Population-level neural coding for higher cognition

Xiaoxiong Lin

Graduate School of
Systemic Neurosciences

LMU Munich

Dissertation at the
Graduate School of Systemic Neurosciences
Ludwig-Maximilians-Universität München
April 2023

Supervisor:
Prof. Dr. med. Simon Jacob
Translational NeuroTechnology
Department of Neurosurgery
Technical University of Munich

First Reviewer:	Prof. Dr. med. Simon Jacob
Second Reviewer:	Prof. Dr. Christian Leibold
External Reviewer:	Dr. Torben Ott

Date of Submission: 26 April 2023
Date of Defense: 11 September 2023

To my loving mother ...

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor, Prof. Dr. med. Simon Jacob, for his unwavering guidance, support, and encouragement throughout my PhD journey. His expertise, patience, and optimism have been instrumental in helping me develop and complete this research.

I am also grateful to the members of my thesis committee, Prof. Dr. Christian Leibold and Prof. Dr. Moritz Grosse-Wentrup, for their insightful feedback and constructive suggestions that have contributed significantly to the development of this work.

I would like to extend my appreciation to the Graduate School of Systemic Neuroscience (GSN) for providing essential resources, fostering a supportive research environment, and offering initial financial assistance for my PhD research. Additionally, I am grateful to the GSN administration team, especially Lena Bittl, for their guidance and support during the challenging transition period when I changed research projects.

I would like to extend special thanks to my dear friend and fellow neuroscientist, Weiwei Chen. Her thought-provoking discussions directly influenced the methodology employed in this thesis, and her invaluable input during the editing process greatly enhanced the final work.

On a personal note, I would like to express my gratitude to my colleagues at the Translational NeuroTechnology Laboratory for creating a joyful and inspiring work environment every day.

Abstract

Higher cognition encompasses advanced mental processes that enable complex thinking, decision-making, problem-solving, and abstract reasoning. These functions involve integrating information from multiple sensory modalities and organizing action plans based on the abstraction of past information. The neural activity underlying these functions is often complex, and the contribution of single neurons in supporting population-level representations of cognitive variables is not yet clear.

In this thesis, I investigated the neural mechanisms underlying higher cognition in higher-order brain regions with single-neuron resolution in human and non-human primates performing working memory tasks. I aimed to understand how representations are arranged and how neurons contribute to the population code.

In the first manuscript, I investigated the population-level neural coding for the maintenance of numbers in working memory within the parietal association cortex. By analyzing intra-operative intracranial micro-electrode array recording data, I uncovered distinct representations for numbers in both symbolic and nonsymbolic formats.

In the second manuscript, I delved deeper into the neuronal organizing principles of population coding to address the ongoing debate surrounding memory maintenance mechanisms. I unveiled sparse structures in the neuronal implementation of representations and identified biologically meaningful components that can be directly communicated to downstream neurons. These components were linked to subpopulations of neurons with distinct physiological properties and temporal dynamics, enabling the active maintenance of working memory while resisting distraction. Lastly, using an artificial neural network model, I demonstrated that the sparse implementation of temporally modulated working memory representations is preferred in recurrently connected neural populations such as the prefrontal cortex.

In summary, this thesis provides a comprehensive investigation of higher cognition in higher-order brain regions, focusing on working memory tasks involving numerical stimuli. By examining neural population coding and unveiling sparse structures in the neuronal implementation of representations, our findings contribute to a deeper understanding of the mechanisms underlying working memory and higher cognitive functions.

Table of contents

Chapter 1

Introduction

In his *Dioptrics*, Descartes made astute observations about neural representations. He stated, "not only do the images of objects form on the back of the eye, but they also pass beyond to the brain." This idea was based on the physiological structure of nerves: "these small fibers... do not crowd or impede each other in any way, and are extended from the brain to the extremities of all parts which are capable of any sensation, in such a way that, however slightly we touch and move the spot in these places where any one of the fibers is attached, we also move at the same instant the place in the brain from which it comes" (Fig. 1.1). Descartes thus suggested that the image transmitted to the brain must bear some resemblance to the one on the retina, but not as a direct copy of the retinal image. Instead, it represents various qualities that the object possesses. He further cautioned, "We must not think that it is by means of this resemblance that the picture makes us aware of the objects - as though we had another pair of eyes to see it, inside our brain." (Descartes, 1965).

Despite the insights about visual functions, Descartes quickly overlooked his own warning when he pondered upon high-level cognition and made his famous assertion that the pineal gland connects the soul, and thus, perceives the image projected onto it and ultimately operates the activity of human body.

The challenge of mechanistically understanding high-level brain functions still resonates today. Low-level cognitive functions, such as the early processes of visual perception, have been mechanistically reduced to a series of processing steps that detect hierarchically organized visual features, with each step making only minor transformations to the previous representation (Lindsay, 2021). However, higher up on the functional hierarchy, much remains to be explained about what is represented. In fact, the framing of "representation" may not even be suitable for high-level brain functions, as an intelligent system does not necessarily need representations to perform complex tasks (Hayes, 1981; Brooks, 1991). Fur-

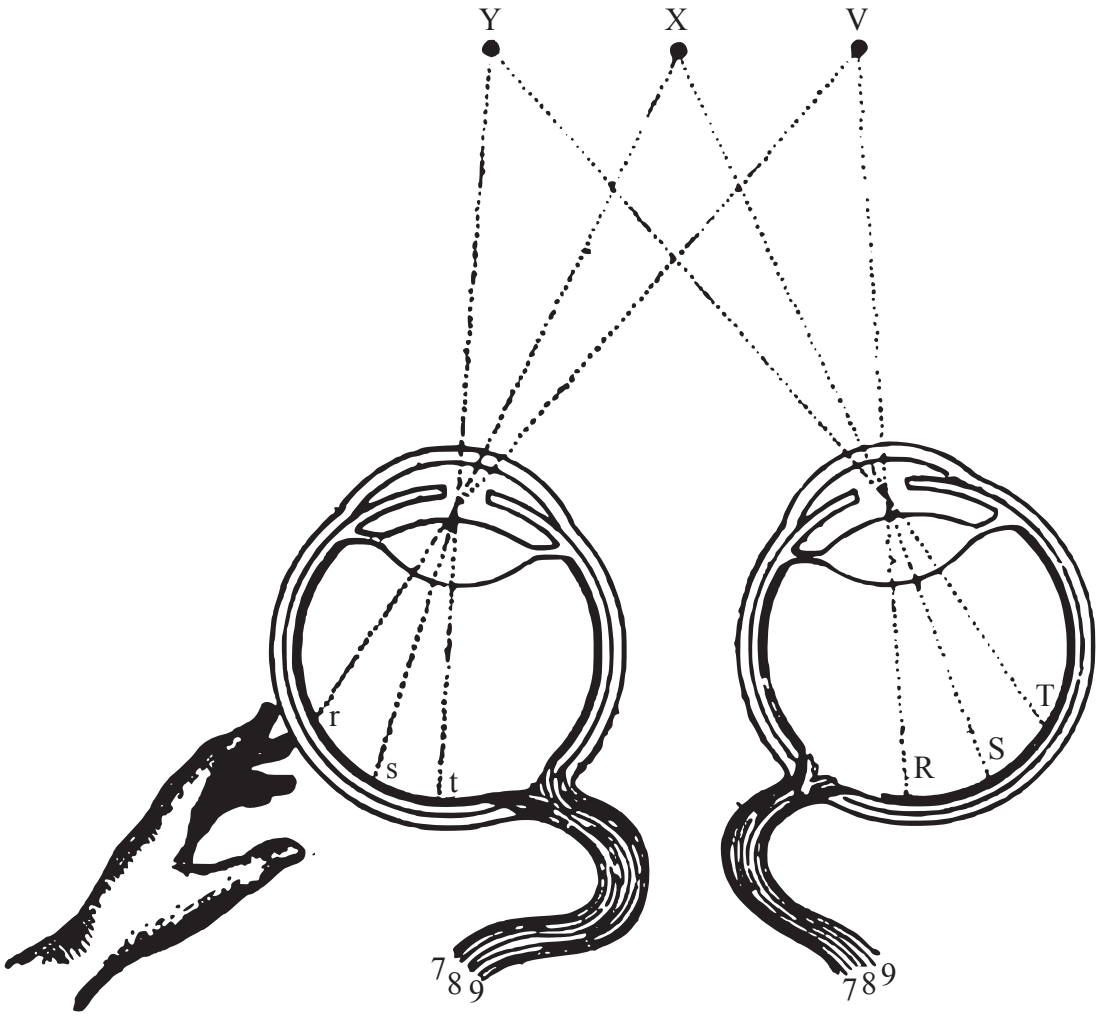


Fig. 1.1 The information at V, X, Y are registered at retinal locations R, S, T that project through nerve fibers 7,8,9 respectively. Figure adapted from Descartes (1965, The fifth discourse)

thermore, although we now know that the pineal gland does not hold the key to coordinating and controlling behavior as once believed, our understanding has only progressed to attributing these functions to certain higher-order cortices, namely the prefrontal cortex (Miller and Cohen, 2001; Quintana and Fuster, 1999; Buschman and Miller, 2022). Trying to explain the mechanism of higher cognition while treating these higher-order cortices as a black box is committing the same *Homunculus Fallacy* as Descartes did, i.e., the circular argument of explaining brain function by postulating a miniature human capable of these complex functions in a subset of the brain (Kenny, 1971). A detailed dissection of neuronal organization and neural dynamics in the higher-order cortices is necessary.

In this introduction, I begin by outlining higher cognition from functional and physiological perspectives. I then explore the neuronal organization underlying higher cognition, examining how external stimuli and internal cognitive variables are represented in local neuronal populations. Finally, I identify the key challenges in the study of higher cognition, which will be addressed in the subsequent results chapter.

1.1 Higher cognition

High-level brain function is often defined in the context of the *Perception-Action Cycle* (Fuster, 1990), as illustrated in Figure 1.2. Organisms continually engage with their environment, obtaining sensory information that is subsequently integrated through a series of processes. Decisions are made, and actions are planned and executed based on this processed information, modifying the environment and producing new sensory input. Although low-level sensory and motor functions directly relate to the observable environment, high-level functions, more distant from the environment, rely on low-level functions. Sensory and motor processes interact at every level. Simple, well-trained behaviors form Perception-Action Cycles via lower-level processes, while complex and novel behaviors necessitate higher-level processes. At the top of this hierarchy, sensory information is integrated across modalities and abstracted to provide behavioral context that is then maintained until a motor plan is formulated. In complex behavior, retrospective maintenance and prospective planning are intertwined, sharing contingencies across multiple time steps (Fuster, 2001).

High-level brain function, therefore, is characterized by the following traits: (1) an extended temporal integration and organization window, not strictly adhering to stimulus occurrence or motor onset (Ehrlich and Murray, 2022); (2) intricate neuronal responses, abstract in nature, resulting from both multi-modal integration and the intermingling of sensory and motor representation, difficult to attribute to a single semantic interpretation (Rigotti et al., 2013; Bernardi et al., 2020); and (3) executive control, involving the selection of appropriate

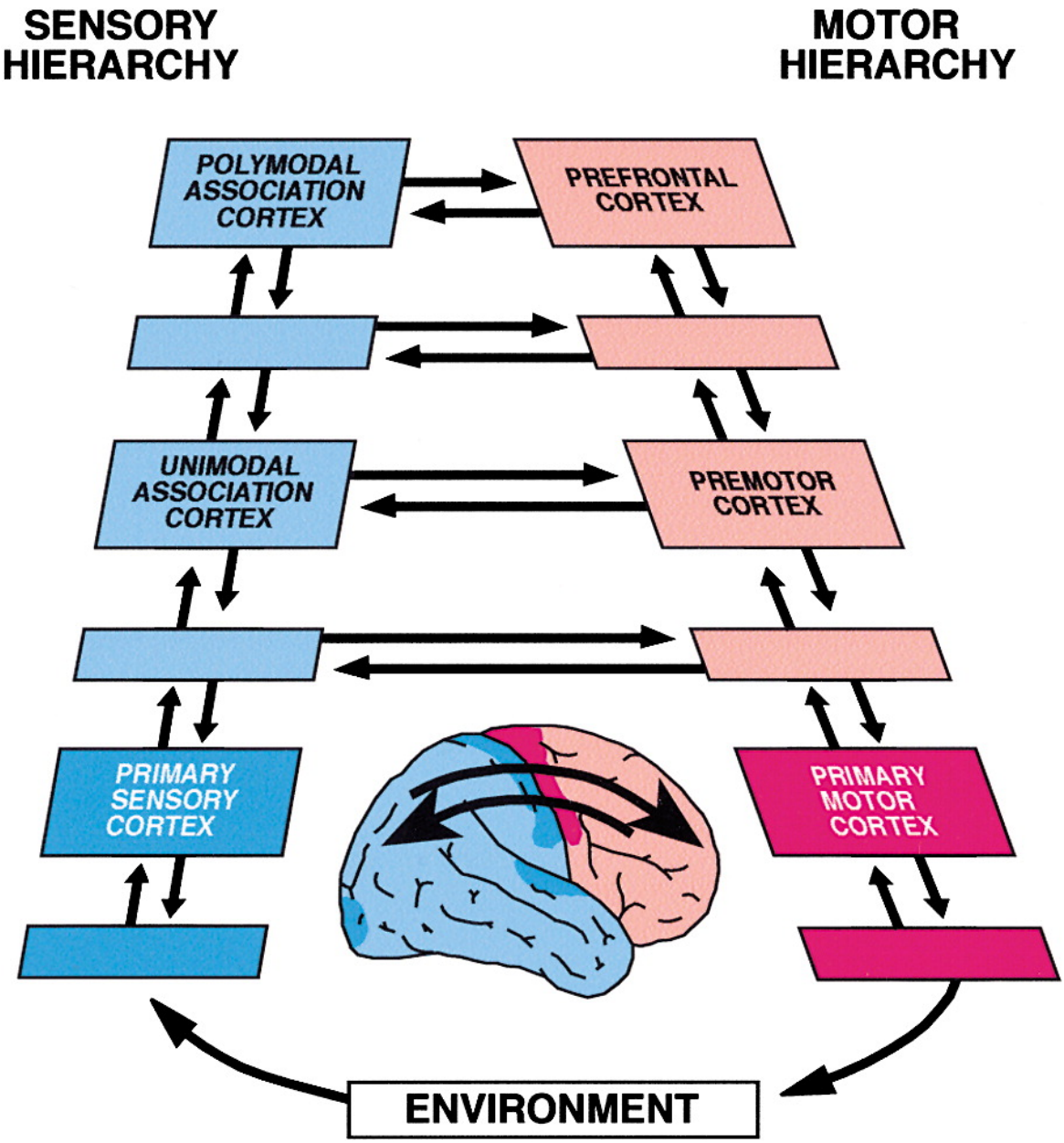


Fig. 1.2 Intermediate areas or subareas of labeled cortex are represented by unlabeled rectangles. All arrows indicate connective pathways identified in monkeys. The human brain image emphasizes reciprocal connectivity between posterior and frontal cortex. Figure adapted from Fuster (2001) with permission.

motor plans and the adaptation of external stimulus representations according to internal objectives (Buschman and Miller, 2022; Cavanagh et al., 2018).

These traits are particularly prominent in working memory. From a cognitive science perspective, working memory updates the concept of short-term memory that can manifest in either phonological or visual-spatial forms, with the impairment of one not influencing the other (Baddeley, 1992). Besides short-term storage, working memory also includes a central executive component to actively maintain and coordinate storage. Working memory is essential for numerous cognitive skills, and performance on working memory tasks is predictive of reading, comprehension, and reasoning abilities (Miller, Lundqvist, and Bastos, 2018; Buschman and Miller, 2022).

1.2 Anatomical bases of higher cognition

The functional hierarchy of the brain is deeply ingrained in its anatomical hierarchy. The ventral stream of the primate visual system offers an example, with visual information captured by the retina, relayed to the thalamus, and subsequently reaching the primary visual cortex (V1). From V1, visual information is transmitted to higher-order visual areas, such as V2, V4, and inferotemporal lobe (IT), which respond to a series of distinct visual features with increasing complexity such as oriented lines, figure ground separation and object identity (Felleman and Van Essen, 1991; Roe et al., 2012; Kravitz et al., 2013).

Although the feedforward structure becomes less clear beyond sensory cortices, the relative position of a cortical area in the hierarchy can still be determined based on the connectome obtained from viral tracing. Using the connecome, the cortex is divided into clusters with stronger within-cluster connections than between them. Then the direction of information flow between clusters can be inferred from the terminal layer, with feedforward connections ending in layer 4 of the target area (driving) and feedback connections ending in other layers of the target area (modulating). Clusters higher in the hierarchy are characterized by sending more feedback projections than feedforward projections, such as prefrontal, premotor, and other association areas (Harris et al., 2019).

Local physiological properties can also delineate the cortical hierarchy in species where viral tracing is not readily available: myelin content density gradually decreases from V1 to the prefrontal cortex; local circuit excitability varies across cortical regions, with higher spine density in higher cortical hierarchy positions, leading to more persistent activity crucial for temporal integration function (Wang, 2020; Murray et al., 2014); the ratio between input-modulating somatostatin-expressing interneurons and output-modulating parvalbumin-expressing interneurons increases with the position in the cortical hierarchy (Wang, 2020);

dopamine D1 receptor density increases along the cortical hierarchy, essential for receiving inhibitory signals to filter out distracting stimuli in working memory tasks (Froudist-Walsh et al., 2021). Numerous other physiological properties also partially correlate with the cortical hierarchy, such as neuron density, pyramidal cell size, myelin content in grey matter, cortical thickness and laminar differentiation (Amunts and Zilles, 2015).

In all the hierarchy-related measures mentioned, the prefrontal cortex (PFC) stands out as the typical high-level cortical area. It provides more feedback, has higher local excitability and has a higher D1 receptor ratio. The PFC, defined as the part of the cerebral cortex that receives projections from the mediodorsal nucleus of the thalamus (Fuster, 2015), serves as a platform where information from diverse brain systems can be integrated and processed within relatively localized circuitry (Miller and Cohen, 2001). The lateral and mid-dorsal PFC receive direct input from an array of secondary (association) cortices. The dorsolateral PFC (particularly brodmann area 46) is interconnected with high-level motor areas, including the supplementary motor area and premotor area, as well as the cerebellum and basal ganglia, which are responsible for automating behavior (Bates and Goldman-Rakic, 1993). The orbital and medial PFC are intricately connected with medial temporal limbic structures crucial for long-term memory, affect, and motivation processing (Barbas and De Olmos, 1990). Furthermore, all regions of the PFC and their subdivisions are strongly interconnected (Miller and Cohen, 2001).

The functional significance of PFC in higher cognition is demonstrated by neuropsychological symptoms in human and non-human primates with lesions. It can be categorized into three PFC clusters: orbital/inferior, medial/cingulate, and lateral regions of the PFC. Orbital PFC lesions often lead to dramatic personality changes, impulsiveness, disinhibition, and attention disorders (MacFall et al., 2001; Izquierdo, Suda, and Murray, 2005). Medial PFC lesions, including the anterior portion of the cingulate gyrus, result in a loss of spontaneity, difficulty initiating movements and speech, apathy, and issues with concentrating attention (Di Pellegrino, Ciaramelli, and Ladavas, 2007; Ostlund and Balleine, 2005). Lastly, lateral PFC damage is characterized by an inability to formulate and execute plans and action sequences, leading to dysexecutive syndrome and a loss of supervisory attentional control (Tanji and Hoshi, 2008). Compared to other clusters, a lesion in lateral PFC is the most detrimental to the higher-order temporal integration functions such as organizing and executing behavior, speech, and reasoning (Fuster, 2001).

The posterior parietal cortex (PPC) also ranks high in the cortical hierarchy as a key component of association cortices. The PPC is anatomically and functionally interconnected with the PFC (Quintana and Fuster, 1999). Instead of being exclusively sensory or motor in nature, the PPC integrates inputs from various brain regions, such as somatosensory, auditory,

visual, motor, cingulate, and prefrontal cortices, while also integrating proprioceptive and vestibular signals from subcortical areas (Whitlock, 2017). Notably, PPC neurons near the intraparietal sulcus respond to numerical quantity regardless of the detailed perceptual features (Nieder, Freedman, and Miller, 2002), demonstrating high-level abstraction ability. The parietal-prefrontal circuit is essential to the executive aspects of the perception-action cycle, responsible for motor planning, decision-making, forward state estimation, and relative-coordinate representations (Andersen and Cui, 2009; Quintana and Fuster, 1999).

1.3 Representations in neuronal populations

As is warned by Kenny, 1971, mere localization does not constitute an explanation for the function. To comprehend the mechanisms of higher cognition, a more detailed examination of neuronal organization, interaction, and computation is required.

It is generally believed that the brain has a certain functional modularity. At a coarse scale, the cortex is divided into sensory, motor and association cortices in the previous sections. At a smaller scale, cortical neurons are organized into columns, which are functionally similar modules of neurons arranged vertically or radially in the cortex, with a horizontal diameter of around 50 μm for minicolumns or 300 μm for macrocolumns. The columnar structure is most extensively studied in sensory cortices, such as the visual cortex, where column positions preserve the topology of the corresponding retinal input (Lund, Angelucci, and Bressloff, 2003; Molnár and Rockland, 2020; Ringach et al., 2016), just as Descartes posited (Fig. 1.1). Columns with similar features, for instance, similar orientations, are arranged adjacently within the cortex (Kremkow et al., 2016). This feature map is also preserved throughout the visual pathway, even in inferior temporal lobe, where the features represent abstract and less intuitive dimensions in visual object space (Bao et al., 2020).

However, columnar structures and functional maps are less well-defined in higher motor and association areas, (Molnár and Rockland, 2020; Constantinidis and Qi, 2018), suggesting a lack of modularity. The cells in higher-order cortices often exhibit complex response properties that simultaneously reflect different cognitive variables, and that are not topologically organized. The response of a PFC neuron, for example, may be correlated with variables of the sensory stimuli, task rule, motor response or any combination of these. This phenomenon, known as *mixed selectivity*, is thought to enable flexible output and serves as a hallmark of the PFC (Rigotti et al., 2013).

Given the absence of a clear relationship between single neurons and task variables, it is crucial to analyze task variable representations in the neuronal population in these higher-order cortices. This is typically achieved by summarizing population activity in population

state space, with modes (latent variables) extracted based on neuronal covariance, forming a latent subspace (Cunningham and Yu, 2014). The latent activity subspace can be constructed to reflect the specific variables in question, such as working memory content (Murray et al., 2017). Instead of finding the latent variables that give rise to the observed population activity, alternatively, decoding approaches try to predict the task variables using the population activity. The generalizability of decoders can be tested across time points or contexts, to investigate the stability of neuronal population's representations of task variables (Parthasarathy et al., 2017; Parthasarathy et al., 2019; Cavanagh et al., 2018; Bernardi et al., 2020).

The population coding in higher-order cortices differs from classical population coding often described in lower-level sensory systems, such as the wind direction coding in crickets, where neurons tuned to two cardinal directions represent wind direction with their vector sum (Dayan and Abbott, 2005). In contrast, higher-order cortices exhibit less clear single-neuron tuning (Rigotti et al., 2013). Their population coding is also not to accurately represent external stimuli, but rather to transform sensory input into suitable motor plans (Ehrlich and Murray, 2022). Therefore, the investigation of higher-order cortices' functions should emphasize the dynamics of neuronal states, rather than their passive representations of task variables. For instance, in working memory, attractor dynamics are often used to explain the mechanism of maintaining memory content during delay periods (Wimmer et al., 2014). This usually manifests as persistent activity in the absence of sensory input either in state space (Murray et al., 2017) or at the single-neuron level (Fuster, 2001). Context-dependent decision-making processes coincide with the convergence of state trajectory to an appropriate feature axis corresponding to the context (Mante et al., 2013). Rotational dynamics have been observed in tasks with serially structured trial stages (Libby and Buschman, 2021) or tasks requiring periodic movement (Michaels, Dann, and Scherberger, 2016).

1.4 Numerical cognition: gateway to complex functions

Numerical cognition, a critical component of higher cognition, exemplifies the integration of sensory information across modalities and the abstraction from tangible object properties (Nieder and Dehaene, 2009). It encompasses three major concepts: numerical quantity, numerical order, and the concept of nominal numbers (e.g., bus number 3) (Wiese, 2003b). These concepts in numerical cognition are often presented as dissociable processes; for instance, in human subjects, quantity judgment between adjacent numbers is slower than for distant numbers, while order judgment between adjacent numbers is faster than for distant numbers (Turconi, Campbell, and Seron, 2006). However, these concepts share common

physiological circuits and processes (Dehaene et al., 2003), and their development for abstract thinking may rely on shared linguistic foundations (Wiese, 2003a).

All these aspects of numerical cognition are fundamental to apprehending the structure of complex tasks and organizing behaviors. Accurate numerical quantity cognition underpins reward estimation (Roitman, Brannon, and Platt, 2007; Cazettes et al., 2023) and the temporal duration of task periods (Meck, Church, and Gibbon, 1985), which could reflect behavioral costs (Masset et al., 2020) or guide motor planning (Niemi and Näätänen, 1981). The order in which a stimulus is presented could determine its behavioral relevance (Jacob and Nieder, 2014; Parthasarathy et al., 2017) and the cued behavioral context (Cavanagh et al., 2018). Nominal number cognition indicates a subject's ability to assign identities to items within a set, forming the basis for associating attributes with these items.

Numerical cognition serves as an excellent springboard for investigating complex higher cognitive functions. It is deeply involved with humans' sophisticated linguistic and logical abilities (Gordon, 2004; Wiese, 2003a) and underpins numerous advancements in human civilizations. Complex arithmetic operations and the recognition of symbolic numbers in humans share evolutionary and physiological origins with non-verbal number cognition (Halberda and Feigenson, 2008), which is present in many species and crucial for their survival (Wilson, Hauser, and Wrangham, 2001; Hauser, Carey, and Hauser, 2000). The neural response for numerical stimuli in humans and animals can be seamlessly connected (Nieder, Freedman, and Miller, 2002; Piazza et al., 2007; Nieder, Wagener, and Rinnert, 2020). This allows us to delve into complex brain functions from a straightforward starting point, utilizing the numerous experimental tools available in animal models.

In comparison to other higher cognitive functions, neuronal representations for numerical stimuli are relatively more straightforward. Individual neurons tuned to specific numerical quantities can be found in PFC and PPC (Nieder and Dehaene, 2009). Neurons selective for larger numbers display broader tuning curves, with the spread of tuning curves remaining constant across neurons on a logarithmic number scale(Nieder and Miller, 2003). This organization of neuronal tuning curves reflects the Weber-Fechner law, exhibiting a structure fundamentally similar to more tractable low-level sensory processes.

Various studies have compared neuronal representations of symbolic and nonsymbolic numbers. Non-human primates have shown the ability to associate Arabic numerals with nonsymbolic numbers. Neurons tuned to Arabic numerals demonstrate tuning curves akin to those of classic nonsymbolic number-tuned neurons. Some neurons are tuned to numbers in both symbolic and nonsymbolic formats, which are more abundant in the PFC compared to the PPC (Diester and Nieder, 2007). In humans, blood-oxygen-level-dependent (BOLD) signals in the PPC respond to the magnitude of deviation from adapted numerical stimuli,

irrespective of the presentation format (Piazza et al., 2007). Nevertheless, differences exist in the neuronal representations of numbers in different formats. Symbolic numbers are represented more categorically than nonsymbolic numbers in the human medial temporal lobe (Kutter et al., 2018). The impact of format on the neuronal representation of numbers in the human PFC and PPC at the single-neuron resolution remains an area for further investigation.

1.5 The challenges

The functional and physiological complexity of higher cognition presents several challenges for research. Firstly, higher cognition requires appropriate tasks to be probed. The cognitive complexity should be high enough such that the automated low-level functions are not sufficient and that higher cognition must be involved (Fuster, 2001). For example, the tasks that involve stimuli with certain levels of abstractness, require holding information for an extended period and have complex task structures that reflect more than passive maintenance, are better suited. Consequently, this puts a constraint on the choice of model organism, often necessitating experiments with (non-human) primates.

Secondly, neural recordings should have sufficient resolution and reflect relevant physiological activity for higher cognition. Electroencephalography (EEG) and functional magnetic resonance imaging (fMRI) may not reveal the intricate dynamics and computations in local circuits underlying higher cognition. Invasive methods, such as extracellular recordings, can provide more detailed insights.

Thirdly, access to higher-order brain regions is crucial. This is often difficult in human experiments, as implanting electrodes into the human brain poses ethical concerns, and recording sites are often determined based on medical requirements rather than research questions.

Finally, to understand the neural mechanisms of higher cognition, our concept of population coding needs to be updated. The neuronal activity in higher-order cortices is not purely stimulus-driven. It does not always follow stimulus-onset and can exhibit diverse temporal modulations (Jacob and Nieder, 2014). A static view of population response patterns is insufficient. Constructing the stimulus coding subspace using temporally averaged activity may obscure neurons' selectivity and bias interpretation. Furthermore, the relationship between single-neuron coding properties and population-level representations needs clarification. Among the many possible mechanisms derived from the same population dynamics, adhering to physiology helps narrow down the hypothesis space.

In this thesis, I focus on the working memory of number stimuli - the high-level temporal integration function and high-level abstract cognition. I aim to determine what is represented in neuronal populations in higher-order cortices during working memory tasks, how these representations are arranged, and how neurons contribute to the population code.

In the first manuscript (collaborative work), we advanced the investigation of brain function using acute micro-electrode array recordings in patients undergoing awake tumor surgery. This approach enabled access to large areas of the cortex including parietal association area with single-unit resolution. The main contribution of this thesis involved examining the population-level neural coding for number stimuli in a working memory task through a decoding approach. I found that the representation of numbers in symbolic and nonsymbolic forms displayed distinct dimensionality and geometrical constructs in the delay period after stimulus offset.

In the second manuscript, I aimed to describe and exploit the neuronal organizing principle of population coding to address the debate surrounding working memory mechanisms—whether memory content is maintained via continuous or sequential representations. The framework I proposed harnessed the physiological principles of neuronal organization, enabling the dissection of complex and often ambiguous representations in higher-order cortices into components that maintain connections to individual neurons.

Chapter 2

Results

2.1 Neuronal representation for numbers in human working memory

Manuscript 1: Human acute microelectrode array recordings with broad cortical access, single-unit resolution and parallel behavioral monitoring

Authors: Viktor M. Eisenkolb, Lisa M. Held, Alexander Utzschmid, **Xiao-Xiong Lin**, Sandro M. Krieg, Bernhard Meyer, Jens Gempt, Simon N. Jacob

Author contributions

V.M.E., B.M., J.G. and S.N.J. conceived the study and designed the experiments. S.K. and J.G. performed the surgeries and implanted the arrays. V.M.E. and S.N.J. collected the data. V.M.E., L.M.H., A.U. and X.-X.L. analyzed the data and prepared the figures. S.N.J. wrote the manuscript with contributions from V.M.E., L.M.H. and A.U. All authors edited the manuscript.

Note: Figure 6J-L and Figure S2 comprise all the analyses I performed for this manuscript.

Human acute microelectrode array recordings with broad cortical access, single-unit resolution and parallel behavioral monitoring

Viktor M. Eisenkolb[1,2], Lisa M. Held[1], Alexander Utzschmid[1], Xiao-Xiong Lin[1,3], Sandro M. Krieg[2], Bernhard Meyer[2], Jens Gempt[2]†, Simon N. Jacob[1,2]†*

[1]Translational Neurotechnology Laboratory, Department of Neurosurgery, Klinikum rechts der Isar, Technical University of Munich, Ismaninger Str. 22, 81675 Munich, Germany.

[2]Department of Neurosurgery, Klinikum rechts der Isar, Technical University of Munich, Ismaninger Str. 22, 81675 Munich, Germany.

[3]Graduate School of Systemic Neurosciences, Ludwig-Maximilians-University Munich, Großhaderner Straße 2 82152 Planegg-Martinsried, Germany

† These authors contributed equally to this work

* Correspondence: simon.jacob@tum.de

Summary

There are vast gaps in our understanding of the organization and operation of the human nervous system at the level of individual neurons and their networks. Here, we report reliable and robust acute multi-channel recordings using planar microelectrode arrays (MEA) implanted intracortically in awake brain surgery with open craniotomies that grant access to large parts of the cortical hemisphere. We obtained high-quality extracellular neuronal activity at the microcircuit, local field potential level, and at the cellular, single-unit level. Recording from parietal association cortex, a region rarely explored in human single-unit studies, we demonstrate applications on these complementary spatial scales and describe travelling waves of oscillatory activity as well as single-neuron and neuronal population responses during numerical cognition including operations with uniquely human number symbols. Intraoperative MEA recordings are practicable and can be scaled up to explore cellular and microcircuit mechanisms of a wide range of human brain functions.

Introduction

There are vast gaps in our understanding of the organization and operation of the human nervous system at the level of individual neurons and their networks. Limited opportunities to directly access the human brain call for multidisciplinary collaborations that combine expertise in neuroscience and clinical medicine to invasively measure neuronal activity with single-unit resolution [1]. This approach has been most fruitful in patients with medically intractable epilepsy implanted with microwire bundles [2-8] and in patients with movement disorders undergoing deep brain stimulation (DBS) [9-11]. Two crucial challenges persist, however, in the investigation of the cellular and circuit physiology of human brain functions. First, epilepsy and DBS surgeries do not provide comprehensive brain coverage, leading to strong focusing of current human single-unit studies on the medial temporal lobe (MTL) and on small circumscribed regions of the frontal lobe. Second, reliable and robust recording technology is still lacking, meaning that clinicians must be trained on increasingly complex devices that necessitate significant modifications to standardized and proven surgical procedures [12,13].

Broad access to the human cortex in large patient groups combined with easy-to-implement methods would greatly accelerate progress in researching the neuronal basis of human brain functions. Here, we demonstrate acute recordings from planar multi-channel microelectrode arrays (Utah MEAs) implanted intracortically in patients operated awake for the removal of left-hemispheric brain tumors. Tumor surgeries with open craniotomies expose large areas of cortex and allow for flexible placement of recording devices, meaning that electrode positions can be adapted to research questions - not vice versa. Awake surgeries with intraoperative functional mapping minimize the risk of postoperative deficits by delineating functionally important regions and thus increase the precision of tumor resection [14]. Patients undergoing awake surgery can perform a wide variety of tasks tapping into sensorimotor functions, visuospatial functions, language and other higher cognitive functions [15]. Penetrating, intracortical MEAs are widely used for chronic measurements of single-unit and population activity in non-human primates [16,17] and have shown potential for clinical applications [18,19] as well as for neurorestorative brain-computer-interfaces (BCIs) in humans [20-25].

Despite these successes, acute intraoperative MEA recordings to investigate human brain functions have not been reported. Cortical microtrauma and neuronal 'stunning' are believed to prohibit measurements with these devices shortly after implantation [26,27].

In this study, we show that these obstacles can be overcome with appropriate choice of the arrays' geometrical configuration. We hypothesized that the degree of tissue impact, and thus the quality of acquired neuronal signals, would depend on the number of implanted electrodes, and in particular the electrode density: increased electrode spacing (lower density) might result in larger pressure at the individual electrode tip during implantation (given the same force applied to the back of the array) and thus allow for faster and less traumatic cortical penetration. We therefore systematically compared

higher density MEAs (standard array, 96 electrodes with 400 μm spacing) and lower density MEAs (custom array, 25 electrodes with 800 μm spacing). We found that all implanted arrays recorded high-quality extracellular signals at the microcircuit level (local field potentials, LFPs). MEAs with increased electrode spacing, however, outperformed standard arrays with higher densities and also captured activity at the cellular, single-unit level. To demonstrate applications on these complementary spatial scales, we describe oscillatory dynamics in the form of waves of activity travelling across human parietal association cortex, a region rarely explored in human single-unit studies, and investigate single-neuron mechanisms of numerical cognition including operations with uniquely human symbolic quantities. Our findings demonstrate that intraoperative MEA recording technology is suited to provide the high-volume recordings necessary to advance translational research on the cellular and microcircuit basis of a wide range of human brain functions.

