How we imagine: insights from single neuron recordings in the human brain

Thesis by

Varun Spenta Wadia

In Partial Fulfillment of the Requirements for
the degree of
Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2024
(Defended August 4th, 2023)

Varun Wadia
ORCID: 0009-0009-5401-5367

DEDICATION

To my grandparents,

Rustom Dhunjisha Wadia,
Nergis Rustom Wadia,
Chevdichi Kamanat Ramachandran,
Jayasree Ramachandran.

Only one of whom lives to see this day, but all of whom instilled discipline, curiosity, and a respect for intellectual pursuit in their children, which kicked off a long chain of events that eventually led to this work. I can only hope that they would be proud.

ACKNOWLEDGEMENTS

I have imagined what this section might look like many times over the last 6 years. I will admit it is rather overwhelming to finally be writing these words. The list of people that have been instrumental in getting me to this point is very long. I will do my best to cover them all.

I would be remiss if I did not, first and foremost, thank my advisors Doris Tsao and Ueli Rutishauser: Doris, for giving me the courage to ask big questions, for pushing me to develop a strong enough work ethic to tackle said questions, for agreeing to open an entirely new research direction in her lab to support my independent ideas, and for generally being a huge inspiration as a thinker and person; Ueli, for always being incredibly generous with his time and attention, for keeping my academic ship steady through life's ups and downs, for teaching me almost all the neuroscience I know, and for becoming a very dear friend in the process.

I am also indebted to my other committee members, Markus Meister and Ralph Adolphs, for the knowledge they have passed on to me both in the classroom and in conversation.

This work stands upon the shoulders of Le Chang, Pinglei Bao, and Liang She. Le's landmark paper in 2017 distilled the axis model to understand facial recognition, Pinglei extended that work to general objects, and Liang has done seminal work to understand how familiarity affects those representations. All 3 have been incredibly helpful and this work would not have been possible without them.

A little help at the right time is much more valuable than a lot of help when one is not ready. The following people made a difference in key moments along the way and deserve a special mention: Francisco Luongo, Srinivas Chivukula, Sumner Norman, Tomas Aquino, Janis Hesse, Juri Minhxa, Jonathan Daume, Brooks Fu, and Hristos Courellis. Francisco for mentoring me during my early years and encouraging me to join the Tsao lab; Sri for opening my eyes to the world of invasive human recordings and getting my project off the ground; Sumner for being both a scientific, and emotional support from day 1 — helping me write better code, answering my math questions, and most importantly, getting me to believe in myself; Tomas for patiently teaching me the details of human electrophysiology from scratch; Janis for being a life support when I was drowning in work I did not know how to do; Juri for

always being willing to patiently teach me concepts and for encouraging me to read more broadly and develop as a thinker; and Jonathan, Brooks, and Hristos for unending technical insight towards the end and making the back half of graduate school a lot of fun.

Special thanks goes to all of my wonderful friends.

I am in an absolutely blessed position wherein it is difficult to name everyone I have formed a deep bond with during my time here, though they can roughly be divided into:

- My home away from home: Alexandros, Jess, Carl, Maria, Vinnie, Noah, Sumner, and Zsófi.
- My fabulous band: Andrew, Cam, David, James, and Janis.
- My BBE community: Frank, Guru, Yuelin, Jialiang, Krithika, Sarah, Whitney, Han, Abdullah, Kelly, David, Adi, Erin, Joe, Uriah, Marky Sikora, and Helen O'Connor.
- The International Students Program office, especially Laura Flower-Kim.
- The Dining staff: Karly, Lucy, Miguel, and José.
- The Janitorial staff: Sonya, Rob, Junior, and many others.
- My childhood bhonduz: Ajan, Rohit, Shawn, Tanooj, Tushar, and Vijit who are my first and last line of defense when life gets difficult.
- Other wonderful people I met along the way: Becca, Maegan, Noosha, and Xuan.
- The Tsao & Rutishauser lab members

Others that deserve thanks include:

Vipul Mehra and Centia D'Souza (now Centia Mehra). High school teachers who took a restless, arrogant teenager and put him on a path to maturity and self-improvement.

Professor Vidita Vaidya, who planted the seed of neuroscience in me 16 years ago. Professors Richmond Thompson, Mary Lou Zeeman, Stephen Naculich, and Ian Ellwood, who nurtured that sapling several years later.

Our patients, who show a lot of kindness and grace in moments of immense personal difficulty to make this scientific endeavor possible.

The fantastic clinical team at Cedars-Sinai: Doctors Adam Mamelak, Chrystal Reed, Jeffrey Chung, and Lisa Bateman.

Finally, my wonderful family: my parents Leena and Spenta, and my older sister Neha who embodies the best of my parents and friends rolled into one. To have parents and an older sibling that love you unconditionally, that you trust completely, and that you can talk to about anything is an absolute privilege that I did not earn, but will always be grateful for. I learnt from a very young age that my parents were scientists, and while it was admittedly exasperating to have tiny questions met with incredibly long answers full of mechanistic detail as a child, over time I learnt a valuable lesson: to understand how something works is to fully appreciate its beauty. I would not change a single thing about my upbringing for the world, and I am acutely aware of how rare that sentiment is.

ABSTRACT

As human beings interact with the world, we sample it, build representations of its underlying structure, and subsequently use that knowledge for future reasoning - also called modeling the world. This model is generative, allowing our knowledge of the world to influence our perception and in extreme cases induce perception in the absence of any external stimulus. This phenomenon called mental imagery is a remarkable cognitive ability that allows us to remember previous experiences, imagine new ones, make plans, and solve problems. In the visual domain, our ability to generate visual percepts without external stimulation is the basis of our memory for experiences, as episodic memories are simply a subset of all possible experiences we could imagine. Animal studies have yielded rich insight into bottom-up visual processing, from the first discovery of neurons that respond to complex objects like faces to determining the precise code for general objects in macaque inferotemporal (IT) cortex. However, the neural mechanisms of internally generated top-down processing have been much more elusive. Here we present findings on the mechanisms of visual imagery at single neuron resolution. We approached deciphering visual imagery by first laying out coding principles for object perception and then directly comparing responses during viewing to subsequent imagery of those images. We recorded 384 visually responsive neurons in inferotemporal (IT) cortex of 12 epilepsy patients as they viewed and subsequently imagined carefully parametrized visual objects. We verified that neurons in IT cortex are 'axis tuned', i.e. as in macaques they represent visual objects by encoding specific axes that span a high dimensional object feature space. 218/384 visually responsive neurons (~58%) were axis tuned, and the axis model explained more variance than other models tested. Armed with this code for visual objects we examined neural responses during pure imagery in the same neurons. We demonstrate robust reactivation of individual neurons across the brain (~35% of neurons across the brain and ~50% of neurons in IT cortex) and a recapitulation of viewing stimulus preference during pure visual imagery in IT. By first uncovering the code for visual objects and examining it during imagery, we demonstrate that neurons in IT cortex subserve visual imagery by reinstating visual context. This study marks the first detailed exploration of visual perception and imagery in the human brain.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS AND/OR TABLES

Number *Page*

NOMENCLATURE

IT: Inferotemporal cortex

SNR: Signal to noise ratio

PSTH:. Peri-stimulus time histogram

fMRI: Functional magnetic resonance imaging

PCA: Principal component analysis

ISI: Inter-spike interval.

TTL: Transistor-transistor logic

GAN: Generative adversarial network

FFA: Fusiform face area

ACC: Anterior cingulate cortex

Pre-SMA: Pre-supplementary motor area

AMY: Amygdala

HIPP:. Hippocampus

CMT: Centro-medial hypothalamus

OFC: Orbito-frontal cortex

AINS: Anterior insula

PHG: Parahippocamal gyrus

Chapter 1

INTRODUCTION

Intelligent beings possess an internal model of the world. This model is developed through active interactions with the environment and is evident at a very young age. In humans, various psychological studies have demonstrated that infants just a few months of age have thoughts structured around both the intuitive physics governing objects and their interactions[1–3], and the intuitive psychology enabling inference of an agent's goals[4,5]. This model is also generative, influencing perception by filling in the gaps left by sparse data[6–11] and allowing for imagination of new possibilities, experiences, or outcomes in the absence of any sense data. Such a model, while present in other intelligent species such as crows[12–14] and chimpanzees[4,15–17], is highly enriched and extended in humans.

We claim that modelling the world involves 3 key steps, the first being core object recognition. It is impossible to model the world without knowing what is in it. Therefore, an understanding of model building necessarily requires an understanding of how we segment the world into the objects and agents we interact with, remember, dream about, and refer to. Humans are primarily visual creatures with a large swath of the primate brain being dedicated to the processing of visual information[18]. Thus, in most people this parsing of the world occurs in the visual domain.

The second step to building a model of the world is developing an understanding of all possible interactions of objects and agents (within and between groups), the aforementioned 'intuitive physics' and 'intuitive psychology' respectively[19]. Even from a very young age humans can understand how objects relate to each other and imagine various possible outcomes concerning object interaction. For example, we understand that a jenga tower which is top-heavy is less stable than one that is not. If we were told it got pushed from a given angle, we could predict which way it would fall and which pieces might hit the ground first. This is made possible by the fact that our model is generative, allowing us to integrate our knowledge of intuitive physics with our current sense data and use it to generate a hypothetical updated state of the jenga tower in the next instant (or the previous one). The same logic applies to agents, we can infer someone else's intentions just from their actions because we have knowledge of how intentions relate to actions. For example, if your friend ignores you as they walk by you on the street you can generate many possible explanations for the action: they were distracted and didn't notice you, they are upset with you, they didn't recognize you, etc.

Finally, the third step is to learn language and use language to learn everything else that does not have an obvious physical manifestation[20,21]. Essentially our model of the world is a bridge between the targets of perception, the substrates of action planning, and the substrates of language. In order to understand how we develop this model we claim that we need to investigate the neural mechanisms of all 3 of the previously outlined steps *and* isolate the generative processes that make them fast, flexible, and learnable with low training data.

This thesis presents an attempt to understand how human beings form such a model of the world by tackling step 1 of model building: core object recognition and the top down, generative processes that make it fast, flexible, and reliable even when the sensory data is noisy and ambiguous. Our investigations have focused on a region of the brain on the underside of the temporal lobe called inferotemporal (IT) cortex as this region has been shown over many decades to subserve visual object perception in macaques and humans[18,22–26], with lesions to it causing profound agnosias[24–26].

Our vision is an active process that goes far beyond pattern recognition. Visual perception is a closed loop that includes both feedforward and feedback (or bottom-up and top-down) pathways wherein an incoming image is broken down into a feature representation via the feedforward visual pathway, and the feedback pathway then attempts to synthesize the input image from its initial feedforward representation[7–10]. Essentially our visual perception is a form of inference wherein we fit an internal model to external reality that aids in the assignment of labels to objects. Mathematically, our internal models' influence on our perception can be codified as a prior in a Bayesian inference paradigm[8,9].

Take an input reaching the eyes or an image description 'I' and the scene that caused that input 's', our brains are trying to compute $P(s \mid I)$ or the most probable scene description that gave rise to the observed I, which using Bayes' rule can be computed by

$$P(s \mid I) = \frac{P(I \mid s)\, P(s)}{P(I)}$$

Here $P(I)$ are the image regularities which arise from the similarity between natural scenes like the statistics of edges and the distribution of contrasts. $P(I \mid s)$ is the probability of the image features given the state of the scene (or the 'likelihood' of the image features) and $P(s)$ is probability of the scene before perceiving the image features (or the 'prior' — given by the model).

Mental imagery is an extreme case of top-down processing in which the visualization of an object occurs in the absence of any viewed features, providing an excellent experimental paradigm to isolate the neural signature of a top-down signal. Animal studies have yielded rich insight into bottom-up processing of objects from the first discovery of individual neurons[22] and functionally connected patches[23] that respond to whole complex objects, to determining the precise code for faces[24] and general objects[25] in inferotemporal (IT) cortex. However, the neural mechanisms of internally generated top-down processing have been much more elusive, largely due to the immense difficulty involved in training an animal to reliably and repeatedly generate a specific mental image.

Thus, humans are the ideal test subjects in which to isolate this generative pathway as we can use language as a powerful tool to manipulate imagination. However, most previous studies of mental imagery in humans have been limited by an inability to record individual neurons as humans carry out mental imagery tasks. The proxies recorded by noninvasive imaging techniques have complicated relationships to underlying neural activity[31,32] and have very low spatial resolution. For example, a single fMRI voxel contains tens of thousands of neurons but a plethora of animal electrophysiology studies have shown that neighboring neurons, spaced less than on hundredth of a millimeter apart, can have very different response properties.

Neurosurgical patients implanted to detect the location of their focal epilepsy provide a rare glimpse into the human brain at high resolution[32–34]. Such settings have been leveraged to investigate visual responses to objects in IT cortex[35–37] with one even demonstrating reactivation of visually responsive neurons during imagery[37]. However, the precise code for visual objects and whether that code is respected during imagery has remained unknown until this writing.

Working in these clinical settings takes great care and an understanding that the patients' needs always come first. Given the difficulty in obtaining this data and the value it holds, Chapter II is a brief outline of how we start with continuous voltage traces recorded in the patient room during an experiment and end with claims about how individual neurons are modulated by task variables. In Chapter III we discuss findings pertaining to the code for visual objects in human IT cortex. We record 431 individual neurons in 12 patients as they view 500 carefully parametrized objects with highly varied features and use this data to examine the bottom-up code for visual objects. Chapter IV then ventures into the unknown territory of purely top-down processing: visual imagery. We record 173 individual neurons across 5 patients as they viewed and subsequently imagined a subset of the 500 objects used to investigate the visual code. We find robust reactivation of neurons across the brain, though a particularly large fraction of visually responsive neurons in IT cortex reactivated (~50%) in a manner that respected the visual code. In Chapter V we summarize the main findings of the thesis and briefly discuss future experiments that are underway to target steps 2 (contextual modulation of sensory representations) and 3 (language encoding) of model building respectively.

References:

1. Spelke, E.S., Breinlinger, K., Macomber, J., and Jacobson, K. (1992). Origins of knowledge. Psychol. Rev. *99*, 605–632.

2. Spelke, E.S. (2000). Core knowledge. Am. Psychol. *55*, 1233–1243.

3. Spelke, E.S., and Kinzler, K.D. (2007). Core knowledge. Dev. Sci. *10*, 89–96.

4. Warneken, F., and Tomasello, M. (2006). Altruistic helping in human infants and young chimpanzees. Science *311*, 1301–1303.

5. Warneken, F., and Tomasello, M. (2007). Helping and Cooperation at 14 Months of Age. Infancy *11*, 271–294.

6. Miller, E.K. (1999). Straight from the top. Nature Publishing Group UK. 10.1038/44291.

7. Rao, R.P., and Ballard, D.H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat. Neurosci. *2*, 79–87.

8. Kersten, D., Mamassian, P., and Yuille, A. (2004). Object perception as Bayesian inference. Annu. Rev. Psychol. *55*, 271–304.

9. Yuille, A., and Kersten, D. (2006). Vision as Bayesian inference: analysis by synthesis? Trends Cogn. Sci. *10*, 301–308.

10. Nguyen, T., Tsao, D., and Anandkumar, A. (2020). Neural networks with recurrent generative feedback. Advances in.

11. Kuhl, P.K., Ramírez, R.R., Bosseler, A., Lin, J.-F.L., and Imada, T. (2014). Infants' brain responses to speech suggest analysis by synthesis. Proc. Natl. Acad. Sci. U. S. A. *111*, 11238–11245.

12. Von Bayern, A.M.P., Heathcote, R.J.P., Rutz, C., and Kacelnik, A. (2009). The role of experience in problem solving and innovative tool use in crows. Curr. Biol.

13. Kenward, B., Rutz, C., Weir, A.A.S., and Kacelnik, A. (2006). Development of tool use in New Caledonian crows: inherited action patterns and social influences. Anim. Behav. *72*, 1329–1343.

14. Rutz, C., Bluff, L.A., Reed, N., Troscianko, J., Newton, J., Inger, R., Kacelnik, A., and Bearhop, S. (2010). The ecological significance of tool use in New Caledonian crows. Science *329*, 1523–1526.

15. Boesch, C., and Boesch, H. (1990). Tool use and tool making in wild chimpanzees. Folia Primatol. *54*, 86–99.

16. Tomasello, M., Davis-Dasilva, M., Camak, L., and Bard, K. (1987). Observational learning of tool-use by young chimpanzees. Hum. Evol. *2*, 175–183.

17. Biro, D., Inoue-Nakamura, N., Tonooka, R., Yamakoshi, G., Sousa, C., and Matsuzawa, T. (2003). Cultural innovation and transmission of tool use in wild chimpanzees: evidence from field experiments. Anim. Cogn. *6*, 213–223.

18. Ungerleider, L.G., and Haxby, J.V. (1994). "What" and "where" in the human brain. Curr. Opin. Neurobiol. *4*, 157–165.

19. Lake, B.M., Ullman, T.D., Tenenbaum, J.B., and Gershman, S.J. (2017). Building machines that learn and think like people. Behav. Brain Sci. *40*, e253.

20. Spelke, E.S. (2003). What makes us smart? Core knowledge and natural language. Language in mind: Advances in the study of.

21. Spelke, E.S. (2017). Core Knowledge, Language, and Number. Lang. Learn. Dev. *13*, 147–170.

22. Kanwisher, N., McDermott, J., and Chun, M.M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. J. Neurosci. *17*, 4302–4311.

23. Ishai, A., Ungerleider, L.G., and Haxby, J.V. (2000). Distributed neural systems for the generation of visual images. Neuron *28*, 979–990.

24. Shelton, P.A., Bowers, D., Duara, R., and Heilman, K.M. (1994). Apperceptive visual agnosia: a case study. Brain Cogn. *25*, 1–23.

25. Ridley, R.M., Warner, K.A., Maclean, C.J., Gaffan, D., and Baker, H.F. (2001). Visual agnosia and Klüver–Bucy syndrome in marmosets (Callithrix jacchus) following ablation of inferotemporal cortex, with additional mnemonic effects of immunotoxic lesions of cholinergic projections to medial temporal areas. Brain Res. *898*, 136–151.

26. Arguin, M. (1996). Shape Integration for Visual Object Recognition and Its Implication in Category-Specific Visual Agnosia. Vis. cogn. *3*, 221–276.

27. Gross, C.G., Rocha-Miranda, C.E., and Bender, D.B. (1972). Visual properties of neurons in inferotemporal cortex of the Macaque. J. Neurophysiol. *35*, 96–111.

28. Tsao, D.Y., Freiwald, W.A., Knutsen, T.A., Mandeville, J.B., and Tootell, R.B.H. (2003). Faces and objects in macaque cerebral cortex. Nat. Neurosci. *6*, 989–995.

29. Chang, L., and Tsao, D.Y. (2017). The Code for Facial Identity in the Primate Brain. Cell *169*, 1013-1028.e14.

30. Bao, P., She, L., McGill, M., and Tsao, D.Y. (2020). A map of object space in primate inferotemporal cortex. Nature *583*, 103–108.

31. Logothetis, N.K. (2002). The neural basis of the blood-oxygen-level-dependent functional magnetic resonance imaging signal. Philos. Trans. R. Soc. Lond. B Biol. Sci. *357*, 1003–1037.

32. Engel, A.K., Moll, C.K.E., Fried, I., and Ojemann, G.A. (2005). Invasive recordings from the human brain: clinical insights and beyond. Nat. Rev. Neurosci. *6*, 35–47.

33. Fried, I., Wilson, C.L., Maidment, N.T., Engel, J., Jr, Behnke, E., Fields, T.A., MacDonald, K.A., Morrow, J.W., and Ackerson, L. (1999). Cerebral microdialysis combined with single-neuron and electroencephalographic recording in neurosurgical patients. Technical note. J. Neurosurg. *91*, 697–705.

34. Fried, I., Rutishauser, U., Cerf, M., and Kreiman, G. (2014). Single Neuron Studies of the Human Brain: Probing Cognition (MIT Press).

35. Axelrod, V., Rozier, C., Malkinson, T.S., Lehongre, K., Adam, C., Lambrecq, V., Navarro, V., and Naccache, L. (2019). Face-selective neurons in the vicinity of the human fusiform face area. Neurology *92*, 197–198.

36. Axelrod, V., Rozier, C., Malkinson, T.S., Lehongre, K., Adam, C., Lambrecq, V., Navarro, V., and Naccache, L. (2022). Face-selective multi-unit activity in the proximity of the FFA modulated by facial expression stimuli. Neuropsychologia *170*, 108228.

37. Khuvis, S., Yeagle, E.M., Norman, Y., Grossman, S., Malach, R., and Mehta, A.D. (2021). Face-Selective Units in Human Ventral Temporal Cortex Reactivate during Free Recall. J. Neurosci. *41*, 3386–3399.

Chapter II

SCIENTIFIC OPPORTUNISM: Neurosurgical patients provide a rare glimpse into the human mind at high resolution.

The human brain has been studied (largely indirectly) with intense fascination for a very long time. Our minds and the brains that give rise to them are the basis of everything we experience in life. Our percepts, thoughts, intentions, feelings, and actions are all governed by the electrical activity of neurons in the brain. Thus, an understanding of the complex phenomena that make up the human condition necessarily requires an understanding of the neural activity that accompanies them.

Most of our understanding of how individual neurons work, or how they work in concert to give rise to behavior stems from causal manipulations of neural circuitry in animal models. While these have shed light on a plethora of behavioral phenomena, there remain a few out-of-reach frontiers that require a deep dive into the human brain such as language, imagination, or the subjective feeling of a conscious percept (qualia) to name a few.

Until very recently our best techniques for studying the human brain lacked the spatiotemporal resolution required to make direct inference about how individual neurons subserve complex phenomena. The proxies being recorded via non-invasive imaging techniques have complicated relationships to neural signals and often remain poorly understood[1,3]. However, neurosurgical patients being treated for ailments such as brain tumors, Parkinson's disease, or focal epilepsy provide the opportunity to peer into the human brain at high resolution[4] which in turn allows for a deeper understanding of the phenomena that make us human.

Overview of electrodes used:

This thesis will exclusively focus on data collected from patients implanted to localize their focal epilepsy. During their treatment process patients are implanted with stereo-EEG depth electrodes that consist of macroscopic electrode contacts along the shank and microwires that protrude from the end of the electrode[3–5,7] (Figure 2.1A). These microwires have a very high impedance (50-500kΩ) and thus allow for the recording of individual action potentials. The high sampling rate is necessary as sampling rates below 16khz can miss important aspects of individual action potential waveforms[5]. Upon implantation patients stay in the epilepsy monitoring unit for several days until they have enough seizures to allow the clinical team to discern the location of their seizure foci via the activity pattern on the electrodes. All the findings in this thesis are made possible by patients consenting to do experiments during this period, and for that the scientific community owes them a debt we will never be able to repay.

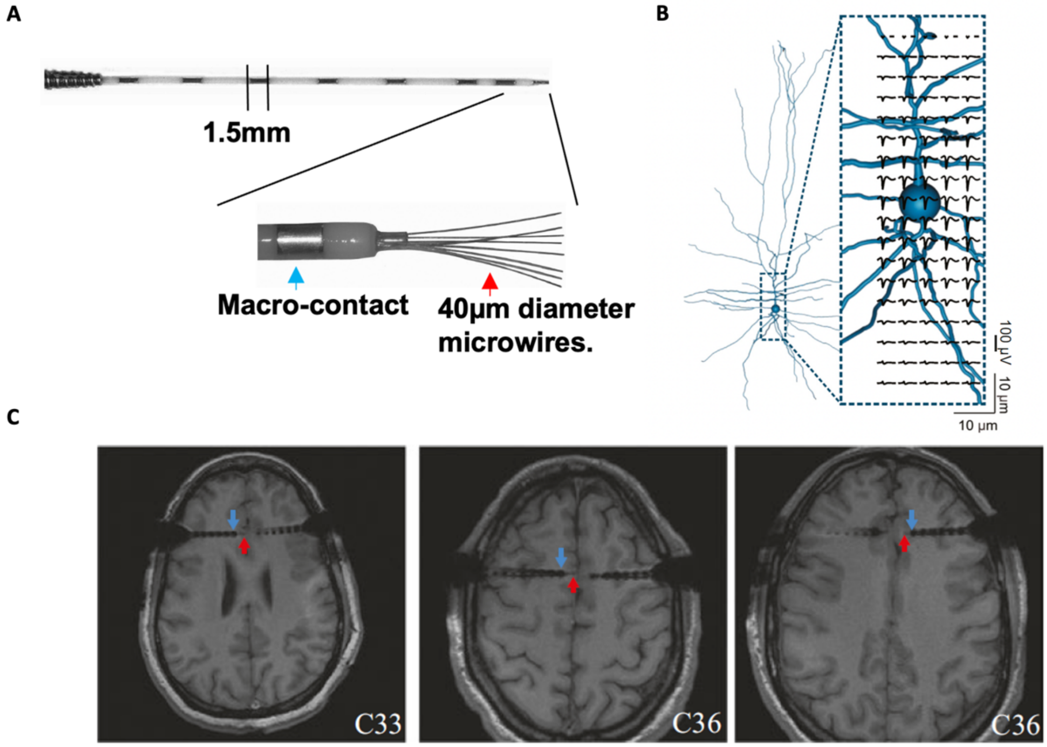


Figure 2.1. Research modification to clinical electrodes allow for recording of individual neurons in the human brain.

