UNIVERSITY OF CALIFORNIA

Los Angeles

Prefrontal Cortical Dynamics in Social Behavior and Dyadic Interaction

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy in
Neuroscience

by

Lyle Kingsbury

2021

© Copyright by

Lyle Kingsbury

2021

ABSTRACT OF THE DISSERTATION

Prefrontal Cortical Dynamics in Social Behavior and Dyadic Interaction

by

Lyle Kingsbury
Doctor of Philosophy in Neuroscience
University of California, Los Angeles, 2021
Professor Weizhe Hong, Committee Chair

Human beings are fundamentally social animals – our daily interactions with others profoundly shape our experience, and in the best of circumstances, are among the most rewarding aspects of life. By the same token, negative social experience and prolonged social isolation can breed loneliness, anxiety, and depression. Our sociality is something that we share with most of the animal kingdom, and it is defined ultimately by our biology and our evolutionary history. Yet despite our deep and intuitive social impulses, our understanding of the biological processes in the brain that shape social perception, cognition, and behavior is still remarkably limited. One of the most important goals of modern neuroscience is to clarify the biological logic of our social experience, and in doing so, to develop a richer understanding of ourselves and our place in the world.

The two studies presented in this dissertation are a step toward this goal. They aim to investigate how the brain transforms social information into behavioral decisions and how it coordinates the dynamic of social interaction when individuals engage with one another. Before describing the studies, I will first provide a discussion in chapter 1 of concepts in social neuroscience, giving a brief history of how the field has evolved out of the separate strands of

ethology and cognitive neuroscience. Chapter 1 will also contain an overview of the basic neurophysiology of sensory perception and social behavior that will help to contextualize the questions addressed in the experiments and discussion of the results. In chapter 2, I will discuss methodological considerations in social neuroscience, including approaches to measure and manipulate brain activity as well as important statistical methods used to analyze relationships between social behavior and the underlying neural processes.

In chapters 3 and 4, I will present two studies which each investigate the involvement of cortical dynamics in social functioning from a distinct perspective. In chapter 3, I will describe experiments performed in mice using microendoscopic calcium imaging that explore how one important social sensory feature – sex identity (male vs. female) – is encoded in the cortex of the brain. Linking measurements of neural encoding to behavior, we found that internal representations of sex identity are sexually dimorphic across male and female animals, and that in males, cortical representations of sex predict their preference toward opposite-sex interaction. Using activity-dependent optogenetic manipulations, we found that these cortical representations of sex identity bi-directionally modulate the behavioral preference of the animal toward male or female-directed interaction. This study is among the first to demonstrate a causal role for functionally defined neuronal populations in behavior, and the first to do so in a way that links native encoding of a social variable to control of social behavior.

In chapter 4, I will describe experiments that investigate neural activity dynamics in the prefrontal cortex of mice while they engage in natural, dyadic interactions. This study focuses on analyzing the synchronization of neural activity across brains of interacting animals, which we identified as an emergent multi-animal neural correlate of social interaction. By analyzing the activity profiles of individual neurons in the brain, we traced the emergence of inter-brain synchronization from specific subsets of neurons that encode the behavior of each subject animal and its social partner. In aggregate, the responses of these neurons give rise to a regional activity signal that is synchronized across animals and predicts their future interaction and social

relationships. This study is among the very first to examine multi-animal neural dynamics in animals, and it sets a foundation for deeper mechanistic investigation of neural processes that coordinate interaction across individuals.

The dissertation of Lyle Kingsbury is approved.

Alcino Silva

Dean Buonomano

Sotirios Masmanidis

Weizhe Hong, Committee Chair

University of California, Los Angeles

2021

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

Chapter 1

Chapter 3

Chapter 4

ACKNOWLEDGEMENTS

I offer my deepest gratitude to my primary advisor Weizhe Hong, who imprinted in me by his own example the absolute highest measures of integrity and intellectual rigor. True progress depends on these virtues – my thanks will be a lifelong effort toward their ideal expression. With seemingly endless patience, and always with kindness, he challenged me to push myself beyond what I thought possible. Whatever the future may bring for me, within or outside of the world of science, I will hold the experiences and lessons of these most formative years close to my heart.

Several members of the Hong lab made the last five years of my life incredibly enjoyable and stimulating. I would like to offer my heartfelt thanks to Shan Huang, who I worked closely with for years on the work described in this dissertation and who was an important role model for me as I learned to think deeply about designing experiments and learned to manage the stresses and challenges of the research process. I would like to thank Tara Raam and Patrick Chen as teachers of life, and for their spirit of debate which expanded my mind on issues ranging from science and philosophy to politics and religion. And I would also like to thank Jun Wang and Letizia Ye for their infectious enthusiasm and their good humor as I slowly learned to navigate the dual roles of student and mentor.

My parents share in this achievement much more than they probably know. My graduate training challenged me deeply on many levels, and I would not have flourished, or perhaps even finished, if it were not for the limitless love and support of my parents, Robin and Peter Kingsbury. They have always encouraged me to follow my passions, whatever they have been.

Regarding the work described in this dissertation, I would like to acknowledge that Chapter 3 consists of a manuscript published in *Neuron* in 2020 entitled ‘Cortical Representations of Conspecific Sex Shape Social Behavior’. The work was authored by me and Shan Huang, and was co-authored by Tara Raam, Letizia S. Ye, Don Wei, Rongfeng K. Hu, Li Ye, and Weizhe Hong, all of whom have consented to the inclusion of the work in this dissertation. Chapter 4 consists of a manuscript published in *Cell* in 2019 entitled ‘Correlated Neural Activity and

Encoding of Behavior Across Brains of Socially Interacting Animals’. This work was authored by me and Shan Huang, and co-authored by Jun Wang, Ken Gu, Peyman Golshani, Ye Emily Wu, and Weizhe Hong, and all have consented to the inclusion of the work in this dissertation.

Finally, I would like to acknowledge the sources of funding that made this work possible. The Interdepartmental Neuroscience Program at UCLA provided funding for research and related travel. The research described in this dissertation was supported by grants awarded to Weizhe Hong, including an NIH grant (R01-NS113124), a NARSAD Young Investigator grant, a Sloan Research Fellowship, a Searle Scholars Award, a Brain Research Foundation Grant, and a Packard Foundation Fellowship. The research was also supported by grants and fellowships awarded to me, including an institutional training grant (T32-NS058280), a Ruth-Kirschstein National Research Service Award (F31- MH117966), an ARCS foundation award, and a travel award from the Brain Research Institute at UCLA. Without the support of both institutional and federal funds, the scientific studies described in this dissertation and the technical, pedagogical, and professional training that enabled them would not have been possible. I am deeply grateful.

VITA

EDUCATION AND WORK EXPERIENCE

Ph.D. Candidate in Neuroscience 2015 – Present
University of California, Los
Angeles (UCLA) | GPA 3.6
Thesis advisor: Dr. Weizhe Hong

B.A. in Biological Sciences 2013 – 2015
(Bioinformatics concentration)
Hunter College (CUNY) | GPA 3.8

Eugene Lang College – The New School 2011 – 2012
(Interdisciplinary Science)

HONORS AND AWARDS

Harold M. Weintraub Graduate Student Award 2021
Annual award recognizing outstanding achievement during graduate study in the biological sciences

UCLA Samuel Eiduson Student Lecture Award 2020
Annual award given to a UCLA neuroscience PhD student for outstanding thesis work

F31 Ruth L. Kirschstein National Research Service Award 2018 – 2021

T32 NINDS/NIH Training Program in Neural Microcircuits 2017 – 2018

Achievement Awards for College Scientists (ARCS) Fellowship 2016 – 2019

Hunter College Else Seringhaus Award in Biological Sciences 2015
Given for excellence in biology research and coursework

Most Outstanding Research Award 2014

Thomas Hunter Honors Program 2013 – 2015
Honors program curriculum at Hunter College

PUBLICATIONS

- **Kingsbury, L.***, Huang, S.*, Wang, J., Gu, K., Golshani, P., Wu, Y. E., & Hong, W. (2019). Correlated Neural Activity and Encoding of Behavior across Brains of Socially Interacting Animals. ***Cell***, *178*(2), 429-446.e16. (*Equal Contribution) 2019

- **Kingsbury, L.***, Huang, S.*, Raam, T., Ye, L. S., Wei, D., Hu, R., Ye, L. & Hong, W. (2020). Cortical Representations of Conspecific Sex Shape Social Behavior. ***Neuron****.* (*Equal Contribution) 2020

- **Kingsbury, L.** & Hong, W. A Multi-Brain Framework for Social Interaction. (2020). ***Trends in Neurosciences****.* 2020

- Urbanski, M. M., **Kingsbury, L.**, Moussouros, D., Kassim, I., Mehjabeen, S., Paknejad, N., & Melendez-Vasquez, C. V. (2016). Myelinating glia differentiation is regulated by extracellular matrix elasticity. ***Scientific Reports***, *6*(1), 33751. 2016

- Lin, A., Vajdi, A., Kushan-Wells, L., Helleman, G., Hansen, L. P., Jonas, R. K., Jalbrzikowski M., **Kingsbury, L.**, Raznahan, A., Bearden, C. E. (2020). Reciprocal copy number variations at 22q11.2 produce distinct and convergent neurobehavioral impairments relevant for Schizophrenia and Autism Spectrum Disorder. ***Biological Psychiatry***. 2020

CONFERENCE PRESENTATIONS

- *Poster Presentation:*
  Computational and Systems Neuroscience (Cosyne) 2020
- *Poster Presentation:*
  Society for Neuroscience (48th) 2019
- *Poster Presentation:*
  Conference on Collective Computation in Biological Systems (Janelia HHMI) 2018
- *Poster Presentation:*
  Computational Neuroscience Workshop, RIKEN BSI 2017
- *Poster Presentation:*
  Neural Microcircuits Symposium, UCLA 2017

CHAPTER ONE

Concepts in Social Neuroscience

1.1: Why do social neuroscience?

Human beings are social animals. We are deeply embedded in a world that we have co-created, and we are continuously shaped by our culture and our interactions with others. In the end, our life experience depends to a large degree on the strength of our bonds, on how we cultivate and navigate them, and on how they influence our choices over a lifetime [1]. When we have intimate social connections, we feel understood, supported, resilient, and engaged. When we are neglected or isolated from others, we become anxious, unmoored, and listless [1]. These facts of human existence are subjective and instinctual, but they are also empirical, rooted ultimately in the biology of the brain. Lack of social connection is a common feeling in people who lose their lives in “deaths of despair,” and is both a cause and consequence of psychiatric illnesses like depression [2–4]. If we want to secure human flourishing and dampen these sources of suffering, then we must gain an understanding of our social nature, of how our social embeddedness affects our experiences and our choices, that is grounded in the vocabulary of empirical science.

Our social nature shapes not only our personal experience, but also how we build societies, organize political structures, and coordinate collective action. Indeed, to whatever degree it is true that man is essentially a “political animal,” this is because of our basic social impulse. Many of the problems that we face collectively, including existential threats such as global climate change and nuclear war, turn on our ability to make collective decisions and coordinate individual choices. If we can penetrate deeper into some understanding of how our decisions are conditioned by our social context, we will be able to better anticipate our limits, our potentialities, and the unique challenges of collective action. In the long term, this may be necessary to secure the survival of the human species.

Despite its deep importance, our social nature is not unique to us. In fact, all animals that sexually reproduce – that is to say, almost all species on planet earth – engage in social interaction in one form or another. We are not alone in this regard, and some of the biological processes that imbue us with feelings of empathy and loneliness are shared by species throughout the animal kingdom, extending beyond our primate and mammalian cousins to birds, reptiles, and even insects [5–9]. The lens of biological science gives us an opportunity to explore these common social experiences and dissect the physical logic that underlies them [7]. Research into how the brain orchestrates social behaviors and coordinates interaction can reveal conserved principles of neural function that shed light on our own experience and our own behavior [10]. It is with this spirit that social neuroscientists focus their investigations of the biology of sociality across diverse species, from flies to humans, aiming to bring clarity to our most ancient instincts [5,10–13].

The studies presented in this dissertation are aimed at understanding the processes in the brain that integrate social information, shape behavioral choices, and coordinate social interaction. I have explored these in the brain of the mouse, a social creature that, despite many differences, has a brain that is very much like our own. The focus of these investigations in the mouse has allowed for precise measurements, incisive experiments, and mechanistic insights that could not be obtained from human subjects with modern technology [14,15]. It is my belief that the knowledge gained from these studies moves us one step closer to a full description of the social brain, and that from this, we can take away a deeper understanding of ourselves.

1.2: A short history of social neuroscience

Social neuroscience, in the broadest sense, is the study of how biological processes in the brain enable animals to engage with other members of their species, interact, and coordinate their behavior toward common goals [10,16,17]. In its present conception, social neuroscience takes a multi-level view of the biology that underlies social behavior and cognition, drawing from genetic, molecular, network and behavioral modes of analysis. Historically, the field has its roots in ethology [7], which is interested in biological and evolutionary descriptions of animal behavior. Although not specifically focused on social behavior, ideas developed in the field of ethology have played an important role in shaping how social neuroscientists frame questions about animal behavior and how they address those questions methodologically.

One of the most important ethologists (and considered one of the founders of the field), Niko Tinbergen, was particularly influential in the development of social neuroscience and its modern conception. In 1963, Tinbergen outlined "four questions" which he suggested as an organizational scheme for the scientific investigation of animal behavior [18,19]. Paraphrased below, Tinbergen's four questions are:

1) *What is a behavior's adaptive function?* This question asks about the role that a particular behavior plays in how an animal interfaces with its present environment. It may be, but is not necessarily, informed by an evolutionary perspective. A behavior's adaptive function is not static but may change over time.

2) *What is a behavior's evolutionary history?* This question asks how a behavior came about through the process of evolution by natural selection with environmental pressure. What type of pressures and responses shaped its development over generations?

3) *How is a behavior controlled mechanistically?* This question asks about the biological (neural) processes that control how a behavior is regulated and expressed in the here and now, regardless of its history, its function, or how it may change in the future.

4) *How did a behavior develop biologically?* This question asks how a behavior was shaped over the course of an organism's lifetime. To what degree is it learned from experience, and to what degree is it innate? How may it continue to change in response to environmental influences?

Tinbergen proposed these "four questions" as a framework to guide experimental inquiry. While pragmatic considerations limit our ability to pose more than one of these questions at a time, they were intended to highlight complementary perspectives and to produce more comprehensive and integrated descriptions of behavior [18,19]. The imprint of this conceptual structure can be clearly recognized in the directions of modern social neuroscience and its emphasis on neural mechanisms (particularly questions three and four) [10]. While all four questions have not received equal attention over the years since Tinbergen outlined his questions, the interdisciplinary nature of the modern discipline aligns well with the spirit of his suggestion.

In addition to setting up a conceptual framework that has held up over decades, Tinbergen also helped to elucidate some important themes in the organization of animal behavior that are still influential. For example, he theorized about a hierarchical organization of innate and instinctual behaviors that proceed through a tree-like decision process with mutual inhibition between behavioral drives at the same level [7]. This also led to the related idea that functionally related behaviors may be "grouped together" in the brain and implemented by overlapping neural circuits. While the brain has proven to be far more complex, modern analysis using genetic tools has found some support, as well as mechanistic clarity, for this type of hierarchical organization [10,20,21].

As neuroscience evolved, the adoption of tools to measure and manipulate electrical activity in the brains of behaving animals opened a new level of depth to the analysis of the biology underlying social behavior. Using stimulation electrodes, researchers were able to link specific regions of the brain to behavioral drives and expression of stereotyped behavioral patterns. One example of this was the identification of the hypothalamic attack area, a subregion of the hypothalamus that, when stimulated (in cats and in rats), produced aggressive behavior such as hissing and biting [22,23]. While this level of spatial resolution is coarse by modern standards, such experiments helped to generate new hypotheses about how specific social behaviors may be controlled by dedicated brain structures. Over the last few decades, this "brain mapping" approach has been immensely fruitful and has led to greater understanding of the mechanisms controlling diverse behaviors such as aggression [24,25], mating [26,27], and parenting [20,28,29]. In many ways, this framework is still driving modern research programs. Modern tools such as optogenetics have allowed researchers to probe brain function more incisively and to define functional structures with far greater spatial resolution [30,31]. For example, more recent studies have refined the "hypothalamic attack area" by clarifying the specific role of Esr1+ (estrogen-receptor expressing) neurons in the ventrolateral portion of the ventromedial hypothalamus (VMHvl) in regulating aggressive behavior [25,27,32]. Still, despite the success of this conceptual framework, the question of whether, and to what degree, the neural mechanisms that control social behaviors are localized in the brain is largely unclear [10]. In many cases, several anatomically distinct circuits have been shown to play some role in a specific behavior. These may form a connected pathway or network of nodes that coordinate to control some behavioral process. Alternatively, the biological control of some behaviors may be distributed across anatomically distinct structures. One aim of modern social neuroscience is to bring clarity to this issue [10,12].

The theoretical idea that control of social behaviors may be anatomically localized also coincided with a related idea that some brain structures may be specialized for general social functions [13,33,34]. This “social brain” hypothesis was initially formulated in the context of human cognitive neuroscience based on observations that tasks engaging social processes (such as mentalization) were linked to activation of a consistent set of brain regions, including the amygdala, parts of temporal association cortex, and parts of the prefrontal cortex [16,34,35]. The difficulty with this idea was noted early on. Even if one can identify patterns of neural activity that correlate with social stimuli or engagement in social tasks, it is difficult to experimentally parse whether there is anything unique about social processing, or whether those tasks simply engage more primitive cognitive functions like attention and motivation in a particular way to meet the demands of the social setting [35]. Still, the idea has been influential, and even though it originally emerged out of work on human cognition, the vocabulary of the “social brain” is now infused in discussions of social behavior across diverse species [10,17,33]. This intersection between ethology and cognitive neuroscience has produced the modern conception of social neuroscience as an interdisciplinary field which aims to coordinate research across species and settings to develop a mechanistic description of the brain processes that underlie social functions [10,16,19].

1.3: The biology of social interaction and social decision making

1.3.1: Social interaction as a dynamic feedback loop

What makes a behavior “social”? And what, if anything, is unique about social behavior? In the broadest sense, social behaviors are those behaviors that structure interaction between conspecifics and coordinate their actions toward common goals [10]. In many animals, these goals are evolutionary in nature, typically directed toward reproduction and species survival, and are deeply instinctual. In humans, social behavior can be directed toward much more abstract goals and can require coordination across large numbers of people across long timescales [36]. In all cases though, social behavior is unique in its specific directedness toward other members of the same species, and because of this, often depends on communication using species-specific cues (such as language in humans). In addition to this, because social goals depend on coordination between individuals, social behaviors are also unique in their embeddedness in a context of mutual interaction [37–39]. This context of interaction can shape the decision process for an individual in dramatic ways, making social behaviors highly complex in comparison to non-social behavior.

To illustrate this point, consider a decision made by a monkey to eat a piece of fruit on the forest floor. In this non-social setting, the decision to eat or not eat the fruit can be thought of as a computation of optimal action given the animal's sensory inputs from the environment and internal state (a *sensorimotor transformation* – **Figure 1.1A**). The monkey may decide to eat the fruit simply because she is hungry, food is present, and there are no competing behavioral drives. Since the environment is relatively stable in time, the monkey does not worry about whether something surprising will happen when she takes a bite. The environment, devoid of other agents, is relatively predictable, and so the choice is conditioned on the animal's state and little else.

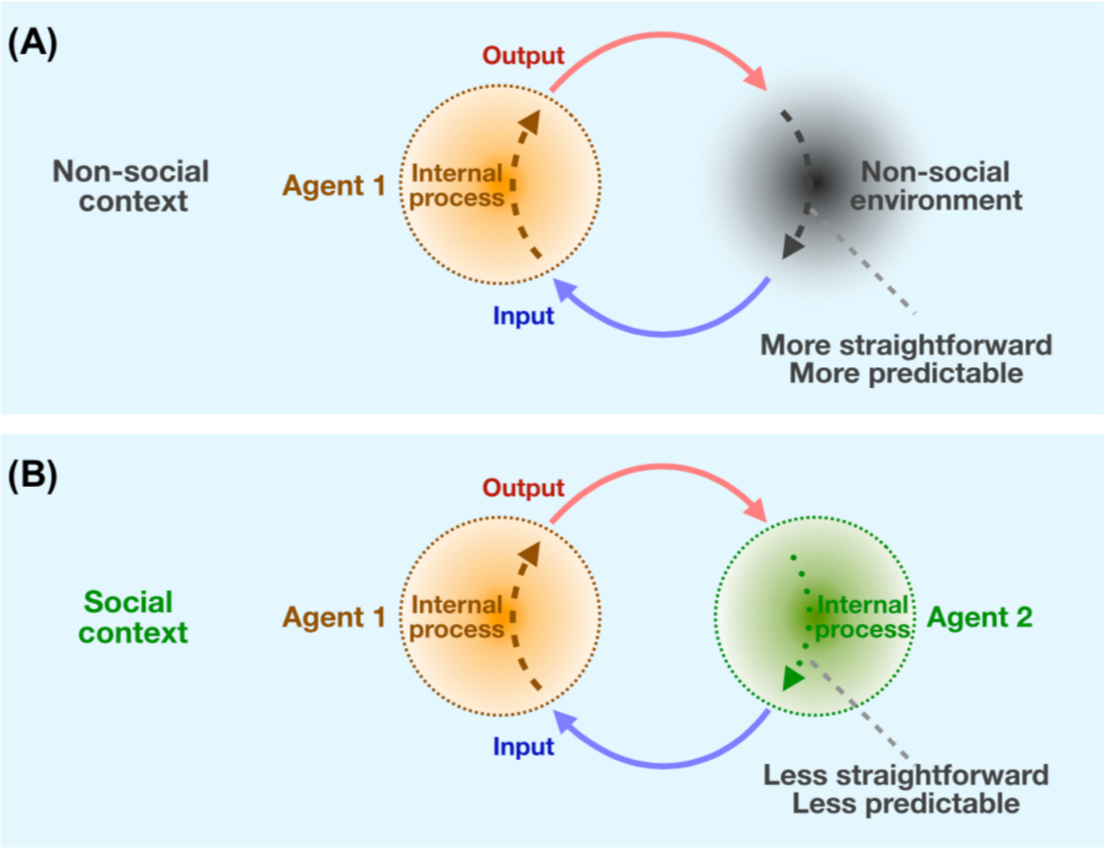

Figure 1.1: Social interaction as a dynamic feedback loop. Illustration of the decision-making process of an agent in a non-social (A) versus social (B) context. For simplicity, the social context is illustrated for a two-agent situation. The agent makes choices based on a sensorimotor transformation that maps an input space to specific behavioral outputs. In the non-social context, the agent uses feedback from the environment which is relatively stable and predictable. By contrast, the social context couples the agent directly to another agent whose internal states (e.g., goals, beliefs, etc.) are hidden. Behavior and behavioral responses from the interacting partner are highly unpredictable, creating a more complex decision-making processes that engages mentalization, dynamic prediction, and other cognitive processes.

By contrast, social interaction engages animals and their decision processes in a fundamentally different way (**Figure 1.1B)**. In a social setting, the environment is not stable because it contains another agent (or several agents) making choices that are not easily predictable [36,40]. During dyadic interaction, two agents become coupled to one another such that each one's actions become part of the immediate sensory space of the other. Because of this mutual coupling of sensation and action, each agent's decisions are contingent upon the immediate context of another agent's actions and potential for future action [17]. This extra degree of complexity demands a wider scope of consideration for behavioral choices and their outcomes: *If I engage in this behavior, then how will my social partner react? What is my social partner going to do next? What are they trying to achieve in this interaction? Are we acting collaboratively or competitively toward our goals?*

While animals like monkeys and mice may not explicitly ask these questions, humans certainly do [36,40]. And even still, this type of mentalization process illustrates the important feature of social interaction that *is* consistent across species – that social decisions are made contingently and collaboratively, and so the interactive process is co-created and coordinated through dynamic feedback between agents [17,39]. What does this mean for the neuroscientific study of social interaction? At least three important implications bear consideration.

First, because of the interactive setting, we should consider that social decisions are typically made in a very rich, multimodal sensory context. While some species may depend more heavily on specific sensory channels for social communication, in almost all species, social decisions depend on more than one modality [10]. Mice, for example, use olfactory channels to communicate information about sex, age, and hormonal states, but auditory and tactile cues are also critical for interaction [10]. Primate interactions depend on gestures and visual cues, but primates also make heavy use of auditory channels to communicate through vocalizations and language (in the case of humans), and express both affiliative and aggressive behaviors through physical touch. Thus, when we study social interaction, it is important to consider – and if possible, to measure – the multitude of channels through which animals may communicate (more on this in chapter 2). And when investigating the neural mechanisms of sensory processing, we should consider that the neural encoding of social features (such as social rank, for example) may depend on integration across several modalities [41,42].

Second, we should consider that the interactive context – the presence of another agent who responds dynamically to one's choices – is itself a force that shapes how social decisions are made [17,38,39]. Because of this, it is important to consider how different experimental settings that preserve more or less of this natural behavioral dynamic may engage the brain in distinct ways. For example, the neural processing of a non-social stimulus, such as a piece of food, may be fundamentally different in the presence of another mouse who represents competition. Of course, human choices are also shaped dramatically by knowledge of social context (think: how do we behave when we are alone vs. in social setting?), and in many cases, the social context itself engages specific brain processes. The effect of interactive context is an important consideration for experimental design. More constraint on a subject and isolation from social partners may afford more control to a researcher and reduce potential confounds. It may also obscure processes that would be engaged during real interaction, which places limitations on the generalizability of observations to more ecological settings [17,38,39].

Finally, we should consider that just as the social behaviors we observe do not occur in isolation, the neural processes that underlie them do not either. With a more reductive framing, animals engaged in interaction can be thought of as brains influencing one another through a limited communication channel of behavior (**Figure 1.2A**), and many aspects of this brain-to-brain interaction may not be readily observable to an outside viewer (**Figure 1.2B**) [39,43]. Indeed, many processes that play an important role in shaping social interactions, such as internal states like motivation and attention, are not observable through simple behavioral measurement [8,44]. Direct examination of brain activity and the relationship between neural signals across individuals may therefore provide insight into how brain processes are coordinated across individuals and how this dynamic relates to the evolution of an interaction [17,39,43,45–47].

