) can decompose the relationship between two variables into the contribution of each pair of outcomes to the overall correlation. I calculate the PMI between participants' behavior and stimulus samples to decompose task-relevant features into the features that support different behavioral responses. Figure 4-1 shows the task-relevant features obtained by MI between stimulus samples and the participant's behavior and the decomposed Nuns-unique, Voltaire-unique as well as common features.

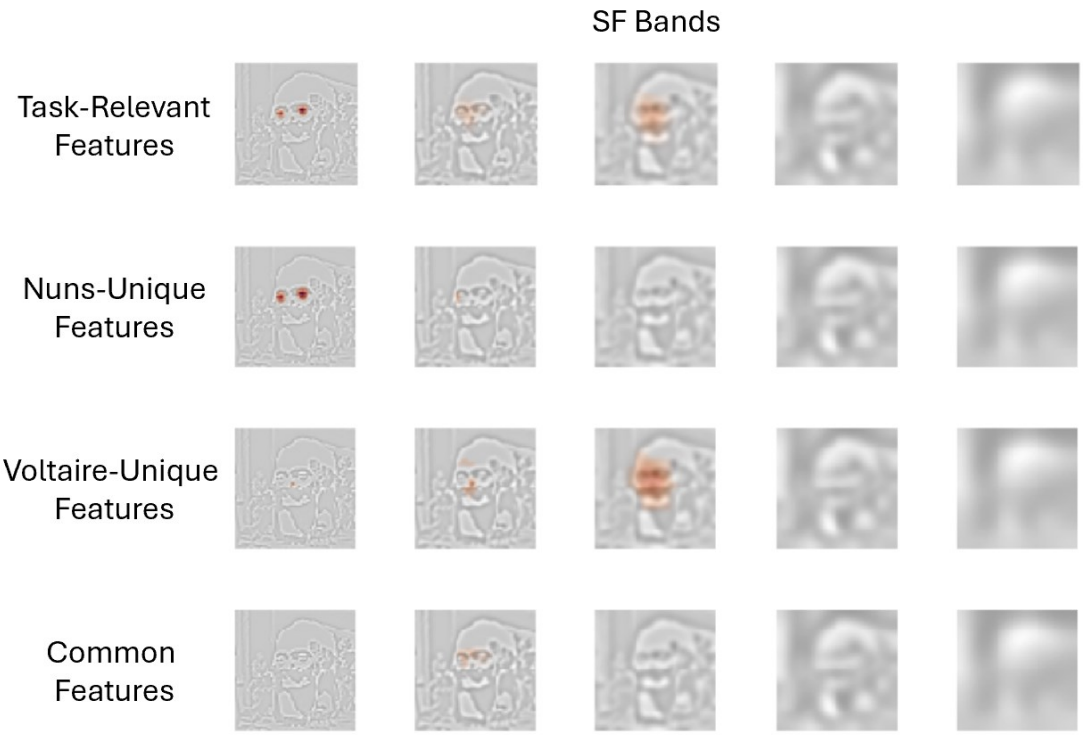

Figure 4-1 PMI decomposition of task relevant features. Each row represents a feature. Each column represents a spatial frequency (SF). The first row shows task relevant features obtained from MI between stimulus samples and the participant's behavior. The second row shows Nuns-unique features. The third row shows Voltaire-unique features. The fourth row shows common features for both Nuns and Voltaire responses.

Single-trial task-relevant features

Samplewise Mutual Information (SMI) can decompose the relationship between two variables into the contribution of each trial to the overall correlation (i.e., single-trial relationship). I calculate the SMI between participants' behavior and stimulus samples for every pixel to decompose task-relevant features into the features that support participant's behavioral responses in each individual trial. Left panel of Figure 4-2 displays three example trials showing the stimulus images presented to the participants. These stimulus images were sampled using the Bubbles technique (Gosselin & Schyns, 2001). The right panel shows the task-relevant features (classification images) derived for each trial using SMI. Specifically, the SMI values on the images reveal which regions or features of the image drove the participants' response in that particular trial.

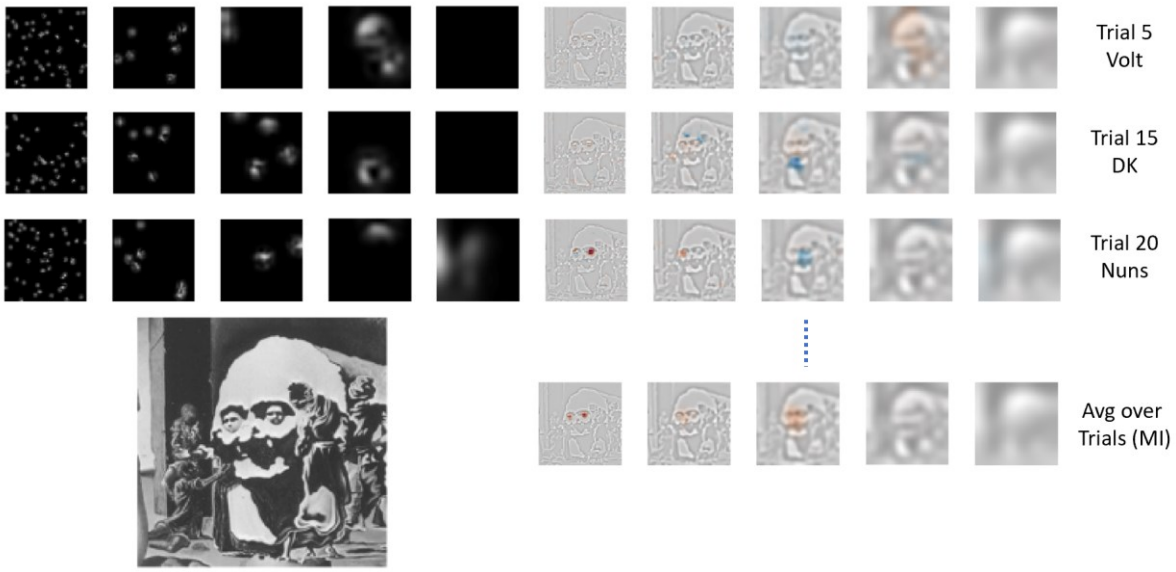

Figure 4-2 Single-trial task relevant features. On the left panel, I selected three trials where participants gave three different behavior responses. One the right panel, red blobs (i.e. positive SMI values) on Images show which features are relevant to the participant's behavior in that trial, which is obtained from SMI between stimulus samples of each pixel and the participant's behavior. The average of SMI values over trials equals to MI.

SMI-NMF decomposition of task relevant features

Non-negative Matrix Factorization (NMF) is a popular algorithm used for decomposition and dimensionality reduction of non-negative data, such as images. NMF aims to express the original matrix as a linear combination of basis vectors (also known as components or features) with non-negative coefficients (Lee & Seung, 1999). I applied the NMF algorithm on matrix of single-trial task-relevant features obtained with SMI to decompose features into the local parts (a local group of pixels) that collectively support participant's behavior from trial to trial (see Figure 4-3). Results show that NMF is able to decompose task-relevant features into parts of eyes, nose, mouth, which have been proven to be represented in participants' brain activity (Zhan, Ince, et al., 2019).