(A) A photograph of the micro-macro depth electrodes used during epilepsy monitoring. Named as such due to the presence of macro contacts (low impedance) to record over large areas for seizure detection, and microwires (high impedance) that pass through the inner shaft and protrude out of the end that can detect individual action potentials.

(B) Extracellular potential of a reconstructed human neuron from the middle temporal gyrus (Cell ID 569844159 in https://celltypes.brain-map.org). The shape of the extracellular action potential that would be recorded from various locations relative to the cell body are overlayed at those locations. The rapid decay of amplitude is noteworthy. Adapted from Mosher et al. 2020; Meisenhelter & Rutishauser, 2022.

(C) Post-operative MRI of a patient showing the localization of depth electrodes in the medial frontal cortex. The blue arrow marks the most medial macro contact and the red arrow marks the location of the microwires. Adapted from Fu & Rutishauser, 2023.

The patient room, stimulus presentation, and data acquisition

All tasks were implemented on Dell laptop with a 15 inch screen using Matlab's psychtoolbox[6]. The stimulus presentation computer is connected to the data acquisition system via a TTL cable (Figure 2.2B) and the tasks are programmed to include TTL pulses at meaningful points to make possible the alignment of neural data with the various task phases later. The data acquisition system used is the Atlas system from Neuralynx Inc. Signals from the microwires are amplified on the head with small pre-amplifiers that attach to the microelectrode and monitored broadband (0.1 – 9khz bandpass filter) throughout the experiment session. All parsing of the continuous data into a list of putative neurons and the timestamps of their action potentials (also called 'spikes') occurs offline after the experiment is complete. The methodology for such data processing or 'spike sorting' is briefly reviewed in the next section.

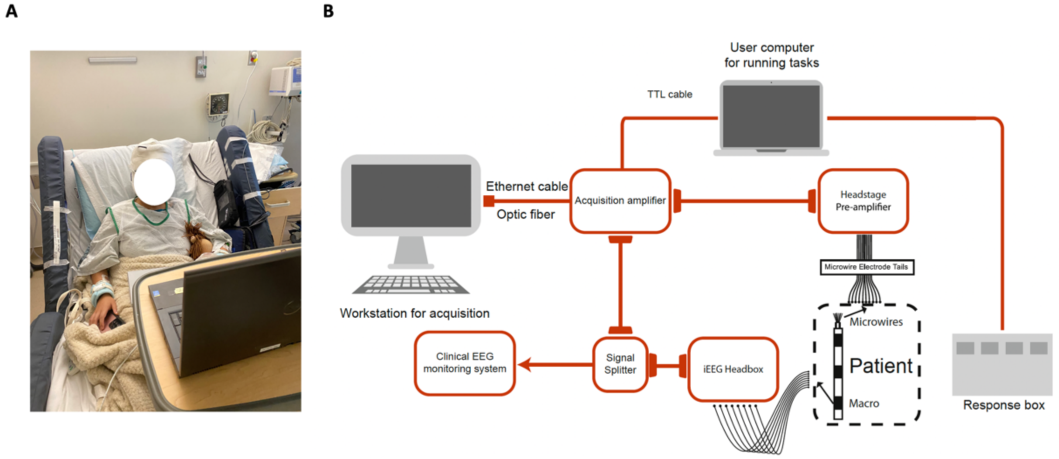


Figure 2.2. The patient room.

(A) A photograph of a patient undergoing testing. Stimuli are presented on a 15-inch Dell laptop, which is connected to the Atlas recording system via a transistor-transistor logic (TTL) cable to synchronize the stimulus presentation with the neural data.

(B) Schematic of the set up used to record simultaneous intracranial EEG and single neuron activity from epilepsy patients. Adapted from Fu & Rutishauser, 2023.

Parsing continuous voltage traces into meaningful units of information: basics of spike sorting

This section provides a technical perspective on all the procedures required for an experiment to succeed in providing scientific insight after the acquisition of continuously sampled data from the microwires. Notably, this section pertains to the time after data has been stored by acquisition system.

The microwires described earlier record extracellular voltage at a particular point in space as a function of time. This signal is well known to be a superposition of the currents generated by synaptic events, action potentials, and even antidromic action potentials[3]. As such, decomposing the extracellular signal into the components that give rise to it is a very difficult problem. However, the occurrence of action potentials near the tip of the microwires leads to the waveform being clearly distinguishable from the background voltage and with a high enough sampling rate one can distinguish different neurons via differences in the recorded waveforms. In rodents it is estimated that electrodes can distinguish extracellular spikes from neurons located as far as 140μm away from the tip of the electrode[8]. At the time of this thesis writing there are very few quantitative studies of the relationship between spike amplitude/shape and distance to the electrode in the human brain. The few that do exist rely on the fitting of multicompartment models to *in-vitro* whole cell recordings and using those to predict the shapes of extracellular waveforms as a function of electrode position[9,10]. To calibrate these models to ground truth one would need to record intracellularly and extracellularly from various positions around the neuron simultaneously, which is impossible *in-vivo* and has not been attempted in human tissue *in-vitro* to our knowledge at this time. These detailed analyses have been carried out in rodents[12,13] and since the size of neurons in rodents and humans are roughly the same it is reasonable to assume that the basic properties of extracellular recordings are applicable to human neurophysiology as well[5].

The pipeline for assigning spikes to putative neurons begins with filtering the signal to remove low frequency content from the raw trace, followed by spike detection via thresholding the filtered signal, and finally a template matching process to cluster similar waveforms as having come from a single putative neuron. At this stage it is important to note that this is an under constrained inference problem and there are many possible solutions. As such many algorithms have been developed to sort spikes ranging from manual to fully automatic[10], with each new one comparing its performance to the ones before it[14–20].

Filtering

Given that individual action potentials are characterized by a large but transient spike in voltage, thresholding is a key step in the spike detection process. In order to ensure that such thresholding captures only action potentials and not any other fluctuations in the extracellular voltage the first step in most spike sorting approaches is to filter the data to remove all low frequency content (< 300hz) from the raw trace (Figure 2.3A). Note that for analyses done offline it is important to not filter while the experiment is being done as this filtering introduces distortions in the waveform[21], this can be avoided offline by using a zero-phase non causal digital filter[22]. All analysis described in this thesis was conducted offline after the experiment session concluded.

Spike detection

The simplest way to detect spikes is by thresholding the bandpass filtered signal. However, as the amplitude of spikes is a function of distance from the electrode tip some recorded neurons have amplitudes that are closer to the noise floor and may be missed in such cases.

As such, a few other techniques have been developed to more effectively detect spikes. Some involve using the energy of the signal[17,23], others use wavelet based methods[16], with more recent approaches even using neural networks[19]. In the former case, the signal is convolved with a rectangular kernel that has the approximate width of a spike (1ms). The kernel acts as a matched filter that will suppress events of differing widths while amplifying events of a similar width. This energy signal is then thresholded to detect spikes[17] (Figure 2.3A, bottom panel). Early implementations of the latter wavelet-based methods used wavelets from well-studied families like the *bior* family that resemble spike waveforms while other more modern methods opt to learn the best basis in an unsupervised manner[24]. Wavelets yield better detection compared to simple rectangular kernels used[24–26] by the former energy-signal methods though these methods are much more computationally expensive[5].

Choosing an appropriate detection method based on the circumstances of the experiment such as the need to detect spikes online vs offline, power requirements, computational cost, and detection performance is crucial to the success of experiments in clinical settings and an important responsibility as the data is a real gift and all possible bits of information need to be gleaned from it. All neurons reported in this thesis were detected and sorted using the Osort algorithm[17] which uses energy signal methods for spike detection. After detection, spikes are aligned to a common reference frame so their shapes can be compared to all other waveforms detected on a given wire before clustering of waveforms into putative neurons (sorting) is completed.

Spike sorting

The relative position of the tip of the electrode and the cell body of the neuron directly affects the ion flow that the electrode will detect and thus shape of the detected waveform[2,9,27] (Figure 2.1B). Spike sorting aims to identify features that maximally separate waveforms from different neurons and cluster similar spikes that likely arose from the same neuron in an unsupervised manner. The most commonly used feature is the shape of the waveform, wherein spikes are either represented as points in an N-dimensional space where N is the number of samples that make up the waveform or as points in some lower dimensional space that captures most of the variance. Such lower dimensional spaces are built utilizing dimensionality reduction techniques such as principal component analysis (PCA).

Once spike features have been identified and parametrized, unsupervised partitioning of the space is done to separate waveforms from different putative neurons. Some approaches make assumptions about the underlying distribution of the data while others rely purely on heuristics computed directly from the data. The Osort algorithm used in this thesis was originally written to run online so it detects each incoming spike and assigns it to a cluster immediately, computing a distance metric between clusters to keep track of the different ones automatically.

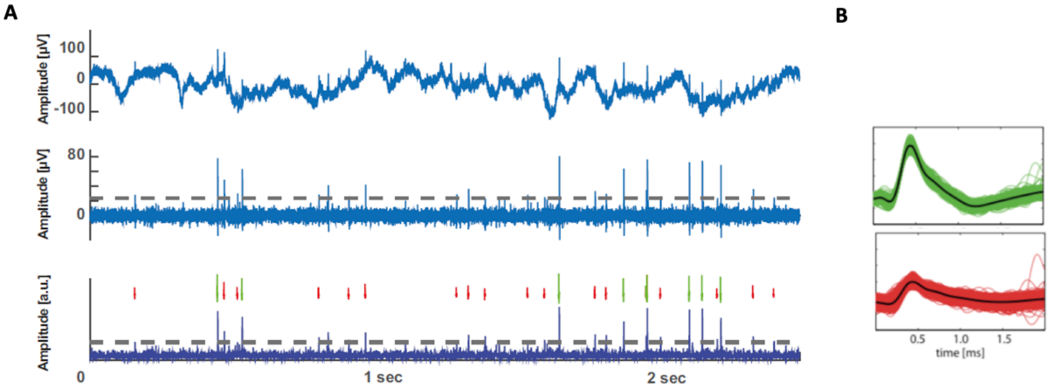


Figure 2.3. Decomposing extracellular signals into action potentials from putative neurons. Adapted from Minhxa et al., 2018.

(A) (Top) Example broadband recording from a single microwire (channel) in the human amygdala with bipolar referencing. (Middle) Same broadband signal bandpass filtered between 300-3000hz. (Bottom) The signal used for spike detection by Osort: the local energy computed using a 1ms kernel (roughly the width of an action potential) which amplifies all actional potential-like events and suppresses broader fluctuations greatly increasing the signal-to-noise of action potentials with respect to baseline.

(B) The two putative single neurons detected on this channel.

Quality metrics

The problem of identifying which action potential came from which neuron is a very challenging problem without a single best solution. Due to such inherent uncertainty all spike sorting approaches rely on the computation of various other quantitative metrics to assess the quality of sorting, and to delineate artifacts and neuronal spikes. These metrics can broadly be separated into two categories: single unit measures and comparisons between different units.

Single unit measures

These measures include the mean waveform, the spike times, and their derivatives. In the case of the mean waveform one also examines its variance and its signal-to-noise ratio (SNR). In the case of spike times useful derivative metrics include the distribution of the inter-spike interval (ISI) and the autocorrelation of spike times.

The SNR for a spike quantifies how different the waveform is from the background. It is useful to compute this for all spikes and then examine the variance of the SNR across all spikes in a given cluster. A large variance could imply the clustering of waveforms from multiple neurons or the inclusion of artifacts.

Osort clusters spikes based purely on properties of the waveform, thus the distribution of ISI is an independent metric that can be used to evaluate the sorting result. In particular, the refractory period of neurons should result in a dip in the autocorrelation of spike times[28] and very short ISIs (< 3ms) should be rare in a well isolated unit[29].

Separability of multiple units

Determining whether two units are truly distinct is difficult to do based on the waveforms alone. Osort implements a projection test to provide evidence of two units being separate. The projection test is based on the observation that, for two units to be statistically separable there needs to be a minimum distance between them that is a function of the amount of background noise[15]. If two units are separated by less than the required distance, it does not necessarily mean they are not different but that they cannot be distinguished at that level of noise. The projection test is a pairwise metric that defines the distance between a pair of units as a multiple of this minimally distinguishable distance. On average this is seen to be a fairly high number (> 10)[29] so setting a reasonable cutoff for this value (in this work we used 5) aids in determining whether two clusters should be considered different or not.

All spike sorting algorithms require a manual curation of the clusters generated, and occasional post processing before the final putative neurons from the recorded channels are chosen.

Inference:

After sorting is complete, we are left with a collection of putative neurons recorded in that session and for each neuron, a list of times at which it fired an action potential. That list of times is aligned to the relevant task phases via comparison to the time stamps of the TTL pulses sent to the Atlas acquisition system during the experiment which allows us to subsequently infer the activity of each neuron in relation to task variables.

Discussion:

The confluence of advances in electrode fabrication, surgical techniques to safely implant those electrodes, and computer systems for data collection, processing, and storage have resulted in neurosurgical patients providing a glimpse into the inner workings of the human mind at unprecedented resolution.

Working in such a clinical setting takes a lot of care, and an understanding that the patient's needs always come first. These patients show a lot of kindness and grace in times of immense personal difficulty to undergo testing that allows us to uncover the neural mechanisms of complex phenomena that were previously out of reach.

In this Chapter we have briefly described the mechanics of recording and processing electrophysiological data from the human brain. This is meant to provide the reader with a general overview of how we start with continuously recorded voltage traces and end with claims about the activity of individual neurons in response to task conditions. This should place the results in the next Chapters in the appropriate context but is in no way meant to be comprehensive. For a more detailed description of each of the topics covered here please see Chapter 6 in reference[5].

References:

1. Logothetis, N.K. (2002). The neural basis of the blood-oxygen-level-dependent functional magnetic resonance imaging signal. Philos. Trans. R. Soc. Lond. B Biol. Sci. *357*, 1003–1037.

2. Buzsáki, G., Anastassiou, C.A., and Koch, C. (2012). The origin of extracellular fields and currents--EEG, ECoG, LFP and spikes. Nat. Rev. Neurosci. *13*, 407–420.

3. Engel, A.K., Moll, C.K.E., Fried, I., and Ojemann, G.A. (2005). Invasive recordings from the human brain: clinical insights and beyond. Nat. Rev. Neurosci. *6*, 35–47.

4. Fried, I., Wilson, C.L., Maidment, N.T., Engel, J., Jr, Behnke, E., Fields, T.A., MacDonald, K.A., Morrow, J.W., and Ackerson, L. (1999). Cerebral microdialysis combined with single-neuron and electroencephalographic recording in neurosurgical patients. Technical note. J. Neurosurg. *91*, 697–705.

5. Fried, I., Rutishauser, U., Cerf, M., and Kreiman, G. (2014). Single Neuron Studies of the Human Brain: Probing Cognition (MIT Press).

6. Misra, A., Burke, J.F., Ramayya, A.G., Jacobs, J., Sperling, M.R., Moxon, K.A., Kahana, M.J., Evans, J.J., and Sharan, A.D. (2014). Methods for implantation of micro-wire bundles and optimization of single/multi-unit recordings from human mesial temporal lobe. J. Neural Eng. *11*, 026013.

7. Brainard, D.H. (1997). The Psychophysics Toolbox. Spat. Vis. *10*, 433–436.

8. Buzsáki, G. (2004). Large-scale recording of neuronal ensembles. Nat. Neurosci. *7*, 446–451.

9. Mosher, C.P., Wei, Y., Kamiński, J., Nandi, A., Mamelak, A.N., Anastassiou, C.A., and Rutishauser, U. (2020). Cellular Classes in the Human Brain Revealed In Vivo by Heartbeat-Related Modulation of the Extracellular Action Potential Waveform. Cell Rep. *30*, 3536-3551.e6.

10. Nandi, A., Chartrand, T., Van Geit, W., Buchin, A., Yao, Z., Lee, S.Y., Wei, Y., Kalmbach, B., Lee, B., Lein, E., et al. (2022). Single-neuron models linking electrophysiology, morphology, and transcriptomics across cortical cell types. Cell Rep. *41*, 111659.

11. Gold, C., Henze, D.A., Koch, C., and Buzsáki, G. (2006). On the origin of the extracellular action potential waveform: A modeling study. J. Neurophysiol. *95*, 3113–3128.

12. Gold, C., Henze, D.A., and Koch, C. (2007). Using extracellular action potential recordings to constrain compartmental models. J. Comput. Neurosci. *23*, 39–58.

13. Chung, J.E., Magland, J.F., Barnett, A.H., Tolosa, V.M., Tooker, A.C., Lee, K.Y., Shah, K.G., Felix, S.H., Frank, L.M., and Greengard, L.F. (2017). A Fully Automated Approach to Spike Sorting. Neuron *95*, 1381-1394.e6.

14. Lewicki, M.S. (1998). A review of methods for spike sorting: the detection and classification of neural action potentials. Network *9*, R53-78.

15. Pouzat, C., Mazor, O., and Laurent, G. (2002). Using noise signature to optimize spike-sorting and to assess neuronal classification quality. J. Neurosci. Methods *122*, 43–57.

16. Quiroga, R.Q., Nadasdy, Z., and Ben-Shaul, Y. (2004). Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering. Neural Comput. *16*, 1661–1687.

17. Rutishauser, U., Schuman, E.M., and Mamelak, A.N. (2006). Online detection and sorting of extracellularly recorded action potentials in human medial temporal lobe recordings, in vivo. J. Neurosci. Methods *154*, 204–224.

18. Gibson, S.P. (2012). Neural Spike Sorting in Hardware: From Theory to Practice.

19. Radmanesh, M., Rezaei, A.A., Jalili, M., Hashemi, A., and Goudarzi, M.M. (2022). Online spike sorting via deep contractive autoencoder. Neural Netw. *155*, 39–49.

20. Lee, J.H., Carlson, D.E., Shokri Razaghi, H., Yao, W., Goetz, G.A., Hagen, E., Batty, E., Chichilnisky, E.J., Einevoll, G.T., and Paninski, L. (2017). YASS: yet another spike sorter. Adv. Neural Inf. Process. Syst. *30*.

21. Quian Quiroga, R. (2009). What is the real shape of extracellular spikes? J. Neurosci. Methods *177*, 194–198.

22. Minxha, J., Mamelak, A.N., and Rutishauser, U. (2018). Surgical and electrophysiological techniques for single-neuron recordings in human epilepsy patients. Extracellular recording.

23. Choi, J.H., Jung, H.K., and Kim, T. (2006). A new action potential detector using the MTEO and its effects on spike sorting systems at low signal-to-noise ratios. IEEE Trans. Biomed. Eng. *53*, 738–746.

24. Shalchyan, V., Jensen, W., and Farina, D. (2012). Spike detection and clustering with unsupervised wavelet optimization in extracellular neural recordings. IEEE Trans. Biomed. Eng. *59*, 2576–2585.

25. Nenadic, Z., and Burdick, J.W. (2005). Spike detection using the continuous wavelet transform. IEEE Trans. Biomed. Eng. *52*, 74–87.

26. Wiltschko, A.B., Gage, G.J., and Berke, J.D. (2008). Wavelet filtering before spike detection preserves waveform shape and enhances single-unit discrimination. J. Neurosci. Methods *173*, 34–40.

27. Meisenhelter, S., and Rutishauser, U. (2022). Probing the human brain at single-neuron resolution with high-density cortical recordings. Neuron *110*, 2353–2355.

28. Gabbiani, F., and Koch, C. (1998). Principles of spike train analysis. Methods in neuronal modeling *12*, 313–360.

29. Rutishauser, U., Schuman, E.M., and Mamelak, A.N. (2008). Activity of human hippocampal and amygdala neurons during retrieval of declarative memories. Proc. Natl. Acad. Sci. U. S. A. *105*, 329–334.

Chapter III

VISION:
The feedforward code for visual objects in the human brain

Introduction:

This thesis presents work attempting to make inroads into the problem of how intelligent beings model the world around them. A crucial first step to modeling the world is knowing what objects are in it[1], a fact that is easy to appreciate as objects are what we interact with, remember, refer to, and dream about. The work described in this Chapter is concerned with this first stage of model building — core object recognition. Consistent with the computational complexity of this problem and its central role in experience, a large swath of the primate temporal lobe — inferotemporal (IT) cortex — is dedicated to this problem[2], with lesions to this area causing specific visual agnosias including prosopagnosia.

A long-standing challenge in visual neuroscience has been to understand how IT cortex encodes the identity of a complex visual object, a problem that until recently did not have any clear solution, with some even positing that an explicit model of IT neurons may be impossible to achieve[3]. However recent work in macaques has demonstrated an explicit code for faces[4] and subsequently general objects[5] in IT. In this coding scheme neurons act as linear projectors, projecting incoming stimuli parametrized as points in a continuous high dimensional feature space onto specific preferred axes. These axes are defined by the weightings of a small set of independent parameters spanning the object space.

However, while object recognition in macaque IT has been the subject of such detailed scrutiny, next to nothing is known about how human IT neurons encode visual objects. Two earlier explorations report face and other category selective neurons in the fusiform face area (FFA)[6,7] with a third finding that the change in features caused by changing facial expression modulates face cell responses[8]. While these findings highlight that the answer to how we perform object recognition lies in the responses of IT neurons just as it does in monkeys, they shed very little light on the precise code for objects.

In this Chapter we present findings related to the neural code for visual objects in human IT cortex. Using the techniques described in Chapter II we record the responses of several hundred neurons in human IT cortex to a large, diverse set of visual objects. We find that the code for object encoding is the same as in macaques, and that the activity of deep layers of convolutional neural networks trained to perform object classification explain a significant and roughly similar amount of variance in IT responses.

Results:

Neurons in IT cortex show diverse visual response profiles

We began our exploration into how human IT neurons encode visual objects by recording responses of those neurons to a large number of objects with varying features using a rapid screening task. Patients sequentially viewed a series of 500 images (4 repetitions each for a total of 2000 total trials), occasionally answering a catch question that appeared at random intervals and pertained to the image that came just before the question. Images stayed on screen for 250ms and the inter trial interval was jittered between 100-150ms (Figure 3.1A). Despite the rapid presentation rate, patients generally got most of the catch questions correct indicating that the stimuli were carefully attended to (Figure 3.1E).

We recorded 431 IT neurons in 42 sessions across 12 patients. Visual responsivity was assessed on a trial-by-trial basis using a Poisson burst metric[9] that detects the precise time during the trial at which the spike rate departs significantly from baseline. Out of the 431 neurons recorded 384 were found to have significantly increased responses upon onset of visual stimulus. Human IT neurons had an average response latency of 132 +/- 1.18ms and showed a number of different response profiles, including canonically described category neurons (Figure 3.1F, left), anti-category neurons (Figure 3.1C), neurons characterized by an initial suppression of activity (Figure 3.1F, middle), and neurons that distinguish categories via a latency code (Figure 3.1F, right).

A computational scheme to parametrize visual objects

Despite establishing that human IT neurons (somewhat predictably) show robust and reliable visual responses, the question of how they encode specific features of objects remains open. In order to solve this problem, it is necessary to precisely map between a *quantitative description* of the features of a given object and the precise firing pattern it elicits in IT. In other words, matching the responses of neurons to object features is impossible without a computational scheme to parametrize the features. Thankfully, deep networks trained to do object classification do exactly this, performing classification by keeping track of the shape and appearance of the inputted object[3]. Thus, one can take a pre-trained deep network used for object classification such as AlexNet[10], pass in a stimulus image and use the unit activations of one of the fully-connected layers (eg. fc6) as a quantitative description of the image features.

Armed with this powerful solution we could describe arbitrarily shaped, complex visual objects as points in a high dimensional feature space. Now the question to pursue is – how does the neural activity map onto such a space?