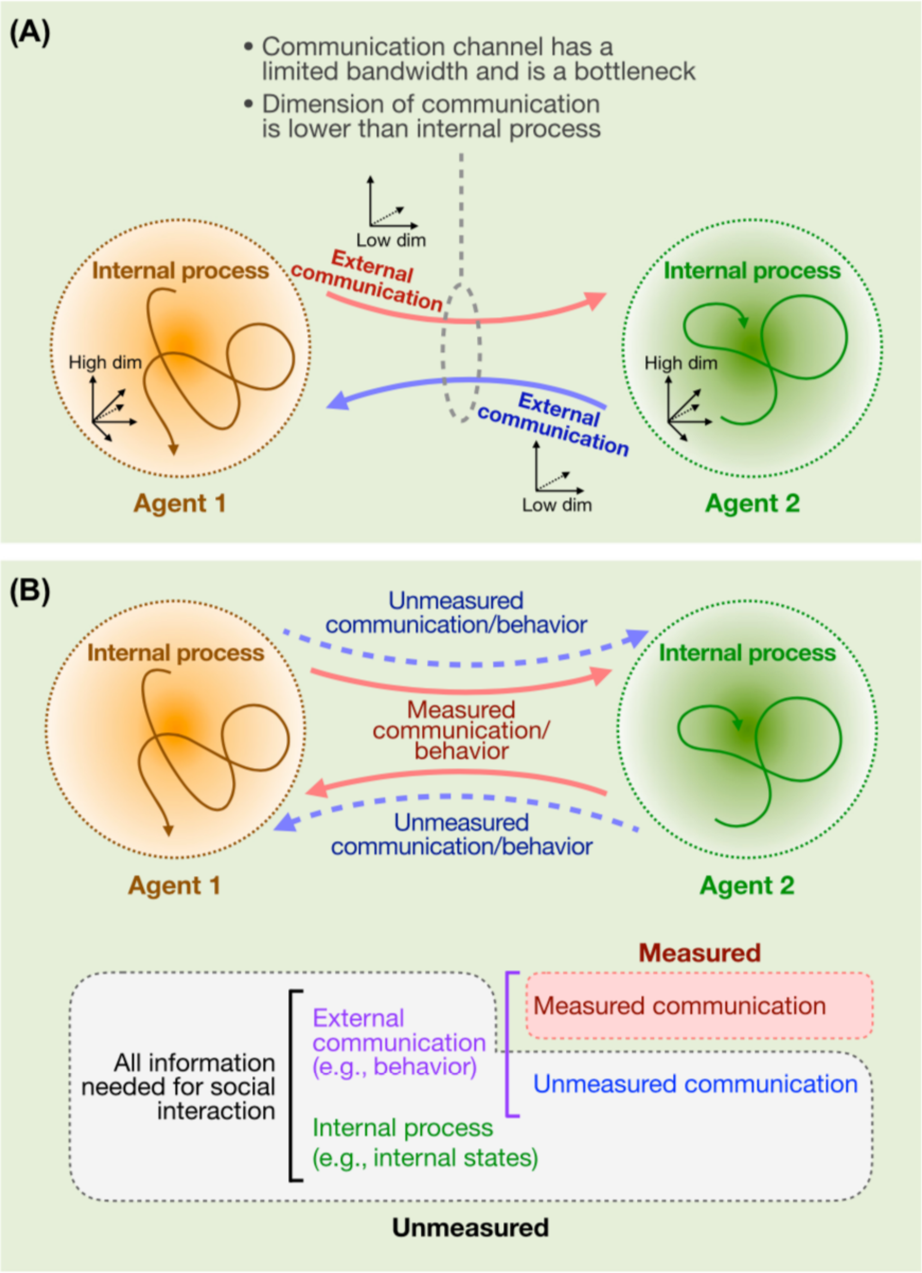
(A)
• Communication channel has a limited bandwidth and is a bottleneck
• Dimension of communication is lower than internal process
Low dim
External communication
Internal process
High dim
Agent 1
Internal process
High dim
Agent 2
External communication
Low dim
(B)
Unmeasured communication/behavior
Internal process
Internal process
Measured communication/ behavior
Unmeasured communication/behavior
Agent 1
Agent 2
Measured
Measured communication
All information needed for social interaction
External communication (e.g., behavior)
Unmeasured communication
Internal process (e.g., internal states)
Unmeasured

Figure 1.2: Neural systems communicate through a behavioral bottleneck. Illustration of social communication between two interacting agents. (A) As the behavioral outputs of each agent form part of the other's input space, interacting agents become coupled in an integrated system (Figure 1.1). The full range of neural processes that shape behavioral decisions span a higher-dimensional space than that of expressed behavior, and communication between agents is therefore restricted by a channel with limited bandwidth (the communication bottleneck). (B) From a third-person perspective, observation of only a limited part of external communication (explicit behavior measured by experimenters) provides an impoverished view that lacks information about the underlying neural processes. Direct measure of the neural processes and their dynamical relationships across agents may provide additional information about unmeasured variables and the interaction itself. Abbreviations: High dim, high-dimensional; Low dim, low-dimensional.

1.3.2: Social decisions in the brain

In order to understand social interaction at a mechanistic level, we need to understand the processes in the brain that lead to social behavioral decisions. This is clearly a complex problem [10,36,48]. Yet with all of the considerations from the previous section in mind, it can also be useful (and often is necessary) to simplify in order to operationalize questions about how the brain works. In this spirit, I will offer a rough framework for thinking about the process of social decision-making in the brain, and I will describe in broad terms what is currently known about the underlying neurophysiology. This context will help to frame the studies described in chapters 3 and 4 and will suggest some sense of the territory still unexplored. In very broad strokes, a social decision (and in fact any decision) can be thought of as a process that progresses in three stages – *sensory transduction, internal integration, and behavioral expression* – which transform sensory inputs from the environment into behavioral changes. In reality, the boundaries between these "stages" are imprecise and probably highly permeable, but the attempt to separate them is useful and can bring some clarity to our thinking about the overall process.

1) *Sensory transduction:* The brain encounters, transduces, and processes sensory inputs from the external environment. This sensory stage depends on sense organs, such as the eyes, ears, and nose, to transmit spatiotemporal patterns of light, air density, and chemical compounds to the central nervous system by conversion into patterns of neural activity [49]. Neural circuits in sensory regions of the brain receive and operate on these signals to extract and transform internal representations of sensory features [50]. For example, a mouse may encounter olfactory and pheromonal cues by sniffing another mouse, and these cues may be integrated to extract

complex features like sex identity [51] or social status [42]. Humans and other primates integrate visual and auditory information to identify others based on their face and voice.

2) *Internal integration:* Neural representations that contain social information are then integrated with other processes in the brain. Some brain processes, called “internal states,” are not expressed explicitly through behavior but may nonetheless shape behavior and physiology in important ways [8,44,52] Common examples of internal states include attention, motivation, and anxiety – in humans these are typically associated with some subjective experience or emotion. In animals, they cannot be measured explicitly but only inferred through their effects on behavior and physiology. Internal states like anxiety can interact with processed sensory inputs to influence behavior. For example, an encounter with an unfamiliar conspecific may cause an animal to run away if it is in a highly anxious state. However, in a state of low anxiety, an encounter with the same conspecific may instead lead to approach behavior, exploration, and further interaction. Thus, internal integration lends flexibility and contingency to behavioral decisions based on the animal’s immediate social context and state [10,44].

3) *Behavioral expression:* Integration of sensory inputs with internal state processes culminates in the expression of a particular action or a change in behavior. Some sensory cues can trigger innate, reflexive responses, whereas others can shape behavior on a longer timescale through changes in internal state [10,44]. For example, if a male mouse is in a highly aggressive state, then introduction of an intruder may trigger an immediate attack response, leading to more fighting behavior. By contrast, repeated defeat in conflicts between mice may lead to a sustained

"depressive" state in the defeated mouse that changes its motivation, social appetite, and sleep patterns [53]. Although these two outcomes are expressed in distinct ways and over different timescales, both the acute attack response and sustained defeat state represent behavioral changes that result from social experience.

1.3.3: Sensory transduction

The first stage in the process of social decision-making is transduction of sensory inputs from the environment and extraction of relevant social features and context. As discussed previously, social interaction often involves communication along several sensory modalities, and to varying degrees (in different species), they are all important for the normal social decision process. In this section, I will describe the basic neurophysiology for the visual, olfactory, and auditory pathways, with particular attention on the mouse brain. For context, I will also describe some of the similarities and differences between the rodent and primate physiology.

1.3.3.1: Vision

In rodents and primates, visual information from the environment is transduced by neurons in the *retina* of the eye. Retinal ganglion cells form the main output carrying visual information to the rest of the brain, synapsing primarily in the *lateral geniculate nucleus* (LGN) of the thalamus, but also sending small projections to the hypothalamus and superior colliculus [49]. LGN neurons carrying visual information from the retina send projections to the cortex, primarily synapsing on neurons in the *primary visual cortex* (V1). Like neurons in the retina, neurons in V1 show response patterns that are largely restricted to specific areas of the visual field (the *receptive field*), and some show stronger responses to oriented edges of luminance contrast (orientation tuning). While these response patterns are topographically organized in the primate brain (representationally tiling the visual space across the cortex), the mouse brain does not show clear topographic organization, although it does show some spatial clustering of orientation tuning [54]. This marks an early stage of feature extraction and transformation along the visual pathway, whereby localized luminance signals in the retina are combined to produce responses in downstream neurons that encode more complex visual features [50].

In the primate brain, this organization continues along a roughly defined hierarchy of visual areas that receive inputs from V1. For example, in area V2 (the secondary visual cortex), neurons that respond to specific object borders and colors emerge [55], and in V4, neurons respond to different types of curvature, depth, and texture, with typically larger receptive fields [50,56]. Visual information is a very important modality for social communication in primates, including in humans. For example, there is considerable evidence for specialization of certain cortical regions in the primate brain for processing information about faces of other individuals. Neurons in the monkey *inferotemporal cortex* (IT) of the temporal lobe show selective responses to facial features [57], and this “face patch” area corresponds anatomically with a region of the human brain in the *fusiform gyrus* that responds specifically to faces [58,59].

In rodents, less is known about the importance of visual processing for social interaction. Structurally, the boundaries between different visual cortical areas are less well defined, and neural responses to complex visual features are spatially graded across the entire posterior cortex [60]. The rodent visual system does not appear to specifically encode social stimuli, as is observed in the primate temporal cortex [57]. Still, information about movement is likely to contribute to social decision-making. For example, stimulation of hypothalamic neurons in mice triggers attack behavior, even to a mirror or a moving glove, suggesting that coarse visual features may be integrated with other modalities to control some social behavior decisions [10,25].

1.3.3.2: Olfaction

Olfaction plays an important role in social communication, and especially so in rodents. In the mouse, odor and pheromonal cues coordinate innate behaviors such as aggression and mating. The *main olfactory bulb* (MOE) and *vomeronasal organ* (VNO) of the accessory olfactory system mediate odor and pheromonal sensation. This early step of sensory transduction is necessary for discrimination of conspecific sex and other social sensory features. Neurons in the VNO respond selectively to odor cues derived from specific social stimuli, including male and female pheromones, pup odors, and predator cues [61,62]. Because of their role in transducing social sensory cues, MOE and VNO neurons are also necessary for sex-typical social behavior in both male and female mice. For example, legions of the MOE disrupt social behavior [63], and knockdown of TRP2 receptors (which mediate transduction in VNO neurons) results in increased male aggression toward females, indiscriminate mounting behavior toward both sexes, and reduced parenting behavior in females [64]. While MOE neurons project to the piriform cortex, olfactory tubercle, and cortical amygdala, VNO neurons send projections through the accessory olfactory bulb to two main downstream structures: the *medial amygdala* (MeA), and the *bed nucleus of the stria terminalis* (BNST) [10]. Neurons in both the MeA and BNST show selective responses to male vs. female conspecifics, indicating that populations of neurons in these regions can represent sex identity [65–67]. Discriminability of sex identity at the neural level in MeA and

BNST, as well as in the *ventromedial hypothalamus* (VMH), correlates with increased social behavior, including social investigation and aggressive behavior [65,68]. These observations suggest that internal representations of sex identity may shape behavioral decisions.

In primates, including in humans, olfactory information is routed from the olfactory epithelium in the nose via the olfactory tract to the piriform cortex, the hypothalamus, and the amygdala [49]. Like in the mouse, amygdala neurons in the primate brain have been shown to encode complex social information, including social attention and the rank status of other individuals [69,70], although the role of amygdala and hypothalamic structures in encoding sex-specific social information is not as clear.

1.3.3.3: Audition

In addition to visual and olfactory cues, auditory signaling is also an important communication channel for social interaction in many species. Humans are unique in the animal kingdom in their use of language, which allows for the rapid vocal communication of detailed plans, complex thoughts, and emotional states. However, many other species utilize vocalizations as signals to communicate and coordinate behavior. Rodents, for example, communicate using ultrasonic vocalizations to send distress calls, initiate play, and coordinate mating behavior [71]. Sensation and perception of complex auditory signals is therefore crucial for social decision making.

In rodents and primates, auditory information is extracted in the *cochlea* of the ear and transduced into neural signals by hair cells which respond to external air vibrations as they propagate through endolymph fluid [49]. Neurons from the cochlear nuclei carrying auditory information project to the inferior colliculus of the midbrain, which in turn sends projections to the *medial geniculate nucleus* (MGN) of the thalamus. MGN neurons send projections to the cortex, primarily targeting neurons in the *primary auditory cortex* (A1) [49]. Much like in the visual system [50], the auditory system is organized in a roughly hierarchical architecture which extracts acoustic features of increasing complexity. For example, in both mice and primates, including humans, neurons in A1 show tuning to specific acoustic features [49,72]. Topographic organization of frequency representations (tonotopy) emerges from relatively early stages of processing, present already in the colliculus and extending to MGN and A1 [73]. Neurons in the auditory cortex also extract more complex acoustic features, including time-varying patterns of tones and specific "syllables" used for vocal communication. In the rodent, for example, neurons in A1 are tuned to specific vocalizations that are relevant for coordinating social behavior [71,74]. In addition, A1 neurons in females enhance responsiveness to distress calls from pups, and these plastic changes are important for pup retrieval behavior [75,76].

In monkeys, neurons in the auditory cortex and the superior temporal sulcus (STS) respond to species-specific vocalizations, and in humans, parts of A1 and A2 show specific responses to human voice and speech. In both monkeys and in humans, neurons in early auditory cortex (the auditory belt) and STS send projections to parts of the ventrolateral *prefrontal cortex* (PFC), which is thought to play an important role in processing communication-relevant sounds. In particular, the human inferior frontal gyrus (which includes the famous Broca's area) has been linked to speech and language processes. While relatively little is known about the role of frontal cortex in mouse auditory perception, parts of the mouse PFC have been linked to the production of ultrasonic vocalizations, suggesting a role for this region in integration of auditory cues and generation of vocalizations.

1.3.4: Internal Integration

Following the sensory transduction pathways outlined above, we have now traced the detection of sensory signals from the environment through initial processing stages along the visual, olfactory, and auditory pathways. All of these sensory modalities (in addition to touch) are used for social communication in mice and in primates. Indeed, many important social features (such as sex identity social status) are integrated across multiple distinct modalities [10,77,78]. In order to affect behavior, these internal representations of social sensory features must also be integrated with internal state variables to be contextualized. In this section, I will discuss this process of internal integration and give an overview of some of the circuit processes that are involved. I will focus the discussion on three behavioral processes that are most relevant for the work presented in chapters 3 and 4. These are *aggression, social status and dominance behavior*, and *mating and sexual behavior*.

1.3.4.1: Aggression

Aggressive behavior between conspecifics is ubiquitous across the animal kingdom from flies and rodents to humans [21]. In mice, aggression between males serves to settle territorial disputes, to mediate conflict over resources and mating opportunities, and to establish social status and dominance hierarchies [79,80]. In females, defensive aggression is observed during pup rearing toward intruders or other threats. Humans show many elaborated forms of aggressive behavior, ranging from non-violent, verbal microaggression to homicide and war to settle disputes and establish or enforce status/dominance relationships.

Several different neural circuits in the brain have been found to play roles in regulating aggressive behavior. One of the earliest demonstrations of a circuit for aggression was the "hypothalamic attack area," a region of the hypothalamus that, when stimulated electrically in rats and cats, was observed to trigger aggressive behavior such as hissing and biting [22,23]. Using molecular and genetic tools in mice, studies in recent years have clarified the underlying physiology. Studies using photostimulation of specific subpopulations of neurons in the ventrolateral portion of the *ventromedial hypothalamus* (VMHvl – a region contained within the classical hypothalamic attack area) have clarified its specific role in the control of aggression [25,32]. VMHvl neurons are active during attack behavior in both male and female mice, and their activity predicts the latency to and duration of future aggressive bouts. Interestingly, some of these neurons are also active during anticipation of aggressive interaction [81], suggesting that they may encode an internal representation of aggressive state.

In addition to the VMH, other regions in the hypothalamus and amygdala have also been implicated in the control of aggressive behavior. The posterodorsal portion of the *medial amygdala* (MeApd), which receives direct pheromonal information, projects both directly and indirectly (via the BNST) to the VMHvl. As in the VMHvl, specific subpopulations of neurons in the MeApd have been shown to control aggression; photostimulation of MeApd inhibitory neurons causes acute attack behavior, and photoinhibition of the same neurons can cause immediate cessation of ongoing attack behavior [24]. Interestingly, both the MeApd and the VMHvl also contain neurons that respond selectively to cues from male and female conspecifics (encoding sex identity) [65,68]. This suggests that that the MeApd and VMHvl may integrate these sensory cues to regulate aggressive state. Indeed, discrimination between sex cues at the neural level in MeApd and in VMHvl correlates with aggressive behavior in males [65,68]. Other circuits that are interconnected with the MeA and VMH have also been implicated in aggression, including the *medial preoptic area* (MPOA) and the *periaqueductal grey* (PAG) [10]. In primates, including in humans, the VMH and amygdala are linked to aggression. In monkeys, temporal lobe legions which include the amygdala have been observed to cause a reduction in aggressive behavior. In humans, activity (measured by fMRI) in the hypothalamus and amygdala is higher when subjects act aggressively, and electrical stimulation of the amygdala induces subjective reports of anger.

Subcortical circuits that integrate sensory and internal state variables to modulate aggression also appear to be regulated by cortical inputs. In mice, stimulation of excitatory neurons in the *medial prefrontal cortex* (mPFC) reduces the intensity of attack behavior, and inhibition of the same neurons increases aggression [82]. Consistent with this regulatory role in the mouse, legions of the frontal lobes in primates, including in humans, have also been associated with changes in aggressive behavior. Taken together, a network of interconnected limbic and subcortical circuits, which include the MeApd, BNST, VMHvl, MPOA, and PAG, shape aggressive behavior through control of an internal aggressive state [21]. In mice, this state depends in part on integration of sensory cues from the environment, and particularly on sex-specific social cues from conspecifics transmitted through the olfactory pathway. Aggression is also regulated by other circuits, including cortical inputs from the mPFC, but the precise nature of this regulation is currently not well understood.

1.3.4.2: Social status and dominance

In many social species, including in mice and primates, individuals in a group organize into stable social hierarchies to manage limited resources, settle conflicts, and coordinate group behavior. While in some species, social hierarchies emerge to serve highly specific functions, other species use social hierarchies to organize many different aspects of group behavior, and hierarchical behavior can vary dramatically across males and females within a species [83,84]. In general, Individuals learn their rank (or social status) in a hierarchy from observation or as a result of conflicts (not always physical) with other individuals in the group [79,85]. The social status of an individual (whether they are relatively *dominant* or *subordinate*) can be an important determinant of his/her behavior toward other conspecifics, as well as susceptibility to challenges and stressors. In male mice for example, high social status is associated with greater aggression toward male conspecifics, more courtship calls toward females, and more access to limited resources [86,87]. In other rodents, it is also associated with greater resilience to symptoms of stress and anhedonia that are triggered by social defeat following inter-male conflict [88,89]. However, the relationship between social status, internal state and behavior varies widely across species and context [83]. For example, in some monkey species, high status males have greater access to reproductive opportunities, but have elevated levels of stress hormones, possibly because they are constantly warding off challengers [90]. Even in humans, measures of perceived societal status are anticorrelated with symptoms of depression and anxiety, suggesting some level of psychological resilience derived from status [90].

Although social hierarchies are a ubiquitous feature of social life in many species, relatively little is understood about the neurobiology and circuit mechanisms underlying social status and the expression of dominance (or subordinate) behaviors. In crustaceans and cichlid fish, serotonergic (5-HT) signaling is linked to dominance behavior [91,92], and in monkeys, the size of the dorsal Raphé nucleus (the primary source of 5-HT neurons in the brain) correlates with social rank [93]. This suggests that the serotonin system may play a role in processes that encode social status and regulate expression of dominance behaviors in some species [41,83]. In rodents, the role of 5-HT signaling is unclear; however, recent studies have begun to dissect specific circuits involved in dominance behavior. In male mice, photostimulation of excitatory inputs from the thalamus to the mPFC increases social status [86], and single neurons in the mPFC respond during effortful behaviors in an inter-male dominance test [46,87]. Interestingly, higher social status is associated with greater synaptic strength in the thalamus-mPFC synapse, a connection that is strengthened following winning in effortful conflict. This role for frontal cortex in the mouse is broadly consistent with studies in humans and non-human primates. Activity measured in the rostromedial PFC of humans using fMRI is elevated when subjects think about status relationships between individuals [78,94], and single neurons in the monkey orbitofrontal cortex have been found to correlate with viewing of high-status social familiars [70].

Less is known about how sensory cues and internal state variables are integrated to form a representation of social status that is stable over time. Because status is a property of individuals, recognition of status relationships between individuals likely depends in part on social memory. In mice, the CA2 region of the hippocampus supports memory of previous encounters with conspecifics [95]. One speculation is that indirect projections from CA2 to the mPFC via the ventral CA1 region of the hippocampus may provide information about previous social encounters that could guide status-related behavior. Projections from mPFC to downstream subcortical targets may also control the expression of dominance behavior in particular contexts. One recent study found a specific role for mPFC projections to the lateral hypothalamus in dominance behavior during a competition for food [96]. In monkeys, information about relative status is encoded in the amygdala as well as in the PFC, suggesting that this region may also be relevant for the expression of dominance behavior in primates [70].

1.3.4.3: Mating and sexual behavior

Mating behavior is essential for all sexually reproducing species, as it forms the basis for the interactions between opposite sex conspecifics that result in offspring and species survival. In many species, mating interactions proceed through distinct phases, from courtship behavior to precopulatory exploration and culminating in copulation [10]. Related to mating behavior are other repertoires of behavior that serve to coordinate mate selection, such as the establishment of social hierarchies [79,83,90], and to facilitate offspring survival, such as parenting behavior [10]. Thus, while mating involves a unique subset of social behaviors that are often stereotyped within a species, it is connected in important ways with other types of social behavior that facilitate reproduction and species survival more generally [7].

In mice, several limbic and hypothalamic regions have been implicated in the control of both male and female mating behavior. Many regions involved in mating receive input carrying pheromonal information from the VNO, suggesting that integration of these sensory cues is critical for coordination of mating behavior in mice. In females for example, neurons in the posteroventral region of the MeA (MeApv) are necessary for lordosis behavior [97], as are neurons in the downstream dorsomedial VMH (VMHdm) that receive direct and indirect inputs from the

MeApv [27]. In males, photostimulation experiments show that MeApd neurons control mounting behavior [24], and specific subpopulations of neurons in the VMHvl are active during mounting (but not during other social behaviors, such as attack) [27,68]. These findings suggest that the MeA and its downstream targets are involved in the integration of pheromonal sensory cues to coordinate mating behavior in both sexes. Consistent with this, the normal male preference for interaction with females depends on neural signaling in the MeA and its regulation by oxytocin [98], and neurons in the BNST (downstream of the MeA) are active during mating and are also required for opposite sex preference [66]. Additionally, lesioning or silencing neural activity in the MPOA, another projection target of the MeA, results in reduced mating behavior in males, suggesting that normal mating behavior depends on integration of signals across a network of interconnected hypothalamic circuits [28,99]. Consistent with these results from circuit analysis in rodents, many of the same limbic and hypothalamic regions have been implicated in sexual behavior in primates, including in humans. For example, fMRI studies have identified activity patterns in the human hypothalamus that correlate with sexual identity [100], and in both men and women, amygdala activation is correlated with sexual responsivity [101,102]. Amygdala legions in humans have also been associated with abnormal sexual behavior, such as hypersexualized states [103].

Like all motivated behaviors, sexual behavior requires not only processing relevant sensory information but also the integration of this with reward and motivational states to generate a behavioral drive. Perhaps not surprisingly, the classic mesolimbic reward system has been implicated in the control of mating behavior in mice and primates. The ventral tegmental area (VTA), which contains much of the dopaminergic input to the ventral striatum and cortex, receives inputs from the MeA, BNST, and MPOA. VTA dopamine signaling has been linked to sexual desire in humans, and activity in the nucleus accumbens (NAc/ventral striatum, a major projection target of VTA dopamine neurons) is correlated with sexual desire and arousal [101]. In both male and female mice, dopamine release in the VTA-NAc pathway is increased during anticipation of sexual contact [104], and manipulation of this pathway bidirectionally modulates time spent investigating opposite sex conspecifics. And in male mice, dopamine signaling in the NAc in necessary for opposite sex preference [105], supporting a role for mesolimbic circuits in transforming integrated sensory cues (in the MeA and hypothalamic network) into motivated behavioral drive.

Finally, in addition to integration of sensory cues and mesolimbic control of sexual motivation, cortical circuits also appear to regulate mating behavior. In female mice, oxytocin signaling in the medial prefrontal cortex (mPFC) mediates social and sexual interest in male conspecifics [106,107]. Little is known about the potential role of prefrontal circuits in male sexual behavior, but the control of male social status by mPFC neurons may play an indirect role in mating, as dominant males are more reproductively successful [42,86,87]. In humans, prefrontal regions may participate in control of sexual behavior and sexual inhibition. Societally concordant social behavior depends on shaping thoughts and actions in a highly context-fluid manner, which demands strong executive and attentional control. Consistent with this, patients with legions in the orbitofrontal cortex sometimes show hypersexuality, which may be interpreted as a disinhibition of sexual drive [103,108].