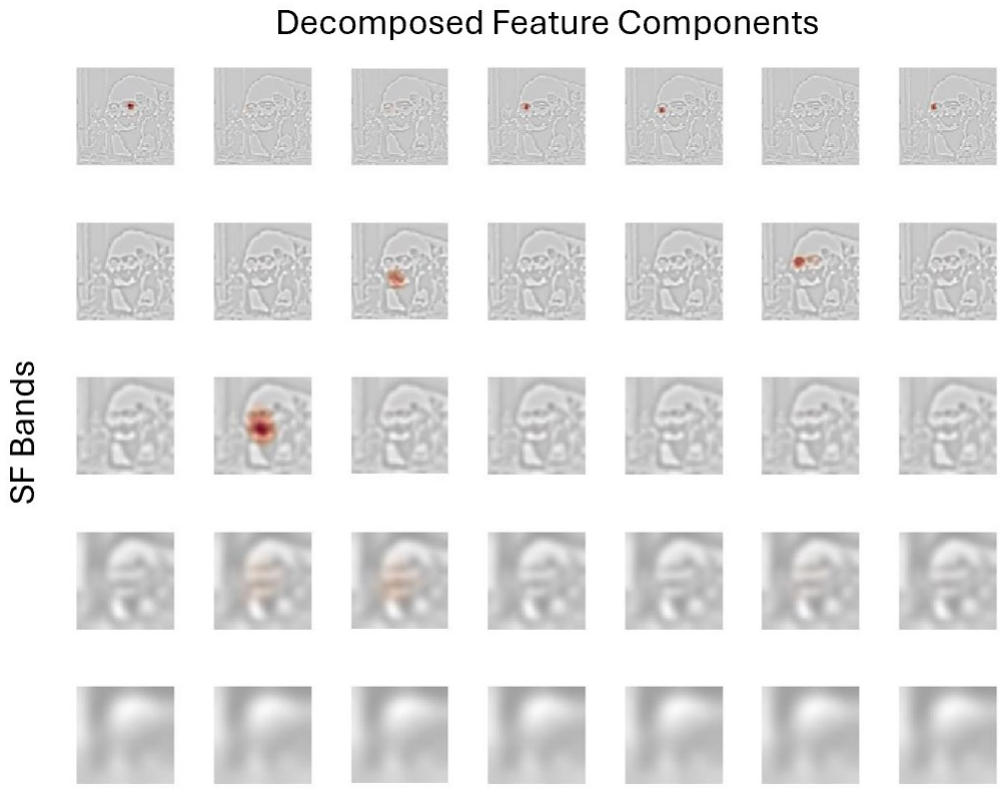

Figure 4-3 SMI-NMF decomposition of task relevant features. By applying NMF on SMI matrix, task-relevant features are decomposed into local features (components). Each column represents a feature (component). Each row represents a spatial frequency (SF). NMF-SMI decomposes the complex task-relevant features into local features with specific semantic meaning (i.e. left or right eye, mouth, broad face).

SMI-NMF features predict the participant's behavior better

I computed MI between participants' behavior and each feature to show how much information in participants' behavior can be predicted or explained by different kinds of features. The results show that the local features obtained from NMF-SMI decomposition can better predict participants' behavior (see Figure 4-4).

	Participant 1	Participant 2	Participant 3	Participant 4	Participant 5
NMF-SMI	0.8982	0.8985	0.6155	0.8131	0.7594
PMI	0.4709	0.3789	0.3458	0.4298	0.4132
MIvsDK	0.4888	0.3954	0.2736	0.4487	0.4015
NMF-MEG	0.4411	0.4143	0.3314	0.4404	0.4643
Ent(Beh)	1.5569	1.4973	1.4583	1.4666	1.4930

Figure 4-4 MI between participants' behavior and each feature. Different rows show MI between participants' behavior and features obtained by different methods. The last row shows how much

information (unit: bit) in total is carried by participants' behavior. This figure reveals how much information about behavior can be predicted or explained by different features. Results show that NMF-SMI features can predict participants' behavior better.

4.3 Methods

Participants

Five right-handed observers with normal (or corrected to normal) vision participated in the experiment. Informed consent was obtained from all observers and ethical approval from the University of Glasgow Faculty of Information and Mathematical Sciences Ethics Committee.

Experiment

The ambiguous portion of Dali's Slave Market with the Disappearing Bust of Voltaire was cropped to retain the bust of Voltaire and the two nuns. The image was presented at 5.72° × 5.72° of visual angle on a projector screen (image size was 256 × 256 pixels). Bubble masks made of randomly placed Gaussian apertures sampled information from the cropped image to create a different sparse stimulus for each trial (see Figure 4-5). The Dali image was decomposed into five independent Spatial Frequency (SF) bands of one octave each, with cutoffs at 128 (22.4), 64 (11.2), 32 (5.6), 16 (2.8), 8 (1.4), 4 (0.7) cycles per image (cycles per degree of visual angle). For each SF band, a bubble mask was generated from a number of randomly located Gaussian apertures (the bubbles), with a standard deviation of 0.13, 0.27, 0.54, 1.08, and 2.15 degrees respectively. The image content of each SF band was sampled by multiplying the bubble masks and underlying greyscale pixels at that SF band, and summed the resulting pixel values across SFs to generate the actual stimulus image. The stimulus remained on the screen until the observer depressed one of three possible response keys, according to which aspect of the image they perceived: "the nuns" (N), "Voltaire" (V), or "don't know" (DK). A fixation cross was presented from 500 ms prior and until stimulus onset and observers were instructed to maintain fixation during each trial. The total number of Gaussian apertures remained constant throughout the task, ensuring that equivalent amounts of visual information was presented on each trial, at a level (60 bubbles) found previously to maintain "don't know" responses at 25% of the total number of responses30. Since the underlying image was always the same, all analysis was performed on the bubble masks controlling visibility. For analysis, I down-sampled (bilinear interpolation) the bubble masks to a resolution of 64 × 64 pixels.

Each trial started with a fixation cross displayed for 500 ms at the center of the screen, immediately followed by a stimulus generated as explained above that remained until response. Observers were instructed to maintain fixation during each trial, and to respond by pressing one of three keys ascribed to each response choice. Stimuli were presented in runs of 150 trials, with randomized inter-trial intervals of 1.5–3.5s (mean 2s). Observers performed 4–5 runs in a single day session with short breaks between runs. Observers completed the experiment over 4–5 days.

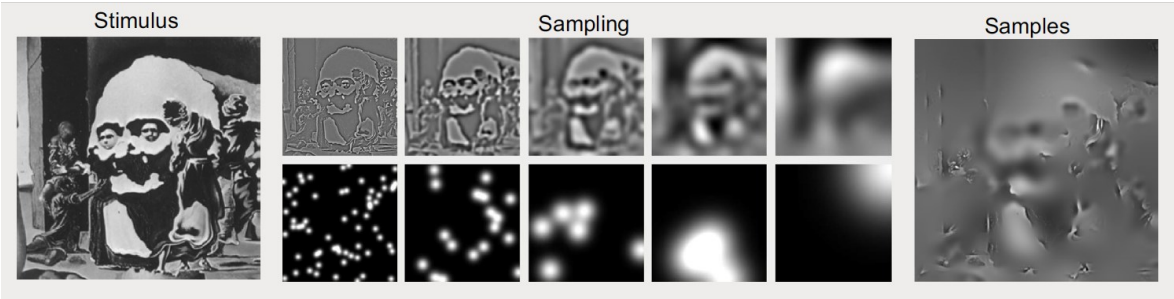

Figure 4-5 The stimulus in DALI experiment.

Pointwise Mutual Information (PMI)

Mutual information (MI) is the most general statistical test of dependence between two variables (Cover & Thomas, 1991; R. A. A. Ince et al., 2017). Here I consider the discrete formulation for categorical variables. Mutual information has a number of advantages as a statistical tool for practical data analysis. It makes minimal assumptions on the form of any dependence, is defined for any number of discrete classes and has a meaningful effect size in bits. The MI between two discrete variables *S* and *R* is usually defined as:

$$I(S;R) = \sum_{s\in S, r\in R} p(s,r)\log_2\frac{p(s,r)}{p(s)p(r)}$$

Note that this is equivalent to expectation with respect to the joint distribution *P(S,R)* of a function (Bouma, 2009; Church & Hanks, 1990; Cover & Thomas, 1991; R. A. A. Ince, 2017; Lizier et al., 2014; Wibral et al., 2015):

$$i(s;r) = \log_2\frac{p(s,r)}{p(s)p(r)} = \log_2\frac{p(r|s)}{p(r)}$$

So that

$$I(S;R) = \langle i(s;r) \rangle_{P(S,R)} = \sum_{s \in S, r \in R} p(s,r) i(s;r)$$

The value of the function *i(s;r)* for specific values of *s*, *r* has been termed the *pointwise* (Bouma, 2009; Church & Hanks, 1990) or the *local* (Lizier et al., 2008) mutual information. While *I(S;R)* is always greater than or equal to zero, the pointwise terms *i(s;r)* can be either positive or negative. Note that $i(s;r)$ for a specific *r* and *s* is positive when $p(r|s) > p(r)$, and negative when $p(r|s) < p(r)$.

Samplewise Mutual Information (SMI)

Samplewise Mutual Information (SMI) is simply the PMI evaluated at the specific pair of values obtained in each sample. For example, if the experiment consists of a set of trials, where the stimulus presented on trial t is s_t and the response on the same trial is r_t, then the SMI for that trial is given by $SMI(t) = i(s_t, r_t)$. Then the overall MI is the mean of the SMI values over trials:

$$I(S;R) = \frac{1}{N_t} \sum_{t} SMI(t)$$

4.4 Discussion

In an experiment where participants can perceive either two nuns or a Voltaire bust from a static ambiguous image, I identified the features that support each perception, as well as the features that can support both perceptions by calculating the PMI between participants' behavior and stimulus samples. I also employed the SMI to calculate the single-trial relationship between participant's behavior and stimulus samples. The results reveal the task-relevant features on each trial—specifically, the image features that drive or influence the participant's response in that trial. Subsequently, I applied Non-negative Matrix Factorization (NMF) algorithms (Lee & Seung, 1999) to decompose the SMI matrix. The decomposition results reveal the local components (i.e. a group of local pixels that collectively influence the participant's behavior) of the task-relevant features. Results show that NMF is able to decompose task relevant features into parts of eyes, nose, mouth, similar

to stimulus features represented in participants' brain activity. By identifying these local components, I effectively pinpointed the minimum units that drive the participant's responses.