At this point it is important to note that this is not a new problem in systems neuroscience. Since the very first discovery of neurons that respond more strongly to whole, complex objects like faces than to simple edges in the primate brain[11], attempts

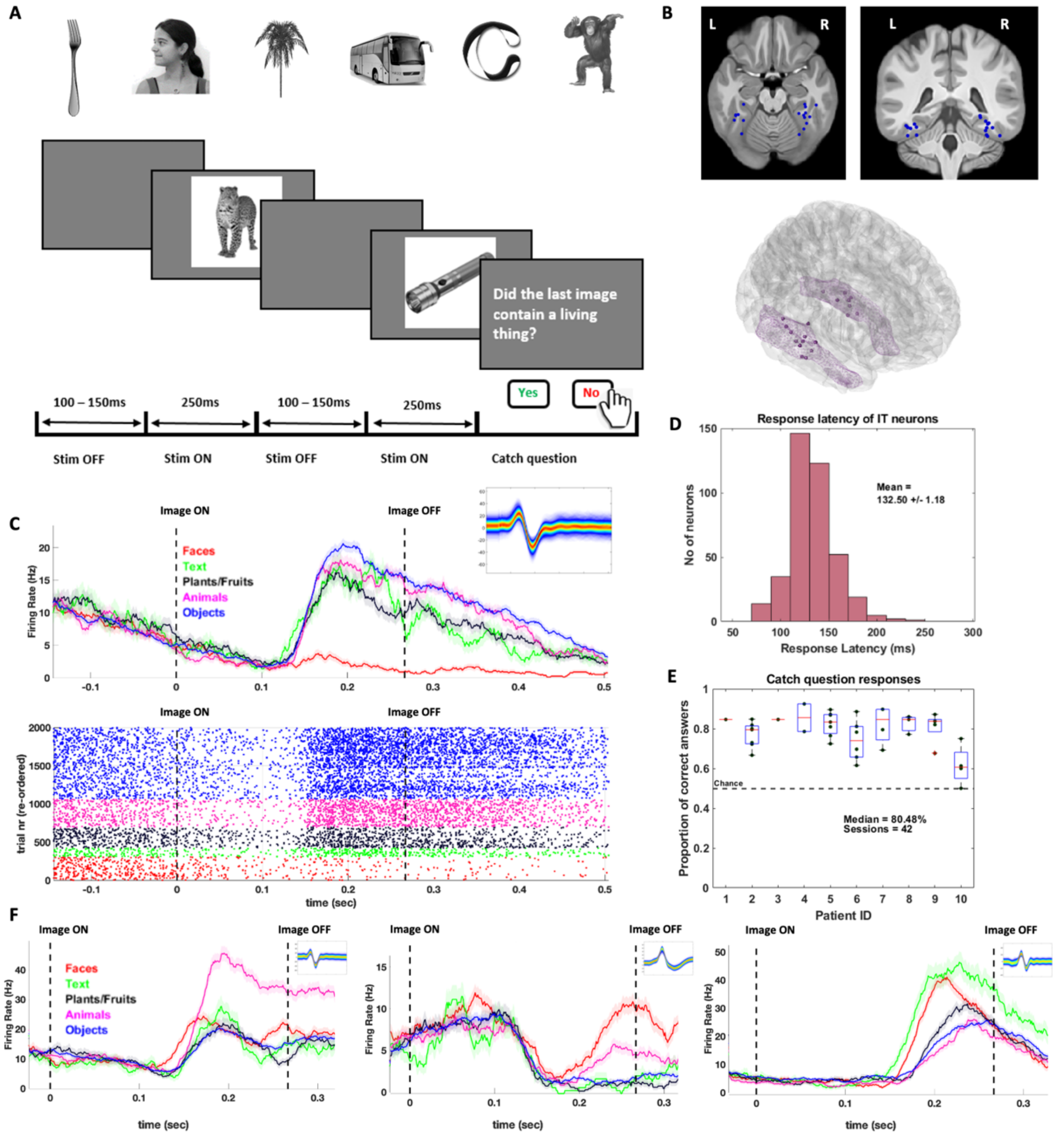


Figure 3.1. Screening task elicits robust visual responses in human IT.

(A) (Top row) Example stimuli used in screening task. All images were grayscale and had no background. Images were displayed on a stimulus computer 1 meter away and subtended 6-7 visual degrees. (Bottom) Schematic of screening task utilized. Background subtracted images were displayed on a gray screen for 250ms with the inter-trial interval jittered between 100-150ms at random intervals (min interval: 1 trial, max interval: 80 trials) a yes-no catch question would appear pertaining to the image that came just before it.

(B) Recording locations of the 18 microwire bundles (left and right) that contained at least one well isolated neuron in human IT cortex across all 12 patients. See Montreal Neurological Institute coordinates in Table 3.1. Each dot represents the location of one microwire bundle (8 channels).

(C) An example neuron recorded during screening. This neuron was responsive to all categories except faces. The computed response latency is 161ms.

(D) Distribution of response latencies for all 384 visually responsive IT neurons. Human IT neurons have a mean response latency of 132.5ms.

(E) Summary of catch question responses for all sessions, grouped by patient. On average patients answered catch questions correctly despite the rapid stimulus presentation, implying the stimuli were closely attended to.

(F) Neurons in human IT cortex show diverse response profiles. (Left) A strong category selective neuron, showing a lower response to any of the non-preferred categories. (Middle) A response profile characterized by an initial suppression of activity. (Right) A neuron that distinguishes it's preferred from non-preferred category using a latency code along with a rate code.

have been made to systematically analyze their activity[12,13] in order to uncover the relationship between complex features and neural firing rates. However, it took the combination of advances in computation and the discovery of anatomically defined, functionally connected face patches[14,15] for the answer to eventually be uncovered in the primate brain a few decades later[4].

As the field grappled with this question multiple theories for how neural activity mapped onto high dimensional feature spaces were put forward in the intervening years. One of them being an 'exemplar based' model, which posits that object recognition is subserved by units tuned to specific exemplars, i.e. points in object space, and that responses of such units to objects further and further away from the exemplar would decrease monotonically as a function of that distance in feature space[16,17]. The other being an axis model, wherein units are tuned not to specific points but specific *directions* in feature space such that the response of a unit to an incoming stimulus is a function of how far along the units preferred direction that stimulus is[17]. As time went by evidence for the axis model mounted, with the exemplar model being unable to explain certain psychological effects like the caricature effect[18] and the finding that primate IT neurons showed ramp shaped tuning to feature values[19].

Eventually the axis model was verified as the computational scheme by which neurons in IT cortex of macaques were responding to incoming faces[4] by showing that the neuron's firing rate was a function of projection value of the incoming stimulus onto its preferred axis, examining the tuning along an orthogonal axis and demonstrating that it was flat, and finally generating stimuli along a neurons preferred and orthogonal axes and demonstrating a methodical increase in firing rate along the preferred axis and no increase along the orthogonal axes[4,5]. Though what about human IT neurons?

Human IT neurons are tuned to specific axes in object space

We find that human IT neurons use the exact same coding scheme as macaque IT neurons, showing a monotonically increasing firing rate to stimuli that were further along their preferred axes while having the same firing rate to images along an axis orthogonal to their preferred axis. An example neuron showing axis tuning can be seen in Figure 3.2C-E. Individual neurons are considered axis tuned if in addition to expected tuning along both axes the correlation between projection value onto the preferred axis and firing rate is significant as compared to a shuffle distribution. We find that a majority (~57%) of visually responsive neurons are significantly axis tuned (Figure 3.2B&H).

Axis tuning also made clear the response pattern of some previously confusing neurons, an example of which is shown in Figure 3.2C. This neuron seemed to have a strong category response, however it also appeared to respond strongly to specific stimuli in other categories (see rows highlighted in yellow, Figure 3.2C). Indeed, when the neurons most and least preferred stimuli are examined (Figure 3.2D) it becomes apparent that there is no semantic label capable of cleanly delineating between the two. However, when examining the projection value of those stimuli the pattern becomes clear: there

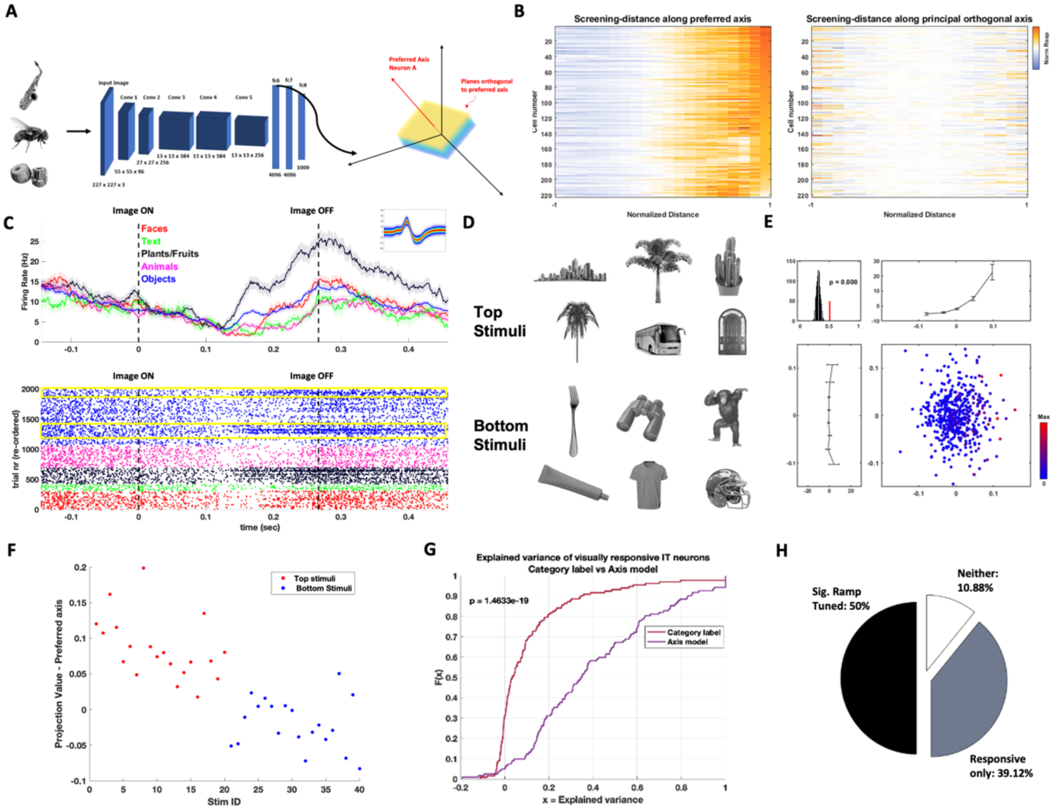


Figure 3.2. Axis tuning in human IT neurons.

(A) Schematic of the stimulus parametrization and axis computation procedure. Individual stimuli are parametrized as points in a high dimensional feature space that is built from the unit activations of AlexNet's fc6 layer. Neurons' preferred axes are computed in this space by projecting the features onto the mean subtracted responses leading to a large null space for each neuron.

(B) Population summary showing the ramp tuning of 218/384 visually responsive units. (Left) Plot showing the increase in normalized response of each neuron as distance along the preferred axis increases. (Right) Corresponding plot for the principal orthogonal axis (orthogonal axis capturing the most variation) showing no change at all in response as a function of distance.

(C) The axis model helps explain previously confusing neuronal tuning. Category tuning has long been used to study visually responsive neurons in the human brain. However, we recorded neurons that had robust category responses at first glance but more complicated responses when examined further. (C1) An example neuron with confusing category tuning. A closer look at its raster betrays robust responses to a few out of category objects (highlighted by yellow box).

(D) The top (most preferred) and bottom (least preferred) stimuli for the neuron shown in (C), no semantic category can cleanly delineate between them.

(E) Axis plot for the same neuron. The 2D scatter plot is a plot of all stimuli projected onto the preferred and orthogonal axes and colored by the neuron's response. The top line plot is the binned firing rate for all stimuli as one moves along the preferred axis. The side (vertical) line plot is the same binned firing rate for the orthogonal axis. The top left inset shows the correlation of projection value along the preferred axis and firing rate (red line) as compared to a shuffled distribution.

(F) Examining the projection values of the previously mentioned top and bottom stimuli (shown in D) reveals a systematic relationship between the two.

(G) The axis model explains much more variance than the category label in human IT neurons.

(H) A chart showing the distribution of axis tuned and visually responsive neurons out of all neurons recorded.

is a linear relationship between the firing rate and the projection value. At a population level, this plays out via the axis model explaining much more variance than the category label (Figure 3.2G).

If neurons in IT are indeed axis tuned that implies their responses can be approximated as a linear combination of object features, with the slopes of the ramps corresponding to the weights[4]. Then for a population of neurons $\vec{R} = C \cdot \vec{F} + \overrightarrow{C_o}$ where $\vec{R}$ is the response vector of the different neurons, C is the weight matrix, F is the vector of object feature values, and C_o is the offset vector. If this is the case, then by simply inverting this equation one should be able to linearly decode the object features from the population activity[20–22].

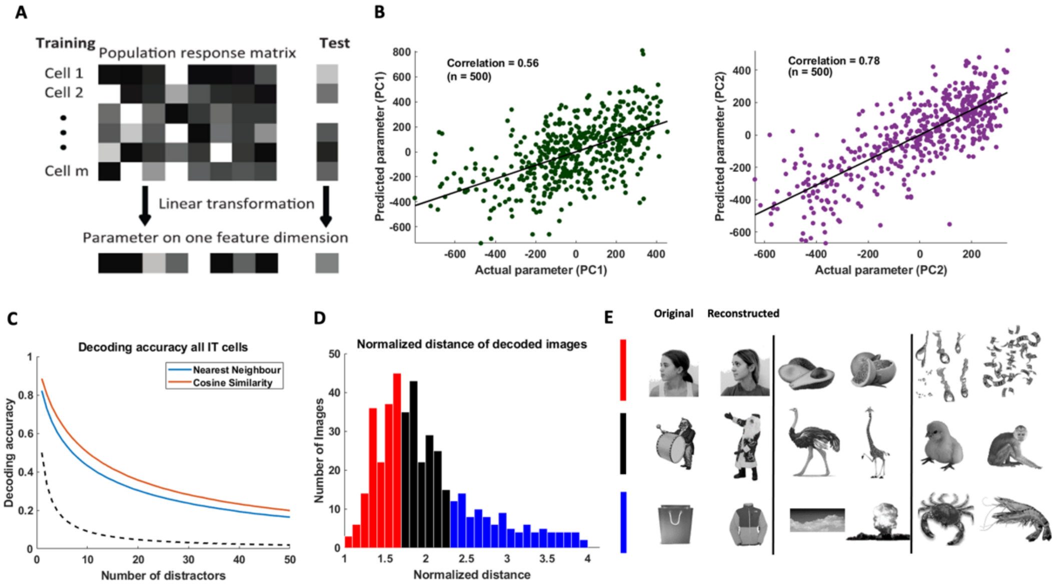


Figure 3.3. Decoding object features via linear regression.

(A) Diagram illustrating decoding model. We used responses to all but one object (500-1 = 499) to determine the transformation between responses and feature values by linear regression, before using that transformation to predict the feature values of the held out object. Adapted from Chang & Tsao, 2017.

(B) Model predictions using human IT data are plotted against the actual feature values for the first (left) and second (right) dimensions of object space.

(C) Decoding accuracy as a function of the number of distractor objects drawn randomly from the stimulus set using two different distance metrics (Euclidean distance & Cosine similarity). For each model, different sets of features were first linearly decoded from population responses, then the distances between decoded and actual features in each feature space were computed to determine decoding accuracy. The black dashed line represents the decoding accuracy one would expect by chance.

(D) Distribution of normalized distances between reconstructed feature vectors and the best-possible reconstructed feature vectors for 437/500 images (see methods) to quantify decoding accuracy across the population. The normalized distance takes into account the fact that the object images used for reconstruction did not include any of the object images shown to the patients. A normalized distance of 1 means the reconstruction found the best solution possible.

(E) Images were split into tertiles based on normalized distance. Examples of the reconstructions in the first tertile (top row), second (middle row), and third (bottom row) as compared to the original stimulus images being reconstructed.

We used this fact to decode object identity using the responses of the neurons. We used leave-one-out cross-validation to learn the linear transform that maps responses to features (Figure 3.3A). The representation turned out to be quite rich, containing a lot of information about object identity allowing objects to be identified among many distractors (Figure 3.3C). As a proxy for 'reconstructing' the object we searched a large auxiliary object database for the object with the feature vector closest to that decoded from the neural activity. A normalized distance in feature space between the best possible and actual reconstruction (see methods) was computed to quantify the decoding accuracy across all stimuli. For 437/500 images the normalized distance was very small (< 4), implying that the neural responses captured many of the fine feature details of the original objects, which was verified via side-by-side comparisons of the original images with the 'reconstructions' (Figure 3.3D).

Finally, to hand the axis model its sternest test we ran a 'closed-loop' synthetic image screen (see methods). We used the morning session to perform our object screen and compute axes for all neurons recorded, then used a pre-trained GAN specifically designed to invert the responses of AlexNet[23,24] to systematically generate images corresponding to evenly spaced points along both the preferred and orthogonal axes. We then returned to the patient room for an afternoon session and rescreened with the synthetic images added to the original stimulus set. Neurons recorded in the morning and afternoon are matched offline using an algorithm that checks for similarity in selectivity, waveform shape, burst index, response latency, and significance in axis tuning (see methods).

In such an experiment the axis model makes two testable predictions: the first is that one should see a monotonic increase in firing rate for images sampled along the preferred axis (and no change along an orthogonal axis), and the second is that if one generates images that go *beyond* the maximum projection value of the original stimulus images these should serve as 'super-stimuli' driving the neuron to a higher firing rate than any of the stimulus images[24]. In fact, if one samples a grid of images in the space spanned by the preferred axis and the principal orthogonal axis one should see changes in tuning only in objects varied along the preferred axis. Both predictions were verified in our data. For a given neuron, the most preferred stimulus with the highest projection value and highest response was a t-shirt (Figure 3. 4A), while the least preferred stimulus was a helicopter. The corresponding super-stimulus and anti-stimulus are shown and, as expected, bear striking resemblance to the most and least preferred stimulus images. The axis plot with synthetic stimuli overlaid (preferred axis – yellow, orthogonal axis – green) demonstrates the expected monotonic increase along the preferred axis and no increase along the orthogonal axis (Figure 3.4B). Moreover, this plot clearly shows that the firing rate to the synthetic stimuli with larger projection values than the stimulus images are substantially higher only along the preferred axis. Figure 3.4D shows an axis plot for the same neuron, only plotting the grid of synthetic images sampled in the space spanned by the preferred and orthogonal axes (Figure 3.4C) showing tuning only along the preferred axis, and nearly identical responses to images varied along the orthogonal axis.

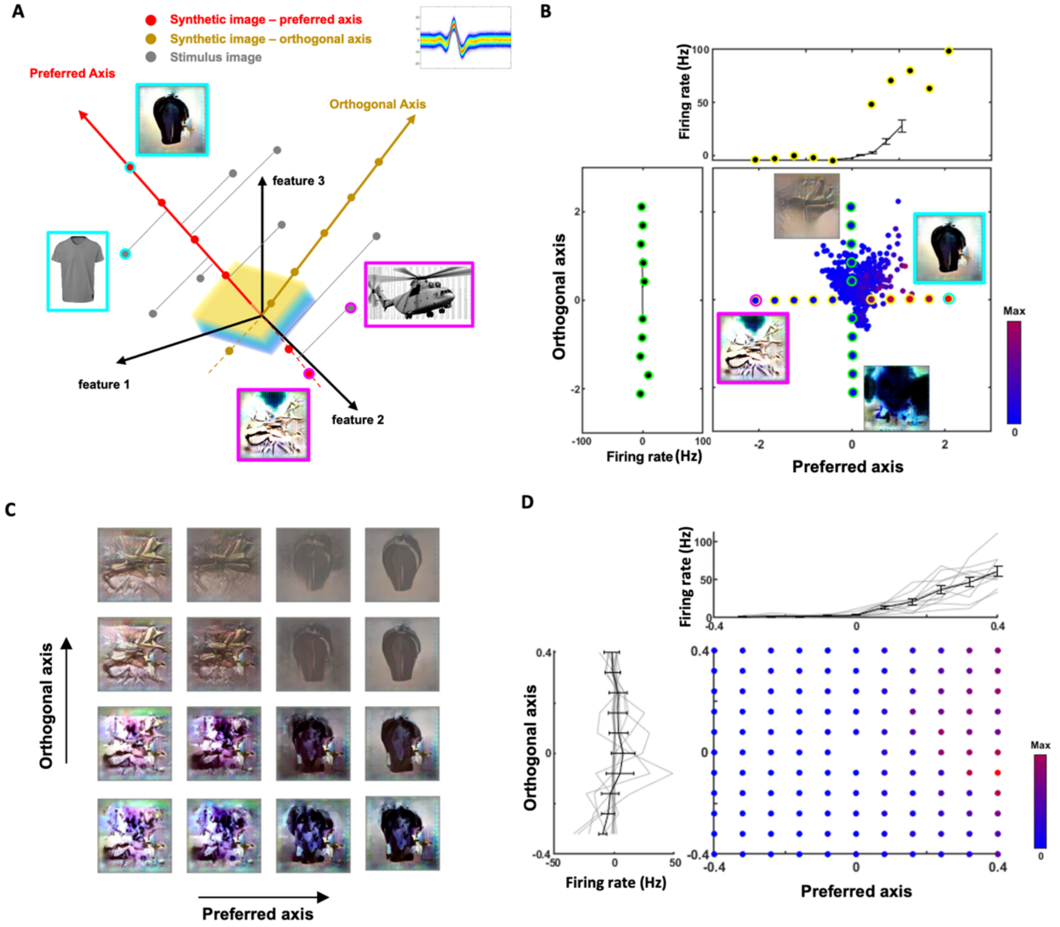


Figure 3.4. Generating super stimuli for an IT neuron using the axis model and a GAN.

(A) Schematic illustrating the generation of super stimuli. For a given neuron, images were parametrized via AlexNet's fc6 layer, then points are evenly sampled along the preferred and orthogonal axes ensuring that the max projection value of the chosen points exceeds the max projection value along both axes of the stimulus images. These feature vectors are subsequently fed into a GAN trained to invert AlexNet's responses to generate the synthetic images.

(B) Axis plot of an example neuron showing the positions of generated stimuli sampled along the preferred axis (yellow) and orthogonal axis (green) relative to the other stimulus images. The maximum and minimum projection valued images are displayed along each axis. The vertical and horizontal line plots are the binned firing rate of the stimulus images as one moves along each axis with the generated images overlaid on top. The systematic increase along the preferred axis with an almost identical response to all images along the orthogonal axis is clearly visible.

(C) To give the axis model its sternest test we sampled a grid of images at even intervals in the space spanned by the preferred and orthogonal axes.

(D) Axis plot of the neuron showing tuning to only changes in images along the preferred axis.

Axis code is independent of the specific convolutional network used to parametrize stimuli

A key ingredient to these results about neurons being linear projectors in feature space is the space itself. If one does PCA directly one the pixels, something that has been attempted for faces many times we get a series of 'eigen faces'[25,26] that bear little resemblance to the actual face. Hence if the construction of the high dimensional feature space is the crucial step it begs the question: is there something special about the latent space of AlexNet? Are the axis tuning results an artifact of the specific network used to parametrize the objects? Nowadays there exist a plethora of deep convolutional neural network models that are capable of performing object recognition at or beyond human capability[27].

We set out to compare several such models for object recognition. Particularly, their ability to explain the response of neurons in IT to general objects. The models tested include AlexNet[10], both VGG-16 and 19[28], VGG-Face[29], the eigen object model, and 4 CORNet models[30,31]. These particular models were chosen as they have very good object recognition capabilities, and all parametrize objects in different ways. The eigen object model is a parallel to the eigen face model referenced earlier consisting of principal components analysis done the pixel-level representation of the images. AlexNet, VGG-16/19, and the CORNet family of networks are trained to classify images into 1000 object categories with varying architectures. AlexNet has an 8 layer structure with 5 convolutional and 3 fully connected layers. VGG-16 and 19 have 16 and 19 layers respectively with 3 fully connected layers and the rest being convolutional layers. VGG models are also known for leveraging the smallest possible receptive field in their convolutional layers (3x3). The CORNet family of networks consists of 4 networks: CORNet-Z being purely feedforward; CORNet-R including some recurrence which has been shown to be essential for object recognition in the primate visual system[32]; CORNet-RT has the same structure as R but includes 'biological unrolling' wherein the input at time t+1 in layer n is the same as input to layer n-1 at time t so that information flows through the layers sequentially[30]; and finally CORNet-S, which is the most complicated of architectures including recurrent and skip connections between the layers[31]. Despite the individual differences all 4 networks have architectures inspired by the primate visual system with layers corresponding to V1, V2, V4, and IT. VGG-Face has the same architecture as VGG-16 but is trained to identify 2622 celebrities.