Results

Intraoperative MEA implantation

Awake surgeries with open craniotomies enable direct, controlled investigations of human brain functions while the patients are alert and can perform tasks of varying complexity [15] (Fig. 1A). Craniotomies overlap in particular over the motor cortical regions and over the posterior frontal lobes (Fig. 1B). They can extend anteriorly to the frontal pole and posteriorly to the parieto-occipital junction, dorsally to the inter-hemispheric fissure (midline) and ventrally to the temporal lobe. Typical craniotomies expose large regions of cortex (several tens of cm^2), yielding broad access to the human brain. Infrared thermal imaging during a representative surgery verified that physiological temperatures are maintained at the cortical surface (Fig. 1C).

We performed a total of 13 acute microelectrode array (MEA) implantations in patients undergoing surgery for brain tumor resection (one array per patient), eight of which were operated awake (Table 1). Except for the procedures related to the array implantation, the course of the surgery was not changed. Following skin incision, preparation and opening of the skull and dura mater, but before awakening the patient from anesthesia, we placed the array's pedestal next to the craniotomy, anchored it with skull screws and positioned the MEA over the target cortical area (Fig. 1D). Reference wires were inserted under the dura. We intended for the implantation site to lie as remotely as possible from the bulk tumor tissue but still within the pre-operatively determined resection area. The array was then pneumatically inserted and covered with saline irrigated strips (Fig. 1E) until explantation, typically when tumor resection started. With established and practiced procedures, the implantation could be performed in less than ten minutes. We encountered no adverse clinical events in connection to MEA implantation or recordings, neither during the surgery nor during routine patient follow-up over several months to years.

For each participant, the implantation site was reconstructed using intraoperative photographic documentation as well as pre-operative structural MR imaging. Three implantations were located in frontal cortex and ten in parietal cortex (Table 1). Examples of implantations in the middle frontal gyrus, the supramarginal gyrus and the angular gyrus are shown (Fig. 1F).

We histologically analyzed three implantations (Table 1). Grids of electrode tracts could be clearly identified from the penetration of the pia mater along the course of the shafts to - in some instances - the tip of the electrode (Fig. 1G). The majority of the electrode tracts reached deeper cortical layers. In two patients, cortical tissue surrounding the electrodes showed no structural abnormalities across the entire array. In one patient, we observed small microbleedings without a space-occupying effect along several electrode tracts as well as in deep cortical layers [26,27] (Fig. 1H). However, these changes were strictly confined to the vicinity of the electrodes. We did not detect any pathology distant from the implantation site.

In sum, implantation of intracortical MEAs in patients undergoing awake brain surgery is safe and practicable, achieving broad and direct access to the neuronal networks of the human cortical left hemisphere.

Figure 1

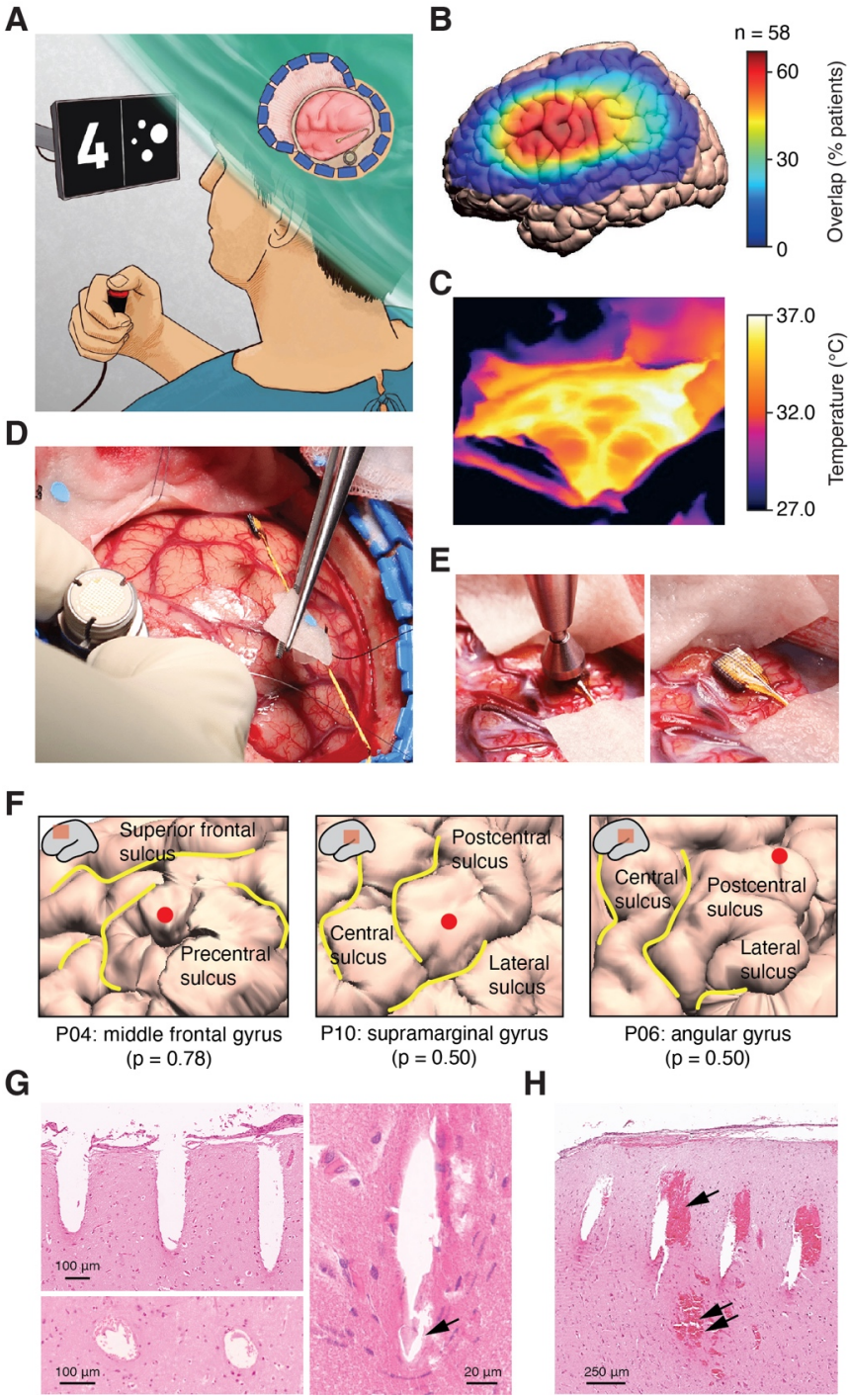


Fig 1. Awake brain surgery and intraoperative microelectrode array implantation. *(A) Schematic of awake brain surgery providing access to the human cortex for microelectrode recordings in*

*participants who can perform cognitive tasks. (**B**) Overlap of craniotomy locations in neurosurgical patients operated awake for the removal of left-hemispheric brain tumors (n = 58 surgeries performed in our department over the course of five years) projected onto the ICBM template brain. (**C**) Infrared thermal imaging of the cortical surface during a typical craniotomy procedure. (**D**) Placement of the microelectrode array in preparation of implantation. (**E**) Pneumatic insertion of the microelectrode array into cortex. (**F**) Cortical surface reconstruction of the implantation site in three example participants. The probability of implantation in the specified gyrus is given according to the JuBrain probabilistic cytoarchitectonic map. (**G**) Histological sections of an example implantation site showing electrode tracts as they penetrate the pia mater (top left, longitudinal section), along the electrode shaft (bottom left, axial section) and at the electrode tip (right, arrow). (**H**) Histological section of a different implantation site showing microhemorrhages along the electrode tracts (single arrow) and in deeper cortical layers (double arrow).*

Extracellular signal quality on MEAs with differing geometrical configurations

In the group of patients operated for awake tumor resection, we discontinued the anesthesia following MEA implantation. We began recording wide-band extracellular activity (Fig. 2A) as soon as the patients were alert and able to engage in conversation with the clinical team and prior to cortical electrostimulation for mapping of language-associated areas. Typically, the arrays had been settling for 30 to 40 minutes. We emphasize that the surgery was not prolonged by this time period; we merely used the awakening time to allow for the signals to develop and stabilize.

We first sought to evaluate the ability to detect the activity of individual neurons (i.e. spikes), present in the high frequency signal components (high-pass filter 250 Hz; Fig. 2B-F). We compared two different MEA configurations: a standard, higher-density array with 400 µm electrode spacing (pitch) and 96 active channels on a 10x10 grid and a custom, lower-density array with 800 µm pitch and 25 channels (Fig. 2C left and right, respectively). Electrode lengths were 1.5 mm for both array types. We performed four implantations with each array type (Table 1). Technical difficulties with grounding (P08, higher-density array) and a medical complication not related to the implantation (P12, lower-density array) did not allow us to advance to neuronal recording in two surgeries. In one case, we observed an abrupt drop in signal quality a few minutes into data acquisition (P13, lower-density array), prompting us to omit this data set from in-depth analysis. Qualitatively, prior to the unexplained event, the recording was not different from the other lower-density recordings.

The likelihood of recording spiking activity varied significantly between array configurations. In an example higher-density array, spiking activity of sufficiently high amplitudes for subsequent waveform sorting was present in only a few channels (Fig. 2D, left). In contrast, in an example lower-density array, spikes were detected on all electrodes (Fig. 2D, right). SNRs in this array were stable across the

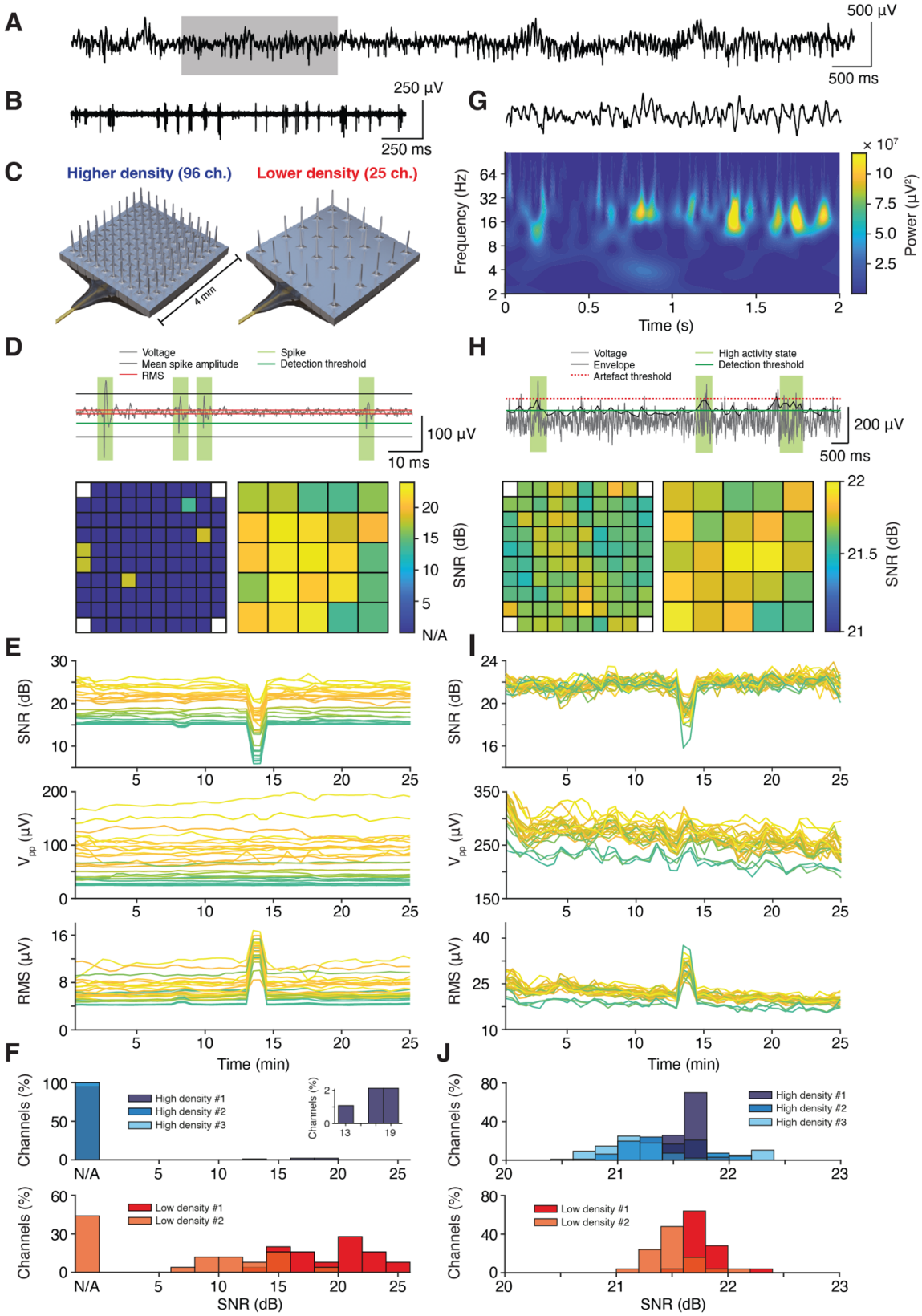
entire recording (25 minutes), with the exception of a single large electrical artefact leading to an increase in noise (Fig. 2E; Fig. S1A, B). This did not impact spike amplitudes, however, which remained stable during data acquisition. Across all successful recordings, this pattern was reproduced (Fig. 2F): in three consecutive implantations with the higher-density array (five implantations including two anesthetized participants, Table 1), we did not observe appreciable spiking activity (2 % of channels). In three consecutive implantations with the lower-density array (one recording not shown due to early termination, see above), we obtained spikes on the majority of channels (78 % of channels; $p < 0.001$, Fisher's exact test higher-density vs. lower-density arrays). In the event that spiking activity could be recorded, SNRs were comparable (mean 17.1 ± 0.9 dB and 16.8 ± 0.8 dB for higher-density and lower-density arrays, respectively; $p = 0.91$, two-tailed Wilcoxon test).

Next, we evaluated the quality of LFPs, a measure of local network activity, i.e. the low-frequency component of our extracellular recordings (low-pass filter 250 Hz; Fig. 2G-J). Epochs of increased LFP activity were readily detected in both higher-density and lower-density arrays and across all channels (Fig. 2H; same example arrays as in Fig. 2D). In both array configurations, SNRs were high and displayed spatial clusters of similar signal strength. In the lower-density array, the clusters of high spiking SNR and high LFP SNR overlapped. As for the spiking activity, LFP signals were stable across the recording session and affected only momentarily due to a single electrical artefact (Fig. 2I; Fig. S1A, B). Across all successful recordings, LFP SNRs were very uniform across channels (mean 21.5 ± 0.1 dB and 21.7 ± 0.03 dB for higher-density and lower-density arrays, respectively; Fig. 2J).

Overall, electrical artefacts could be well controlled during intraoperative data acquisition. Very rarely, we observed a single high-amplitude 'pop' across all electrodes that disrupted recordings for a few hundred milliseconds until the signals settled again (Fig. S1A, B). Such electrode 'pops' have been reported with sudden changes in impedance, likely related to the recording system electrostatically discharging when in contact with a liquid such as blood [28]. 50 Hz line noise and its harmonics were regularly present in the recordings (Fig. S1C, D), but could be efficiently removed by offline filtering. Good grounding (i.e. strong connection of the pedestal to the skull) significantly reduced the hum. Bad choice of grounding, in contrast, lead to signal contamination, e.g., by facial

muscle activity (Fig. S1E, F).

Figure 2



182

Fig 2. Extracellular neuronal signals recorded from microelectrode arrays with different densities. *(**A**) Wide-band extracellular voltage signal recorded at an individual electrode (10 s trace). (**B**) High-pass filtered signal showing extracellular spiking activity in the section highlighted in (A) (2 s trace). (**C**) CAD drawings of the standard higher-density microelectrode array (left, 96 active channels) and of the custom lower-density microelectrode array (right, 25 active channels) used for intraoperative recordings. (**D**) Top: Schematic of the procedure for identifying spikes in high-pass filtered voltage signals. Bottom: Session-averaged SNR of a representative higher-density and a lower-density array (left and right, respectively). (**E**) Time course of spike SNR (top), peak-to-peak amplitude (middle) and RMS noise (bottom) across the entire session (bin width 60 s, step 30 s) recorded with the lower-density array in (D). Note the brief increase in noise and reduction in SNR in the middle of the recording. (**F**) Distribution of spike SNR values obtained from electrodes in higher-density and lower-density recordings (top and bottom, respectively). (**G**) Low-pass filtered signal showing oscillatory LFP activity in the section highlighted in (A) (2 s trace). (**H**) Top: Schematic of the procedure for quantifying SNR in low-pass filtered voltage signals. Bottom: Session-averaged SNR of a representative higher-density and a lower-density array (left and right, respectively; same arrays as in (D)). (**I**) Time course of LFP SNR (top), peak-to-peak amplitude in high activity states (middle) and RMS in low activity states (bottom) across the entire session (bin width 60 s, step 30 s; amplitude and RMS determined within the same bins) recorded with the lower-density array in (D). Note the same deflections in LFP noise and SNR as in the spike-filtered signal in (E). (**J**) Distribution of LFP SNR values obtained from electrodes in higher-density and lower-density recordings (top and bottom, respectively).*

To determine whether single units could be isolated from the population (multi-unit) spiking activity (Fig. 3A), we sorted the thresholded waveforms. Distinct waveform clusters representing well-isolated single units were separated from noise (Fig. 3B, C) with little to no loss of spikes around the detection threshold (false negatives, Fig. 3D; less than 5 % of spikes in 74 % of units), no contamination by spikes violating the refractory period (false positives, Fig. 3E; less than 1 % of spikes in all units), stable firing rates throughout the recording session (Fig. 3F) and little to no mixing of spikes between different clusters (Fig. 3G). Following this procedure, single units could be isolated on the majority of electrodes in the example lower-density array (Fig. 3H), with two or more single units present on multiple channels. Across all analyzed recordings, single units were rarely picked up by the higher-density arrays (2 % of channels) but frequently isolated on the lower-density arrays (62 % of channels; $p < 0.001$, Fisher's exact test higher-density vs. lower-density arrays). On lower-density array electrodes with sortable spikes, we recorded on average 1.6 single units per electrode.

Figure 3

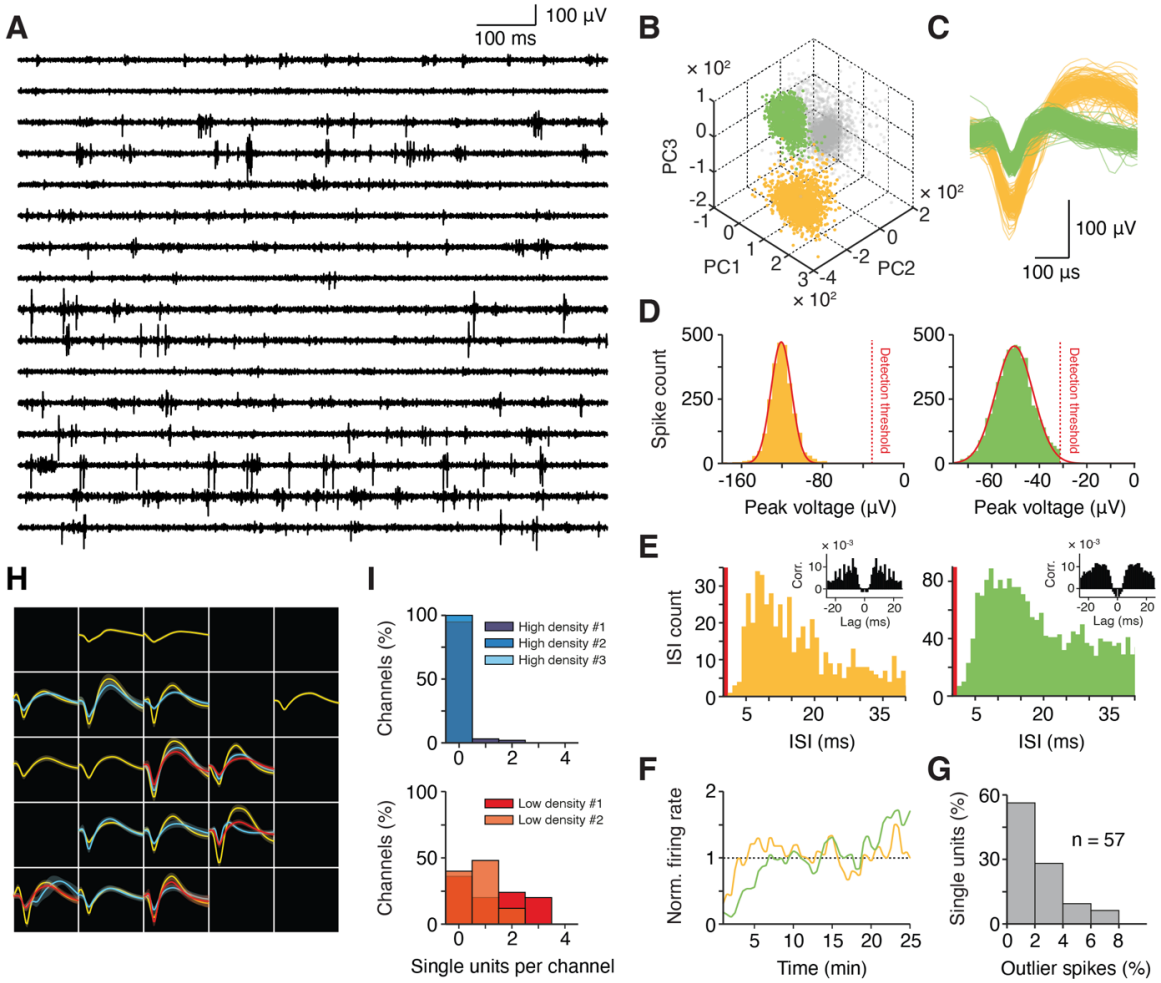


Fig 3. Isolation of single units from intraoperative microelectrode recordings. *(**A**) High-pass filtered extracellular voltage signals from selected electrodes of the same array (P10; 1 s traces). (**B**) Principal component decomposition of thresholded waveforms recorded on an individual channel showing two distinct waveform clusters (yellow, green) separated from noise (gray). (**C**) Waveforms of the single units isolated by PCA in (B). (**D**) Distribution of waveform negative peak (trough) voltages for the two example units with gaussian fits and the selected detection threshold. (**E**) Distribution of inter-spike-intervals (ISI) for the two example units together with spike train autocorrelograms (insets). The refractory period (ISI < 1 ms) is marked in red. (**F**) Firing rates of the two example units across the entire recording session, normalized to a unit's session-averaged activity. (**G**) Distribution of the percentage of spikes per unit that are assigned to different waveform clusters and thus considered outliers (n = 57 sorted units in all recordings). (**H**) Average single unit waveforms recorded from a lower-density microelectrode array. Bands indicate standard deviation across waveforms. Channels with multi-unit activity, but no well-isolated single units, are black. (**I**) Distribution of channels with*

well-isolated activity of one or more single units recorded from higher-density and lower-density arrays (top and bottom, respectively).

While single neurons represent the brain's elementary processing units, it is increasingly recognized that temporal coordination and synchronization of neuronal activity across distances is crucial in particular for higher cognitive functions [29]. Given their planar, grid-like configuration with well-defined spatial relationships between individual electrodes, MEAs are ideally suited to investigate the lateral propagation of activity in cortical networks. Several studies with chronic MEA recordings have reported waves of oscillatory brain activity that travel across the non-human primate and human cortex [30-33] and could reflect higher-order organization of neuronal processing in space and time [34]. Examination of oscillatory beta activity (20 ± 1.5 Hz) in a higher-density recording showed LFP peaks temporally shifted across neighboring electrodes with ordered progression of activity from one side of the array to the other (top to bottom in Fig. 4A). At each timepoint, LFP phases across the array could be approximated by a linear plane with non-zero slope aligned to the direction of activity propagation, in agreement with the notion of a travelling wave. We extracted and characterized such travelling waves in 500 ms epochs following presentation of visual stimuli (sample numbers, see Fig. 5) for both theta (6 - 9 Hz) and beta LFP bands (15 - 35 Hz; Fig. 4B-E). Waves travelled in preferred directions ($p < 0.001$ in theta and beta, Hodges-Ajne test for nonuniformity) that were frequency-band-specific (Fig. 4B). A second modal direction almost opposing the dominant primary direction suggested a spatial propagation axis (Fig. 4B), in line with intracranial EEG and ECoG recordings [35-37] and during ictal discharges in patients with epileptic seizures [38,39]. With increasing oscillatory frequency, travelling waves were detected less often (Fig. 4C) and showed higher propagation velocities (theta mean 0.57 m/s, beta mean 2.40 m/s; Fig. 4D), again matching data from chronic MEA recordings (e.g. in non-human primate prefrontal cortex [30]). Spatial phase gradients fit the plane model well in both frequency bands (measured by Phase-Gradient Directionality, PGD; theta mean 0.72, beta mean 0.62; Fig. 4E). For comparison, we conducted the same analysis in a lower-density recording (Fig. 4F-J). In this participant, beta waves dominated (Fig. 4H) with steeper phase gradient slopes indicating slower propagation speeds (theta mean 0.23 m/s, beta mean 0.96 m/s; Fig. 4I). Overall, travelling waves were again reliably detected (PGD theta mean 0.72, beta mean 0.71; Fig. 4J) and obeyed the same regularities as in the higher-density recording.

Figure 4

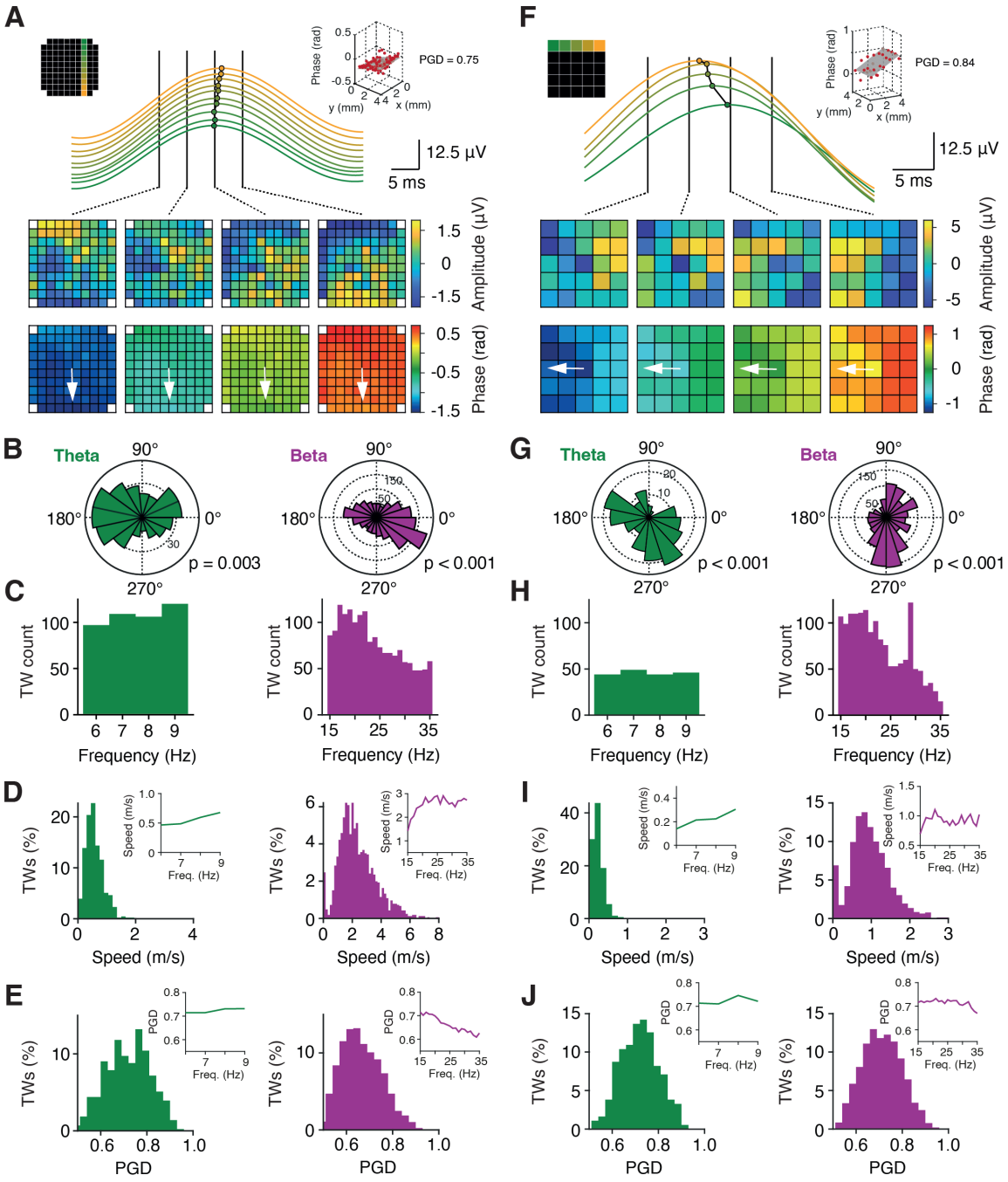


*Fig 4. **Propagation of waves of oscillatory activity across microelectrode arrays.** (A) Example travelling wave recorded on a higher-density array. Top: peaks of LFP beta activity (20 ± 1.5 Hz) are temporally shifted across neighboring electrodes, illustrating the propagation of neural activity. Middle: demeaned LFP activity (amplitude) across the array at four example timepoints. Bottom: phase gradient across the array per timepoint. The arrow indicates the direction of wave propagation (from top to bottom). Inset: linear plane fitted to the phase gradient across the array at one example timepoint.*

*(**B-E**) Distribution of travelling wave (TW) directions (B), count per frequency bin (C), speed (D) and plane model goodness-of-fit (PGD, E) in the theta (6 - 9 Hz, left) and beta (15 - 35 Hz, right) band in 500 ms epochs following the presentation of visual stimuli (sample numbers, see Fig. 5). Insets in (D) and (E) show frequency-resolved speed and PGD, respectively. p-values in (B) are given for Hodges-Ajne test for nonuniformity. (**F-J**) Same layout for travelling waves recorded on a lower-density array. TW, travelling waves; PGD, phase gradient directionality.*

In sum, our neurophysiological signal analysis showed that acquisition of multi-channel extracellular neuronal activity via intracortically implanted MEAs is feasible in the setting of awake brain surgery with its tight clinical and procedural constraints. Mesoscale network (LFP) activity for studying both local and propagating neuronal oscillations was obtained in high quality in every recording, while the extent of microscale spiking activity and yield of single units depended on the array configuration and favored the use of MEAs with increased electrode spacing.

Probing higher cognitive functions in awake brain surgery

In parallel to neuronal data acquisition, we administered a task to the participants to probe the human number sense, a higher-level cognitive function of the parietal and (lateral) prefrontal association cortex that enables us to represent and manipulate abstract numerical categories [40]. The frontoparietal cortex has undergone disproportionate expansion in human evolutionary history, but is hardly ever targeted in single unit studies with DBS or epilepsy patients.

All six patients with recordings from either higher-density or lower-density arrays (Figs. 2 and 3) performed a delayed-match-to-sample task requiring them to memorize a visually presented sample number and compare it to a subsequently presented test number (Fig. 5A). Stimuli were presented either in nonsymbolic notation (sets of dots, numerosities) or in symbolic notation (Arabic numerals), allowing us to investigate the neuronal coding of and mapping between 'non-verbal' number, which animals have access to, and 'verbal' number, which is unique to humans. In half of the nonsymbolic trials, dot diameters were selected at random. In the other half, dot density and total occupied area were equated across stimuli. This visual variation in the presented images ensured that subjects processed the numerical information contained in the stimuli and that low-level, non-numerical visual features could not systematically influence task performance [41].

Four patients performed well in all conditions, whereas two patients (P07 and P09, higher-density arrays) did not exceed chance level in the nonsymbolic (dot) trials and were excluded from further analysis. There was only a small reduction in intra-operative response accuracy compared with pre-operative training levels ($p = 0.04$, one-tailed *t*-test; Fig. 5B) and a small increase in intra-operative

response times ($p = 0.23$, one-tailed *t*-test per participant; $p < 0.001$, one-tailed Wilcoxon test with pooled trials; Fig. 5C). Following a brief 'warm-up' period, all patients maintained high performance levels throughout the recording session and completed between 200 and 300 trials (Fig. 5D).

The patients' task performance was qualitatively very similar during pre-operative training and intra-operative recording and not distorted (compare Fig. 5E, F with Fig. 5G, H). Errors were more frequent during surgery, in nonsymbolic trials and for larger numbers ($p_{setting} = 0.02$, $p_{notation} = 0.003$, $p_{number} = 0.01$, 3-factorial ANOVA; Fig. 5E, G). Behavioral tuning functions (Fig. 5F, H) showed that participants correctly matched sample and test stimuli in particular for small numbers (peak of each curve), while accuracy dropped with increasing number. In non-match trials, the percentage of errors depended on the numerical distance between sample and test (distance effect; fewer errors for larger distances) and on the absolute magnitudes of the compared numbers (size effect; fewer errors for small numbers). Together, these results show that all key behavioral signatures of numerical cognition were captured by the task administered to the participants.