In the context of reverse correlation study, task-relevant features (i.e. classification images) are regarded as mental representations—visual stimulus features that are mentally represented in the participant's brain while performing specific visual categorization tasks (Brinkman et al., 2017; Murray, 2011). Therefore, the local components derived from the NMF decomposition can be viewed as the minimum units of mental representations processed in the brain.

It is worth noting that the same analysis can be applied to brain activity data to better characterize single-trial stimulus features represented in the brain. However, due to the significantly larger volume of brain activity data compared to the behavioral data, this method encounters computational challenges in practice, which requires further exploration in future work.

It is assumed that people do classification tasks by matching visual stimuli to mental templates/representations (Brinkman et al., 2017). Chapter 2 shows that, in the time window of P3 ERP component (associated with decision making and attention), brain represents similar stimulus contents to mental templates (i.e., task relevant features obtained from reverse correlation analysis), supporting this hypothesis. However, chapter 2 uses a stimulus with simple features for each categorization task. This chapter decomposes the more complexed task-relevant features from the perception of Dali's ambiguous painting experiment (Bonnar et al., 2002) into parts carrying specific semantic meaning (e.g., eyes, nose, mouth), which have been proven to be represented in participants brain activity (Zhan, Ince, et al., 2019).

In Chapter 2, I demonstrated how the brain actively transforms the complex visual inputs into task-relevant features at ~300 ms post-stimulus. For tasks where participants rely on more complex features, it is also crucial to investigate whether the brain similarly transforms complex visual inputs into task-relevant features with refined components around the same time window. This hypothesis remains to be tested in future work. To verify this, it will be necessary to apply the feature decomposition methods with NMF and SMI to brain activity data, as discussed earlier.

5 General Discussion

One of the most influential model in cognitive neuroscience posits the brain as an information processing system (Marr & Ullman, 2010). In this model, identifying the specific visual information (i.e. the mental representation of the stimulus features) processed by the brain is crucial for understanding the neural mechanisms underlying behavior. Although numerous studies have attempted to approximate the mental representation of stimulus features using either task-relevant features obtained from behavior data (Brinkman et al., 2017; Murray, 2011) or neural representations (Schyns et al., 2009; M. L. Smith et al., 2004; Zhan, Ince, et al., 2019), there is still a lack of understanding how the brain transforms the internal representation of stimulus features depending on the task at hand (K. Kay et al., 2023).

One significant reason for the challenges in understanding neural representations is the complexity of these representations (Biederman, 1987; Logothetis & Sheinberg, 1996), which has not been adequately considered. Due to the brain's attentional mechanisms (Evans et al., 2011; Shiffrin & Gardner, 1972), it does not represent all contents of a visual input but selectively processes task-relevant features. These features are not fixed or inherent. That means, the brain does not simply passively receive and then choose which features to process or suppress. Instead, the brain actively extracts and constructs features from the visual input based on the task at hand (Schyns et al., 1998; Schyns & Rodet, 1997). Investigating this dynamic process requires fine-grained, high-dimensional control over the stimuli (Gosselin & Schyns, 2001; Murray, 2011), the use of multitask experimental paradigms (Harel et al., 2014) along with efficient data analysis methods (R. A. A. Ince et al., 2017).

In this thesis, I employed the Bubbles (Gosselin & Schyns, 2001) method to control the visual stimuli. Bubbles control the visibility of each pixel of stimulus images. In a visual categorization task, the participant can correctly perform the task only when the task-relevant features are revealed by the Bubbles. This approach allows us to identify which features in the image support the participant’s categorization behavior. Since every pixel is controlled, I can observe pixels collectively influence the participant's behavior, thereby flexibly identifying the features within the image. A similar approach can be used to observe which pixels collectively influence brain activity, that is, which pixels are collectively represented by brain activity. This enables the flexible revelation of the represented feature content.

For the multitask, I designed an experiment comprising four 2-Alternative-Forced-Choice (AFC) categorization tasks applied to the same realistic, complex city street scene images randomly sampled with Bubbles procedure (Gosselin & Schyns, 2001). Each task required the participant to use different features from the images, leading to a process in the brain where the same visual input is transformed into different features. This process involves a combination of bottom-up visual processing and top-down task-related and attentional information (DiCarlo et al., 2012; Evans et al., 2011; Harel et al., 2014; VanRullen & Thorpe, 2001).

I then employed information theory (Cover & Thomas, 2012; R. A. A. Ince et al., 2017) methods as an efficiency framework for quantifying the relationships between two or more variables. This efficiency is crucial given the high-dimensional nature of both the stimulus data and the brain activity data. A key advantage of this approach is its ability to measure relationships regardless of whether the variables are discrete, continuous, or a mix of both. Moreover, the relationships are measured on a common scale (bits), facilitating the comparison of results.

Through this experimental approach, I provided a descriptive model of how the brain gradually transforms high-dimensional, complex visual input into task-relevant features through occipital-ventral pathway by three distinct stages. This model also highlights the crucial role of interactions between the prefrontal cortex (PFC) and the occipital-ventral visual pathway in this dimension-reducing transformations depending on tasks (Johnston & Everling, 2006; Roy et al., 2010).

In this experiment, participants complete each task using relatively simple features, but many real-world visual tasks depend on more complex features. In such cases, the task-relevant features identified by reverse correlation methods may not necessarily represent the smallest processing units in the brain. To identify these minimal units, I introduced a measure based on information theory, the Samplewise Mutual Information (SMI), which quantifies the single-trial relationship between two variables. By calculating the SMI between stimulus samples and the participant's behavior, I identified the features that support participant's response in each single trial. These features, referred to as single-trial task-relevant features, were then analyzed using clustering algorithms (i.e. Non-negative Matrix Factorization, NMF) to learn the patterns among them (Lee & Seung, 1999). This allowed for the decomposition of complex task-relevant features into finer components. The components identified through this method correspond to meaningful features such as eyes, mouth, and

broader facial regions. These refined feature components can better predict the participant's behavior than the features directly derived from reverse correlation analysis (Bonnar et al., 2002; Zhan, Ince, et al., 2019).

Broader implications

My research focuses on understanding how the brain transforms complex visual inputs into low-dimensional, task-relevant features. A key factor motivating this work is the brain's attentional mechanisms (Evans et al., 2011; Harel et al., 2014; Kastner & Pinsk, 2004; Shiffrin & Gardner, 1972), which allow it to selectively process only a subset of information rather than all visual input. The transformation in this thesis is closely related to the mechanisms of attention.

It’s important to emphasize that attention itself is a phenomenon we aim to explain through neuroscience. Theories and evidence suggest that visual categorization abilities of brain depend on features (DiCarlo et al., 2012; Schyns et al., 1998; Schyns & Rodet, 1997; VanRullen & Thorpe, 2001). However, these features are neither fixed nor inherent in the stimuli. Rather, they are subjectively created by the brain by the demands of the visual task at hand. My data show that the visual content of images represented within the brain progressively transforms across three stages into task-relevant features with the involvement of synergistical interactions between the occipital-ventral pathway and the PFC cortex (Duan et al., 2024). This supports the idea that the brain actively constructs features based on the task.

Moreover, the evidence from visual information decoding (Harel et al., 2014) suggests that neural representations are not purely bottom-up reflections of stimuli but rather complex representations integrating both stimulus information and top-down information of task context. My results align with this view and demonstrate two key characteristics that result in such task-dependent representations: 1. The visual feature content represented by neural activity is highly task-dependent, which is demonstrated as the result of three stages of transformations. 2. Even for the same feature, the internal neural state that encodes the external feature states is also task-dependent, demonstrated as the result of opponent representations (Duan et al., 2024) of the same feature in chapter 2.

The methodologies I have developed in this thesis are not limited to human brain data. They can also be applied to deep network models that also integrate attentional mechanisms (i.e.

transformers) and learn contextual representations (Vaswani et al., 2017). By employing the same techniques, we can investigate how deep networks transform complex inputs into features. Given that attention mechanisms have significantly enhanced the performance of deep networks, I anticipate that similar processes of feature transformation will be observable in these models. This approach provides a way to compare the information processing strategies of the human brain and deep networks (Schyns et al., 2022) in a visually intuitive manner.