1.3.5: Behavioral expression

Integration of social sensory inputs with the internal state of the animal allows specific behavioral decisions to be selected in a flexible, context-dependent manner, which is highly adaptive. Still, in any circumstance, only a small subset of possible behavioral changes can be expressed at a time. In this section, I will discuss how sensory transduction and internal integration lead to the selection of specific behavioral choices or internal state changes.

For innate behaviors such as aggression and mating/sexual behavior, a network of interconnected limbic and subcortical circuits appears to be involved in integrating relevant social cues and encoding internal states that drive those behavioral patterns. These involve primarily the MeA, VMH, BNST, and MPOA (described in the previous section). In this sense, these structures clearly play some role in the expression of social behaviors – photostimulation experiments suggest that activation of parts of the MeA and VMH can increase aggression and mating behavior [24,25], and silencing experiments indicate that activity in these brain regions is also necessary for the expression of normal behavior. However, despite the involvement of these circuits in encoding/controlling behavioral drive, downstream circuits may play a more specific role in the gating and expression of specific behavioral choices. For example, the periaqueductal gray (PAG), a region that receives input from the VMHvl and other hypothalamic regions, appears to be important for the coordination of attack motor patterns in male mice. Distinct from neurons the VMHvl that show highly mixed responses to social sensory cues and during aggressive behavior, neurons in PAG are specifically active during attack behavior and exhibit time-locked responses during biting [109]. Silencing of PAG neurons in males decreases aggression, and in females, PAG legions increase aggression but decrease lordosis behavior. Similarly, amygdala inputs to the PAG are necessary to coordinate defensive behavior across species [110,111], and VMHdm projections to the PAG are necessary to coordinate escape behavior [112]. These observations suggest that, although there may be differences in how behavioral output is gated between males and females, the PAG plays a critical role in transforming signals from the limbic/hypothalamic network into directed behavioral choices.

Cortical circuits also play a role in gating behavioral choices. The mPFC, which sends projections to the PAG and other subcortical regions including the VTA and NAc, provides an anatomical substrate for top-down cognitive control of behavioral gating [113]. In mice, neurons in the mPFC that project to dorsal PAG are preferentially active during shock but not during reward, suggesting that prefrontal input may help to coordinate responses to threats [112]. Neurons in the mPFC are active during social approach behavior, indicating possible involvement in the decision to engage in social interaction [46,114]. In addition, prefrontal circuits may play a role in the expression of social behavioral changes over a longer timescale. For example, social defeat is a state of decreased social preference and anhedonia in mice that can be triggered by repeated defeat in antagonistic encounters [53]. The behavioral expression of this internal state, and specifically the increase in social avoidance, is mediated by inputs to the PAG from mPFC [115]. mPFC projections to the Raphé nucleus also control effortful behavior in response to behavioral challenge [116], and over longer timescales, activity in mPFC neurons is correlated with increased resilience to defeat symptoms [117]. Together, these observations suggest that the mPFC plays a role in the modulation of selected behavior (e.g., coordinating escape in response to threat), as well as in shaping internal states that affect behavior over longer timescales.

Together, these three stages of sensory transduction, internal integration, and behavioral expression capture the core processes of decision making in the brain. Still, it bears repeating that these “stages” are not completely separable and not necessarily sequential. As is clear from the discussion, many brain regions and circuits appear to play a role in more than one, or even in all three, of these stages [10,118]. For some behaviors, we may discover that there is some anatomical boundary between the circuits that process information leading to a decision and the circuit(s) that implement behavioral selection and expression [119]. In other cases, however, we may find that the process is fundamentally a gradual one, with no sharp discontinuity between integration of information and the decision point. Future research can shed more light on the underlying logic of how neural circuits implement complex behavioral choices, and how exactly these processes of transduction, integration, and expression are intertwined.

1.4: References

1. Shenk, J. W. What makes us happy? *The Atlantic* (2009).

2. Kawachi, I. & Berkman, L. F. Social ties and mental health. *J. Urban Heal.* **78**, 458–467 (2001).

3. Kumar, P. *et al.* Increased neural response to social rejection in major depression. *Depress. Anxiety* **34**, 1049–1056 (2017).

4. Cacioppo, J. T. & Hawkley, L. C. Perceived social isolation and cognition. *Trends in Cognitive Sciences* **13**, 447–454 (2009).

5. Lee, C. R., Chen, A. & Tye, K. M. The neural circuitry of social homeostasis: Consequences of acute versus chronic social isolation. *Cell* **184**, 1500–1516 (2021).

6. Matthews, G. A. & Tye, K. M. ANNALS OF THE NEW YORK ACADEMY OF SCIENCES Neural mechanisms of social homeostasis. *Ann. N.Y. Acad. Sci* doi:10.1111/nyas.14016

7. Tinbergen, N. *The study of instinct*. (1951).

8. Anderson, D. J. & Adolphs, R. A framework for studying emotions across species. *Cell* **157**, 187–200 (2014).

9. Preston, S. D. & de Waal, F. B. M. Empathy: Its ultimate and proximate bases. *Behav. Brain Sci.* **25**, 1–20; discussion 20-71 (2002).

10. Chen, P. & Hong, W. Neural Circuit Mechanisms of Social Behavior. *Neuron* **98**, 16–30 (2018).

11. Anderson, D. J. & Adolphs, R. A framework for studying emotions across species. *Cell* **157**, 187–200 (2014).

12. Adolphs, R. Conceptual Challenges and Directions for Social Neuroscience. *Neuron* **65**,

752–767 (2010).

13. Stanley, D. A. & Adolphs, R. Toward a Neural Basis for Social Behavior. *Neuron* **80**, 816–826 (2013).

14. Bennett, A. J. & Ringach, D. L. Animal Research in Neuroscience: A Duty to Engage. *Neuron* **92**, 653–657 (2016).

15. Ringach, D. L. The use of nonhuman animals in biomedical research. in *American Journal of the Medical Sciences* **342**, 305–313 (Lippincott Williams and Wilkins, 2011).

16. Ochsner, K. N. & Lieberman, M. D. The emergence of social cognitive neuroscience. *Am. Psychol.* **56**, 717–34 (2001).

17. Schilbach, L. *et al.* Toward a second-person neuroscience. *Behav. Brain Sci.* **36**, 393–414 (2013).

18. Bateson, P. & Laland, K. N. Tinbergen's four questions: An appreciation and an update. *Trends in Ecology and Evolution* **28**, 712–718 (2013).

19. Tinbergen, N. On aims and methods of Ethology. *Z. Tierpsychol.* **20**, 410–433 (1963).

20. Kohl, J. *et al.* Functional circuit architecture underlying parental behaviour. *Nature* **556**, 326–331 (2018).

21. Anderson, D. J. Optogenetics, Sex, and Violence in the Brain: Implications for Psychiatry. *Biol. Psychiatry* **71**, 1081–1089 (2012).

22. Hess, W. R. & Brügger, M. Das subkortikale Zentrum der affektiven Abwehrreaktion [The subcortical center for affective defense reactions]. *Helv. Physiol. Pharmacol. Acta* **1**, 33–52 (1943).

23. Kruk, M. R. Hypothalamic attack: A wonderful artifact or a useful perspective on escalation

and pathology in aggression? A viewpoint. *Curr. Top. Behav. Neurosci.* **17**, 143–188 (2014).

24. Hong, W., Kim, D.-W. & Anderson, D. J. Antagonistic control of social versus repetitive self-grooming behaviors by separable amygdala neuronal subsets. *Cell* **158**, 1348–1361 (2014).

25. Lin, D. *et al.* Functional identification of an aggression locus in the mouse hypothalamus. *Nature* **470**, 221–226 (2011).

26. Kimchi, T., Xu, J. & Dulac, C. A functional circuit underlying male sexual behaviour in the female mouse brain. *Nature* **448**, 1009–1014 (2007).

27. Lee, H. *et al.* Scalable control of mounting and attack by Esr1+ neurons in the ventromedial hypothalamus. *Nature* **509**, 627–632 (2014).

28. Wu, Z., Autry, A. E., Bergan, J. F., Watabe-Uchida, M. & Dulac, C. G. Galanin neurons in the medial preoptic area govern parental behaviour. *Nature* **509**, 325–330 (2014).

29. Chen, P. B. *et al.* Sexually Dimorphic Control of Parenting Behavior by the Medial Amygdala. *Cell* **176**, 1206-1221.e18 (2019).

30. Yizhar, O., Fenno, L. E., Davidson, T. J., Mogri, M. & Deisseroth, K. Optogenetics in Neural Systems. *Neuron* **71**, 9–34 (2011).

31. Luo, L., Callaway, E. M. & Svoboda, K. Genetic Dissection of Neural Circuits. *Neuron* **57**, 634–660 (2008).

32. Falkner, A. L., Dollar, P., Perona, P., Anderson, D. J. & Lin, D. Decoding ventromedial hypothalamic neural activity during male mouse aggression. *J. Neurosci.* **34**, 5971–84 (2014).

33. Adolphs, R. The Social Brain: Neural Basis of Social Knowledge. *Annu. Rev. Psychol.* **60**, 693–716 (2009).

34. Spunt, R. P. & Adolphs, R. A new look at domain specificity: Insights from social neuroscience. *Nature Reviews Neuroscience* **18**, 559–567 (2017).

35. Adolphs, R. Investigating the cognitive neuroscience of social behavior. *Neuropsychologia* **41**, 119–126 (2003).

36. Rilling, J. K. & Sanfey, A. G. The Neuroscience of Social Decision-Making. *Annu. Rev. Psychol.* **62**, 23–48 (2011).

37. Schilbach, L. *et al.* Toward a second-person neuroscience. *Behav. Brain Sci.* **36**, 393–414 (2013).

38. Redcay, E. & Schilbach, L. Using second-person neuroscience to elucidate the mechanisms of social interaction. *Nat. Rev. Neurosci.* **20**, 495–505 (2019).

39. Kingsbury, L. & Hong, W. A Multi-Brain Framework for Social Interaction. *Trends Neurosci.* (2020). doi:10.1016/j.tins.2020.06.008

40. Sanfey, A. G. Social Decision-Making: Insights from Game Theory and Neuroscience. *Science (80-.).* **318**, 598–602 (2007).

41. Zink, C. F. *et al.* Know Your Place: Neural Processing of Social Hierarchy in Humans. *Neuron* **58**, 273–283 (2008).

42. Wang, F. *et al.* The mouse that roared: neural mechanisms of social hierarchy. *Trends Neurosci.* **37**, 674–82 (2014).

43. Montague, P. R. *et al.* Hyperscanning: simultaneous fMRI during linked social interactions. *Neuroimage* **16**, 1159–64 (2002).

44. Kennedy, A. *et al.* Internal States and Behavioral Decision-Making: Toward an Integration of Emotion and Cognition. *Cold Spring Harb. Symp. Quant. Biol.* **79**, 199–210 (2014).

45. King-Casas, B. *et al.* Getting to Know You: Reputation and Trust in a Two-Person Economic Exchange. *Science (80-.).* **308**, 78–83 (2005).

46. Kingsbury, L. *et al.* Correlated Neural Activity and Encoding of Behavior across Brains of Socially Interacting Animals. *Cell* **178**, 429-446.e16 (2019).

47. Zhang, W. & Yartsev, M. M. Correlated Neural Activity across the Brains of Socially Interacting Bats. *Cell* **178**, 413-428.e22 (2019).

48. Wallace, K. J. & Hofmann, H. A. Decision-making in a social world: Integrating cognitive ecology and social neuroscience. *Curr. Opin. Neurobiol.* **68**, 152–158 (2021).

49. E. R., K., J. H., S. & T. M., J. *Principles of neural science*. (New York: McGraw-Hill, Health Professionals Division, 2000).

50. Marr, D. *Vision: a computational investigation into the human representation and processing of visual information Vision: a Computational Investigation Into the Human Representation and Processing of Visual Information*. (The MIT Press, 1982).

51. Li, Y. & Dulac, C. Neural coding of sex-specific social information in the mouse brain. *Current Opinion in Neurobiology* **53**, 120–130 (2018).

52. Kennedy, A. *et al.* Stimulus-specific hypothalamic encoding of a persistent defensive state. *Nature* **586**, 730–734 (2020).

53. Chaouloff, F. Social stress models in depression research: what do they tell us? *Cell Tissue Res.* **354**, 179–90 (2013).

54. Ringach, D. L. *et al.* Spatial clustering of tuning in mouse primary visual cortex. *Nat. Commun.* **7**, (2016).

55. Livingstone, M. & Hubel, D. Segregation of form, color, movement, and depth: Anatomy,

physiology, and perception. *Science (80-.).* **240**, 740–749 (1988).

56. Pasupathy, A., Popovkina, D. V. & Kim, T. Visual Functions of Primate Area V4. *Annual Review of Vision Science* **6**, 363–385 (2020).

57. Tsao, D. Y., Schweers, N., Moeller, S. & Freiwald, W. A. Patches of face-selective cortex in the macaque frontal lobe. *Nat. Neurosci.* **11**, 877–879 (2008).

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

59. Baldauf, D. & Desimone, R. Neural mechanisms of object-based attention. *Science* **344**, 424–7 (2014).

60. Minderer, M., Brown, K. D. & Harvey, C. D. The Spatial Structure of Neural Encoding in Mouse Posterior Cortex during Navigation. *Neuron* **102**, 232-248.e11 (2019).

61. Luo, M., Fee, M. S. & Katz, L. C. Encoding Pheromonal Signals in the Accessory Olfactory Bulb of Behaving Mice. *Science (80-.).* **299**, 1196–1201 (2003).

62. Ben-Shaul, Y., Katz, L. C., Mooney, R. & Dulac, C. In vivo vomeronasal stimulation reveals sensory encoding of conspecific and allospecific cues by the mouse accessory olfactory bulb. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 5172–5177 (2010).

63. Keller, M., Douhard, Q., Baum, M. J. & Bakker, J. Destruction of the main olfactory epithelium reduces female sexual behavior and olfactory investigation in female mice. *Chem. Senses* **31**, 315–323 (2006).

64. Stowers, L., Holy, T. E., Meister, M., Dulac, C. & Koentges, G. Loss of Sex Discrimination and Male-Male Aggression in Mice Deficient for TRP2. *Science (80-.).* **295**, 1493–1500 (2002).

65. Li, Y. *et al.* Neuronal Representation of Social Information in the Medial Amygdala of

Awake Behaving Mice. *Cell* **171**, 1176-1190.e17 (2017).

66. Bayless, D. W. *et al.* Limbic Neurons Shape Sex Recognition and Social Behavior in Sexually Naive Males. *Cell* **176**, 1190-1205.e20 (2019).

67. Bergan, J. F., Ben-Shaul, Y. & Dulac, C. Sex-specific processing of social cues in the medial amygdala. *Elife* **3**, e02743 (2014).

68. Remedios, R. *et al.* Social behaviour shapes hypothalamic neural ensemble representations of conspecific sex. *Nature* **550**, 388–392 (2017).

69. Mosher, C. P., Zimmerman, P. E. & Gothard, K. M. Neurons in the Monkey Amygdala Detect Eye Contact during Naturalistic Social Interactions. *Curr. Biol.* **24**, 2459–2464 (2014).

70. Munuera, J., Rigotti, M. & Salzman, C. D. Shared neural coding for social hierarchy and reward value in primate amygdala. *Nat. Neurosci.* **21**, 415–423 (2018).

71. Wöhr, M. & Schwarting, R. K. W. Affective communication in rodents: Ultrasonic vocalizations as a tool for research on emotion and motivation. *Cell and Tissue Research* **354**, 81–97 (2013).

72. Kato, H. K., Asinof, S. K. & Isaacson, J. S. Network-Level Control of Frequency Tuning in Auditory Cortex. *Neuron* **95**, 412-423.e4 (2017).

73. Romero, S. *et al.* Cellular and Widefield Imaging of Sound Frequency Organization in Primary and Higher Order Fields of the Mouse Auditory Cortex. *Cereb. Cortex* **30**, 1603–1622 (2020).

74. Kim, H. & Bao, S. Experience-dependent overrepresentation of ultrasonic vocalization frequencies in the rat primary auditory cortex. *J. Neurophysiol.* **110**, 1087–1096 (2013).

75. Marlin, B. J., Mitre, M., D'Amour, J. A., Chao, M. V. & Froemke, R. C. Oxytocin enables

maternal behaviour by balancing cortical inhibition. *Nature* **520**, 499–504 (2015).

76. Schiavo, J. K. *et al.* Innate and plastic mechanisms for maternal behaviour in auditory cortex. *Nature* **587**, 426–431 (2020).

77. Zhou, T., Sandi, C. & Hu, H. Advances in understanding neural mechanisms of social dominance. *Curr. Opin. Neurobiol.* **49**, 99–107 (2018).

78. Utevsky, A. V & Platt, M. L. Status and the brain. *PLoS Biol.* **12**, e1001941 (2014).

79. Williamson, C. M., Lee, W. & Curley, J. P. Temporal dynamics of social hierarchy formation and maintenance in male mice. *Anim. Behav.* **115**, 259–272 (2016).

80. Stagkourakis, S. *et al.* A neural network for intermale aggression to establish social hierarchy. *Nat. Neurosci.* **21**, 834–842 (2018).

81. Falkner, A. L., Grosenick, L., Davidson, T. J., Deisseroth, K. & Lin, D. Hypothalamic control of male aggression-seeking behavior. *Nat. Neurosci.* **19**, 596–604 (2016).

82. Takahashi, A., Nagayasu, K., Nishitani, N., Kaneko, S. & Koide, T. Control of intermale aggression by medial prefrontal cortex activation in the mouse. *PLoS One* **9**, e94657 (2014).

83. Chiao, J. Y. Neural basis of social status hierarchy across species. *Curr. Opin. Neurobiol.* **20**, 803–809 (2010).

84. van den Berg, W. E., Lamballais, S. & Kushner, S. A. Sex-Specific Mechanism of Social Hierarchy in Mice. *Neuropsychopharmacology* **40**, 1364–1372 (2015).

85. Stagkourakis, S. *et al.* A neural network for intermale aggression to establish social hierarchy. *Nat. Neurosci.* **21**, 834–842 (2018).

86. Zhou, T. *et al.* History of winning remodels thalamo-PFC circuit to reinforce social dominance. *Science (80-.).* **357**, 162–168 (2017).

87. Wang, F. *et al.* Bidirectional Control of Social Hierarchy by Synaptic Efficacy in Medial Prefrontal Cortex. *Science (80-.).* **334**, (2011).

88. Cooper, M. A., Clinard, C. T. & Morrison, K. E. Neurobiological mechanisms supporting experience-dependent resistance to social stress. *Neuroscience* **291**, 1–14 (2015).

89. Morrison, K. E. *et al.* Maintenance of dominance status is necessary for resistance to social defeat stress in Syrian hamsters. *Behav. Brain Res.* **270**, 277–86 (2014).

90. Sapolsky, R. M. The Influence of Social Hierarchy on Primate Health. *Science (80-.).* **308**, 648–652 (2005).

91. Loveland, J. L., Uy, N., Maruska, K. P., Carpenter, R. E. & Fernald, R. D. Social status differences regulate the serotonergic system of a cichlid fish, Astatotilapia burtoni. *J. Exp. Biol.* **217**, 2680–2690 (2014).

92. Edwards, D. H., Issa, F. A. & Herberholz, J. The neural basis of dominance hierarchy formation in crayfish. *Microsc. Res. Tech.* **60**, 369–376 (2003).

93. Noonan, M. P. *et al.* A Neural Circuit Covarying with Social Hierarchy in Macaques. *PLoS Biol.* **12**, e1001940 (2014).

94. Ligneul, R., Obeso, I., Ruff, C. C. & Dreher, J.-C. Dynamical Representation of Dominance Relationships in the Human Rostromedial Prefrontal Cortex. *Curr. Biol.* **26**, 3107–3115 (2016).

95. Hitti, F. L. & Siegelbaum, S. A. The hippocampal CA2 region is essential for social memory. *Nature* **508**, 88–92 (2014).

96. Padilla-Coreano, N. *et al.* A cortical-hypothalamic circuit decodes social rank and promotes dominance behavior. (2020). doi:10.21203/rs.3.rs-94115/v1

97. Ishii, K. K. *et al.* A Labeled-Line Neural Circuit for Pheromone-Mediated Sexual Behaviors in Mice. *Neuron* **95**, 123-137.e8 (2017).

98. Yao, S., Bergan, J., Lanjuin, A. & Dulac, C. Oxytocin signaling in the medial amygdala is required for sex discrimination of social cues. *Elife* **6**, (2017).

99. Hull, E. M. & Dominguez, J. M. Sexual behavior in male rodents. *Horm. Behav.* **52**, 45–55 (2007).

100. Brunetti, M. *et al.* Hypothalamus, sexual arousal and psychosexual identity in human males: A functional magnetic resonance imaging study. *Eur. J. Neurosci.* **27**, 2922–2927 (2008).

101. Wise, N. J., Frangos, E. & Komisaruk, B. R. Brain Activity Unique to Orgasm in Women: An fMRI Analysis. *J. Sex. Med.* **14**, 1380–1391 (2017).

102. Ferretti, A. *et al.* Dynamics of male sexual arousal: Distinct components of brain activation revealed by fMRI. *Neuroimage* **26**, 1086–1096 (2005).

103. Kühn, S. & Gallinat, J. Neurobiological Basis of Hypersexuality. in *International Review of Neurobiology* **129**, 67–83 (Academic Press Inc., 2016).

104. Micevych, P. E. & Meisel, R. L. Integrating neural circuits controlling female sexual behavior. *Frontiers in Systems Neuroscience* **11**, (2017).

105. Beny-Shefer, Y. *et al.* Nucleus Accumbens Dopamine Signaling Regulates Sexual Preference for Females in Male Mice. *Cell Rep.* **21**, 3079–3088 (2017).

106. Nakajima, M., Görlich, A. & Heintz, N. Oxytocin modulates female sociosexual behavior through a specific class of prefrontal cortical interneurons. *Cell* **159**, 295–305 (2014).

107. Moore, K. M. *et al.* Glutamate Afferents From the Medial Prefrontal Cortex Mediate

Nucleus Accumbens Activation by Female Sexual Behavior. *Front. Behav. Neurosci.* **13**, (2019).

108. Rodriguez-Nieto, G., Sack, A. T., Dewitte, M., Emmerling, F. & Schuhmann, T. Putting out the blaze: The neural mechanisms underlying sexual inhibition. *PLoS One* **14**, (2019).

109. Falkner, A. L. *et al.* Hierarchical Representations of Aggression in a Hypothalamic-Midbrain Circuit. *Neuron* **106**, 637-648.e6 (2020).

110. Li, Y. *et al.* Hypothalamic Circuits for Predation and Evasion. *Neuron* **97**, 911-924.e5 (2018).

111. Terburg, D. *et al.* The Basolateral Amygdala Is Essential for Rapid Escape: A Human and Rodent Study. *Cell* **175**, 723-735.e16 (2018).

112. Lefler, Y., Campagner, D. & Branco, T. The role of the periaqueductal gray in escape behavior. *Current Opinion in Neurobiology* **60**, 115–121 (2020).

113. Ko, J. Neuroanatomical Substrates of Rodent Social Behavior: The Medial Prefrontal Cortex and Its Projection Patterns. *Front. Neural Circuits* **11**, 41 (2017).

114. Lee, E. *et al.* Enhanced Neuronal Activity in the Medial Prefrontal Cortex during Social Approach Behavior. *J. Neurosci.* **36**, 6926–36 (2016).

115. Franklin, T. B. *et al.* Prefrontal cortical control of a brainstem social behavior circuit. *Nat. Neurosci.* **20**, 260–270 (2017).

116. Warden, M. R. *et al.* A prefrontal cortex–brainstem neuronal projection that controls response to behavioural challenge. *Nature* **492**, 428 (2012).

117. Morrison, K. E., Bader, L. R., McLaughlin, C. N. & Cooper, M. A. Defeat-induced activation of the ventral medial prefrontal cortex is necessary for resistance to conditioned defeat. *Behav. Brain Res.* **243**, 158–64 (2013).

118. Raam, T. & Hong, W. Organization of neural circuits underlying social behavior: A consideration of the medial amygdala. *Curr. Opin. Neurobiol.* **68**, 124–136 (2021).

119. Evans, D. A. *et al.* A synaptic threshold mechanism for computing escape decisions. *Nature* **558**, 590–594 (2018).

CHAPTER TWO

Methods in Social Neuroscience

2.1: Measurement of the brain and behavior

2.1.1: The relationship between brain and behavior

Social neuroscience investigates how biological processes within the brain give rise to social behavior and social interaction. Because of this focus, many of the driving questions in the field are essentially relational – that is, they are concerned with the relationship *between* the brain and behavior, rather than either of these in isolation [1,2] As in any scientific field, inquiry into the relationship between different processes depends on both *measurement* and *manipulation*. In order to generate theories and hypotheses about how the brain shapes social behavior, we need to measure brain processes and behavior, and then analyze how they are related; in order to test our hypotheses, we need to manipulate the brain or the animal's behavior, and then observe the consequences.

In this section, I will discuss the methods used in social neuroscience to measure brain processes and animal behavior, as well as considerations and challenges that are important in the study of social interaction. In the following section, I will outline a conceptual framework for the analysis of neural and behavioral data from social neuroscience experiments, highlighting three main directions of inquiry: *dimensionality, relationality, and temporality*. Finally, I will touch on methods for manipulating neural activity in the brain and considerations for experimental design in the context of studying social interaction.

2.1.2: Measuring processes in the brain

In order to understand the processes in the brain that shape social behavior, we need to measure them. Depending on the questions driving an experiment, this can involve measurement of gene expression, synaptic plasticity, hormonal and neuromodulatory signals, and in many cases, the activity of neurons in the brain. There are several different approaches currently available to measure neural activity which each carry distinct advantages and considerations. In human subjects, where invasive methods often cannot be used (except in rare cases with some clinical subjects [3,4]), common methods include functional magnetic resonance imaging (fMRI), which measures the oxygenation level of blood in different parts of the brain as a proxy for neural activity [5–7], and electroencephalography (EEG), which measures changes in electrical potential of groups of neurons at the cortical surface of the brain [8,9]. fMRI and EEG each permit different levels of spatial and temporal resolution, and together are very powerful methods for recording brain activity in human subjects and mapping behavior and cognitive processes to specific regions of the brain. However, both methods lack cellular-level resolution, and so make it difficult to pin down mechanisms of neural processes with a high degree of precision.