Figure 5

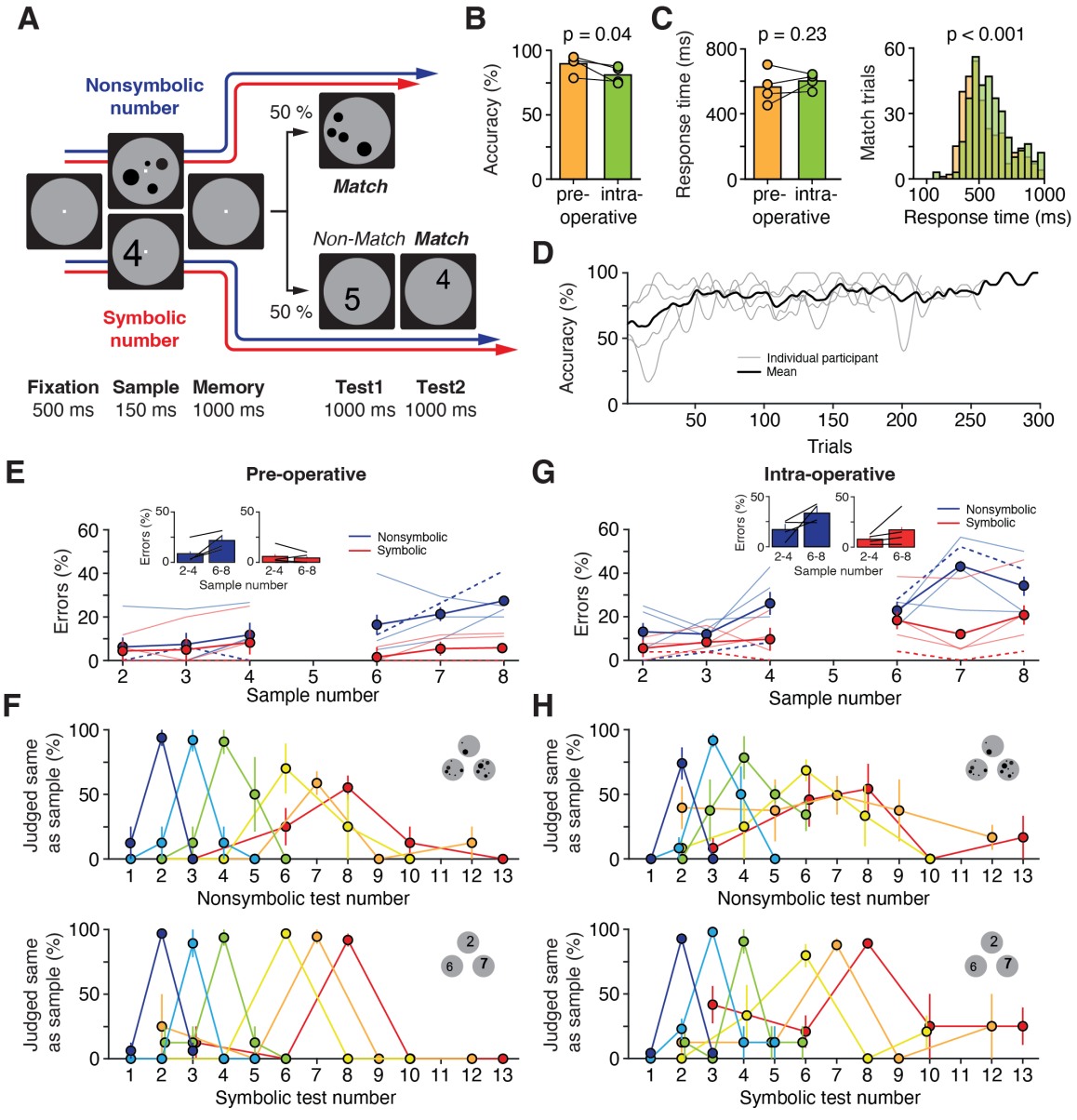


Fig 5. Preoperative and intraoperative cognitive performance in patients undergoing awake brain surgery. *(**A**) Delayed-match-to-number task. Participants memorized the number of the sample stimulus and compared it to a subsequently presented test number. Trials were presented either in nonsymbolic notation (sets of dots, numerosities) or in symbolic notation (Arabic numerals). (**B**) Preoperative and intraoperative task performance (n = 4 participants; one-tailed t-test). (**C**) Preoperative and intraoperative response times in match trials on a per-participant basis (left) and pooled across trials (right) (one-tailed t-tests). (**D**) Time courses of intraoperative task performance across sessions. (**E**) Percentage of errors during preoperative behavioral testing plotted as a function of sample number and stimulus notation. Inset: performance pooled across small numbers (2-4) and*

*large numbers (6-8). Error bars indicate SEM across participants. Dashed lines mark single-subject data for P10 (see Figs. 6, 7) (**F**) Preoperative behavioral tuning functions for trials with numbers presented in nonsymbolic and symbolic notation (top and bottom, respectively). Performance is shown for all sample-test-combinations. The peak of each curve represents the percentage of correct match trials, and other data points mark the percentage of errors in non-match trials. Error bars indicate SEM across participants. (**G**) Same layout as in (E) for intraoperative testing. (**H**) Same layout as in (F) for intraoperative testing*

Human neuronal coding of number at the micro- and mesoscale level

Extracellular recordings in the non-human primate frontoparietal cortex suggest that single units tuned to individual numerosities give rise to numerical cognitive abilities [41-43]. The human neuronal code for number in these brain areas, however, is not known. A recent study found single neurons responsive to Arabic numerals in the inferior posterior parietal cortex of two participants implanted for the development of a motor brain-computer-interface, but did not investigate nonsymbolic number representations [44]. Leveraging the flexibility in array placement and high-quality data obtained with MEA recordings from open craniotomies, we illustrate here a potential application of this method by exploring - in parietal cortex (inferior parietal lobule, IPL) of an example participant (P10) - the neuronal correlates of the human number sense at the single-neuron and neuronal network level.

In nonsymbolic trials, an example single unit strongly increased its firing rate after presentation of the sample stimulus (Fig. 6A, left). The increase was graded and a function of sample numerosity with peak activity for 7 and 8 dots. This unit's firing rates were smaller and more transient in trials with symbolic number, but showed a similar graded response (Fig. 6A, right). Average firing rates in the 500 ms epoch following sample presentation confirmed significant tuning to nonsymbolic number, but failed to reach significance in symbolic trials due to the distinct temporal activity profile (Fig. 6B). Thus, this single unit carried information (ω^2 percent explained variance) about sample notation and numerosity (Fig. 6C). Similar responses were found in a different example single unit recorded on a neighboring electrode (Fig. 6D-F). An example multi-unit measured on a different electrode of the same array was tuned to nonsymbolic number 1 (Fig. 6G, left). This unit also showed a congruent response in trials with symbolic numbers, albeit with distinct dynamics and a more categorical coding of small versus large numbers (Fig. 6G, right and Fig. 6H, I).

To provide a population-wide perspective on number coding, we trained a linear discriminant analysis (LDA) decoder to separate small from large numerosities using the entire spiking activity recorded across the array (Fig. 6J-L). In trials with nonsymbolic number, decoding accuracy was high and peaked (86 %) after sample presentation, matching the single unit responses. Cross-temporal training and decoding showed a dynamically evolving code across the memory delay with reduced off-diagonal

359 accuracy (Fig. 6J). In trials with symbolic number, decoding was less accurate (62 % peak) and only
360 possible in the first half of the memory delay, again matching single unit responses (Fig. 6K). The
361 results of cross-notation decoding (training on nonsymbolic number, testing on symbolic number) were
362 qualitatively similar with decoding accuracy bounded by the weaker coding of symbolic number
363 compared to nonsymbolic number (Fig. 6L). Furthermore, to investigate the difference between
364 nonsymbolic and symbolic number coding, we trained a decoder to separate all 6 numbers (chance level
365 16.7 %) with the dimensionality used for decoding systematically manipulated (Fig. S2A, B). We found
366 the decoding accuracy for nonsymbolic number peaked with one dimension while decoding accuracy
367 for symbolic number peaked with two dimensions (Fig. S2C, D). The difference can be understood with
368 the geometrical structure used to represent numbers in the neuronal population, with nonsymbolic
369 numbers represented on a line, signifying magnitude and symbolic numbers each represented more
370 idiosyncratically (Fig. S2E, F).

Figure 6

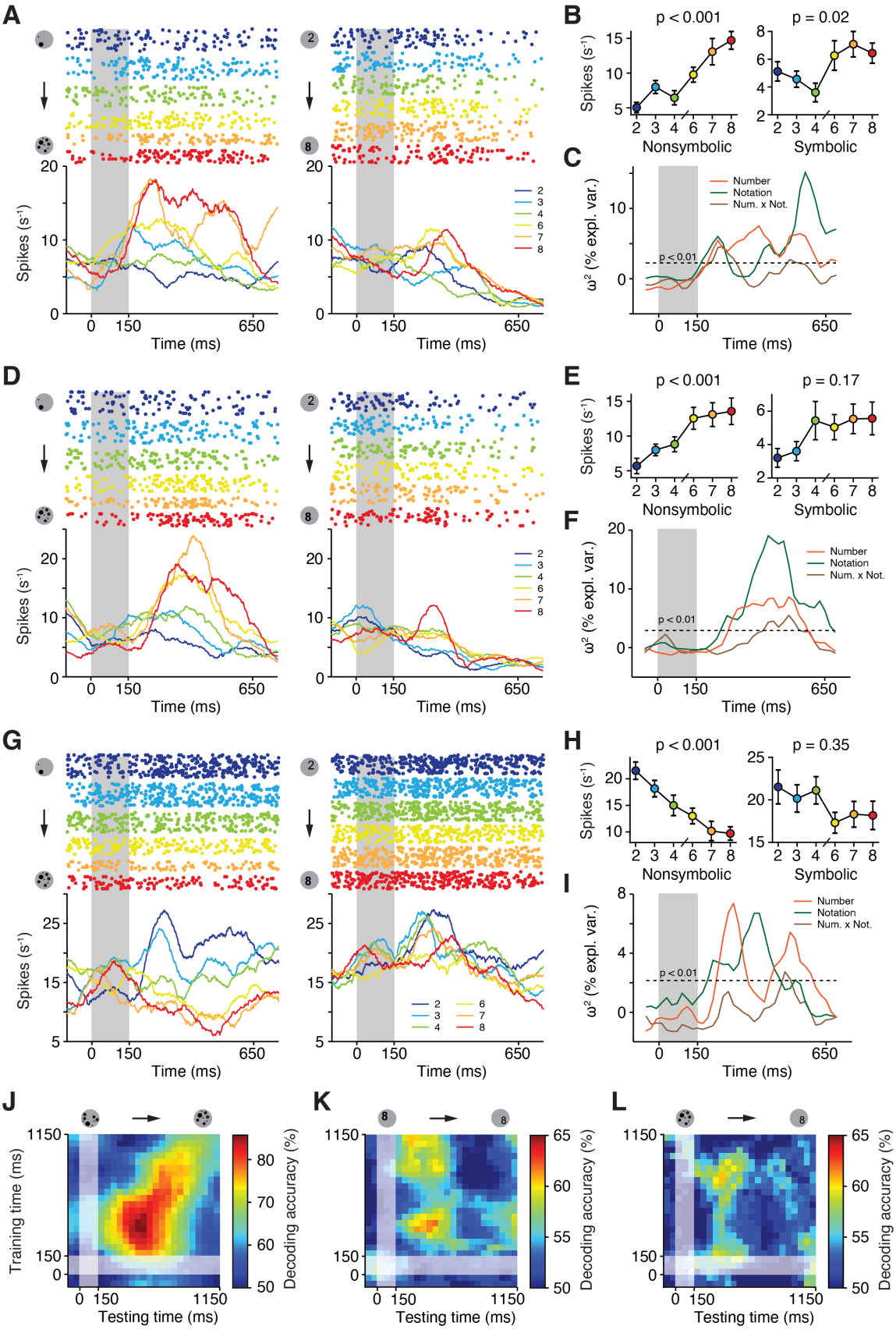

A
B
p < 0.001
p = 0.02
Spikes (s⁻¹)
Nonsymbolic
Symbolic
C
Number
Notation
Num. x Not.
ω² (% expl. var.)
p < 0.01
Time (ms)
D
E
p < 0.001
p = 0.17
F
G
H
p < 0.001
p = 0.35
I
J
K
L
Training time (ms)
Testing time (ms)
Decoding accuracy (%)

Fig 6. Single unit and neuronal population coding of nonsymbolic and symbolic number. *(**A**) Spike raster plots and spike-density histograms (smoothed using a 150 ms Gaussian window) for an example single unit recorded in the inferior parietal lobe. Trials are sorted by sample numerosity and by stimulus notation (left: nonsymbolic, right: symbolic). Sample presentation is highlighted. (**B**) Firing rate of the neuron in (A) in the 500 ms epoch following presentation of nonsymbolic and symbolic sample numerosities (left and right, respectively; one-factorial ANOVA). (**C**) Sliding-window ω^2 percent explained variance (two-factorial ANOVA) quantifying the information about sample number and notation as well as their interaction contained in the firing rate of the neuron in (A) in correct trials. Dashed line marks the significance threshold (p = 0.01; shuffle distribution). (**D-F**) Same layout as in (A-C) for a different single unit recorded on a neighboring channel on the same microelectrode array. (**G-I**) Same layout as in (A-C) for a multi-unit recorded on a neighboring channel on the same microelectrode array. (**J**) Cross-temporal LDA decoding of nonsymbolic number (small, i.e. 2-4, versus large, i.e. 6-8) in the 1000 ms memory epoch following sample presentation using spiking activity (multi-units) on all channels of the microelectrode array. Sample presentation is highlighted. (**K**) Same layout as in (J) for symbolic number. (**L**) Same layout as in (J) for cross-notation decoding. The decoder was trained in trials with nonsymbolic numerosities and tested in trials with symbolic numerosities.*

We then directly compared the microscale neuronal activity elicited during the task with mesoscale network responses. At the same electrode on which the number-tuned single unit shown in Fig. 6A-C was recorded, LFP power varied strongly with sample number and notation (and their interaction) in particular in the gamma band (45 - 100 Hz; ω^2 percent explained variance; Fig. 7A). However, in contrast to the early changes in spiking activity, sample selectivity measured by LFPs increased only 150 ms after sample offset (compare e.g. Fig. 7A left with Fig. 6A left). In the 500 ms epoch following sample number presentation, gamma power increased monotonically with numerosity in nonsymbolic trials, but did not vary with symbolic number ($p < 0.001$ and $p = 0.46$, respectively, one-factorial ANOVA; Fig. 7B top). On two neighboring channels (same electrodes on which units shown in Fig. 6D-F and Fig. 6G-I were recorded) a qualitatively similar pattern was found ($p < 0.001$ and $p = 0.02$, respectively, one-factorial ANOVA; Fig. 7C, D top), albeit with a clear spatial gradient. Beta responses, in contrast, were spatially more uniform, underscoring the local nature of gamma activity and the potentially distinct functional reach of the analyzed frequency bands (Fig. 7B-D bottom). Of note, while not all units in Fig. 6 were tuned to the same preferred numerosity, LFP power scaled uniformly with numerosity across electrodes (compare Fig. 6G left with Fig. 7D top; Fig. S3). Numerosity-responsive electrodes were spatially clustered with overlap of sites selected using LFP activity and sites selected using (multi-unit) spiking activity (Fig. S3). Analysis of propagating oscillatory activity across the array also showed that, at equal strength, travelling waves were faster for larger numerosities (Fig. 7E).

Figure 7

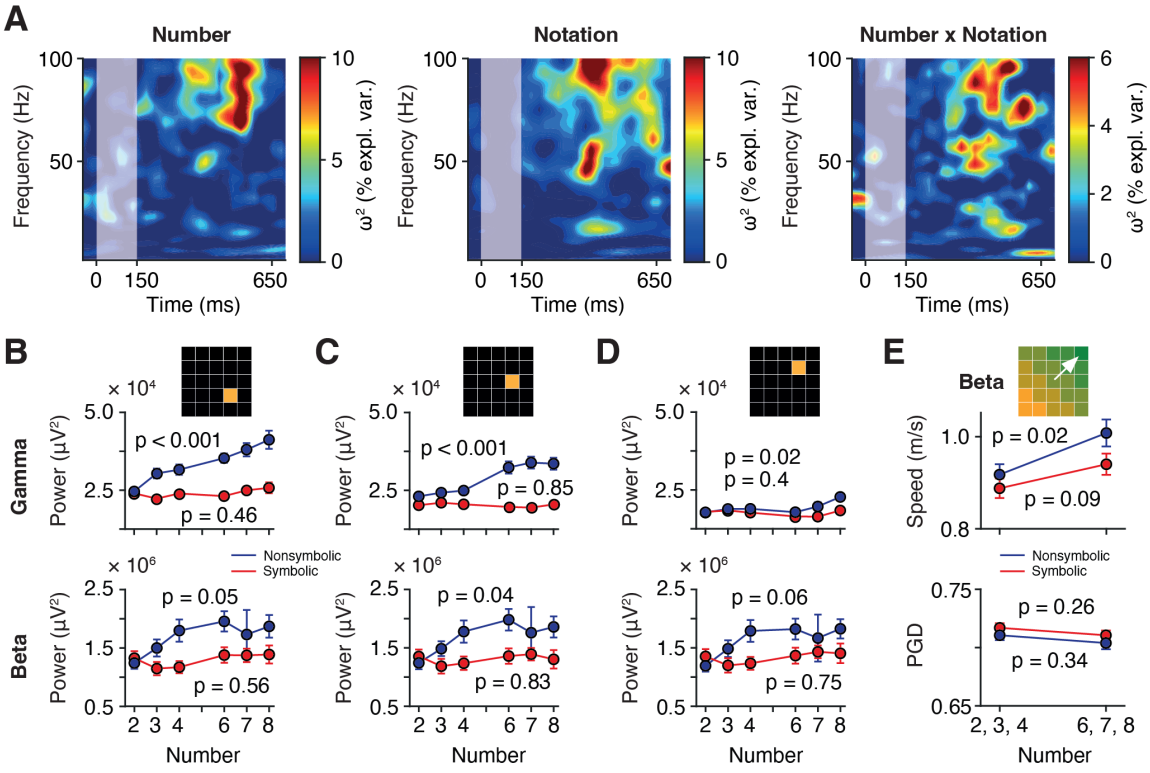


*Fig 7. Local and propagating oscillatory neuronal activity during number coding. (A) Sliding-window ω^2 percent explained variance (two-factorial ANOVA) quantifying the information about sample number (left) and notation (middle) as well as their interaction (right) contained in the LFP power spectrum of an example single channel on a lower-density array (same channel as in Fig. 6A-C) in correct trials. Sample presentation is highlighted. (**B**) LFP power in the gamma (45 - 100 Hz, top) and beta (15 - 35 Hz, bottom) band in the 500 ms epoch following sample number presentation as a function of sample number in nonsymbolic and symbolic notation. Same channel as in (A). p-values are given for one-factorial ANOVA. (**C**) Same layout as in (B) for a neighboring single channel. (**D**) Same layout as in (C) for a neighboring single channel. (**E**) Speed (top) and goodness-of-fit (PGD, bottom) of LFP beta band travelling waves propagating across the array in the 500 ms epoch following sample number presentation for small (2-4) and large (6-8) numbers in nonsymbolic and symbolic notation. p-values are given for one-factorial ANOVA.*

Our proof-of-concept results suggest that, first, the human parietal cortex harbors single units that are tuned to number, establishing a previously missing link to the non-human primate animal model. Second, at the single-neuron level, nonsymbolic set sizes are coded with graded and continuous responses, displaying no sign of a discontinuity in activity that might signal the presence of different

426 neuronal representations for small and large numerosities. A well-studied behavioral signature of the
427 approximate (nonsymbolic) number system, subitizing denotes the accurate apprehension of small
428 numbers of items at a glance (evidenced by a disproportionate increase in errors for larger numerosities
429 in nonsymbolic, but not symbolic notation; single-subject data for P10 [dashed lines] in Fig. 5E, G) and
430 is thought to indicate different representational systems for small and large quantities [45]. In our example
431 participant, we found no evidence for subitizing at the neuronal level. Our findings therefore rather
432 argue that the representation of small and large quantities emerges from a single system [46]. Third,
433 symbolic numbers are coded with distinct temporal dynamics and more categorical responses than
434 nonsymbolic quantities, in line with recent findings in the human MTL [6]. However, the number code
435 partially generalizes across notations with number-congruent responses for nonsymbolic and symbolic
436 stimuli. Fourth, spiking activity and oscillatory activity reflect distinct aspects of numerical information
437 processing in the local microcircuit, with LFPs possibly capturing in particular the network's load-
438 dependent activity state.

Discussion

We found that intracortically implanted MEAs are suitable for acute recordings of human brain activity at both meso- and microscale resolution (Figs. 2-4). All arrays acquired LFPs (synaptic network activity) with high fidelity. Increasing the interelectrode spacing also allowed us to record responses from populations of single units. The devices can be used in awake surgeries with large open craniotomies, providing broad access to the cortex (Fig. 1) in patients who achieve close to normal levels of cognitive performance (Fig. 5). We illustrated a potential application by exploring the neuronal correlates of human numerical cognition in parietal cortex (Figs. 6, 7), a brain region that is typically inaccessible in DBS or epilepsy surgery, i.e. in procedures that so far have produced the vast majority of intracranial data tapping into the neuronal underpinnings of human cognitive functions.

We believe the comparative ease with which MEA recordings can be introduced into the operating room and incorporated into established neurosurgical procedures to be their greatest advantage. Positioning of the array and implantation can be completed within ten minutes. After insertion, the arrays 'float' on cortex. No extra manipulators or electrode holders are required [12,13]. The arrays readily follow brain movements, yielding stable recordings without the need for additional mechanical stabilization [9,10]. Slight shifts of the skull in awake participants and above all vertical displacements of the cortex during brain pulsations pose a major challenge when externally secured probes are used that occupy a different spatial reference frame than the tissue they record from, necessitating elaborate post-acquisition motion correction [12,13]. Furthermore, penetrating MEAs are robust, have a well-documented safety profile and are used with equipment that has been validated for sterilization and re-use. There is no risk of shank breakage, no inadvertent deposition of electrode material in brain tissue, and no need to perform piotomies to allow entry of the device into cortex as with more delicate (e.g. Neuropixels) probes [12,13]. Good grounding could be reliably achieved either by anchoring the pedestal to the skull or by establishing a strong connection to the head frame. Both configurations were effective in our experience and sufficient to reduce electrical hum and noise to levels that enable high-quality extracellular recordings despite an environment full of potential sources of interference. We did not find it necessary to turn off suction, lighting, warming blankets or any other piece of medical equipment during recording.

The arrays' grid-like electrode arrangement allows for dense sampling of neuronal activity in the horizontal plane, i.e. from a patch of cortex. There is rapidly mounting interest in the mechanisms by which propagating neuronal activity, e.g. in form or travelling waves (Fig. 4), mediates intercortical information transfer [30-33,35-37]. In contrast to microwire bundles with their irregularly placed electrode tips or linear probes that record from one single cortical column, MEAs with their well-defined planar geometry are ideally suited to address such questions. Spatial coverage may be extended even further by the addition of ECoG grids, which can be placed directly on top of MEAs, or intracranial stereo EEG leads [47-49]. Lastly, using MEAs in open craniotomy surgeries where the implanted tissue is resected (as

in our participants) opens up the possibility of complementing the *in vivo* recordings with *in vitro* physiological or histological analyses to explore structural-functional relationships in neural circuit organization [50].

MEAs with increased interelectrode spacing (25 channels) recorded on average more than one well-isolated single unit per channel (Fig. 3). Per patient and recording session, this yield is similar to semi-chronic recordings in epilepsy patients (2 to 3 neurons per microwire bundle with up to 10 bundles implanted per patient [2,6]). Acute DBS recordings from prefrontal cortex (10 to 20 neurons per participant [9,10]) or midbrain structures (fewer than 10 neurons per participant [11,51]) yield less. Efforts are currently underway to establish acute intracranial recordings with high-density linear probes (Neuropixels), which have been reported to pick up between several tens of neurons in open craniotomies [13] to a few hundred units in DBS burr holes [12]. Critical technical challenges are still to be met, but these probes could eventually provide a valuable addition to the armamentarium of intraoperative recording devices from which the neurophysiologist and neurosurgeon can chose depending on the particular research question and clinical setting.

The arrays' geometrical configuration was a crucial determinant of spiking activity SNR (Fig. 2). This is likely a consequence of the electrodes' comparatively large footprint (thickness 180 - 200 µm near the base), the main disadvantage of the MEAs used in this study. Lower-density arrays produce less cortical trauma, thereby increasing the chances of measuring single unit activity shortly after array insertion. Our histological analyses showed microhemorrhages in some [26,27], but not all implantations of standard 96 channel arrays. Cortical neuronal 'stunning' might therefore be an important reason for the very low single unit yield in higher-density arrays. Fittingly, unit activity in our recordings only appeared after several minutes and continued to develop until data acquisition began when the patient was fully awake, a time period significantly longer than recently reported for thinner linear probes [12,13]. A second limitation of the described setup is the difficulty in precisely controlling pneumatic array insertion. Whether the inserter wand is stabilized by a dedicated holder or manually (we preferred the latter to expedite implantation), the inherent variability in inserter positioning will significantly affect the forces that the electrode pad experiences during implantation, much unlike micromanipulator-controlled implantations of e.g. linear probes. Imperfect alignment of the inserter with the array could disproportionately impact implantations of higher-density arrays and in older patients [26], where optimal forces are required to overcome the increased resistance to insertion from the pial meninges and brain tissue. We found it best to place the inserter into direct contact with the array, applying very gentle downward pressure to eliminate dead space between the electrode tips and cortical surface (Fig. 1). This approach resulted in complete array insertions and reproduceable signals for both higher-density and lower-density arrays (Fig. 2).

High-volume recordings are necessary to accelerate progress in our understanding of the neuronal basis of human brain functions. Awake surgeries for tumor resection are performed at many medical centers.

We have shown here that these procedures are as suitable for acquiring cellular resolution data from the human brain as DBS or epilepsy surgeries. As any other probe in the expanding palette of multichannel recording devices [12,13], intracortical MEAs do not promise a fail-safe or turn-key solution. However, the technology is more mature and more lenient in the intraoperative setting where clinical constraints considerably limit options for optimizing the recording setup and neuronal signal quality. Once mastered, it can also be effectively put to use in chronic (e.g. BCI) applications where MEAs represent the gold-standard for intracranial sensors. Human single-unit recordings are multidisciplinary endeavors, for which all stakeholders must advance beyond their comfort zones. The methods we describe here can stimulate productive collaborations between neuroscientists and clinicians and propel forward the exploration of the unique neural computations performed by the human brain.

Limitations of the study

For ethical reasons, invasive human recordings are necessarily confined to brain areas with potential pathological changes. We did not systematically assess array placement in relation to the tumor. But given our surgical planning procedure with intraoperative MRI-guided neuronavigation and inspection of the cortical and vascular anatomy prior to implantation, we are confident that the tumor was distant enough from the recording site in all cases. This notion is confirmed by the absence of tumor cell infiltration into the tissue surrounding the electrodes in our histological analyses (Fig. 1). Although we did not randomize the implanted array type per patient (we performed consecutive implantations with the higher-density array before switching to the lower-density array), we do not think it likely that the surgical team's experience influenced our results. We did not observe a gradual improvement in (spiking) signal quality across the implantations. Instead, there was a disruptive increase in unit activity when we changed from the 96-channel to the 25-channel array. Continued efforts are warranted, in any case, to increase the currently small sample sizes and to further explore the effect of varying surgical expertise, implantation sites and array geometries on the quality of intraoperatively acquired extracellular neuronal signals.

Acknowledgments

This study was supported by grants from the German Research Foundation (DFG JA 1999/5-1), the European Research Council (ERC StG MEMCIRCUIT, GA 758032) and the Else Kröner-Fresenius Foundation (TUM Doctorate Program Translational Medicine) to S.N.J., a grant from the Technical University of Munich (TUM Innovation Network Neurotech) to S.N.J. and J.G., and by a grant from the German Research Foundation to J.G. (GE 3008/3-1). We would like to thank Doris Droese for technical assistance with intraoperative recordings and Sandra Baur, Claire Delbridge and Friederike Liescher-Starnecker for preparing histological sections. We are especially indebted to our patients for their willingness to participate in this research.

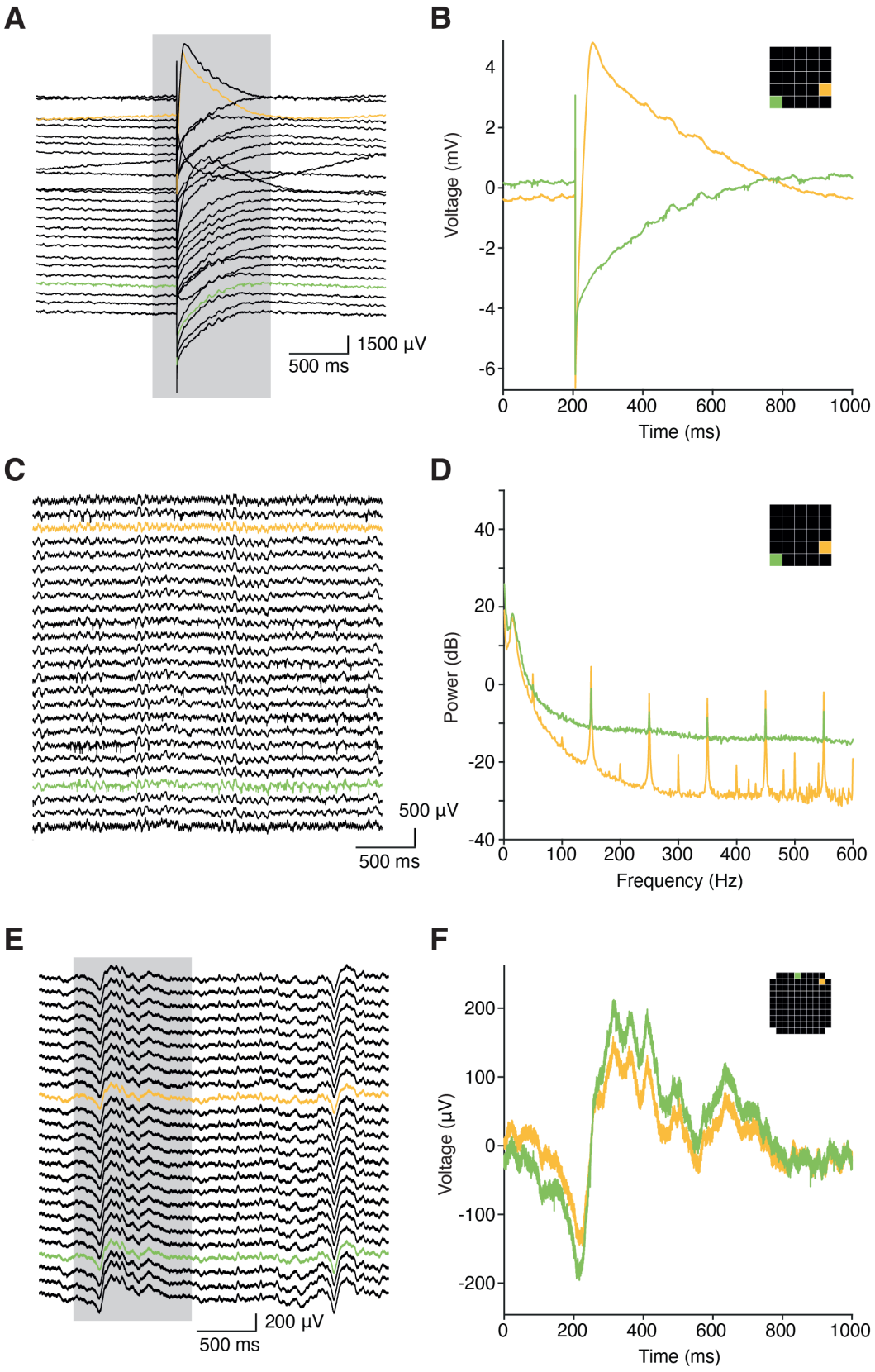
Author contributions

V.M.E., B.M., J.G. and S.N.J. conceived the study and designed the experiments. S.K. and J.G. performed the surgeries and implanted the arrays. V.M.E. and S.N.J. collected the data. V.M.E., L.M.H., A.U. and X.L. analyzed the data and prepared the figures. S.N.J. wrote the manuscript with contributions from V.M.E., L.M.H. and A.U. All authors edited the manuscript.

Declaration of interests

The authors declare no competing interests.

Supplemental Figure S1



555
556 ***Fig. S1. Example electrical artefacts during intraoperative recording.*** *(**A, B**) Single large-amplitude*
557 *electrode 'pop' with prolonged voltage settling time in a lower-density array recording. Note the*
558 *voltage scale and compare to subsequent panels. Two representative channels are highlighted in (B)*
559 *together with their location on the MEA grid (inset). (**C, D**) Line noise (50 Hz) and its harmonics in*
560 *the same recording as in (A, B). (**E, F**) Contamination of the ground in a higher-density array*
561 *recording by frontal facial and ocular muscle activity leading to intermittent slow artefacts.*
562

Supplementary Figure S2

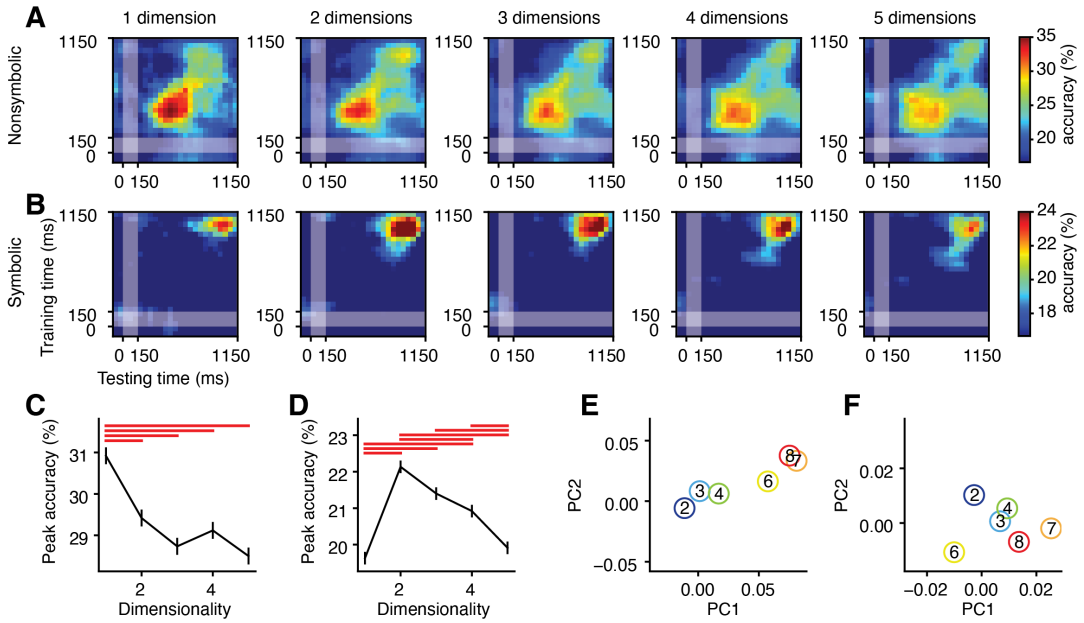


Fig. S2. Dimensionality of number coding. (A) cross-temporal decoding for nonsymbolic number (2, 3, 4, 6, 7 and 8) with decoder's dimensionality controlled. (B) same as (A), but for symbolic number (C) Peak decoding accuracy (smoothed with Gaussian filter with sigma of 100ms) for nonsymbolic number using with different dimensionalities. Error bars: stardard deviation of mean. Red lines: significant difference between the decoding accuracy using different number of dimensions ($p < 0.01$, with Bonferroni correction). (D) same as (C), but for symbolic number. (E) the representational geometry of nonsymbolic numbers during the time window of peak accuracy (500 ms - 800 ms). (F) same as (E), but for nonsymbolic numbers during 800 ms – 1100 ms.

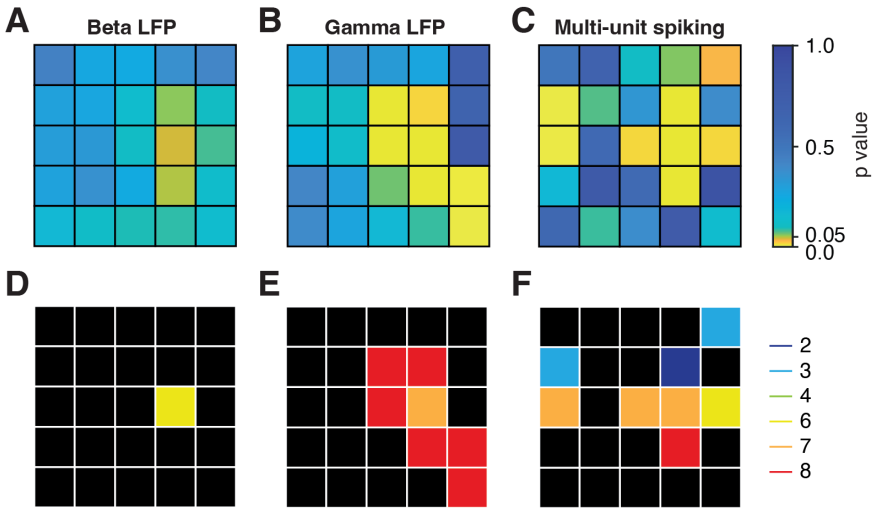


Fig. S3. Spatial clustering of numerosity representations. *(**A, B, C**) P values of one-factorial ANOVA quantifying the degree of numerosity-selectivity per electrode site in the 500 ms epoch following sample number presentation (nonsymbolic trials) using either beta or gamma LFP power (A, B) or multi-unit spiking activity (C). (**D, E, F**) Preferred numerosity at the selective electrode sites ($p < 0.05$) determined by highest power (D, E) or firing rate (F) averaged per numerosity across the 500 ms epoch following sample number presentation.*

.