Future work

In the context of this thesis, I define the transformation of visual content represented in neural activity as the macroscopic behavior of the brain, conceptualized as an information-processing machine. If this macroscopic behavior can be characterized as thoroughly as human behavior, it could then be utilized as a powerful tool to study the cognitive system of the human brain. By relating neural activity patterns to brain behaviors, in a manner analogous to existing methodologies that relate neural activity to human behaviors (Greene et al., 2023). This perspective allows us to bridge the gap between neural activity (physical properties of brain) and cognitive function (emerging properties of mind), though a middle level of macroscopic behavior of the brain. Further efforts are required to address this challenge. Although this ambitious goal was not achieved in this thesis, the proposed methods bring us closer to a solution.

In this thesis, I tested the internal transformation of simplest feature manifolds with 2D images. While the real world is 3D. Considering the time, the real visual input is at least 4D. Like 2D images, the internal representation of 3D/4D visual inputs are a low-dimensional manifold of them. Testing the actual feature manifolds and their transformations for 3D/4D visual input and generative models (Schyns et al., 2022; Zhan, Garrod, et al., 2019) are needed in future work.

Moreover, the information theory framework employed in this thesis has significant potential for further development and application. For instance, Figure 5-1 show that Conditional Mutual Information (CMI) reveals representational effects that are not captured by the Mutual Information (MI) analysis. It indicates that this representation depends on its own past stage, which reflects a bottom-up procure in neural dynamics. Besides, Figure 5-2 show that co-information is able to identify the complex network interaction patterns for neural sources. Further exploration of the exact feature contents communicated by these

networks can be conducted. For example, an important question that arises is whether the communication between different neural sources serves merely for transmitting features, or if, during this process of communication, there is also a transformation of the visual content. These analyses need to be formalized and applied to large-scale datasets in future work.

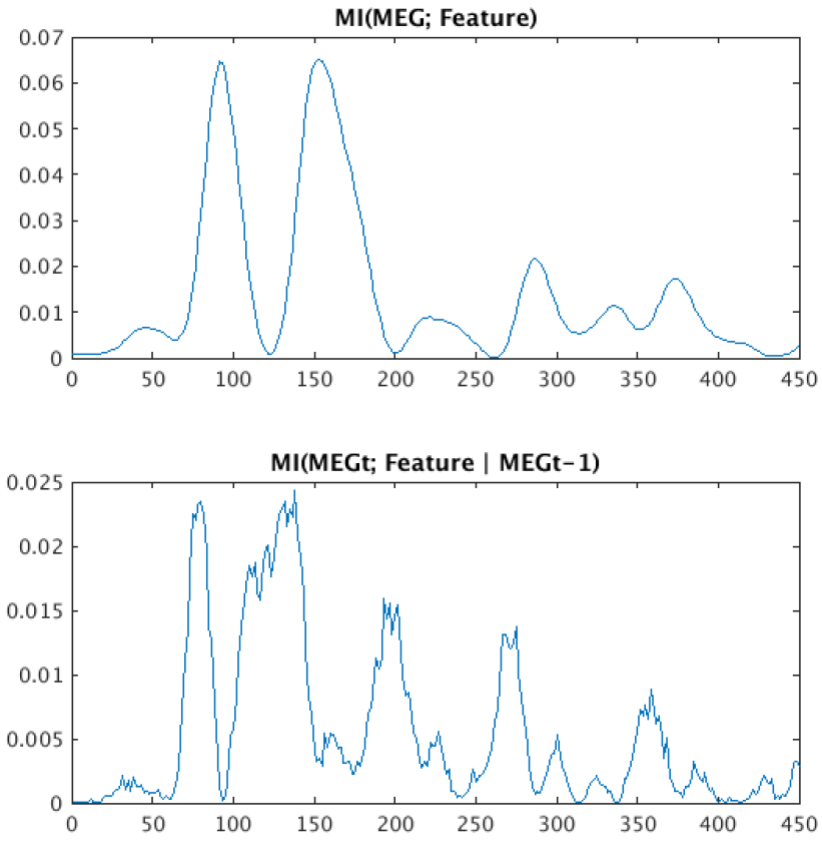

Figure 5-1 Self Conditional Mutual Information (CMI) reveals effects hidden in MI analysis. The Conditional Mutual Information (CMI) analysis (bottom panel) reveals representational effects that are not captured by the Mutual Information (MI) analysis (top panel). It reflects that the neural representation in this source depends on its own past state.

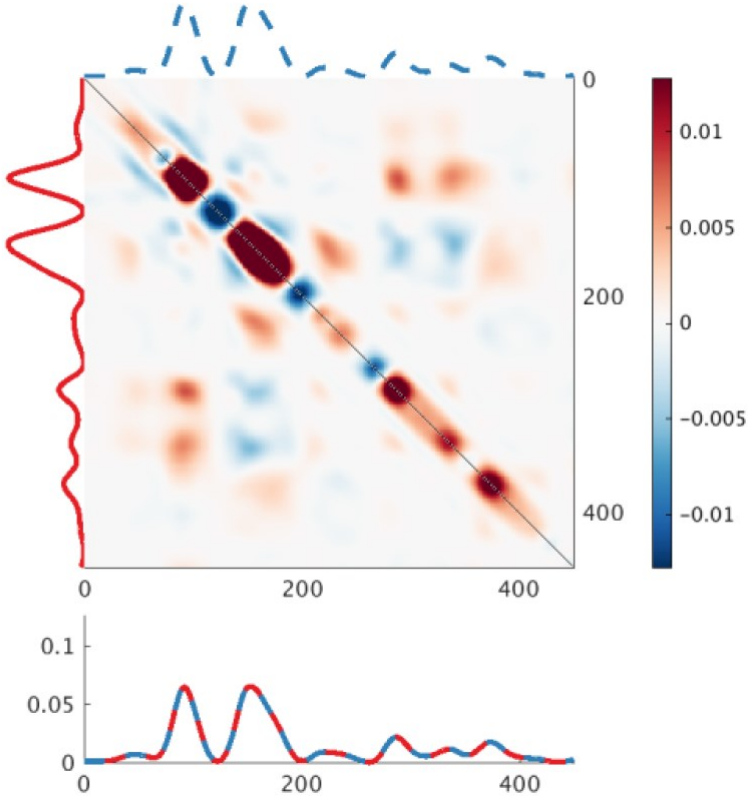

Figure 5-2 Co-information reflects source self-interaction patterns The co-information between each pair of time points from a source reveals local redundant representations, local synergistic representations, and patterns of cross-temporal redundancy and synergy.

The methods employed in this paper are primarily data driven. Within this framework, incorporating additional experimental manipulations could yield more meaningful and robust results. For instance, the transfer of task-relevant features from the occipital to the ventral stream, and the transformations that occur during this transfer, might be the crucial mechanisms underlying conscious visual processing and working memory (Duan et al., 2024). To test this hypothesis, one could introduce behavioral manipulations within the same experimental paradigm. By controlling the presentation time of stimuli and using masking techniques to inhibit participants' conscious perception of visual content, we could observe whether these neural processes diminish, thereby validating these proposed mechanisms.

Reference

Ahumada, A., Jr., & Lovell, J. (1971). Stimulus Features in Signal Detection. *The Journal of the Acoustical Society of America*, *49*(6B), 1751–1756. https://doi.org/10.1121/1.1912577

Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417–423. https://doi.org/10.1016/S1364-6613(00)01538-2

Baillet, S. (2017). Magnetoencephalography for brain electrophysiology and imaging. *Nature Neuroscience*, *20*(3), 327–339. https://doi.org/10.1038/nn.4504

Baker, B., Lansdell, B., & Kording, K. P. (2022). Three aspects of representation in neuroscience. *Trends in Cognitive Sciences*, *26*(11), 942–958.

Bassett, D. S., & Sporns, O. (2017). Network neuroscience. *Nature Neuroscience*, *20*(3), 353–364. https://doi.org/10.1038/nn.4502

Bentin, S., Allison, T., Puce, A., Perez, E., & McCarthy, G. (1996). Electrophysiological Studies of Face Perception in Humans. *Journal of Cognitive Neuroscience*, *8*(6), 551–565. https://doi.org/10.1162/jocn.1996.8.6.551

Bentin, S., Taylor, M. J., Rousselet, G. A., Itier, R. J., Caldara, R., Schyns, P. G., Jacques, C., & Rossion, B. (2007). Controlling interstimulus perceptual variance does not abolish N170 face sensitivity. *Nature Neuroscience*, *10*(7), 801–802. https://doi.org/10.1038/nn0707-801

Benwell, C. S. Y., Beyer, R., Wallington, F., & Ince, R. A. A. (2019). History biases reveal novel dissociations between perceptual and metacognitive decision-making. *bioRxiv*, 737999. https://doi.org/10.1101/737999

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, *94*(2), 115–147. https://doi.org/10.1037/0033-295X.94.2.115

Bonnar, L., Gosselin, F., & Schyns, P. G. (2002). Understanding Dali's *Slave Market with the Disappearing Bust of Voltaire*: A Case Study in the Scale Information Driving Perception. *Perception*, *31*(6), 683–691. https://doi.org/10.1068/p3276

Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, 31–40.