Invasive recording techniques, which can be used more readily in animal subjects, offer high enough spatial resolution to capture the activity of individual neurons in vivo. Probably the most common method used to measure single neuron activity in vivo is electrode recording. Electrodes implanted into the brain can measure changes in the electrical potential across the plasma membrane of individual neurons, and implants with multiple electrode arrays can measure simultaneous activity from populations of neurons in the same brain region. Using this method, researchers have been able to identify individual neurons that show increased activity during specific types of behavioral decisions [10–13] or during detection of specific social sensory cues [14–17]. Recent developments in electrode recording technology have enabled much higher yield and access to multiple regions of the brain simultaneously, opening investigation of large scale dynamics across the brain and between regions within the brain [18–21].

Optical methods have also been developed to record activity from populations of neurons [22,23]. Calcium imaging is a technique that combines microscopy with genetically encoded calcium indicators to measure fluorescence changes in individual neurons that are correlated with intracellular calcium concentrations [24]. Because action potentials produce rapid changes in intracellular calcium (through opening of calcium channels in the plasma membrane), these optically recorded changes in calcium concentration can be used as a proxy for neural activity [22,25]. As calcium imaging is an optical method, it also contains information about the spatial arrangement of neurons in the imaging field of view, and if used in combination with molecular markers, can give information about cell identity [26–30]. Using two-photon calcium imaging, one can resolve activity of hundreds to thousands of neurons, opening investigation into the dynamics and structure of population responses during behavior. These advantages come at a cost, however. Since the fluorescence signal is a measure of calcium concentration, which has slower dynamics than changes in electrical potential, calcium imaging cannot explicitly capture spiking activity. Thus, at present, electrophysiological methods are favored for investigation of hypotheses that involve direct examination of spiking activity.

In the context of social neuroscience, we are often interested in how processes in the brain relate to natural social behavior during interactions with conspecifics. Thus, techniques that enable recordings of large populations of neurons in freely behaving animals are desired. In recent years, technology has been developed to adapt optical methods to freely moving animals through the use of miniaturized light microscopes that can resolve calcium fluorescence through a lens implanted in the subject animal's brain [31–33]. These "microendoscopes" enable investigation of large-scale neural population activity during natural animal behavior, and are therefore ideal for neural recordings in social neuroscience experiments [30,34,35]. However, because the microscopy hardware must be small and lightweight, two-photon imaging is difficult to achieve using such as setup (though this limitation may be resolved in the near future [36,37]), and the quality of fluorescence images acquired with "single-photon" imaging is relatively poor. Owing to these considerations, it may still be preferred for some applications to use electrophysiological methods rather than microendoscopes, especially if high temporal resolution and/or spiking activity is desired.

In the studies presented in chapters 3 and 4, microendoscopes were used to record large-scale population activity from the prefrontal cortex of mice during unconstrained social interaction. This method has been previously applied in social contexts [30,34,35], and in our case, was chosen because of a preference for high yield recordings of hundreds of neurons at the cost of temporal resolution that could be afforded using electrodes. Because the social behavior events under examination occur over a timescale on the order of seconds, high resolution spiking data was not a priority when designing the experimental protocol and methods. However, electrode recordings, especially across multiple brain regions in the same animal [18,19], may be very useful in the future to investigate single-neuron coding principles and inter-regional activity patterns more closely.

2.1.3: Measuring animal behavior

In any experiment, it is critical to make careful measurements of the phenomena of interest and to quantify them precisely – this is necessary to identify patterns, formulate theories, and test predictions. In the context of social neuroscience, careful measurement of behavior is particularly important because social interaction is highly complex [38], and social behavior highly variable [2].

As discussed in chapter 1, part of the reason that social behavior is complex is that interaction and communication between animals often takes place across more than one sensory modality. This means that without substantial prior knowledge, we often do not know exactly which measured variables correspond to the most relevant communication channels. For example, two mice may use olfactory and tactile cues to coordinate behavior, but these may not be relevant to the same degree or at the same time. Communication channels may also change as an interaction evolves. For example, mice may initiate an interaction with investigatory sniffing and then proceed to physical engagement with vocalizations. Without measuring this communication space comprehensively, we may miss important information about the interaction and how it relates to processes in the brain.

In addition to this, because animal behavior is shaped by many different factors, it is also highly variable across time and across individuals. For example, as an animal's internal state (motivation, hunger, etc.) changes over minutes and hours, its decisions during similar social engagements may be very different [2,39]. Different animals may also have very different life experiences due to their previous interactions with conspecifics and other environmental factors. While this behavioral variability may sometimes be an experimental nuisance, it can also reveal important patterns in behavior that may shed light on the underlying biological mechanisms. Without measuring and analyzing this variability, we may miss patterns that could lead to new theories and hypotheses.

With these considerations in mind, there are of course practical limitations on how much we can reasonably measure in a single experiment. One way expand our description of behavior, outside of expanding the repertoire of measurement tools, is to refine current approaches to capture more and to do so more objectively [40]. Traditionally, measurement of animal behavior has typically been done using video capture and hand scoring of behavioral events by researchers. While this can be effective, variability in subjective measurement can also lead to inconsistencies in how behaviors are defined, and manual scoring can be prohibitively time consuming for some experiments. Recent advances in pattern recognition and machine vision technologies have enabled automated tracking of animals and in some cases automated identification of their behavior [41–43]. These methods allow for objective quantification of animal behavior in place of manual scoring, dramatically increase the efficiency of behavior analysis, and allow for precise measure of subtle behavioral changes that may be informative but are difficult to identify manually.

2.2: Analytical methods in social neuroscience

A typical set of data obtained from a social neuroscience experiment contains measurements of events in the environment, which include a description of the subject animal's behavior, and measurements of biological processes in the animal's brain, often of neural activity. Such complex datasets open a very wide range of questions that can potentially give insight into how processes in the brain give rise to behavior. In this section, I will outline a conceptual framework (loosely inspired by Tinbergen's four questions) to organize investigation of the brain and behavior using datasets obtained from measuring both. At the highest level, this framework considers the "question space" to be spanned by three basic directions of inquiry: *dimensionality* (how are variables related *within* a set of data?), *relationality* (how are variables related *between* sets of data?), and *temporality* (how does the structure of the data change over time?). While these three directions are not exhaustive and are an obvious simplification, I believe they provide a useful way to organize inquiry and help to highlight the relationships between different types of questions, their limitations, and their assumptions. In the following section, I will then give a more technical explanation of some of the important methods used in the studies presented in this dissertation (chapters 3 and 4).

2.2.1: Dimensionality

The first direction in this question space, dimensionality, is concerned with the number of variables that the researcher has measured within a dataset and the structure of their inter-relationships. For datasets of different sizes, different types of questions may be posed, and different methods may be necessary [44]. In some experiments, we may only be interested in one behavioral variable (such as choice in a task) or neural variable (such as the activity of a neuron). In these cases, standard approaches for analysis of time-series data may be sufficient. In other cases, however, we may record dozens of behavioral events or hundreds to thousands of neural signals at the same time [20,21]. Such datasets are called "high dimensional," because the number of dimensions along which the data can vary (the number of variables) is large [20,21,45]. Datasets with many variables can pose unique challenges because they are difficult to visualize, and we often do not know beforehand which variables will be interesting or informative to look at. However, in many cases, measured signals are not completely independent from one another, but are in fact inter-related and structured in some way [46]. Movement along this direction of inquiry from 1-dimensional data into multivariable datasets involves asking questions about how the measured signals are inter-related and applying methods to quantify those relationships to understand the data, and the system that generated it, more deeply.

Consider, for example, a set of data containing the weights and heights of children ranging from ages five to ten. We could analyze these variables separately. But we may also observe that weight and height are highly correlated, as they are both driven by one common underlying factor – age. This insight may lead us to view the data in a different way and form new hypotheses about how it relates to the systems under investigation. By the same token, *correlational structure* may exist in much larger datasets like the ones we get from large scale neural recordings [20,21]. In the context of the brain, this correlational structure may reflect physical relationships between neurons (reflecting synaptic connections, for example), architecture that constrains neural activity (such as common input from another brain region), or involvement in related computational processes. Generally speaking, strong *covariance* among multiple signals may reflect some underlying *factor* that governs their collective dynamics. To understand the nature of such a system, these underlying factors are useful to know about. Thus, methods for analyzing the correlation structure across multiple variables, and for identifying common underlying factors that can account for that structure, are incredibly useful for working with high dimensional data [46,47].

One class of methods for identifying underlying factors or components of the data that explain the covariance between many measured signals is called *dimensionality reduction* – so called because these methods seek to compactly describe high-dimensional datasets using a smaller number of variables [46]. *Principal component analysis* (PCA, described in detail in the next section) is perhaps the simplest and most intuitive method for reducing the dimensionality of data. Other methods, which define components based on how well they capture covariance with some other measured variable or variables (such as between neural activity and behavior), are also useful for certain questions. In the studies presented in this dissertation (chapters 3 and 4), dimensionality reduction is used to visualize the activity of many neurons using a small number of dimensions. It is also used to analyze how patterns of activity across populations of neurons change when an animal is exposed to different stimuli or makes behavioral decisions.

Finally, multivariate datasets can also be examined to understand how multiple variables relate to some other variable of interest. For example, multivariate regression (discussed below) can be used to quantify the relationship between the activity of single neurons and all the measured sensory and behavioral events in the external environment.

2.2.2: Relationality

Many questions in social neuroscience are inherently relational [1,2]. The second direction in this question space is concerned with the degree of relational structure between sets of data. That is, rather than analyzing the inter-relationships *within* one dataset (as above, with the concept of dimensionality), relational analyses consider the relationships *between* different types of data, with varying degrees of association strength. To illustrate, consider an experiment in which we have recorded activity from one hundred neurons in the brain of a mouse engaged in social interaction. On one extreme, we can pose questions about the structure of the neural data by itself (no relationality). We can analyze the statistics of a single neuron's activity (one variable; low dimensionality), or we can analyze the covariance structure of all the neurons and investigate their underlying factors or dynamic (many variables; high dimensionality). However, because we also measured the animal's behavior, we can also ask questions about how the neural activity corresponds to behavioral decisions – about the relational structure between brain and behavior [1]. Further movement along this direction of inquiry brings questions about the strength of association between datasets, with certain types of analyses probing directed relationships between variables or suggesting causal relationships between variables [48–51]. In the most typical case, this direction of inquiry concerns the relationship between brain activity and animal behavior, but this framing can also be used to investigate the relationship between neural dynamics in different parts of the brain [20,21,52], or across the brains of multiple individuals [6,53,54]. This approach is used throughout chapter 4 to quantify the correlation of neural activity across brains of interacting animals [53].

One commonly used set of statistical methods for quantifying relational structure is *regression analysis*. A regression between some variable Y (the *dependent/outcome* variable) and another variable X (the *independent*/*predictor* variable) quantifies the degree to which changes in X explain changes in Y. In settings with multiple variables (high dimensionality), multivariate regression can be used to quantify the association between a set of explanatory variables and one or more dependent variables. In the context of systems neuroscience, regression analysis is often used to measure the relationship between neural activity and external variables. In general, two classes of questions motivate this kind of analysis: 1) *What events in the external world are driving changes in brain activity?* 2) *What information about the external world can be extracted from brain activity?*

The first question relates to what are called *encoding models*, because they attempt to quantify the degree to which some external variable is "encoded" in brain activity. This type of analysis can yield insight into a neuron's physiology, its connectivity with other neural structures, and the types of computational processes it may participate in. Regression analysis is one method that can be used to quantify the association between neural activity and external variables and can thus be used as a type of encoding model. In the studies described in this dissertation (chapters 3 and 4), regression encoding models are used to examine how activity in neurons is associated with social interaction, features of the social environment, and behavioral decisions.

The second type of question relates to *decoding models*, which examine how much information about some measured variable is contained within a neuron or neural population's activity pattern. In a decoding model, some external stimulus or behavioral variable is modeled as a function of neural activity. Encoding models and decoding models are therefore two sides of the same coin, distinct in concept but closely related in their formulation. Regression, in one form or another, is also commonly used as a type of decoding model [17,34,35] as are some dimensionality reduction methods that are based on relational structure between variables [34,35,55].

2.2.3: Temporality

The third direction in this space of inquiry is temporality, which is concerned with how properties of variables, or the relationship between different types of data (relational structure), change over time. Analysis of temporal structure can be applied to any type of data that is measured across multiple time points. On one extreme, some measured signals may have properties that are unchanging over time or are changing so slowly that they can be treated as stationary. For example, a single neuron in the visual cortex may have very consistent orientation tuning over the course of a recording session. This relational structure – the relationship between the neuron's activity and a particular external variable – could thought of as stationary. In the other extreme, some processes may be highly dynamic [56,57]. Neurons may quickly change how they respond to sensory inputs during learning, for example, and animal behavior may evolve dramatically over the course of an interaction. In general, changes in the structure of measured data can take place over many different timescales and may occur continuously or discretely [56]. Some analytical methods treat properties under investigation as stationary in the sense that they do not explicitly account for temporal structure – two examples of this are PCA and simple linear regression. Other methods attempt to account for temporal structure by modeling changes in correlational structure across trials (for example, using Tensor Component Analysis [58]), or by modeling changes in relational structure (for example, using dynamic Generalized Linear Models [59]). In chapter 3, temporal structure in single neuron tuning to social stimuli is explored using regression across different epochs of a recording session [60].

2.3: Technical explanation of applied methods

2.3.1: Dimensionality reduction

Principal Component Analysis (PCA) is a method that seeks to identify components (linear combinations of the original variables) that explain (in a least-squares sense) the most variance in the data possible while maintaining an orthogonality constraint (forming an orthonormal basis) [61]. Because the principal components of a matrix X are orthogonal, they can be thought of as a "rotation" of the original coordinate system to form a new set of coordinates that best capture the most information in the data. The principal components can be found from an eigenvalue decomposition of the covariance matrix of X, Σ_X, where the normalized eigenvectors of $X^T X \propto \Sigma_X$ are the principal components of X, and are arranged in order of descending eigenvalues. In the studies described in this dissertation, PCA is used as a dimensionality reduction tool to visualize stimulus-evoked population responses and to analyze different patterns of population activity.

In a setting where we are interested in identifying components of the data that best capture variance in some other variable or variables of interest, different methods may be more appropriate. For example, *Fisher's Linear Discriminant* (FLD, also Linear Discriminant Analysis) provides a method for dimensionality reduction that maximizes the contrast in the data between samples belonging to two or more labeled classes. It is therefore a supervised multivariate method that, unlike PCA, is explicitly based on relational structure between datasets. Consider, for example, if we wanted to find the component (linear combination of neurons) of largest contrast between neural population activity evoked by two distinct types of stimuli. FLD identifies the component $\widehat{w}$ that, when used as a projection axis, maximizes the ratio of the between-class covariance $\Sigma_{between}$ to the within-class covariance Σ_{within}. Maximization of this ratio ($\frac{\Sigma_{between}}{\Sigma_{within}}$, the Rayleigh Quotient) can be thought of as maximizing the distance between the sample means for each class *per unit variance* in that direction. In the binary case, the optimal discriminant dimension is given by $\widehat{w} = {\Sigma_{within}}^{-1} * (\mu_2 - \mu_1)$, where μ_k denotes the mean over samples belonging to class k. FLD can be generalized to deal with more than two classes by optimizing the same ratio. In the multi-class case (where k > 2), the set of dimensions that maximize contrast between all classes is given by W, the set of eigenvectors that solve the generalized eigenvalue problem $\Sigma_{between}W = \Lambda\Sigma_{within}W$. The $k - 1$ eigenvectors associated with the $k - 1$ nonzero eigenvalues form a set of dimensions that maximizes projected between-class variance and minimizes within-class variance. In chapters 3 and 4, FLD is used primarily as a tool for decoding models, where the projection of population activity onto the first 1-2 dimensions is used to discriminate sensory cues and behavioral decisions of the animal.

Another method that can be used for dimensionality reduction, related to both PCA and FLD, is called partial least-squares (or projection to latent structures) regression (PLS). PLS is similar to FLD in that it is fundamentally based on relational structure between datasets, but instead of maximizing distances between classes (one dependent variable), it seeks to maximize covariance between a set of independent and dependent variables. In the studies in chapters 3 and 4, it is mostly used to visualize neural population data on components that contrast responses associated in time with different external events. It is also used as a processing step to prevent overfitting prior to construction of decoding models using FLD (PLS-FLD/PLS-LDA) [34].

2.3.2: Regression

Regression methods are commonly used to measure the association between different variables or sets of variables, and so fundamentally are tools to explore relational structure. In the setting of systems neuroscience, a typical question is to ask how well a neuron's activity can be explained by some set of external variables, such as an animal's sensory experience and/or behavioral decisions. To illustrate, consider that we want to use regression to model the response of a neuron (Y) as a function of one or more behavioral variables measured during the experiment – that is, we want to construct an encoding model $Y = f(X)$.

In the most basic form of regression, *Linear Regression*, the dependent variable Y is estimated as a scalar multiple of X plus some offset with gaussian error ($Y = X * \beta + \varepsilon$) [61]. The coefficient β is a parameter that is fit to maximize the variance explained in Y by X, or, equivalently, to minimize the sum of squared error between the true Y and our estimate of Y given X. In this univariate case, the coefficient β is given by the covariance between X and Y, normalized by the variance of X ($\beta = cov(X, Y)/var(X)$). It provides information about the strength and direction of association between the variables of interest.

In a multivariate setting with n independent variables (x_1, x_2, …, x_n), Y is estimated as a linear combination of the variables with gaussian error ($Y = x_1\beta_1 + x_2\beta_2 + \ldots + x_n\beta_n + \varepsilon$, or in matrix form, $Y = X\hat{\beta} + \varepsilon$). As in simple linear regression, the coefficient vector $\hat{\beta}$ is fit to minimize the total squared error between the model estimate and the true dependent variable Y. This least-squares optimization problem has an analytical solution given by the normal equation, $\hat{\beta} = (X^TX)^{-1}(X^TY)$ [61]. Intuitively, this can be thought of as the covariance of each predictor variable x_n with the response variable Y, scaled by the covariance structure of the predictor matrix X (the multivariate generalization of the solution for the univariate case). With a multiple regression model, one can ask questions about the relationship between a set of predictor variables X and the response variable Y. In general, one can interpret the coefficient β_n that is fit to predictor x_n as the expected change in Y with a unit change in x_n, so the direction and magnitude of each coefficient β_n gives information about the strength and direction of its association with Y, while also accounting for the relationships among the predictors.

A further augmentation of multiple linear regression, commonly used as an encoding model in neuroscience [27,57,59], is the *Generalized Linear Model* (GLM). The GLM models a response variable Y as a linear combination of predictor variables passed through some function g called the *link function* with an error term: $Y = g^{-1}(X\hat{\beta}) + \varepsilon$. The choice of the link function is typically based on an expectation of some particular relationship between X and Y, and/or an observation or expectation of non-gaussian distributed data and noise. Because single neuron activity time series are typically not gaussian distributed and spike generation can be idealized as a Poisson process, a log link function can be chosen to model single neuron activity using a GLM [57]. This gives the *Poisson Regression* model:

$$Y = e^{X\beta} + \varepsilon$$

In the setting of a decoding question, where we want to model some external variable as a function of neural activity, other link functions may be more appropriate. For example, for binary-valued response variables (like the presence or absence of a particular behavior), the logistic function can be used as a link function to convert the output of the linear term into a variable in the range [0 1] (interpreted as a likelihood) that can be thresholded to yield a binary output. This version of the GLM, also known as *Logistic Regression*, can be written as:

$$Y = \frac{e^{X\beta}}{1 + e^{X\beta}} + \varepsilon$$

2.3.2: Decoding models

In contrast to encoding models, which quantify how well neural activity can be explained by some set of predictors (usually external measured variables), decoding models quantify how well some measured variable (usually stimulus or behavior) can be explained by neural activity. Among the simplest decoding models are forms of dimensionality reduction and regression, described above. FLD, for example, can be used to find components of neural data that best capture the contrast across different samples of data that are labeled by association with some other variable. This method can therefore be used to identify a neural component (or set of components) that best discriminate between two or more types of events (for example, different types of animal behavior) [35,55]. The projection of the population data onto these components can then be used as a decoding model to quantify how well neural activity patterns can predict behavior. This approach is used to formulate decoding models of animal behavior in both chapters 3 and 4 [53,60].

Regression methods can also be used as decoding models. For example, if the variable to be decoded takes on binary values, then logistic regression (described above) may be a good choice for a decoding model. Other related methods, such as support-vector machines (SVM) [17] and Bayesian decoder models [62,63], can be useful under some circumstances. For example, methods such as PCA and the SVM can be easily adapted to capture non-linear relationships using kernel methods. In general, methods that can more flexibly capture complex non-linear relationships, such as deep neural networks, may provide greater predictive power at the cost of interpretability [64,65]. In order to retain the interpretation of events decoded from neural population activity as a biologically plausible linear readout between neural circuits, linear methods are used throughout chapters 3 and 4 for decoder analyses.

2.4: Manipulating brain activity

Analysis of the relationship between brain activity and behavior helps to refine hypotheses about the mechanistic logic of neural structures and circuits. However, observations based on measured data alone are essentially correlational. Even if some observations may suggest causal relationships between neural processes and behavior, controlled experimentation is required to establish this. In the most typical social neuroscience application, we would like to manipulate activity in the brain with as much precision as possible, guided by a hypothesis about the effects, and then measure the effect this manipulation has on animal behavior.

Brain activity can be manipulated using several different methods. Classic experiments using electrical current to stimulate activity in the brain have provided some of the initial insights about the role of specific regions in behavior that form the foundation of modern investigation [66,67]. This method offers very high temporal specificity, but suffers from relatively poor spatial precision, as electrical stimulation can affect local neurons as well as axonal fiber tracts that spread the stimulation nonspecifically to other brain regions. Pharmacological approaches allow researchers to infuse neuromodulator agents which can transiently alter neural activity in relatively localized regions of the brain. For example, muscimol (a GABA receptor agonist) infusion can be used to temporarily silence a population of neurons, and infusion of agonists or antagonists of neuromodulators (such as dopamine and 5-HT) can be used to test the effects of these signaling molecules in behaving animals. While these approaches can be very specific and informative for certain hypotheses, they are also relatively slow, and so obscure information about neuronal function on timescales faster than tens of seconds.

Over the past decade, optical approaches to manipulating neural activity have gained widespread application because they allow for a high degree of spatial and/or molecular specificity with high temporal specificity [23,68,69]. Optogenetics is based on manipulating light-activated ion channels, derived from photosensitive bacteria and virally introduced into neurons in vivo, that either trigger (photostimulation) or silence (photoinhibition) action potentials in response to light [70]. Using genetic and molecular methods, expression of these light-activated channels can be targeted to specific subpopulations of neurons based on their location in the brain, their connectivity with other brain regions, their expression of specific molecular markers, or their natural activity profile [23,71]. This degree of target specificity allows very precise manipulation of neuron activity and has allowed researchers to map functions of sensory processing, internal integration, and behavioral expression to highly specific circuit elements.

In the context of social neuroscience, we want to be able to manipulate the brain over a range of different timescales. While pharmacological methods may be useful to investigate how different hormones or neuromodulators affect behavior over longer timescales, optogenetics provides a way to manipulate circuitry involved in the decision-making process with sub second precision [72–74]. Because natural social interaction is not structured explicitly by the experimenter, interesting behavioral decisions may be made volitionally by animals in real time. Thus, precise manipulations of brain activity that can be introduced in response to or in anticipation of natural interaction and behavioral decisions are ideally suited to studying unstructured behavior. In addition, the use of activity-dependent promoters to control expression of ion channels allows manipulations to be targeted to populations of neurons based on their response properties in vivo [28,29,75,76] This permits mechanistic investigation of how specific neurons that are defined functionally (and possibly also anatomically) contribute to some behavioral or cognitive process. In chapter 3 of this dissertation, an activity-dependent optogenetics approach is used to manipulate neurons that encode specific social sensory cues, allowing us to test the causal role of native neural representations on behavior [60].

2.5: References

1. Jazayeri, M. & Afraz, A. Navigating the Neural Space in Search of the Neural Code. *Neuron* **93**, 1003–1014 (2017).

2. Chen, P. & Hong, W. Neural Circuit Mechanisms of Social Behavior. *Neuron* **98**, 16–30 (2018).

3. Stangl, M. *et al.* Boundary-anchored neural mechanisms of location-encoding for self and others. *Nature* **589**, 420–425 (2021).

4. Jamali, M. *et al.* Single-neuronal predictions of others' beliefs in humans. *Nature* **591**, 610–614 (2021).

5. King-Casas, B. *et al.* Getting to Know You: Reputation and Trust in a Two-Person Economic Exchange. *Science (80-.).* **308**, 78–83 (2005).

6. Montague, P. R. *et al.* Hyperscanning: simultaneous fMRI during linked social interactions. *Neuroimage* **16**, 1159–64 (2002).

7. Stephens, G. J., Silbert, L. J. & Hasson, U. Speaker-listener neural coupling underlies successful communication. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 14425–14430 (2010).

8. Sänger, J., Müller, V. & Lindenberger, U. Directionality in hyperbrain networks discriminates between leaders and followers in guitar duets. *Front. Hum. Neurosci.* **7**, 234 (2013).

9. Babiloni, F. *et al.* Hypermethods for EEG hyperscanning. in *2006 International Conference of the IEEE Engineering in Medicine and Biology Society* 3666–3669 (IEEE, 2006). doi:10.1109/IEMBS.2006.260754

10. Zhou, T. *et al.* History of winning remodels thalamo-PFC circuit to reinforce social

dominance. *Science (80-.).* **357**, 162–168 (2017).

11. Wang, F. *et al.* Bidirectional Control of Social Hierarchy by Synaptic Efficacy in Medial Prefrontal Cortex. *Science (80-.).* **334**, (2011).

12. Lee, E. *et al.* Enhanced Neuronal Activity in the Medial Prefrontal Cortex during Social Approach Behavior. *J. Neurosci.* **36**, 6926–6936 (2016).

13. Lin, D. *et al.* Functional identification of an aggression locus in the mouse hypothalamus. *Nature* **470**, 221–226 (2011).