584 ***Table 1. Study participants.***

585

ID	Sex	Age	Tumor location	Procedure	State	Array location	Channels	Spikes	Single units (Multi units)	Behavior	Notes
P01	F	68	right frontal	histology	anesthetized	inferior parietal cortex	96				
P02	M	54	right parietal	histology	anesthetized	inferior parietal cortex	96				
P03	M	62	right parietal	histology	anesthetized	inferior parietal cortex	96				
P04	M	56	left frontal	setup testing and recording	anesthetized	middle frontal gyrus	96	no	0 (0)		
P05	F	75	left central	setup testing and recording	anesthetized	superior frontal gyrus	96	no	0 (0)		
P06	M	57	left parietal	recording	awake	angular/supramarginal gyrus	96	(yes)	7 (5)	number task	
P07	M	73	left parietal	recording	awake	angular /supramarginal gyrus	96	no	0 (0)	number task	performance non-symbolic trials ↓
P08	F	55	left parietal	recording	awake	inferior parietal cortex	96				no data acquisition bad ground
P09	M	51	left fronto-parietal	recording	awake	middle frontal gyrus	96	no	0 (0)	number task	performance non-symbolic trials ↓
P10	M	32	left temporal	recording	awake	supramarginal/angular gyrus	25	yes	32 (25)	number task	
P11	M	67	left frontal	recording	awake	supramarginal/angular gyrus	25	yes	18 (14)	number task	
P12	M	71	left insular	recording	awake	angular/supramarginal gyrus	25				no data acquisition intracerebral hemorrhage (unrelated to implantation)
P13	F	59	left central	recording	awake	supramarginal/postcentral gyrus	25	yes	N/A (N/A)	number task	spiking activity as in P10 and P11 prior to sudden SNR drop

STAR Methods

RESOURCE AVAILABILITY

Materials availability

This study did not generate new unique reagents.

Data and code availability

- All data reported in this paper will be shared by the lead contact upon request.
- This paper does not report original code.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANTS

We included 13 participants in this study with intracerebral tumors (mainly glioblastoma) referred to our department for surgical resection (Table 1). All study procedures were conducted in accordance with the Declaration of Helsinki guidelines and approved by institutional review board (IRB) of the Technical University of Munich (TUM) School of Medicine (528/15 S). Participants were enrolled after giving informed consent. The scientific aims of this study had no influence on the decision to operate. With the exception of array implantation, the course of the surgery was not altered.

METHOD DETAILS

Multielectrode arrays and implantation procedure

Per participant, one Neuroport IrOx planar multielectrode array (Blackrock Neurotech) was implanted. In nine patients, we implanted the standard array with 96 wired (active) electrodes on a 10x10 grid (1.5 mm electrode length, interelectrode spacing 400 µm). In four patients, we implanted a custom array with 25 channels, which was produced by removal of every second row and column from the standard array (interelectrode spacing 800 µm; Fig. 2c). The modifications were performed by the array manufacturer (Blackrock Neurotech; purchase orders for custom arrays are accepted). The array's pedestal was first anchored to the skull adjacent to the craniotomy. The array was then positioned on the cortical surface of the to-be-implanted gyrus guided by MRI-neuronavigation (Brainlab, Germany). Care was taken to avoid prominent vascular structures, which in some cases prompted us to deviate from the preoperatively determined implantation site by a few millimeters. References wires were inserted under the dura.

The array was implanted pneumatically following the manufacturer's guidelines (Blackrock Neurotech). We found that introducing a dedicated external wand holder was inconvenient, and that positioning of the holder unnecessarily prolonged the implantation procedure. We therefore secured the wand manually such that it touched the array's dorsal pad and brought the electrode tips into contact with the pia. Insertion was performed with a single pulse (20 psi, pulse width 3.5 ms). We did not systematically explore different insertion pressure or pulse width settings. The array was then covered with saline irrigated strips and left to settle. Anesthesia was discontinued in patients planned for awake tumor resection.

All equipment in contact with the patient (inserter wand, trigger, tubing, headstages, cabling) was re-sterilized (Steris V-Pro) and used in multiple surgeries.

In all participants, the implantation site was chosen to lie within the resection area surrounding the tumor. In some cases, however, intraoperative evaluation determined that the implanted tissue could not be safely resected, so that the array was removed from the brain tissue prior to closure of the dura and the craniotomy. In three participants (P01, P02 and P03), the resected implantation region was formalin-fixed with the array *in situ* and processed further for histological analysis (hematoxylin eosin staining).

Cortical surfaces were reconstructed from individual participants' structural MRI using BrainSuite [52]. The implantation site was marked manually, guided by intraoperative neuronavigation data and photographic documentation. Individual MRI scans were then normalized to the MNI-152 template in SPM12 (Wellcome Center Human Neuroimaging). The macroanatomical cortical area corresponding to the implantation site was determined with the JuBrain SPM anatomy toolbox (Forschungszentrum Jülich).

Neurophysiological recordings

We recorded intraoperative neuronal data in eight awake participants. All eight participants underwent the same procedures before, during and after recordings. Extracellular voltage signals were acquired using either analog patient cable headstages in combination with a front-end amplifier (P04, P05, P06, P07 and P09) or digital Cereplex E128 headstages connected to digital hubs (P10, P11 and P13) as part of a 128-channel NSP system (NeuroPort Biopotential Signal Processing System, Blackrock Neurotech). Settings for signal amplification, filtering and digitization were identical in both setups (high-pass 0.3 Hz, low-pass 7.5 kHz, sampling rate 30 kHz, 16-bit resolution).

We did not find it necessary to switch between the two reference wires, both of which provided high-quality reference signals in all cases. However, particular attention was paid to achieving a strong ground connection via the pedestal. Long skull screws (6 mm) in combination with intermittent

irrigation of the pedestal's base where it contacted the skull produced the best results. Impedances were checked after array implantation and in most surgeries were initially higher than the upper bound of the normal range (80 kΩ for IrOx electrodes), but continued to normalize over the course of several tens of minutes. We attributed this to improving electrical conductivity at the pedestal-skull interface. Additional ground connections were not necessary and could even contaminate signals if placed badly (e.g. subdermal needles in the vicinity of musculature).

Behavioral task and stimuli

Six participants performed a delayed-match-to-number task during neuronal recording. MonkeyLogic 2 (NIMH) running on a dedicated PC was used for experimental control and behavioral data acquisition. Behavioral time stamps were transmitted to the NSP system for parallel logging of neuronal data and behavioral events.

We familiarized participants with the task ahead of the surgery and allowed them to complete multiple training trials. Participants viewed a 12" monitor positioned 40 - 50 cm in front of them. They were instructed to maintain eye fixation on a central white dot and pressed a button on a hand-held device to initiate a trial. Stimuli were presented on a centrally placed gray circular background subtending approx. 9,4 ° of visual angle. Following a 500 ms pre-sample period, a 150 ms sample stimulus was shown. In nonsymbolic trials, 2, 3, 4, 6, 7 or 8 randomly arranged black dots specified the corresponding numerosity. In symbolic trials, black Arabic numerals (Arial, 40 - 56 pt) were shown. The participants were required to memorize the sample number for 1,000 ms and compare it to the number of dots (in nonsymbolic trials) or the Arabic numeral (in symbolic trials) presented in a 1,000 ms test stimulus. If the quantities matched (50 % of trials), participants released the button (correct Match trial). If the quantities were different (50 % of trials), the participants continued to push the button until the matching quantity was presented in the subsequent image (correct Non-match trial). Match and non-match trials and nonsymbolic and symbolic trials were pseudo-randomly intermixed. New stimuli were generated for each participant and recording.

Behavioral performance

Behavioral tuning functions were used to describe the percentage of trials (y axis) for which a test stimulus (x axis, units of numerical distance to sample number) was judged as being equal in number to the sample. A numerical distance of 0 denotes match trials; the data point represents the percentage of correct trials. As the numerical distance increases, there is less confusion of the test with the sample number; the data points represent the percentage of error trials. Tuning curves were calculated separately for trials with nonsymbolic stimuli and for trials with symbolic stimuli.

686

687 **Spiking activity and single unit quality metrics**

688 Raw signals were filtered (250 Hz high-pass, 4-pole Butterworth), and spike waveforms were manually
689 separated from noise using Offline Sorter (Plexon). Signal-to-noise ratio (SNR) was calculated as

690 $$SNR = 20 * log_{10}(\frac{V_{PP}}{V_{RMS}})$$

691 where V_{pp} is the mean peak-to-peak spike amplitude of a given channel and V_{RMS} is the root-mean-
692 square (RMS) voltage

693 $$V_{RMS} = \sqrt{\frac{1}{N}\sum_{n=1}^{N} x_n^2}$$

694 with x_n being individual voltage values (Fig. 2D top). Spike SNR was calculated across the entire
695 recording session (Fig. 2D bottom) or in sliding windows (Fig. 2E; 60 s bins, 30 s steps).

696 Thresholded waveforms were manually sorted into clusters of single units (Offline Sorter). We
697 estimated the rate of false negatives (missed spikes) by fitting a gaussian to the distribution of spike
698 troughs (Fig. 3D). Autocorrelograms (Fig. 3E) were calculated by shifting a unit's spike train in steps
699 of 1 ms over a range of 1 to 25 ms. To determine the percentage of outlier spikes (Fig. 3G) [53], each
700 spike was considered as a point on a 2D plane spanned by the first two principal components that were
701 used for spike sorting. For each spike, the Mahalanobis distance to the corresponding cluster's average
702 waveform was calculated. A chi-square distribution was then fitted to the distribution of distances [54]. If
703 the likelihood of a given spike to belong to this distribution was lower than a fixed threshold (the inverse
704 of the total number of spikes in the given cluster), it was considered an outlier spike.

705

706 **Local field potentials and quality metrics**

707 Data was processed using the FieldTrip toolbox [55]. Raw signals were filtered (1.5 Hz high-pass, 1-pole
708 Butterworth; 250 Hz low-pass, 3-pole Butterworth), and line noise was removed (2-pole Butterworth
709 band-stop filters of ± 0.2 Hz at 50 Hz and harmonics). LFP traces were then visually inspected for large-
710 amplitude artefacts, which were excluded from further analysis.

711 Spectral transformation was performed with the additive superlet method [56]. SNR was calculated in
712 sliding windows (60 s bins, 30 s steps) and then averaged across windows for the session-SNR (Fig. 2H
713 bottom) or presented as time-resolved data (Fig. 2I). For each bin and channel, states of high and low
714 LFP activity were identified and used for signal and noise estimators, respectively (Fig. 2H top) [57,58].
715 High and low activity states were derived from the smoothed LFP amplitude envelope (100 ms

averaging window) obtained through complex Hilbert transform. Any timepoints of the smoothed envelope that fell outside of three standard deviations of its distribution were marked as artefacts and automatically assigned to the noise intervals. The mean of the smoothed envelope, excluding artefact timepoints, served as a detection threshold for high activity states. Thus, epochs of the smoothed envelope surpassing the threshold for at least 400 ms were considered states of high activity, whereas all others counted as low activity states [57]. SNR was then calculated as

$$SNR = 20 * log_{10}\left(\frac{\frac{1}{N_{High}}\sum_{n=1}^{n=N_{High}} \mathrm{PP(High}_n)}{\frac{1}{N_{Low}}\sum_{n=1}^{n=N_{Low}} \mathrm{RMS(Low}_n)}\right),$$

where N_{High} and N_{Low} are the number of high or low activity states, respectively, PP (peak-to-peak amplitude) is the difference between the highest and lowest voltage reading during a given high activity state and RMS is

$$RMS = \sqrt{\frac{1}{N}\sum_{n=1}^{N} x_n^2}$$

with x_n being individual voltage values of an interval of low activity.

The Power-Spectral-Density (PSD) was calculated using Welch's method. Specifically, across five minutes of the recording (0:30 to 5:30 min), modified periodograms in 3-s bins (smoothed using a Hamming window) with 50 % overlap were obtained by Fast Fourier transform (FFT) and averaged [59].

Travelling waves

We assumed the simplest form of travelling waves, a planar wave with linear phase gradient [33]. First, zero-phase bandpass filters (± 1.5 Hz) were applied for each frequency of interest (theta: 6 to 9 Hz; beta: 15 to 35 Hz, in steps of 1 Hz) and every channel. We then applied the Hilbert transform (Hlb) to the resulting signal (V) to obtain the instantaneous phase φ(x,y,t) of each time point (t) and channel position (x,y)

$$V(t,x,y) + iHlb[V(x,y,t)]) = a(x,y,t)e^{i\varphi(x,y,t)}$$

Instantaneous phases were unwrapped and de-noised [60]. Next, a plane model was fit to the data using linear regression. The plane was modelled as

$$\varphi(t,x,y) = b_x(t)x + b_y(t)y + \varphi_c(t)$$

With $b_x(t)$ and $b_y(t)$ being the slope of the plane in the x-direction and y-direction at time t, respectively, and $\varphi_c(t)$ the constant phase shift at time t. The model's goodness-of-fit was expressed by the Phase-

744 Gradient Directionality (PGD) [33]. PGD is the Pearson correlation between the predicted and actual phase
745 and is given by

746 $$PGD(t) = \frac{\sum_i^{N_{ch}}((\varphi(t,x_i,y_i) - \overline{\varphi}(t))(\hat{\varphi}(t,x_i,y_i) - \bar{\hat{\varphi}}(t)))}{\sqrt{\sum_i^{N_{ch}}(\varphi(t,x_i,y_i) - \overline{\varphi}(t))^2 \sum_i^{N_{ch}}(\hat{\varphi}(t,x_i,y_i) - \bar{\hat{\varphi}}(t))^2}}$$

747 with $\overline{\varphi}$ being the average and $\hat{\varphi}$ the predicted phase.

748 When zero fell outside the 99th percentile of at least one of the coefficients' b_x or b_y confidence intervals
749 and PGD was bigger than 0.5, a moment in time was considered for travelling wave-like activity [33]. The
750 direction [60] and speed [33] of the travelling wave-like activity were then calculated as

751 $$direction(t) = arctan(\frac{b_y(t)}{b_x(t)})$$

752 $$speed(t) = \frac{\omega(t)}{\sqrt{b_x(t)^2 + b_y(t)^2}}$$

753 with ω(t) being the instantaneous angular velocity.

754 A travelling wave epoch was defined by non-zero slopes in the phase gradient with a PGD > 0.5 for a
755 minimum length of 5 ms and a maximal average change in direction of 3 deg/ms. Polar distributions
756 (10° bins) that showed a second peak reaching 25 % or more of the distribution's modal value and that
757 significantly differed from uniformity (Hodges-Ajne test) were considered bidirectional.

758

759 **Neuronal information**

760 To quantify the information about sample number and notation that was carried by a neuron's spiking
761 rate, we used the ω^2 percent explained variance measure [42]. ω^2 reflects how much of the variance in a
762 neuron's firing rate can be explained by a given factor. It was calculated in sliding windows (100 ms
763 bins, 20 ms steps) using

764 $$\omega^2 = \frac{SS_{Groups} - df * MSE}{SS_{Total} + MSE}$$

765 where the individual terms are derived from a two-way categorical ANOVA: SS_{Groups} denotes the sum-
766 of-squares between groups (numbers), SS_{Total} the total sum-of-squares, *df* the degrees of freedom, and
767 *MSE* the mean squared error. The number of trials in each group was balanced. Balancing was
768 accomplished by stratifying the number of trials in each group to a common value: A random subset of
769 trials was drawn (equal to the minimum trial number across groups) and the statistic was calculated.
770 This process was repeated 25 times, and the overall statistic was taken to be the mean of the stratified
771 values. Significance thresholds were determined by randomly shuffling the association between spiking

rates and trial type (number and notation) during the pre-sample epoch (500 ms). This process was repeated 1,000 times, and the significance threshold was set to the 99th percentile of the cumulative distribution ($p < 0.01$).

For task information contained in LFPs, we calculated ω^2 in sliding windows (5 ms bins, 0.25 ms steps, 1 Hz bins, 1 Hz steps) using spectral power derived as described above.

Linear discriminant analysis

Unsorted (multi-unit) spikes were aggregated into firing rates using Gaussian windows with 50 ms sigma and 50 ms step size. Trials were grouped for small numbers (2, 3, 4) and large numbers (6, 7, 8). A procedure of 7-fold cross validation with 7 repetitions was used, resulting in 49 training and testing set pairs. At every time step, an LDA decoder (Scikit-learn package in Python) was trained on the activity of the current time step in the training set and tested on all the time steps in the testing set in order to investigate how well the code generalizes across different timesteps. Decoding accuracy is given as the average across test trials. LDA finds the component that maximizes the Mahalanobis distance between the centroids of small and large number classes. The algorithm assumes equal within-class covariance in different classes. Shrinkage of the empirical covariance matrix was applied by averaging the empirical covariance matrix with a diagonal matrix, discounting the spurious covariation between units. The amount of shrinkage was determined by the Ledoit-Wolf lemma [61].

Dimensionality controlled version of LDA decoding was done by projecting data on the top n dimensions that preserves the Mahalanobis distance and finding the closest class centroid (all classes were used without grouping) to the test population response in this subspace. Peak accuracy was found by first smoothing the cross-temporal decoding accuracy matrix with Gaussian filter (100ms, or 2 time steps sigma) then finding the highest accuracy value. This prevents transient noise from dominating the result. Repeated measure T test was done over 150 repetitions of 7-fold cross validation between all pairs of peak accuracy with different dimensions (with Bonferroni correction).

QUANTIFICATION AND STATISTICAL ANALYSIS

All data analysis was performed with MATLAB (Mathworks) and Python.

801 **KEY RESOURCES TABLE**

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
MATLAB	MathWorks	RRID: SCR_001622
Python Programming Language	Python website	RRID: SCR_008394
MonkeyLogic 2	NIMH	N/A
Offline Sorter	Plexon	RRID: SCR_000012
FieldTrip toolbox	FieldTrip website	RRID: SCR_004849
BrainSuite	BrainSuite website	RRID: SCR_006623
SPM	SPM website	RRID: SCR_007037
JuBrain SPM anatomy toolbox	fz-juelich website	N/A
Other		
Microelectrode arrays	Blackrock Neurotech	N/A

802

803 **References**

804 1. Cash, S.S., and Hochberg, L.R. (2015). The emergence of single neurons in clinical neurology.
805 Neuron *86*, 79-91. 10.1016/j.neuron.2015.03.058.
806 2. Fu, Z., Beam, D., Chung, J.M., Reed, C.M., Mamelak, A.N., Adolphs, R., and Rutishauser, U.
807 (2022). The geometry of domain-general performance monitoring in the human medial frontal
808 cortex. Science *376*, eabm9922. 10.1126/science.abm9922.
809 3. Minxha, J., Adolphs, R., Fusi, S., Mamelak, A.N., and Rutishauser, U. (2020). Flexible
810 recruitment of memory-based choice representations by the human medial frontal cortex.
811 Science *368*, eaba3313. 10.1126/science.aba3313.
812 4. Kaminski, J., Sullivan, S., Chung, J.M., Ross, I.B., Mamelak, A.N., and Rutishauser, U. (2017).
813 Persistently active neurons in human medial frontal and medial temporal lobe support working
814 memory. Nat Neurosci *20*, 590-601. 10.1038/nn.4509.
815 5. Rutishauser, U., Ross, I.B., Mamelak, A.N., and Schuman, E.M. (2010). Human memory
816 strength is predicted by theta-frequency phase-locking of single neurons. Nature *464*, 903-907.
817 10.1038/nature08860.
818 6. Kutter, E.F., Bostroem, J., Elger, C.E., Mormann, F., and Nieder, A. (2018). Single Neurons in
819 the Human Brain Encode Numbers. Neuron *100*, 753-761 e754. 10.1016/j.neuron.2018.08.036.
820 7. Kornblith, S., Quian Quiroga, R., Koch, C., Fried, I., and Mormann, F. (2017). Persistent
821 Single-Neuron Activity during Working Memory in the Human Medial Temporal Lobe. Curr
822 Biol *27*, 1026-1032. 10.1016/j.cub.2017.02.013.
823 8. Sheth, S.A., Mian, M.K., Patel, S.R., Asaad, W.F., Williams, Z.M., Dougherty, D.D., Bush, G.,
824 and Eskandar, E.N. (2012). Human dorsal anterior cingulate cortex neurons mediate ongoing
825 behavioural adaptation. Nature *488*, 218-221. 10.1038/nature11239.
826 9. Jamali, M., Grannan, B.L., Fedorenko, E., Saxe, R., Baez-Mendoza, R., and Williams, Z.M.
827 (2021). Single-neuronal predictions of others' beliefs in humans. Nature *591*, 610-614.
828 10.1038/s41586-021-03184-0.
829 10. Jamali, M., Grannan, B., Haroush, K., Moses, Z.B., Eskandar, E.N., Herrington, T., Patel, S.,
830 and Williams, Z.M. (2019). Dorsolateral prefrontal neurons mediate subjective decisions and
831 their variation in humans. Nat Neurosci *22*, 1010-1020. 10.1038/s41593-019-0378-3.
832 11. Zaghloul, K.A., Blanco, J.A., Weidemann, C.T., McGill, K., Jaggi, J.L., Baltuch, G.H., and
833 Kahana, M.J. (2009). Human substantia nigra neurons encode unexpected financial rewards.
834 Science *323*, 1496-1499. 10.1126/science.1167342.
835 12. Paulk, A.C., Kfir, Y., Khanna, A.R., Mustroph, M.L., Trautmann, E.M., Soper, D.J., Stavisky,
836 S.D., Welkenhuysen, M., Dutta, B., Shenoy, K.V., et al. (2022). Large-scale neural recordings
837 with single neuron resolution using Neuropixels probes in human cortex. Nat Neurosci *25*, 252-
838 263. 10.1038/s41593-021-00997-0.
839 13. Chung, J.E., Sellers, K.K., Leonard, M.K., Gwilliams, L., Xu, D., Dougherty, M.E., Kharazia,
840 V., Metzger, S.L., Welkenhuysen, M., Dutta, B., and Chang, E.F. (2022). High-density single-
841 unit human cortical recordings using the Neuropixels probe. Neuron *110*, 2409-2421 e2403.
842 10.1016/j.neuron.2022.05.007.
843 14. Sanai, N., Mirzadeh, Z., and Berger, M.S. (2008). Functional outcome after language mapping
844 for glioma resection. N Engl J Med *358*, 18-27. 10.1056/NEJMoa067819.
845 15. Mandonnet, E., and Herbet, G. (2021). Intraoperative Mapping of Cognitive Networks
846 (Springer).
847 16. Chen, X., Wang, F., Fernandez, E., and Roelfsema, P.R. (2020). Shape perception via a high-
848 channel-count neuroprosthesis in monkey visual cortex. Science *370*, 1191-1196.
849 10.1126/science.abd7435.
850 17. Mitz, A.R., Bartolo, R., Saunders, R.C., Browning, P.G., Talbot, T., and Averbeck, B.B. (2017).
851 High channel count single-unit recordings from nonhuman primate frontal cortex. J Neurosci
852 Methods *289*, 39-47. 10.1016/j.jneumeth.2017.07.001.
853 18. Schevon, C.A., Tobochnik, S., Eissa, T., Merricks, E., Gill, B., Parrish, R.R., Bateman, L.M.,
854 McKhann, G.M., Jr., Emerson, R.G., and Trevelyan, A.J. (2019). Multiscale recordings reveal
855 the dynamic spatial structure of human seizures. Neurobiol Dis *127*, 303-311.
856 10.1016/j.nbd.2019.03.015.

19. Truccolo, W., Donoghue, J.A., Hochberg, L.R., Eskandar, E.N., Madsen, J.R., Anderson, W.S., Brown, E.N., Halgren, E., and Cash, S.S. (2011). Single-neuron dynamics in human focal epilepsy. Nat Neurosci *14*, 635-641. 10.1038/nn.2782.
20. Willett, F.R., Avansino, D.T., Hochberg, L.R., Henderson, J.M., and Shenoy, K.V. (2021). High-performance brain-to-text communication via handwriting. Nature *593*, 249-254. 10.1038/s41586-021-03506-2.
21. Pandarinath, C., Nuyujukian, P., Blabe, C.H., Sorice, B.L., Saab, J., Willett, F.R., Hochberg, L.R., Shenoy, K.V., and Henderson, J.M. (2017). High performance communication by people with paralysis using an intracortical brain-computer interface. Elife *6* (*e18554*). 10.7554/eLife.18554.
22. Hochberg, L.R., Serruya, M.D., Friehs, G.M., Mukand, J.A., Saleh, M., Caplan, A.H., Branner, A., Chen, D., Penn, R.D., and Donoghue, J.P. (2006). Neuronal ensemble control of prosthetic devices by a human with tetraplegia. Nature *442*, 164-171. 10.1038/nature04970.
23. Aflalo, T., Kellis, S., Klaes, C., Lee, B., Shi, Y., Pejsa, K., Shanfield, K., Hayes-Jackson, S., Aisen, M., Heck, C., et al. (2015). Decoding motor imagery from the posterior parietal cortex of a tetraplegic human. Science *348*, 906-910. 10.1126/science.aaa5417.
24. Fernandez, E., Alfaro, A., Soto-Sanchez, C., Gonzalez-Lopez, P., Lozano, A.M., Pena, S., Grima, M.D., Rodil, A., Gomez, B., Chen, X., et al. (2021). Visual percepts evoked with an intracortical 96-channel microelectrode array inserted in human occipital cortex. J Clin Invest *131* (*e151331*). 10.1172/JCI151331.
25. Flesher, S.N., Collinger, J.L., Foldes, S.T., Weiss, J.M., Downey, J.E., Tyler-Kabara, E.C., Bensmaia, S.J., Schwartz, A.B., Boninger, M.L., and Gaunt, R.A. (2016). Intracortical microstimulation of human somatosensory cortex. Sci Transl Med *8*, 361ra141. 10.1126/scitranslmed.aaf8083.
26. Fernandez, E., Greger, B., House, P.A., Aranda, I., Botella, C., Albisua, J., Soto-Sanchez, C., Alfaro, A., and Normann, R.A. (2014). Acute human brain responses to intracortical microelectrode arrays: challenges and future prospects. Front Neuroeng *7*, 24. 10.3389/fneng.2014.00024.
27. House, P.A., MacDonald, J.D., Tresco, P.A., and Normann, R.A. (2006). Acute microelectrode array implantation into human neocortex: preliminary technique and histological considerations. Neurosurg Focus *20*, E4. 10.3171/foc.2006.20.5.5.
28. Colachis, S.C.t., Dunlap, C.F., Annetta, N.V., Tamrakar, S.M., Bockbrader, M.A., and Friedenberg, D.A. (2021). Long-term intracortical microelectrode array performance in a human: a 5 year retrospective analysis. J Neural Eng *18* (*0460d7*). 10.1088/1741-2552/ac1add.
29. Fries, P. (2015). Rhythms for Cognition: Communication through Coherence. Neuron *88*, 220-235. 10.1016/j.neuron.2015.09.034.
30. Bhattacharya, S., Brincat, S.L., Lundqvist, M., and Miller, E.K. (2022). Traveling waves in the prefrontal cortex during working memory. PLoS Comput Biol *18*, e1009827. 10.1371/journal.pcbi.1009827.
31. Sato, T.K., Nauhaus, I., and Carandini, M. (2012). Traveling waves in visual cortex. Neuron *75*, 218-229. 10.1016/j.neuron.2012.06.029.
32. Takahashi, K., Saleh, M., Penn, R.D., and Hatsopoulos, N.G. (2011). Propagating waves in human motor cortex. Front Hum Neurosci *5*, 40. 10.3389/fnhum.2011.00040.
33. Rubino, D., Robbins, K.A., and Hatsopoulos, N.G. (2006). Propagating waves mediate information transfer in the motor cortex. Nat Neurosci *9*, 1549-1557. 10.1038/nn1802.
34. Muller, L., Chavane, F., Reynolds, J., and Sejnowski, T.J. (2018). Cortical travelling waves: mechanisms and computational principles. Nat Rev Neurosci *19*, 255-268. 10.1038/nrn.2018.20.
35. Das, A., Myers, J., Mathura, R., Shofty, B., Metzger, B.A., Bijanki, K., Wu, C., Jacobs, J., and Sheth, S.A. (2022). Spontaneous neuronal oscillations in the human insula are hierarchically organized traveling waves. Elife *11* (*e76702*). 10.7554/eLife.76702.
36. Zhang, H., Watrous, A.J., Patel, A., and Jacobs, J. (2018). Theta and Alpha Oscillations Are Traveling Waves in the Human Neocortex. Neuron *98*, 1269-1281 e1264. 10.1016/j.neuron.2018.05.019.

911 37. Zhang, H., and Jacobs, J. (2015). Traveling Theta Waves in the Human Hippocampus. J
912 Neurosci *35*, 12477-12487. 10.1523/JNEUROSCI.5102-14.2015.
913 38. Liou, J.Y., Smith, E.H., Bateman, L.M., McKhann, G.M., Goodman, R.R., Greger, B., Davis,
914 T.S., Kellis, S.S., House, P.A., and Schevon, C.A. (2017). Multivariate regression methods for
915 estimating velocity of ictal discharges from human microelectrode recordings. J Neural Eng *14*,
916 044001. 10.1088/1741-2552/aa68a6.
917 39. Smith, E.H., Liou, J.Y., Davis, T.S., Merricks, E.M., Kellis, S.S., Weiss, S.A., Greger, B.,
918 House, P.A., McKhann, G.M., 2nd, Goodman, R.R., et al. (2016). The ictal wavefront is the
919 spatiotemporal source of discharges during spontaneous human seizures. Nat Commun *7*,
920 11098. 10.1038/ncomms11098.
921 40. Nieder, A. (2016). The neuronal code for number. Nat Rev Neurosci *17*, 366-382.
922 10.1038/nrn.2016.40.
923 41. Jacob, S.N., Hahnke, D., and Nieder, A. (2018). Structuring of Abstract Working Memory
924 Content by Fronto-parietal Synchrony in Primate Cortex. Neuron *99*, 588-597 e585.
925 10.1016/j.neuron.2018.07.025.
926 42. Jacob, S.N., and Nieder, A. (2014). Complementary roles for primate frontal and parietal cortex
927 in guarding working memory from distractor stimuli. Neuron *83*, 226-237.
928 10.1016/j.neuron.2014.05.009.
929 43. Nieder, A., Diester, I., and Tudusciuc, O. (2006). Temporal and spatial enumeration processes
930 in the primate parietal cortex. Science *313*, 1431-1435. 10.1126/science.1130308.
931 44. Rutishauser, U., Aflalo, T., Rosario, E.R., Pouratian, N., and Andersen, R.A. (2018). Single-
932 Neuron Representation of Memory Strength and Recognition Confidence in Left Human
933 Posterior Parietal Cortex. Neuron *97*, 209-220 e203. 10.1016/j.neuron.2017.11.029.
934 45. Piazza, M., Fumarola, A., Chinello, A., and Melcher, D. (2011). Subitizing reflects visuo-
935 spatial object individuation capacity. Cognition *121*, 147-153.
936 10.1016/j.cognition.2011.05.007.
937 46. Cheyette, S.J., and Piantadosi, S.T. (2020). A unified account of numerosity perception. Nat
938 Hum Behav *4*, 1265-1272. 10.1038/s41562-020-00946-0.
939 47. Tong, A.P.S., Vaz, A.P., Wittig, J.H., Inati, S.K., and Zaghloul, K.A. (2021). Ripples reflect a
940 spectrum of synchronous spiking activity in human anterior temporal lobe. Elife *10* (*e68401*).
941 10.7554/eLife.68401.
942 48. Vaz, A.P., Wittig, J.H., Jr., Inati, S.K., and Zaghloul, K.A. (2020). Replay of cortical spiking
943 sequences during human memory retrieval. Science *367*, 1131-1134. 10.1126/science.aba0672.
944 49. Chiang, C.H., Won, S.M., Orsborn, A.L., Yu, K.J., Trumpis, M., Bent, B., Wang, C., Xue, Y.,
945 Min, S., Woods, V., et al. (2020). Development of a neural interface for high-definition, long-
946 term recording in rodents and nonhuman primates. Sci Transl Med *12* (*eaay4682*).
947 10.1126/scitranslmed.aay4682.
948 50. Loomba, S., Straehle, J., Gangadharan, V., Heike, N., Khalifa, A., Motta, A., Ju, N., Sievers,
949 M., Gempt, J., Meyer, H.S., and Helmstaedter, M. (2022). Connectomic comparison of mouse
950 and human cortex. Science *377*, eabo0924. 10.1126/science.abo0924.
951 51. Zaghloul, K.A., Weidemann, C.T., Lega, B.C., Jaggi, J.L., Baltuch, G.H., and Kahana, M.J.
952 (2012). Neuronal activity in the human subthalamic nucleus encodes decision conflict during
953 action selection. J Neurosci *32*, 2453-2460. 10.1523/JNEUROSCI.5815-11.2012.
954 52. Shattuck, D.W., and Leahy, R.M. (2002). BrainSuite: An automated cortical surface
955 identification tool. Medical Image Analysis *6*, 129-142. 10.1016/s1361-8415(02)00054-3.
956 53. Meirhaeghe, N., Sohn, H., and Jazayeri, M. (2021). A precise and adaptive neural mechanism
957 for predictive temporal processing in the frontal cortex. Neuron *109*, 2995-3011 e2995.
958 10.1016/j.neuron.2021.08.025.
959 54. Hill, D.N., Mehta, S.B., and Kleinfeld, D. (2011). Quality metrics to accompany spike sorting
960 of extracellular signals. J Neurosci *31*, 8699-8705. 10.1523/JNEUROSCI.0971-11.2011.
961 55. Oostenveld, R., Fries, P., Maris, E., and Schoffelen, J.M. (2011). FieldTrip: Open source
962 software for advanced analysis of MEG, EEG, and invasive electrophysiological data. Comput
963 Intell Neurosci *2011*, 156869. 10.1155/2011/156869.
964 56. Moca, V.V., Barzan, H., Nagy-Dabacan, A., and Muresan, R.C. (2021). Time-frequency super-
965 resolution with superlets. Nat Commun *12*, 337. 10.1038/s41467-020-20539-9.

966 57. Compte, A., Reig, R., Descalzo, V.F., Harvey, M.A., Puccini, G.D., and Sanchez-Vives, M.V.
967 (2008). Spontaneous high-frequency (10-80 Hz) oscillations during up states in the cerebral
968 cortex in vitro. J Neurosci *28*, 13828-13844. 10.1523/JNEUROSCI.2684-08.2008.
969 58. Suarez-Perez, A., Gabriel, G., Rebollo, B., Illa, X., Guimera-Brunet, A., Hernandez-Ferrer, J.,
970 Martinez, M.T., Villa, R., and Sanchez-Vives, M.V. (2018). Quantification of Signal-to-Noise
971 Ratio in Cerebral Cortex Recordings Using Flexible MEAs With Co-localized Platinum Black,
972 Carbon Nanotubes, and Gold Electrodes. Front Neurosci *12*, 862. 10.3389/fnins.2018.00862.
973 59. Zilio, F., Gomez-Pilar, J., Cao, S., Zhang, J., Zang, D., Qi, Z., Tan, J., Hiromi, T., Wu, X.,
974 Fogel, S., et al. (2021). Are intrinsic neural timescales related to sensory processing? Evidence
975 from abnormal behavioral states. Neuroimage *226*, 117579.
976 10.1016/j.neuroimage.2020.117579.
977 60. Woods, B. (2011). Spatio-temporal Patterns in Multi-Electrode Array Local Field Potential
978 Recordings. arXiv *1501.00230v1*.
979 61. Ledoit, O., and Wolf, M. (2004). A well-conditioned estimator for large-dimensional
980 covariance matrices. Journal of Multivariate Analysis *88*, 365-411. 10.1016/s0047-
981 259x(03)00096-4.