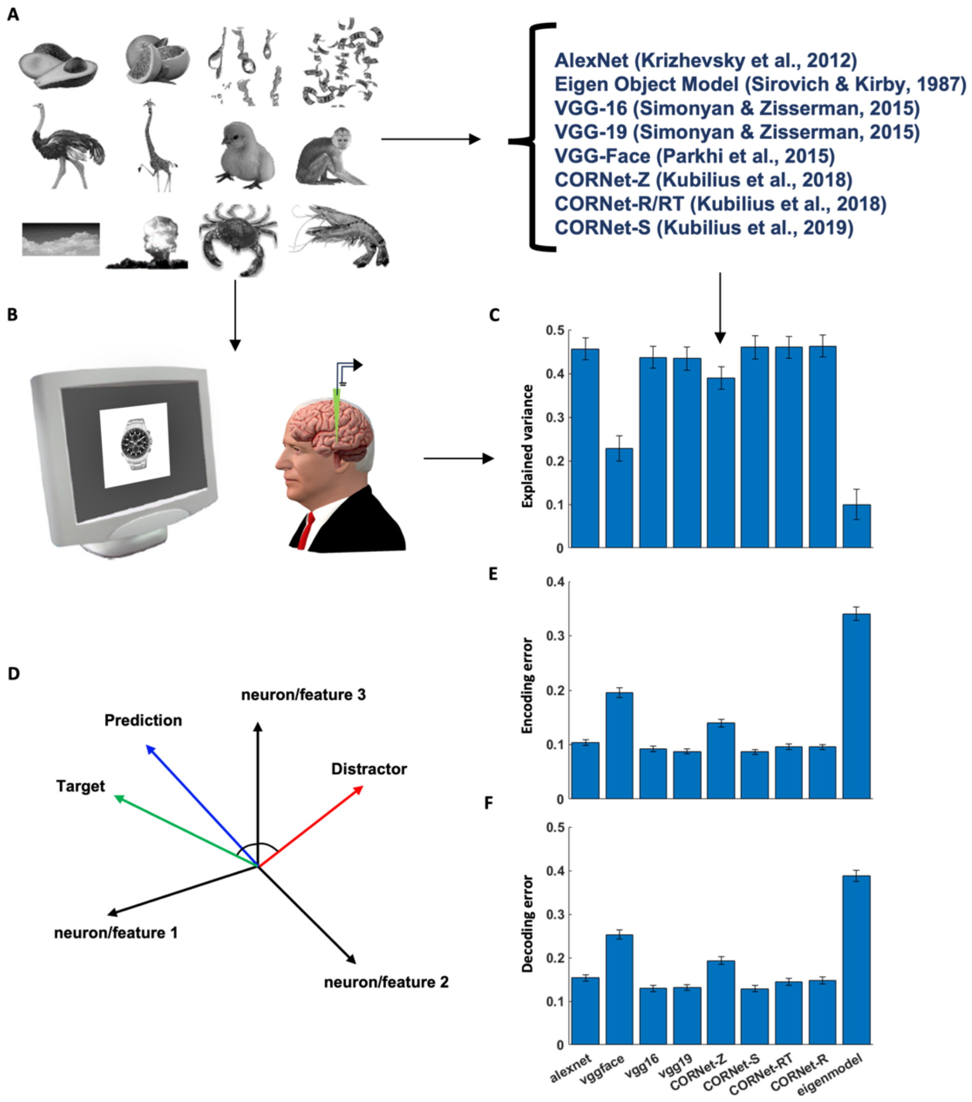


Figure 3.5. Comparing how well various models explained human IT responses to object images.

(A) The 500 grayscale stimulus images used in the screening task (taken from www.freepngs.com) were parametrized using 8 different models from 4 different model families (AlexNet, the eigen object model, VGG, and CORNet). The same number of features were extracted from units of the different models using principal components analysis (PCA for comparison).

(B) Responses from IT neurons were recorded as patients viewed these objects 4 times each. For recording locations and task schematic see Figure 3. 1.

(C) The explained variances for each model after 50 features were extracted using PCA. For each neuron, explained variance was normalized by the explainable variance (see methods). Error bars represent SEM for the recorded neurons. The eigen model, VGG-Face, and CORNet-Z performed worse than the other models with no significant differences between the rest.

(D) The various models were compared with respect to how well they could predict the neuronal responses or the object features. In both cases a leave-one-out procedure was used to learn and test the transformation between responses and features. (D1) To quantify encoding error for example, for each object we compared predicted responses to individual objects in the neural state space to the actual responses to that object and a distractor object. If the angle between the predicted response and the actual response was smaller than the angle between the predicted response and the distractor the encoding was considered correct. To quantify decoding error, we reversed the roles of the neural responses and the object features and decoded object features before comparing the decoded features to the actual features for a given object and a distractor object.

(E) Encoding error across all models.

(F) Decoding error across all models. The eigen model, VGG-Face, and the purely feedforward CORNet-Z had larger encoding/decoding errors than the other models, consistent with them explaining less variance as well.

To quantify the ability of each network to explain human IT responses we learnt a linear mapping between the features of each model and the neural responses[33]. As we did in our earlier axis tuning computations and to avoid overfitting we reduced dimensionality of the feature representations via principal-component analysis (PCA) yielding M features for each object and model. In our main analyses M=50, we used leave-one-out cross validation and for each neuron fit the responses of N-1 images to the M features via linear regression before predicting the response of the neuron to the Nth image using the same linear transform. The explained variance by the linear transform was used as an initial measure of goodness-of-fit. Beyond this we also computed the encoding and decoding error for each neuron with every model. Encoding error is computed as follows: the observed population vector to each object was compared to the predicted population response vector and the observed population response vector to a random other object in the set. If the angle between the observed and predicted response to the chosen object was smaller than the angle between the predicted response and the distractor, the prediction was considered correct (Figure 3.5D). Decoding error is computed via the same method except the feature vector of the object is predicted from the neural responses. In other words, the role of the neural responses and the model features is reversed (see methods). A model is considered to explain neural responses well if it has high explained variance and low encoding/decoding error. We found that with the exception of VGG-Face and the eigen object model that performed significantly worse, there was no significant difference in explained variance between the models (Figure 3.5C; AlexNet vs VGG-Face, p = 2.2825e-07; AlexNet vs eigen model, p = 3.8617e-21). The most complicated CORNet, CORNet-S got the absolute highest values and the purely feedforward CORNet-Z got the lowest though the differences are not significant (p = 0.2182).

Up to this point, for comparing the deep feedforward models (AlexNet and VGG) we have used the unit activations of fc6 before ReLu, allowing the unit activations to contain negative numbers. An opinion in the field is that it is better to use activations in the fully connected layers after ReLu as this seems more biologically plausible owing to the impossibility of a negative neural firing rate. A layer-by-layer comparison finds that for AlexNet and the VGG-object networks found no significant difference in model performance between either of the last fully connected layers before or after ReLu (Figure 3.S1), however a comparison between the 4 layers of the CORNet family shows a large difference between the layers with the ultimate layer corresponding to IT showing the highest performance by far (Figure 3.S1).

Discussion:

We present the first exploration of object recognition at single neuron resolution in human IT cortex. In this Chapter we have provided a detailed outline of the previous work done on core object recognition at single neuron resolution in macaques, discussed innovations in arbitrary object parametrization (via deep networks trained to do object classification), made clear that the former two points *together* were required to make this study possible, and finally that human IT neurons use exactly the same axis code as macaques to represent visual objects.

Since the discovery of the fusiform face area in humans[34] a hotly debated topic in the field has been whether faces, given their social importance are 'special'. In other words, is the machinery for processing faces in humans more specialized than that for general objects? This notion is supported by behavioral experiments of face recognition[35] (see box 1 in reference), however our results suggest that at the apperceptive level wherein features are stitched together to create a percept this is not the case. Further studies will be required to assess whether this holds at the associative level too.

That being said, the real value of this work lies in the future studies it makes possible. Mental imagery has been the subject of intense scrutiny for many decades and popularized by the so-called 'imagery debate' between psychologists Steven Kosslyn and Zenon Pylyshyn (amongst others) about whether mental imagery relied on depictive (pictorial; with a structure similar to sensory perception) or propositional representations[36]. Our current results on the feedforward representation of visual objects provide a vessel one can use to charter unknown territory: the single neuron mechanisms of visual imagery.

Methods:

Participants

Patients with drug-resistant epilepsy

12 patients with drug-resistant epilepsy volunteered for this study and gave their informed consent. The institutional review board of Cedars-Sinai Medical Center approved all protocols. The task was conducted while patients stayed in the hospital after implantation of depth electrodes for monitoring seizures. The location of the implanted electrodes was solely determined by clinical needs. The neural results were analyzed across all 12 patients.

Task

Screening

Patients viewed a set of 500 object stimuli with varying features (taken from www.freepngs.com) 4 times each for a total of 2000 trials in a shuffled order. Each image stayed on screen for 250ms and the inter-trial interval consisting of a blank screen was jittered between 100-150ms. The task was punctuated with yes/no 'catch' questions pertaining to the image that came right before the question requiring the patients to pay close attention in order to answer them correctly. Catch questions occurred between 2 to 80 trials after the previous one. Patients responded to the questions using a RB-844 response pad (https://cedrus.com/rb_series/). During the synthetic image screens, the synthetic images would be added to the original stimulus set and the task parameters remained unchanged.

Electrophysiology

We recorded broadband signals from inferotemporal cortex in addition to the standard epileptic targets (amygdala, hippocampus, anterior cingulate cortex, pre-supplementary motor area, and orbitofrontal cortex) using micro-macro electrodes manufactured by Ad-Tech (https://adtechmedical.com/epilepsy#depth-electrodes). All analyses pertain to signals recorded on the 8 microwires protruding from the end of the electrode and sampled at 32khz.

Spike sorting and quality metrics

Signals were bandpass filtered offline in the range of 300-3000hz with a zero phase lag filter before spike detection. Spike detection was carried out by the semiautomated template matching algorithm Osort[40]. The spike quality metrics returned by the algorithm were examined before final clusters were curated.

Electrode localization

Electrode localization was based on postoperative MRI/computed tomography (CT) scans. We co-registered postoperative and preoperative MRIs using Freesurfer's mri_robust_register[41]. To summarize recording locations across participants we aligned each participant's preoperative scan to the CITI168 template brain in MNI152 coordinates[42] using a concatenation of an affine transformation and symmetric image normalization (SyN) diffeomorphic transform[43]. The MNI coordinates of the microwires from a given electrode shank were marked as one location. MNI coordinates of microwires with putative neurons detected from all participants were plotted on a template brain for visualization (Figure 3.1B).

Data Analyses

Visual responsiveness classification and response latency computation

We computed such a single trial latency by using a Poisson spike-train analysis. This method detects points of time in which the observed inter-spike intervals (ISI) deviate significantly from that assumed by a constant-rate Poisson process. This is done by maximizing a Poisson surprise index[9]. The firing rate of the neuron during the inter-trial interval was used to set the baseline rate for the Poisson process. Spikes from a window of 50-350ms after stimulus onset were included. The statistical threshold for detecting onset was $p < 0.001$. The response latency of the neuron was taken to be the average latency across all responsive trials. A neuron was considered visually responsive if it satisfied the following criteria: either one of the groups had a response that was 3.5 standard deviations above baseline (category neurons) or the response to a single stimulus was 10 standard deviations above baseline (sparse, concept neuron-like coding) and the average firing rate during the stimulus presentation period was above 0.5hz.

Axis computation

Preferred axis

The preferred axis of each neuron was computed using the spike triggered average (STA). The neural response vector was computed by binning spikes elicited by each stimulus in a 250ms window starting from the response latency of the neuron — necessarily restricting analysis to visually responsive neurons.

Once the neural response vector was computed the STA is defined as:

$$P_{sta} = (\vec{r} - \bar{r})F$$

where $\vec{r}$ is the 1 x n neural response vector to n objects, $\bar{r}$ is the mean firing rate, and F is an n x d matrix of features with each row corresponding to the features for a given object.

Principal orthogonal axis

The orthogonal axis seen in all plots is the principal orthogonal axis. This is defined as the axis orthogonal to the preferred axis along which there is the most variation. For each neuron, the preferred axis was computed. The component along the preferred axis (P) was subsequently subtracted from all object feature vectors in F leaving a matrix of orthogonal feature vectors. Succinctly, for a given feature vector $\overrightarrow{f_d}$ in feature space we computed

$$\overrightarrow{f_{d-1}} = \overrightarrow{f_d} - (\overrightarrow{f_d} \cdot \frac{P}{|P|^2})P$$

Then principal component analysis was performed on this set of n vectors f_{d-1}, and the first principal component is chosen as the principal orthogonal axis.

Quantifying significance of axis tuning

For each neuron after the preferred axis was computed we examined the correlation between the firing rate response to the stimuli and their projection value along the preferred axis. This correlation value was recomputed after shuffling the features (1000 repetitions) and the original value was compared to this bootstrap distribution. If the original value was greater than 99% of the shuffled values the neuron was considered axis tuned.

Decoding analysis

We find that neurons in human IT cortex are performing linear projection onto specific preferred axis in object space. As such, their responses can be well modeled by the equation

$$\vec{R} = C \cdot \vec{F} + \overrightarrow{C_o}$$

where $\vec{R}$ is the population response vector to a given image, C is the weight matrix for different neurons, $\vec{F}$ is the vector of object feature values, and $\overrightarrow{C_o}$ is the offset vector. Thus the decoding analysis was carried out by inverting this equation giving us

$$\vec{F} = \vec{R} \cdot C' + \overrightarrow{C'_o}$$

We used the responses of all but one of the objects (500 – 1 = 499) to fit C' and $\overrightarrow{C'_o}$. These were then plugged into the equation to predict the feature vector of the last object. Decoding accuracy was quantified by randomly selecting a subset of object images that included the actual feature vector of the decoded object from the total set of 500 and compared their feature vectors to the predicted feature vector of the decoded object by Euclidean distance or cosine similarity. If the actual feature vector closest to the predicted feature vector is of the object being decoded ('target') the decoding is considered correct. This procedure is repeated 100 times for each of the 500 images with a varying number of distractors to get an aggregate measure of decoding accuracy (Figure 3.3C).

Object 'reconstruction'

To generate images that reflect the features encoded in the neural responses we gathered images from an auxiliary database and passed 17,856 background free images through AlexNet. The images were then projected into the space built by the 500 stimulus objects. None of these ~18k images had been shown to the patients. For each stimulus image the feature vector decoded from the neural activity was compared to the feature vectors of the large stimulus set. The object in the large image set with the smallest Euclidean distance to the decoded feature vector was considered the 'reconstruction' of that stimulus image[5].

To account for the fact that the large object set did not contain any images shown to the patients, which sets a limit on how good the reconstruction can be, we computed a 'normalized distance' to quantify the reconstruction accuracy for each object. We defined the normalized reconstruction distance for an image as

$$Normalized\ distance\ = \frac{|V_{recon} - V_{original}|}{|V_{best\ possible\ recon} - V_{original}|}$$

where V_{recon} is the feature vector reconstructed from neuronal responses, $V_{original}$ is the feature vector of the image presented to the patients, and $V_{best\ possible\ recon}$ is the feature vector of the best possible reconstruction (image in the large set with the closest distance to $V_{original}$). A normalized distance of 1 means the decoded image is the best reconstruction possible.

Generation of synthetic stimuli

The axis model provides a very clear relationship between images and responses for individual neurons. In essence, images with increasing projection values onto a neuron's preferred axis will show increasing firing rates. This implies that if one computes a neuron's preferred axis and then evenly samples points along it and generates images from those points, those images will elicit systematically increasing responses from the neuron. This also implies that if one generates an image from a point further along the axis than any of the stimulus images used to compute the neurons axis, that image will act as a super stimulus and drive the neuron to a higher firing rate than any of the stimulus images.

To test these predictions we ran the screening task in one session, computed the axes for the neurons recorded, sampled points along the preferred and orthogonal axes and fed those vectors back into a pre-trained GAN[23] to generate the synthetic stimuli. We then went back to the patient room and re-ran the screening task with the synthetic images added in. The neurons from the morning and afternoon were matched (see below) and the responses of the neurons to the synthetic stimuli were recorded.

Matching neurons from morning and afternoon sessions

Given the nature of recording neurons in a clinical setting wherein you can only record a few neurons at a time it is common practice to run a screening task in the morning in order to determine the stimuli that drive those particular neurons before using those stimuli in other tasks related to decision making, memory, etc. In such cases it is always assumed that you are recording from the same neuron a few hours later but how can you be sure?

One way to convince yourself of this is to re-run the same screen in the afternoon and assess the selectivity of the neuron. However, if you have multiple neurons having roughly similar selectivity this method is inadequate by itself. In order to meet this challenge we examined multiple features of the neurons recorded in the morning and afternoon. Our algorithm would compute the selectivity vector (rank ordered list of stimulus number for the neuron), the waveform, the burst index which is a measure of how many bursts per unit time the neuron discharges[40], the computed response latency of the neuron, and whether or not the neuron was axis tuned (binary variable). The selectivity vectors were compared using cosine distance (matlab's 'pdist' function), and the waveforms by Euclidean distance.

Each neuron in the morning session was compared to all neurons in the afternoon session that were in the same region (left or right IT) and had the same binary value of 'axis tuned or not'. The afternoon neurons were then rank ordered in every category with first place in a given category giving an afternoon neuron a score of p where p = number of afternoon neurons being compared to the one morning neuron, second place giving a score of p-1 all the way until the last place neuron in a given category receiving a score of 1. The scores of all afternoon neurons were then summed and the algorithm would assign the afternoon neuron with the max score to be the morning neurons 'match'. This procedure is then repeated in the reverse direction, i.e. each afternoon neuron is compared to all morning neurons. Pairs that were bijective were automatically returned as 'matches' and all others were marked out for manual curation. Manual curation was carried out by examining the shape of the peri-stimulus time histogram of the category response of both neurons.

Model comparisons

Extraction of features from stimulus images

Each stimulus image was fed into one of the following models to extract the corresponding features:

Eigen object model:

PCA was performed on the original images of the 500 stimulus objects and the top 50 PCs were extracted to compare with other models.

AlexNet:

We used a pre-trained MATLAB implementation of AlexNet (download the Deep Learning Toolbox and use 'net = alexnet'). This is an 8 layer deep convolutional neural network with 5 convolutional layers and 3 fully connected layers, trained to classify images into 1000 object categories.

VGG Family:

We used pre-trained MATLAB implementations of VGG-16, a 16 layer deep convolutional neural network that contains 16 layers with 13 convolutional layers and 3 fully connected layers trained to classify images into 1000 object categories[28], VGG-Face which has the same structure as VGG-16 but is trained to recognize the faces of 2622 celebrities[29], and VGG-19 that has 19 layers (16 convolutional and 3 fully connected) trained on the same task as VGG-16[28].

CORNet Family:

We used a pre-trained PyTorch implementation of CORNet. The CORNet family contains 3 architectures: CORNet-Z, CORNet-R, and CORNet-S. Each architecture includes 4 main layers that correspond to V1, V2, V4, and IT. CORnet-Z is the simplest model and is purely feedforward. CORNet-R takes the otherwise feedforward network and introduces recurrent dynamics within each area. CORNet-S is the most complex containing within area recurrent connections, skip connections and the most convolutional layers. Our plots include a CORNet-RT plot which refers to a version of CORNet-R that does biological temporal unrolling[30] (see fig 2 in reference).

The parameters of AlexNet and VGG can be gotten form MATLAB's Deep Learning Toolbox. The CORNets were downloaded from (https://github.com/dicarlolab/CORnet).

Explained variance computation

To quantify the explained variance for each neuron, a leave-one-out cross validation approach was used. The responses to 499 objects were used to fit a linear regression model using the PCs of the features, and the response of this neuron to the left out object were predicted using the same linear transform. In this manner we could produce a predicted response for all images. Note that the computation of variance explained by the category label was done in this manner as well, replacing the PC features with the vector of category labels.

To set an upper bound for the explained variance different trials of responses to the stimuli were randomly split into two halves. The Pearson correlation (r) between the average responses from two half-splits across images was calculated and corrected using the Spearman-Brown correction:

$$r' = \frac{2r}{(1+r)}$$

The square of r' was considered the upper bound or explainable variance.

The reported results are the ratio of explained to explainable variance in the 183/218 axis tuned neurons that had a > 10% explainable variance.

Encoding/Decoding error computation

For the encoding analysis, the response of each neuron was first zscored, then the same procedure as the explained variance computation was followed to obtain a predicted response to every single object. To quantify prediction accuracy, we examined the angle between the predicted population response to each object and its actual population response (target) or the population response to a different object (distractor). If the angle between the predicted response vector and the distractor was smaller than the angle between the predicted response vector and the target this was counted as an error. Overall encoding error was quantified as the average errors across 1000 pairs of target and distractor objects.

For the decoding analysis we used exactly the same procedure, however the roles of the neural responses and the object features were reversed. We first normalized each dimension of object features to have zero mean and unit variance, then for each image using a leave-one-out procedure to fit a linear transform using the responses to 499 images and then predicting the features of the left out image. Decoding error was computed as the average decoding error across all target and distractor pairs in feature space.

References:

1. Lake, B.M., Ullman, T.D., Tenenbaum, J.B., and Gershman, S.J. (2017). Building machines that learn and think like people. Behav. Brain Sci. *40*, e253.

2. Tanaka, K. (1996). Inferotemporal cortex and object vision. Annu. Rev. Neurosci. *19*, 109–139.

3. Yamins, D.L.K., Hong, H., Cadieu, C.F., Solomon, E.A., Seibert, D., and DiCarlo, J.J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc. Natl. Acad. Sci. U. S. A. *111*, 8619–8624.

4. Chang, L., and Tsao, D.Y. (2017). The Code for Facial Identity in the Primate Brain. Cell *169*, 1013-1028.e14.

5. Bao, P., She, L., McGill, M., and Tsao, D.Y. (2020). A map of object space in primate inferotemporal cortex. Nature *583*, 103–108.

6. Axelrod, V., Rozier, C., Malkinson, T.S., Lehongre, K., Adam, C., Lambrecq, V., Navarro, V., and Naccache, L. (2019). Face-selective neurons in the vicinity of the human fusiform face area. Neurology *92*, 197–198.

7. Khuvis, S., Yeagle, E.M., Norman, Y., Grossman, S., Malach, R., and Mehta, A.D. (2021). Face-Selective Units in Human Ventral Temporal Cortex Reactivate during Free Recall. J. Neurosci. *41*, 3386–3399.

8. Axelrod, V., Rozier, C., Malkinson, T.S., Lehongre, K., Adam, C., Lambrecq, V., Navarro, V., and Naccache, L. (2022). Face-selective multi-unit activity in the proximity of the FFA modulated by facial expression stimuli. Neuropsychologia *170*, 108228.

9. Hanes, D.P., Thompson, K.G., and Schall, J.D. (1995). Relationship of presaccadic activity in frontal eye field and supplementary eye field to saccade initiation in macaque: Poisson spike train analysis. Exp. Brain Res. *103*, 85–96.

10. Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2012). Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. *25*.

11. Gross, C.G., Rocha-Miranda, C.E., and Bender, D.B. (1972). Visual properties of neurons in inferotemporal cortex of the Macaque. J. Neurophysiol. *35*, 96–111.

12. Bruce, C., Desimone, R., and Gross, C.G. (1981). Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. J. Neurophysiol. *46*, 369–384.

13. Desimone, R., Albright, T.D., Gross, C.G., and Bruce, C. (1984). Stimulus-selective properties of inferior temporal neurons in the macaque. J. Neurosci. *4*, 2051–2062.

14. Tsao, D.Y., Freiwald, W.A., Knutsen, T.A., Mandeville, J.B., and Tootell, R.B.H. (2003). Faces and objects in macaque cerebral cortex. Nat. Neurosci. *6*, 989–995.

15. Tsao, D.Y., Moeller, S., and Freiwald, W.A. (2008). Comparing face patch systems in macaques and humans. Proc. Natl. Acad. Sci. U. S. A. *105*, 19514–19519.

16. Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. Q. J. Exp. Psychol. A *43*, 161–204.

17. Riesenhuber, M., and Poggio, T. (1999). Hierarchical models of object recognition in cortex. Nat. Neurosci. *2*, 1019–1025.

18. Tsao, D.Y., and Freiwald, W.A. (2006). What's so special about the average face? Trends Cogn. Sci. *10*, 391–393.

19. Freiwald, W.A., Tsao, D.Y., and Livingstone, M.S. (2009). A face feature space in the macaque temporal lobe. Nat. Neurosci. *12*, 1187–1196.

20. Kay, K.N., Naselaris, T., Prenger, R.J., and Gallant, J.L. (2008). Identifying natural images from human brain activity. Nature *452*, 352–355.

21. Cowen, A.S., Chun, M.M., and Kuhl, B.A. (2014). Neural portraits of perception: reconstructing face images from evoked brain activity. Neuroimage *94*, 12–22.

22. Nestor, A., Plaut, D.C., and Behrmann, M. (2016). Feature-based face representations and image reconstruction from behavioral and neural data. Proc. Natl. Acad. Sci. U. S. A. *113*, 416–421.

23. Dosovitskiy, A., and Brox, T. (2016). Generating images with perceptual similarity metrics based on deep networks. Adv. Neural Inf. Process. Syst. *29*.

24. Ponce, C.R., Xiao, W., Schade, P.F., Hartmann, T.S., Kreiman, G., and Livingstone, M.S. (2019). Evolving Images for Visual Neurons Using a Deep Generative Network Reveals Coding Principles and Neuronal Preferences. Cell *177*, 999-1009.e10.

25. Sirovich, L., and Kirby, M. (1987). Low-dimensional procedure for the characterization of human faces. J. Opt. Soc. Am. A *4*, 519–524.

26. Turk, M.A., and Pentland, A.P. (1991). Face Recognition Using Eigenfaces. https://www.cin.ufpe.br/~rps/Artigos/Face%20Recognition%20Using%20Eigenfaces.pdf.

27. Taigman, Y., Yang, M., Ranzato, M., and Wolf, L. (2014). DeepFace: Closing the gap to human-level performance in face verification. In 2014 IEEE Conference on Computer Vision and Pattern Recognition (IEEE), pp. 1701–1708.