Bracci, S., & Op De Beeck, H. P. (2023). Understanding Human Object Vision: A Picture Is Worth a Thousand Representations. *Annual Review of Psychology*, *74*(1), 113–135. https://doi.org/10.1146/annurev-psych-032720-041031

Brignani, D., Lepsien, J., & Nobre, A. C. (2010). Purely endogenous capture of attention by task-defining features proceeds independently from spatial attention. *NeuroImage*, *51*(2), 859–866. https://doi.org/10.1016/j.neuroimage.2010.03.029

Brinkman, L., Todorov, A., & Dotsch, R. (2017). Visualising mental representations: A primer on noise-based reverse correlation in social psychology. *European Review of Social Psychology*, *28*(1), 333–361. https://doi.org/10.1080/10463283.2017.1381469

Broadbent, D. E. (1957). A mechanical model for human attention and immediate memory. *Psychological Review*, *64*(3), 205–215. https://doi.org/10.1037/h0047313

Broadbent, D. E. (1958). *Perception and communication* (pp. v, 340). Pergamon Press. https://doi.org/10.1037/10037-000

Buchsbaum, G., Gottschalk, A., & Barlow, H. B. (1983). Trichromacy, opponent colours coding and optimum colour information transmission in the retina. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, *220*(1218), 89–113. https://doi.org/10.1098/rspb.1983.0090

Carrasco, M., & Barbot, A. (2019). Spatial attention alters visual appearance. *Current Opinion in Psychology*, *29*, 56–64. https://doi.org/10.1016/j.copsyc.2018.10.010

Church, K. W., & Hanks, P. (1990). Word Association Norms, Mutual Information, and Lexicography. *Comput. Linguist.*, *16*(1), 22–29.

Cichy, R. M., & Kaiser, D. (2019). Deep Neural Networks as Scientific Models. *Trends in Cognitive Sciences*, *23*(4), 305–317. https://doi.org/10.1016/j.tics.2019.01.009

Cichy, R. M., Pantazis, D., & Oliva, A. (2014). Resolving human object recognition in space and time. *Nature Neuroscience*, *17*(3), 455–462. https://doi.org/10.1038/nn.3635

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. Wiley New York.

Cover, T. M., & Thomas, J. A. (2012). *Elements of Information Theory*. John Wiley & Sons.

Çukur, T., Nishimoto, S., Huth, A. G., & Gallant, J. L. (2013). Attention during natural vision warps semantic representation across the human brain. *Nature Neuroscience*, *16*(6), 763–770. https://doi.org/10.1038/nn.3381

Dailey, M., Cottrell, G. W., & Reilly, J. (2001). California facial expressions, CAFE. *Unpublished Digital Images, University of California, San Diego, Computer Science and Engineering Department*.

Daube, C., Xu, T., Zhan, J., Webb, A., Ince, R. A. A., Garrod, O. G. B., & Schyns, P. G. (2021). Grounding deep neural network predictions of human categorization behavior in understandable functional features: The case of face identity. *Patterns*, *2*(10), 100348. https://doi.org/10.1016/j.patter.2021.100348

De Melo, C. M., Torralba, A., Guibas, L., DiCarlo, J., Chellappa, R., & Hodgins, J. (2022). Next-generation deep learning based on simulators and synthetic data. *Trends in Cognitive Sciences*, *26*(2), 174–187. https://doi.org/10.1016/j.tics.2021.11.008

Dehaene, S., Charles, L., King, J.-R., & Marti, S. (2014). Toward a computational theory of conscious processing. *Current Opinion in Neurobiology*, *25*, 76–84. https://doi.org/10.1016/j.conb.2013.12.005

Desimone, R., & Duncan, J. (1995). Neural Mechanisms of Selective Visual Attention. *Annual Review of Neuroscience*, *18*(1), 193–222. https://doi.org/10.1146/annurev.ne.18.030195.001205

DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences*, *11*(8), 333–341. https://doi.org/10.1016/j.tics.2007.06.010

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How Does the Brain Solve Visual Object Recognition? *Neuron*, *73*(3), 415–434. https://doi.org/10.1016/j.neuron.2012.01.010

Duan, Y., Zhan, J., Gross, J., Ince, R. A. A., & Schyns, P. G. (2024). Pre-frontal cortex guides dimension-reducing transformations in the occipito-ventral pathway for categorization behaviors. *Current Biology*, *34*(15), 3392-3404.e5. https://doi.org/10.1016/j.cub.2024.06.050

Evans, K. K., Horowitz, T. S., Howe, P., Pedersini, R., Reijnen, E., Pinto, Y., Kuzmova, Y., & Wolfe, J. M. (2011). Visual attention. *WIREs Cognitive Science*, *2*(5), 503–514. https://doi.org/10.1002/wcs.127

Finn, E. S., Huber, L., & Bandettini, P. A. (2020). Higher and deeper: Bringing layer fMRI to association cortex. *Progress in Neurobiology*, 101930. https://doi.org/10.1016/j.pneurobio.2020.101930

Frangou, P., Emir, U. E., Karlaftis, V. M., Nettekoven, C., Hinson, E. L., Larcombe, S., Bridge, H., Stagg, C. J., & Kourtzi, Z. (2019). Learning to optimize perceptual decisions through suppressive interactions in the human brain. *Nature Communications*, *10*(1), Article 1. https://doi.org/10.1038/s41467-019-08313-y

Friston, K. (2010a). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. https://doi.org/10.1038/nrn2787

Friston, K. (2010b). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. https://doi.org/10.1038/nrn2787

Gescheider, G. A. (2013). *Psychophysics: The Fundamentals*. Taylor & Francis. https://books.google.co.uk/books?id=gATPDTj8QoYC

Gosselin, F., & Schyns, P. G. (2001). Bubbles: A technique to reveal the use of information in recognition tasks. *Vision Research*, *41*(17), 2261–2271. https://doi.org/10.1016/S0042-6989(01)00097-9

Graham, N., & Wolfson, S. S. (2004). Is there opponent-orientation coding in the second-order channels of pattern vision? *Vision Research*, *44*(27), 3145–3175. https://doi.org/10.1016/j.visres.2004.07.018

Greene, A. S., Horien, C., Barson, D., Scheinost, D., & Constable, R. T. (2023). Why is everyone talking about brain state? *Trends in Neurosciences*, *46*(7), 508–524. https://doi.org/10.1016/j.tins.2023.04.001

Grill-Spector, K., & Weiner, K. S. (2014). The functional architecture of the ventral temporal cortex and its role in categorization. *Nature Reviews Neuroscience*, *15*(8), 536–548. https://doi.org/10.1038/nrn3747

Gross, J., Baillet, S., Barnes, G. R., Henson, R. N., Hillebrand, A., Jensen, O., Jerbi, K., Litvak, V., Maess, B., Oostenveld, R., Parkkonen, L., Taylor, J. R., van Wassenhove, V., Wibral, M., & Schoffelen, J.-M. (2013). Good practice for conducting and reporting MEG research. *NeuroImage*, *65*, 349–363. https://doi.org/10.1016/j.neuroimage.2012.10.001

Grossman, M., Smith, E. E., Koenig, P., Glosser, G., DeVita, C., Moore, P., & McMillan, C. (2002). The Neural Basis for Categorization in Semantic Memory. *NeuroImage*, *17*(3), 1549–1561. https://doi.org/10.1006/nimg.2002.1273

Hanks, T. D., & Summerfield, C. (2017). Perceptual Decision Making in Rodents, Monkeys, and Humans. *Neuron*, *93*(1), 15–31. https://doi.org/10.1016/j.neuron.2016.12.003

Harel, A., Kravitz, D. J., & Baker, C. I. (2014). Task context impacts visual object processing differentially across the cortex. *Proceedings of the National Academy of Sciences*, *111*(10). https://doi.org/10.1073/pnas.1312567111

Hari, R., Parkkonen, L., & Nangini, C. (2010). The brain in time: Insights from neuromagnetic recordings. *Annals of the New York Academy of Sciences*, *1191*(1), 89–109. https://doi.org/10.1111/j.1749-6632.2010.05438.x

Hebart, M. N., Bankson, B. B., Harel, A., Baker, C. I., & Cichy, R. M. (2018). The representational dynamics of task and object processing in humans. *eLife*, *7*, e32816. https://doi.org/10.7554/eLife.32816

Henderson, J. M., & Hayes, T. R. (2017). Meaning-based guidance of attention in scenes as revealed by meaning maps. *Nature Human Behaviour*, *1*(10), Article 10. https://doi.org/10.1038/s41562-017-0208-0