2.2 Neuronal implementation of representational geometry in prefrontal cortex

Manuscript 2: The neuronal implementation of representational geometry in primate prefrontal cortex

Authors: Xiao-Xiong Lin, Andreas Nieder, Simon N. Jacob

Author contributions:

X.-X.L. conceived the study and performed the analyses with contributions from S.N.J. A.N. and S.N.J. designed the experiments and collected the data. X.-X.L. and S.N.J. wrote the manuscript and prepared the figures. All authors edited the manuscript.

FRONT MATTER

Title

- The neuronal implementation of representational geometry in primate prefrontal cortex
- Neuronal implementation of representational geometry

Authors

Xiao-Xiong Lin[1,2], Andreas Nieder[3], Simon N. Jacob[1]*

Affiliations

[1] Translational Neurotechnology Laboratory, Department of Neurosurgery, Klinikum rechts der Isar, Technical University of Munich, Germany
[2] Graduate School of Systemic Neurosciences, Ludwig-Maximilians-University Munich, Germany
[3] Animal Physiology, University of Tübingen, Germany

* Correspondence: simon.jacob@tum.de

Abstract

Modern neuroscience has seen the rise of a population-doctrine that represents cognitive variables using geometrical structures in activity space. Representational geometry does not, however, account for how individual neurons implement these representations. Here, leveraging the principle of sparse coding, we present a framework to dissect representational geometry into biologically interpretable components that retain links to single neurons. Applied to extracellular recordings from the primate prefrontal cortex in a working memory task with interference, the identified components revealed disentangled and sequential memory representations including the recovery of memory content after distraction, signals hidden to conventional analyses. Each component was contributed by small subpopulations of neurons with distinct electrophysiological properties and response dynamics. Modelling showed that such sparse implementations are supported by recurrently connected circuits as in prefrontal cortex. The perspective of neuronal implementation links representational geometries to their cellular constituents, providing mechanistic insights into how neural systems encode and process information.

Teaser

Geometrical structures that describe working memory activity in neuronal populations are dissected into neuron-specific components

MAIN TEXT

Introduction

For decades, the dominant approach to understanding neural systems has been to characterize the role and contributions of individual neurons. In a recent paradigm shift, the concept of high-dimensional activity spaces that represent cognitive and other variables at the level of neuronal populations has taken the center stage and sidelined the single-neuron perspective (*1*, *2*). These population representations capture multi-neuron activity in different behavioral task conditions in the form of geometrical structures (*3*, *4*). Representational geometry provides a complete description of the information encoded by and processed in a neuronal population. It does not, however, account for how individual neurons – the nuts and bolts of brain processing – give rise to the representations and the operations performed on them (*5*) because there is no direct connection between informational representation and biological implementation at the cellular and circuit level.

In constructing representational geometries, the choice of coordinate system, that is the set of components that capture the population activity, is arbitrary. The question then arises what the most meaningful coordinate system is to represent the data. In principal component analysis (PCA), a widely used method for dimensionality reduction, the principal components (PCs) capture the neuronal activity's variance, but they are not designed to yield biologically interpretable aspects of the representational geometry. Identifying coordinate systems that are rooted in biology is particularly relevant in association cortices where neurons often have mixed-selective responses that are not easily interpreted as the representation of any single stimulus or task variable alone (*3*, *6*). Neuronal signals in association cortices also show complex temporal dynamics and task-dependent modulations that reflect distinct sensory and memory processing stages (*7–9*). During working memory, for example, behaviorally relevant target items are maintained in online storage and must be protected against interfering distractors (*8*, *9*). However, depending on which coordinate system is used to express the representational geometry, the same task-related neuronal activity could be interpreted in one of two ways: either as components representing the target in each task epoch individually, suggesting a memory mechanism built on sequential relay of target information among components (*10*), or, alternatively, as components that represent the target across task epochs, suggesting a memory mechanism of continuous representation of target information by the same components (*11*).

The biological implementation of representations points to how components are accessed and information is communicated. Unlike the units in neuronal network models, *in vivo* neurons are subject to anatomical and physiological constraints. There are approximately 10^{10} neurons in the human brain and 10^9 in a hypothetical functional module such as the dorsolateral prefrontal cortex (PFC) (*12*, *13*). A pyramidal cortical neuron has on the order of 10^4 dendritic spines (*14*). Thus, given the disproportion between the low number of possible connections and the large number of potentially informative neurons, a neuron downstream of the PFC can only 'read out' from a small fraction of neurons in this region. That is, it cannot access arbitrary components of the representational geometry. Instead, it would be more efficient and biologically plausible to read out

components that a few neurons predominantly contribute to, that is the components with a sparse neuronal implementation.

Here, we present a framework that exploits the structure in the representational geometry's neuronal implementation. We show that this approach yields unbiased components of population activity that retain links to individual neurons. We performed data dimensionality reduction on extracellular multi-channel recordings from the non-human primate PFC by leveraging sparsity constraints in order to identify components that are contributed mainly by small subpopulations of strongly coding neurons (sparse component analysis, SCA) (*15*, *16*). We found that the activities on these components nontrivially matched the working memory task sequence performed by the animals, revealing separate sensory and memory components including a previously hidden component, namely the recovery of memory content after distraction. Notably, each component was made up of non-overlapping subpopulations of neurons with distinct electrophysiological properties and temporal dynamics. Finally, neuronal network modelling showed that recurrent connectivity as in the PFC favors such sparse implementations over non-structured Gaussian implementations. The framework and findings presented here bridge the gap between the single-neuron doctrine and the neuronal population doctrine (*1*, *2*) and establish the perspective of neuronal implementation as an important complement to representational geometry.

Results

Different neuronal implementations may underlie the same representational geometry

Representational geometry abstracts the information coded by a population of neurons from their individual tuning profiles (*5*). It specifies the pairwise distances between task-related collective neuronal responses, but no longer reflects the exact pattern of firing rates. This approach defines a stimulus-representing subspace. To illustrate, the representations for two stimuli A and B in PC space separate, rotate and collapse back to the origin (**Fig. 1a**).

The same stimulus-representing subspace can be defined with arbitrary sets of components. Components can be chosen to capture specific aspects of the representation, e.g., to continuously distinguish between stimuli (**Fig. 1b**), or to distinguish between stimuli at different time points (**Fig. 1c**). Note that in the former example, the components align with the PCs, while in the latter they do not. Various studies have followed this approach, selecting the components e.g. such that they express representations sequentially (*17*) or such that they each correspond to a particular task variable of interest (*18*, *19*).

Neuronal activity can be reconstructed by the weighted sum of components. Every neuron has a set of weights quantifying its relation to the different components, i.e. its loadings on the components. The loadings of neurons on the PCs visualize their positions in implementation space (**Fig. 1d-f**), where the loadings along any axis correspond to a component in representation space with the same orientation (**Fig. 1a-c**). The structure in the implementation space, i.e., the distribution of loadings across neurons, can be exploited to identify a unique, non-arbitrary set of components that

emphasizes biological plausibility of stimulus coding over enforcing possibly unjustified priors.

Representational geometry is invariant to the rotation of neuronal coordinates (*20*). Different neuronal implementations may therefore underlie the same representational geometry. We first consider the scenario of a Gaussian (dense) distribution of loadings (**Fig. 1d**), where the standardized moments (e.g., skewness and kurtosis) are constant, meaning there are no differences in these distributional statistics across axis orientations. We define the sparsity index (SI; **Fig. 1d**, top inset) to denote the sparsity of the implementation along a given axis. SI is proportional to a distribution's kurtosis. If SI is constant across axis orientations, neurons do not preferentially align to any axes.

Next, we consider a sparse distribution (**Fig. 1e**). Most neurons lie around the origin of the coordinate system. However, because SI is not constant (**Fig. 1e**, top inset), we can find the sparse components that strongly coding neurons align to. In the present case, these sparse axes correspond to the components in representational space that code the difference between stimulus A and B continuously (with one of the components reversing between epochs; compare **Fig. 1e** with **Fig. 1b**). Importantly, sparse distributions can exist for arbitrary axis orientations. For example, strongly coding neurons could align to the components that sequentially represent the stimulus information at time point 1 and time point 2 (compare **Fig. 1f** with **Fig. 1c**).

Although both scenarios are characterized by sparse neuronal implementations, we note that they have fundamentally different implications for readout, lending particular importance to the positioning of sparse axes orientations. Continuous readout (**Fig. 1b** and **e**, component 1) is stable, but not optimized for either time point 1 or time point 2, whereas sequential readouts (**Fig. 1c** and **1f**) are more precise at the respective time points, but not stable across time points.

In summary, the perspective of neuronal implementation offers a way to connect representational geometries to their cellular constituents, revealing mechanistic insights into how a neural system encodes, processes and relays information.

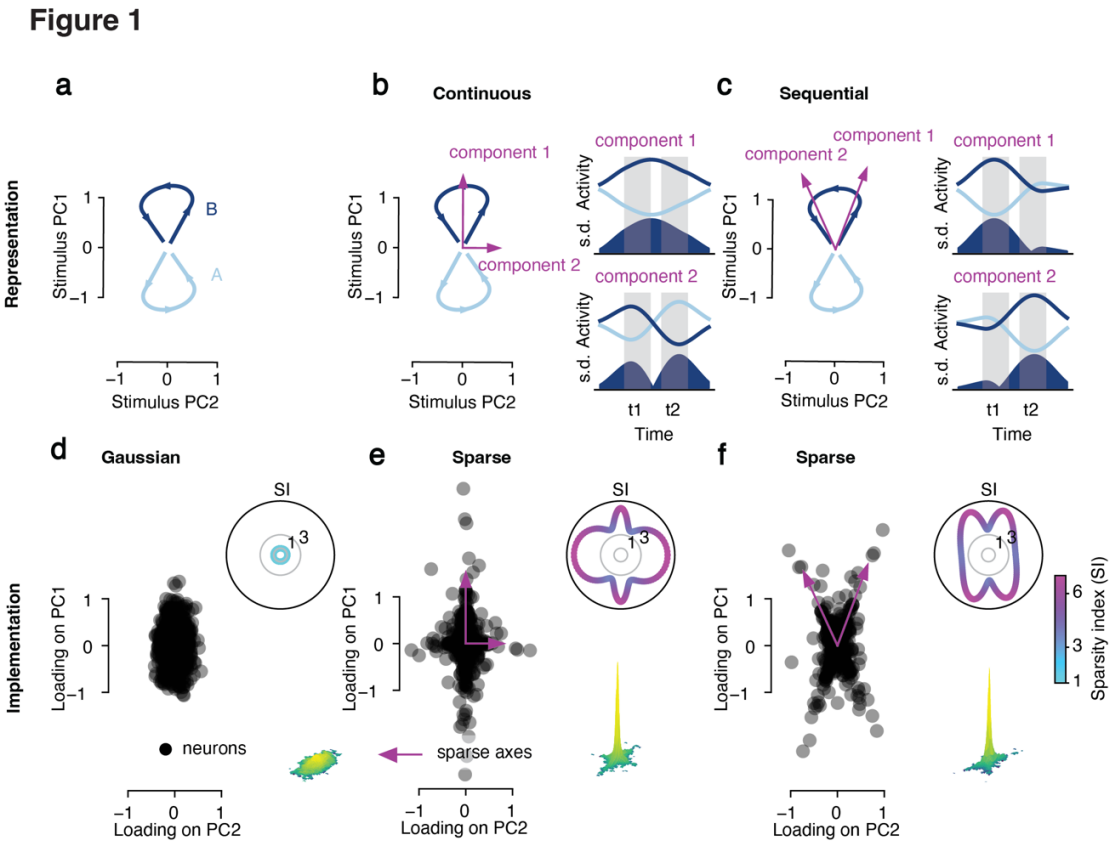


Fig. 1. Different neuronal implementations of the same representational geometry

*(**a**) Representational geometry for two trials with stimuli A and B on the plane specified by stimulus PC1 and PC2. Time runs along the individual trajectories. (**b**) Left: example pair of components that express the representational geometry (magenta arrows). Right: activities on the corresponding components and standard deviation (s.d.) across components as a measure of amount of information carried by them. Components are aligned with the PCs. (**c**) Same layout as in (b) for a non-aligned pair of components. (**d-f**) Neuronal implementation underlying the representational geometry in (a-c), specified by the distribution of neuronal loadings on the stimulus PCs. Insets: sparsity index (SI) of all axis orientations in the space spanned by PC1 and PC2. Axes with high SI (sparse axes, magenta arrows) in (e) and (f) correspond to the components 1 and 2 in (b) and (c), respectively.*

The neuronal implementation of working memory

With this framework, we now examine neuronal implementation of working memory, a core cognitive function for online maintenance and manipulation of information in the absence of sensory inputs. Extracellular multi-channel recordings were performed in the lateral PFC of two monkeys trained on a delayed-match-to-numerosity task, requiring them to memorize the number of dots (i.e., numerosity) in a visually presented sample and resist an interfering distracting numerosity (*9*) (**Fig. 2a**). A total of 467 single units recorded across 78 sessions were included in the analysis. Spike rates were binned,

averaged across conditions of the same type and demixed into their constituent parts (**Fig. 2b**) (*21*). Because the task design was balanced (i.e., all sample-distractor combinations were included), the different task variables were statistically independent of each other. Demixing therefore allowed to isolate and analyze signal components that would otherwise be overshadowed by signals that dominate the raw firing rates. Across neurons, the neuronal activities coding for trial time, sample numerosity, distractor numerosity and the sample-distractor interaction accounted for 72.7 %, 8.7 %, 5.8 % and 12.9 % of the total variance, respectively (**Fig. 2b**).

We first focused on the representation of the sample numerosity throughout the trial, the crucial function for completing the task (**Fig. 2c**). In PC space, the representations of different numerosities (1 and 4 visualized here) started to separate, marking an increase of the information during sample presentation. Then the representations rotated and returned to the origin. Similar representational changes have been reported previously (*10*, *22*, *23*).

The distribution of loadings of individual neurons onto the first three PCs was highly non-Gaussian ($p < 0.001$; Henze-Zirkler multivariate normality test; **Fig. 2d**). Accordingly, the sparsity index (SI) was not uniform across all axis orientations (**Fig. 2d**). Using sparse component analysis (SCA) that identifies components with sparse distributions of neuronal loadings (sparse components, SCs), we found three SCs that optimally decomposed the sample numerosities' representational geometry. The SCs displayed temporally well-defined active periods that matched the task structure and tiled the duration of a trial (**Fig. 2e**). Intuitively, they correspond to components for sensory encoding, memory maintenance and memory recovery following distraction, in accord with the scenario of sequential representations (cp. to **Fig. 1c** and **f**).

To control for the possibility that noise in non-sparse implementations is mistaken for structure by SCA, we created substitute datasets with random Gaussian implementations (i.e., Gaussian distributions of neuronal loadings) while keeping the representational geometry intact and then systematically compared the original SCs with the substitute SCs (example substitute SCs in **Fig. 2f**). First, the sparsity parameter β (fit to the distribution of loadings on the SCs) was smaller for all three original SCs than for the substitutes ($p < 0.001$ for all three SCs; permutation test with $n = 3 \times 1000$ permutations; **Fig. 2g**), confirming the presence of structure in the implementation. Second, the activities on the SCs showed temporally restricted sample representations with shorter spread ($p < 0.002$; permutation test with $n = 1000$ permutations; same as for Fig. 2i-k; **Fig. 2h**), less temporal overlap with other SCs ($p < 0.003$; **Fig. 2i**), and less reversal of sample numerosity tuning ($p < 0.030$; **Fig. 2j**) than the substitutes, suggesting that the observed SC activity was more sequential than to be expected with a random implementation. Third and finally, the SCs were closer to orthogonal than the substitutes ($p < 0.019$; **Fig. 2k**), demonstrating that the observed implementation is more efficient than a random implementation.

In summary, the neuronal implementation of the sample numerosities' representational geometry was structured and sparse. The activities on the sparse components

demonstrated sequential rather than continuous coding of working memory content, indicating that the change of behavioral demands in the course of the trial triggers a switching of informative subpopulations.

Figure 2

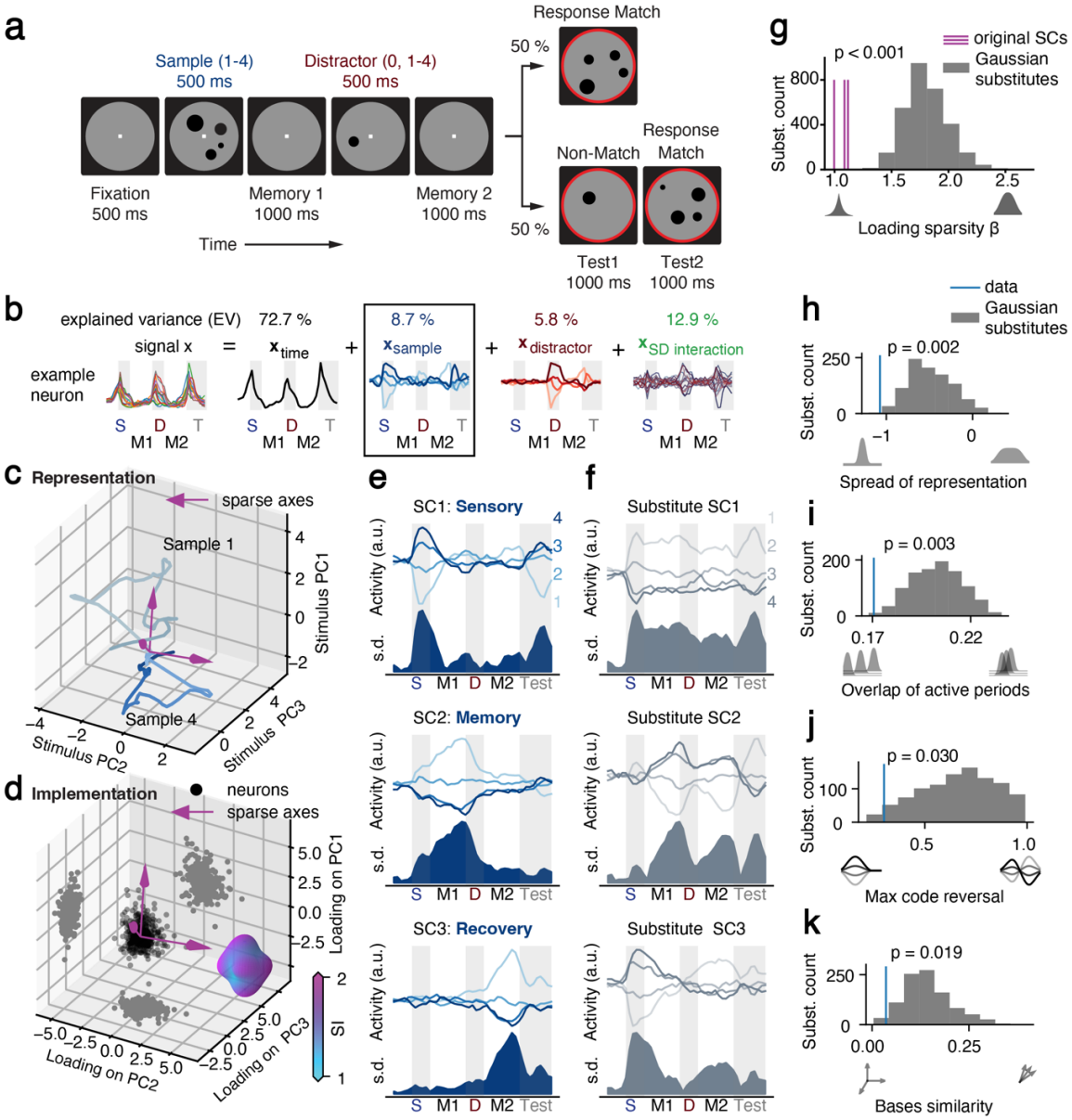


Fig. 2. The neuronal implementation of working memory

*(**a**) Delayed-match-to-numerosity task with distractors. (**b**) Demixing procedure separating the activity of each neuron into the parts coding time, sample numerosity, distractor numerosity and sample-distractor interaction. The sample coding part is used for the following analyses. Top: percentage of explained variance for each part. (**c**) Representational geometry for sample numerosities 1 and 4 in PC space, averaged across trials of the same condition. (**d**) Loadings of all recorded neurons on the top three PCs (black dots) including distributions projected onto the planes formed by PC pairs*

*(gray dots). Sparse axes (magenta arrows; determined by SCA) have high SI. Inset: surface plot of SI for all axes in the space. (**e**) Activity of the three identified sparse components (SCs), averaged across trials for each sample numerosity condition (top; numbers indicate sample numerosity) and relative information across conditions measured as standard deviation (s.d.). (**f**) SCs of an example substitute dataset with non-structured Gaussian implementation. (**g**) Sparsity β of the neuronal loadings on the SCs (fit to generalized normal distribution) for the original data and the substitute datasets (permutation test with n = 3×1000 permutations). (**h-k**) Activity measures for the SCs of the original data and the substitute datasets (permutation test with n = 1000 permutations).*

The effect of distraction on sample numerosity representations

The lack of a component that continuously represented the behaviorally relevant sample numerosity throughout the trial was unexpected. We therefore investigated the influence of distraction on sample number coding.

First, we applied SCA to the demixed distractor coding part of the data (**Fig. 3a**, top). Two SCs were obtained that were sequentially active during presentation and maintenance of the distractor numerosity, respectively (**Fig. 3a**, bottom). These components resembled the sensory and memory sample coding SCs (cp. to **Fig. 2e**), suggesting that target and distracting information initially occupied similar resources despite their distinct behavioral relevance. Supporting this hypothesis, we found strongly overlapping neuronal loadings between sample SCs and distractor SCs (cosine similarity; 0.69 and 0.57 for the sensory and memory components, respectively; **Fig. 3b**) with displacement of sample information by distractor information as the trial evolved (**Fig. S1a**, top and middle). However, in contrast to the sample sensory and memory components, the sample recovery SC was unique and did not share loadings with any other SC (Fig. 3b). Furthermore, the sample recovery SC was not influenced by distractor information and carried sample information until test numerosity presentation (**Fig. S1a**, bottom). To correctly complete a trial, more activity in the sample sensory and recovery SCs was required when the trial contained a distractor than when a trial without a distractor was presented (**Fig. S1b**). Conversely, distractors led to reduced sample activity in the memory component.

Second, we applied SCA to the sample-distractor interaction part of the data. One SC was identified. Its activity was most pronounced when the sample and distractor numerosity were the same (**Fig. S2**). The neuronal loadings on this SC did not overlap with the loadings on sample or distractor SCs (**Fig. 3b**), suggesting that the boost in numerosity information was generated by a dedicated subpopulation responding to a repeated presentation of the same number, instead of changing the activity of the sample representing neurons.

Together, these results indicate a (partially) shared capacity for sample and distractor representations during the sensory input and subsequent memory delay stages. The invasion of distractor information forced the recruitment of an extra component, the recovery component, to maintain sample information in working memory.

So far, all analyses were performed on separated (demixed) representations. We next investigated whether sample and distractor information could be equally disentangled using SCA alone without demixing the numerosity coding signal (**Fig. 3c**). SCA performed on firing rates averaged across the second memory delay recovered two sparse components that each selectively captured sample and distractor information (**Fig. 3d**). The corresponding representational geometry was grid-like with clearly factorized sample and distractor information that each aligned well to one SC (**Fig. 3e**). Notably, this alignment was non-trivial and not enforced by our analytical method, arguing that the PFC spontaneously disentangles target and distractor representations in working memory. The underlying implementation showed clear sparse structure in the neuronal loadings onto these components (**Fig. 3f**).

For comparison, PCA, which is insensitive to the neuronal implementation, was unable to recover factorized components (**Fig. 3g**). The grid-like geometry was still largely preserved, but it did not align with the PCs (**Fig. 3h**). In contrast to SCA, PCA did not identify the components with the sparsest loadings (**Fig. 3i**).

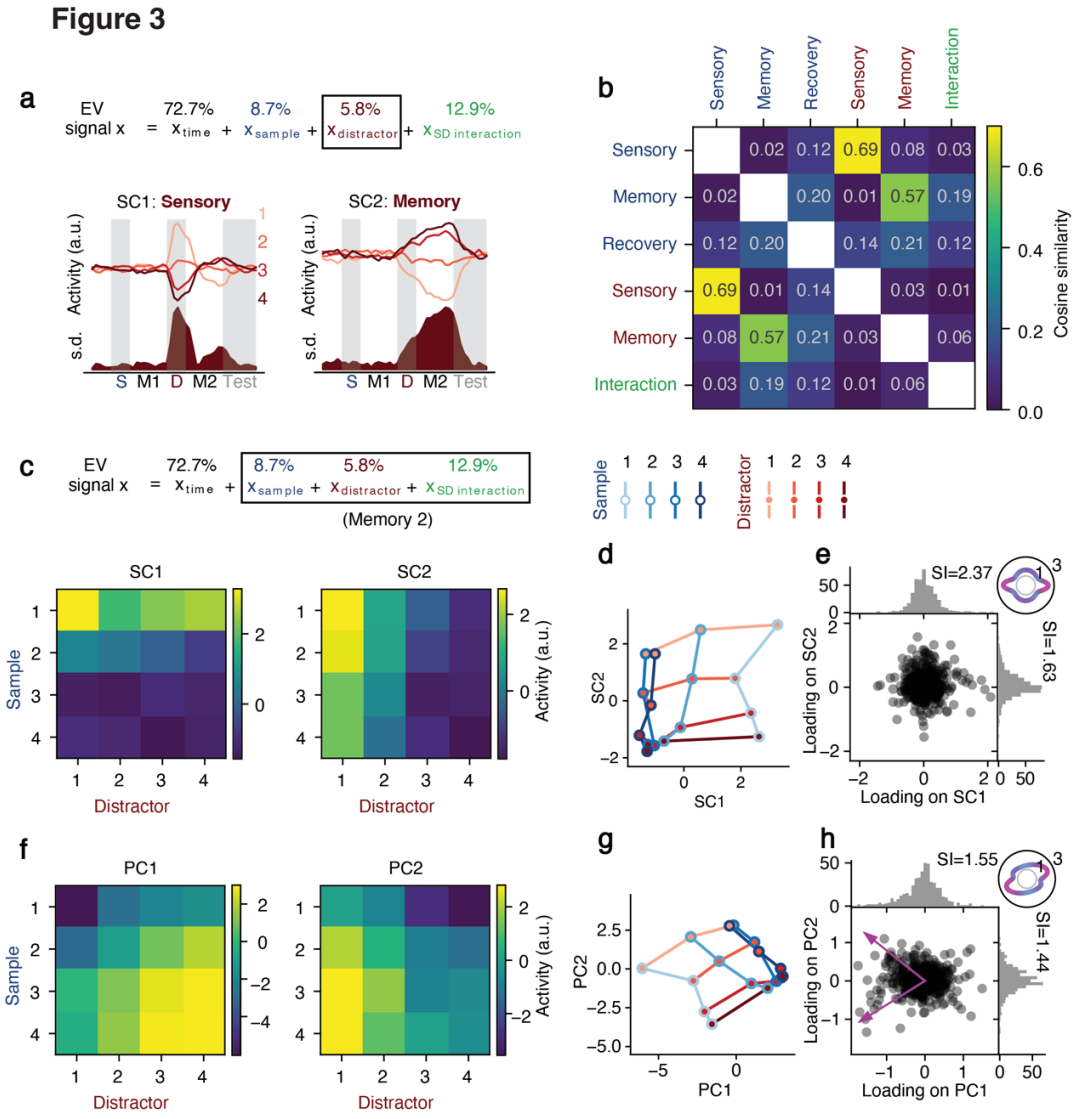


Fig. 3. The effect of distraction on sample representations

*(**a**) Top: the demixed distractor representing part used in the analysis. Bottom: distractor numerosity sparse components (SCs). Numbers indicate distractor numerosity. (**b**) Cosine similarity between loadings of sample numerosity SCs (blue), distractor numerosity SCs (red) and the sample-distractor interaction SC (green). (**c**) Activity of the two SCs identified using firing rates averaged across the second memory delay for all sample-distractor combinations without demixing the stimulus presentations. (**d**) Representational geometry in SC space. Blue and red colors indicate sample and distractor numerosity, respectively. (**e**) Neuronal loadings on the 2 SCs. Dots: joint distribution in SC space. Histograms: marginal distribution of neuronal loadings on SC1 and SC2. Inset: SI for all axes. (**f-h**) Same layout as in (c-e) but for PCs. Magenta arrows in (H) indicate sparse axes.*

Subpopulations of neurons dominate working memory representations

Next, we investigated whether the implementation was sparse enough to be able to reliably reconstruct the population-level sample representation using only a small fraction of neurons. We performed cross-temporal linear discriminant analysis (LDA) to decode sample numerosity at a given time point in the trial using training data from a different time point (**Fig. 4**). Decoding accuracy therefore quantifies the degree to which the representation is transferable. With four numerosities, chance level accuracy is 25 %. Using the entire population of 467 recorded neurons, we found a highly dynamic code with good within-epoch transfer, but very little generalization across epochs, in particular from the first to the second memory delay (**Fig. 4a**). In line with our previous results, this finding suggests that working memory representations are non-uniform and that distinct, complementary processes are required to protect behaviorally relevant information from interference.

We selected the neurons that contributed most to the previously identified SCs (loading on the SC larger than two standard deviations; **Fig. 4b**). 36, 28 and 28 single neurons passed the criterion for the sensory, memory and recovery SC, respectively. Although each subpopulation comprised only 6 to 8 % of the entire recorded population, these 'dominant neurons' explained 88 %, 82 % and 87 % of their respective component's variance (sum of squares of dominant neurons' loadings over sum of squares of all neurons' loadings). Overlapping membership in two subpopulations was very rare (no more than three neurons in any SC pair; **Fig. 4b**).

Cross-temporal LDA using only the dominant neurons showed a very similar sample numerosity decoding pattern as with the entire population (**Fig. 4c**, cp. with **Fig. 4a**), confirming that the decoder previously relied mainly on this small subset of neurons. The sensory subpopulation contributed to decoding in particular during the sample and test numerosity presentation, but showed very little activity in the memory epochs (**Fig. 4d**, top). The memory subpopulation dominated in the first delay, but surprisingly was not involved in sample coding during the second delay (**Fig. 4d**, middle). Instead, after distraction, the recovery subpopulation was exclusively responsible for carrying sample information (**Fig. 4d**, bottom). This suggests that these neurons crucially contribute to shielding working memory information from interference (see also **Fig. S1**).

Figure 4

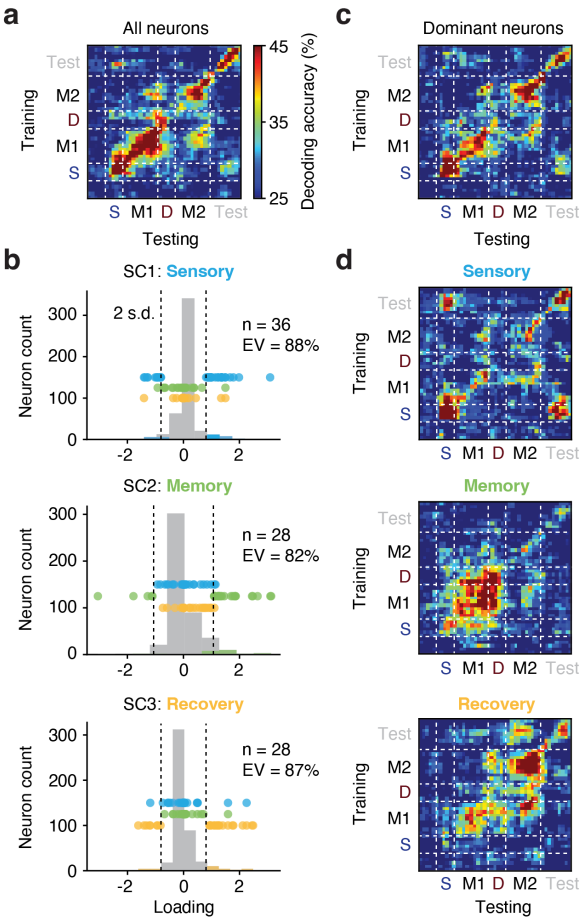


Fig. 4. Subpopulations of neurons dominating working memory coding

*(**a**) Accuracy of cross-temporal linear discriminant analysis (LDA) decoding of sample numerosity using all recorded neurons (y axis: training, x axis: testing). (**b**) Neuronal loadings on the three identified sample numerosity SCs. Colored dots indicate the 'dominant' neurons selected in each SC (cut-off: two s.d.). The percentage of variance explained within each SC is given for each subpopulation. (**c**) Accuracy of cross-temporal LDA decoding of sample numerosity using only the dominant neurons. Compare to (a). (**d**) Sample numerosity decoding accuracy using the dominant subpopulations of each SC. Same color scale in (a), (c) and (d).*

Subpopulation-specific electrophysiological properties

Above, we identified dominant neurons based on their stimulus selectivity. We now investigated whether their different roles in representing sample information were possibly mirrored by distinct electrophysiological properties.

First, we calculated the across-trial similarity (Pearson correlation) between each neuron's activity at different time points in the fixation period in order to derive the intrinsic time scale, a measure considered to index a neuron's ability to maintain memory

traces (*24*). Representative neurons from all three subpopulations are shown (**Fig. 5a**). The example recovery neuron had a significantly larger spread from the diagonal than the sensory and memory neuron, i.e., its activity in distant time points was more strongly correlated, thus signifying a longer time constant (**Fig. 5a**, bottom panel). For each subpopulation, an exponential decay was fitted to the mean correlation coefficient across neurons (**Fig. 5b**). The recovery subpopulation had the largest time constant τ (165 ms, 127 ms, and 338 ms for sensory, memory and recovery neurons, respectively). The distribution of τ values in the recovery population also stood out from the distributions observed in subsampled subpopulations of PFC neurons, whereas the sensory and memory neurons' distributions were not significantly different ($p = 0.874$, $p = 0.455$, $p = 0.002$ for sensory, memory and recovery subpopulations, respectively; KL-divergence with bootstraps; **Fig. 5c**).

Next, we investigated spike train statistics using the inter-spike intervals (ISI) measured during the neurons' entire recording lifetime. The coefficient of variation (CV) measures the irregularity of a spike train (**Fig. 5d**). CVs of all recorded neurons were larger than 1 (i.e., more irregular than a Poisson process) with a gradual increase of spiking irregularity across the sensory, memory and recovery subpopulations. CVs in the recovery neuron population were significantly larger than in the sensory subpopulation ($p = 0.030$, two-tailed *t*-Test; **Fig. 5d**). The local variation (LV) measures local ISI differences and complements CV, which is a global measure. LVs in all dominant neurons were smaller than 1 (i.e., less local variation than a Poisson process) and significantly lower than in the non-coding PFC population ($p < 0.001$, two-tailed *t*-Tests; **Fig. 5e**).

Notably, these distinct electrophysiological properties were not involved in the original selection of subpopulations and therefore lend support to the notion that the implementation structure carries biological meaning.