28. Simonyan, K., and Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv [cs.CV].

29. Parkhi, O., Vedaldi, A., and Zisserman, A. (2015). Deep face recognition. BMVC 2015 - Proceedings of the British Machine Vision Conference 2015.

30. Kubilius, J., Schrimpf, M., Nayebi, A., Bear, D., Yamins, D.L.K., and DiCarlo, J.J. (2018). CORnet: Modeling the Neural Mechanisms of Core Object Recognition. bioRxiv, 408385. 10.1101/408385.

31. Kubilius, J., Schrimpf, M., Kar, K., Rajalingham, R., Hong, H., Majaj, N., Issa, E., Bashivan, P., Prescott-Roy, J., Schmidt, K., et al. (2019). Brain-like object recognition with high-performing shallow recurrent ANNs. Adv. Neural Inf. Process. Syst. *32*.

32. Kar, K., Kubilius, J., Schmidt, K., Issa, E.B., and DiCarlo, J.J. (2019). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. Nat. Neurosci. *22*, 974–983.

33. Chang, L., Egger, B., Vetter, T., and Tsao, D.Y. (2021). Explaining face representation in the primate brain using different computational models. Curr. Biol. *31*, 2785-2795.e4.

34. Kanwisher, N., McDermott, J., and Chun, M.M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. J. Neurosci. *17*, 4302–4311.

35. Hesse, J.K., and Tsao, D.Y. (2020). The macaque face patch system: a turtle's underbelly for the brain. Nat. Rev. Neurosci. *21*, 695–716.

36. Kosslyn, S.M., and Pylyshyn, Z. (1995). Image and brain: the resolution of the imagery debate. J. Cogn. Neurosci. *7*, 415–420.

37. O'Craven, K.M., and Kanwisher, N. (2000). Mental imagery of faces and places activates corresponding stimulus-specific brain regions. J. Cogn. Neurosci. *12*, 1013–1023.

38. Kamitani, Y., and Tong, F. (2005). Decoding the visual and subjective contents of the human brain. Nat. Neurosci. *8*, 679–685.

39. Miyashita, Y., and Chang, H.S. (1988). Neuronal correlate of pictorial short-term memory in the primate temporal cortexYasushi Miyashita. Nature *331*, 68–70.

40. Rutishauser, U., Schuman, E.M., and Mamelak, A.N. (2006). Online detection and sorting of extracellularly recorded action potentials in human medial temporal lobe recordings, in vivo. J. Neurosci. Methods *154*, 204–224.

41. Reuter, M., Rosas, H.D., and Fischl, B. (2010). Highly accurate inverse consistent registration: a robust approach. Neuroimage *53*, 1181–1196.

42. Pauli, W.M., Nili, A.N., and Tyszka, J.M. (2018). A high-resolution probabilistic in vivo atlas of human subcortical brain nuclei. Sci Data *5*, 180063.

43. Avants, B., Duda, J.T., Kim, J., Zhang, H., Pluta, J., Gee, J.C., and Whyte, J. (2008). Multivariate analysis of structural and diffusion imaging in traumatic brain injury. Acad. Radiol. *15*, 1360–1375.

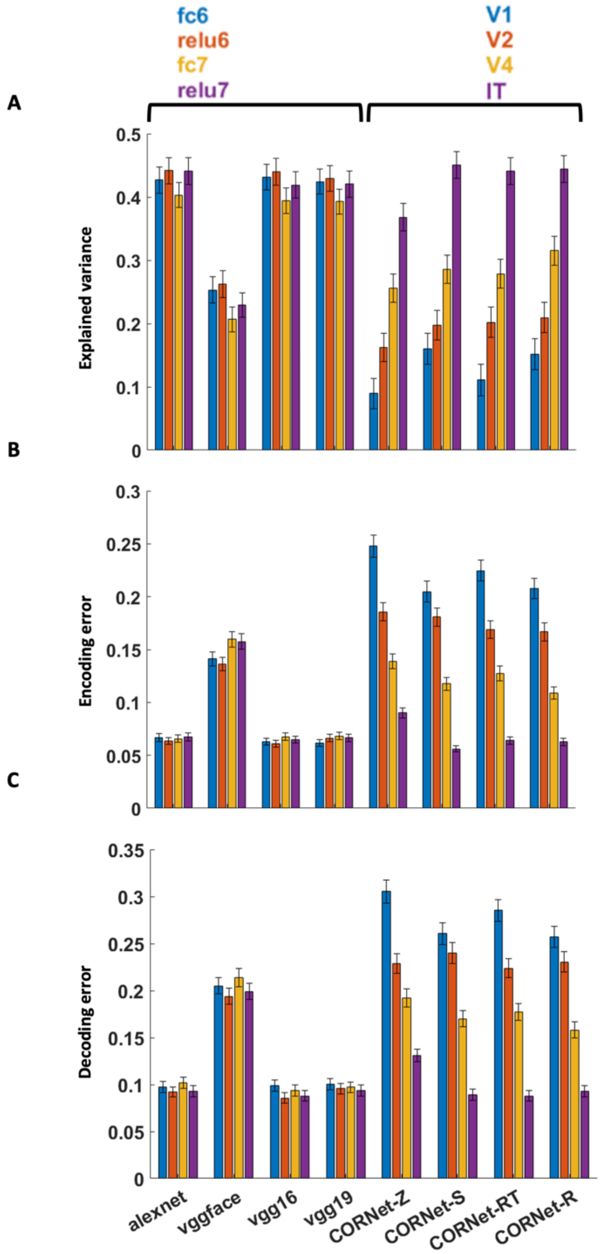
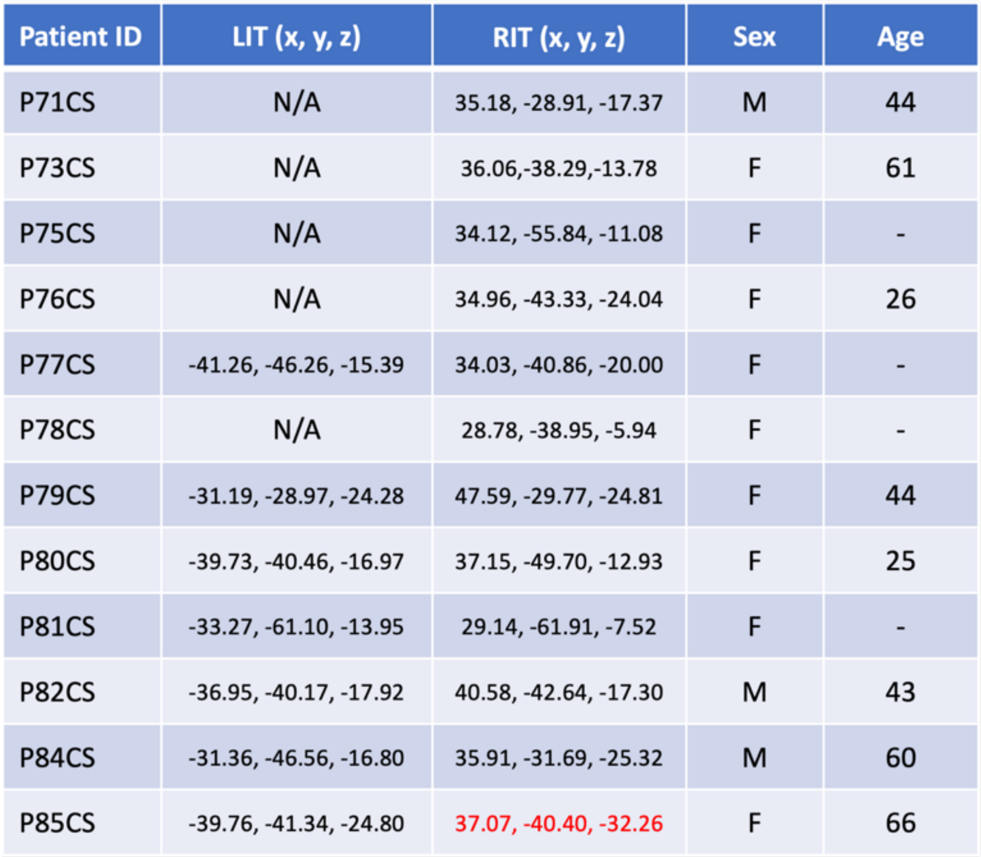


Figure 3.S1. Comparing explained variance, encoding, and decoding error across layers of the various models. Related to Figure 3.5.

(A) Explained variance for the two fully connected layers pre and post relu in AlexNet, VGG-Face, VGG-16, VGG-19 and for the output layers of V1, V2, V4, and IT layers in the CORNets. While the performance across the fully connected layers is similar and not affected by ReLu activation, the performance varies greatly across the 4 layers in the CORNets with IT outperforming the other layers substantially.

(B) Encoding error across the different layers, results are consistent with the explained variance.

(C) Decoding error across the different layers.



Patient ID	LIT (x, y, z)	RIT (x, y, z)	Sex	Age
P71CS	N/A	35.18, -28.91, -17.37	M	44
P73CS	N/A	36.06,-38.29,-13.78	F	61
P75CS	N/A	34.12, -55.84, -11.08	F	-
P76CS	N/A	34.96, -43.33, -24.04	F	26
P77CS	-41.26, -46.26, -15.39	34.03, -40.86, -20.00	F	-
P78CS	N/A	28.78, -38.95, -5.94	F	-
P79CS	-31.19, -28.97, -24.28	47.59, -29.77, -24.81	F	44
P80CS	-39.73, -40.46, -16.97	37.15, -49.70, -12.93	F	25
P81CS	-33.27, -61.10, -13.95	29.14, -61.91, -7.52	F	-
P82CS	-36.95, -40.17, -17.92	40.58, -42.64, -17.30	M	43
P84CS	-31.36, -46.56, -16.80	35.91, -31.69, -25.32	M	60
P85CS	-39.76, -41.34, -24.80	37.07, -40.40, -32.26	F	66

Table 3.1. MNI coordinates of microwire bundles and associated patient metadata. Related to Figure 3.1.

In cases where details are not known (for example, age) they are left blank. Multiple patients only got right posterior temporal electrodes and are thus labeled ‘N/A’ on the left. The coordinate labeled in red is one from where we recorded 0 responsive units.

Chapter IV

IMAGINATION: Reactivation of the feedforward code for visual objects

Introduction:

Chapter III focused on uncovering the precise code for visual objects — understanding how we convert visible object features into mental representations. However, our internal model enables the opposite transformation as well — to generate sensory percepts starting from mental representations. This remarkable cognitive ability, known as mental imagery, allows us to imagine new sensations and experiences[1,2], remember old ones[3], make plans[4], and make sense of sparse data in sensation, action, or language[5,6]. For example, blurry black dots moving in a line are likely ants if I am standing 6 feet tall but are likely cars on a highway if I am 10,000 feet in the air. In a clinical setting, uncontrolled mental imagery could contribute to post-traumatic stress disorder[7] while being able to detect imagery could help reveal conscious awareness in locked-in patients[8].

Given its far reaching consequences mental imagery has been studied with intense fascination for many decades[1,9–13]. While mental imagery can cover all five senses and even emotional states, visual imagery research has tended to dominate as is also the case with perception research[2]. This is understandable, as a large part of the primate brain is dedicated to visual processing in line with its importance to our lives[14] a point re-iterated in the last Chapter. Animal studies and the work described in Chapter III have together yielded rich insight into the bottom up processing of visual objects[15–19], however the neural mechanisms of visual imagery have remained much more elusive.

This is due to animal studies being hampered by the inability to instruct animals to reliably and repeatedly generate a mental image and previous studies in human subjects, even those recording individual neurons[20,21], being hampered by the code for visual objects being unknown. In the same manner that we cannot appreciate the magic of Shakespeare without first understanding basic English, we need to first understand the visual code for objects before we can attempt to unpack how they are imagined.

In this study, armed with a detailed understanding of the feedforward code for visual objects in human brain (described in Chapter III) and the ability to record individual neurons in human IT with high fidelity (described in Chapter II) we present new findings on the neural mechanisms of visual imagery, namely that it is remarkably similar to visual perception and that activation during imagery at the single neuron level is not localized to any one hemisphere or region.

Results:

A long standing hypothesis in the field has been that long term memories are stored in the same cortical populations that encode sensory stimuli[22] and therefore sensory neurons subserve imagery by reinstating sensory context. In particular, it has been proposed that visual memory is supported by the activity of neurons in IT cortex[20,23] though a detailed understanding of how widespread this reactivation is or whether the reactivation is structured in a manner that respects the visual code is unknown.

Neurons in IT cortex reactivate during imagery

We recorded 173 IT neurons in 11 sessions across 5 patients during a cued imagery task (Figure 4.1A). The imagery task involved patients viewing and subsequently visualizing from memory 6-8 visual objects (see methods). Out of the 173 neurons recorded 147 were found to be visually responsive via Poisson burst metric[24] (see Chapter III methods), and 92 were axis tuned.

We find robust reactivation of neurons in human IT cortex upon imagery. As in vision, neurons showed a diverse range of reactivation profiles during imagery, some reactivating sparsely to a specific stimulus (Figure 4.1B) others reactivating to imagery of multiple stimuli in a graded manner (Figure 4.1E), and a few that were highly active during imagery but did not participate in visual perception (Figure 4.1D) as predicted by early studies of patients with brain damage that displayed intact imagery despite profound visual agnosia which indicates a shared but not identical mechanism for both[25–27]. We find that a substantial minority of neurons (75/173, 43%) are active during imagery (see methods).

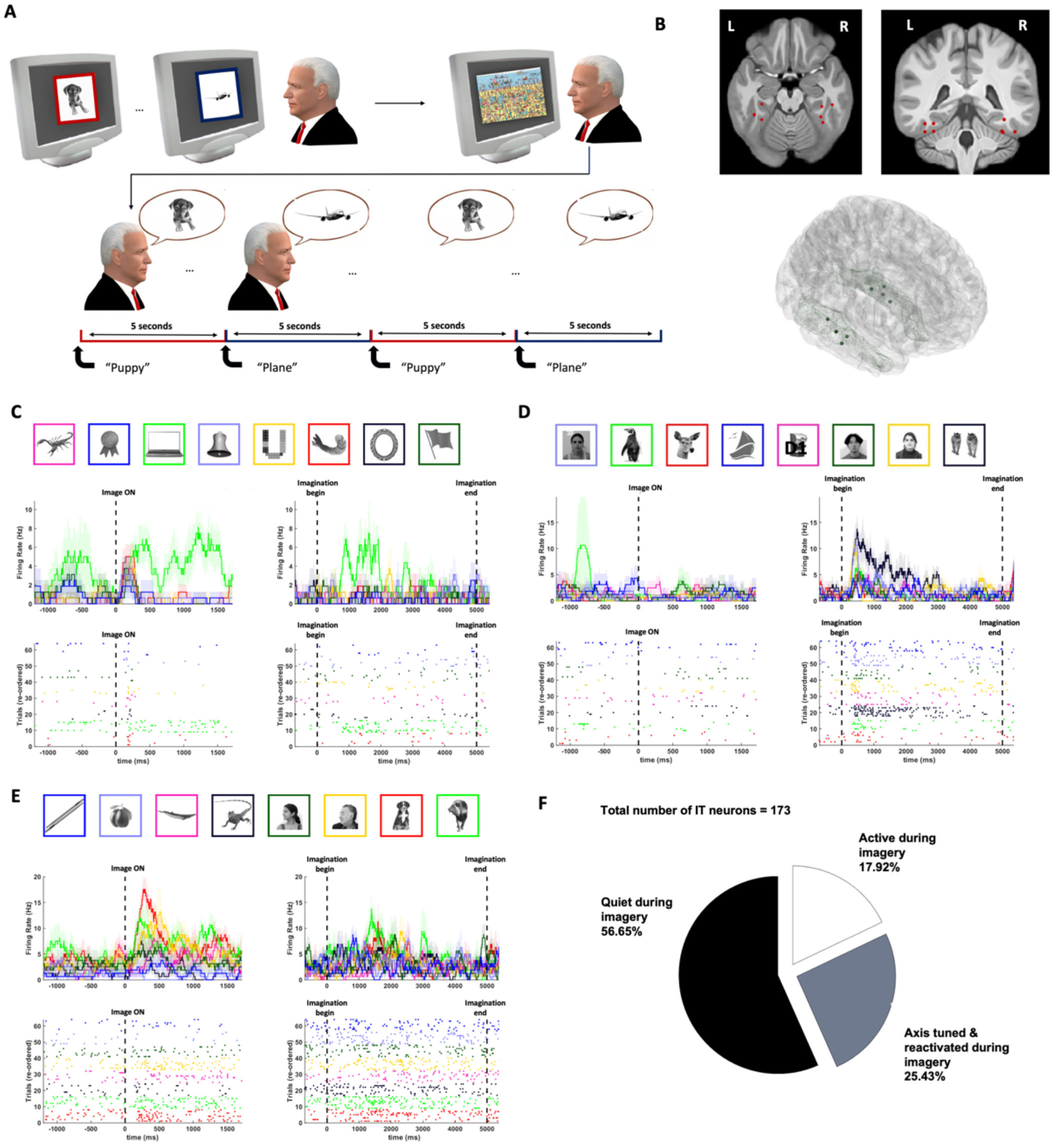


Figure 4.1. Cued Imagery task reveals robust reactivation of IT neurons.

(A) Schematic of the cued imagery task utilized. 6-8 of the background subtracted grayscale screening images were utilized, with 2 images used per trial. Each trial consisted of an 'encoding' period where patients viewed the stimuli multiple times each, a 'distraction' period wherein patients did a visual search for 30s, and finally a cued imagery period where patients would visualize each image 4 times for 5s each in an alternating fashion prompted verbally by the experimenter.

(B) Recording locations of the 8 microwire bundles (left and right) that contained at least one well isolated neuron in human IT cortex across all 5 patients. See Montreal Neurological Institute coordinates in Table 1. Each dot represents the location of one microwire bundle (8 channels).

(C) As in viewing, neurons in human IT cortex show diverse response during imagery. An example neuron reactivating during imagery. This neuron's preferred stimulus was the laptop during encoding (left column) and it reactivated robustly during imagery of the laptop (right column).

(D) A small number of neurons recorded were quiet during encoding but strongly active during imagery.

(E) Example of an axis tuned unit that reactivated to multiple stimuli in a graded manner during imagery.

(F) Distribution of neurons recorded that were active during imagery of objects.

Reactivation occurs in a structured manner that recapitulates sensory context

In order to truly understand mental imagery it is necessary to compare mental representations of visual objects to a concrete mechanism for sensory processing. In Chapter III we presented the axis model as a model for object recognition in human IT neurons. Thus, in order to compare representations it was necessary to establish axis tuning in the recorded neurons, sample stimuli in a principled way that have specific response differences in viewing, and investigate whether or not those firing rate differences are preserved when the same stimuli are imagined.

Our workflow to accomplish this goal was as follows: a morning screening session was conducted using the 500 object images that had varied features (see Chapter III methods), axes were computed for the axis tuned neurons, then for an exemplar neuron 3-4 pairs of stimuli were chosen that had increasing projection value along the neurons preferred axis (Figure 4.2). Each pair had roughly similar projection value onto the preferred axis but were spread out along the orthogonal axis. Note that the stimuli would be spread out in perfect pairs for only one exemplar neuron, however pooling the positions of all imagined stimuli across patients demonstrates a substantial spread across both axes nonetheless (Figure 4.3E). The axis model would predict that during viewing stimuli spread across the orthogonal axis would elicit roughly the same response from the neuron but along the preferred axis there would be a systematic increase in response as a function of the projection value which was verified (Figure 4.2C&E). We find that same structured response plays out during imagery. Individual neuron examples can be seen in Figure 4.2B-E. Examining the population of axis tuned neurons recorded during imagery finds a significant correlation between projection value onto the neurons' preferred axes and responses during imagery with no such correlation along the orthogonal axes (Figure 4.3B&C). Lastly, the distribution of correlation coefficients between viewing and imagery in axis tuned neurons that reactivated during imagery shows a high average value (0.38; Figure 4.3A) with 21/44 reactive neurons being significantly correlated as compared to a shuffled distribution ($p < 0.05$). Taken together these findings indicate that neurons in human IT cortex support mental imagery through a substantial minority of those active in viewing reinstating sensory context during imagery.

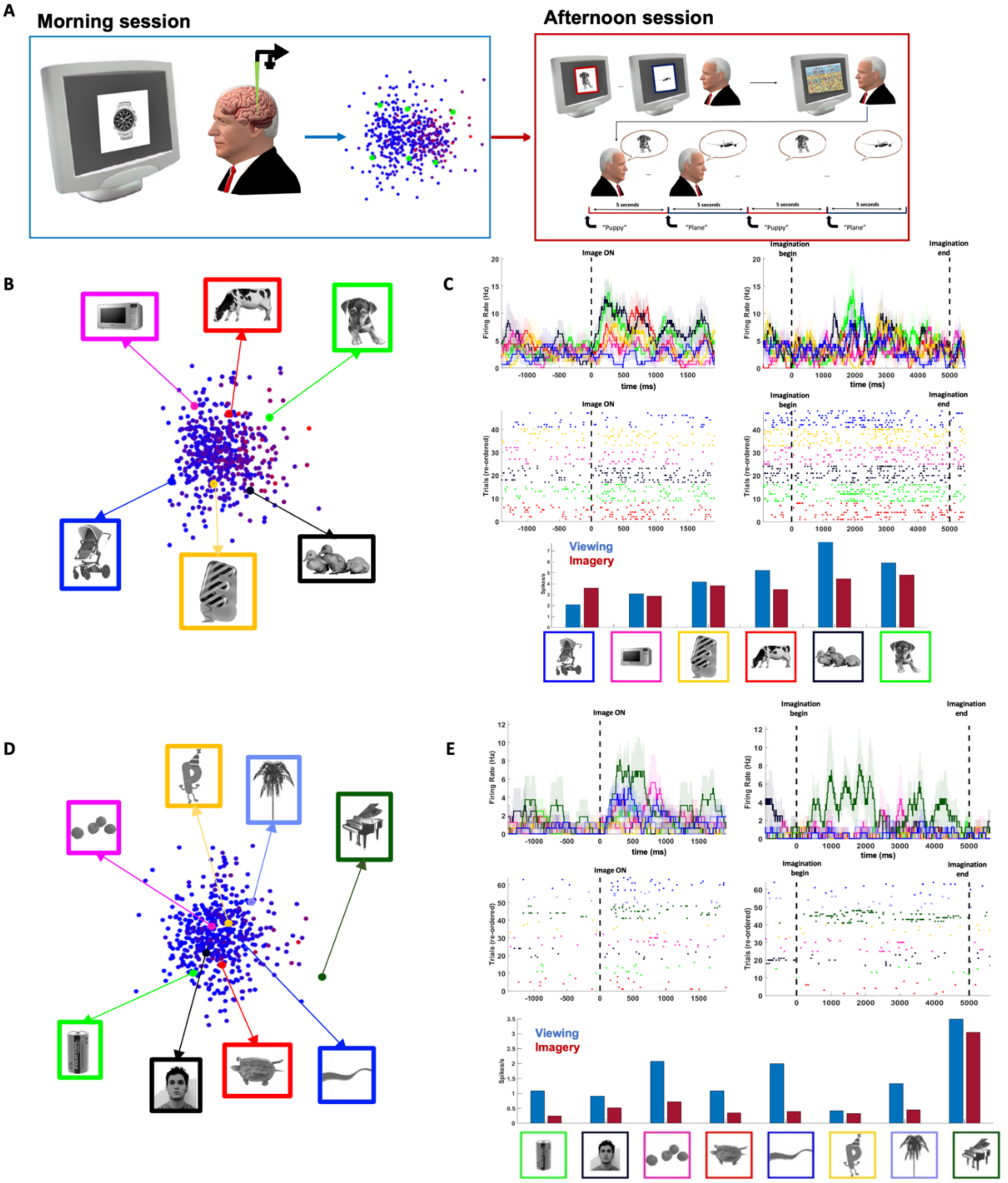


Figure 4.2. Neurons in IT cortex reactivate in a structured manner that recapitulates visual context.

(A) Schematic of the workflow used to investigate structured reactivation. A morning screening session was conducted wherein axis tuned neurons were identified. 6-8 stimuli were chosen that had some spread along the preferred and principal orthogonal axes, then the cued imagery task was carried out in the afternoon followed by another screening session, which would help to match the neurons recorded in the morning and afternoon sessions.

(B) An axis plot for an example neuron showing the 6 stimuli chosen for cued imagery.

(C) The responses of the neuron in (B) during encoding/viewing (left column) and imagery (right column). The bottom panel shows average spike counts for viewing and imagery of the stimuli, arranged in order of increasing projection value along the preferred axis.

(D) Axis plot for another example neuron demonstrating reactivation in a manner similar to visual context. Overall, 44/92 axis unted neurons reactivated during imagery.

(E) Responses of the neuron in (D) during encoding and imagery. The neurons response aligns with the projection value onto the preferred axis.