Herculano-Houzel, S. (2009). The human brain in numbers: A linearly scaled-up primate brain. *Frontiers in Human Neuroscience*, *3*. https://doi.org/10.3389/neuro.09.031.2009

Hillebrand, A., & Barnes, G. R. (2005). Beamformer Analysis of MEG Data. In *International Review of Neurobiology* (Vol. 68, pp. 149–171). Elsevier. https://doi.org/10.1016/S0074-7742(05)68006-3

Huber, L., Finn, E. S., Chai, Y., Goebel, R., Stirnberg, R., Stöcker, T., Marrett, S., Uludag, K., Kim, S.-G., Han, S., Bandettini, P. A., & Poser, B. A. (2021). Layer-dependent functional connectivity methods. *Progress in Neurobiology*, *207*, 101835. https://doi.org/10.1016/j.pneurobio.2020.101835

Humphreys, G. W. (2016). Feature confirmation in object perception: Feature integration theory 26 years on from the Treisman Bartlett lecture. *The Quarterly Journal of Experimental Psychology*, *69*(10), 1910–1940. https://doi.org/10.1080/17470218.2014.988736

Huth, A. G., Nishimoto, S., Vu, A. T., & Gallant, J. L. (2012). A Continuous Semantic Space Describes the Representation of Thousands of Object and Action Categories across the Human Brain. *Neuron*, *76*(6), 1210–1224. https://doi.org/10.1016/j.neuron.2012.10.014

Ince, R. (2017). Measuring Multivariate Redundant Information with Pointwise Common Change in Surprisal. *Entropy*, *19*(7), 318. https://doi.org/10.3390/e19070318

Ince, R. A. A. (2017). Measuring Multivariate Redundant Information with Pointwise Common Change in Surprisal. *Entropy*, *19*(7), 318. https://doi.org/10.3390/e19070318

Ince, R. A. A., Giordano, B. L., Kayser, C., Rousselet, G. A., Gross, J., & Schyns, P. G. (2017). A statistical framework for neuroimaging data analysis based on mutual information estimated via a gaussian copula: Gaussian Copula Mutual Information. *Human Brain Mapping*, *38*(3), 1541–1573. https://doi.org/10.1002/hbm.23471

Ince, R. A. A., Jaworska, K., Gross, J., Panzeri, S., Rijsbergen, N. J. van, Rousselet, G. A., & Schyns, P. G. (2016). The Deceptively Simple N170 Reflects Network Information Processing Mechanisms Involving Visual Feature Coding and Transfer Across Hemispheres. *Cerebral Cortex*, *26*(11), 4123–4135. https://doi.org/10.1093/cercor/bhw196

Ince, R. A. A., Jaworska, K., Gross, J., Panzeri, S., Van Rijsbergen, N. J., Rousselet, G. A., & Schyns, P. G. (2016). The Deceptively Simple N170 Reflects Network Information Processing Mechanisms Involving Visual Feature Coding and Transfer Across Hemispheres. *Cerebral Cortex*, *26*(11), 4123–4135. https://doi.org/10.1093/cercor/bhw196

Ince, R. A. A., Jaworska, K., Gross, J., Panzeri, S., van Rijsbergen, N. J., Rousselet, G. A., & Schyns, P. G. (2016). The Deceptively Simple N170 Reflects Network Information Processing Mechanisms Involving Visual Feature Coding and Transfer Across Hemispheres. *Cerebral Cortex*, *26*(11), 4123–4135. https://doi.org/10.1093/cercor/bhw196

Ince, R. A. A., Kay, J. W., & Schyns, P. G. (2020). Bayesian inference of population prevalence. *bioRxiv*, 2020.07.08.191106. https://doi.org/10.1101/2020.07.08.191106

Ince, R. A. A., Kay, J. W., & Schyns, P. G. (2022). Within-participant statistics for cognitive science. *Trends in Cognitive Sciences*, *26*(8), 626–630. https://doi.org/10.1016/j.tics.2022.05.008

Ince, R. A. A., van Rijsbergen, N. J., Thut, G., Rousselet, G. A., Gross, J., Panzeri, S., & Schyns, P. G. (2015). Tracing the Flow of Perceptual Features in an Algorithmic Brain Network. *Scientific Reports*, *5*(1), 17681. https://doi.org/10.1038/srep17681

Ince, R. A., Paton, A. T., Kay, J. W., & Schyns, P. G. (2021). Bayesian inference of population prevalence. *eLife*, *10*, e62461. https://doi.org/10.7554/eLife.62461

Itier, R. J., & Preston, F. (2018). Increased Early Sensitivity to Eyes in Mouthless Faces: In Support of the LIFTED Model of Early Face Processing. *Brain Topography*, *31*(6), 972–984. https://doi.org/10.1007/s10548-018-0663-6

Jack, R. E., & Schyns, P. G. (2017). Toward a Social Psychophysics of Face Communication. *Annual Review of Psychology*, *68*(1), 269–297. https://doi.org/10.1146/annurev-psych-010416-044242

Jaworska, K., Yan, Y., van Rijsbergen, N. J., Ince, R. A., & Schyns, P. G. (2022). Different computations over the same inputs produce selective behavior in algorithmic brain networks. *eLife*, *11*, e73651. https://doi.org/10.7554/eLife.73651

Johnston, K., & Everling, S. (2006). Neural Activity in Monkey Prefrontal Cortex Is Modulated by Task Context and Behavioral Instruction during Delayed-match-to-sample and Conditional Prosaccade—Antisaccade Tasks. *Journal of Cognitive Neuroscience*, *18*(5), 749–765. https://doi.org/10.1162/jocn.2006.18.5.749

Ju, H., & Bassett, D. S. (2020). Dynamic representations in networked neural systems. *Nature Neuroscience*, *23*(8), 908–917. https://doi.org/10.1038/s41593-020-0653-3

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception. *The Journal of Neuroscience*, *17*(11), 4302. https://doi.org/10.1523/JNEUROSCI.17-11-04302.1997

Kanwisher, N., & Wojciulik, E. (2000). Visual attention: Insights from brain imaging. *Nature Reviews Neuroscience*, *1*(2), 91–100. https://doi.org/10.1038/35039043

Kastner, S., & Pinsk, M. A. (2004). Visual attention as a multilevel selection process. *Cognitive, Affective, & Behavioral Neuroscience*, *4*(4), 483–500. https://doi.org/10.3758/CABN.4.4.483

Kaufmann, J. M., Schweinberger, S. R., & Burton, A. M. (2008). N250 ERP Correlates of the Acquisition of Face Representations across Different Images. *Journal of Cognitive Neuroscience*, *21*(4), 625–641. https://doi.org/10.1162/jocn.2009.21080

Kay, K., Bonnen, K., Denison, R. N., Arcaro, M. J., & Barack, D. L. (2023). Tasks and their role in visual neuroscience. *Neuron*, S0896627323002180. https://doi.org/10.1016/j.neuron.2023.03.022

Kay, K. N. (2011). Understanding Visual Representation by Developing Receptive-Field Models. In N. Kriegeskorte & G. Kreiman (Eds.), *Visual Population Codes: Toward a Common Multivariate Framework for Cell Recording and Functional Imaging* (p. 0). The MIT Press. https://doi.org/10.7551/mitpress/8404.003.0009

Kay, K. N., Weiner, K. S., & Grill-Spector, K. (2015). Attention Reduces Spatial Uncertainty in Human Ventral Temporal Cortex. *Current Biology*, *25*(5), 595–600. https://doi.org/10.1016/j.cub.2014.12.050

Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K. A., Cichy, R. M., Hauk, O., & Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, *116*(43), 21854–21863. https://doi.org/10.1073/pnas.1905544116

Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: Integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, *17*(8), 401–412. https://doi.org/10.1016/j.tics.2013.06.007

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., & Bandettini, P. A. (2008). Matching Categorical Object Representations in Inferior

Temporal Cortex of Man and Monkey. *Neuron*, *60*(6), 1126–1141. https://doi.org/10.1016/j.neuron.2008.10.043

Lancaster, J. L., Woldorff, M. G., Parsons, L. M., Liotti, M., Freitas, C. S., Rainey, L., Kochunov, P. V., Nickerson, D., Mikiten, S. A., & Fox, P. T. (2000). Automated Talairach atlas labels for functional brain mapping. *Human Brain Mapping*, *10*(3), 120–131. https://doi.org/10.1002/1097-0193(200007)10:3<120::aid-hbm30>3.0.co;2-8

Lawrence, S. J. D., Formisano, E., Muckli, L., & de Lange, F. P. (2019). Laminar fMRI: Applications for cognitive neuroscience. *NeuroImage*, *197*, 785–791. https://doi.org/10.1016/j.neuroimage.2017.07.004

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788–791. https://doi.org/10.1038/44565