Figure 5

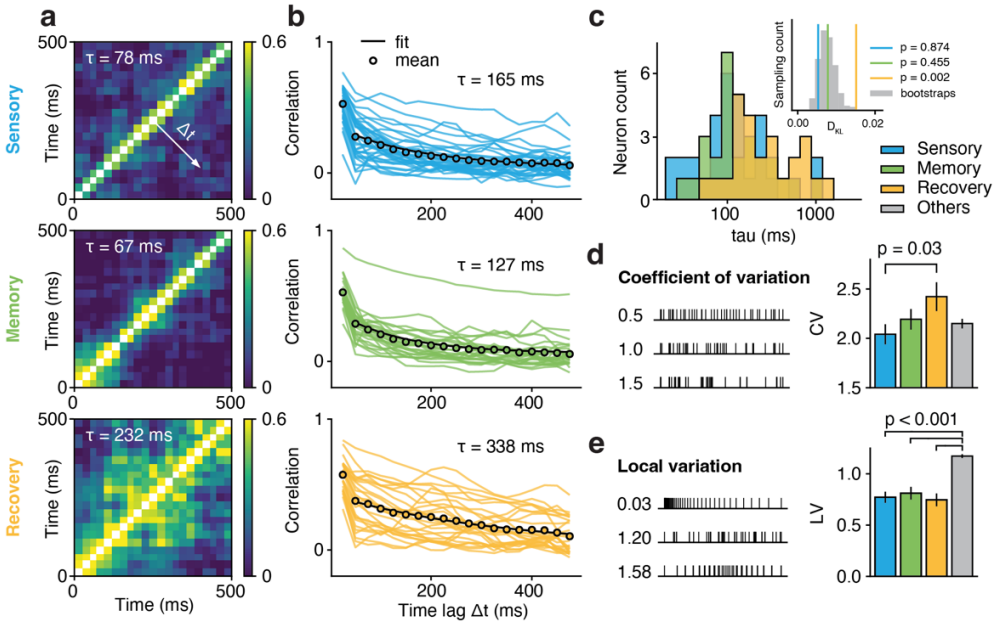


Fig. 5. Subpopulation-specific electrophysiological properties

*(**a**) Between-timepoint Pearson correlations of the trial-to-trial fluctuation of firing rates in the fixation epoch for the three dominant subpopulations. (**b**) Auto-correlograms obtained by averaging across diagonal offsets in (a). Auto-correlograms of individual neurons are given (single lines) together with the subpopulation average and the fitted exponential decay (black dots and line, respectively). (**c**) Distribution of fitted decay constants of individual neurons in each dominant subpopulation. Inset: Kullback-Leibler divergence (D_{KL}) between the distribution of each subpopulation and the whole population (null distribution for significance testing created with n = 1000 bootstraps from the whole population). (**d**) Coefficient of variation (CV) of inter-spike intervals (ISI) of the dominant subpopulations and the non-dominant other neurons (two-tailed t-Test). Left: example spike trains for different CVs. (**e**) Same layout as in (d) for the local variation (LV) of ISI.*

Subpopulation-specific temporal dynamics and representation of context

There was no perceptual cue in the working memory task specifying the difference between sample and distractor. This forced the animals to internally keep track of a trial's temporal evolution. To investigate whether temporal dynamics and context played a role in supporting the subpopulation-specific stimulus representations, we next analyzed the temporal part of the demixed signal and visualized condition-averaged activity trajectories in each of the dominant subpopulations (**Fig. 6a**).

In the sensory subpopulation, the trajectory followed a periodic, quasi-circular course (**Fig. 6a**, top panel). The first and second memory epochs overlapped almost entirely. This indicates that the sensory neurons did not distinguish between the time periods after

sample and after distractor presentation. The trajectory of the memory subpopulation was less periodic, but intertwined in the first and second memory epochs (**Fig. 6a**, middle panel). In contrast, the trajectory of the recovery subpopulation was less intertwined, with most time points distinguishable from each other, especially the first and second memory epochs, signifying a better representation of the contextual difference following sample and distractor presentation (**Fig. 6a**, bottom panel).

Overlap of the memory epochs in the sensory and memory subpopulations could be due to the limitations of a linear projection and the emphasis of PCA on global structure. We therefore performed non-linear embedding using t-SNE (**Fig. 6b**). This analysis revealed comparable structures as the linear projection, with the first and second memory epochs separated only in the recovery neuron subpopulation.

To further investigate the temporal evolution of neuronal activity, we measured the Euclidean distances between individual time points in each subpopulation (full space; **Fig. 6c**). All distance matrices displayed a strong diagonal, reflecting the fact that close-by time points were represented similarly. Notably, there were also strong offset diagonals in the sensory subpopulation, meaning that activity in these neurons repeated with a cycle of about 1.5 s. Furthermore, activity in the sensory and memory epochs differed most in this subpopulation. These patterns were present, albeit weaker, in the memory subpopulation, but absent in the recovery neurons. We quantified periodicity for each neuron by computing the relative power of 1/1.5 s (0.67 Hz) activity and its harmonics normalized to the power of the full frequency spectrum (**Fig. 6d**). Compared to randomly sampled subpopulations of PFC neurons, the sensory subpopulation and the recovery subpopulation showed significantly different (higher and lower, respectively) periodicity ($p < 0.001$, $p = 0.051$, $p = 0.043$ for sensory, memory and recovery subpopulations, respectively; KL-divergence with bootstraps; **Fig. 6d** inset).

Neuronal activity is not static and temporally independent. Instead, firing rates at every time point depend on previous time points. To characterize the dynamical properties of the recorded PFC population in more detail, we used the measure of tangling (*25*). Tangling measures the extent to which the velocity (direction and speed) of a given state on a trajectory diverges from the velocity of its neighboring states (**Fig. 6e**), reflecting the level of unpredictability and instability (chaos) in the system. High tangling means a small disturbance in the current state would lead to large changes in the next state (difference of derivatives of neighboring points). The instability or inability to determine the next state from the current state (i.e., high tangling) indicates that other neuronal populations or external stimuli may drive the trajectory. Consequently, tangling was increased following the onset and offset of sensory input in all three subpopulations. Tangling was highest, however, in the sensory subpopulation and lowest in the recovery subpopulation (sensory vs. memory, $p < 0.001$; memory vs. recovery, $p = 0.013$; two-tailed *t*-Test across all trial time points; **Fig. 6f**).

In summary, these results suggest that the subpopulation of recovery neurons keeps a record of time and temporal context, which could contribute to these neurons' ability to separate sample and distracting information. In contrast, the sensory subpopulation - and

the memory subpopulation to a lesser degree - is characterized by its strong input-driven temporal dynamics, which is consistent with these neurons' passive representation of numerosity regardless of it being behaviorally relevant (sample) or irrelevant (distractor).

Figure 6

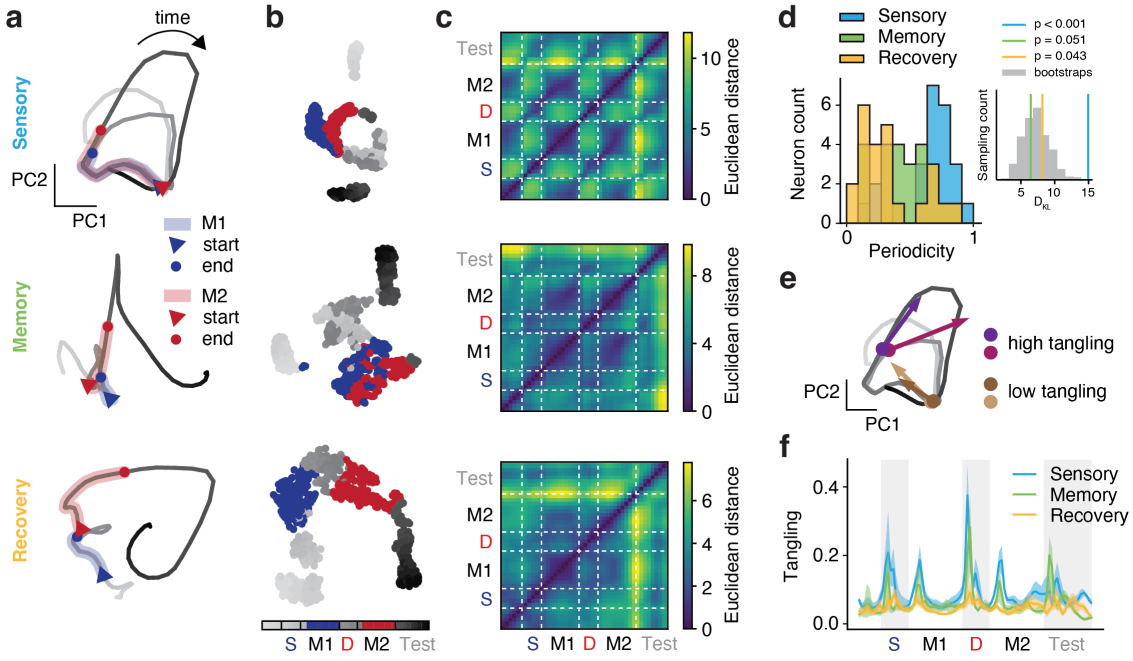


Fig. 6. Subpopulation-specific temporal dynamics

*(**a**) Temporal part of the demixed neuronal activity, averaged across conditions, of each dominant subpopulation projected onto their respective top two PCs. Time runs along the individual trajectories (bin width 50 ms). First and second memory delay are marked in blue and red, respectively. (**b**) Full signal averaged within each condition and embedded in 2D t-SNE space. Bins as in (a). (**c**) Euclidean distances between timepoints on the trajectory in (a) of each subpopulation. (**d**) Distribution of periodicity (relative power of 1/1.5 Hz and harmonics) of individual neurons in each subpopulation. Inset: Kullback-Leibler divergence (D_{KL}) between the distribution of each subpopulation and the whole population (null distribution for significance testing created with n = 1000 bootstraps from the whole population). (**e**) Example timepoints on the trajectory of the sensory subpopulation with high and low tangling. (**f**) Time resolved tangling of the trajectory of each subpopulation.*

Recurrent connectivity favors sparse implementations

The implementation underlying the temporal evolution of neuronal representations is not arbitrary, but must be derived from the dynamical system of constituent neurons and their anatomical connectivity pattern. The PFC is a highly recurrent, rather than purely feed-forward, brain region (*26*). If biological structure and resource efficiency indeed

favor sparse implementations, these should be better captured by recurrently connected networks than non-structured Gaussian implementations.

To address this hypothesis, we constructed a recurrent neural network model (RNN) to reproduce the target (to-be-fitted) firing rate sequences of each sample-distractor combination (**Fig. 7a**). The model consists of 467 neurons (to match the recorded population) receiving inputs of stimulus information according to the task structure. The model learns the recurrent connectivity W among the neurons. W summarizes the influence of the current time point's firing rates $\boldsymbol{r}$ on the firing rates of the next time point. An indicator vector $\boldsymbol{n}$ (one non-zero entry) represents the sample and distractor numerosity, activating the numerosity-specific input in I to the entire neuronal population. To reflect the absence of an explicit visual cue that differentiates between sample and distractor in the task design, sample and distractor numerosity share the same input channel (I, $\boldsymbol{n}$). The contextual difference is left for the model to resolve. The intercept term $\boldsymbol{b}$ captures the baseline activity of each neuron.

We first trained the model on the original dataset and visualized the trajectory of the output averaged across all conditions (**Fig. 7b**). The model reproduced the original dataset well, capturing 85.7 % of total variance. Next, we created substitute datasets with altered implementations of numerosity representations ($x_{sample} + x_{distractor} + x_{SD\ interaction}$) for the model to fit. The temporal part of the demixed data was unchanged. Three different implementations were created: first, a non-structured Gaussian distribution of neuronal loadings and no alignment to any components (cp. **Fig. 1d**); second, a distribution with the same degree of sparsity as the original data, but with sparse axes randomly rotated to align to other components (cp. **Fig. 1e**); third, a substitute with the same sparse distribution of neuronal loadings as in the original data (cp. **Fig. 1f**).

The model captured an increasing proportion of variance of the full signal across the three substitutes ($p < 0.001$; one-way ANOVA; **Fig. 7c**). The absolute differences in explained variance were comparatively small (left axis), but remarkable in relation to the variance of the manipulated signal (right axis) and given that the representational geometry was unchanged and identical for all substitutes (cp. **Fig. 1**). A comparable result was obtained for the explained variance of the numerosity coding part ($p < 0.001$; one-way ANOVA; **Fig. 7d**).

Taken together, these results demonstrate that sparse implementations of working memory representations are favored by recurrent circuits, the characteristic wiring motif of association cortices such as the PFC.

Figure 7

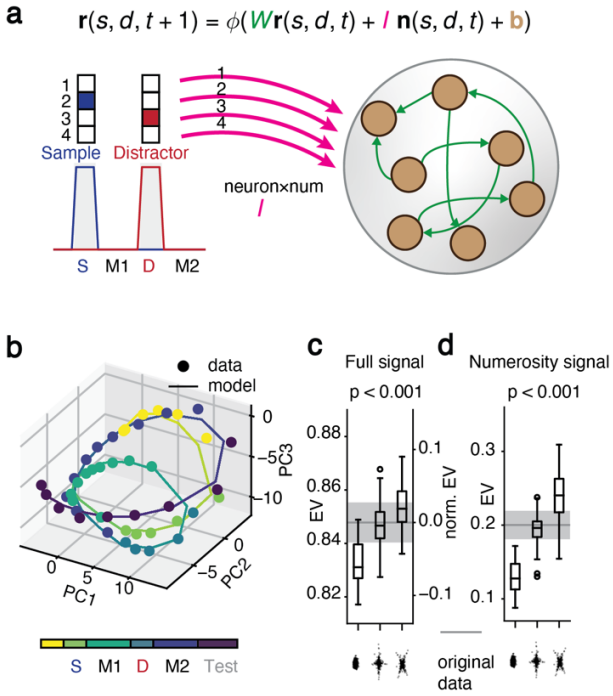


Fig. 7. Recurrent neural network modeling

*(**a**) RNN model governing equation and structure. Magenta and green arrows indicate numerosity-specific inputs and connectivity weights to be trained, respectively. (**b**) Model fit (solid trajectory) to original data (dots) averaged across all conditions. (**c**) Percentage of variance of the full signal explained by the model for non-structured Gaussian implementations of numerosity representations (left bar), sparse implementations with random orientations of sparse axes (middle bar) and sparse implementations with the same orientation of sparse axes as in the original data (right bar). Left and right axis show explained variance relative to the full signal and to the manipulated signal, respectively (one-way ANOVA across substitutes). (**d**) Same layout as in (c) for the percentage of variance of the numerosity signal explained by the model.*

Discussion

We presented a framework to examine the contributions of individual neurons to population-level responses in representation space and to utilize its implementation structure. We identified heavy-tailed, i.e., sparse distributions of neuronal loadings on components that captured disentangled and sequential memory representations including the recovery of memory content after distraction. The switching of working memory components circumvented interference. These components could be traced to small subpopulations of neurons with distinct electrophysiological properties and temporal dynamics. Modelling showed that such sparse implementations with sequentially active components are supported by recurrently connected networks.

Bridging population activity and neuronal implementation

Population-level activity and representational geometry were previously studied without forming direct links to individual neurons (*3–5*, *27*). However, while single-neuron selectivity measures have the advantage of being more easily connected to biological properties such as cell type, receptor expression and axonal projection targets, they are typically chosen based on intuition and past experience and only partially or indirectly reflect the full representational space (*9*, *28*).

Our sparse component analysis (SCA) framework (**Fig.1**) combines the advantages of both perspectives. It builds on representational geometry for a comprehensive account of the data and then links the relevant coding dimensions in the activity space to populations of strongly contributing neurons, which allows relating the population-wide activity patterns to tangible physiological measures.

Implementation reveals biologically relevant dimensions in activity space

Without respecting implementation, selecting components in activity space for further analysis is arbitrary. It is often done post-hoc after visualizing the top PCs, or by relying on the heuristics of 'what should be coded' in the system (*3*, *17*, *18*). This approach becomes problematic when the dimensionality is too high or when too many variables are involved.

By exploiting neuronal implementation, SCA identifies activity components in an un-biased and non-arbitrary way. SCA can therefore capture a more complete set of stimulus-associated variables (dimensions), most notably the temporal modulation of stimulus coding. This reduces bias otherwise introduced by selecting specific time windows, across which neuronal activity is averaged, and acknowledges the role of different response dynamics for information coding (*19*, *29*). Furthermore, incorporating temporal modulation renders analyses more robust to noise (*30*), which is usually Gaussian and could hide the structure in implementation.

The implementation's sparse structure is a result of biological constraints regarding the connections among individual neurons. The approximately 10^4 dendritic spines on each cortical neuron (*14*) define an upper limit for the number of neurons it could read out from. The 10^9 neurons in a cortical region such as human PFC (*12*, *13*), and even sub-

modules with one to two magnitudes fewer neurons, therefore cannot be reached directly. The addition of one connection step would allow reaching the majority of PFC neurons, but at the cost of producing a layer of 10^4 to 10^5 neurons that are dedicated exclusively to feeding the single hypothetical downstream neuron. This is prohibitively inefficient. In such polysynaptic chains, it is more likely that meaningful representations have already emerged in intermediate layers as a result of direct connections from the source region. This notion is also in line with the high dimensionality and non-linear mixed selectivity characteristic of PFC, which allow for direct linear readout of complex representations without further computations (*6*).

Neurons share inputs and have local recurrent connections, which are particularly pronounced in association cortices such as the PFC (*26*), resulting in more similar firing patterns among neurons within cortical regions. Consequently, neurons might display activity that is weakly correlated to some components of the representational geometry even though they do not participate in the readout. This emphasizes the importance of truncating neurons with weak loadings and enforcing sparsity constraints for estimating potential readout connections (**Fig. 4**) and motivates the use of dynamical systems modelling to validate correlative measures (**Fig. 7**).

Working memory persistence without neuronal persistence

Applied to working memory maintenance in the face of distraction, our framework uncovered an unexpected sequential representation of numerosity information across multiple task epochs (**Fig. 2**). This result was neither encouraged nor guaranteed by SCA. This suggests that the readout of memory content from the PFC is optimized for accuracy in each behavioral context rather than optimized for stability across time periods. The distractor occupied the same resources as the sample numerosity with regard to the sensory and memory component (**Fig. 3**), forcing behaviorally relevant information to be shifted to the recovery component following distraction. Thus, working memory content was maintained by distinct mechanisms before and after interference (**Fig. 4**).

The subpopulation of recovery neurons was characterized by electrophysiological properties that set these neurons apart from the other populations and could render them particularly suited to working memory storage. Their longer intrinsic timescales (**Fig. 5**) suggest more stable memory retention (*24*, *31*). These neurons also distinguished between sample and distractor contexts, which is crucial for determining what information to keep and what information to discard (**Fig. 6**). The contextual signal was additively mixed with the numerosity coding signal in these neurons, but might still act as gain modulation for numerosity information given the neuronal input-output non-linearity (*32*).

Representing memory content by sequentially active subpopulations is advantageous. With relay of information, a result of locally feed-forward connectivity, a network can maintain multiple inputs from previous time points and show more resistance to noise (*33*). Furthermore, the PFC might be non-linearly mixing context and memory representations in all possible ways, expanding dimensionality to enable flexible readout

(*6*). Extensive training could have strengthened the non-linear mixture of second memory epoch context and sample numerosity representations that was most important in the current task, with the PFC retaining other mixtures (e.g. the component coding for sample numerosity in the first memory epoch) for other behavioral demands. In this view, the subpopulation of memory neurons could function as a more passive short-term memory storage oblivious to the behavioral relevance of the memorized information.

Introducing distraction into the memory delay unmasked the crucial role of recovery neurons for working memory maintenance, which would have been hidden in simpler tasks. This highlights the importance of including richer temporal structure, multiple processing stages and behavioral perturbation into cognitive task designs to enable dissection of higher-order brain functions in finer detail and sampling from the full spectrum of underlying mechanisms.

Alternative implementation structures

We focused here on detecting sparse structure in the representational geometry's neuronal implementation, which is linked to the standardized moment of kurtosis. Consequently, the loading distributions have both positive and negative heavy tails. Reading out a given sparse component thus requires both excitatory and inhibitory connections. However, long-range corticocortical projections are mainly excitatory. This means that other selection criteria that capture non-symmetrical structure such as the standardized moment of skewness should also be explored (*34*, *35*).

Structure could be in the form of disjointed cell clusters (*28*) or a mixture of Gaussians (*32*). However, if present, these structures would not dissect the representational geometry, as they do not have a one-to-one relation to the dimensions in the activity space. Our neuronal implementation followed a unimodal Laplace distribution (Fig. 2g) instead of a multimodal distribution.

Structure can also be investigated when there are no prior assumptions about the underlying distributions of neuronal loadings. For example, given that neuronal firing is energy-consuming and non-negative, possibly encouraging neurons to align to the dimensions of the representational geometry that have shorter ranges of variation, non-uniform distributions of the number of selective neurons across different dimensions can arise (*36*). However, because all neurons are counted equally, structure probed non-parametrically could potentially be clouded by the large number of weakly coding (non-dominant) neurons and thus difficult to detect, in particular in PFC (*3*).

Relation of SCA to other linear dimensionality reduction methods

Different linear dimensionality reduction methods based on L2 reconstruction loss will yield comparable representational geometries, but they will not find the same projections of the representational geometry, i.e., the same components or the same coordinate system in which the data is expressed. The principle components of PCA are conveniently orthogonal and ranked by variance (*37*), but usually neither correspond to task-related components nor align to the activity of individual neurons (*38*). Truncating the smaller PCs provides denoised signal as a preprocessing step for independent

component analysis (ICA) that can infer the independent sources in the signal space (*39*). Its most common form, fastICA, enforces sparsity constraints on the activity of the components, reflecting an assumption about the activity (*40*). In contrast, in SCA the sparsity constraint is on the neuronal implementation, i.e., the potential readout weights corresponding to the mixing matrix in ICA, reflecting an assumption about the connectivity.

Neuronal representations must be communicated. Information that cannot be accessed by other neurons does not exist. In order to understand complex neural systems such as the PFC where we lack clear priors about the signal sources, it is paramount to exploit the circuit and wiring motifs that underlie the observed activity patterns.

Materials and Methods

Subjects

Two adult male rhesus monkeys (*Macaca mulatta*, 12 and 13 years old) were used for this study. All experimental procedures were in accordance with the guidelines for animal experimentation approved by the national authority, the Regierungspräsidium Tübingen. A detailed description is provided elsewhere (*8*, *9*). Monkeys were implanted with two right-hemispheric recording chambers centered over the principal sulcus of the lateral prefrontal cortex (PFC) and the ventral intraparietal area (VIP) in the fundus of the intraparietal sulcus. This study reports on the PFC data.

Task and stimuli

The animals grabbed a bar to initiate a trial and maintained eye fixation (ISCAN, Woburn, MA) within 1.75°of visual angle of a central white dot. Stimuli were presented on a centrally placed gray circular background subtending 5.4° of visual angle. Following a 500 ms pre-sample (pure fixation) period, a 500 ms sample stimulus containing 1 to 4 dots was shown. The monkeys had to memorize the sample numerosity for 2,500 ms and compare it to the number of dots (1 to 4) presented in a 1,000 ms test stimulus. Test stimuli were marked by a red ring surrounding the background circle. If the numerosities matched (50 % of trials), the animals released the bar (correct Match trial). If the numerosities were different (50 % of trials), the animals continued to hold the bar until the matching number was presented in the subsequent image (correct Non-match trial). Match and non-match trials were pseudo-randomly intermixed. Correct trials were rewarded with a drop of water. In 80 % of trials, a 500 ms interfering numerosity of equal numerical range was presented between the sample and test stimulus. The interfering numerosity was independent from either the sample or test numerosity and therefore not useful for solving the task. In 20 % of trials, a 500 ms gray background circle without dots was presented instead of an interfering stimulus, i.e., trial length remained constant (control condition, blank). Trials with and without interfering numerosities were pseudo-randomly intermixed. Stimulus presentation was balanced: a given sample was followed by all interfering numerosities with equal frequency, and vice versa. Throughout the monkeys' training on the distractor task, there was never a condition where a stimulus appearing at the time of the distractor was task-relevant.

Low-level, non-numerical visual features could not systematically influence task performance (*9*, *41*):in half of the trials, dot diameters were selected at random. In the other half, dot density and total occupied area were equated across stimuli. CORTEX software (NIMH, Bethesda, MD) was used for experimental control and behavioral data acquisition. New stimuli were generated before each recording session to ensure that the animals did not memorize stimulus sequences.

Electrophysiology

Up to eight 1 MΩ glass-insulated tungsten electrodes (Alpha Omega, Israel) per chamber and session were acutely inserted through an intact dura with 1 mm spacing. Single units were recorded at random; no attempt was made to preselect for particular response

properties (*9*). Signal amplification, filtering, and digitalization were accomplished with the MAP system (Plexon, Dallas, TX). Waveform separation was performed offline (Plexon Offline Sorter).

Data analysis tools

Data analysis was performed with Python using custom scripts based on packages NumPy, SciPy, sci-kit learn, TensorFlow2, PyTorch, Matplotlib and Plotly.

Preprocessing

Single units were included in the analysis if they were recorded in at least 4 correct trials of each task condition (meaning each unique sample and distractor numerosity combination). This resulted in 467 neurons across 78 sessions recorded in the PFC. Trials without distractors were not included in the analyses unless specified otherwise.

Unless specified otherwise, the firing rates were binned in a Gaussian window with sigma of 50 ms and step of 100 ms, aligned to the start of the fixation period. The data were then organized into a neuron-by-condition-by-timepoint tensor. Each tensor entry was normalized by the standard deviation across trials (within each condition).

Demixing

Given the independence of the task variables sample numerosity (s), distractor numerosity (d) and trial time (t), the neuronal activity can be directly factorized into parts for each variable and their interaction:

$$x = \bar{x} + \bar{x_t} + \bar{x_s} + \bar{x}_d + \bar{x}_{st} + \bar{x}_{dt} + \bar{x}_{sd} + \bar{x}_{sdt} \quad (1)$$

Because the stimulus response is also modulated by time, each part was grouped together with its interaction with time (*21*):

$$x_{time} = \bar{x_t} \quad (2)$$

$$x_{sample} = \bar{x_s} + \bar{x}_{st} \quad (3)$$

$$x_{distractor} = \bar{x}_d + \bar{x}_{dt} \quad (4)$$

$$x_{sd\ interaction} = \bar{x}_{sd} + \bar{x}_{sdt} \quad (5)$$

Visualization of representation and implementation space

For a data matrix X where each column vector x is the demixed activity of a neuron, the singular value decomposition was taken:

$$X = U\Sigma V^T \quad (6)$$

where U and V are unitary matrices and Σ is a diagonal matrix with ordered singular values. The first n columns of $U\Sigma$ are the PCs that were used to visualize the representational geometry. The first n columns of $V\Sigma$ are loadings on the PCs that were used to visualize the implementation space.

Within this subspace an arbitrary component can be specified with $U\Sigma P_{:,1}$ ($P_{:,1}$ being a column vector from a unitary matrix P), with the orientation of this component given by $P_{:,1}$. The loadings on this component will be the first row of $(U\Sigma P)^+ X = P^T V^T$, that is ${P_{:,1}}^T V^T$. This way, the loadings are visualized with the same orientation $P_{:,1}$. in implementation space as their corresponding component in representation space. The sparsity index of the neuronal loadings on component $U\Sigma P_{:,1}$ is then:

$$SI\left(P_{:,1}\right) = \frac{kurtosis\left({P_{:,1}}^T V^T\right)}{3} \tag{7}$$

$$kurtosis(\boldsymbol{x}) = \langle(\boldsymbol{x} - \overline{\boldsymbol{x}})^4\rangle / \langle(\boldsymbol{x} - \overline{\boldsymbol{x}})^2\rangle^2 \tag{8}$$

Sparse component analysis

Following the formulation of sparse coding (*15*, *16*, *42*), sparse component analysis (SCA) reduces the dimensionality of the dataset and extracts the unique components by enforcing a sparse penalty on neuronal loadings:

$$Loss = \left\| X - \sum_{i=1}^{k} \overrightarrow{u_i}\, \overrightarrow{v_i}^T \right\|_{frobenius} + \alpha \sum_{i=1}^{k} \|\overrightarrow{v_i}\|_1 + \beta \sum_{i=1}^{k} \|\overrightarrow{v_i}\|_2^2 \tag{9}$$

$$where\ \|\overrightarrow{u_i}\| = 1$$

The loss function is defined as the sum of the reconstruction loss and the regularizations. Data X is organized as a n firing instances by p neurons matrix. X is then approximated by k firing activity vectors $\vec{u}$ and their corresponding neuronal loadings $\vec{v}$. Parameter α controls the strength of L1-regularization that encourages sparsity of the loadings. Parameters α and k were determined by a cross-validated grid search. β was set at 0.01 to smooth the loss landscape and make the result stable across random initializations.

Substitute data for SCA

Substitute data were created for the demixed sample coding part X of the data (Fig. 2). For the singular value decomposition $X = U\Sigma V^T$, $U\Sigma$ specifies the representational geometry (see above). Operations were performed on V only.

A random unitary matrix R with the size of the number of neurons was drawn from a Haar distribution. The original matrix V was replaced with $V' = VR$. V' is also a unitary matrix, meaning that this manipulation will not change the geometries but will rotate them to random axes. In other words, it will linearly combine the loadings including those on the components with very low variance, which will render the substitute distribution of loadings on the sample numerosity components close to Gaussian. The substitute data is then:

$$X' = U\Sigma {V'}^T = XR \tag{10}$$

Measures of sparse component activity

$\overrightarrow{u_i}$ in SCA specifies the activity of the sparse component i. The following measures of the set of $\overrightarrow{u_i}$ were compared between the original dataset and its substitutes (n = 1000).

Spread of representation. The standard deviation of $\overrightarrow{u_i}$ across different numerosity conditions k at each time point was used to define the relative (normalized) information at that time point. Specifically, each $\overrightarrow{u_i}$ was first reshaped into a condition-by-timepoint matrix Y^i. Then the information in component i at time point t is given by:

$$Z_{i,t} = \sqrt{\langle (Y^i_{k,t} - \langle Y^i_{k,t} \rangle_k)^2 \rangle_k} \tag{11}$$

The skewness of the information across time points was calculated for each component and averaged across components as follows:

$$Skew_i = \langle (Z_{i,t} - \overline{Z_{i,t}})^3 \rangle_t / \langle (Z_{i,t} - \overline{Z_{i,t}})^2 \rangle_t^{3/2} \tag{12}$$

Positively skewed Z indicates a long tail in the distribution of information across time points, corresponding to few time points having high information. Conversely, a smaller or even negative skewness implies there are more high information timepoints than low information time points, making the high information more spread out across time points. We define the spread of representation as the negative skewness:

$$Spread = -\langle Skew_i \rangle_i \tag{13}$$

Overlap of active periods. The dot product of the information of every pair of components i and j was taken and averaged across pairs:

$$Overlap = \langle Z_{i,t} Z^T_{j,t} \rangle \tag{14}$$

Maximum tuning reversal. A given component i may show changes of tuning to sample numerosities during the course of a trial. Its tuning at time t is specified by $Y^i_{:,t}$. For each component i, the dot product similarity of tunings between timepoint pairs was specified in the non-diagonal entries in $C^i = {Y^i}^T Y^i$, where the diagonal entries are the strength of the tuning at each time point. C^i was then normalized to the strongest tuning: $C^{i'} = C^i / \max(C^i)$. The most negative entry in $C^{i'}$ was then the degree of reversal in this component. $Reversal_i = -\min(C^{i'})$. It would reach the maximum of 1 when tuning at a given time point is the complete reversal of the strongest tuning. It would be close to 0 when the tuning does not reverse. The maximum tuning reversal is then the largest reversal in a set of SCs:

$$Max\ tuning\ reversal = \max_i Reversal_i = \max_i [-\min\left(\frac{{Y^i}^T Y^i}{\max({Y^i}^T Y^i)}\right)] \tag{15}$$

Component similarity. Let U_{sca} be the concatenation of activity $\vec{u}_i$ and V_{sca} the concatenation of loadings $\vec{v}_i$ of the sparse component i. The data matrix can be expressed as $X = U_{sca} V^T_{sca} + \epsilon$. ϵ denotes the noise term. Then it follows $U^+_{sca}(X - \epsilon) = V^T_{sca}$. The pseudoinverse U^+_{sca} can be viewed as a linear transform of the original data. Since all the activities $\vec{u}$ have unit length, larger loadings would be required to express an arbitrary

geometry when the activities are correlated, meaning lower efficiency. The component similarity is measured by the product of the singular values of U_{sca}. Formally, if the singular value decomposition gives $U_{sca} = U\Sigma V^T$, then

$$Similarity = \prod_i \Sigma_{i,i} \quad (16)$$

The similarity can also be viewed as the determinant of the transformation matrix from arbitrary orthogonal bases to the bases of U_{sca}.

Numerosity information in different components

The standard deviation $Z_{i,t}$ for all time points t specifies the evolution of normalized information within this component. But since $\vec{u}_i$ in component i has unit length, this measure does not allow for direct comparisons between components (see above). To allow for such comparisons (Fig. S1), the norm of $\vec{v_i}$ is therefore applied to $Z_{i,t}$ as a scaling factor:

$$Information = \|\vec{v_i}\| Z_{i,t} \quad (17)$$

Linear discriminant analysis decoding

Neurons recorded in different sessions were stitched together. To account for the different number of trials recorded per neuron, a criterion was set to ensure there were at least 1.5 times more trials than neurons. This resulted in 228 neurons with at least 385 trials each. Removing incorrect trials and selecting the minimum number of trials recorded per condition and neuron left 118 trials per neuron. Trials of the same condition were then randomly selected for each repetition of the analysis.

Multi-class linear discriminant analysis (LDA; sci-kit learn package) was used for decoding because of its advantageous property of accounting for data covariance. LDA assumes the same covariance in every class. It finds the projection that preserves the Mahalanobis distance between classes and predicts the label of a new data point by its Mahalanobis distance to the class centroid. Shrinkage of the measured covariance matrix was performed by averaging with a diagonal matrix. The strength of shrinkage was determined following the Ledoit-Wolf lemma (*43*).

Decoding accuracy, i.e., the ratio of correctly predicted trials, was averaged across 7 repetitions of 7-fold cross-validation.

Spike train statistics

Firing rates were binned in a Gaussian window with sigma of 12.5 ms and step of 25 ms.

Correlation, autocorrelation and intrinsic timescales were determined as described elsewhere (*24*). The firing rate of each neuron n at timepoint t of trial i is expressed as $x_{n,i,t}$. The Pearson correlation between timepoints $t1$ and $t2$ is then:

$$r_n(t1,t2) = \frac{\left\langle \left(x_{n,i,t1} - \langle x_{n,i,t1}\rangle_i\right)\left(x_{n,i,t2} - \langle x_{n,i,t2}\rangle_i\right)\right\rangle_i}{\left\langle \left(x_{n,i,t1} - \langle x_{n,i,t1}\rangle_i\right)^2\right\rangle_i^{1/2} \left\langle \left(x_{n,i,t2} - \langle x_{n,i,t2}\rangle_i\right)\right\rangle_i^{1/2}} \quad (18)$$

Autocorrelation is defined as:

$$AC_n(\Delta t) = \langle r_n(t0, t0+\Delta t)\rangle_{t0} \quad (19)$$

To account for the refractoriness and adaptation at small time lags, fitting started at the time lag where the autocorrelation function had dropped most strongly. Neurons with the strongest drop after 400 ms were discarded (6 neurons). The autocorrelation was then fitted with an exponential decay:

$$AC(\Delta t) = A[\exp(-\Delta t/\tau) + B] \quad (20)$$

Parameters A and B were constrained in $[0,1]$ and τ was constrained from 10 ms to 2000 ms. The autocorrelation function of 8 neurons could not be fitted. The neurons with τ fitted below 20 ms (20 neurons) or above 1600 ms (25 neurons) were excluded because of the biologically unrealistic fit. This left 408 neurons. Very few neurons were excluded in the dominant subpopulations (2, 2, and 1 neurons for the sensory, memory and recovery subpopulation, respectively).

The inter-spike intervals (ISI) were determined for the entire session. The coefficient of variation (CV) measures the global variation of a neuron's ISI and is defined as:

$$CV = s.d.(ISI)/\langle ISI\rangle \quad (21)$$

In contrast to CV, local variation (LV) measures the local ISI change (*44*). It is defined as:

$$LV = \frac{3}{n-1}\sum_{i=1}^{n-1}(ISI_i - ISI_{i+1})^2/(ISI_i + ISI_{i+1})^2 \quad (22)$$

CV and LV are both expected to be 1 for spiking activity following a Poisson process. CV and LV would be 0 for perfectly regular firing and larger than 1 for more irregular firing than by a Poisson process.