Responses in IT neurons during imagery capture the feature details of objects

In Chapter III we discussed how if neurons in IT are indeed acting as linear projectors in feature space then their responses can be approximated by a linear combination of object features and that by inverting the equation we were able to decode object features using the neurons' responses. If neurons are indeed respecting the visual code then in principle we should be able to decode the features of the imagined objects using the same method.

We attempted to decode the fine feature details of the objects being imagined using the responses of the axis tuned IT neurons during imagery. We used leave-one-out cross validation to learn the linear transform between the imagined responses and the features (Figure 4.4A). We reduced the number of features to be less than the number of neurons simultaneously recorded in each session (to keep the polynomial well-conditioned) and the representations were impressively rich allowing the objects imagined in a given session to be identified amongst many distractors (Figure 4.4B). We subsequently attempted to 'reconstruct' objects by searching a large auxiliary object database for the objects with the feature vector closest to that decoded from neural activity. A normalized distance in feature space between the best possible reconstruction and actual reconstruction pooled across all sessions (see methods) was computed to quantify the decoding accuracy and can be seen in Figure 4.4C. The model performed remarkably well — with low normalized distances and the reconstructions capturing the fine feature details of the objects imagined (Figure 4.4D).

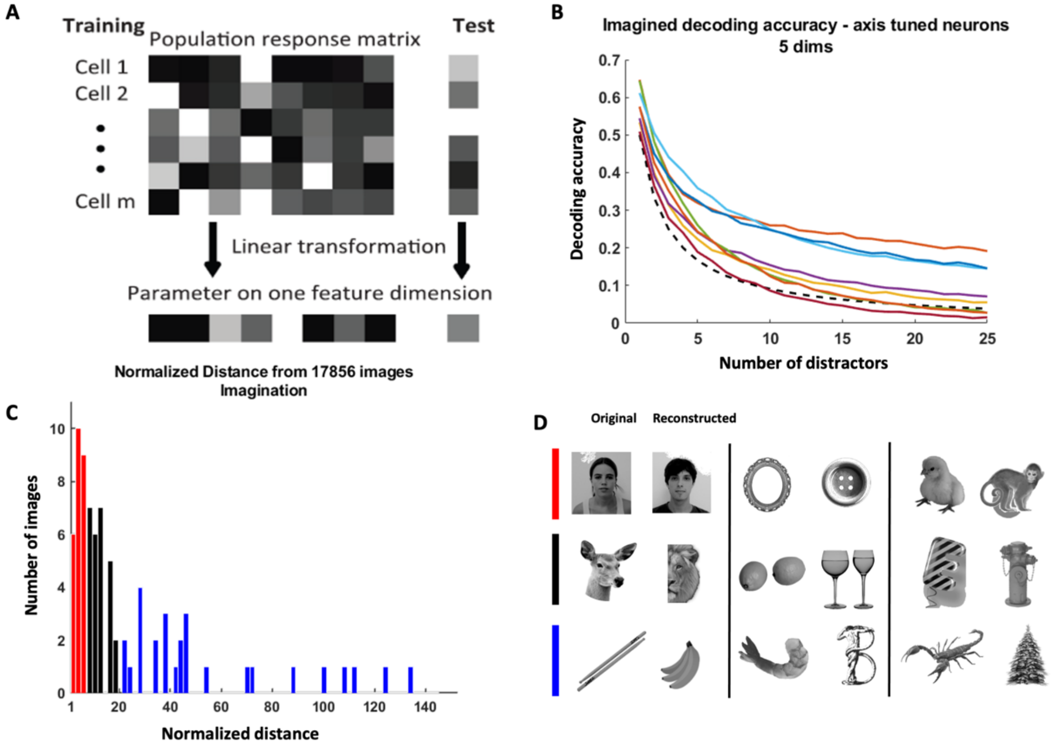


Figure 4.3. Decoding imagined object features via linear regression

(A) Diagram illustrating decoding model. We used responses to all but one object to determine the transformation between imagined responses and feature values by linear regression, before using that transformation to predict the feature values of the held out object.

(B) Decoding accuracy as a function of the number of distractor objects drawn randomly from the stimulus set. For each model, different sets of features were first linearly decoded from population responses, then the distances between decoded and actual features in each feature space were computed to determine decoding accuracy. The black dashed line represents the decoding accuracy one would expect by chance. Instead of decoding 50 features for each object as we did in viewing we reduced the number of dimensions to 5 in order to keep the polynomial well conditioned. Each line corresponds to the simultaneously recorded neurons in each session.

(C) Distribution of normalized distances between reconstructed feature vectors and the best-possible reconstructed feature vectors for the visualized images (see methods) to quantify decoding accuracy across the population. The normalized distance takes into account the fact that the object images used for reconstruction did not include any of the object images shown to the patients. A normalized distance of 1 means the reconstruction found the best solution possible.

(D) Images were split into tertiles based on normalized distance. Examples of the reconstructions in the first tertile (top row), second (middle row) and third (bottom row) as compared to the original stimulus images being reconstructed.

A note on the heterogeneity of mental representations across people

There is a large body of evidence to support the notion that the subjective vividness of visual imagery varies greatly between individuals[1,10,28], with some individuals demonstrating a complete inability to generate a mental image (aphantasia)[29] while others have photorealistic mental images (hyperphantasia). Moreover, various neuroimaging studies have shown differences in fMRI bold signals — both intensity in early visual areas[30] and functional connectivity between areas[10,31] — between subjects that reported different amounts of vividness in imagery. Eventually growing evidence of these differences led to the conclusion that examining mental imagery at the group level with the current tools (fMRI and psychophysics) was not appropriate — leading to the end of the imagery debate[32,33]. It is natural at this point to wonder why the debate went on for as long as it did, given that an understanding of these differences dates all the way back to the 1800s[28], though that particular curiosity is beyond the scope of this thesis.

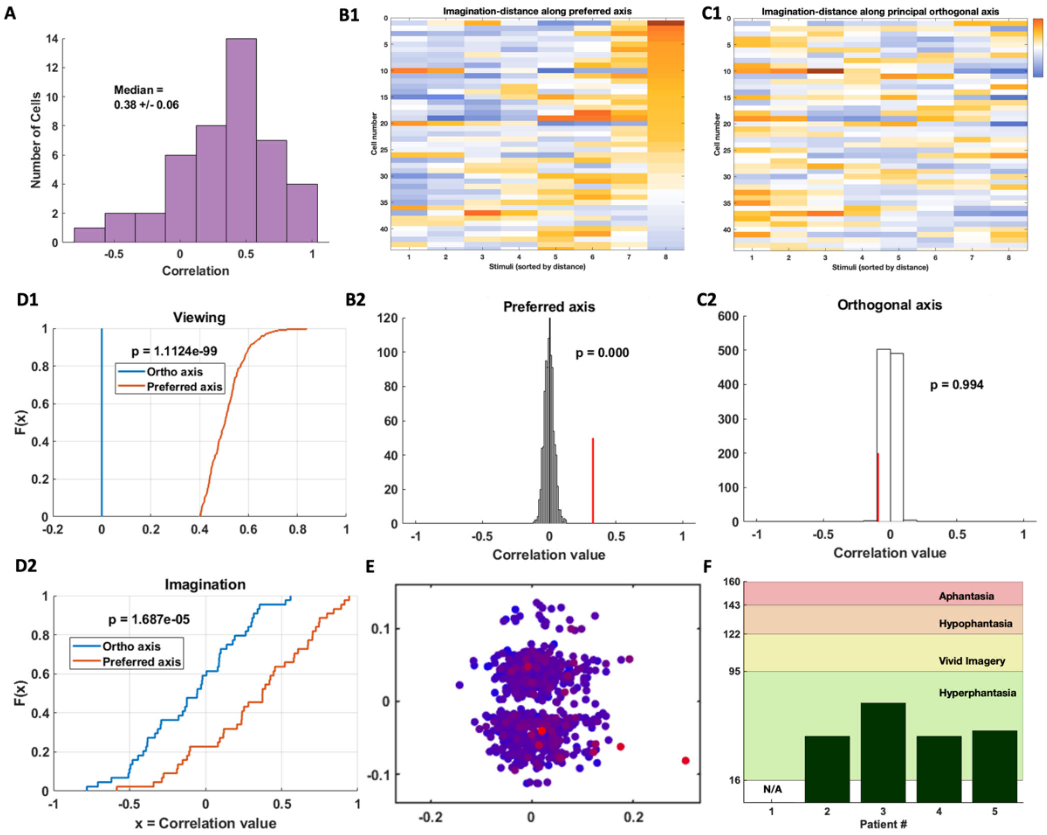


Figure 4.4. Population summary of reactivated IT neurons

(A) Distribution of correlation coefficients between viewing and imagery for neurons that reactivated. The responses to each stimulus in encoding and imagery were average across trials and the Pearson correlation coefficient was computed between those 2 vectors for each neuron.

(B) (B1) Population summary showing that axis tuning is respected during imagination in all 44 neurons that reactivated and (B2) that the correlation between imagined firing rates and projection value is significant when compared to a shuffled distribution.

(C) C1 and C2 are the same as B1 and B2 but for the principal orthogonal axis.

(D) (D1) CDF of correlation between projection value and firing rate for the preferred axis (red) and orthogonal axis (blue) during screening and (D2) during imagery. There is no correlation between them along the orthogonal axis but a strong correlation along the preferred axis in both cases.

(E) Plot showing the overall spread of chosen stimuli for imagery across all recorded neurons. There is a reasonable spread between the preferred and orthogonal axes.

(F) VVIQ scores of the 4/5 patients that responded to the VVIQ questionnaire. All patients recorded from in this study were 'hyperphantasic'. The survey is now conducted before any experiment sessions, however for the first 2 patients it had to be filled out retroactively and we were unable to get in contact with the very first one.

In order to understand whether there is a correlation between the data discussed here and subjective vividness 4/5 patients also completed the Vividness of Visual Imagery Questionnaire (VVIQ)[34], a standard in the cognitive science field for mapping out subjective vividness of mental images. These responses were recorded in person starting with the 3rd patient in this study, and attempts to retroactively collect responses were unsuccessful in one patient. Remarkably, all the patients discussed here had very high scores in the vividness scale, with every single one of them falling into the 'hyperphantasic' category (Figure 4.3F). Even the one left out patient who did not fill out the VVIQ is a visual artist and had reported very clear visualization capabilities during the experiment sessions. Thus, it is important to admit that perhaps the strong recapitulation of sensory context demonstrated in this Chapter might be restricted to those with very strong visual imagery capabilities.

Reactivation across other brain regions

Previous neuroimaging studies that conducted detailed investigations of visual imagery identified a few regions that seemingly support imagery including frontal regions, the ventral temporal cortex, parietal regions, and early visual areas (V1)[1,2], and reported a strong hemispheric asymmetry with the left hemisphere being the dominant driver of our imagery capabilities[9,13,25,35–37]. However, this is inconsistent with our current findings at the single neuron level. In fact, we find that re-doing the analyses summarized in Figure 4.3 for left and right IT neurons separately shows that the results hold in both hemispheres independently (Figure 4.S1).

Lastly, despite the purported localization of human visual imagery capabilities to specific regions we find robust activations during imagery in at least some neurons in every single region we recorded from with a seemingly even distribution of reactive neurons in both hemispheres (Table 4.2). In total across all regions, we recorded 366/1019 imagery neurons. As in IT, we observed a wide variety of reactivation responses with some neurons showing an increase in firing rate that was selective for specific stimuli (Figure 4.4A1, A2), others showing non-selective increases in response during imagery (Figure 4.4A4), and some surprisingly being initially suppressed by imagery (Figure 4.4A3). The interpretation of these brain wide imagery activations and their response diversity is left to future studies.

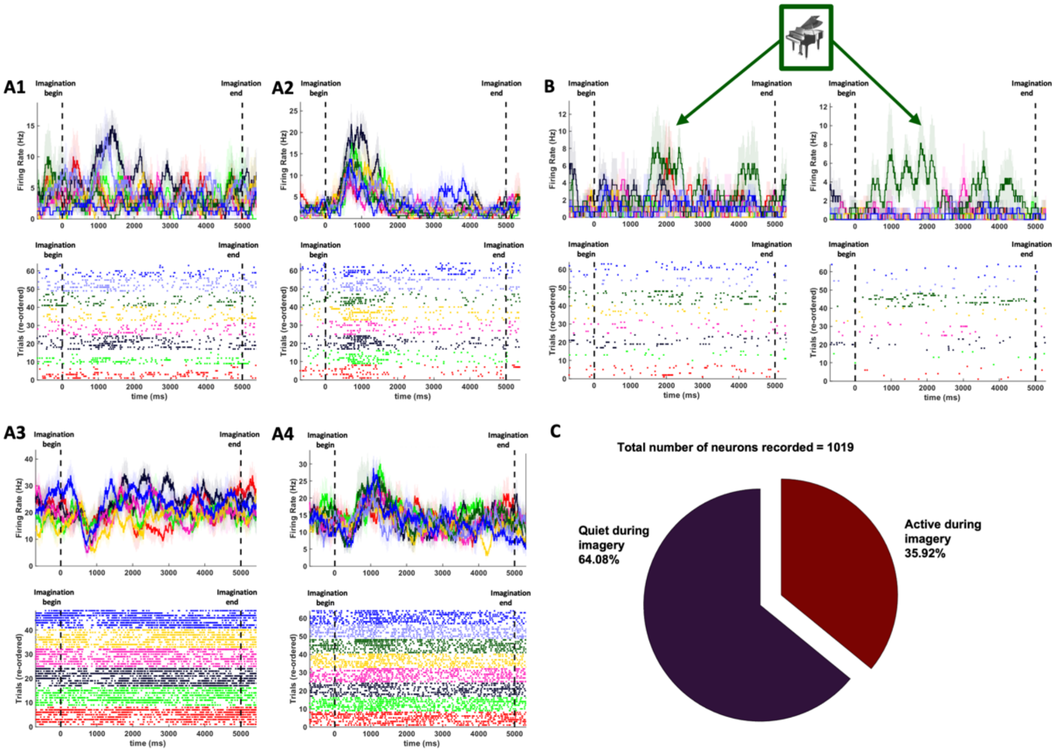


Figure 4.5. Neurons activate across the brain during imagery.

(A) Examples of the diverse responses recorded from neurons across the brain during imagery. (A1) A neuron in the right amygdala showing robust activity during imagery. (A2) Neuron in the right centro-medial nucleus of the thalamus. (A3) Neuron in the left amygdala that showed an initial suppression of activity during imagery. (A4) Neuron in the left pre-supplementary motor area showing a non-selective increase in activity during imagery.

(B) Two example neurons recorded simultaneously, one in the left orbitofrontal cortex (left) and one in left IT (right) that were active during the imagery of a piano. The IT neuron is the same as the one shown in Figure 2D&E. Analyses to determine functional connectivity are underway.

(C) Distribution of the total number of neurons across the brain that were active during imagery. For a detailed breakdown per region and hemisphere see Table 2.

Discussion:

Our internal model being generative has lots of consequences. It allows us to generate percepts in the absence of any external stimuli which supports fast and flexible learning via simulation of outcomes[4,38], visual inference in cases where the data is uncertain, planning and exploration of new sensations, and is the basis of all of our memory for experiences. The long-standing hypothesis in the field was that sensory areas sub serve imagery by reinstating sensory context. However while this was seen in many neuroimaging studies[1,10,39], the evidence at single neuron resolution was slim restricted to demonstrating the reactivation of a few neurons[20,23].

We address both of these limitations by recording a large number of neurons in IT cortex, mapping out the feedforward code for visual objects and demonstrate the reactivation of this code during imagery. We find a substantial number of neurons (75/173 total, 44/92 axis tuned) reactivate during imagery in a manner that respects the visual code, as evidenced by the high correlation between viewed and imagined responses (Figure 4.3A). In the case of axis tuned neurons we see that the axis code is respected during imagery as well (Figure 4.3) indicating that reactivation is strong and structured.

At this stage it is important to point out that a lot of the results detailed in this Chapter rest on the quantitative method for computing activity during imagery and thus labeling an IT neuron 'reactive'. Current methods have treated individual stimuli as categories and computed reactivation in a manner that tracks a large increase in mean firing rate across the imagination period or a differential activation of the various stimuli during imagery. These methods do not allow us to analyze reactivation on an individual trial level. Moreover, this Chapter conspicuously lacks connectivity analyses, i.e. attempts to determine the origin of top down signals that would trigger such reactivation in IT. Further analyses in both domains are underway.

Regardless, we have provided concrete proof of the generative visual pathway, and demonstrated its interaction with the feedforward one. This constitutes an important first step to understanding many complex phenomena that are important to us — like the representation of individual episodic memories — that have been out of reach until now.

Methods:

Participants

Patients with drug-resistant epilepsy

5 patients with drug-resistant epilepsy volunteered for this study and gave their informed consent. The institutional review board of Cedars-Sinai Medical Center approved all protocols. The task was conducted while patients stayed in the hospital after implantation of depth electrodes for monitoring seizures. The location of the implanted electrodes was solely determined by clinical needs. The neural results were analyzed across all 5 patients.

Task

Screening

Patients viewed a set of 500 object stimuli with varying features (taken from www.freepngs.com) 4 times each for a total of 2000 trials in a shuffled order. Each image stayed on screen for 250ms and the inter-trial interval consisting of a blank screen was jittered between 100-150ms. The task was punctuated with yes/no 'catch' questions pertaining to the image that came right before the question requiring the patients to pay close attention in order to answer them correctly. Catch questions occurred between 2 to 80 trials after the previous one. Patients responded to the questions using a RB-844 response pad (https://cedrus.com/rb_series/).

Cued Imagery

Patients viewed a set of 6-8 object stimuli chosen from the 500 used for screening (taken from www.freepngs.com). Each trial focused on 2 images and had an encoding period, a visual search distraction period, and a cued imagery period. During encoding patients would see the 2 images 4 times each in a shuffled order. Each image stayed on screen for 1.5s and the inter-image interval was 500-800ms. After this encoding period a visual search puzzle was presented (puzzles created by artist Gergely Dudas https://thedudolf.blogspot.com/) and stayed on screen for 30s. After reporting via button press whether or not they were able to find the object in the puzzle, patients began the cued imagery period. During cued imagery patients would close their eyes and imagine the stimuli in the trial in an alternating fashion for 4 repetitions of 5s each (40s continuous imagery period). Patients were cued to switch every 5s by verbal cue. After the imagery period patients would begin the next trial via button press when ready. Every image was present in 2 trials, leading to 8 repetitions of both encoding and imagery for each image.

Workflow

To examine the interaction between imagery and the feedforward code for objects we structured our experiment days as follows:

A morning session was conducted to identify axis tuned neurons. Then 6-8 stimuli were chosen for the imagery task that had some spread along both the preferred and orthogonal axes. For an exemplar neuron, they were chosen in pairs such that each pair had increasing projection values along the preferred axis but the within-pair images only differed along the orthogonal axis. Note that for the other neurons that had axes pointing in different directions in object space the stimuli did not end up nicely spread out in pairs — however on average we got good spread along both axes (Figure 4.3E).

After axis tuned neurons were identified and the stimuli chosen we conducted the cued imagery task, followed by another screening session immediately after. The second screening was used match the neurons from the morning and afternoon sessions (see Chapter III methods).

Electrophysiology

We recorded broadband signals from inferotemporal cortex in addition to the standard epileptic targets (amygdala, hippocampus, anterior cingulate cortex, pre-supplementary motor area, and orbitofrontal cortex) using micro-macro electrodes manufactured by Ad-Tech (https://adtechmedical.com/epilepsy#depth-electrodes). All analyses pertain to signals recorded on the 8 microwires protruding from the end of the electrode and sampled at 32khz.

Spike sorting and quality metrics

Signals were bandpass filtered offline in the range of 300-3000hz with a zero phase lag filter before spike detection. Spike detection was carried out by the semiautomated template matching algorithm Osort[40]. The spike quality metrics returned by the algorithm were examined before final clusters were curated.

Electrode localization

Electrode localization was based on postoperative MRI/computed tomography (CT) scans. We co-registered postoperative and preoperative MRIs using Freesurfer's mri_robust_register[41]. To summarize recording locations across participants we aligned each participant's preoperative scan to the CITI168 template brain in MNI152 coordinates[42] using a concatenation of an affine transformation and symmetric image normalization (SyN) diffeomorphic transform[43]. The MNI coordinates of the microwires from a given electrode shank were marked as one location. MNI coordinates of microwires with putative neurons detected from all participants were plotted on a template brain for visualization (Figure 4.1B).

Data Analyses

Reactivation metric

An important part of the work described in this Chapter is identifying neurons that reactivate. As such it is important to have a quantitative measure for 'reactivation'. For our data we identified reactivated neurons as follows: we first computed the neurons baseline firing rate then set a threshold 2 standard deviations above it. If a neuron during imagery had an average value above that threshold for any stimulus or a 1 way ANOVA between the groups during imagery was significant ($p < 0.05$) the neuron was considered active during imagery.

Correlation of viewed and imagined responses

For neurons that were active during imagery we computed the correlation between viewed and imagined responses. To compute the viewed response to each stimulus, we computed the neurons response latency via a Poisson burst method (see Chapter III methods), then collected spikes in a window of 500ms after the response onset of the neuron and averaged across repetitions of a given stimulus. To compute the imagined response we simply averaged the response across the 5s time period for every trial and further averaged across repetitions of a given stimulus. The Pearson correlation (r) was then computed between these 2 vectors.

Vividness of Visual Imagery Questionnaire (VVIQ)

The VVIQ was taken from https://davidfmarks.net/vividness-of-visual-imagery-questionnaire-vviq/. For patients 3-5 the questionnaire was printed out and taken to the patient room the day before experiments, and the patients attempted the various imagery tests and filled out the questionnaire with the help of the experimenter. For patients 1-2 a link to an online version was emailed to them and patient 2 wrote back with their scores.

Decoding analysis

In Chapter II we find that neurons in human IT cortex are performing linear projection onto specific preferred axis in object space during viewing. As the correlation between projection value onto these axes and imagined firing rates was significant (Figure 4.3, Figure 4.S1) we assume neurons are doing the same linear projection during imagery. As such, their responses can be well modeled by the equation

$$\vec{R} = C \cdot \vec{F} \ + \overrightarrow{C_o}$$

where $\vec{R}$ is the population response vector to a given image, C is the weight matrix for different neurons, $\vec{F}$is the vector of object feature values and $\overrightarrow{C_o}$ is the offset vector. Thus the decoding analysis was carried out by inverting this equation giving us

$$\vec{F} = \vec{R} \cdot C' \ + \overrightarrow{C'_o}$$

We used the responses of all but one of the objects imagined in a given session (6/8 – 1 = 5/7) to fit C' and $\overrightarrow{C'_o}$. These were then plugged into the equation to predict the feature vector of the last object. Decoding accuracy was quantified by randomly selecting a subset of object images that included the actual feature vector of the decoded object from the total set of 500 used for screening and compared their feature vectors to the predicted feature vector of the decoded object by Euclidean distance. If the actual feature vector closest to the predicted feature vector is of the object being decoded ('target') the decoding is considered correct. This procedure is repeated 100 times for each of the 6 or 8 images with a varying number of distractors to get an aggregate measure of decoding accuracy (Figure 4.4B). Note that these analyses had to be done in a session-by-session manner as the stimuli imagined during each session were different, thus each line in Figure 4.4B corresponds to a different session. The number of feature dimensions for this analysis was reduced from 50 to 5 in order to keep the polynomial well-conditioned (number of simultaneously recorded neurons > number of feature dimensions to be decoded).

Object 'reconstruction'

To generate images that reflect the features encoded in the neural responses we gathered images from an auxiliary database and passed 17,856 background free images through AlexNet. The images were then projected into the space built by the 500 stimulus objects. None of these ~18k images had been shown to or imagined by the patients. For each stimulus image the feature vector decoded from the neural activity was compared to the feature vectors of the large stimulus set. The object in the large image set with the smallest Euclidean distance to the decoded feature vector was considered the 'reconstruction' of that imagined image[19].

To account for the fact that the large object set did not contain any images shown to the patients, which sets a limit on how good the reconstruction can be, we computed a 'normalized distance' to quantify the reconstruction accuracy for each object. We defined the normalized reconstruction distance for an image as

$$Normalized\ distance \ = \frac{|V_{recon} - V_{original}|}{|V_{best\ possible\ recon} - V_{original}|}$$

where V_{recon} is the feature vector reconstructed from neuronal responses, $V_{original}$ is the feature vector of the image presented to the patients, and $V_{best\ possible\ recon}$ is the feature vector of the best possible reconstruction (image in the large set with the closest distance to $V_{original}$). A normalized distance of 1 means the decoded image is the best reconstruction possible. As with the earlier decoding, this analysis was done in a session-by-session basis and the normalized distances computed for each session were pooled together to make Figure 4.4C.

References:

1. Pearson, J., Naselaris, T., Holmes, E.A., and Kosslyn, S.M. (2015). Mental Imagery: Functional Mechanisms and Clinical Applications. Trends Cogn. Sci. *19*, 590–602.

2. Pearson, J. (2019). The human imagination: the cognitive neuroscience of visual mental imagery. Nat. Rev. Neurosci. *20*, 624–634.