Lin, R., Meng, X., Chen, F., Li, X., Jensen, O., Theeuwes, J., & Wang, B. (2024). Neural evidence for attentional capture by salient distractors. *Nature Human Behaviour*, *8*(5), 932–944. https://doi.org/10.1038/s41562-024-01852-5

Lizier, J. T., Prokopenko, M., & Zomaya, A. Y. (2008). Local information transfer as a spatiotemporal filter for complex systems. *Physical Review E*, *77*(2), 026110. https://doi.org/10.1103/PhysRevE.77.026110

Lizier, J. T., Prokopenko, M., & Zomaya, A. Y. (2012). Local measures of information storage in complex distributed computation. *Information Sciences*, *208*, 39–54. https://doi.org/10.1016/j.ins.2012.04.016

Lizier, J. T., Prokopenko, M., & Zomaya, A. Y. (2014). A Framework for the Local Information Dynamics of Distributed Computation in Complex Systems. In M. Prokopenko (Ed.), *Guided Self-Organization: Inception* (pp. 115–158). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-53734-9_5

Logothetis, N. K., & Sheinberg, D. L. (1996). Visual Object Recognition. In *Annual Review of Neuroscience* (Vol. 19, Issue Volume 19, 1996, pp. 577–621). Annual Reviews. https://doi.org/10.1146/annurev.ne.19.030196.003045

Luck, S. J., & Kappenman, E. S. (Eds.). (2012). *Oxford handbook of event-related potential components*. Oxford University Press.

Luppi, A. I., Mediano, P. A. M., Rosas, F. E., Holland, N., Fryer, T. D., O'Brien, J. T., Rowe, J. B., Menon, D. K., Bor, D., & Stamatakis, E. A. (2022). A synergistic core for human brain evolution and cognition. *Nature Neuroscience*, *25*(6), 771–782. https://doi.org/10.1038/s41593-022-01070-0

Malcolm, G. L., Nuthmann, A., & Schyns, P. G. (2014). Beyond Gist: Strategic and Incremental Information Accumulation for Scene Categorization. *Psychological Science*, *25*(5), 1087–1097. https://doi.org/10.1177/0956797614522816

Margalit, E., Jamison, K. W., Weiner, K. S., Vizioli, L., Zhang, R.-Y., Kay, K. N., & Grill-Spector, K. (2020). Ultra-high-resolution fMRI of Human Ventral Temporal Cortex Reveals Differential Representation of Categories and Domains. *Journal of Neuroscience*, *40*(15), 3008–3024. https://doi.org/10.1523/JNEUROSCI.2106-19.2020

Marr, D., & Ullman, S. (2010). *Vision: A computational investigation into the human representation and processing of visual information*. MIT Press.

Martínez, A., Anllo-Vento, L., Sereno, M. I., Frank, L. R., Buxton, R. B., Dubowitz, D. J., Wong, E. C., Hinrichs, H., Heinze, H. J., & Hillyard, S. A. (1999). Involvement of striate and extrastriate visual cortical areas in spatial attention. *Nature Neuroscience*, *2*(4), Article 4. https://doi.org/10.1038/7274

Martinez-Cancino, R., Heng, J., Delorme, A., Kreutz-Delgado, K., Sotero, R. C., & Makeig, S. (2018). Measuring transient phase-amplitude coupling using local mutual information. *NeuroImage*. https://doi.org/10.1016/j.neuroimage.2018.10.034

Mashour, G. A., Roelfsema, P., Changeux, J.-P., & Dehaene, S. (2020). Conscious Processing and the Global Neuronal Workspace Hypothesis. *Neuron*, *105*(5), 776–798. https://doi.org/10.1016/j.neuron.2020.01.026

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review*, *88*(5), 375–407. https://doi.org/10.1037/0033-295X.88.5.375

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon Statistical Significance. *The American Statistician*, *73*(sup1), 235–245. https://doi.org/10.1080/00031305.2018.1527253

Moore, T., & Zirnsak, M. (2017). Neural Mechanisms of Selective Visual Attention. *Annual Review of Psychology*, *68*(1), 47–72. https://doi.org/10.1146/annurev-psych-122414-033400

Moran, J., & Desimone, R. (1985). Selective Attention Gates Visual Processing in the Extrastriate Cortex. *Science*, *229*(4715), 782–784. https://doi.org/10.1126/science.4023713

Murray, R. F. (2011). Classification images: A review. *Journal of Vision*, *11*(5), 2–2. https://doi.org/10.1167/11.5.2

Naitzat, G., Zhitnikov, A., & Lim, L.-H. (n.d.). *Topology of deep neural networks*.

Nemrodov, D., Anderson, T., Preston, F. F., & Itier, R. J. (2014). Early sensitivity for eyes within faces: A new neuronal account of holistic and featural processing. *NeuroImage*, *97*, 81–94. https://doi.org/10.1016/j.neuroimage.2014.04.042

Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping*, *15*(1), 1–25. https://doi.org/10.1002/hbm.1058

Nichols, T., & Hayasaka, S. (2003). Controlling the familywise error rate in functional neuroimaging: A comparative review. *Statistical Methods in Medical Research*, *12*(5), 419–446. https://doi.org/10.1191/0962280203sm341ra

Niedenthal, P. M., Mermillod, M., Maringer, M., & Hess, U. (2010). *The Simulation of Smiles (SIMS) model: Embodied simulation and the meaning of facial expression*. https://doi.org/10.18452/23061

Nieuwenhuis, S., Aston-Jones, G., & Cohen, J. D. (2005). Decision making, the P3, and the locus coeruleus—Norepinephrine system. *Psychological Bulletin*, *131*(4), 510–532. https://doi.org/10.1037/0033-2909.131.4.510

Noesselt, T., Hillyard, S. A., Woldorff, M. G., Schoenfeld, A., Hagner, T., Jäncke, L., Tempelmann, C., Hinrichs, H., & Heinze, H.-J. (2002). Delayed Striate Cortical Activation during Spatial Attention. *Neuron*, *35*(3), 575–587. https://doi.org/10.1016/S0896-6273(02)00781-X

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57. https://doi.org/10.1037/0096-3445.115.1.39

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2010). FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data. *Computational Intelligence and Neuroscience*, *2011*, e156869. https://doi.org/10.1155/2011/156869

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data. *Computational Intelligence and Neuroscience*, *2011*, 1–9. https://doi.org/10.1155/2011/156869

Pernet, C. R., Sajda, P., & Rousselet, G. A. (2011). Single-Trial Analyses: Why Bother? *Frontiers in Psychology*, *2*. https://doi.org/10.3389/fpsyg.2011.00322

Poldrack, R. A. (2017). Precision Neuroscience: Dense Sampling of Individual Brains. *Neuron*, *95*(4), 727–729. https://doi.org/10.1016/j.neuron.2017.08.002

Poldrack, R. A. (2021). The physics of representation. *Synthese*, *199*(1–2), 1307–1325. https://doi.org/10.1007/s11229-020-02793-y

Popivanov, I. D., Schyns, P. G., & Vogels, R. (2016). Stimulus features coded by single neurons of a macaque body category selective patch. *Proceedings of the National Academy of Sciences*, *113*(17), E2450–E2459. https://doi.org/10.1073/pnas.1520371113

Ratcliff, R., Philiastides, M. G., & Sajda, P. (2009). Quality of evidence for perceptual decision making is indexed by trial-to-trial variability of the EEG. *Proceedings of the National Academy of Sciences*, *106*(16), 6539–6544. https://doi.org/10.1073/pnas.0812589106

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372–422. https://doi.org/10.1037/0033-2909.124.3.372

Recchia, G., & Jones, M. N. (2009). More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis. *Behavior Research Methods*, *41*(3), 647–656. https://doi.org/10.3758/BRM.41.3.647

Rhodes, G., Jeffery, L., Boeing, A., & Calder, A. J. (2013). Visual coding of human bodies: Perceptual aftereffects reveal norm-based, opponent coding of body identity. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(2), 313–317. https://doi.org/10.1037/a0031568

Rhodes, S., & Cowan, N. (2018). Attention in working memory: Attention is needed but it yearns to be free. *Annals of the New York Academy of Sciences*, *1424*(1), 52–63. https://doi.org/10.1111/nyas.13652

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

Rossion, B., Joyce, C. A., Cottrell, G. W., & Tarr, M. J. (2003). Early lateralization and orientation tuning for face, word, and object processing in the visual cortex. *NeuroImage*, *20*(3), 1609–1624. https://doi.org/10.1016/j.neuroimage.2003.07.010

Rousselet, G. A., Ince, R. A. A., Rijsbergen, N. J. van, & Schyns, P. G. (2014). Eye coding mechanisms in early human face event-related potentials. *Journal of Vision*, *14*(13), 7–7. https://doi.org/10.1167/14.13.7