14. Bergan, J. F., Ben-Shaul, Y. & Dulac, C. Sex-specific processing of social cues in the medial amygdala. *Elife* **3**, e02743 (2014).

15. Yao, S., Bergan, J., Lanjuin, A. & Dulac, C. Oxytocin signaling in the medial amygdala is required for sex discrimination of social cues. *Elife* **6**, (2017).

16. Mosher, C. P., Zimmerman, P. E. & Gothard, K. M. Neurons in the Monkey Amygdala Detect Eye Contact during Naturalistic Social Interactions. *Curr. Biol.* **24**, 2459–2464 (2014).

17. Munuera, J., Rigotti, M. & Salzman, C. D. Shared neural coding for social hierarchy and reward value in primate amygdala. *Nat. Neurosci.* **21**, 415–423 (2018).

18. Jun, J. J. *et al.* Fully integrated silicon probes for high-density recording of neural activity. *Nature* **551**, 232–236 (2017).

19. Juavinett, A. L., Bekheet, G. & Churchland, A. K. Chronically implanted neuropixels probes enable high-yield recordings in freely moving mice. *Elife* **8**, (2019).

20. Allen, W. E. *et al.* Thirst regulates motivated behavior through modulation of brainwide neural population dynamics. *Science (80-.).* **364**, eaav3932 (2019).

21. Stringer, C. *et al.* Spontaneous behaviors drive multidimensional, brainwide activity.

Science (80-.). **364**, 255 (2019).

22. Yang, W. & Yuste, R. In vivo imaging of neural activity. *Nature Methods* **14**, 349–359 (2017).

23. Luo, L., Callaway, E. M. & Svoboda, K. Genetic Dissection of Neural Circuits. *Neuron* **57**, 634–660 (2008).

24. Chen, T. W. *et al.* Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* **499**, 295–300 (2013).

25. Carrillo-Reid, L., Yang, W., Kang Miller, J., Peterka, D. S. & Yuste, R. Imaging and Optically Manipulating Neuronal Ensembles. *Annu. Rev. Biophys.* **46**, 271–293 (2017).

26. Kamigaki, T. & Dan, Y. Delay activity of specific prefrontal interneuron subtypes modulates memory-guided behavior. *Nat. Neurosci.* **20**, 854–863 (2017).

27. Pinto, L. & Dan, Y. Cell-Type-Specific Activity in Prefrontal Cortex during Goal-Directed Behavior. *Neuron* **87**, 437–450 (2015).

28. Jennings, J. H. *et al.* Interacting neural ensembles in orbitofrontal cortex for social and feeding behaviour. *Nature* **565**, 645–649 (2019).

29. Kim, C. K. *et al.* Molecular and Circuit-Dynamical Identification of Top-Down Neural Mechanisms for Restraint of Reward Seeking. *Cell* **170**, 1013-1027.e14 (2017).

30. Murugan, M. *et al.* Combined Social and Spatial Coding in a Descending Projection from the Prefrontal Cortex. *Cell* **171**, 1663-1677.e16 (2017).

31. Ghosh, K. K. *et al.* Miniaturized integration of a fluorescence microscope. *Nat. Methods* **8**, 871–8 (2011).

32. Resendez, S. L. & Stuber, G. D. In vivo calcium imaging to illuminate neurocircuit activity dynamics underlying naturalistic behavior. *Neuropsychopharmacology* **40**, 238–9 (2015).

33. Aharoni, D. & Hoogland, T. M. Circuit investigations with open-source miniaturized microscopes: Past, present and future. *Frontiers in Cellular Neuroscience* **13**, 141 (2019).

34. Remedios, R. *et al.* Social behaviour shapes hypothalamic neural ensemble representations of conspecific sex. *Nature* **550**, 388–392 (2017).

35. Li, Y. *et al.* Neuronal Representation of Social Information in the Medial Amygdala of Awake Behaving Mice. *Cell* **171**, 1176-1190.e17 (2017).

36. Zong, W. *et al.* Miniature two-photon microscopy for enlarged field-of-view, multi-plane and long-term brain imaging. *Nat. Methods* **18**, 46–49 (2021).

37. Zong, W. *et al.* Fast high-resolution miniature two-photon microscopy for brain imaging in freely behaving mice. *Nat. Methods* **14**, 713–719 (2017).

38. Kingsbury, L. & Hong, W. A Multi-Brain Framework for Social Interaction. *Trends Neurosci.* (2020). doi:10.1016/j.tins.2020.06.008

39. Kennedy, A. *et al.* Internal States and Behavioral Decision-Making: Toward an Integration of Emotion and Cognition. *Cold Spring Harb. Symp. Quant. Biol.* **79**, 199–210 (2014).

40. Anderson, D. J. & Perona, P. Toward a Science of Computational Ethology. *Neuron* **84**, 18–31 (2014).

41. Mathis, A. *et al.* DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nat. Neurosci.* **21**, 1281–1289 (2018).

42. Robie, A. A., Seagraves, K. M., Egnor, S. E. R. & Branson, K. Machine vision methods for analyzing social interactions. *J. Exp. Biol.* **220**, 25–34 (2017).

43. Hong, W. *et al.* Automated measurement of mouse social behaviors using depth sensing, video tracking, and machine learning. *Proc. Natl. Acad. Sci.* **112**, E5351–E5360 (2015).

44. Paninski, L. & Cunningham, J. P. Neural data science: accelerating the experiment-analysis-theory cycle in large-scale neuroscience. *Current Opinion in Neurobiology* **50**, 232–241 (2018).

45. Fusi, S., Miller, E. K. & Rigotti, M. Why neurons mix: high dimensionality for higher cognition. *Curr. Opin. Neurobiol.* **37**, 66–74 (2016).

46. Cunningham, J. P. & Yu, B. M. Dimensionality reduction for large-scale neural recordings. *Nature Neuroscience* **17**, 1500–1509 (2014).

47. Kobak, D. *et al.* Demixed principal component analysis of neural population data. *Elife* **5**, (2016).

48. Schippers, M. B., Roebroeck, A., Renken, R., Nanetti, L. & Keysers, C. Mapping the information flow from one brain to another during gestural communication. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 9388–9393 (2010).

49. Leong, V. *et al.* Speaker gaze increases information coupling between infant and adult brains. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 13290–13295 (2017).

50. Sänger, J., Müller, V. & Lindenberger, U. Directionality in hyperbrain networks discriminates between leaders and followers in guitar duets. *Front. Hum. Neurosci.* **7**, (2013).

51. Müller, V., Sänger, J. & Lindenberger, U. Hyperbrain network properties of guitarists playing in quartet. *Ann. N. Y. Acad. Sci.* **1423**, 198–210 (2018).

52. Schmitt, L. I. *et al.* Thalamic amplification of cortical connectivity sustains attentional control. *Nature* **545**, 219–223 (2017).

53. Kingsbury, L. *et al.* Correlated Neural Activity and Encoding of Behavior across Brains of

Socially Interacting Animals. *Cell* **178**, 429-446.e16 (2019).

54. Zhang, W. & Yartsev, M. M. Correlated Neural Activity across the Brains of Socially Interacting Bats. *Cell* **178**, 413-428.e22 (2019).

55. Grewe, B. F. *et al.* Neural ensemble dynamics underlying a long-term associative memory. *Nature* **543**, 670–675 (2017).

56. Rule, M. E., O'Leary, T. & Harvey, C. D. Causes and consequences of representational drift. *Current Opinion in Neurobiology* **58**, 141–147 (2019).

57. Driscoll, L. N., Pettit, N. L., Minderer, M., Chettih, S. N. & Harvey, C. D. Dynamic Reorganization of Neuronal Activity Patterns in Parietal Cortex. *Cell* **170**, 986-999.e16 (2017).

58. Williams, A. H. *et al.* Unsupervised Discovery of Demixed, Low-Dimensional Neural Dynamics across Multiple Timescales through Tensor Component Analysis-Dimensional Neural Dynamics across Multiple Timescales through Tensor Component Analysis. *Neuron* **98**, 1–17 (2018).

59. Roy, N. A., Bak, J. H., Akrami, A., Brody, C. D. & Pillow, J. W. Extracting the dynamics of behavior in sensory decision-making experiments. *Neuron* **109**, 597-610.e6 (2021).

60. Kingsbury, L. *et al.* Cortical Representations of Conspecific Sex Shape Social Behavior. *Neuron* (2020). doi:10.1016/j.neuron.2020.06.020

61. Strang, G. *Introduction to Linear Algebra*. (Wellesley-Cambridge Press, Wellesley, M, 2009).

62. Taxidis, J. *et al.* Differential Emergence and Stability of Sensory and Temporal Representations in Context-Specific Hippocampal Sequences. *Neuron* **108**, 984-998.e9

(2020).

63. Shuman, T. *et al.* Breakdown of spatial coding and interneuron synchronization in epileptic mice. *Nat. Neurosci.* **23**, 229–238 (2020).

64. Minderer, M., Brown, K. D. & Harvey, C. D. The Spatial Structure of Neural Encoding in Mouse Posterior Cortex during Navigation. *Neuron* **102**, 232-248.e11 (2019).

65. Saxe, A., Nelli, S. & Summerfield, C. If deep learning is the answer, what is the question? *Nature Reviews Neuroscience* **22**, 55–67 (2021).

66. Kruk, M. R. Hypothalamic attack: A wonderful artifact or a useful perspective on escalation and pathology in aggression? A viewpoint. *Curr. Top. Behav. Neurosci.* **17**, 143–188 (2014).

67. Hess, W. R. & Brügger, M. Das subkortikale Zentrum der affektiven Abwehrreaktion [The subcortical center for affective defense reactions]. *Helv. Physiol. Pharmacol. Acta* **1**, 33–52 (1943).

68. Yizhar, O., Fenno, L. E., Davidson, T. J., Mogri, M. & Deisseroth, K. Optogenetics in Neural Systems. *Neuron* **71**, 9–34 (2011).

69. Boyden, E. S., Zhang, F., Bamberg, E., Nagel, G. & Deisseroth, K. Millisecond-timescale, genetically targeted optical control of neural activity. *Nat. Neurosci.* **8**, 1263–1268 (2005).

70. Deisseroth, K. & Hegemann, P. The form and function of channelrhodopsin. *Science* **357**, (2017).

71. Luo, L., Callaway, E. M. & Svoboda, K. Genetic Dissection of Neural Circuits: A Decade of Progress. *Neuron* **98**, 256–281 (2018).

72. Yizhar, O. Optogenetic Insights into Social Behavior Function. *Biol. Psychiatry* **71**, 1075–1080 (2012).

73. Anderson, D. J. Optogenetics, Sex, and Violence in the Brain: Implications for Psychiatry. *Biol. Psychiatry* **71**, 1081–1089 (2012).

74. Riga, D. *et al.* Optogenetic dissection of medial prefrontal cortex circuitry. *Front. Syst. Neurosci.* **8**, 230 (2014).

75. Kawashima, T. *et al.* Functional labeling of neurons and their projections using the synthetic activity-dependent promoter E-SARE. *Nat. Methods* **10**, 889–895 (2013).

76. Ye, L. *et al.* Wiring and Molecular Features of Prefrontal Ensembles Representing Distinct Experiences. *Cell* **165**, 1776–1788 (2016).

CHAPTER THREE

Cortical Representations of Conspecific Sex Shape Social Behavior

3.1: Abstract

A central question related to virtually all social decisions is how animals integrate sex-specific cues from conspecifics. Using microendoscopic calcium imaging in mice, we find that sex information is represented in the dorsal medial prefrontal cortex (dmPFC) across excitatory and inhibitory neurons. These cells form a distributed code that differentiates the sex of conspecifics and is strengthened with social experience. While males and females both represent sex in the dmPFC, male mice show stronger encoding of female cues, and the relative strength of these sex representations predicts sex preference behavior. Using activity-dependent optogenetic manipulations of natively active ensembles, we further show that these specific representations modulate preference behavior toward males and females. Together, these results define a functional role for native representations of sex in shaping social behavior, and reveal a neural mechanism underlying male- vs. female-directed sociality.

3.2: Introduction

In order to navigate the social world, animals must integrate environmental and social cues from conspecifics to make decisions that secure their survival, health, and reproductive vitality. Representation and discrimination of conspecific sex—the recognition of another as male or female—is critical for social interaction, including behavioral decisions to preferentially engage with same- vs. opposite-sex conspecifics [1,2]. Previous work in mice has revealed that encoding of conspecific sex depends on chemosensory signaling and involves subcortical nodes including the amygdala and the hypothalamus [2–4], leading to a canonical view that processing of sex information depends primarily on subcortical circuits. However, how conspecific sex is represented beyond subcortical processing is poorly understood. More importantly, despite advances in identifying specific subcortical areas that encode conspecific sex, it is not clear whether representations at the level of natively active neural ensembles actually influence animal behavior. As neural representations of male and female may not precisely map onto molecularly separable cell types [5,6], direct manipulation of natively active neurons is necessary to understand how sex representations affect behavior. To date, this has not been explored and presents a critical gap in understanding how the brain shapes social decisions.

Recent work has illuminated the role of the medial prefrontal cortex (mPFC) in encoding general social information and in shaping social behavior such as dominance and sexual behavior [7–11]. As sex recognition is deeply embedded in these behavioral processes, such findings raise the question of whether cortical circuits play a role in sex recognition and the control of sex-typical behaviors that drive social interaction. In this study, we employ in vivo microendoscopic calcium imaging in freely behaving mice to explore the involvement of dorsal mPFC (dmPFC) ensembles in the encoding of conspecific sex and control of social behavior. We find that conspecific sex is represented in the dmPFC in a distributed population code that recruits both excitatory and inhibitory subpopulations. Male but not female mice show a bias toward encoding of female cues, and this relative strength of sex representation in males predicts their sex preference behavior. Finally, activation of these same ensembles modulates their preference towards male vs. female, defining a functional role for native sex representations in shaping behavior.

3.3: Materials and methods

3.3.1: Experimental model and subject details

All experiments were carried out in accordance with the NIH guidelines and approved by the UCLA institutional animal care and use committee (IACUC). Subject mice were male and female C57BL6/J and male Vgat-Cre mice ordered from Jackson Laboratories at 8-12 weeks of age and 24-30 g of weight. Mice were maintained in a 12 h:12h light/dark cycle (lighted hours: 10:00 pm – 10:00 am) with food and water *ad libitum*. All mice used for calcium imaging were individually housed for three weeks prior to experiments. Mice used for behavior experiments were individually housed for at least one week prior to experiments. All experiments were performed during the dark cycle of the animals.

3.3.2: Surgical procedures

3.3.2.1: Viral injections for imaging and GRIN lens implantations

Mice were anaesthetized with 1.0 to 2.0% isoflurane. Viral injections and lens implantation in the dorsomedial prefrontal cortex (dmPFC; also prelimbic cortex, PL) were done as previously described [7]. Specifically, we bilaterally injected 300 nl (on each side) of virus (AAV1.Syn.GCaMP6f.WPRE.SV40 for non-specific neuron imaging, AAV1-CaMKII-GCaMP6f-WPRE for CaMKII$^+$ cells, or 500 nl (on each side) of AAV1-Syn-FLEX-GCaMP6f for Vgat$^+$ cells) (Penn Vector Core) at 30 nl min^{-1} into the dmPFC using the stereotactic coordinates (AP: +2.0 mm, ML: ± 0.3mm, DV: −1.8mm to bregma skull surface). 1 week after injection, a 1.9mm diameter circular craniotomy was centered at the coordinates (AP: +2.0 mm, ML: 0.0 mm), and the GRIN lens (Edmund Optics; 1.8mm) was implanted above the injection site at a depth of -1.6mm ventral to the bregma skull surface and secured to the skull using super glue and dental cement. Mice were given one subcutaneous injection of Ketoprofen (4mg/kg) on the same day of surgery and Ibuprofen in drinking water (30mg/kg) starting on surgery day for 4 days. Mice were individually housed after surgery for three weeks. Then, the microscope together with a plastic baseplate were placed on top of the lens. We adjusted the position of the microscope until the cells and blood vessels appeared sharp in the focal plane and secured this position using dental cement.

The typical distance between the GRIN lens implant and the imaging focal plane is ~200 um. Although some tissue above the dmPFC must be removed in order to implant the lens, we have established in our previous work that the implant in the dmPFC does not disrupt normal social interaction [7]. We have also compared sex preference behavior in implanted animals with that of un-implanted animals (Figure 3.S4B) and found that implanted and un-implanted animals do not show significant differences in sex preference behavior.

For non-specific and CaMKII$^+$ cell imaging, we used wild type mice, and for GABAergic cell imaging, we used Vgat-Cre mice. All mice were handled and habituated to wearing the head-mounted microscope for at least 4 days before imaging experiments. While CaMKII promoter in AAV labels ~20% of inhibitory neurons in somatosensory cortex [12], 1-2% of CaMKII$^+$ cells express GABAergic cell markers in the mPFC [13]. This suggests that, while CaMKII$^+$ cells are predominantly excitatory in the prefrontal cortex, a very small fraction could be inhibitory.

3.3.2.2: Viral injections for activity-dependent labeling experiments

For activity-dependent optogenetics experiments, mice were bilaterally injected with 650 nL (each side) of a mixture of two viruses (AAV5-E-SARE-CreER [14] and AAV2-EF1a-DIO-hChR2-EYFP; ratio 1:5) at 40 nl min^{-1} into dmPFC using the stereotactic coordinates (AP: +2.0 mm, ML: ± 0.3mm, DV: −1.8mm to bregma skull surface). Optic fiber ferrules were implanted 4mm above the injection site (DV: −1.4mm to bregma skull surface). In order to create enough space for attachment of both optic fibers, the left ferrule was implanted at an angle 20 degrees from the midline DV axis, to the same depth of -1.4mm to bregma skull surface. Implants were secured to the skull using super glue and dental cement. For recovery, mice were given one subcutaneous injection of Ketoprofen (4mg/kg) on the same day of surgery and Ibuprofen in drinking water (30mg/kg) starting on surgery day for 4 days. Mice were individually housed after surgery. The same procedure was followed for the E-SARE validation experiment, except that 450 nL of a mixture of AAV5-E-SARE-CreER and AAV2-EF1a-DIO-EYFP (ratio 1:5) was used and that no fiber ferrules were implanted.

3.3.3: Histological analysis and E-SARE validation

After calcium imaging and optogenetics experiments were completed, mice were transcardially perfused with 4% paraformaldehyde (PFA). Brains were post-fixed in 4% PFA overnight at 4 °C and cryo-protected for 48-72 h in 30% sucrose at 4°C before freezing in OCT on dry ice. 50-μm coronal sections were obtained using a cryostat, and sections were stained with DAPI (1:5,000 dilution) and mounted on slides. Images were acquired using a fluorescence microscope (Invitrogen EVOS FL Auto 2 Imaging System or Leica DM6 automated microscope) to confirm the position of lens implantation and expression of GCaMP6f (Figure 3.1C) or E-SARE-CreER-driven ChR2-EYFP (Figures 3.S5-S6).

For analysis of overlap between male/female-induced E-SARE-CreER and male/female-induced Fos/Arc (Figures 3.S5C-E), coronal sections were obtained at 40 μm using a cryostat (Leica) at -20°C. Sections containing the dmPFC were washed in 1X PBS, blocked with 10% Normal Donkey Serum for 1 h at room temperature, and incubated with rabbit anti-Arc at 1:500 and rabbit anti-Fos at 1:500. The following day, sections were incubated with donkey anti-rabbit Alexa 568 antibody at 1:500. Images were acquired using a confocal microscope (Zeiss LSM 880). Overlap between EYFP$^+$ cells and Fos/Arc$^+$ cells was measured and quantified using CellProfiler 3.0 [15]. For each condition, ≥10 sections from four independently injected hemispheres were analyzed and quantified.

For analysis of neural projections in E-SARE-CreER labeled cells (Figure 3.S6), coronal sections were obtained at 35 μm using a cryostat (Leica) at -20°C. Sections were washed in 1X PBS, blocked with 10% normal donkey serum for 1 h at room temperature, and incubated in chicken anti-GFP antibody at 1:500 overnight. The following day, sections were incubated with donkey anti-chicken Alexa 488 antibody at 1:500. Images were acquired using a fluorescence microscope (Leica DM6 automated microscope). Using FIJI (ImageJ), fluorescence intensity was measured for dmPFC injection sites, as well as the following axon terminal sites: nucleus accumbens, dorsomedial striatum, dorsolateral striatum, basolateral amygdala, lateral hypothalamus, and ventral tegmental area. For each region of interest, area (square microns) and raw intensity were measured for six to eight hemisections. The total area and total raw intensity were computed as a sum of each value for all hemisections, and total raw intensity was normalized to total area. Finally, each axon terminal region was normalized to the dmPFC injection site.

3.3.4: Behavior assays

3.3.4.1: Home cage social investigation assay

For imaging during home cage social investigation, animals were outfitted with the head-mounted microscope and briefly habituated for 2-3 minutes in their home cage. Subject animals were then presented manually with 8 different male and female conspecifics, as well as a novel object (a toy mouse), for 20-30 second periods in which they were permitted to freely investigate the stimuli. For each session, there were a total of 8 presentations of male conspecifics (8 unique and novel animals), 8 presentations of female conspecifics (8 unique and novel animals), and 8 presentations of the toy. Conspecifics and the toy were presented in a pseudorandomized order such that no type of stimulus was presented twice in a row. Sessions typically lasted 12-15 minutes. In order to preserve naturalistic behavior of subject animals as much as possible, each bout of social interactions is allowed for around 20-30 s, as opposed to a strictly fixed window. In each bout, animals were permitted to initiate investigate events at will (they were not cued or forced to investigate stimulus animals) and ongoing investigation events were, as much as possible, not artificially ended by experimenters.

As each stimulus animal is presented for a much shorter period of time (20-30 s) compared to a typical resident-intruder assay, subject animals rarely display aggressive or mating behavior toward presented conspecifics; it usually takes a longer interaction time for these behaviors to occur. Because of these conditions, we did not observe aggressive or mating behavior during the 20-30 s presentations of stimulus animals. Therefore, while aggressive and mating behaviors are part of the natural behaviors directed towards different conspecific sexes, the encoding of male or female conspecifics is not due to the presence or absence of aggressive or mating behavior.

3.3.4.2: Neural responses to odor cues and juvenile animals

In order to examine how sex-encoding dmPFC neurons respond to sex-specific odor cues, we performed additional sessions of the social investigation assay (described above) that were immediately followed by presentation of soiled bedding material gathered from cages with novel male or female conspecifics (Figure 3.S1I). Soiled bedding was presented to subject animals for 20-30 s intervals during which subjects were permitted to freely investigate and sniff bedding material. Male and female bedding were alternately presented four separate times. All bedding investigation events were manually annotated.

We also explored how sex-encoding neurons respond during investigation of male and female juveniles (Figure 3.S1J). For these experiments, we performed the social investigation assay as described above, and immediately followed this with alternating presentations of 4 novel male juveniles and 4 novel female juveniles (~8 weeks of age). Juvenile investigation events were manually annotated using the same criteria as adult investigation events.

3.3.4.3: Two-chamber social exploration assay

For imaging in the two-chamber assay, animals were outfitted with the head-mounted microscope, briefly habituated in their home cage for 2-3 minutes, and then placed in a 24'' x 47'' two-chamber arena with novel male and female conspecifics placed in opposing corners underneath barred pencil cups (Figure 3.4A). The subject animals were then allowed to freely move about the environment and investigate social stimuli at will for the duration of the session.

For experiments described in Figures 3.4A-B, two 20-minute sessions were performed in immediate succession and the positions of the male and female conspecifics were swapped in each one. The bottoms of pencil cups were fitted with petri dishes in order prevent excrement and odor cues from contaminating the floor of the arena.

For experiments described in Figures 3.4C-H, two-chamber sessions were 30 minutes long and stimulus animals were not altered in any way throughout the experiment. These experiments immediately followed home cage social investigation sessions, and the microscope, LED power and fluorescence collection settings were not changed across the two sessions. For experiments presented in Figure 3.4H, datapoints from the first 20 mins were presented in Figure 3.4H (so that comparison between conditions is consistent with other analyses in Figure 3.4I-K). Datapoints from the full 30 mins session was presented in Figure 3.S3E.

For experiments described in Figures 3.4I-K, two-chamber sessions were 20 minutes long. Each animal underwent one recording session in which male and female conspecifics were placed beneath the cups, and one recording session in which non-social novel objects were placed beneath the cups, on separate days.

For behavior experiments using non-implanted animals (Figures 3.S4C-F), two-chamber sessions were 30 minutes long using male/female, male/empty cup, female/empty cup, or 2 empty cups. For experiments used to analyze the stability of sex preference over time, animals were first habituated to the social chamber on the day prior to experiments, and then underwent two consecutive days of two-chamber experiments using novel conspecifics on both days. In each session, animals were first habituated to the environment using empty cups for 10 minutes. Social stimuli were then introduced under the cups, and animals explored for 20 minutes.

3.3.5: Analysis of animal behavior

For the home cage social investigation assay, behavior videos were recorded with a video camera at 20 frames per second (fps) and manually annotated frame by frame to identify onset and offset times for individual investigation bouts. Investigation bouts were considered to be epochs of at least one second where the subject animal displayed directed interaction with the stimulus and its head was facing and in physical contact with the stimulus. Animals displayed similar frequencies of social vs. toy investigation but spent significantly more time investigating social stimuli per-bout than the toy (Figures 3.S1A-B).

For two-chamber experiments, behavior videos were recorded at 20 fps and the location and pose of the animal were automatically tracked using DeepLabCut (DLC) [16], a neural network framework that we trained using manually annotated video frames. For each frame, we extracted DLC output for the animal's nose, left ear, right ear, and tail coordinates. We considered social investigation events to be periods lasting at least 1 second where the animal's head was within 3 inches of the center of the cup, and the angle between its head and the vector between its head and the cup was < 90 degrees. Behavior annotations were converted into binary vectors that denote precisely when animals are engaged in social investigation for downstream analysis.

In all behavior experiments, we measured the bias between male vs. female interaction using a sex preference score, computed as the fractional difference in male vs. female investigation time (T♂-T♀) / (T♂+T♀). In the home cage social investigation assay, male and female investigation time were each normalized by the total amount of time male and female conspecifics were presented in the session. While we observed a modest female preference in our behavior experiments, consistent with previous reports, there was large variability in preference scores in individual animals that was greater than expected just from random movement (Figure 3.S4C). Sex preference scores of individual animals were also consistent across two days of testing (Figure 3.S4E) with different stimulus animals, indicating stability of the behavioral state. We observed no difference in sex preference behavior in animals wearing the microendoscope and animals without any implant (Figure 3.S4B).