Kullback-Leibler divergence

KL divergence measures the difference between two distributions. For the analyses of intrinsic time scales and periodicity, KL divergence was calculated between the distribution of statistic x for the entire population P and that of sub-samples Q (either dominant subpopulations or bootstrap subsamples). It is given by:

$$D_{KL}(P\|Q) = -\sum_{\mathrm{x}} P(\mathrm{x}) \cdot \log\, Q(\mathrm{x})/P(\mathrm{x}) \quad (23)$$

To create the null distribution of D_{KL}, 27 neurons (comparable to the number of neurons in the dominant subpopulations after exclusion of neurons in which no autocorrelation function could be fitted) were randomly sampled from the PFC population 1000 times.

Temporal dynamics

Periodicity. The Fourier transform of the demixed temporal part of the firing rate of each neuron is given by:

$$PSD(f) = DFT(x_{time}(t)) \tag{24}$$

Then, the periodicity was defined as the ratio between the power of the harmonics of 1/1.5 Hz (reflecting the onset of visual input at regular spacing of 1.5 s) and the power of all frequencies:

$$Periodicity = \sum_{i \in Z^+} PSD(i\frac{2}{3}) \Big/ \sum_{f} PSD(f) \tag{25}$$

Tangling. Tangling reflects the smoothness and stability of the flow field around the vicinity of state x_t on a trajectory (*25*). It is given by:

$$Q(t) = \max_{t'} \frac{\|\dot{\boldsymbol{x}}_t - \dot{\boldsymbol{x}}_{t'}\|^2}{\|\boldsymbol{x}_t - \boldsymbol{x}_{t'}\|^2 + \epsilon} \tag{26}$$

It specifies the maximum difference between the derivative at state x_t and the derivative at other states $\boldsymbol{x}_{t'}$, normalized by their Euclidean distance. A small constant ϵ was added to avoid numerical error when the two states were too close.

Recurrent neural network

A recurrent neural network (RNN) model was implemented using the PyTorch neural network module. The model has the formulation:

$$\boldsymbol{r}(s, d, t+1) = \phi(W\boldsymbol{r}(s, d, t) + I\boldsymbol{n}(s, d, t) + \boldsymbol{b}) \tag{27}$$

$\boldsymbol{r}$ is the firing rate of units in the condition of sample numerosity s and distractor numerosity d at time point t. ϕ is the non-linear activation function, chosen to be a rectified linear unit (ReLu) to respect the biological characteristics of non-negative firing rates with high upper limits. W is the within-population connectivity matrix. I is the input matrix with the dimensions of 467 (total number of units) by 4 (number of numerosities). A column $I_{:,a}$ is the input to the units when numerosity a is being presented. $\boldsymbol{n}$ is an indicator vector with the entry $\boldsymbol{n}_a$ corresponding to the presented numerosity being 1 and all other entries being 0. $\boldsymbol{b}$ is the intercept. W, I and $\boldsymbol{b}$ are the parameters to be trained. Formally, $\boldsymbol{n}$ as a function of trial type specified by s and d and time point t is defined by:

$$\boldsymbol{n}(s, d, t) = \boldsymbol{m}(s) \cdot mask_{[0.5,1)}(t) + \boldsymbol{m}(d) \cdot mask_{[2,2.5)}(t) \tag{28}$$

$$where \qquad \boldsymbol{m}(x) = \left[\mathbf{1}_{\{1\}}(x), \mathbf{1}_{\{2\}}(x), \mathbf{1}_{\{3\}}(x), \mathbf{1}_{\{4\}}(x)\right]^T$$

$$mask_A(t) = \mathbf{1}_A(t * 0.1)$$

$$\mathbf{1}_A(x) \coloneqq \begin{cases} 1, x \in A \\ 0, x \notin A \end{cases}$$

$\boldsymbol{m}$ maps a numerosity to the corresponding one-hot vector. $mask_A(t)$ indicates the time (0.1 s steps) when the corresponding stimulus is presented. $\mathbf{1}_A(x)$ is an ancillary indicator function to define $\boldsymbol{m}$ and $mask$.

The model was trained to produce the whole sequence of firing rates $\boldsymbol{r}(s,d,t)$ in order to match the target data $\boldsymbol{x}_{s,d,t}$, given the initial firing rate in the fixation period $\boldsymbol{r}(s,d,0)$ and the input $\boldsymbol{n}(s,d,t)$. The loss function is defined as:

$$Loss(W,I,\boldsymbol{b}) = \sum_{s,d,t}\left[\boldsymbol{r}(s,d,t) - \boldsymbol{x}_{s,d,t}\right]^2 + \lambda\|W\|_1 + \lambda\|I\|_1 \tag{29}$$

$$\boldsymbol{r}(s,d,t_0) = \boldsymbol{x}_{s,d,t_0} \tag{30}$$

The coefficient λ controls the strength of regularization and was determined by a grid search with cross validation.

The prediction of the later timepoints relies on the quality of the prediction of the early timepoints. If the training was done only by giving the first timepoint, convergence would be difficult to achieve and learning heavily biased towards reproducing early timepoints in the data. To overcome this possible instability, the model was trained in a recursive fashion by first using every timepoint as the initial firing rate, training the model to predict the following timepoints and gradually increasing the number of timepoints the model needs to predict. As such, at each iteration i, the temporal sequence $x_{s,d,t}$ was reorganized into $T-i$ chunks of length $i+1$, $\langle \boldsymbol{x}_{s,d,t_0}, \dots, \boldsymbol{x}_{s,d,t_0+i}\rangle, t_0 \in \langle 1, \dots, T-i\rangle$, with the first firing rate in each chunk as initial firing rate and the rest as target to be fit by the model.

Variance explained by RNN

The variance explained by the model was determined by the difference between the model's predicted trajectory and the trajectory of the original data normalized to the difference between a reference trajectory (constant activity set to the first entry of the fixation period) and the trajectory of the original data:

$$EV = 1 - \sum_{s,d,t}\left[\boldsymbol{r}(s,d,t) - \boldsymbol{x}_{s,d,t}\right]^2 / \sum_{s,d,t}\left[\boldsymbol{x}_{s,d,t_0} - \boldsymbol{x}_{s,d,t}\right]^2 \tag{31}$$

The normalized EV (Fig. 7c, right axis) was defined as the difference between a substitute's EV and the original data's EV, divided by the percentage of the manipulated variance (numerosity coding signal, 27.4 %; cp. Fig. 2b). EV for the numerosity signal (Fig. 7d) was calculated by replacing both $\boldsymbol{r}(s,d,t)$ and $\boldsymbol{x}_{s,d,t}$ with their demixed numerosity representing parts.

Substitute data for RNN

In order not to distort the strong connection between sample and distractor numerosity coding (e.g., Fig. 3b, Fig. S1), the loadings of these two parts of the data and their interaction were shuffled together to create three types of substitute datasets. The RNN model was then trained on the substitutes.

922 *Gaussian distribution of loadings.* The Gaussian substitutes were created as described for
923 SCA, except for that singular value decomposition was performed on $X_{sample} +$
924 $X_{distractor} + X_{sd\_interaction} = X_{all} - X_t = U\Sigma V^T$.

925 *Sparse distribution with random alignment.* For k dimensions of the numerosity coding
926 part of the data (determined by cross validation), a $k \times k$ unitary matrix R was randomly
927 drawn from a Haar distribution and combined with an identity matrix I to create $R' =$
928 $\begin{pmatrix} R & 0 \\ 0 & I \end{pmatrix}$. Then, $V' = VR'$ was substituted for V. This leaves the sparse structure in the
929 original k dimensional numerosity representing subspace intact, but rotates the sparse
930 structure in $V_{:,1:k}$ to random orientations.

931 *Sparse distribution with original alignment*. The rows of $V_{:,1:k}$, i.e., the neuronal identities,
932 were permuted by substituting $V' = (V_{permute,1:k}, V_{:,k+1:p})$ for V.

933
934
935 **References**

936 1. D. L. Barack, J. W. Krakauer, Two views on the cognitive brain. *Nat Rev Neurosci*. **22**,
937 359–371 (2021).

938 2. S. Saxena, J. P. Cunningham, Towards the neural population doctrine. *Current Opinion in*
939 *Neurobiology*. **55**, 103–111 (2019).

940 3. S. Bernardi, M. K. Benna, M. Rigotti, J. Munuera, S. Fusi, C. D. Salzman, The Geometry
941 of Abstraction in the Hippocampus and Prefrontal Cortex. *Cell*. **183**, 954-967.e21 (2020).

942 4. G. Okazawa, C. E. Hatch, A. Mancoo, C. K. Machens, R. Kiani, Representational
943 geometry of perceptual decisions in the monkey parietal cortex. *Cell*. **184**, 3748-3761.e18
944 (2021).

945 5. N. Kriegeskorte, X.-X. Wei, Neural tuning and representational geometry. *Nature Reviews*
946 *Neuroscience*. **22**, 703–718 (2021).

947 6. M. Rigotti, O. Barak, M. R. Warden, X.-J. Wang, N. D. Daw, E. K. Miller, S. Fusi, The
948 importance of mixed selectivity in complex cognitive tasks. *Nature*. **497**, 585–590 (2013).

949 7. S. E. Cavanagh, J. P. Towers, J. D. Wallis, L. T. Hunt, S. W. Kennerley, Reconciling
950 persistent and dynamic hypotheses of working memory coding in prefrontal cortex. *Nat*
951 *Commun*. **9**, 3498 (2018).

952 8. S. N. Jacob, D. Hähnke, A. Nieder, Structuring of Abstract Working Memory Content by
953 Fronto-parietal Synchrony in Primate Cortex. *Neuron*. **99**, 588-597.e5 (2018).

954 9. S. N. Jacob, A. Nieder, Complementary Roles for Primate Frontal and Parietal Cortex in
955 Guarding Working Memory from Distractor Stimuli. *Neuron*. **83**, 226–237 (2014).

956 10. A. Parthasarathy, C. Tang, R. Herikstad, L. F. Cheong, S.-C. Yen, C. Libedinsky, Time-
957 invariant working memory representations in the presence of code-morphing in the lateral
958 prefrontal cortex. *Nature Communications*. **10**, 4995 (2019).

959 11. C. Tang, R. Herikstad, A. Parthasarathy, C. Libedinsky, S.-C. Yen, Minimally dependent
960 activity subspaces for working memory and motor preparation in the lateral prefrontal
961 cortex. *Elife*. **9**, e58154 (2020).

962 12. E. Courchesne, P. R. Mouton, M. E. Calhoun, K. Semendeferi, C. Ahrens-Barbeau, M. J.
963 Hallet, C. C. Barnes, K. Pierce, Neuron Number and Size in Prefrontal Cortex of Children
964 With Autism. *JAMA*. **306**, 2001–2010 (2011).

965 13. S. Herculano-Houzel, K. Catania, P. R. Manger, J. H. Kaas, Mammalian Brains Are Made
966 of These: A Dataset of the Numbers and Densities of Neuronal and Nonneuronal Cells in
967 the Brain of Glires, Primates, Scandentia, Eulipotyphlans, Afrotherians and Artiodactyls,
968 and Their Relationship with Body Mass. *Brain Behav Evol*. **86**, 145–163 (2015).

969 14. G. Eyal, M. B. Verhoog, G. Testa-Silva, Y. Deitcher, R. Benavides-Piccione, J. DeFelipe,
970 C. P. J. de Kock, H. D. Mansvelder, I. Segev, Human Cortical Pyramidal Neurons: From
971 Spines to Spikes via Models. *Front. Cell. Neurosci.* **12**, 181 (2018).

15. P. Georgiev, F. Theis, A. Cichocki, H. Bakardjian, Sparse component analysis: a new tool for data mining. *Data mining in biomedicine*. **7**, 91–116 (2007).

16. B. A. Olshausen, D. J. Field, Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*. **381**, 607–609 (1996).

17. M. C. Aoi, V. Mante, J. W. Pillow, Prefrontal cortex exhibits multidimensional dynamic encoding during decision-making. *Nat Neurosci*. **23**, 1410–1420 (2020).

18. A. Libby, T. J. Buschman, Rotational dynamics reduce interference between sensory and memory representations. *Nat Neurosci*. **24**, 715–726 (2021).

19. V. Mante, D. Sussillo, K. V. Shenoy, W. T. Newsome, Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*. **503**, 78–84 (2013).

20. S. Kornblith, M. Norouzi, H. Lee, G. Hinton, "Similarity of neural network representations revisited" in (PMLR, 2019), pp. 3519–3529.

21. D. Kobak, W. Brendel, C. Constantinidis, C. E. Feierstein, A. Kepecs, Z. F. Mainen, X.-L. Qi, R. Romo, N. Uchida, C. K. Machens, Demixed principal component analysis of neural population data. *eLife*. **5**, e10989 (2016).

22. G. F. Elsayed, J. P. Cunningham, Structure in neural population recordings: an expected byproduct of simpler phenomena? *Nat Neurosci*. **20**, 1310–1318 (2017).

23. J. D. Murray, A. Bernacchia, N. A. Roy, C. Constantinidis, R. Romo, X.-J. Wang, Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex. *Proc Natl Acad Sci USA*. **114**, 394–399 (2017).

24. J. D. Murray, A. Bernacchia, D. J. Freedman, R. Romo, J. D. Wallis, X. Cai, C. Padoa-Schioppa, T. Pasternak, H. Seo, D. Lee, X.-J. Wang, A hierarchy of intrinsic timescales across primate cortex. *Nat Neurosci*. **17**, 1661–1663 (2014).

25. A. A. Russo, S. R. Bittner, S. M. Perkins, J. S. Seely, B. M. London, A. H. Lara, A. Miri, N. J. Marshall, A. Kohn, T. M. Jessell, L. F. Abbott, J. P. Cunningham, M. M. Churchland, Motor Cortex Embeds Muscle-like Commands in an Untangled Population Response. *Neuron*. **97**, 953-966.e8 (2018).

26. J. A. Harris, S. Mihalas, K. E. Hirokawa, J. D. Whitesell, H. Choi, A. Bernard, P. Bohn, S. Caldejon, L. Casal, A. Cho, A. Feiner, D. Feng, N. Gaudreault, C. R. Gerfen, N. Graddis, P. A. Groblewski, A. M. Henry, A. Ho, R. Howard, J. E. Knox, L. Kuan, X. Kuang, J. Lecoq, P. Lesnar, Y. Li, J. Luviano, S. McConoughey, M. T. Mortrud, M. Naeemi, L. Ng, S. W. Oh, B. Ouellette, E. Shen, S. A. Sorensen, W. Wakeman, Q. Wang, Y. Wang, A. Williford, J. W. Phillips, A. R. Jones, C. Koch, H. Zeng, Hierarchical organization of cortical and thalamic connectivity. *Nature*. **575**, 195–202 (2019).

27. S. Chung, L. F. Abbott, Neural population geometry: An approach for understanding biological and artificial neural networks. *Current Opinion in Neurobiology*. **70**, 137–144 (2021).

28. J. Hirokawa, A. Vaughan, P. Masset, T. Ott, A. Kepecs, Frontal cortex neuron types categorically encode single decision variables. *Nature*. **576**, 446–451 (2019).

29. G. Bondanelli, S. Ostojic, Coding with transient trajectories in recurrent neural networks. *PLoS Comput Biol*. **16**, e1007655 (2020).

30. I. M. Johnstone, A. Y. Lu, On Consistency and Sparsity for Principal Components Analysis in High Dimensions. *Journal of the American Statistical Association*. **104**, 682–693 (2009).

31. R. Kim, T. J. Sejnowski, Strong inhibitory signaling underlies stable temporal dynamics and working memory in spiking neural networks. *Nature Neuroscience*. **24**, 129–139 (2021).

32. A. Dubreuil, A. Valente, M. Beiran, F. Mastrogiuseppe, S. Ostojic, "Complementary roles of dimensionality and population structure in neural computations" (preprint, bioRxiv, 2020), , doi:10.1101/2020.07.03.185942.

33. A. E. Orhan, X. Pitkow, "Improved memory in recurrent neural networks with sequential non-normal dynamics." in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020* (2020; https://openreview.net/forum?id=ryx1wRNFvB).

34. V. Koren, A. R. Andrei, M. Hu, V. Dragoi, K. Obermayer, Pairwise Synchrony and Correlations Depend on the Structure of the Population Code in Visual Cortex. *Cell Reports*. **33**, 108367 (2020).

35. M. Román Rosón, Y. Bauer, A. H. Kotkat, P. Berens, T. Euler, L. Busse, Mouse dLGN Receives Functional Input from a Diverse Population of Retinal Ganglion Cells with Limited Convergence. *Neuron*. **102**, 462-476.e8 (2019).

36. J. C. Whittington, W. Dorrell, S. Ganguli, T. E. Behrens, Disentangling with Biological Constraints: A Theory of Functional Cell Types. *arXiv preprint arXiv:2210.01768* (2022).

37. V. Q. Vu, J. Lei, Minimax sparse principal subspace estimation in high dimensions. *Ann. Statist.* **41** (2013), doi:10.1214/13-AOS1151.

38. I. Higgins, L. Chang, V. Langston, D. Hassabis, C. Summerfield, D. Tsao, M. Botvinick, Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons. *Nat Commun*. **12**, 6456 (2021).

39. A. Hyvärinen, E. Oja, Independent component analysis: algorithms and applications. *Neural Networks*. **13**, 411–430 (2000).

40. A. Hyvarinen, "Fast ICA for noisy data using Gaussian moments" in (IEEE, 1999), vol. 5, pp. 57–61.

41. A. Nieder, D. J. Freedman, E. K. Miller, Representation of the Quantity of Visual Items in the Primate Prefrontal Cortex. *Science*. **297**, 1708–1711 (2002).

42. H. Lee, A. Battle, R. Raina, A. Y. Ng, "Efficient sparse coding algorithms" in *Advances in neural information processing systems* (2007), pp. 801–808.

43. O. Ledoit, M. Wolf, A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*. **88**, 365–411 (2004).

44. S. Shinomoto, H. Kim, T. Shimokawa, N. Matsuno, S. Funahashi, K. Shima, I. Fujita, H. Tamura, T. Doi, K. Kawano, N. Inaba, K. Fukushima, S. Kurkin, K. Kurata, M. Taira, K.-I. Tsutsui, H. Komatsu, T. Ogawa, K. Koida, J. Tanji, K. Toyama, Relating Neuronal Firing Patterns to Functional Differentiation of Cerebral Cortex. *PLoS Comput Biol*. **5**, e1000433 (2009).

Acknowledgments

Funding:
German Research Foundation DFG JA 1999/1-1 (S.N.J.)
German Research Foundation DFG JA 1999/5-1 (S.N.J.)
German Research Foundation DFG JA 1999/6-1 (S.N.J.)
European Research Council ERC StG MEMCIRCUIT, GA 758032 (S.N.J.)
German Research Foundation DFG NI 618/10-1 (S.N.J.)
German Research Foundation DFG NI 618/13-1 (A.N.)

Author contributions:
Conceptualization: XXL, SNJ, AN
Methodology: XXL
Formal analysis: XXL
Investigation: SNJ
Visualization: XXL, SNJ
Supervision: SNJ
Writing—original draft: XXL, SNJ
Writing—review & editing: XXL, SNJ, AN

Competing interests:
The authors declare no competing interests.

Supplementary Figures

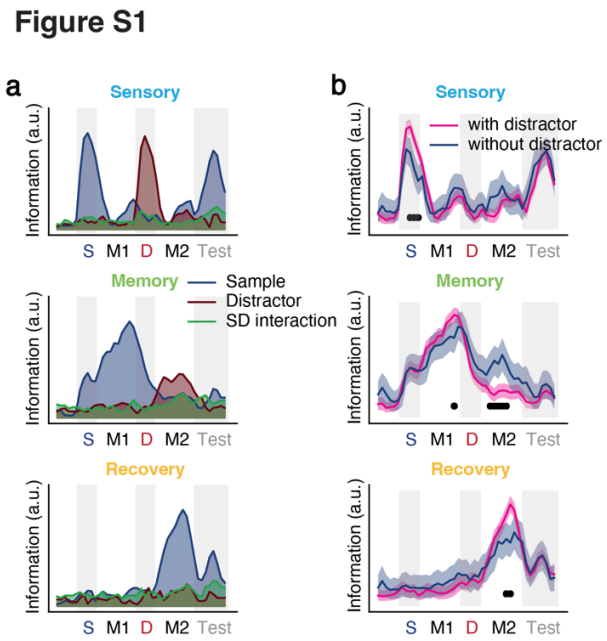


Fig. S1. The effect of distraction on sample numerosity sparse components

*(**a**) Information (standard deviation across conditions) about sample numerosity, distractor numerosity and their interaction in each of the three sample numerosity sparse components (SCs) in trials with a distractor. (**b**) Sample numerosity information as in (a) for the three SCs in trials with and without a distractor. Shaded area indicates [2.5 %, 97.5 %] confidence interval. Black dots indicate timepoints with significant differences ($p < 0.00125$, bootstrap).*

Figure S2

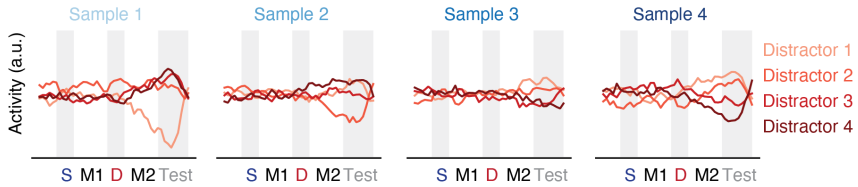


Fig. S2. Sample-distractor interaction sparse component

SCA performed on the demixed sample-distractor interaction part of the data identified one component that optimally reconstructed the data using cross-validation. The activity of this SC is shown for all sample-distractor combinations.

Chapter 3

General Discussion

In this thesis, I investigated higher cognition in higher-order brain regions with single-neuron resolution in primates performing tasks that involve complex temporal coordination and integration. I explored the population-level description of how the network of neurons is organized to represent task variables, offering insights into the neural mechanism underlying higher cognition.

In the first manuscript, I examined working memory representations of number stimuli in humans undergoing brain surgery. The intra-operative intracranial micro-electrode array recording allowed access to a broader range of brain regions, including the parietal association cortex. Capitalizing on the methodological advancements, I probed the population-level neural coding for the working memory maintenance of numbers in the parietal association cortex and discovered distinct representations for numbers in symbolic and nonsymbolic formats.

In the second manuscript, I furthered the investigation with a more complex temporal structure of the behavioral task and expanded coverage of single units. By considering the temporal modulation of multiple task variables represented in the population, I unveiled sparse structures in the neuronal implementation of representations. Such sparse neuronal implementations have often been observed in lower sensory systems but have not been reported in the prefrontal cortex (PFC). Leveraging the sparse structure, I identified biologically meaningful components of the representations that can be directly communicated to downstream neurons. Corroborating the physiological roots of these components, each component was linked to a small subpopulation of neurons, and it was found that these subpopulations have distinct physiological properties and temporal dynamics. These characteristics underlie their capacity to actively maintain working memory while resisting distraction. Lastly, using an artificial neural network model, I demonstrated that sparse implementation of the observed

temporally modulated working memory representations is preferred in recurrently connected neuronal populations such as the prefrontal cortex.

In the following sections, I will discuss the implications of these results in more details.

3.1 Symbolic and nonsymbolic number representations

In the first manuscript, I was able to target numerical cognition in symbolic format (Arabic numbers) that is unique to humans. Behaviorally, subjects showed less confusion between adjacent numbers (such as 2 and 3 or 5 and 6) in symbolic form and overall better performance in these trials. This effect was reflected in the recorded neuronal activity. The neuronal responses showed differences between symbolic and nonsymbolic trials, implying that the format was coded in the brain. Furthermore, neuronal representation for individual symbolic numbers emerged later than nonsymbolic numbers, suggesting additional processing was required for symbolic numbers. Finally, symbolic numbers required higher dimensionality in neuronal representation, allowing each number to be coded more idiosyncratically and reducing confusion with adjacent numbers.

The characteristics of the neuronal representation for numbers in symbolic and nonsymbolic support the direct involvement of the parietal association cortex in higher cognition:

1. Number format was represented. It is in line with the role of higher-order cortices that process not only the external stimuli but also cognitive variables such as task rules and trial types for executive control (Miller and Cohen, 2001).

2. Higher temporal integration functions necessitate a transition of neuronal representations from maintaining sensory input to adopting a format that supports motor output planning (Fuster, 2001). In the current task, a correct behavioral response depends on accurately remembering the exact number, irrespective of its magnitude (confusion with adjacent or distant numbers is equally incorrect). The observed idiosyncratic neuronal representations of symbolic numbers later in the delay period are suitable for subsequent motor planning.

3. From the perspective of mixed selectivity, the representation of symbolic numbers observed in the current study can be interpreted as non-linearly mixing magnitude with other number properties, such as parity or prime factors (Rigotti et al., 2013; Bernardi et al., 2020). A higher-dimensional geometry of neuronal representation allows for linearly reading out any number arbitrarily, thus providing flexible motor output for any number. In contrast, a low-dimensional magnitude code enables generalization to

unseen numbers and favors readout for very small and very large numbers (boundary effect). We observed better behavioral responses when subjects were memorizing small numbers, which is consistent with previous research reporting a Weber-Fechner law for the variability of number cognition in both behavior and neuronal tunings (Nieder and Miller, 2003). The neuronal representation of numbers in the parietal association cortex strikes a balance between extreme low-dimensional and high-dimensional scenarios, optimizing both flexibility and generalizability, consistent with previous reports in other higher-order cortices (Bernardi et al., 2020).

4. Lastly, there are additional meanings of symbolic numbers - instead of quantity, it could signify order (Nieder, Diester, and Tudusciuc, 2006; Nieder and Dehaene, 2009). In daily life, number can also be used as a label (nominal number). The polysemy of symbolic number is typical for higher cognition, requiring integrating information of various modalities. The need to represent the many aspects of symbolic numbers may also underlie their higher dimensionality of neuronal representation than nonsymbolic numbers.

My results for the dimensionality of number coding presented here were limited by the relatively small pool of sampled neurons. It could lead to underestimation of dimensionality. Certain aspects of number coding in working memory may be carried out by small subsets of neurons that are not easily captured with the small sample size. This motivated me to further investigate the questions of population coding for working memory in the second manuscript where more neurons were recorded.

3.2 Neuronal organization in population coding

In the second manuscript, I sought to bridge the gap between the population and single neuron doctrines (Saxena and Cunningham, 2019) by exploiting the structures in the neuronal implementation of working memory. The division of these two perspectives stems from our heuristics of the principle of neuronal organization in a local network that can be epitomized by the fundamental debate of grandmother cell vs. distributed coding in high level visual processing (Gross, 2002; Quiroga et al., 2008). The grandmother cell concept suggests that one neuron in the inferotemporal cortex (IT) responds to only one specific face, such as someone's grandmother. This means the stimuli coded in the neuronal population can be fully reflected in the response of certain single neurons. Conversely, distributed coding posits that the perception of any face is represented by the whole population, with single neurons not tuned to specific faces, therefore meaningful representations only arise at the population level.

While some IT neurons tuned to specific individuals' faces have been reported (Quiroga et al., 2005), it is impossible to prove the grandmother cell proposition, which would require sampling all neurons and all possible faces. Instead of specific faces, it is common to find individual neurons preferentially tuned to certain canonical facial features, such as gender, age, and hair length (Higgins et al., 2021; Freiwald, Tsao, and Livingstone, 2009), which could lead to sparse neuronal representations of faces (Quiroga et al., 2008).

Between the extremes of grandmother cell and distributed population coding, functions could be traced to subsets of neurons in a population, with a population vector indicating the contribution of each neuron. In the second manuscript, dominant subpopulations of neurons with high contributions to working memory components were selected, and they exhibited distinct single-neuron physiological properties. This physiological heterogeneity of neurons in the population may underlie their functional segregation, further supporting the division between dominant subpopulations and the elemental roles of the corresponding components. However, it is not clear whether the dominant neurons and non-dominant neurons for one working memory component are strictly disparate. On one hand, it is possible that only the "dominant" neurons participate in the computation involved with this component, as the non-dominant neurons of a certain component may only show correlation to the corresponding component due to recurrence and local inhibition in the PFC population; on the other hand, the observed distribution of neuronal loading was sparse but continuous, with no absolute criterion for counting a neuron as dominant.

Instead of assigning a subset of neurons to each function, neuronal organization for population coding can be understood as a hierarchical local network with differently ranked roles for each neuron. For example, it has been shown that the count of hippocampal place cells exhibits an exponential decay distribution over their place field size, with fewer cells having large place fields (Zhang et al., 2023). This exponential decay is the hallmark of hierarchy, where less frequent occurrence signifies a higher rank (Zipf's law) (Sharpee, 2019; Zhou, Smith, and Sharpee, 2018). Neurons with large place fields are thus higher ranked in the network. They are connected to a larger number of other place cells and have overlapping place fields with them. These neurons represent the coarse location, while neurons with small place fields represent the fine-grained location.

Such neuronal organization is optimal for encoding both physical space and abstract variable space, as well as for updating the network to adapt to new experiences, such as adding new small place fields to the network as the animal becomes more familiar with the space (Zhang et al., 2023). This concept can be extended to PFC working memory neurons. In my study, the observed neuronal loadings on working memory components followed a Laplace distribution, which is a combination of two exponential distributions on the positive

and negative sides. This distribution of neuronal loadings supports a hierarchical organization in which a small fraction of dominant coding neurons are responsible for fundamental and general computations of working memory representations. The majority of neurons with small absolute loadings were not simply silent but were responsible for computations in other specific contexts or scenarios not optimally probed by our behavioral task.

The question then arises whether this principle of hierarchical organization for population coding is ubiquitous across the brain. The isocortex is built by repeating structures (Wang, 2020). It is reasonable that local hierarchies underlie the macro-scale hierarchical organization across the brain. However, neuronal representations in higher-order cortices are often assumed to be distributed (Rigotti et al., 2013; Bernardi et al., 2020). Higher-order cortices have different physiological and network properties and functions that may make the hierarchical or sparse organization not easily reflected in experiments. First, in experiments regarding visual perception, subjects are usually confronted with a large number of visual stimuli, such as faces, probing the possible stimuli that neurons may best respond to. Neurons in higher-order cortices usually respond to more abstract task variables that are not simply stimulus-driven. The number of those variables that can be probed in a single experiment is limited, due to the repetition required for training and the time needed for complex task structures. We can only access a small fraction of the full task variable space, making it harder to "hit" the variables neurons optimally respond to. Second, temporal integration of both information maintenance and action planning is a key function of higher-order cortices. Neuronal activity usually shows complex temporal modulation instead of simply following stimulus presentation. Yet, temporally modulated neuronal activity is typically averaged in a pre-selected time window (usually during the delay period) when constructing the neural subspace (Murray et al., 2017; Parthasarathy et al., 2017; Bernardi et al., 2020). This may cause the analysis to miss variables with specific temporal modulation that neurons prefer.

3.3 Factorization of neuronal representations

The presence of the aforementioned organizing principles often results in inhomogeneity across the variables represented, meaning that not all aspects of the task variable space are equally represented in the neuronal population. Understanding which specific aspects are implemented differently by neurons can provide insights into the detailed operations within the brain. Neuronal populations, especially in higher-order cortices, tend to concentrate most of their activity variability in a low-dimensional latent space where task variables can be factorized (Bernardi et al., 2020). In this space, neurons preferentially represent disentangled factors that satisfy compositionality, meaning that these factors can be treated more or less

independently and the order of applying them does not matter, such as the color and size of an object (Higgins et al., 2021). This concept of disentangled factors allows for a more efficient and flexible representation of the task variable space, ultimately contributing to the brain's ability to process and adapt to complex information.

My results first highlighted the disentangled sample and distractor number memory in distinct subpopulations. This factorization is crucial because it demonstrates that animals have learned to treat the two stimuli as separate variables, rather than repeated or sequential samplings of one variable. Interestingly, animals also internally formed the distractor representation, even though the distractor number is not relevant to the behavioral output in the current task. This finding suggests that animals spontaneously create factorized representations of input without requiring reinforcement. Such a mechanism may underlie their ability to quickly generalize to new environments and task settings, such as an n-back working memory task.

Moreover, the temporal modulation of sample number memory was factorized into representations in different task periods. At first glance, this may seem counter-intuitive. If we were to treat the representations in PFC as serving only the memory maintenance function, we would expect a single sustained representation throughout the entire delay. However, the factorized temporal modulation of working memory representations indicates that the observed PFC representations support more complex functions than mere maintenance. One possibility is that this temporally factorized representation results from temporal integration with action planning. As new information appears and a new task period begins, organisms might need to update their contingent action plans (Ehrlich and Murray, 2022). Consequently, the significance and output contingency of past sensory input may change in new task periods. In other words, the same past sensory input could correspond to different variables in different contexts, making it beneficial for its representation to be factorized in various task periods. In contrast, the neuronal representation of sample number during the first delay and the neuronal representation of distractor number during the second delay largely overlap, indicating that the PFC treats them as the same cognitive variable, even if they correspond to two separate external stimuli.

Additionally, neuronal disentanglement of task variables in the brain may reflect energy-efficient coding. This efficiency can be attributed to the geometry formed by the combination of variables. For variables that are independently and uniformly distributed within a range, the resulting geometry features corners and edges at the extreme values of these variables. The neural state space also has a corner and edges at neurons' low firing states, constrained by the non-negativity of neuronal firing rates. Optimal use of neuronal firings (reducing the highest firing rate needed) occurs when individual neurons code each independent variable rather

than their linear mixture, fitting the geometry of variables to the boundaries of the neural state space (Whittington et al., 2022). In this context, the disentanglement of sample and distractor number representations at the second delay may stem from the grid-like geometry of their combination. Intriguingly, the neuronal factorization of sample representation in different task periods might also arise from the geometry of its temporal modulation in state space. The trajectory exhibited a sharp turn at the transition of task periods. To efficiently implement this geometry, the turning point should be situated at the corner of the neural state space, corresponding to the scenario when individual neurons contribute the trajectory either before or after the turning point.

3.4 Temporal dynamics of working memory

To factorize memory representations in different task periods, it is necessary to first recognize and register the temporal structure of the task. This temporal organization ability may be central to PFC functions (Fuster, 2001). The majority of the neuronal activity's variability I observed was explained by the factor of trial time, irrespective of number stimuli. This could form the basis of how the PFC registers task structure in order to modulate memory representations. Due to the non-linearity of the input-output relationship in neurons, their sensitivity to small perturbations varies depending on the general activity level (Dubreuil et al., 2022). The trial time signal, which determines the general activity level of neurons, can act as a gating mechanism for the memory signal that is much smaller in scale.

Notably, trial time was represented differently in the three dominant subpopulations. The sensory and memory subpopulations represent trial time periodically, corresponding to the periodic sensory input. Consequently, these two subpopulations followed a temporal structure that is more input-driven. They represented the most recent sensory input regardless of whether it was the sample or distractor number. The trial time signal in recovery subpopulation, on the other hand, could discriminate between the first and the second delays. This subpopulation only represented sample number at the second delay and ignored distractor number. These results support the function of trial time related activity in modulating and controlling the memory activity.

The question then arises: where does the trial time signal come from? Naturally, trial time could be computed based on the decaying trace of sensory input and the accumulating expectation for reward. However, a more intricate representation that involves certain mental construction of the trial structure requires further computation in the local circuitry. The analysis of tangling showed that the trial time activity in the recovery subpopulation could be maintained locally, while the trial time activity in sensory and memory subpopulations

relied on external input. This suggests that the recovery subpopulation may be responsible for internally constructing the trial structure. In this case, the three subpopulations were not solely responsible for different factorized aspects of memory representation, but also belonged to different functional hierarchies, with the recovery subpopulation being more involved in providing control signals. Nevertheless, the current results do not exclude the possibility that other subpopulations could feed different trial time signals to these three dominant memory subpopulations.