3. Tulving, E., and Others (1972). Episodic and semantic memory. Organization of memory *1*, 1.

4. Moulton, S.T., and Kosslyn, S.M. (2009). Imagining predictions: mental imagery as mental emulation. Philos. Trans. R. Soc. Lond. B Biol. Sci. *364*, 1273–1280.

5. Craik, K.J.W. (1967). The Nature of Explanation (CUP Archive).

6. Lake, B.M., Ullman, T.D., Tenenbaum, J.B., and Gershman, S.J. (2017). Building machines that learn and think like people. Behav. Brain Sci. *40*, e253.

7. Mary, A., Dayan, J., Leone, G., Postel, C., Fraisse, F., Malle, C., Vallée, T., Klein-Peschanski, C., Viader, F., de la Sayette, V., et al. (2020). Resilience after trauma: The role of memory suppression. Science *367*. 10.1126/science.aay8477.

8. Owen, A.M., Coleman, M.R., Boly, M., Davis, M.H., Laureys, S., and Pickard, J.D. (2006). Detecting awareness in the vegetative state. Science *313*, 1402.

9. Spagna, A., Hajhajate, D., Liu, J., and Bartolomeo, P. (2021). Visual mental imagery engages the left fusiform gyrus, but not the early visual cortex: A meta-analysis of neuroimaging evidence. Neurosci. Biobehav. Rev. *122*, 201–217.

10. Dijkstra, N., Bosch, S.E., and van Gerven, M.A.J. (2017). Vividness of Visual Imagery Depends on the Neural Overlap with Perception in Visual Areas. J. Neurosci. *37*, 1367–1373.

11. Dijkstra, N., Zeidman, P., Ondobaka, S., van Gerven, M.A.J., and Friston, K. (2017). Distinct Top-down and Bottom-up Brain Connectivity During Visual Perception and Imagery. Sci. Rep. *7*, 5677.

12. Winlove, C.I.P., Milton, F., Ranson, J., and Fulford, J. (2018). The neural correlates of visual imagery: A co-ordinate-based meta-analysis. Cortex.

13. Ishai, A., Ungerleider, L.G., and Haxby, J.V. (2000). Distributed neural systems for the generation of visual images. Neuron *28*, 979–990.

14. Tanaka, K. (1996). Inferotemporal cortex and object vision. Annu. Rev. Neurosci. *19*, 109–139.

15. Gross, C.G., Rocha-Miranda, C.E., and Bender, D.B. (1972). Visual properties of neurons in inferotemporal cortex of the Macaque. J. Neurophysiol. *35*, 96–111.

16. Desimone, R., Albright, T.D., Gross, C.G., and Bruce, C. (1984). Stimulus-selective properties of inferior temporal neurons in the macaque. J. Neurosci. *4*, 2051–2062.

17. Freiwald, W.A., Tsao, D.Y., and Livingstone, M.S. (2009). A face feature space in the macaque temporal lobe. Nat. Neurosci. *12*, 1187–1196.

18. Chang, L., and Tsao, D.Y. (2017). The Code for Facial Identity in the Primate Brain. Cell *169*, 1013-1028.e14.

19. Bao, P., She, L., McGill, M., and Tsao, D.Y. (2020). A map of object space in primate inferotemporal cortex. Nature *583*, 103–108.

20. Khuvis, S., Yeagle, E.M., Norman, Y., Grossman, S., Malach, R., and Mehta, A.D. (2021). Face-Selective Units in Human Ventral Temporal Cortex Reactivate during Free Recall. J. Neurosci. *41*, 3386–3399.

21. Gelbard-Sagiv, H., Mukamel, R., Harel, M., Malach, R., and Fried, I. (2008). Internally generated reactivation of single neurons in human hippocampus during free recall. Science *322*, 96–101.

22. Scoville, W.B., and Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. J. Neurol. Neurosurg. Psychiatry *20*, 11–21.

23. Miyashita, Y., and Chang, H.S. (1988). Neuronal correlate of pictorial short-term memory in the primate temporal cortexYasushi Miyashita. Nature *331*, 68–70.

24. Hanes, D.P., Thompson, K.G., and Schall, J.D. (1995). Relationship of presaccadic activity in frontal eye field and supplementary eye field to saccade initiation in macaque: Poisson spike train analysis. Exp. Brain Res. *103*, 85–96.

25. Farah, M.J., Levine, D.N., and Calvanio, R. (1988). A case study of mental imagery deficit. Brain Cogn. *8*, 147–164.

26. Jankowiak, J., Kinsbourne, M., Shalev, R.S., and Bachman, D.L. (1992). Preserved visual imagery and categorization in a case of associative visual agnosia. J. Cogn. Neurosci. *4*, 119–131.

27. Behrmann, M., Winocur, G., and Moscovitch, M. (1992). Dissociation between mental imagery and object recognition in a brain-damaged patient. Nature *359*, 636–637.

28. Galton, F. (1880). Visualised Numerals. Nature Publishing Group UK. 10.1038/021252a0.

29. Zeman, A., Dewar, M., and Della Sala, S. (2015). Lives without imagery - Congenital aphantasia. Cortex *73*, 378–380.

30. Cui, X., Jeter, C.B., Yang, D., Montague, P.R., and Eagleman, D.M. (2007). Vividness of mental imagery: individual variability can be measured objectively. Vision Res. *47*, 474–478.

31. Fulford, J., Milton, F., Salas, D., Smith, A., Simler, A., Winlove, C., and Zeman, A. (2018). The neural correlates of visual imagery vividness – An fMRI study and literature review. Cortex *105*, 26–40.

32. Kosslyn, S.M. (1995). Image and brain: the resolution of the imagery debate. J. Cogn. Neurosci. *7*, 415–420.

33. Pearson, J., and Kosslyn, S.M. (2015). The heterogeneity of mental representation: Ending the imagery debate. Proc. Natl. Acad. Sci. U. S. A. *112*, 10089–10092.

34. Marks, D.F. (1973). Vividness of Visual Imagery Questionnaire. British Journal of PsychologyJournal of Mental Imagery. 10.1037/t05959-000.

35. Yomogida, Y., Sugiura, M., Watanabe, J., Akitsuki, Y., Sassa, Y., Sato, T., Matsue, Y., and Kawashima, R. (2004). Mental visual synthesis is originated in the fronto-temporal network of the left hemisphere. Cereb. Cortex *14*, 1376–1383.

36. Liu, J., Spagna, A., and Bartolomeo, P. (2022). Hemispheric asymmetries in visual mental imagery. Brain Struct. Funct. *227*, 697–708.

37. Riddoch, M.J. (1990). Loss of visual imagery: A generation deficit. Cogn. Neuropsychol. *7*, 249–273.

38. Lake, B.M., Salakhutdinov, R., and Tenenbaum, J.B. (2015). Human-level concept learning through probabilistic program induction. Science *350*, 1332–1338.

39. O'Craven, K.M., and Kanwisher, N. (2000). Mental imagery of faces and places activates corresponding stimulus-specific brain regions. J. Cogn. Neurosci. *12*, 1013–1023.

40. Rutishauser, U., Schuman, E.M., and Mamelak, A.N. (2006). Online detection and sorting of extracellularly recorded action potentials in human medial temporal lobe recordings, in vivo. J. Neurosci. Methods *154*, 204–224.

41. Reuter, M., Rosas, H.D., and Fischl, B. (2010). Highly accurate inverse consistent registration: a robust approach. Neuroimage *53*, 1181–1196.

42. Pauli, W.M., Nili, A.N., and Tyszka, J.M. (2018). A high-resolution probabilistic in vivo atlas of human subcortical brain nuclei. Sci Data *5*, 180063.

43. Avants, B., Duda, J.T., Kim, J., Zhang, H., Pluta, J., Gee, J.C., and Whyte, J. (2008). Multivariate analysis of structural and diffusion imaging in traumatic brain injury. Acad. Radiol. *15*, 1360–1375.

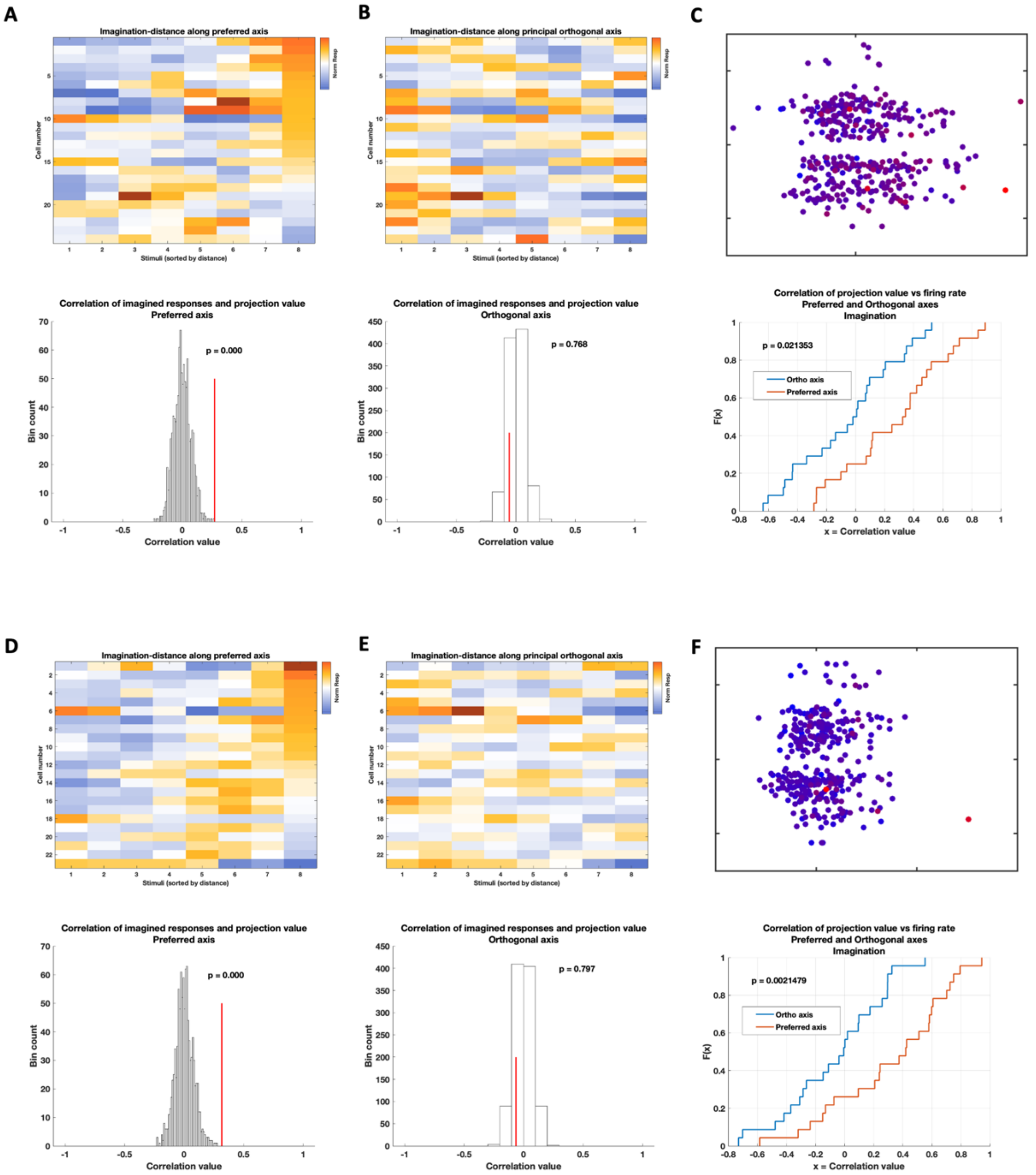
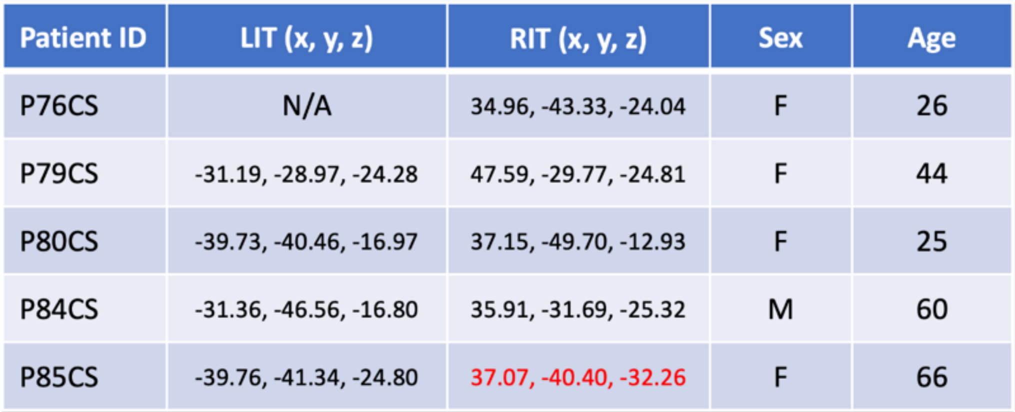


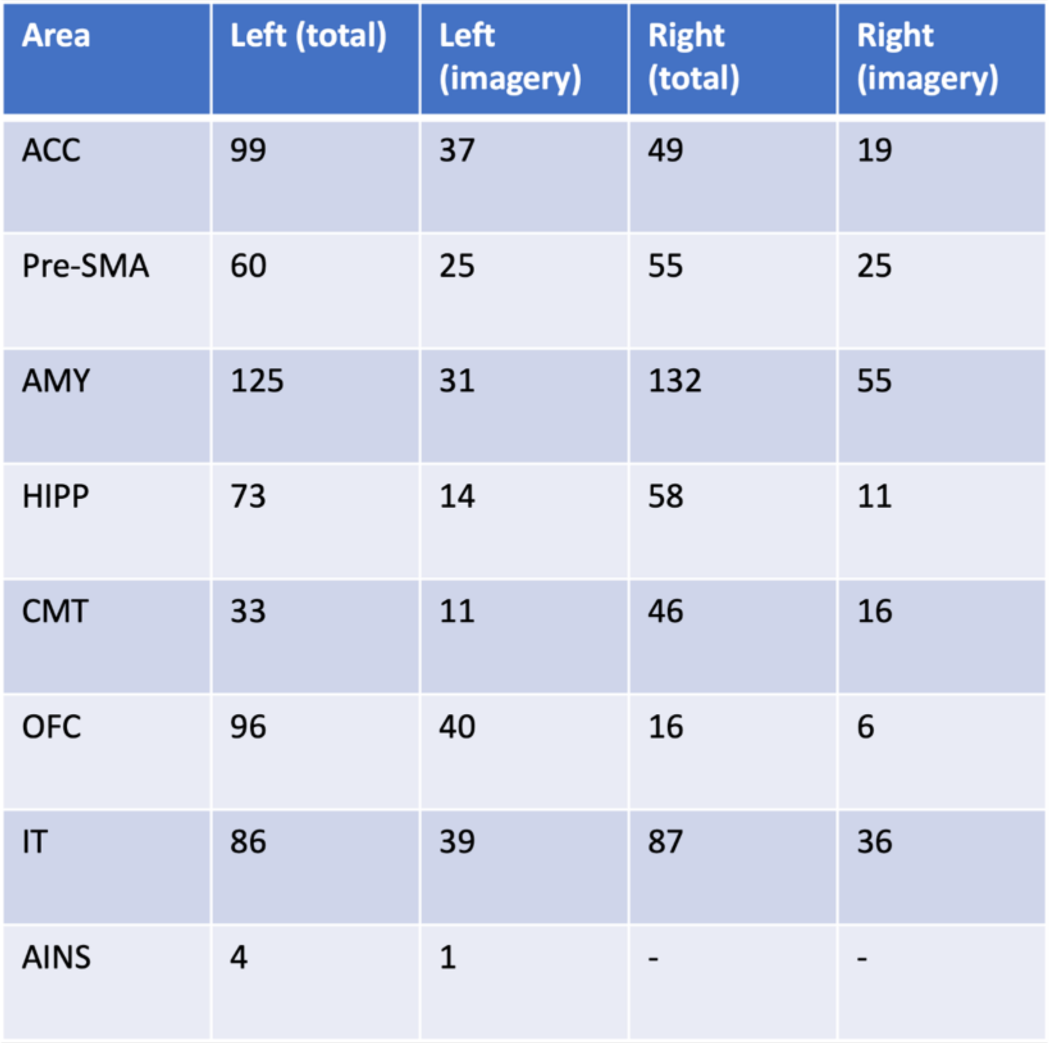
Figure 4.S1. Recapitulation of visual code occurs independently in both hemispheres
(A, B, C) Results showing that axis tuning is respected during imagination for all left IT reactivated neurons.
(D, E, F) Same as A, B, C but for all right IT reactivated neurons.

(A) (Top) Heat plot showing the normalized response of neurons as a function of increasing projection value along the preferred axis (Bottom) The correlation between projection value on the preferred axis and firing rate as compared to a shuffled distribution.
(B) Same as (A). But along the orthogonal axis.
(C) (Top) Plot showing the overall spread of imagined stimuli along the preferred and orthogonal axes for all left IT neurons. (Bottom) CDFs of projection value and firing rate for both axes showing a significant right shift for the CDF of the preferred axis
(D) Same as A for reactivated right IT neurons.
(E) Same as B for reactivated right IT neurons
(F) Same as C for reactivated right IT neurons



Patient ID	LIT (x, y, z)	RIT (x, y, z)	Sex	Age
P76CS	N/A	34.96, -43.33, -24.04	F	26
P79CS	-31.19, -28.97, -24.28	47.59, -29.77, -24.81	F	44
P80CS	-39.73, -40.46, -16.97	37.15, -49.70, -12.93	F	25
P84CS	-31.36, -46.56, -16.80	35.91, -31.69, -25.32	M	60
P85CS	-39.76, -41.34, -24.80	37.07, -40.40, -32.26	F	66

Table 4.1. MNI coordinates of microwire bundles and associated patient metadata. Related to Figure 4.1.
The first patient tested only got right posterior temporal electrodes and is thus labeled ‘N/A’ on the left. The coordinate labelled in red is one where we got 0 responsive units.



Area	Left (total)	Left (imagery)	Right (total)	Right (imagery)
ACC	99	37	49	19
Pre-SMA	60	25	55	25
AMY	125	31	132	55
HIPP	73	14	58	11
CMT	33	11	46	16
OFC	96	40	16	6
IT	86	39	87	36
AINS	4	1	-	-

Table 4.2. Distribution of recorded and reactivated neurons in each area. Related to Figure 4.5. We see robust reactivation of neurons in all recorded brain areas, with even distribution across left and right IT.

Chapter V

CONCLUSIONS AND FUTURE DIRECTIONS:

Summary of main findings:

In this thesis, we present the first detailed exploration of IT cortex at single neuron resolution in the human brain. We build on earlier studies that demonstrate robust responses to visual objects in IT cortex[1–3] and reactivation of those neurons[1] during imagery, sampling a large number of neurons in IT (431) and systematically comparing their responses during the viewing of a large number of objects with varied features. We find that human IT neurons encode specific axes in a high dimensional feature space, similar to macaques[4,5]. We discuss the importance of advances in deep learning for object classification in building said high dimensional spaces — allowing for the parametrization of arbitrary objects. We lay out several experiments and analyses that verify the axis model. We show that tuning along the orthogonal axis in feature space is flat. We use the axis model along with a pre-trained GAN[6] to 1) generate stimuli in the space spanned by the encoded and orthogonal axis and show that tuning only changes in directions parallel to the preferred axis and 2) generate super stimuli and drive the neurons higher than any of the stimulus images. We also demonstrate that while parametrization of the visual objects is essential for the axis model to work, the specific deep convolutional network used to parametrize the images does not affect the results.

Finally, armed with these feedforward coding principles for general objects we investigated single neuron mechanisms of visual imagery. Once again, we recorded a large number of IT neurons (173), verified the axis coding model in these neurons during viewing, and immediately after recorded responses as a subset of those objects (6-8) were imagined. We find robust and structured reactivation of IT neurons during imagery. A substantial minority of recorded IT neurons were active during imagery (43%) with an even larger minority of axis tuned IT neurons reactivating (48%). We find that the responses to different objects recapitulate the visual responses in axis tuned neurons, with the correlation of imagined firing rate and projection value being significant along the encoded axes and insignificant along the orthogonal axes. We use this recapitulation of visual context to decode object features in the midst of several distractors using the same methods as in viewing, and even attempt to ‘reconstruct’ imagined objects by comparing decoded object features to the best possible reconstructions using a large image set (~18k images) that was not used in any experiments. We show that the imagined firing rates capture fine feature details of the objects in question. Finally we note that while previous imaging studies would suggest laterality in the temporal lobe[7–10] and reactivation of a specific fronto-temporal network during imagery[11] we observed no laterality effects in IT and robust activity of at least some neurons in every single recorded region — underlining the importance for single neuron investigations of complex phenomena.

Future directions:

Delving further into single neuron mechanisms of imagery:

Connectivity analyses:

This thesis leads to several follow-up questions, a few of which are discussed in the next section. However, one obvious one is as follows: If one sees robust reactivation of neurons all across the brain during imagery where does the signal originate from? What is the source of the top-down trigger that causes other areas (particularly IT) to be active during imagery? How does the activity across these regions (or a subset of them) support imagery?

Coordinated neural activity across spatially disparate regions plays a vital role in supporting the complex behaviors we seek to understand. There are many techniques to characterize such coupling such as cross-correlation, cross-frequency coupling, or coherence[12].

Cross-correlation

Examining correlations between neurons during a behavior is the simplest way to try and understand how that population of neurons comes together to support that behavior (in our case, imagery). Correlations can also provide information about the functional architecture of a given circuit such as the retina[13], or between cortical areas[14] as experimental evidence accumulated over many years finds that neurons close to each other with similar tuning have higher correlations[15–18] implying that correlations reflect co-fluctuations in the responses of restricted subsets of neurons rather than global fluctuations affecting all neurons[19].

We will examine the cross correlation between neurons in IT cortex during both viewing and imagery to gain a better understanding of how regions are communicating with each other during these behaviors.

Spike field coherence

Coherence is a measure of the phase relationship between two oscillatory signals[12,20] but can also be used to assess the relationship between an oscillatory signal (typically LFP) and neural spike trains[21]. It allows researchers to investigate the enhanced synchronization of neuronal groups activated by a given stimulus or behavior in addition to the increase in neuronal firing rate to the same[22]. Such enhanced synchronization of specific neuronal groups is thought to increase the output onto their postsynaptic target neurons and impose a structured activation pattern[23] aiding in multi-scale communication between areas[24–26]. The added dimension investigation of a temporal code provides has shed light on many important phenomena in the past such as the effects of selective attention[22,27], memory[28], or even the precise coding of position via theta phase precession in the rodent navigation system[29–33].

We will attempt to leverage the temporal as well as rate codes of neurons in IT to understand the origin of the top down signal that triggers reactivation by using pairwise phase consistency, which is a circular statistic similar to SFC but not biased by exact number of trials in a condition or the exact number of spikes of a neuron[34].

Modification of sensory representations by context: effects of familiarity

In a natural environment context guides the use of sensory representations for appropriate action. For example, how to react to someone walking up to you in a restaurant will be affected by whether they are familiar or not, implying that the visual information needs to interact with the contextual variable of familiarity. Behavioral studies have shown that while people are much better at recognizing familiar faces in a blurry or cluttered context than unfamiliar ones[35] the placement of familiar faces in unexpected contexts makes them more difficult to recognize[36].

This begs the question, does the contextual variable affect behavior by directly altering the visual representation or is the interaction of the sensory and contextual variables handled elsewhere? Patient H.M. famously could not form new memories but retained all his old ones after having both hippocampi removed[37] leading to a long-standing hypothesis in the field that long term memories are stored in cortical neurons. Armed with the code for visual objects we will attempt to uncover the network level code for visual memory by examining interaction of memory signals with the axis code in both viewing and imagery.

Task overview

Patients are interviewed about their interests in order to identify people (mostly celebrities) that are familiar to them. After said interview, a set of 100 familiar faces is compiled and added to the previously used screening set of 500 objects. An fMRI localizer is used to determine the location of face selective patches in IT cortex of each patient (Figure 5.1A). The screening task described in Chapter III is conducted as usual, once axis tuned neurons have been found in a given session we will carry out the cued imagery task described in Chapter IV using a subset of familiar faces and low level feature matched unfamiliar faces.

In macaques, neurons in the anterior medial and perirhinal face patches represent familiar faces in a distinct long-latency subspace that is rotated relative to the short-latency subspace used to represent the physical, context-free structure of a face. We will observe the response dynamics in human fusiform face area neurons during viewing to see if familiarity affects the IT axis code in the same way as macaques. After developing a detailed understanding of how the visual code is affected we will examine responses during imagery in an attempt to corroborate or disprove the hypothesis that the imagined responses will load onto the familiar subspace at shorter latency, i.e. in an opposite order to viewing.

Preliminary results

We report that our targeting procedure was successful, yielding 33 face selective neurons out of the 52 total IT neurons recorded (~63%) from 3 sessions with the first patient (Figure 5.1).