In order to examine whether sex preference behavior in males was altered by sexual experience, we measured sex preference in the two-chamber assay on two consecutive days, between which subjects had two 30-minute exposures to novel female conspecifics in their home cage. During these female exposure sessions, mice were freely permitted to engage in mating behavior with females. We found that sex preference scores were not significantly different between sessions, suggesting that sexual experience does not alter preference behavior (Figure 3.S4F).

For Figure 3.1H, both male and female animals were used. For Figures 3.S1C-D and 3.S1H, female animals were used. For all other figures, male animals were used.

3.3.6: Activity-dependent labeling and stimulation of male and female cells

For optogenetics experiments, animals were first virally injected with Tamoxifen-inducible E-SARE-CreER and Cre-dependent ChR2 constructs (described above). 3 weeks after injection, animals underwent a social exposure paradigm in which they interacted with either male or female conspecifics, or they remained in their home cage with no social stimulus. On each of two consecutive days, animals were presented sequentially with 6 novel male or female intruders in their home cage and allowed 10 minutes of interaction time with each one over a total period of 1 hour. For home cage controls, animals were left in their home cage for one hour, and an experimenter briefly placed their hand in the cage once every 10 minutes to control for effects related to human stimulus. At the end of the social exposure session on each day, animals were injected with 20mg/kg of 4OH-tamoxifen (4-TM). Prior to social exposure and E-SARE::ChR2 induction, animals were habituated to the induction room for 1 week, handled for 5 days, and habituated to IP injection with daily saline injections for 4 days.

In order to validate the efficacy and specificity of the activity-dependent labeling strategy, we followed a previously established procedure [14] by measuring the overlap between E-SARE labeled cells and cells that express immediate early genes Fos and Arc following a second exposure (Figures 3.S5C-E). Animals first underwent the E-SARE-CreER induction as described above on two consecutive days. On each day, animals were exposed to either male conspecifics, female conspecifics, or a home cage control over a period of one hour, immediately followed by 4-TM injection (20mg/kg). This was repeated on two consecutive days (one exposure and one injection on each day). Cre-dependent EYFP was used in place of ChR2 to visualize individual cell bodies. Three weeks later, each animal underwent a second exposure with novel male conspecifics or female conspecifics. For the second exposure, four novel animals were presented for 15 min each for a total of one hour. One hour following this second exposure, animals were anaesthetized with isoflurane and were transcardially perfused with 4% paraformaldehyde (PFA). Brains were sectioned to perform histological analysis of E-SARE::EYFP and expression of Fos and Arc.

In order to test whether neural populations active during interaction with male or female conspecifics control the sex preference of animals, we optically stimulated male and female cells during the social two-chamber assay. After habituation to the two-chamber environment for at least one day with empty cups, animals were placed in the same arena with novel male and female conspecifics in either corner, as described above for imaging experiments. After a 10-minute baseline period, dmPFC neurons were stimulated for 10 minutes using a 473 nm laser (4 mW/mm^2). The stimulation applied a protocol of repeated sequences of 3 seconds light-on and 2 seconds light-off, at 20hz with 20 ms light pulses. This stimulation epoch was then followed by another 10-minute baseline period. These experiments were performed 4-5 weeks following social exposure and 4-TM injection to allow time for sufficient ChR2 expression.

Binary vectors denoting male and female investigation were automatically extracted using DeepLabCut [16] (described above). These were then smoothed using a 10-minute moving average window, and (T♂−T♀) / (T♂+T♀) was computed to obtain time courses for sex-preference behavior throughout the experiment. Male and female investigation time was computed and compared across 10-minute windows corresponding to the baseline and stimulation epochs. For each cohort, we performed sham control sessions in which animals were attached to the optic fiber, but no light was delivered (Figures 3.S5I-K). Two sham sessions were performed on different days and behavior for each animal was averaged across sham sessions.

While optogenetic stimulation may have an effect over a timescale of seconds, this temporal resolution is constrained by the experimental assays used to measure this effect. Also, optogenetic stimulation may or may not result in acute changes in behavior (e.g. activating the neurons when the animals are facing away from stimulus animals may not elicit a strong effect). Thus, due to the nature of the behavioral assay and its natural variability, we have to measure this over a period of time. This has been the typical way to operationalize and analyze sex preference behavior in the established literature [17,18].

For experiments using male-exposed animals (stimulating male cells), n = 14 animals, for experiments using female-exposed animals (stimulating female cells), n = 8 animals, and for home cage non-social control experiments, n = 9 animals.

3.3.7: Extraction of calcium signals

3.3.7.1: Motion-correction and preprocessing

During behavior experiments, calcium fluorescence videos were recorded through customized miniature microscopes (UCLA miniscope) at 30Hz using custom-written data acquisition software as previously described [7]. Raw videos from each imaging session were processed using a MATLAB implementation of the NoRMCorre algorithm to correct for motion-induced artifacts across frames [19]. In order to normalize image frames prior to cell sorting, $(F-F0)/F0$ ($\Delta F/F$) was applied to each frame, where $F0$ was the de-trended mean frame. $\Delta F/F$ normalized videos were de-noised using an FFT spatial band-pass filter in ImageJ (v1.52a, U.S. National Institutes of Health), and spatially down sampled by a factor of 2 prior to ROI identification and cell sorting.

For comparison of cell encoding across the home cage and two-chamber assays (Figures 3.4C-H), we performed these two experiments in immediate succession without removing the head-mounted microscope or changing the recording settings (Figure 3.S2F), such that identical cells in the same imaging view were recorded across two assays. Calcium fluorescence videos from these two sessions were concatenated and preprocessed together, and single cell ROIs were segmented (described below) across both experiments. These procedures ensured that cells in both sessions were precisely aligned, allowing analysis of the exact same cells across sessions.

3.3.7.2: Segmentation and ROI identification

We identified putative cell bodies for extraction of neural signals using an established pipeline as previously described and validated [7]. Specifically, we employed an automated ROI detection algorithm that uses principal (PCA) and independent component analysis (ICA) to extract spatial filters based on spatiotemporal correlations among pixels [20]. Independent components were manually inspected to remove components that did not represent cell bodies, and binary thresholding was applied to remove contributions from pixels outside the bounds of putative neurons. Spatial filters were then applied to the *ΔF/F* movie to extract the calcium traces. All traces from recorded cells were manually inspected to ensure quality signals. Specifically, putative neurons that had abnormally shaped cell bodies (abnormally large or small), or that had calcium transients with low signal-to-noise were excluded from further analysis (<5% of all putative neurons were excluded in this manner).

For home cage social investigation imaging experiments in Figure 3.1, a total of 5829 (mean ± SEM = 211 ± 9) single neurons were analyzed (n = 23 males and 5 females). For imaging of GABAergic (Vgat$^+$) neurons, a total of 366 (mean ± SEM = 61 ± 4) neurons were analyzed (9.6% ± 0.3 of all cells, n = 6 males). For imaging of glutamatergic (CaMKII$^+$) neurons, a total of 1373 (mean ± SEM = 229 ± 7) neurons were analyzed (25.4% ± 2.6 of all cells, n = 6 males).

For two-chamber imaging experiments, a total of 6686 (mean ± SEM = 216 ± 6) neurons were analyzed (n = 14 males). For Figures 3.4A-B, n = 7 males. For figures 3.4C-H, n = 10 males. For Figures 3.4I-K, n = 7 males.

For all experiments, a single neuron refers to one calcium trace extracted from an ROI, identified as described above, from one recording session.

3.3.8: Analysis of single cell responses during social interaction

Prior to downstream analysis, all *ΔF/F* calcium traces were z-scored and are presented throughout in units of standard deviation (σ). Responses of single neurons during social interaction events were quantified using ROC (receiver operating characteristic) analysis as previously described [4,7]. Upon application of a binary threshold to the *ΔF/F* signal and comparison with a binary event vector denoting behavior, event detection based on neural activity can be measured using the true positive rate (TPR) and the false positive rate (FPR) over all time-points. Plotting the TPR against the FPR over a range of binary thresholds, spanning the minimum and maximum values of the neural signal, yields an ROC curve that describes how well the neural signal detects behavior events at each threshold (Figure 3.1F). We used the area under the ROC curve (auROC) as a metric for how strongly neurons were modulated during social interaction. For each neuron/behavior category, the observed auROC was compared to a null distribution of 1,000 auROC values generated from constructing ROC curves over randomly permuted calcium signals (traces that were circularly permuted using a random time shift). A neuron was considered significantly responsive ($\alpha = 0.05$) if its auROC value exceeded the 95th percentile of the random distribution (auROC < 2.5th percentile for inhibited responses, auROC > 97.5th percentile for excited responses).

We analyzed the difference in stimulus-evoked activity in dmPFC neurons over the course of the two-chamber session (Figure 3.4H-I, 3.S3E). We normalized differential activity in each epoch by the average activity of that cell. This normalization ensures that changes in differential activity are not due to overall changes in spontaneous activity, but rather reflect changes in the relative responsiveness of neurons to male and female stimuli.

Because some behavior events occur at different frequencies, there may be differences in the effective statistical power for the bootstrap ROC analysis between different event types (e.g. male and female investigation). Although this is unlikely to substantially affect measures of significantly responsive neurons, we performed additional control analyses to confirm that differences in fractions of responsive neurons were not due to unequal class sampling (Figure 3.S2A-D). To this end, we normalized male/female representation by uniformly rescaling each male or female investigation bout for each session (using nearest-neighbor interpolation) by scaling factors $\boldsymbol{s}_{♂} = mean(T_{♂}, T_{♀})/T_{♂}$ and $\boldsymbol{s}_{♀} = mean(T_{♂}, T_{♀})/T_{♀}$ where $T_{♂}$ and $T_{♀}$ are the fractions of total time spent investigating male and female conspecifics. Corresponding epochs of the neural traces were equivalently rescaled. Following this, the total number of frames within a session that correspond to male and female investigation are equalized. The ROC bootstrap analysis was then performed again using these rescaled bouts and calcium traces, and fractions of significant cells for each category were computed. The differences in male- and female-responsive cells observed across all animals and in female-preferring animals were robust in this control analysis (Figures 3.S2A-D, 3.S4G-H).

In order to examine the coding consistency of sex-specific neurons across social investigation and two-chamber assays, we compared their stimulus-evoked activity in response to male vs. female stimuli in both experiments. We found that 72.2% of male-excited cells, 66.5% of female-cells, 64.7% of male-inhibited cells, 72.0% of female-inhibited cells ($p < 0.001$, bootstrap test) showed consistent increased or decreased activity in response to social stimulus investigation.

3.3.9: Cell tuning analysis suing models of single neuron activity

In order to analyze how single dmPFC neurons change their tuning to male and female interaction over time, we used gaussian generalized linear models (GLM) to model calcium activity in individual neurons as a function of several factors (Figure 3.S3F-K). This approach discounts contributions to neural activity that may be explained by other variables such as approach behavior or speed, so that weights fit to social interaction provide a measure of individual neural tuning to social stimuli. We used five variables to model cell activity: male and female interaction events, social approach, the animal's speed, and the decay of overall neural population activity over time. Binary-valued vectors denoting interaction and approach events were first smoothed using an exponential filter (tau = 3 sec), and the animal's speed was smoothed using a 3 sec moving average. For experiments with objects instead of conspecifics in the corners, behavior vectors denoted interaction with the right or left object, defined using the same criteria as described above for male and female interaction. GLM weights were fit using MATLAB (glmfit), and weights corresponding to male/female or object investigation were averaged across stimulus excited cells (defined using ROC analysis). Separate models were constructed for different epochs of the two-chamber session.

3.3.10: Analysis of population dynamics during behavior

3.3.10.1: Principal Component Analysis

To visualize population responses during social interaction (Figure 3.2E), we applied principal component analysis (PCA) to obtain components that capture the covariance of the neural population during interaction events [21]. Calcium traces were first smoothed using a 5-second moving window average. Trial-averaged responses were computed over a time window of 20 seconds after event onset for male and female investigation events and concatenated, and responses for each neuron were formed into a matrix to perform PCA. Population trajectories for individual investigation bouts were then projected onto the first 3 principal components for visualization (Figure 3.3.2E, one example animal), and trial-averaged responses were also projected in thick lines.

3.3.10.2: Strength of population responses

In order to measure population responses associated with male and female investigation (Figure 3.5A-D), we used the Mahalanobis distance, which provides a measure of the separation between two population vectors while accounting for the covariance structure of the underlying distribution. The Mahalanobis distance between two population vectors was computed as:

$$\mathrm{D}_{\mathrm{MAH}}(\mathrm{x},\mathrm{b}) = \sqrt{(\mathrm{x}-\mathrm{b})^{\mathrm{T}}\mathrm{S}^{-1}(\mathrm{x}-\mathrm{b})}$$

where **x** is a population response vector at some timepoint, **b** is the average population vector over all baseline frames (where the animal does not exhibit social investigation behavior), and ***S*** is the covariance matrix computed over all baseline frames. For population response time-courses (Figures 3.5A, C), the Mahalanobis distance for individual investigation bouts was computed over a window of 10 seconds prior to 30 seconds after event onset. Here, we compared the average population responses evoked by male and female investigation and compared these across the top 25% male-preferring and top 25% female-preferring animals (Figures 3.5B, D). For each animal, a population preference score was defined as the fractional difference in average male vs. female response vectors, (R♂-R♀) / (R♂+R♀).

3.3.10.3: Male vs. female population decoder analysis

In order to measure population-level encoding of conspecific sex by dmPFC neurons, we constructed statistical models to predict the sex identity (male vs. female) of events based on population activity. For this, we used binary Fisher's linear discriminant (FLD) classifiers.

For all classifiers used to quantify sex discrimination within the home cage and two-chamber sessions, training sets and test sets were constructed using population vectors evoked during male and female investigation events. Classifiers were constructed separately for each animal and used neurons only from that animal. We used 10-fold cross validation to measure classifier performance. For each cross-validation fold, the test set was a continuous 10% epoch of the data, and the remaining 90% training set was used to construct the model. For each fold, partial least-squares regression (PLS) was used to reduce the dimensionality of the training data, and the top 15 PLS components were retained for FLD analysis. Model performance was measured using the area under the ROC curve (auROC) for test data projected onto the Fisher discriminant. Overall model performance for each animal was calculated as the average over 50 folds where the training and test sets were randomly redrawn, and folds where both male and female events were not represented in the training and test set were dropped. Models were compared with null models constructed using data with randomly permuted calcium traces.

We found that decoders performed well above chance levels when constructed using a mixture of training data from both experiments (continuous 90% epochs) and tested on events from both (Figure 3.4F). For cross-session decoders (Figures 3.S3A-B), the analysis was performed as described above, except the training set comprised male and female investigation events from the entire home cage or two-chamber session, and the test set those from the entire other session. For each animal, the neural populations for both cross-session decoders were down-sampled to contain only the population of cells that were significantly responsive to social stimuli based on ROC analysis. Performance may be slightly higher in decoders trained on two-chamber data and tested using home cage data because more training data is available from the longer two-chamber session. Figure 3.4G shows the projection of mean population vectors, from one example animal, associated with all investigation bouts from both home cage and two-chamber experiments onto FLD components that are defined using data from only the two-chamber session.

For the analyses in Figures 3.3L-M, where the performance of sex discrimination decoders using CaMKII$^+$ neurons is compared with that of GABAergic neurons, neural populations were randomly down-sampled on each cross-validation fold to a subset of N neurons (on the x-axis in Figure 3.3L). 500 iterations of down sampling and cross validation were performed for each ensemble size. Partial least-squares regression for dimensionality reduction was not used for this analysis, as some cell subsets smaller than 15 are sampled (the number of PLS components used in other decoder analyses).

3.3.11: Quantification and statistical analysis

All statistical analyses for this study were conducted using GraphPad Prism (v8.4.0) or custom routines in MATLAB (Mathworks) and are described in the respective figure legends and Methods. All bar plots with error bars represent mean ± SEM; all box and whisker plots represent the median, interquartile range (box), and min and max (whiskers) of the underlying distribution. Statistical significance was defined with $\alpha < 0.05$ using two-tailed tests unless otherwise specified. Full statistics for all ANOVA analyses are reported in Table S1. For comparisons of cell pairwise distance distributions, two-sample Kolmogorov-Smirnov tests were used. Resampling methods based on temporally permuting calcium traces (described in Methods) were used to assess significance of auROC values for social modulation of neural signals and performance of FLD decoders.

3.4: Results

3.4.1: dmPFC neurons encode conspecific sex

In order to first explore whether dmPFC neurons encode male- and female-specific information, we performed *in vivo* microendoscopic calcium imaging [22] of dmPFC neurons during natural investigation of male and female conspecifics (**Figure 3.1A**). To optically record from dmPFC neurons, we injected an adeno-associated virus (AAV) expressing the fluorescent calcium indicator GCaMP6f [23] and implanted a gradient refractive index (GRIN) lens above the dmPFC (**Figure 3.1B**). Expression of GCaMP6f and placement of the lens were confirmed histologically (**Figure 3.1C**). During imaging, we presented each subject animal with 8 novel males and 8 novel females interleaved with novel objects. Each stimulus animal was presented for 20-30 s, during which the subject animal was free to investigate. After imaging sessions, activity signals associated with single cells were extracted using independent component analysis as previously described [7,20] and are reported throughout as relative change in fluorescence ($\Delta F/F$) (**Figures 3.1D-E**). We recorded a total of 5897 dmPFC neurons across 28 animals, including both males and females.

In order to determine whether single neurons selectively encode sex-specific information, we computed a receiver operating characteristic (ROC) curve for each neuron/stimulus (male/female/toy) relationship, which quantifies its stimulus detection strength over a range of binary thresholds (**Figure 3.1F**). Using this approach, we identified a substantial fraction (22%) of dmPFC neurons that showed significant responses during investigation of social stimuli (**Figure 3.1G, H**). Many neurons were selectively tuned to either male or female investigation (**Figures 3.1I-J, 3.S1C-D**). Interestingly, in male animals, a larger fraction of neurons showed increased activity during interaction with female compared to male conspecifics (**Figure 3.1I**), suggesting a bias toward encoding of female cues. However, when we examined this in female mice, we found that comparable fractions of neurons responded to male and female conspecifics (**Figure 3.S1C-D**), suggesting that stronger encoding of female cues is specific to the male dmPFC.

This intriguing observation raises the questions of how female and male conspecifics are differentially encoded in the male brain and how this representation is linked to behavior. Although single cells encoding sex-specific information preferentially responded to male or female cues, male and female cells had overall similar response amplitude and reliability (**Figure 3.S1E-G**), suggesting that their basic response properties are not significantly different. Moreover, male- and female-encoding cells were not selectively tuned to investigation of sex-specific odor cues (**Figure 3.S1I**), suggesting that sex representation is unlikely solely due to olfactory inputs. Lastly, to examine whether these neurons were intermingled or separately clustered within the dmPFC, we analyzed the spatial distributions of sex-encoding cells. We found no differences in the distributions of sex-encoding compared to non-encoding cells (**Figures 3.1K-L**), indicating that sex-encoding cells are not spatially organized. Collectively, these results show that single neurons in the dmPFC encode the multisensory variable of sex, and in males, form a stronger representation of female compared to male cues.

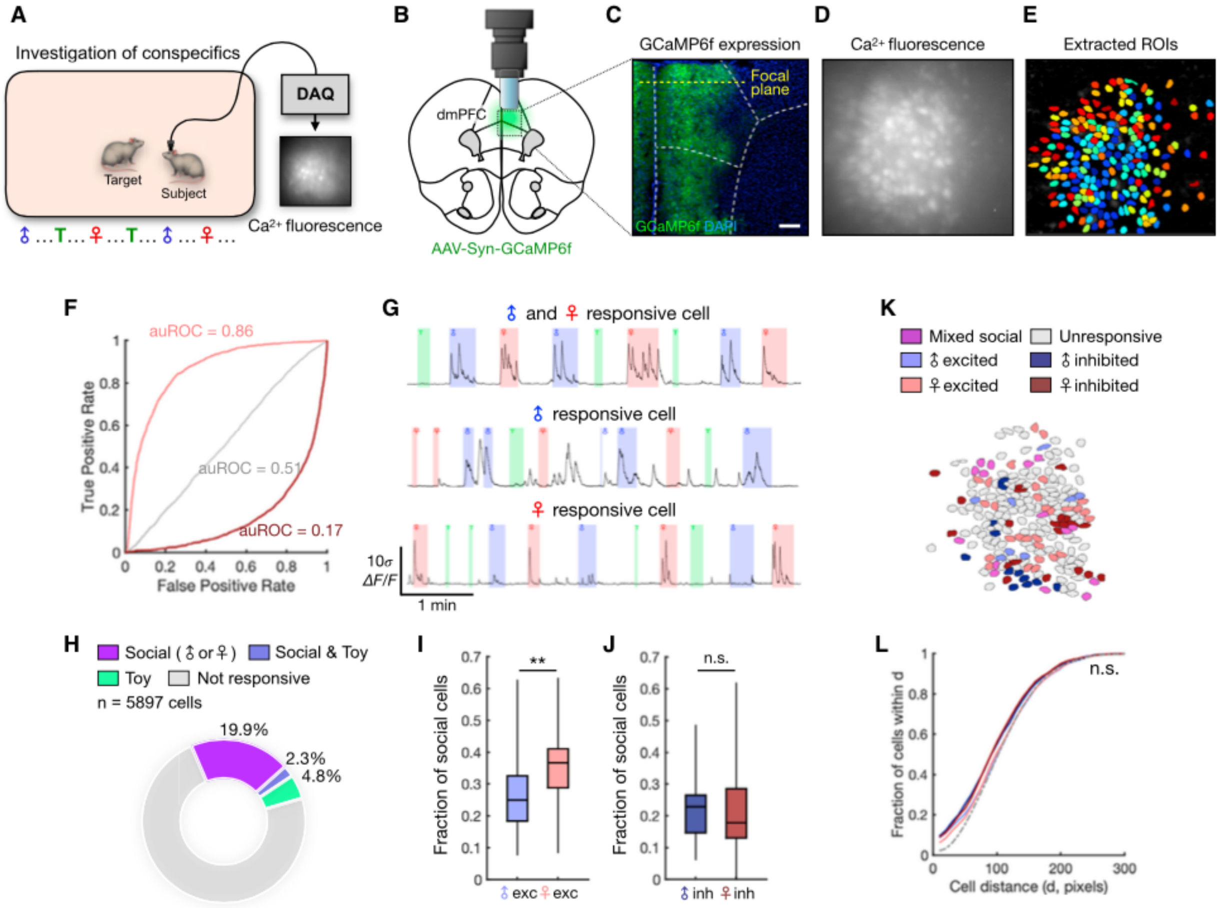

***Figure 3.1: dmPFC neurons encode conspecific sex during natural social interaction*.** **(A)** Schematic of social investigation assay. Eight novel male and eight novel female conspecifics are presented interleaved with a novel object, and the subject animal is allowed to freely investigate. **(B)** Illustration of the microendoscope placed above the dmPFC. **(C)** Example image showing expression of GCaMP6f in neuronal cell bodies. The location of the focal plane is estimated based on the estimated working distance of the lens. **(D)** Imaging field of view showing raw calcium fluorescence from one animal (max projection). **(E)** ROIs corresponding to single neurons extracted from the field of view in (D). **(F)** Receiver operating characteristic (ROC) curves computed from three example neurons that are female excited (auROC = 0.86), female suppressed (auROC = 0.17), or not responsive to female (auROC = 0.51). **(G)** Example calcium traces from social-mixed (top), male (middle), and female (bottom) neurons. **(H)** Distribution of social cells (responding to either male or female interaction) and toy-responsive cells from all recorded animals. **(I)** Fractions of male- and female-excited cells among socially responsive dmPFC neurons recorded from males (p = 0.0071). Mixed cells responding to more than one category constituted 6.29% ± 1.18% (mean ± SEM) of socially responsive cells. We also observed a higher fraction of female-excited cells when normalizing the sampling of male and female investigation. **(J)** Fractions of male- and female-inhibited neurons recorded from male animals (p = 0.59). **(K)** Example field of view showing spatial locations of male- and female-responsive cells. **(L)** Cumulative histogram showing the mean fraction of cells within a given pairwise distance (x axis), compared between subsets of functionally defined neurons. In (H), n = 28 animals (including males and females); (I and J) Mann-Whitney U test, n = 23 animals; (L) two-sample Kolmogorov-Smirnov test, n = 23 animals. **p < 0.01, n.s., not significant. Scale bar, 200 mm.

3.4.2: Population representations in the dmPFC uniquely encode conspecific sex

Single neurons in the dmPFC formed largely non-overlapping subpopulations that responded specifically to male or female investigation (**Figures 3.2A-D, 3.S1K**), suggesting that male and female interaction may elicit unique and separable trajectories of population activity. In order to explore this, we projected single-trial population dynamics during investigation bouts onto principal components. This revealed a clear separation of single-trial responses associated with male vs. female interaction (**Figure 3.2E-F**), indicating that the population may discriminate sex information. Still, while this visualization suggests separable patterns of activity, sex representations may not be temporally stable or general enough to decode sex across time and across novel animals. In order to test whether sex representations are stable and general, we constructed linear discriminant decoders to predict the sex identity of novel animals using population activity. The cross-validated performance of these decoders was well above chance levels (**Figure 3.2G**), indicating that unique and stable response patterns in the dmPFC encode the sex of conspecifics. Importantly, the fact that sex could be decoded across novel animals (**Figure 3.2G**) suggests that different animals of the same sex evoked consistent activity patterns, indicating a general representation of sex itself. Consistent with male mice having more female- than male-responsive cells (**Figure 3.1I**), we also found that male animals showed higher amplitude population responses during female investigation compared to male investigation (**Figure 3.2F**), whereas females did not show this difference (**Figure 3.S1H**). Together, these results suggest that dmPFC neurons encode sex stably at the population level, and that in males, this distributed representation exhibits a bias toward encoding of female cues.