3.5 Active information maintenance

The manuscripts in this thesis have portrayed the sequential representation of working memory across task periods in higher-order cortices, with a possible contextual control signal in the recovery subpopulation to resist distraction. In this section, I will clarify its relation and differences with similar working memory theories.

Working memory maintenance is often thought to be implemented through persistent neuronal activity, which stems from the early discovery of neurons that persistently fire in the delay period (Fuster, 2001). This persistent activity in the absence of sensory input can be modeled by the dynamics of local networks, e.g., the bi-stable states of local circuits (Camperi and Wang, 1998; Brunel and Wang, 2001), and further generalized to continuous variables with a bump attractor model (Wimmer et al., 2014). In this perspective, network dynamics possess several stable fixed points that allow memory to persist, and maintaining memory means keeping neural states stable (Murray et al., 2017). Stable fixed points usually result from symmetric connections. In contrast, an asymmetrically connected network can also maintain memory by relaying it through different network states. This approach may have better resistance to distraction (Orhan and Pitkow, 2020) but requires a fixed delay until readout (Murray et al., 2017). Memory representations in my results were stable within each task period and sequential across periods. This could be due to the combination of the behavioral task and the natural environment. In a natural environment, an organism might not know how long it needs to hold relevant information, so employing stable fixed points should be the default strategy. In our task setting, the animals were trained on a task with a fixed time structure and had to resist distracting stimuli. Therefore, a coarse-grained sequential representation changing only at the transition to a new task period may be more suitable. However, as discussed in previous sections, the sequential representation is not necessarily responsible for maintenance but rather for temporal organization and flexible behavioral output.

Furthermore, it has been proposed that instead of persistent activity, neural systems use intermittent bursts of activity for the maintenance of information during working memory (Buschman and Miller, 2022). Maintaining information during activity-silent periods requires changes in synaptic weight. The sequential representation observed in my current study does not prove or exclude such a possibility. Resolving this debate would require a new methodology with more microscopic specificity.

Working memory is the active, rather than passive, maintenance of behaviorally relevant information, which includes the ability to resist distraction. It is often conceived in the form of filtering out distractions, involving specific types of inter-neurons in local circuitry and specific neuromodulators such as dopamine (Brunel and Wang, 2001; Wang, 2020; Ott, Jacob, and Nieder, 2014). In my study, the distractor was not selectively suppressed but coexisted with sample information in the PFC. Although the recovery subpopulation did not represent the distractor, it also did not hold sample information from the beginning. This presents an unconventional mechanism for resolving working memory tasks with distractors: instead of filtering out distractions, it maintains both sample and distractor information in different information channels and uses the information according to the behavioral context. These results provided a fresh perspective for examining the dorsolateral PFC neuronal population in greater detail, as opposed to previous descriptions that assume the entire population serves a single functional purpose.

3.6 Interpreting sparsity constraint

The primary analyses in my second manuscript were inspired by the anatomical observation that cortical neurons have significantly fewer dendritic spines (approximately 10^4) than the total number of neurons (approximately 10^9) in an upstream area, such as the dorsolateral PFC (Herculano-Houzel et al., 2015; Courchesne et al., 2011; Eyal et al., 2018). This discrepancy limits the number of neurons in the dorsolateral PFC that can directly project to each downstream neuron. To uncover the information communicated by the dorsolateral PFC and identify the neurons responsible for this communication, I employed sparse component analysis (SCA), a methodology that capitalizes on the statistical principle that non-Gaussianity, such as sparsity, leads to identifiable components (Hyvärinen and Oja, 2000; Ganguli and Sompolinsky, 2012). Although several sparsity-based methods share mathematical similarities with the methodology used in my study, their motivations and implications differ substantially.

The most common configuration of independent component analysis (ICA) also utilizes sparsity, but its motivation is to find the sparse source signals from mixed observations (Hy-

varinen, 1999). Capitalizing on the central limit theorem, which states that a mixture of independent random variables tends towards a Gaussian distribution, ICA identifies non-Gaussianity as a signature of independent sources before mixing. Consequently, sparsity is maximized for the inferred source signal, such as a picture or an audio sequence. In my study, the equivalent task would be finding the latent source with maximally sparse activity underlying the recorded neuronal activity. This approach does not encourage sparse loadings/mixing vectors, in contrast to the motivation of sparse component analysis (SCA).

To interpret the analyses in my study within the ICA framework, the population vector specifying the activity across neurons in one condition-timepoint combination should be viewed as the observed "signal vector." In this context, SCA in my study could be understood as finding the independent population vectors that were mixed in each condition-timepoint combination. Preprocessing the data to construct a low-dimensional subspace would be necessary; otherwise, the resulting independent population vectors would trivially be indicator vectors, each with one active neuron. Varimax rotation after principal component analysis (PCA) is conceptually similar to this "population vector ICA." It aims to enforce unique solutions and is commonly applied to improve the interpretability of factors (Kaiser, 1958; Rohe and Zeng, 2020). Unlike ICA, which is usually applied after preprocessing with principal component truncation and whitening, SCA here retains the covariance structure and finds the SCs directly, so the result is not limited by pre-selection of the subspace.

Dictionary learning or sparse coding is another method that utilizes sparsity (Kreutz-Delgado et al., 2003; Olshausen and Field, 1996). However, dictionary learning typically learns an overcomplete or complete set of factors, while in SCA, the dimensionality is significantly reduced. This difference is related to the fact that dictionary learning is usually applied to a vast set of natural image data, where one designs the tuning of hypothetical neurons to make it energy-efficient by having sparse activity. In contrast, SCA is applied to the activity of neurons to find what signal they might be communicating, given the sparsity of synapses, not necessarily leading to less activity in the components.

The mathematical formulation of SCA in my study is most similar to sparse principal component analysis (SPCA) (Zou, Hastie, and Tibshirani, 2006). SPCA requires the factors to be linear projections of the original data, thus only utilizing the covariance among neurons and ignoring the exact activity information. In contrast, SCA could potentially capture factors that are not within the linear span of neuronal activity. This aspect may render SCA more suitable for uncovering the latent factors with complex temporal modulations that are not directly reflected in the recorded neuronal activity.

Sparsity-based methods are useful for optimally compressing and representing a wide range of sensory input. These methods are commonly used as a model of the "design

principle" of efficient coding in sensory systems (Ganguli and Sompolinsky, 2012). In my study, the method was applied to neural recordings during a working memory task with a limited set of sensory stimuli and a focus on temporal modulation. The identified components represent activity patterns that reflect computations in distinct neuronal ensembles, rather than a compressed representation of sensory inputs.

The sparsity constraint is often applied to the weight matrix as an engineering approach for feature/channel selection and addressing the mathematically ill-posed problem arising from high-dimension, low-sample-size data – a common issue in neural imaging studies (Haufe et al., 2014). The physiological meaning of the neural signal determines the implications of a sparse prior on the weights. In extracellular recordings, where recorded units relate directly to single neurons, the sparsity constraint on decoding weights corresponds to the sparsity of connections from recorded neurons to downstream neurons, resulting in activity patterns communicable by the local circuit. For imaging modalities such as functional magnetic resonance imaging (fMRI) and electroencephalography (EEG), the sparsity constraint across voxels or channels does not reflect any macroscale physiological characteristics of the brain and is primarily applied for engineering purposes, with limited physiological implications (Haufe et al., 2014; Friston et al., 2008).

Data interpretation in multivariate analyses can be categorized into two classes: forward/encoding models that generate data from latent processes, and backward/decoding models that extract information from data. Forward/encoding models have been suggested to be more stable and precise when interpreting weight matrices (Haufe et al., 2014). In my study, I formulated SCA as an encoding model instead of analyzing the weights of a linear discriminant analysis decoder. This approach presents a limitation when linking the discovered neuronal implementation to potential readout weights, as sparse encoding weights do not always result in sparse readout weights. An additional condition of orthogonality of activity components in the whitened space must be satisfied, which was observed in my analyses. The distinction between encoding sparsity and decoding sparsity should be carefully considered when applying the current approach in future research.

3.7 Outlook

Understanding the neural mechanisms underlying higher cognition has been an ongoing challenge since Descartes' time. In this thesis, I have delved into the complexities of the brain's representation of task variables with single-neuron resolution in primates performing working memory tasks. In the following paragraphs, I will discuss future research directions and potential avenues to build upon the findings presented in this thesis.

One key area of focus should be the executive control function, which guides the active maintenance of behaviorally relevant information and resists distractions. My results demonstrate that task-structure related temporal dynamics in population activity can be used as a control signal. However, it is still unclear how this control signal is generated and communicated to memory-representing neurons. To understand the executive control function in higher cognition, future research needs to investigate how neuronal ensembles exhibiting control signals interact with other neurons in the local network and how higher-order cortices interact with other brain regions. This will require simultaneous recording of large populations of neurons and computational modeling of these populations to infer possible connection patterns.

Memory maintenance is only the tip of the iceberg of higher cognition. The complex neuronal representations in higher-order cortices need to be considered within a broader context - how organisms interact with their environment - in order to fully reveal their functional significance. Evidence has shown that the PFC can be compared to the recurrent neural network in a reinforcement learning agent, responsible for registering past actions, rewards, and most importantly, the current latent state of the task (Botvinick et al., 2019; Wang et al., 2018). A similar view posits that the persistent activity in the PFC encodes the transition probability of latent states, with the PFC's anatomy being suitable for Bayesian belief updating (Parr et al., 2020). Trial time encoding units and sequential memory coding units, akin to the neurons described in this thesis, can be found in deep reinforcement learning agents performing working memory tasks (Lin and Richards, 2021). Therefore, to appropriately probe the higher cognition functions that are closely intertwined with learning, investigations of both behavior and neural signatures, should extend beyond well-trained stages of a task, focusing on the entire task acquisition and even the generalization to new tasks (Bernklau, 2022).

Experiments need to encompass a wider range of task variables. Current experimental designs for investigating higher cognition probe a limited space of cognitive variables. Expanding the dimensionality of the task variable space may help identify the latent variables neurons optimally respond to and discover the population implementation structure more accurately (Stringer et al., 2019). In sensory-driven systems, this can be achieved by increasing the number of stimuli. For higher cognition, possible approaches are to introduce multi-task settings, apply randomized behavioral perturbations, and record spontaneous behaviors instead of only investigating a limited set of heavily trained behaviors.

Models and statistical tools need to adjust accordingly. First, rather than averaging neuronal activity across trials, methods based on general linear models that allow for spontaneous timing of events are more suitable (Aoi, Mante, and Pillow, 2020). Second, when considering

spontaneous behaviors, it is crucial to extract causality information from numerous aspects that are only quasi-experimentally controlled (Marinescu, Lawlor, and Kording, 2018). Third, brain models need updating to accommodate the possible compositionality of functions across multiple tasks (Yang et al., 2019) and account for the rich influence of the natural environment in shaping behavior (Molano-Mazón et al., 2023).

Future neuronal implementation analyses could benefit from large-scale simultaneous recordings, such as two-photon calcium imaging, which allows tracking neurons across multiple sessions. My study used single units aggregated from various sessions in the well-trained stage. Neuronal implementation could be more accurately identified with simultaneous large-scale neural recordings that preserve noise correlation and allow tracking neurons across sessions. Additionally, investigating the stability of neuronal implementation at various time scales could help disentangle subpopulations of neurons that exhibit similar activity in one learning stage and further unveil the dynamics of neuronal organization.

Finally, theories and analyses should focus on physiological principles of the brain. Descartes correctly posited the preservation of retinotopy in the brain based on nerve connections (see Introduction). Similarly, based on the physiological property of sparse neuronal connections, distinct subpopulations for memory representation were identified in this thesis, updating the intuition that memory is continuously maintained in one population. Generally, our heuristics of possible cognitive variables are often biased by subjective introspection of mental processes. Uncovering a more accurate description of the mind requires building theories upon the physical characteristics of its material essence - the brain (Cornman, 1968). Distilled from various aspects of physiology, several "first principles" have been proposed to guide a systemic understanding of the brain (Chen et al., 2023). This thesis touches on sparse coding, but more inspiration can be found in, for example, criticality of dynamical systems that could underlie PFC's neuronal activity enabling diverse output, and neural plasticity rules such as Hebbian learning that could be crucial in explaining temporal integration functions.

References

Amunts, Katrin and Karl Zilles (2015). "Architectonic mapping of the human brain beyond Brodmann". In: *Neuron* 88.6, pp. 1086–1107.

Andersen, Richard A and He Cui (2009). "Intention, action planning, and decision making in parietal-frontal circuits". In: *Neuron* 63.5, pp. 568–583.

Aoi, Mikio C., Valerio Mante, and Jonathan W. Pillow (Nov. 2020). "Prefrontal cortex exhibits multidimensional dynamic encoding during decision-making". en. In: *Nature Neuroscience* 23.11, pp. 1410–1420. ISSN: 1097-6256, 1546-1726. DOI: 10.1038/s41593-020-0696-5.

Baddeley, Alan (1992). "Working memory". In: *Science* 255.5044, pp. 556–559.

Bao, Pinglei et al. (2020). "A map of object space in primate inferotemporal cortex". In: *Nature* 583.7814, pp. 103–108.

Barbas, H and J De Olmos (1990). "Projections from the amygdala to basoventral and mediodorsal prefrontal regions in the rhesus monkey". In: *Journal of Comparative Neurology* 300.4, pp. 549–571.

Bates, Julianna F and Patricia S Goldman-Rakic (1993). "Prefrontal connections of medial motor areas in the rhesus monkey". In: *Journal of Comparative Neurology* 336.2, pp. 211–228.

Bernardi, Silvia et al. (Nov. 2020). "The Geometry of Abstraction in the Hippocampus and Prefrontal Cortex". en. In: *Cell* 183.4, 954–967.e21. ISSN: 00928674. DOI: 10.1016/j.cell.2020.09.031.

Bernklau, Tobias (2022). "Dopamine signaling across striatal subregions during acquisition of instrumental associations". PhD thesis. lmu.

Botvinick, Matthew et al. (2019). "Reinforcement learning, fast and slow". In: *Trends in cognitive sciences* 23.5, pp. 408–422.

Brooks, Rodney A (1991). "Intelligence without representation". In: *Artificial intelligence* 47.1-3, pp. 139–159.

Brunel, Nicolas and Xiao-Jing Wang (2001). "Effects of neuromodulation in a cortical network model of object working memory dominated by recurrent inhibition". In: *Journal of computational neuroscience* 11, pp. 63–85.

Buschman, Timothy J and Earl K Miller (2022). "Working memory is complex and dynamic, like your thoughts". In: *Journal of cognitive neuroscience* 35.1, pp. 17–23.

Camperi, Marcelo and Xiao-Jing Wang (1998). "A model of visuospatial working memory in prefrontal cortex: recurrent network and cellular bistability". In: *Journal of computational neuroscience* 5, pp. 383–405.

Cavanagh, Sean E. et al. (Dec. 2018). "Reconciling persistent and dynamic hypotheses of working memory coding in prefrontal cortex". en. In: *Nature Communications* 9.1, p. 3498. ISSN: 2041-1723. DOI: 10.1038/s41467-018-05873-3.

Cazettes, Fanny et al. (2023). "A reservoir of foraging decision variables in the mouse brain". In: *Nature Neuroscience*, pp. 1–10.

Chen, Luyao et al. (2023). "AI of Brain and Cognitive Sciences: From the Perspective of First Principles". In: *arXiv preprint arXiv:2301.08382*.

Constantinidis, Christos and Xue-Lian Qi (2018). "Representation of spatial and feature information in the monkey dorsal and ventral prefrontal cortex". In: *Frontiers in Integrative Neuroscience* 12, p. 31.

Cornman, James W (1968). "On the elimination of'sensations' and sensations". In: *The Review of Metaphysics* 22.1, pp. 15–35.

Courchesne, Eric et al. (2011). "Neuron number and size in prefrontal cortex of children with autism". In: *Jama* 306.18, pp. 2001–2010.

Cunningham, John P and Byron M Yu (2014). "Dimensionality reduction for large-scale neural recordings". In: *Nature neuroscience* 17.11, pp. 1500–1509.

Dayan, Peter and Laurence F Abbott (2005). *Theoretical neuroscience: computational and mathematical modeling of neural systems*. MIT press.

Dehaene, Stanislas et al. (2003). "Three parietal circuits for number processing". In: *Cognitive neuropsychology* 20.3-6, pp. 487–506.

Descartes, René (1965). *Discourse on Method, Optics, Geometry, and Meteorology*. Trans. by P.J. Olscamp. Library of liberal arts. Bobbs-Merrill. ISBN: 9780672604591. URL: https://books.google.de/books?id=Wz0PAQAAMAAJ.

Di Pellegrino, Giuseppe, Elisa Ciaramelli, and Elisabetta Ladavas (2007). "The regulation of cognitive control following rostral anterior cingulate cortex lesion in humans". In: *Journal of cognitive neuroscience* 19.2, pp. 275–286.

Diester, Ilka and Andreas Nieder (2007). "Semantic associations between signs and numerical categories in the prefrontal cortex". In: *PLoS biology* 5.11, e294.

Dubreuil, Alexis et al. (2022). "The role of population structure in computations through neural dynamics". In: *Nature neuroscience* 25.6, pp. 783–794.

Ehrlich, Daniel B. and John D. Murray (2022). "Geometry of neural computation unifies working memory and planning". In: *Proceedings of the National Academy of Sciences* 119.37, e2115610119. DOI: 10.1073/pnas.2115610119. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2115610119. URL: https://www.pnas.org/doi/abs/10.1073/pnas.2115610119.

Eyal, Guy et al. (2018). "Human cortical pyramidal neurons: from spines to spikes via models". In: *Frontiers in cellular neuroscience* 12, p. 181.

Felleman, Daniel J and David C Van Essen (1991). "Distributed hierarchical processing in the primate cerebral cortex." In: *Cerebral cortex (New York, NY: 1991)* 1.1, pp. 1–47.

Freiwald, Winrich A, Doris Y Tsao, and Margaret S Livingstone (2009). "A face feature space in the macaque temporal lobe". In: *Nature neuroscience* 12.9, p. 1187.

Friston, Karl et al. (2008). "Multiple sparse priors for the M/EEG inverse problem". In: *NeuroImage* 39.3, pp. 1104–1120.

Froudist-Walsh, Sean et al. (2021). "A dopamine gradient controls access to distributed working memory in the large-scale monkey cortex". In: *Neuron* 109.21, 3500–3520.e13. ISSN: 0896-6273. DOI: https://doi.org/10.1016/j.neuron.2021.08.024. URL: https://www.sciencedirect.com/science/article/pii/S0896627321006218.

Fuster, Joaquin (2015). *The prefrontal cortex*. Academic press.

Fuster, Joaquin M (1990). "Prefrontal cortex and the bridging of temporal gaps in the perception-action cycle." In: *Annals of the New York Academy of Sciences*.

Fuster, Joaquın M (2001). “The prefrontal cortex—an update: time is of the essence”. In: *Neuron* 30.2, pp. 319–333.

Ganguli, Surya and Haim Sompolinsky (2012). “Compressed sensing, sparsity, and dimensionality in neuronal information processing and data analysis”. In: *Annual review of neuroscience* 35, pp. 485–508.

Gordon, Peter (2004). “Numerical cognition without words: Evidence from Amazonia”. In: *Science* 306.5695, pp. 496–499.

Gross, Charles G (2002). “Genealogy of the “grandmother cell””. In: *The Neuroscientist* 8.5, pp. 512–518.

Halberda, Justin and Lisa Feigenson (2008). “Developmental change in the acuity of the" Number Sense": The Approximate Number System in 3-, 4-, 5-, and 6-year-olds and adults.” In: *Developmental psychology* 44.5, p. 1457.

Harris, Julie A. et al. (Nov. 2019). “Hierarchical organization of cortical and thalamic connectivity”. en. In: *Nature* 575.7781, pp. 195–202. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-019-1716-z. URL: https://www.nature.com/articles/s41586-019-1716-z (visited on 01/12/2023).

Haufe, Stefan et al. (2014). “On the interpretation of weight vectors of linear models in multivariate neuroimaging”. In: *Neuroimage* 87, pp. 96–110.

Hauser, Marc D, Susan Carey, and Lilan B Hauser (2000). “Spontaneous number representation in semi–free–ranging rhesus monkeys”. In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 267.1445, pp. 829–833.

Hayes, Patrick J (1981). “The frame problem and related problems in artificial intelligence”. In: *Readings in Artificial Intelligence*. Elsevier, pp. 223–230.

Herculano-Houzel, Suzana et al. (2015). “Mammalian brains are made of these: a dataset of the numbers and densities of neuronal and nonneuronal cells in the brain of glires, primates, scandentia, eulipotyphlans, afrotherians and artiodactyls, and their relationship with body mass”. In: *Brain, Behavior and Evolution* 86.3-4, pp. 145–163.

Higgins, Irina et al. (Dec. 2021). “Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons”. en. In: *Nature Communications* 12.1, p. 6456. ISSN: 2041-1723. DOI: 10.1038/s41467-021-26751-5.

Hyvarinen, Aapo (1999). “Fast ICA for noisy data using Gaussian moments”. In: *1999 IEEE international symposium on circuits and systems (ISCAS)*. Vol. 5. IEEE, pp. 57–61.

Hyvärinen, Aapo and Erkki Oja (2000). “Independent component analysis: algorithms and applications”. In: *Neural networks* 13.4-5, pp. 411–430.

Izquierdo, Alicia, Robin K Suda, and Elisabeth A Murray (2005). “Comparison of the effects of bilateral orbital prefrontal cortex lesions and amygdala lesions on emotional responses in rhesus monkeys”. In: *Journal of Neuroscience* 25.37, pp. 8534–8542.

Jacob, Simon Nikolas and Andreas Nieder (July 2014). “Complementary Roles for Primate Frontal and Parietal Cortex in Guarding Working Memory from Distractor Stimuli”. en. In: *Neuron* 83.1, pp. 226–237. ISSN: 08966273. DOI: 10.1016/j.neuron.2014.05.009.

Kaiser, Henry F (1958). “The varimax criterion for analytic rotation in factor analysis”. In: *Psychometrika* 23.3. publisher: Springer, pp. 187–200. ISSN: 1860-0980.

Kenny, Anthony J.P. (1971). “The Homunculus Fallacy”. In: *Interpretations of Life and Mind: Essays Around the Problem of Reduction*. Ed. by Marjorie Grene. New York: Humanities Press. Chap. 6, pp. 155–165.

Kravitz, Dwight J et al. (2013). “The ventral visual pathway: an expanded neural framework for the processing of object quality”. In: *Trends in cognitive sciences* 17.1, pp. 26–49.

Kremkow, Jens et al. (2016). "Principles underlying sensory map topography in primary visual cortex". In: *Nature* 533.7601, pp. 52–57.

Kreutz-Delgado, Kenneth et al. (2003). "Dictionary learning algorithms for sparse representation". In: *Neural computation* 15.2, pp. 349–396. ISSN: 0899-7667.

Kutter, Esther F et al. (2018). "Single neurons in the human brain encode numbers". In: *Neuron* 100.3, pp. 753–761.

Libby, Alexandra and Timothy J. Buschman (May 2021). "Rotational dynamics reduce interference between sensory and memory representations". en. In: *Nature Neuroscience* 24.5, pp. 715–726. ISSN: 1097-6256, 1546-1726. DOI: 10.1038/s41593-021-00821-9.

Lin, Dongyan and Blake A Richards (2021). "Time cell encoding in deep reinforcement learning agents depends on mnemonic demands". In: *bioRxiv*, pp. 2021–07.

Lindsay, Grace W (2021). "Convolutional neural networks as a model of the visual system: Past, present, and future". In: *Journal of cognitive neuroscience* 33.10, pp. 2017–2031.

Lund, Jennifer S, Alessandra Angelucci, and Paul C Bressloff (2003). "Anatomical substrates for functional columns in macaque monkey primary visual cortex". In: *Cerebral cortex* 13.1, pp. 15–24.

MacFall, James R et al. (2001). "Medial orbital frontal lesions in late-onset depression". In: *Biological psychiatry* 49.9, pp. 803–806.

Mante, Valerio et al. (Nov. 2013). "Context-dependent computation by recurrent dynamics in prefrontal cortex". en. In: *Nature* 503.7474, pp. 78–84. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature12742.

Marinescu, Ioana E, Patrick N Lawlor, and Konrad P Kording (2018). "Quasi-experimental causality in neuroscience and behavioural research". In: *Nature human behaviour* 2.12, pp. 891–898.

Masset, Paul et al. (2020). "Behavior-and modality-general representation of confidence in orbitofrontal cortex". In: *Cell* 182.1, pp. 112–126.

Meck, Warren H, Russell M Church, and John Gibbon (1985). "Temporal integration in duration and number discrimination." In: *Journal of Experimental psychology: animal behavior processes* 11.4, p. 591.

Michaels, Jonathan A, Benjamin Dann, and Hansjörg Scherberger (2016). "Neural population dynamics during reaching are better explained by a dynamical system than representational tuning". In: *PLoS computational biology* 12.11, e1005175.

Miller, Earl K and Jonathan D Cohen (2001). "An integrative theory of prefrontal cortex function". In: *Annual review of neuroscience* 24.1, pp. 167–202.

Miller, Earl K, Mikael Lundqvist, and André M Bastos (2018). "Working Memory 2.0". In: *Neuron* 100.2, pp. 463–475.

Molano-Mazón, Manuel et al. (2023). "Recurrent networks endowed with structural priors explain suboptimal animal behavior". In: *Current Biology*.

Molnár, Zoltán and Kathleen S Rockland (2020). "Cortical columns". In: *Neural Circuit and Cognitive Development*. Elsevier, pp. 103–126.

Murray, John D et al. (Dec. 2014). "A hierarchy of intrinsic timescales across primate cortex". en. In: *Nature Neuroscience* 17.12, pp. 1661–1663. ISSN: 1097-6256, 1546-1726. DOI: 10.1038/nn.3862.

Murray, John D. et al. (Jan. 10, 2017). "Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex". en. In: *Proceedings of the National Academy of Sciences* 114.2, pp. 394–399. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1619449114.

Nieder, Andreas and Stanislas Dehaene (2009). "Representation of number in the brain". In: *Annual review of neuroscience* 32, pp. 185–208.
Nieder, Andreas, Ilka Diester, and Oana Tudusciuc (2006). "Temporal and spatial enumeration processes in the primate parietal cortex". In: *Science* 313.5792, pp. 1431–1435.
Nieder, Andreas, David J. Freedman, and Earl K. Miller (Sept. 6, 2002). "Representation of the Quantity of Visual Items in the Primate Prefrontal Cortex". en. In: *Science* 297.5587, pp. 1708–1711. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1072493.
Nieder, Andreas and Earl K Miller (2003). "Coding of cognitive magnitude: Compressed scaling of numerical information in the primate prefrontal cortex". In: *Neuron* 37.1, pp. 149–157.
Nieder, Andreas, Lysann Wagener, and Paul Rinnert (2020). "A neural correlate of sensory consciousness in a corvid bird". In: *Science* 369.6511, pp. 1626–1629.
Niemi, Pekka and Risto Näätänen (1981). "Foreperiod and simple reaction time." In: *Psychological bulletin* 89.1, p. 133.
Olshausen, Bruno A and David J Field (1996). "Emergence of simple-cell receptive field properties by learning a sparse code for natural images". In: *NATURE* 381, p. 13.
Orhan, A. Emin and Xaq Pitkow (Feb. 10, 2020). "Improved memory in recurrent neural networks with sequential non-normal dynamics". en. In: *arXiv:1905.13715 [cs, stat]*. arXiv: 1905.13715. URL: http://arxiv.org/abs/1905.13715.
Ostlund, Sean B and Bernard W Balleine (2005). "Lesions of medial prefrontal cortex disrupt the acquisition but not the expression of goal-directed learning". In: *Journal of Neuroscience* 25.34, pp. 7763–7770.
Ott, Torben, Simon Nikolas Jacob, and Andreas Nieder (2014). "Dopamine receptors differentially enhance rule coding in primate prefrontal cortex neurons". In: *Neuron* 84.6, pp. 1317–1328.
Parr, Thomas et al. (2020). "Prefrontal computation as active inference". In: *Cerebral Cortex* 30.2, pp. 682–695.
Parthasarathy, Aishwarya et al. (2017). "Mixed selectivity morphs population codes in prefrontal cortex". In: *Nature neuroscience* 20.12, pp. 1770–1779.
Parthasarathy, Aishwarya et al. (2019). "Time-invariant working memory representations in the presence of code-morphing in the lateral prefrontal cortex". In: *Nature communications* 10.1, p. 4995.
Piazza, Manuela et al. (2007). "A magnitude code common to numerosities and number symbols in human intraparietal cortex". In: *Neuron* 53.2, pp. 293–305.
Quintana, Javier and Joaquin M Fuster (1999). "From perception to action: temporal integrative functions of prefrontal and parietal neurons". In: *Cerebral Cortex* 9.3, pp. 213–221.
Quiroga, R Quian et al. (2005). "Invariant visual representation by single neurons in the human brain". In: *Nature* 435.7045, pp. 1102–1107.
Quiroga, R Quian et al. (2008). "Sparse but not 'grandmother-cell'coding in the medial temporal lobe". In: *Trends in cognitive sciences* 12.3, pp. 87–91.
Rigotti, Mattia et al. (May 2013). "The importance of mixed selectivity in complex cognitive tasks". en. In: *Nature* 497.7451, pp. 585–590. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature12160.
Ringach, Dario L et al. (2016). "Spatial clustering of tuning in mouse primary visual cortex". In: *Nature communications* 7.1, p. 12270.
Roe, Anna W et al. (2012). "Toward a unified theory of visual area V4". In: *Neuron* 74.1, pp. 12–29.

Rohe, Karl and Muzhe Zeng (2020). "Vintage factor analysis with varimax performs statistical inference". In: *arXiv preprint arXiv:2004.05387*.

Roitman, Jamie D, Elizabeth M Brannon, and Michael L Platt (2007). "Monotonic coding of numerosity in macaque lateral intraparietal area". In: *PLoS biology* 5.8, e208.

Saxena, Shreya and John P Cunningham (Apr. 2019). "Towards the neural population doctrine". en. In: *Current Opinion in Neurobiology* 55, pp. 103–111. ISSN: 09594388. DOI: 10.1016/j.conb.2019.02.002.

Sharpee, Tatyana O (2019). "An argument for hyperbolic geometry in neural circuits". In: *Current Opinion in Neurobiology* 58.C.

Stringer, Carsen et al. (2019). "High-dimensional geometry of population responses in visual cortex". In: *Nature* 571.7765, pp. 361–365.

Tanji, Jun and Eiji Hoshi (2008). "Role of the lateral prefrontal cortex in executive behavioral control". In: *Physiological reviews* 88.1, pp. 37–57.

Turconi, Eva, Jamie ID Campbell, and Xavier Seron (2006). "Numerical order and quantity processing in number comparison". In: *Cognition* 98.3, pp. 273–285.

Wang, Jane X et al. (2018). "Prefrontal cortex as a meta-reinforcement learning system". In: *Nature neuroscience* 21.6, pp. 860–868.

Wang, Xiao-Jing (2020). "Macroscopic gradients of synaptic excitation and inhibition in the neocortex". In: *Nature Reviews Neuroscience* 21.3, pp. 169–178.

Whitlock, Jonathan R (2017). "Posterior parietal cortex". In: *Current biology* 27.14, R691–R695.

Whittington, James CR et al. (2022). "Disentangling with Biological Constraints: A Theory of Functional Cell Types". In: *arXiv preprint arXiv:2210.01768*.

Wiese, Heike (2003a). "Iconic and non-iconic stages in number development: The role of language". In: *Trends in cognitive sciences* 7.9, pp. 385–390.

— (2003b). *Numbers, language, and the human mind*. Cambridge University Press.

Wilson, Michael L, Marc D Hauser, and Richard W Wrangham (2001). "Does participation in intergroup conflict depend on numerical assessment, range location, or rank for wild chimpanzees?" In: *Animal Behaviour* 61.6, pp. 1203–1216.

Wimmer, Klaus et al. (2014). "Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory". In: *Nature neuroscience* 17.3, pp. 431–439.

Yang, Guangyu Robert et al. (2019). "Task representations in neural networks trained to perform many cognitive tasks". In: *Nature neuroscience* 22.2, pp. 297–306.

Zhang, Huanqiu et al. (2023). "Hippocampal spatial representations exhibit a hyperbolic geometry that expands with experience". In: *Nature Neuroscience* 26.1, pp. 131–139.

Zhou, Yuansheng, Brian H Smith, and Tatyana O Sharpee (2018). "Hyperbolic geometry of the olfactory space". In: *Science advances* 4.8, eaaq1458.

Zou, Hui, Trevor Hastie, and Robert Tibshirani (June 2006). "Sparse Principal Component Analysis". en. In: *Journal of Computational and Graphical Statistics* 15.2, pp. 265–286. ISSN: 1061-8600, 1537-2715. DOI: 10.1198/106186006X113430.

List of publications

Xiaoxiong Lin

April 2023

Note: publication 1 and 2 are part of the thesis.

1. **Lin, Xiao-Xiong**, Nieder, A. & Jacob, S. N. The neurocellular implementation of representational geometry in primate prefrontal cortex. *bioRxiv,* 2023–03 (2023).
2. Eisenkolb, V. M., Held, L. M., Utzschmid, A., **Lin, Xiao-Xiong**, Krieg, S. M., Meyer, B., Gempt, J. & Jacob, S. N. Human acute microelectrode array recordings with broad cortical access, single-unit resolution and parallel behavioral monitoring. *Cell Reports* (in press).
3. Zhang, Z., **Lin, Xiaoxiong** & Bao, Y. Holistic temporal order judgment of tones requires top-down disentanglement. *PsyCh Journal* (2022).
4. Bao, Y., Yang, T., Zhang, J., Zhang, J., **Lin, Xiaoxiong**, Paolini, M., Pöppel, E. & Silveira, S. The "third abstraction" of the Chinese artist LaoZhu: Neural and behavioral indicators of aesthetic appreciation. *PsyCh Journal* **6,** 110–119 (2017).
5. Bao, Y., Yang, T., **Lin, Xiaoxiong** & Pöppel, E. Donders revisited: D iscrete or continuous temporal processing underlying reaction time distributions? *PsyCh Journal* **5,** 177–179 (2016).
6. Wang, L., Bao, Y., Zhang, J., **Lin, Xiaoxiong**, Yang, L., Pöppel, E. & Zhou, B. Scanning the world in three seconds: M ismatch negativity as an indicator of temporal segmentation. *PsyCh journal* **5,** 170–176 (2016).
7. Wang, L., **Lin, Xiaoxiong**, Zhou, B., Pöppel, E. & Bao, Y. Rubberband effect in temporal control of mismatch negativity. *Frontiers in Psychology* **7,** 1299 (2016).
8. Bao, Y., Yang, T., **Lin, Xiaoxiong**, Fang, Y., Wang, Y., Pöppel, E. & Lei, Q. Aesthetic preferences for Eastern and Western traditional visual art: identity matters. *Frontiers in Psychology* **7,** 1596 (2016).

9. Bao, Y., Pöppel, E., Wang, L., **Lin, Xiaoxiong**, Yang, T., Avram, M., Blautzik, J., Paolini, M., Silveira, S., Vedder, A., *et al.* Synchronization as a biological, psychological and social mechanism to create common time: a theoretical frame and a single case study. *PsyCh Journal* **4,** 243–254 (2015).

10. Wang, L., **Lin, Xiaoxiong**, Zhou, B., Pöppel, E. & Bao, Y. Subjective present: a window of temporal integration indexed by mismatch negativity. *Cognitive processing* **16,** 131–135 (2015).