We find that face-selective neurons in the FFA are strongly modulated by familiarity in both directions, with some neurons showing a marked increase in firing rate to familiar faces (Figure 5.2A left vs right) while others show a substantially lower response to familiar faces (Figure 5.2A left vs right). We also see canonical face-selective neurons that are agnostic to whether a face was familiar or not.

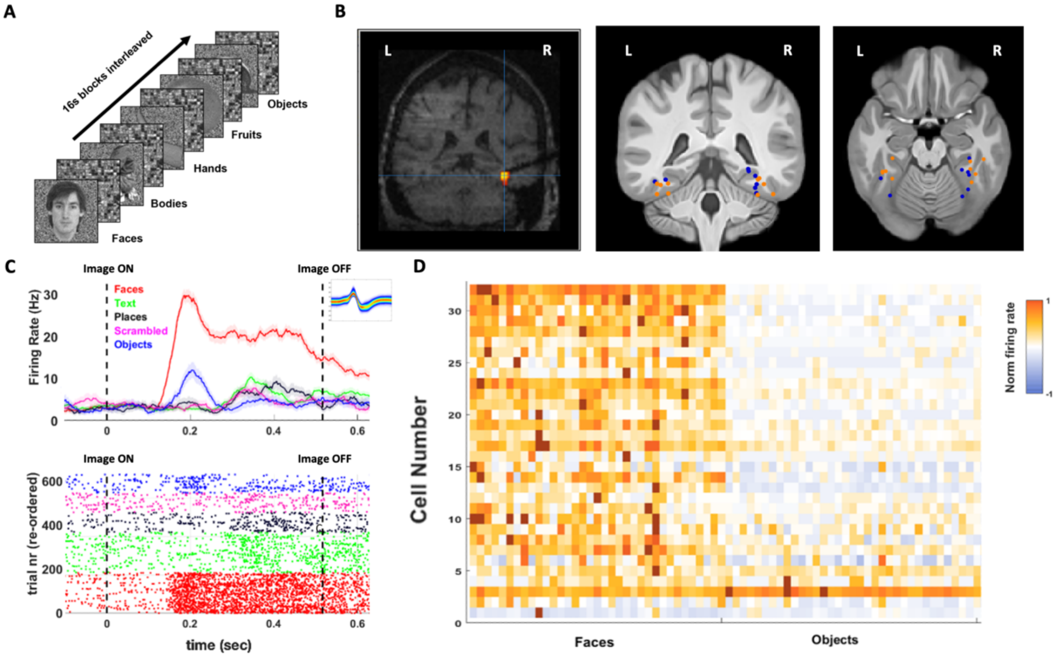


Figure 5.1. Face localizer used to target FFA.

(A) Schematic of the block design face localizer used to target fusiform face area (FFA). Blocks of face, body, hand, fruit, and object images were interleaved with blocks of scrambled images for a total of 32 blocks per round.

(B) (left) Coronal section of MRI showing FFA targeting in an example patient. (middle, right) Recording locations of the microwire bundles that yielded at least 1 responsive unit. The locations that were fMRI targeted are colored orange.

(C) Example face selective unit from the same example patient referred to in (B).

(D) Population summary of all the neurons recorded in a single session from the example patient. The neuron in (C) is the top row in this panel.

Discussion

Previous accounts of IT neuron responses in familiarity vary. A previous study with a single neuron[2] shows a decreased response to familiar faces, while a comprehensive macaque study investigating the interaction of familiarity with the axis code[38] finds that familiar faces are represented in a different *long-latency* subspace that is rotated with respect to familiar faces. It remains to be seen whether the same holds for human IT neurons, however the clear modulation of IT neuron responses to familiarity lends credence to the notion that familiarity alters the visual code directly.

An important caveat to these results is that this particular patient is on the spectrum. Expanding the dataset will be necessary before general claims of face memory in humans can be made.

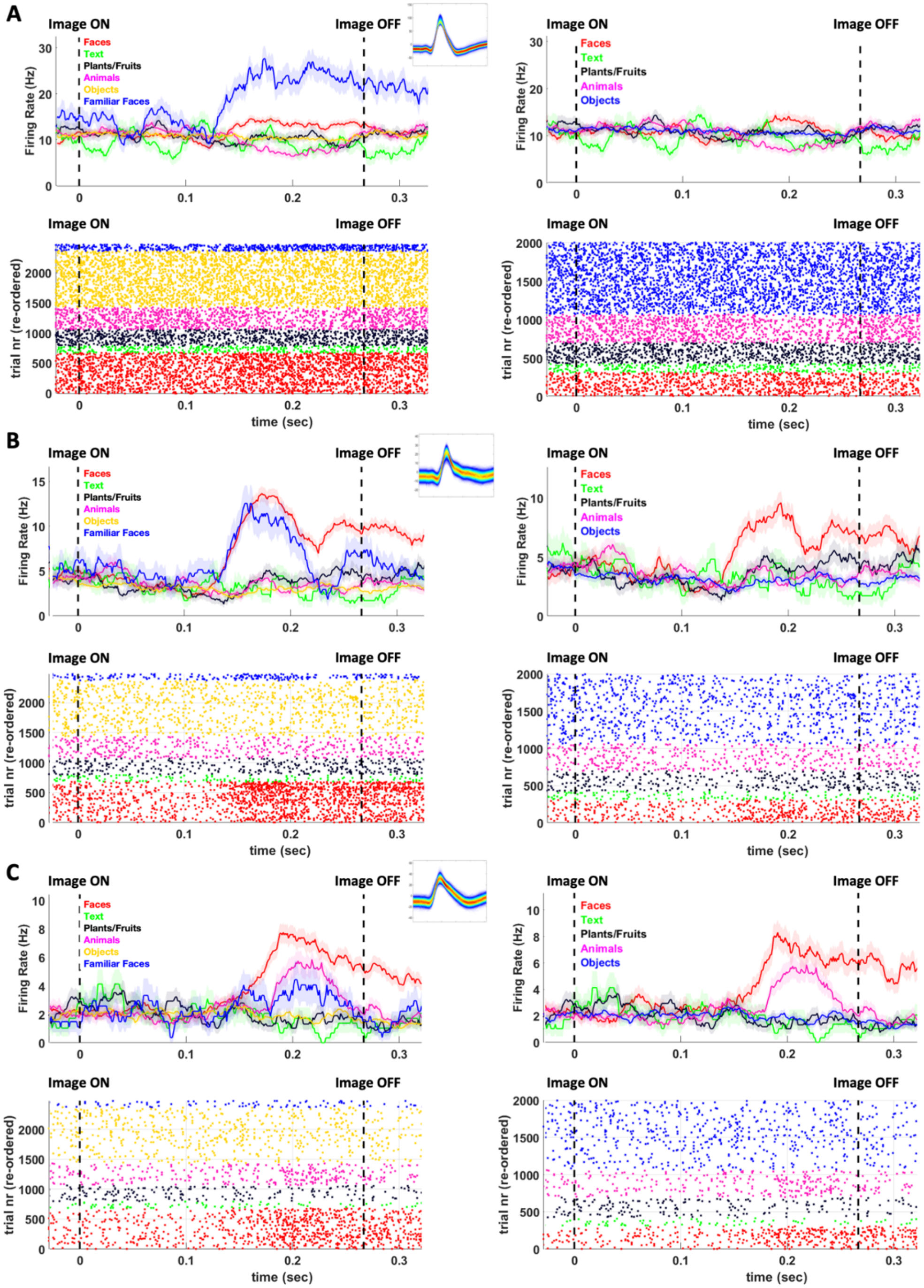


Figure 5.2. FFA neurons show responses modulated by familiarity.

(A) Example face-selective neuron in the FFA showing a markedly increased response to familiar faces. (left) The PSTH of the neuron with familiar faces included. (right) PSTH of the same neuron without familiar faces included.

(B) Example face-selective neuron in the FFA of the same patient that was agnostic to whether a face was familiar or not.

(C) Example face-selective neuron in the FFA that showed a reduced response to familiar faces.

Delving into other aspects of model formation:

As referenced in Chapter I of this thesis, forming a model involves three essential steps: parsing the world into its component parts, i.e. the objects and agents that make up the world, understanding all possible relationships between those parts — called intuitive physics and intuitive psychology in cognitive science[39] — and finally learning and using language to learn everything else that does not have an obvious physical manifestation[40,41].

Thus, understanding how we model the world necessarily requires an understanding of language encoding. Language allows us to sculpt the minds of others and our own into arbitrary (often task relevant) configurations without trial-and-error learning. This is often used to distinguish between animals that are *trainable* and humans that are *programmable*[41]. This jump from trainable to programmable constitutes a large leap in the way we learn, interact, and transmit knowledge. It is important to note here that one does not need language to learn certain complex concepts (for example, justice or fairness) though communicating them becomes almost infinitely harder. As such we are now attempting to understand language encoding in the human brain. Do we find single neurons responses to abstract concepts like justice? Do neurons in the brain track the semantic peaks or the main point in a continuous narrative?

Task overview

6 Patients listened to 6 stories 6 times each with a free recall component in between (after 3 listens of each story) and at the end (Figure 5.3A). The stories used were from the StoryCrop database (https://storycorps.org/) and were chosen to be short (~30s) with only one major event being described in each story.

Preliminary results

We recorded 430 individual neurons across all brain areas (Table 5.1) We find 3 broad categories of responses in the human brain. The first category is neurons that represent auditory boundaries in the story for example, when a new person starts speaking, when the audio switches from conversation to laughter, or even the beginning and end of the story. The second category is neurons that encode the 'semantic peak' of the story. As mentioned earlier these stories were chosen to be short and communicate a single event or answer a single question. We find neurons that respond in a time-locked manner when the main point has been communicated. The final category are neurons that respond to individual words and phrases mainly in the memory centers of the brain (medial temporal lobe) potentially implying that those words and phrases are triggering imagery of concepts in the brain — something that is essential to a lot of learning and communication[41,42]. Examples of the latter 2 types can be seen in Figure 5.3B&C.

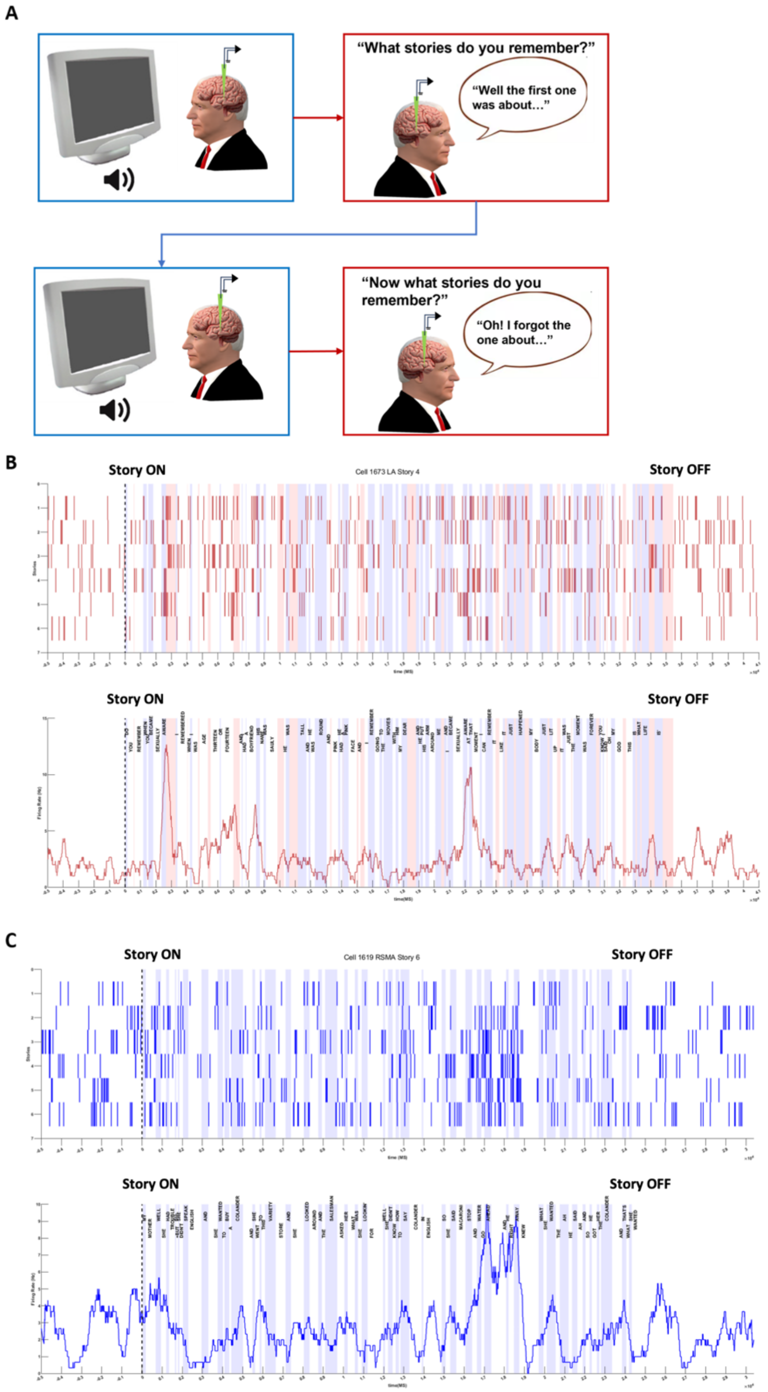


Figure 5.3. Individual neurons record salient timepoints during story listening.

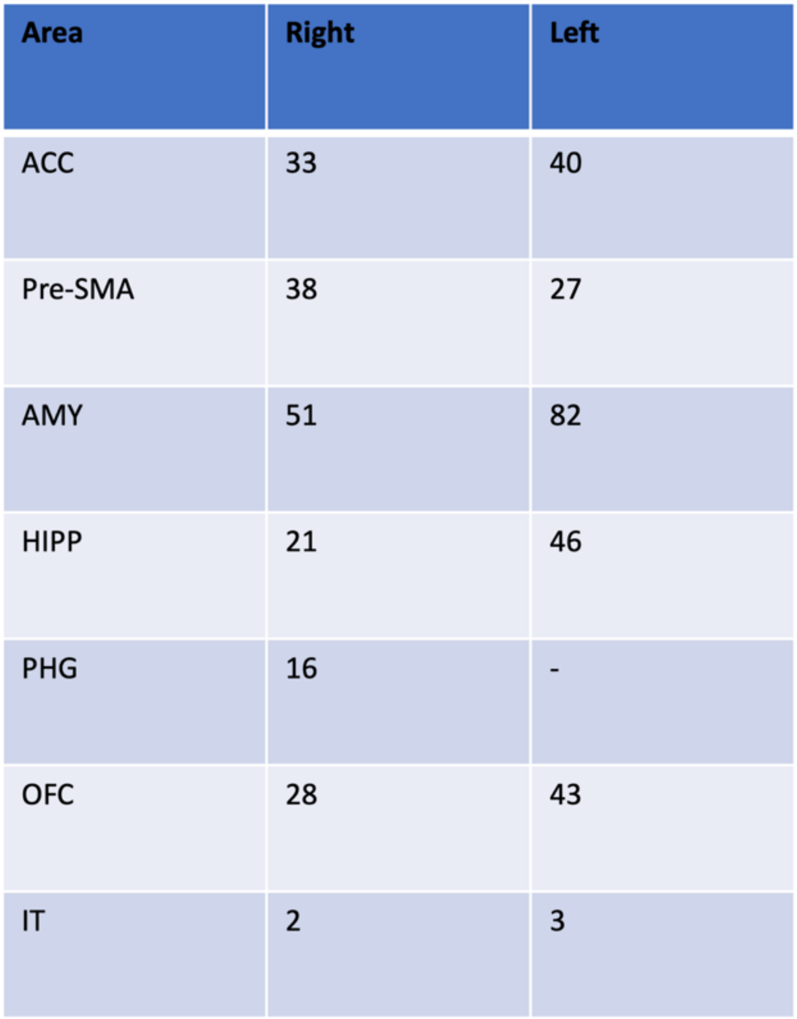
(A) Schematic of the task. Patients listened to 7 stories 3 times each before performing free recall. This structure was repeated twice.

(B) Neuron in the left amygdala of an example patient responding in a time-locked manner to the phrase 'sexually aware'. The transcript is overlayed on the PSTH. The words arranged vertically to save space, with each alternating white/blue bar indicating how long that particular word lasted. Red bars indicate pauses in speech. The phrase occurs twice in the story and neuron responds to both implying a response to the phrase as opposed to simply a saliency effect.

(C) Neuron in the medial frontal cortex (pre-SMA) of an example patient showing a strong ERP to the 'ah-ha' moment in a story. The story describes someone attempting to solve a communication problem and the ERP is time-locked to the moment they are successful.

Discussion

The discovery of neurons that are parsing various aspects of continuous narratives with complex concepts discussed in them was very encouraging. It implies that a model of semantic encoding can be reached by detailed investigation of how neural activity maps onto semantic spaces. This is a direction we intend to pursue further.



Area	Right	Left
ACC	33	40
Pre-SMA	38	27
AMY	51	82
HIPP	21	46
PHG	16	-
OFC	28	43
IT	2	3

Table 5.1. Distribution of recorded neurons in each area during the stories task. Related to Figure 5.3.

References:

1. Khuvis, S., Yeagle, E.M., Norman, Y., Grossman, S., Malach, R., and Mehta, A.D. (2021). Face-Selective Units in Human Ventral Temporal Cortex Reactivate during Free Recall. J. Neurosci. *41*, 3386–3399.

2. Axelrod, V., Rozier, C., Malkinson, T.S., Lehongre, K., Adam, C., Lambrecq, V., Navarro, V., and Naccache, L. (2019). Face-selective neurons in the vicinity of the human fusiform face area. Neurology *92*, 197–198.

3. Axelrod, V., Rozier, C., Malkinson, T.S., Lehongre, K., Adam, C., Lambrecq, V., Navarro, V., and Naccache, L. (2022). Face-selective multi-unit activity in the proximity of the FFA modulated by facial expression stimuli. Neuropsychologia *170*, 108228.

4. Chang, L., and Tsao, D.Y. (2017). The Code for Facial Identity in the Primate Brain. Cell *169*, 1013-1028.e14.

5. Bao, P., She, L., McGill, M., and Tsao, D.Y. (2020). A map of object space in primate inferotemporal cortex. Nature *583*, 103–108.

6. Dosovitskiy, A., and Brox, T. (2016). Generating images with perceptual similarity metrics based on deep networks. Adv. Neural Inf. Process. Syst. *29*.

7. Liu, J., Spagna, A., and Bartolomeo, P. (2022). Hemispheric asymmetries in visual mental imagery. Brain Struct. Funct. *227*, 697–708.

8. Yomogida, Y., Sugiura, M., Watanabe, J., Akitsuki, Y., Sassa, Y., Sato, T., Matsue, Y., and Kawashima, R. (2004). Mental visual synthesis is originated in the fronto-temporal network of the left hemisphere. Cereb. Cortex *14*, 1376–1383.

9. Ishai, A., Ungerleider, L.G., and Haxby, J.V. (2000). Distributed neural systems for the generation of visual images. Neuron *28*, 979–990.

10. Spagna, A., Hajhajate, D., Liu, J., and Bartolomeo, P. (2021). Visual mental imagery engages the left fusiform gyrus, but not the early visual cortex: A meta-analysis of neuroimaging evidence. Neurosci. Biobehav. Rev. *122*, 201–217.

11. Pearson, J. (2019). The human imagination: the cognitive neuroscience of visual mental imagery. Nat. Rev. Neurosci. *20*, 624–634.

12. Lepage, K.Q., Kramer, M.A., and Eden, U.T. (2011). The dependence of spike field coherence on expected intensity. Neural Comput. *23*, 2209–2241.

13. Greschner, M., Shlens, J., Bakolitsa, C., Field, G.D., Gauthier, J.L., Jepson, L.H., Sher, A., Litke, A.M., and Chichilnisky, E.J. (2011). Correlated firing among major ganglion cell types in primate retina. J. Physiol. *589*, 75–86.

14. Alonso, J.M., and Martinez, L.M. (1998). Functional connectivity between simple cells and complex cells in cat striate cortex. Nat. Neurosci. *1*, 395–403.

15. Zohary, E., Shadlen, M.N., and Newsome, W.T. (1994). Correlated neuronal discharge rate and its implications for psychophysical performance. Nature *370*, 140–143.

16. Cohen, M.R., and Maunsell, J.H.R. (2009). Attention improves performance primarily by reducing interneuronal correlations. Nat. Neurosci. *12*, 1594–1600.

17. Lee, D., Port, N.L., Kruse, W., and Georgopoulos, A.P. (1998). Variability and correlated noise in the discharge of neurons in motor and parietal areas of the primate cortex. J. Neurosci. *18*, 1161–1170.

18. Smith, M.A., and Kohn, A. (2008). Spatial and temporal scales of neuronal correlation in primary visual cortex. J. Neurosci. *28*, 12591–12603.

19. Cohen, M.R., and Kohn, A. (2011). Measuring and interpreting neuronal correlations. Nat. Neurosci. *14*, 811–819.

20. Montgomery, S.M., and Buzsáki, G. (2007). Gamma oscillations dynamically couple hippocampal CA3 and CA1 regions during memory task performance. Proc. Natl. Acad. Sci. U. S. A. *104*, 14495–14500.

21. Jarvis, M.R., and Mitra, P.P. (2001). Sampling properties of the spectrum and coherency of sequences of action potentials. Neural Comput. *13*, 717–749.

22. Fries, P., Womelsdorf, T., Oostenveld, R., and Desimone, R. (2008). The effects of visual stimulation and selective visual attention on rhythmic neuronal synchronization in macaque area V4. J. Neurosci. *28*, 4823–4835.

23. Salinas, E., and Sejnowski, T.J. (2001). Correlated neuronal activity and the flow of neural information. Nat. Rev. Neurosci. *2*, 539–550.

24. Bastos, A.M., Usrey, W.M., Adams, R.A., Mangun, G.R., Fries, P., and Friston, K.J. (2012). Canonical microcircuits for predictive coding. Neuron *76*, 695–711.

25. Buzsáki, G., Logothetis, N., and Singer, W. (2013). Scaling brain size, keeping timing: evolutionary preservation of brain rhythms. Neuron *80*, 751–764.

26. Anastassiou, C.A., and Koch, C. (2015). Ephaptic coupling to endogenous electric field activity: why bother? Curr. Opin. Neurobiol. *31*, 95–103.

27. Fries, P., Reynolds, J.H., Rorie, A.E., and Desimone, R. (2001). Modulation of oscillatory neuronal synchronization by selective visual attention. Science *291*, 1560–1563.

28. Pesaran, B., Pezaris, J.S., Sahani, M., Mitra, P.P., and Andersen, R.A. (2002). Temporal structure in neuronal activity during working memory in macaque parietal cortex. Nat. Neurosci. *5*, 805–811.

29. O'Keefe, J., and Recce, M.L. (1993). Phase relationship between hippocampal place units and the EEG theta rhythm. Hippocampus *3*, 317–330.

30. O'Keefe, J., and Burgess, N. (1996). Geometric determinants of the place fields of hippocampal neurons. Nature *381*, 425–428.

31. Buzsaki, G. (2006). Rhythms of the Brain (Oxford University Press).

32. Hafting, T., Fyhn, M., Bonnevie, T., Moser, M.-B., and Moser, E.I. (2008). Hippocampus-independent phase precession in entorhinal grid cells. Nature *453*, 1248–1252.

33. Kovács, K.A. (2020). Episodic Memories: How do the Hippocampus and the Entorhinal Ring Attractors Cooperate to Create Them? Front. Syst. Neurosci. *14*, 559168.

34. Vinck, M., van Wingerden, M., Womelsdorf, T., Fries, P., and Pennartz, C.M.A. (2010). The pairwise phase consistency: a bias-free measure of rhythmic neuronal synchronization. Neuroimage *51*, 112–122.

35. Burton, A.M., Wilson, S., Cowan, M., and Bruce, V. (1999). Face Recognition in Poor-Quality Video: Evidence From Security Surveillance. Psychol. Sci. *10*, 243–248.

36. Laurence, S., Eyre, J., and Strathie, A. (2021). Recognising Familiar Faces Out of Context. Perception *50*, 174–177.

37. Scoville, W.B., and Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. J. Neurol. Neurosurg. Psychiatry *20*, 11–21.

38. She, L., Benna, M.K., Shi, Y., Fusi, S., and Tsao, D.Y. (2021). The neural code for face memory. BioRxiv.

39. Lake, B.M., Ullman, T.D., Tenenbaum, J.B., and Gershman, S.J. (2017). Building machines that learn and think like people. Behav. Brain Sci. *40*, e253.

40. Lupyan, G. (2016). The Centrality of Language in Human Cognition. Lang. Learn. *66*, 516–553.

41. Lupyan, G., and Bergen, B. (2016). How Language Programs the Mind. Top. Cogn. Sci. *8*, 408–424.

42. Spelke, E.S. (2003). What makes us smart? Core knowledge and natural language. Language in mind: Advances in the study of.