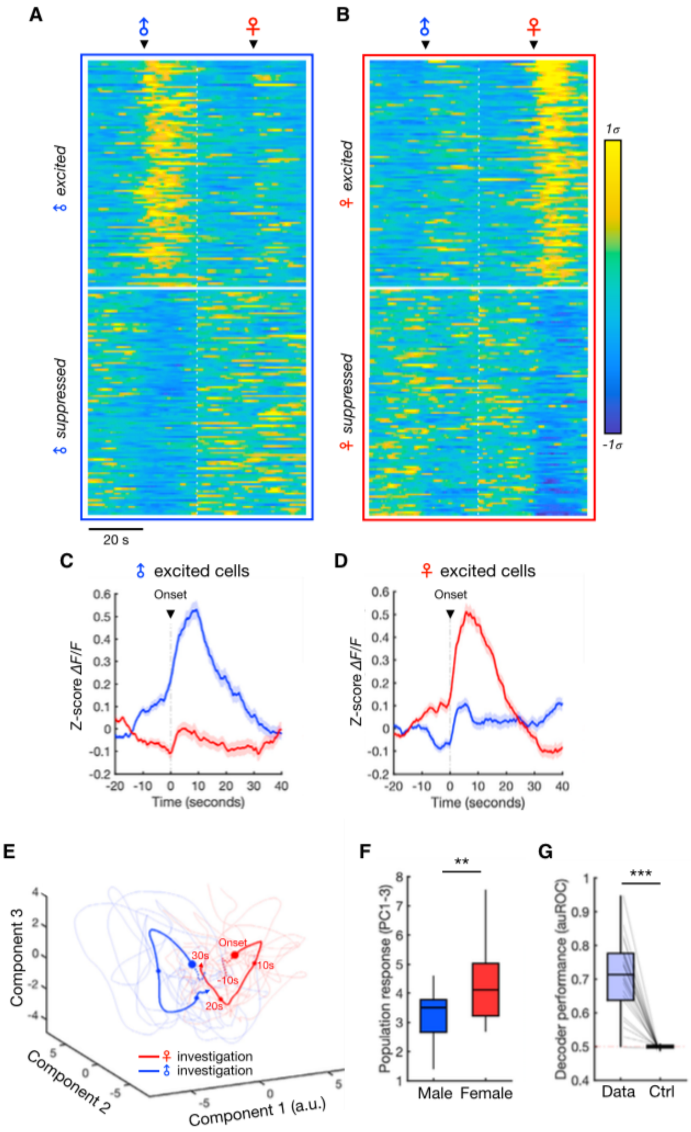

Figure 3.2: Population representations of male and female**. (A and B)** Average responses of male-responsive (A) and female-responsive (B) neurons centered around investigation onset. The top 100 neurons, sorted using rank-ordered auROC values, are shown for each cell type (showing excited or in- hibited responses). **(C and D)** Average responses of all male-excited (C) and female-excited (D) neurons during male and female investigation (mean ± SEM). **(E)** Principal component projection of population dynamics during individual male and female inves- tigation bouts (thin lines) and average responses (thick lines) from one example animal. Dots indicate 10 s time intervals, and dotted lines indicate time before investigation onset. **(F)** Comparison of the population response ampli- tude during male versus female interaction in male animals. Population responses (projected to PC1–3) are measured over the first 10 s of interaction and averaged across bouts within each animal (Mann- Whitney U test, $p = 0.0095$, $n = 23$ animals). **(G)** Performance of Fisher's linear discriminant (FLD) decoders trained to classify male versus female investigation using population dynamics. Cross- validated auROC is compared with performance of null models trained and tested using time-permuted calcium traces (Wilcoxon signed-rank test, $p = 2.7e5$, $n = 23$ animals). ***$p < 0.001$, **$p < 0.001$.

3.4.3: Encoding of conspecific sex across distinct subpopulations

Next, we sought to gain deeper insight into how sex representations are distributed across different dmPFC neuron cell types. While encoding of conspecific sex may be distributed broadly, another possibility is that specific neuronal populations may preferentially encode sex-specific information. To this end, we explored how excitatory and inhibitory subpopulations of dmPFC neurons contribute to the encoding of conspecific sex. We injected an AAV expressing GCaMP6f in either CaMKII$^+$ (predominantly excitatory, **Figure 3.3A**) or Vgat$^+$ (GABAergic, **Figure 3.3B**) neurons and recorded neural activity during the social investigation assay. Overall, we recorded from a total of 1373 CaMKII$^+$ neurons and 336 Vgat$^+$ neurons.

Analysis of single neuron responses using ROC curves showed that both populations contained a substantial fraction of cells encoding social cues (**Figures 3.3C-D**). Consistent with our results above, we found that there were more female- than male-encoding cells in both subpopulations in males (**Figures 3.3E-F, 3.S2C-D**), indicating that the bias in encoding of female cues is preserved across both excitatory and inhibitory cell types. We also found that social cells in each subpopulation were similarly spatially intermixed (**Figure 3.3G**), indicating that sex information is distributed widely and not clustered within either subtype. Interestingly however, despite similarities in spatial distributions and encoding of social information, a higher fraction of Vgat$^+$ compared to CaMKII$^+$ neurons were specifically responsive to conspecific sex (**Figure 3.3H**), suggesting that sex information may be more enriched in GABAergic neurons. This enrichment of sex encoding in GABAergic neurons was preserved across different response categories, and significant among male- and female-excited cells (**Figures 3.3I, 3.S2E**). Lastly, we analyzed how information at the single neuron level in CaMKII$^+$ and Vgat$^+$ ensembles could be read out at the population level to discriminate conspecific sex. To explore this, we constructed linear decoders and found that male vs. female investigation could be decoded from both CaMKII$^+$ and Vgat$^+$ ensembles with above chance accuracy (**Figures 3.3J-K**). However, when we examined the dependence of decoder performance on population size, we observed that

GABAergic neurons, when compared using the same number of neurons, were consistently better at discriminating conspecific sex (**Figures 3.3L-M**). This suggests again that sex information is more strongly encoded within inhibitory neurons. Taken together, these data demonstrate that while conspecific sex is represented more strongly in inhibitory neurons, population-level encoding of conspecific sex recruits both subpopulations of cells, which both exhibit a stronger representation of female compared to male cues.

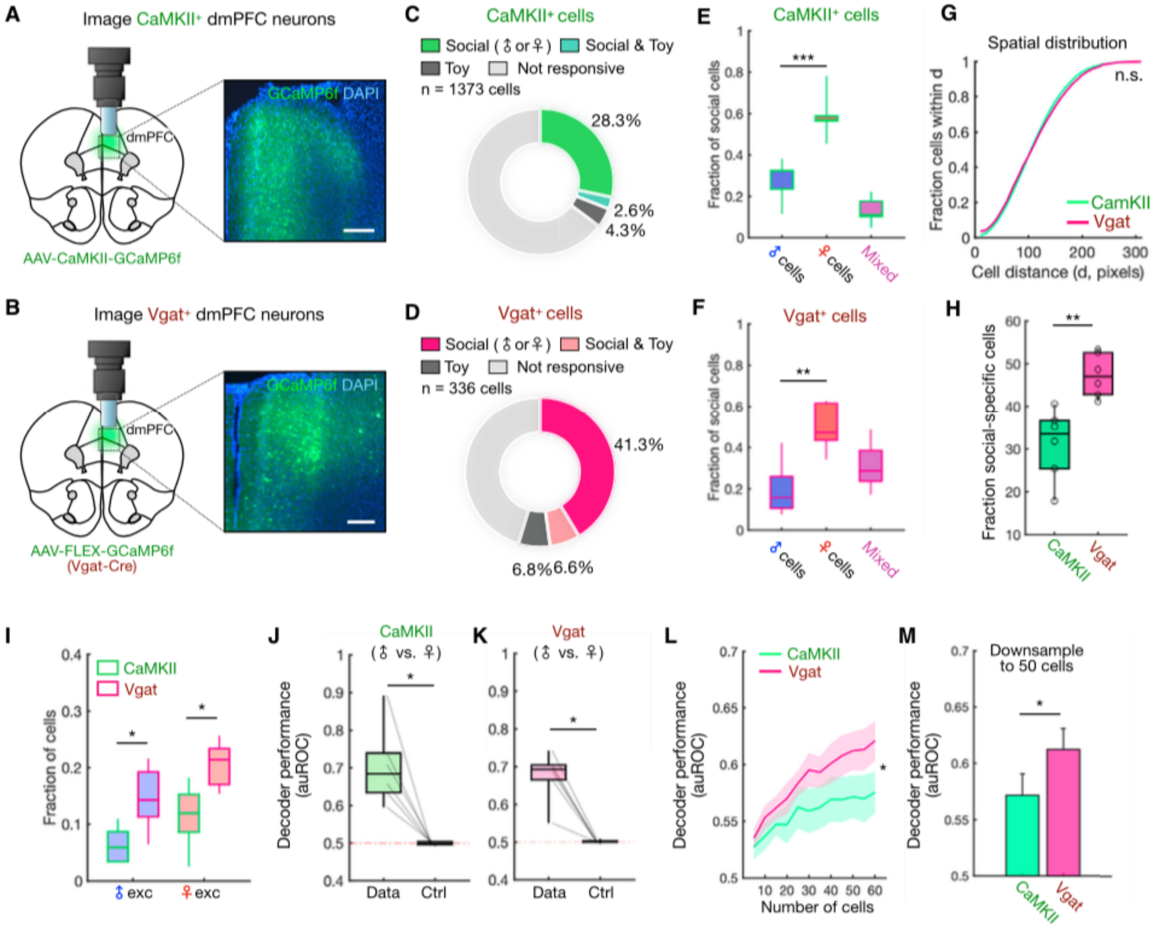

Figure 3.3: dmPFC encoding of sex is distributed across distinct neuronal subpopulations. **(A and B)** Schematic showing lens and microscope placement above dmPFC (left) and expression of GCaMP6f in CaMKII$^+$ or Vgat$^+$ neuron cell bodies (right). CaMKII-GCaMP labeled 25.4% ± 2.6% of dmPFC cells, and Vgat-GCaMP labeled 9.6% ± 0.3% of dmPFC cells (mean ± SEM). **(C and D)** Distribution of social (male and/or female responsive) and toy-responsive neurons among CaMKII$^+$ (n = 1373) or Vgat$^+$ (n = 336) cells. **(E and F)** Comparison of fractions of male- and female-encoding cells within the CaMKII$^+$ or Vgat$^+$ population (E, p = 0.0001; F, p = 0.0015). We also observed a higher fraction of female-encoding cells when normalizing the

sampling of male and female investigation. **(G)** Cumulative histogram showing the mean fraction of cells within a given pairwise distance (x axis) for CaMKII$^+$ and Vgat$^+$ neurons. **(H)** Comparison of the percentage of social cells identified in CaMKII$^+$ versus Vgat$^+$ populations (p = 0.0043). **(I)** Fractions of male- and female-excited cells within the CaMKII$^+$ and Vgat$^+$ populations (male-excited, p = 0.031; female-excited, p = 0.012). **(J and K)** Performance of Fisher's linear discriminant (FLD) decoders trained to classify male versus female investigation using the CaMKII$^+$ or Vgat$^+$ population. Cross-validated auROC is compared with performance of null models trained using time-permuted calcium traces (J, p = 0.0312; K, p = 0.0312). **(L)** Decoder performance in CaMKII$^+$ and Vgat$^+$ cell populations as a function of ensemble size (mean ± SEM, p = 0.040 number/cell-type interaction). **(M)** Decoder performance of CaMKII$^+$ versus Vgat$^+$ cell populations after down-sampling ensembles to 50 neurons (mean ± SEM, p = 0.0411). In (E) and (F), one-way ANOVA followed by Tukey's range test; (G) two-sample Kolmogorov-Smirnov test; (H), (M) Mann-Whitney U test; (I) Tukey's range test; (J and K) Wilcoxon signed-rank test; (L) two-way repeated-measures ANOVA. n = 6 animals for each group (CaMKII$^+$ and Vgat$^+$). ***p < 0.001, **p < 0.01, *p < 0.05, n.s., not significant. Scale bar, 200 mm. See also Figure S2.

3.4.4: Encoding of conspecific sex across distinct contexts

In order to guide behavior in an adaptive way, one may expect representations of sex to generalize across different environments as well as animal identities. To determine whether dmPFC encoding of conspecific sex is general across different environments, we followed the home cage social investigation assay with a two-chamber assay in which animals freely explored novel male and female conspecifics placed in opposing corners in a two-chamber environment (**Figure 3.4A**). As mPFC neurons may also encode spatial information [9], we performed two sessions alternating the locations of male and female stimuli in the arena to dissociate encoding of conspecific sex from location (**Figure 3.4A**). We used DeepLabCut [16] to automatically track the location and pose of subject animals—this was then used to identify social investigation events in an unbiased manner (see Methods). We first confirmed that single dmPFC neurons also encode sex information in the two-chamber context. Individual cells responded to male or female stimuli across sessions with alternating locations, indicating that they encode conspecific sex irrespective of spatial conjunction (**Figure 3.4B**). By precisely mapping neuronal cell bodies across the home cage and two-chamber assays (**Figure 3.S2F**, see Methods), we found that 26% of the recorded cells responded selectively across both assays (**Figures 3.4C-E**), suggesting that single cells in the dmPFC are able to encode conspecific sex consistently across different contexts. To explore whether the population representation within each animal could generalize across contexts, we trained linear decoders to discriminate male vs. female investigation across both sessions. Consistent with our observations of shared encoding at single neuron levels, population decoders could predict sex identity across both contexts (**Figure 3.4F**). Further, male and female investigation events were generally well separated by a discrimination axis defined in only one context (**Figure 3.4G, 3.S3A-B**). Together, these data suggest that dmPFC representations of male and female cues are general across distinct contexts.

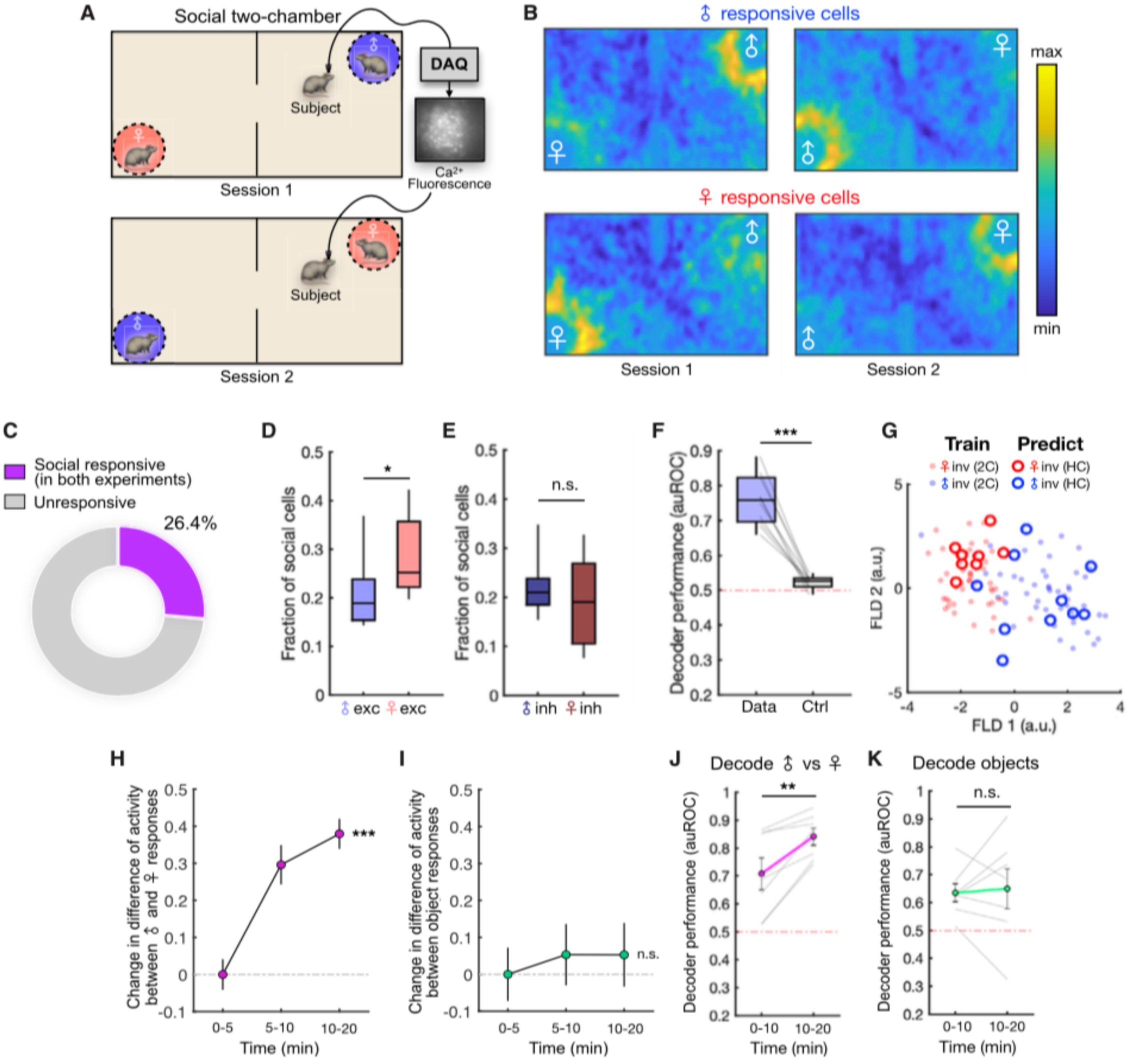

Figure 3.4: Encoding of conspecific sex across contexts and experience-dependent changes. *(***A**) Illustration of the two-chamber social exploration task. Male and female locations are switched across two sequential sessions. **(B)** Neural activity heatmaps showing mean responses of male-excited (top) and female-excited (bottom) neurons during the two-chamber assay. **(C)** Distribution of social (male or female responsive) dmPFC neurons that are significantly responsive across both the home cage social investigation and two- chamber assays. **(D)** Fractions of male- and female-excited cells that consistently encode sex information across both the home cage and two-chamber sessions (p = 0.031). **(E)** Fractions of male- and female-suppressed cells that encode sex information across sessions (p = 0.450). **(F)** Performance of Fisher's linear discriminant (FLD) decoders trained using data from both the social investigation and two-chamber assays and tested on data from both. Cross-validated auROC is compared to control decoders constructed using time-permuted calcium traces (p = 0.002). **(G)** Projection of mean population responses associated with male and female investigation bouts from the home cage (circles) and two-chamber (solid dots) sessions onto FLD components computed from two-chamber data. Performance of cross session decoders using sex-selective cells are shown in Figures 3.S3A and 3.S3B. **(H)** Change in activity difference of male- and female-responsive cells evoked by male versus female investigation during different time epochs in the two- chamber assay (mean ± SEM, p < 0.0001). Stimulus-evoked activity in each epoch is normalized to overall population activity. **(I)** Change in activity difference of object-excited cells during investigation of one object versus the other object (mean ± SEM, p = 0.87). Stimulus-evoked activity in each epoch is normalized to overall population activity. **(J)** Performance of decoders to discriminate male versus female investigation using population dynamics during the first and second half of the two-chamber session (animal locations not switched, p = 0.016). **(K)** Performance of decoders to discriminate investigation of different objects using population dynamics during the first and second half of non-social two- chamber experiments (p = 0.69). In (D) and (E), Mann-Whitney U test, n = 10 animals; (F) Wilcoxon signed-rank test, n = 10 animals; (H and I) one-way ANOVA, n = 291 cells (H), n = 115 cells (I); (J and K) Wilcoxon signed-rank test, n = 7 animals. ***p < 0.001, **p < 0.01, *p < 0.05, n.s. = not significant.

3.4.5: Representation of conspecific sex is strengthened by short-term social experience

Previous work has suggested that in subcortical regions, neural representations of conspecific sex are sharpened over the course of days or weeks [3,4]. Even within the course of one interaction episode, accumulation of social cues may dynamically shape subsequent interaction. We therefore explored whether and how neural representations of sex may change over a faster time scale during a single exploration episode in the two-chamber. Interestingly, we found that the mean activity of sex-encoding cells during investigation of male and female conspecifics changed over the course of the session (**Figures 3.S3C-D**) such that the difference in response strength between the two stimuli grew larger (**Figure 3.4H**). This was not true for cells that responded to non-social objects (**Figure 3.4I**), suggesting that single cell discriminability of sex is specifically sharpened over time. Indeed, population decoders trained and tested during the first and second halves of the session showed significantly higher performance in the second half (**Figure 3.4J**), but not for decoding of two objects (**Figure 3.4K**), suggesting an increase in population-level sex discriminability that depends on social experience. These data suggest that although representations of sex are stable and general, they can also be sharpened by social experience.

In order to further analyze how changes in responses to male and female cues affect neural tuning, we constructed generalized linear models (GLM) to model single neuron activity as a function of behavioral variables, including social investigation (**Figure 3.S3F**). Analysis of the coefficients fit to single neuron GLMs showed a significant increase in tuning to male and female investigation for male and female cells, respectively, over time (**Figures 3.S3G-I**). In contrast, cells that responded to objects did not show an increase in tuning (**Figures 3.S3J-K**), further indicating that encoding of conspecific sex, but not non-social stimuli, is sharpened with social experience. While previous work has found that subcortical encoding of social information can be strengthened over days [3,4], these results suggest that cortical sex representations may be more amenable to short-term sharpening.

3.4.6: Activity of sex-encoding cells predicts behavioral sex preference

Internal representations of sex-specific social cues are thought to play an important role in driving behavior directed toward male or female conspecifics [1,2], raising the possibility that sex representations in the dmPFC may play a role in controlling male- or female-directed sociality. Based on this, we next explored the relationship between sex encoding and sex preference behavior in individual animals.

At the behavioral level, we observed a modest group-level bias in males toward female investigation (**Figure 3.S4A**), consistent with previous literature [17,18]. Interestingly, we found earlier that dmPFC neurons encoded female more strongly in distinct cell types (**Figures 3.1I, 3.2F, 3.3E, and 3.3F**), and across different contexts (**Figure 3.4D**), suggesting that this sex encoding bias may be linked to female preference behavior in males. To rule out the possibility that bias toward female encoding may be due to unequal sampling of investigation events and resulting differences in statistical power, we controlled this analysis by normalizing the representation of male and female investigation (see Methods). We confirmed that a larger fraction of dmPFC neurons encoded females even when behavioral events were precisely equalized (**Figure 3.S2A-D**).

However, beyond this group-level female preference, individual males displayed a wide range of preference scores, with some displaying a preference for males (**Figure 3.S4A**). When social stimuli were absent, the variability in behavioral preference in the two-chamber assay was significantly reduced (**Figure 3.S4C**), indicating that individual preferences for male or female are driven by the presence of conspecifics and not simply by random movement. Moreover, this individual behavioral preference is relatively stable—sex preference of individual animals was significantly correlated within sessions and across consecutive days (**Figures 3.S4D-E**), indicating that sex preference in individuals reflects a consistent behavioral state.

Interestingly, when we compared animals that displayed a stronger preference for female vs. for male interaction, female-preferring animals had significantly more female-responsive neurons than male-preferring animals (**Figure 3.S4G-H**), suggesting that the relative strength of male vs. female representations is linked to behavioral preference. In light of this, we next explored the relationship between population-level encoding of sex and preference behavior. To analyze the strength of male and female representations, we measured the population response amplitude during each male/female investigation event as the Mahalanobis distance between the stimulus-evoked population vector and baseline activity. While female-preferring animals showed a stronger population response during female interaction (**Figures 3.5A-B**), male-preferring animals showed a stronger response during male interaction (**Figures 3.5C-D**). Taken together, these data suggest that a preference state is encoded in the relative activity of neural sex representations that predicts sex preference behavior (**Figure 3.5E**).

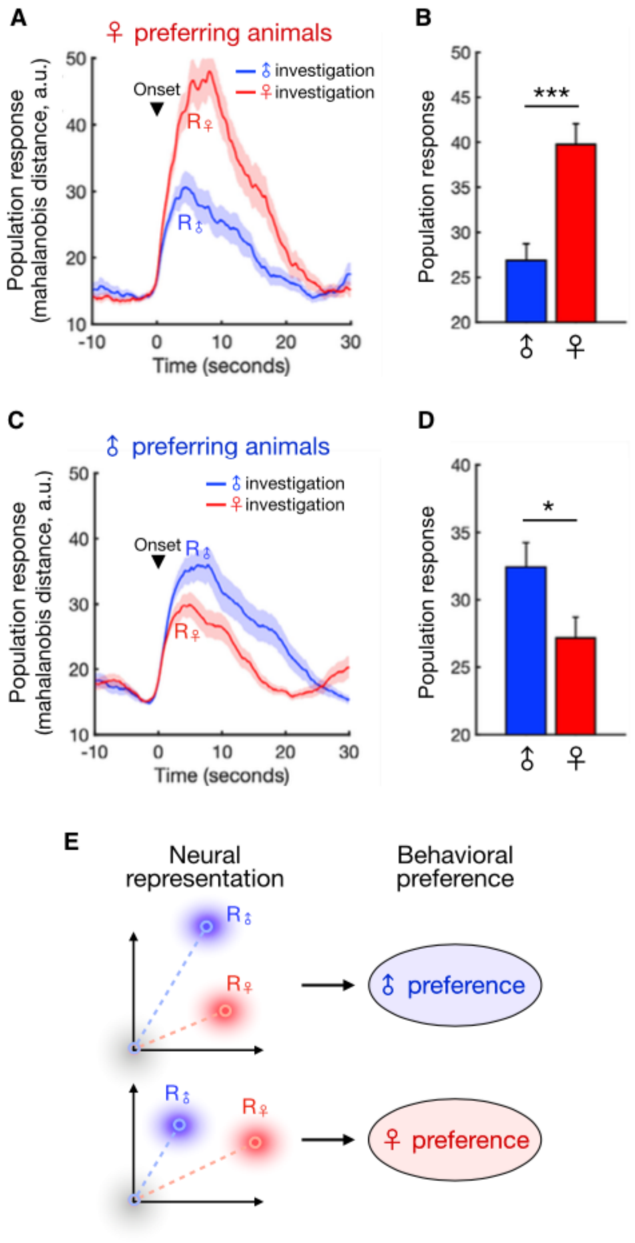

Figure 3.5: dmPFC neurons encode a sex preference state. **(A and C)** Average population response, measured as the Mahalanobis distance between trial response vectors and baseline (see STAR Methods) for male-preferring (A) and female-preferring (C) animals evoked during male or female interaction (mean ± SEM). **(B and D)** Average responses in (A) and (C) quantified over the first 10 s following interaction (mean ± SEM, Mann-Whitney U test, p = 1.26e4, n = 46 male bouts and n = 42 female bouts [B], p = 0.0195, n = 60 male bouts and n = 66 female bouts [D]). **(E)** Schematic showing how the relative strength of sex-specific responses predicts individual sex preference. ***p < 0.001, *p < 0.05.