Rousselet, G. A., Macé, M. J.-M., & Fabre-Thorpe, M. (2004). Animal and human faces in natural scenes: How specific to human faces is the N170 ERP component? *Journal of Vision*, *4*(1), 2–2. https://doi.org/10.1167/4.1.2

Roy, J. E., Riesenhuber, M., Poggio, T., & Miller, E. K. (2010). Prefrontal Cortex Activity during Flexible Categorization. *Journal of Neuroscience*, *30*(25), 8519–8528. https://doi.org/10.1523/JNEUROSCI.4837-09.2010

Rubin, D. C. (2022). A conceptual space for episodic and semantic memory. *Memory & Cognition*, *50*(3), 464–477. https://doi.org/10.3758/s13421-021-01148-3

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, *115*(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Rust, N. C., & Movshon, J. A. (2005). In praise of artifice. *Nature Neuroscience*, *8*(12), 1647–1650. https://doi.org/10.1038/nn1606

Schreiber, T. (2000). Measuring Information Transfer. *Physical Review Letters*, *85*(2), 461–464. https://doi.org/10.1103/PhysRevLett.85.461

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Geiger, F., Schmidt, K., Yamins, D. L. K., & DiCarlo, J. J. (2018). *Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like?* [Preprint]. Neuroscience. https://doi.org/10.1101/407007

Schyns, P. G. (2018). Object Recognition: Complexity of Recognition Strategies. *Current Biology*, *28*(7), R313–R315. https://doi.org/10.1016/j.cub.2018.02.059

Schyns, P. G., Bonnar, L., & Gosselin, F. (2002). Show me the features! Understanding recognition from the use of visual information. *Psychological Science*, *13*(5), 402–409.

Schyns, P. G., Goldstone, R. L., & Thibaut, J.-P. (1998). The development of features in object concepts. *Behavioral and Brain Sciences*, *21*(1), 1–17. https://doi.org/10.1017/S0140525X98000107

Schyns, P. G., Gosselin, F., & Smith, M. L. (2009). Information processing algorithms in the brain. *Trends in Cognitive Sciences*, *13*(1), 20–26. https://doi.org/10.1016/j.tics.2008.09.008

Schyns, P. G., Jentzsch, I., Johnson, M., Schweinberger, S. R., & Gosselin, F. (2003). A principled method for determining the functionality of brain responses. *NeuroReport*, *14*(13), 1665–1669.

Schyns, P. G., & Oliva, A. (1994). From blobs to boundary edges: Evidence for time-and spatial-scale-dependent scene recognition. *Psychological Science*, *5*(4), 195–200.

Schyns, P. G., & Oliva, A. (1999). Dr. Angry and Mr. Smile: When categorization flexibly modifies the perception of faces in rapid visual presentations. *Cognition*, *69*(3), 243–265.

Schyns, P. G., Petro, L. S., & Smith, M. L. (2007). Dynamics of visual information integration in the brain for categorizing facial expressions. *Current Biology*, *17*(18), 1580–1585.

Schyns, P. G., & Rodet, L. (1997). Categorization creates functional features. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(3), 681–696. https://doi.org/10.1037/0278-7393.23.3.681

Schyns, P. G., Snoek, L., & Daube, C. (2022). Degrees of algorithmic equivalence between the brain and its DNN models. *Trends in Cognitive Sciences*, *26*(12), 1090–1102. https://doi.org/10.1016/j.tics.2022.09.003

Self, M. W., van Kerkoerle, T., Goebel, R., & Roelfsema, P. R. (2019). Benchmarking laminar fMRI: Neuronal spiking and synaptic activity during top-down and bottom-up processing in the different layers of cortex. *NeuroImage*, *197*, 806–817. https://doi.org/10.1016/j.neuroimage.2017.06.045

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences*, *104*(15), 6424–6429. https://doi.org/10.1073/pnas.0700622104

Shannon, C. E. (1948). *A Mathematical Theory of Communication*.

Shashidhara, S., Mitchell, D. J., Erez, Y., & Duncan, J. (2019). Progressive Recruitment of the Frontoparietal Multiple-demand System with Increased Task Complexity, Time Pressure, and Reward. *Journal of Cognitive Neuroscience*, *31*(11), 1617–1630. https://doi.org/10.1162/jocn_a_01440

Shiffrin, R. M., & Gardner, G. T. (1972). Visual processing capacity and attentional control. *Journal of Experimental Psychology*, *93*(1), 72–82. https://doi.org/10.1037/h0032453

Slotnick, S. D. (2018). The experimental parameters that affect attentional modulation of the ERP C1 component. *Cognitive Neuroscience*, *9*(1–2), 53–62. https://doi.org/10.1080/17588928.2017.1369021

Smith, F. W., & Muckli, L. (2010). Nonstimulated early visual areas carry information about surrounding context. *Proceedings of the National Academy of Sciences*, *107*(46), 20099–20103. https://doi.org/10.1073/pnas.1000233107

Smith, M. L., Gosselin, F., & Schyns, P. G. (2004). Receptive Fields for Flexible Face Categorizations. *Psychological Science*, *15*(11), 753–761. https://doi.org/10.1111/j.0956-7976.2004.00752.x

Smith, M. L., Gosselin, F., & Schyns, P. G. (2012). Measuring Internal Representations from Behavioral and Brain Data. *Current Biology*, *22*(3), 191–196. https://doi.org/10.1016/j.cub.2011.11.061

Stephan, K. E., Petzschner, F. H., Kasper, L., Bayer, J., Wellstein, K. V., Stefanics, G., Pruessmann, K. P., & Heinzle, J. (2019). Laminar fMRI and computational theories of brain function. *NeuroImage*, *197*, 699–706. https://doi.org/10.1016/j.neuroimage.2017.11.001

Teichmann, L., Hebart, M. N., & Baker, C. I. (2023). *Dynamic representation of multidimensional object properties in the human brain*. https://doi.org/10.1101/2023.09.08.556679

Urai, A. E., Gee, J. W. de, Tsetsos, K., & Donner, T. H. (2019, July 2). *Choice history biases subsequent evidence accumulation*. eLife. https://doi.org/10.7554/eLife.46331

van Moorselaar, D., Foster, J. J., Sutterer, D. W., Theeuwes, J., Olivers, C. N. L., & Awh, E. (2018). Spatially Selective Alpha Oscillations Reveal Moment-by-Moment Trade-offs between Working Memory and Attention. *Journal of Cognitive Neuroscience*, *30*(2), 256–266. https://doi.org/10.1162/jocn_a_01198

van Rijsbergen, N., Jaworska, K., Rousselet, G. A., & Schyns, P. G. (2014). With Age Comes Representational Wisdom in Social Signals. *Current Biology*, *24*(23), 2792–2796. https://doi.org/10.1016/j.cub.2014.09.075

VanRullen, R., & Thorpe, S. J. (2001). The Time Course of Visual Processing: From Early Perception to Decision-Making. *Journal of Cognitive Neuroscience*, *13*(4), 454–461. https://doi.org/10.1162/08989290152001880

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. ukasz, & Polosukhin, I. (2017). Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Wibral, M., Lizier, J. T., & Priesemann, V. (2015). Bits from Brains for Biologically Inspired Computing. *Frontiers in Robotics and AI*, *2*. https://doi.org/10.3389/frobt.2015.00005

Wibral, M., Lizier, J., Vögler, S., Priesemann, V., & Galuske, R. (2014). Local active information storage as a tool to understand distributed neural information processing. *Frontiers in Neuroinformatics*, *8*, 1. https://doi.org/10.3389/fninf.2014.00001

Williams, P. L., & Beer, R. D. (2010). *Nonnegative Decomposition of Multivariate Information* (No. arXiv:1004.2515). arXiv. http://arxiv.org/abs/1004.2515

Yan, Y., Zhan, J., Garrod, O., Cui, X., Ince, R. A. A., & Schyns, P. G. (2023). Strength of predicted information content in the brain biases decision behavior. *Current Biology*, *33*(24), 5505-5514.e6. https://doi.org/10.1016/j.cub.2023.10.042

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: Analysis by synthesis? *Trends in Cognitive Sciences*, *10*(7), 301–308. https://doi.org/10.1016/j.tics.2006.05.002

Zhan, J., Garrod, O. G. B., Van Rijsbergen, N., & Schyns, P. G. (2019). Modelling face memory reveals task-generalizable representations. *Nature Human Behaviour*, *3*(8), 817–826. https://doi.org/10.1038/s41562-019-0625-3

Zhan, J., Ince, R. A. A., van Rijsbergen, N., & Schyns, P. G. (2019). Dynamic Construction of Reduced Representations in the Brain for Perceptual Decision Behavior. *Current Biology*, *29*(2), 319-326.e4. https://doi.org/10.1016/j.cub.2018.11.049