3.4.7: Male and female neural representations modulate sex preference behavior

The strong association between the neural representation of conspecific sex and behavioral preference raises the possibility that ensemble representations may causally influence preference behavior. A typical way to examine causal influence of neural activity is to activate an anatomically or genetically defined subset of neurons [24–26]. However, accumulating evidence points to an increasingly common picture in which one neuronal subpopulation may participate in multiple representations, and one representation may recruit multiple subpopulations [5,6,27]. Indeed, we found that sex representations in the dmPFC are distributed across excitatory and inhibitory subpopulations (**Figure 3.3**). This underscores the necessity of directly controlling neuronal populations defined by natural activity rather than by specific molecular markers.

Therefore, to test the functional role of sex representations in sex preference behavior, we employed an activity-dependent labeling strategy to express ChR2 in specific subsets of dmPFC neurons that were activated by male or female conspecifics (**Figure 3.6A**). We virally injected an AAV carrying E-SARE-CreER to express a Tamoxifen (4-TM)-inducible CreER driven by E-SARE, an activity-induced synthetic promoter that was previously shown to display greater activity-dependent induction compared to immediate early genes such as Fos or Arc [14,28]. We also co-injected an AAV expressing Cre-dependent ChR2. This combination of AAVs allowed us to restrict ChR2 expression to the select neural population that was active at the time of 4-TM injection [14]. Specific dmPFC neurons were labeled following natural exposure to either male or female conspecifics, and then optogenetically re-activated several weeks later in the two-chamber assay to measure sex preference. Histological analysis showed robust expression of ChR2 in animals exposed to social stimuli at the time of 4-TM induction, but not in animals lacking 4-TM injection (**Figures 3.S5A-B**), confirming specific labeling of neurons that were active during social interaction.

In order to further validate the efficacy and specificity of this strategy, we examined the overlap between EYFP$^+$ cells labeled by E-SARE following social or non-social exposure and cells expressing the immediate early genes Fos and Arc following a second exposure to males or females (**Figure 3.S5C**). We found that male-induced E-SARE EYFP$^+$ cells showed a substantially higher overlap with male-induced Fos/Arc$^+$ cells, compared to that of EYFP$^+$ cells induced by females or in home cage controls (**Figure 3.S5D**). Conversely, female-induced E-SARE EYFP$^+$ cells showed a substantially higher overlap with female-induced Fos/Arc$^+$ cells, compared to that of EYFP$^+$ cells induced by males or in home cage controls (**Figure 3.S5E**). This suggests that the E-SARE labeling strategy captures specific ensembles of neurons that encode sex-specific cues across distinct social experiences.

We then optogenetically re-activated neurons that were naturally activated by exposure to females (Figure S5F-H). Stimulation of female cells resulted in an acute change in sex preference behavior toward female-directed interaction (**Figure 3.6B, 3.S5I**). This was driven by a specific increase in the time spent investigating the female conspecific, as male investigation time remained unchanged (**Figures 3.6C-D**). In striking contrast, stimulation of male cells had the opposite effect on sex preference behavior (**Figure 3.6E, 3.S5J**), resulting in an acute bias toward male interaction driven by a specific increase in male-directed investigation (**Figures 3.6F-G**). This suggests that sex preference behavior is modulated by specific subpopulations of neurons that are defined by their native activation in response to male or female stimuli. Indeed, optogenetic re-activation of neurons that were activated during a home cage exposure without male or female stimuli had no effect on preference behavior (**Figures 3.6H-J, 3.S5K**).

Finally, we analyzed the axonal projection patterns of male- and female-activated dmPFC neurons (**Figure 3.S6**). Both male and female cells showed broad projection patterns to several subcortical structures involved in social behavior and motivation. Although we observed a trend toward stronger nucleus accumbens and dorsal striatum projections in female cells (**Figure 3.S6B**), there was no overall difference between the projection patterns of male and female cells. This suggests that their opposing effects on sex preference behavior are unlikely due to gross differences in their axonal projection patterns.

Together, these data demonstrate that two subsets of dmPFC neurons—which are separable subpopulations that are natively activated in response to sex-specific social cues—modulate interaction with male vs. female conspecifics. These results firmly establish a functional role for cortical representations of conspecific sex in sex preference behavior.

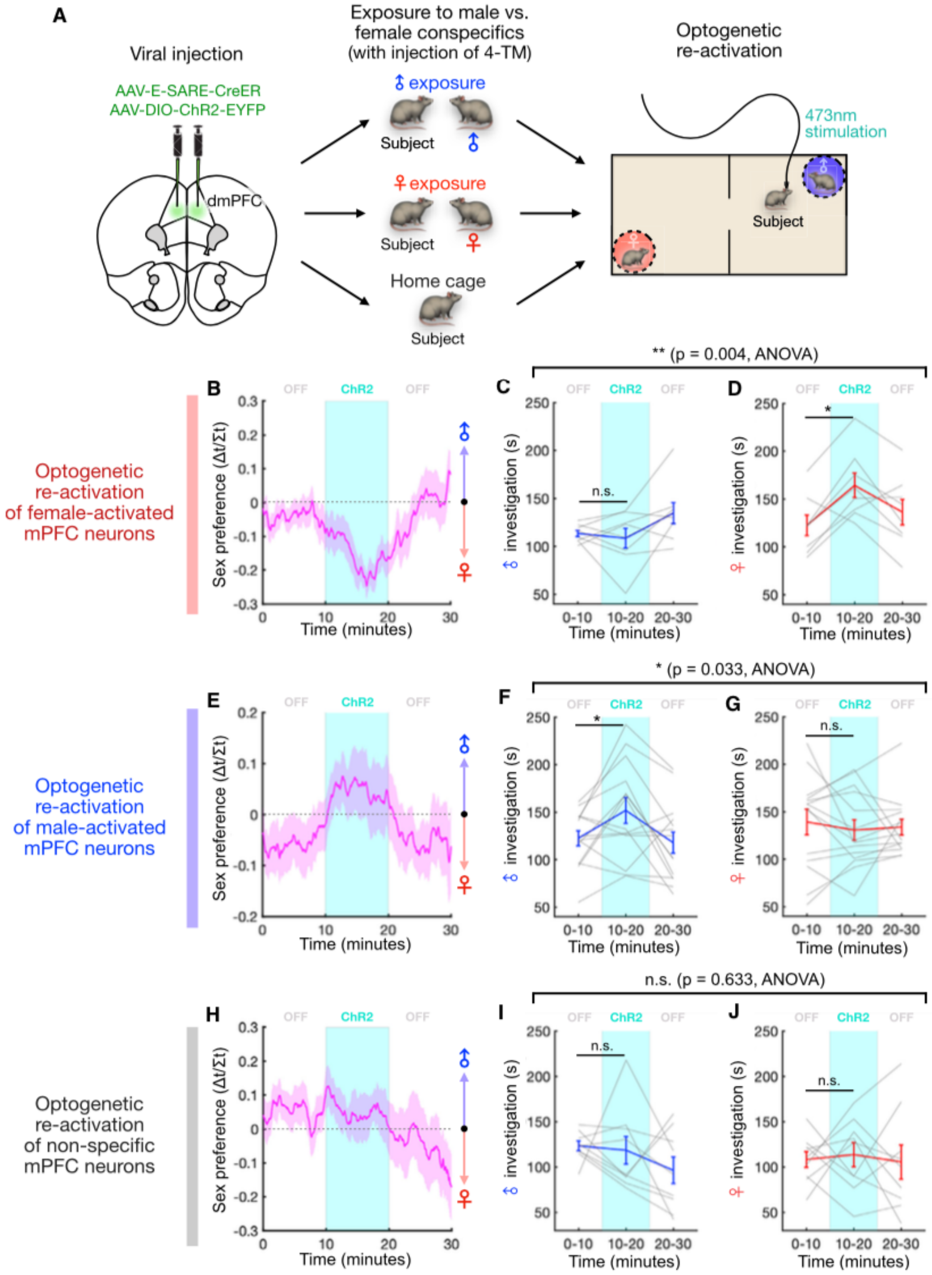

Figure 3.6: Male- and female-activated cells modulate male vs female interaction. **(A)** Experimental paradigm used to test the causal role of male and female representations in sex preference behavior. After viral injection of E-SARE-CreER and Cre-dependent ChR2 constructs, animals interact with either male or female conspecifics followed by tamoxifen injection, and

activity-defined cells are optically stimulated in the social two-chamber environment (see STAR Methods). **(B, E, and H)** Time course of sex preference in the male/female two-chamber session (mean ± SEM). Optical stimulation of female-activated cells induces a bias toward female interaction (B), stimulation of male-activated cells induces a bias toward male interaction (E), and stimulation of non-specific cells induces no bias (H). **(C and D)** Male or female interaction time before, during, and after stimulation of female-activated neurons (mean ± SEM, p = 0.004, light/sex interaction in two- way repeated-measures ANOVA; C, p = 0.84, D, p = 0.016, Wilcoxon signed-rank test; n = 8 animals). **(F and G)** Male or female interaction time before, during, and after stimulation of male-activated neurons (mean ± SEM, p = 0.033, light/sex interaction in two-way repeated-measures ANOVA; F, p = 0.042, G, p = 0.50, Wilcoxon signed-rank test; n = 14 animals). **(I and J)** Male or female interaction time before, during, and after stimulation of non-specific neurons (mean ± SEM, p = 0.63, light/sex interaction in two-way repeated-measures ANOVA; I, p = 0.36, J, p = 0.32, Wilcoxon signed-rank test; n = 9 animals). ***p < 0.001, *p < 0.05, n.s., not significant.

3.5: Discussion

Using microendocsopic calcium imaging, we found that the dmPFC uses a distributed neural code, which recruits both excitatory and inhibitory subpopulations, to represent conspecific sex. While both male and female dmPFC encodes conspecific sex, an overall stronger encoding of female conspecifics is specifically observed in males but not females, and the relative strength of male vs. female representations predicts sex preference behavior. Further, using activity-dependent optogenetic manipulations of natively active ensembles, we found that cortical representations of conspecific sex modulate sex preference behavior in males. Together, these results demonstrate a functional role for neural representations of sex in shaping behavior and present a neural mechanism for cortical control of male- and female-directed sociality.

3.5.1: Cortical representation of conspecific sex

To navigate the social world, animals must be able to recognize conspecific sex, an essential feature that defines social relationships and shapes interaction between conspecifics. Previous studies have contributed to a classic view that processing of sex-specific social cues depends almost exclusively on subcortical circuits [2–4]. In contrast, growing appreciation for the role of prefrontal circuits in social behavior has spurred an effort to explore native mPFC activity during social interaction [7–9]. As mPFC circuits have been implicated in the control of social motivational states [29], social dominance [7,11], and sexual behavior [10], complex social information must be integrated by mPFC neurons for animals to make adaptive decisions. Still, the basic question of whether cortical neurons play a role in sex recognition has not been addressed. Here, we found that a substantial fraction of neurons in the dmPFC of both male and female mice respond preferentially to male vs. female interaction.

Recent work has shown that neurons in the mPFC respond similarly to both male and female odor cues [30]. Our study shows that, during natural interaction with conspecifics, males and females are encoded in largely non-overlapping sets of neurons that form unique population representations. This suggests that natural social stimuli may drive neural responses differently, and perhaps more strongly, than simple odor cues. Furthermore, we found that in the brains of male but not female mice, prefrontal neurons exhibit a bias toward stronger encoding of female cues which is maintained at the population level. This representational dimorphism may reflect distinct mechanisms underlying sex encoding and preference behavior in males vs. females. Indeed, subcortical sex representations, which also show a bias toward encoding of female cues in males, are selectively regulated by oxytocin signaling in males but not females [4,17]. More generally, whether and how cortical sex representations differ from sex encoding in subcortical regions such as the amygdala and hypothalamus [3,4] also deserve future attention. While subcortical representations are shaped over the course of days with social experience, we found that neurons in the dmPFC sharpen their sex-specific tuning on a faster timescale of 10-20 minutes. This may suggest that cortical representations of sex are more flexible, and may be more amenable to conjunctive coding schemes that link social stimuli with contextual or spatial cues [9].

We also leveraged the optical recording approach to investigate how different populations of dmPFC neurons contribute to sex representations. By imaging both excitatory (CaMKII$^+$) and inhibitory (Vgat$^+$) neurons, we found that both subpopulations encode male and female cues and discriminate conspecific sex. Interestingly, while sex information was distributed across both subpopulations, male and female cues were represented more strongly by GABAergic neurons, suggesting that sex-specific social information is enriched in inhibitory neurons. These results illustrate the neural heterogeneity of native sex representations, highlighting the limitations of defining functional circuitry based molecular or anatomical profiles.

3.5.2: Neural representation of behavioral sex preference

Despite our current understanding of how neurons encode sex cues in the brain, how native neural representations are related to social behavior, such as opposite-sex preference, remains poorly understood. Previous studies have looked into the group-level bias that animals display toward interaction with opposite-sex conspecifics [17,18,31]. However, this has left open the question of how male- vs. female-directed social interaction is actually controlled in individual animals. Here, we show that besides the group-level preference for females, individual males naturally exhibit distinct preferences for either male or female interaction. At the neural level, we observed a stronger representation of female in male animals across multiple contexts, and across both excitatory and inhibitory subpopulations, linking neural representations of sex to a group-level behavioral preference toward females. Yet beyond this overall enrichment of female cells in males, dmPFC ensembles also encode individuals' sex preference states in the relative activity of male vs. female population responses. While female-preferring animals showed a stronger neural response to female conspecifics, male-preferring animals in fact showed a stronger response to males. Collectively, these data support the idea that behavioral sex preference is closely linked to a particular representation of conspecific sex in the prefrontal cortex, which separately controls male- vs. female-directed sociability.

3.5.3: Control of sex preference by native ensemble representations

Although some molecularly defined neural populations have been linked to social behaviors, it has been challenging to establish causal roles for native ensemble representations of sex information. In part, this is because neural representations of social features such as conspecific sex are often distributed across distinct neuronal cell types that defy precise molecular or anatomical distinction [5,6,27]. Indeed, we show here that cortical representations of conspecific sex are distributed across both excitatory and inhibitory neural subpopulations. In the context of early sensory and valence processing, activity-dependent labeling has been used to re-activate native neural representations, yielding insights into circuit function that may be missed by more traditional approaches [32–35]. Using activity-dependent labeling coupled with optogenetic stimulation, we show that neurons encoding male and female are directly involved in social behavior, providing a critical missing link between native sex representations and control of behavior. When optogenetically re-activated weeks later, neural populations that were natively activated during interaction with male or female conspecifics acutely and specifically increased male or female interaction, respectively. The fact that each subpopulation only increased social interaction with a specific sex demonstrates that male- and female-directed sociality is regulated by separable prefrontal representations. Together, the activity of these two representations effectively control a sex preference state that may be adjusted to shape behavior. As immediate early gene-based labeling methods cannot capture neural ensembles that show reduced activity during stimulus presentation, how socially-inhibited neurons play a role in sex preference behavior remains an open question for future studies.

Collectively, these results establish a novel role for the prefrontal cortex in encoding of conspecific sex and provide direct evidence of a functional role for native sex representations in shaping male- and female-directed social behavior. These findings provide new insights into how the brain transforms social information into adaptive behavior.

3.6: Supplemental data

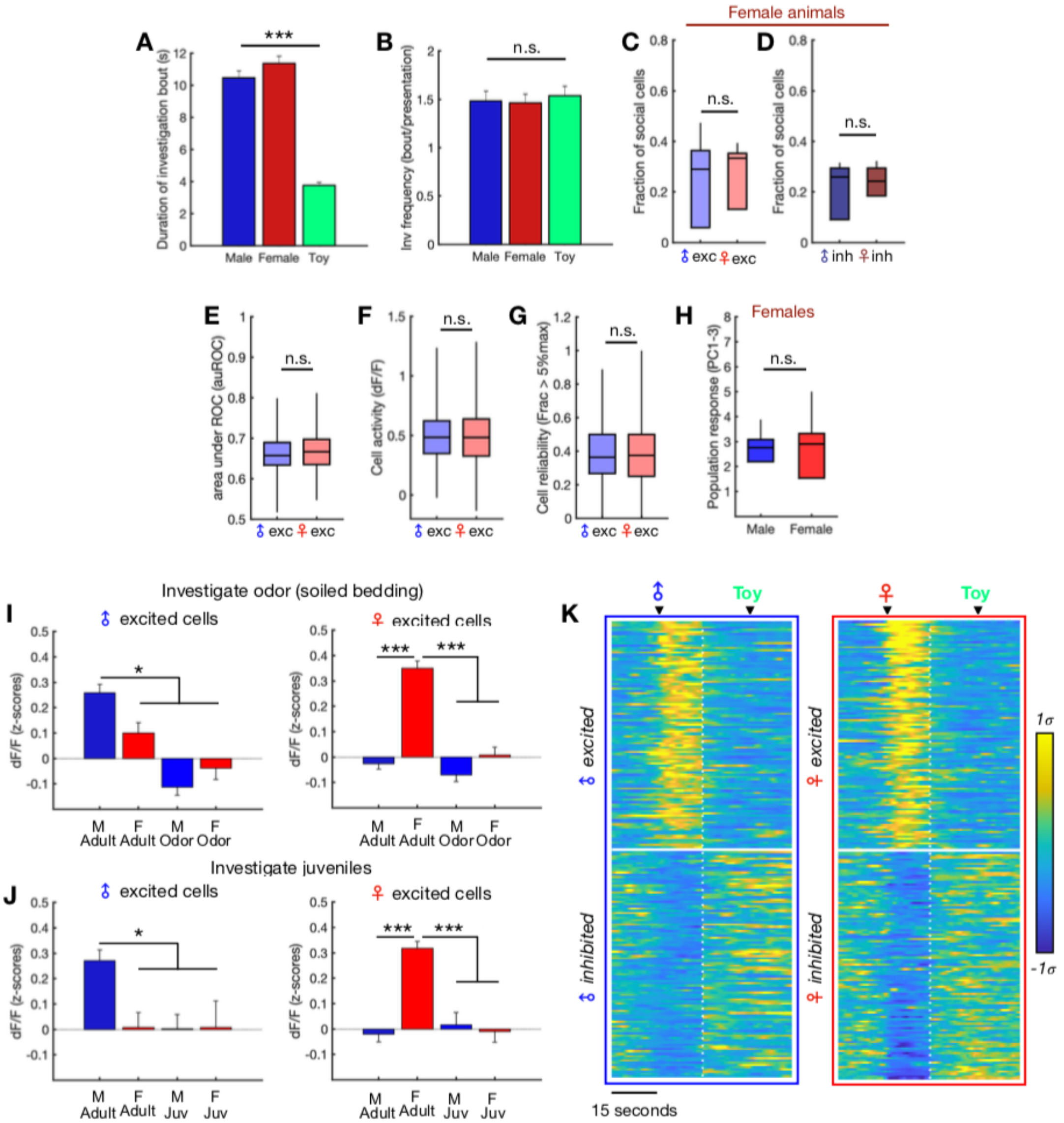

Figure 3.S1: Analysis of social investigation and firing properties of male- and female-encoding cells. **(A)** Comparison of the duration of male, female, and toy investigation bouts in the home cage social investigation assay (mean ± SEM, p = 1.4e-47). **(B)** Comparison of the frequency of male, female, and toy investigation bouts. (mean ± SEM, p = 0.869). **(C-D)** Fractions of male- and female-excited (C) or suppressed (D) cells among socially responsive neurons recorded from female animals (C, p = 1; D, p = 0.89). Mixed cells responding to more than one category constituted 5.31 ± 1.96% (mean ± SEM) of socially responsive cells. **(E)** Distributions of auROC (area under ROC) values for male- and female-excited cells (p = 0.202). **(F)** Response strength of cells during social investigation computed as the average Z-scored ΔF/F activity over all male or female investigation bouts (p = 0.84). **(G)** Response reliability of cells during social investigation computed as the fraction of bouts where the change in activity exceeds 5% of maximum (p = 0.90). **(H)** Comparison of the population response amplitude during male vs. female interaction in female animals. Population responses (projected to PC1-3) are measured over the first 10 seconds of interaction and averaged across bouts within each animal (p =

0.8413). **(I)** Activity of male-excited and female-excited cells during investigation of adults or odors (using soiled bedding from male and female conspecifics) (mean ± SEM). **(J)** Activity of male-excited and female-excited cells during investigation of male or female adult or juvenile conspecifics (mean ± SEM). **(K)** Average responses for male- and female-responsive neurons centered around onset of stimulus (male or female) and toy. The top 100 neurons, sorted using rank- ordered auROC values, are shown for each cell type. In (A-B), one-way ANOVA; (C-D, H) Mann-Whitney U test, n = 5 animals; (E-G) Mann-Whitney U test, n = 228 male-encoding cells and 331 female-encoding cells; (I-J) Tukey's range test, n = 39 male-encoding cells and n = 87 female-encoding cells (I); n = 24 male-encoding cells and n = 51 female-encoding cells (J). ***p < 0.001, *p < 0.05, n.s. = not significant.

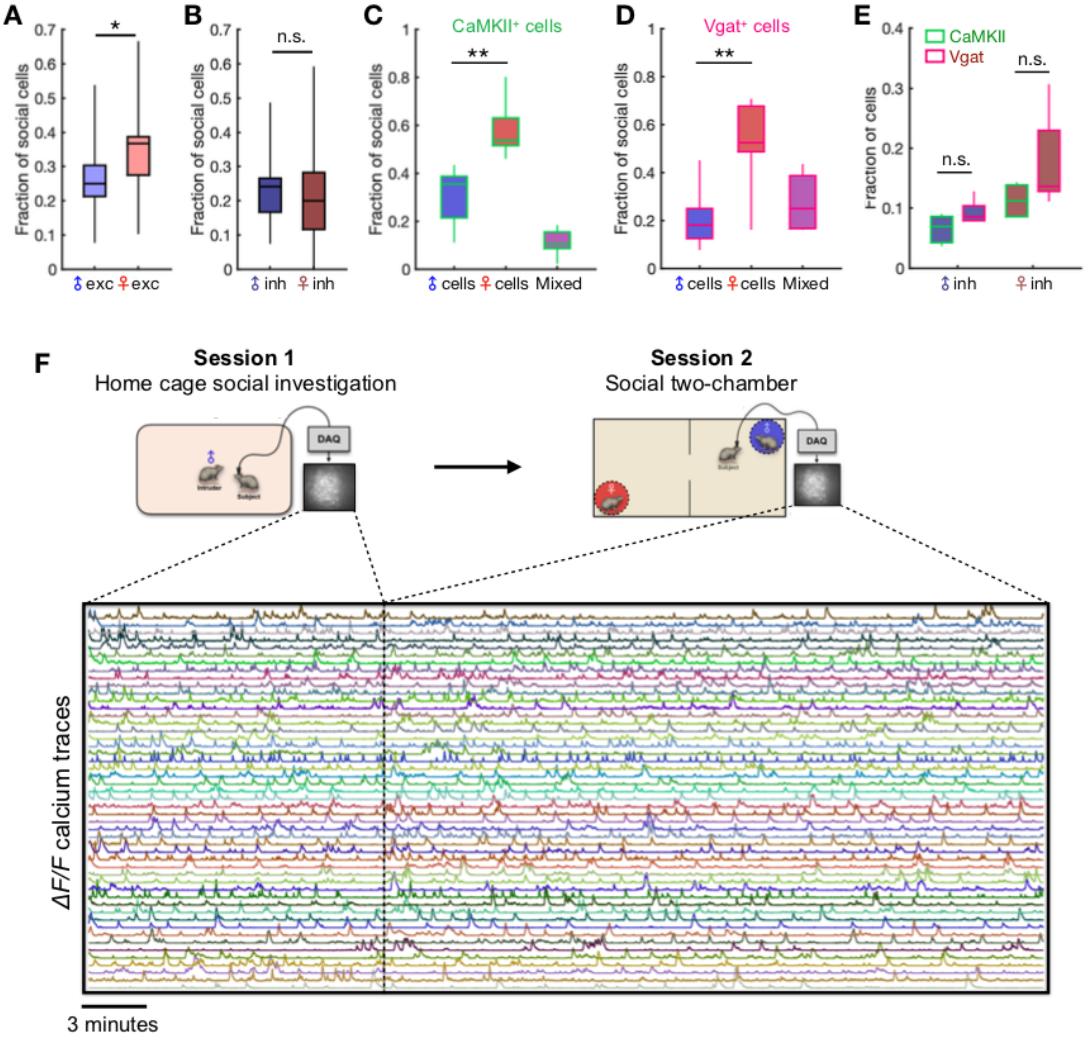

Figure 3.S2: Analysis of sex-encoding cells with equalized behavior sampling. **(A)** Fractions of male- and female-excited (A) or suppressed (B) cells computed using ROC analysis where calcium traces and behavior vectors have been normalized to equalize representation of male and female events (see Methods). This controls for differences that may be due to unequal sampling of behavior variables (A, p = 0.012; B, p = 0.42). **(C-D)** Fractions of male- and female-responsive neurons in the CaMKII^{+} or Vgat^{+} population computed using ROC analysis after equalizing representation of male and female bouts (C, 0.0012; D, p = 0.009). **(E)** Fractions of male- and female-inhibited cells within the CaMKII^{+} and Vgat^{+} populations (p = 0.68 (male-

inhibited), p = 0.096 (female-inhibited)). **(F)** Schematic showing sequential home cage social investigation and two-chamber experiments. The microscope is not removed between sessions, and fluorescence videos from each session are concatenated and processed together so that the same exact neurons are identified across sessions. Image shows 50 example neurons recorded from one animal across both sessions. Example fluorescence images for illustration only. In (A-B) Mann-Whitney U test, n = 23 animals; (C-E) One-way ANOVA followed by Tukey's range test, n = 6 animals per group. **p < 0.01, *p < 0.05, n.s. = not significant.

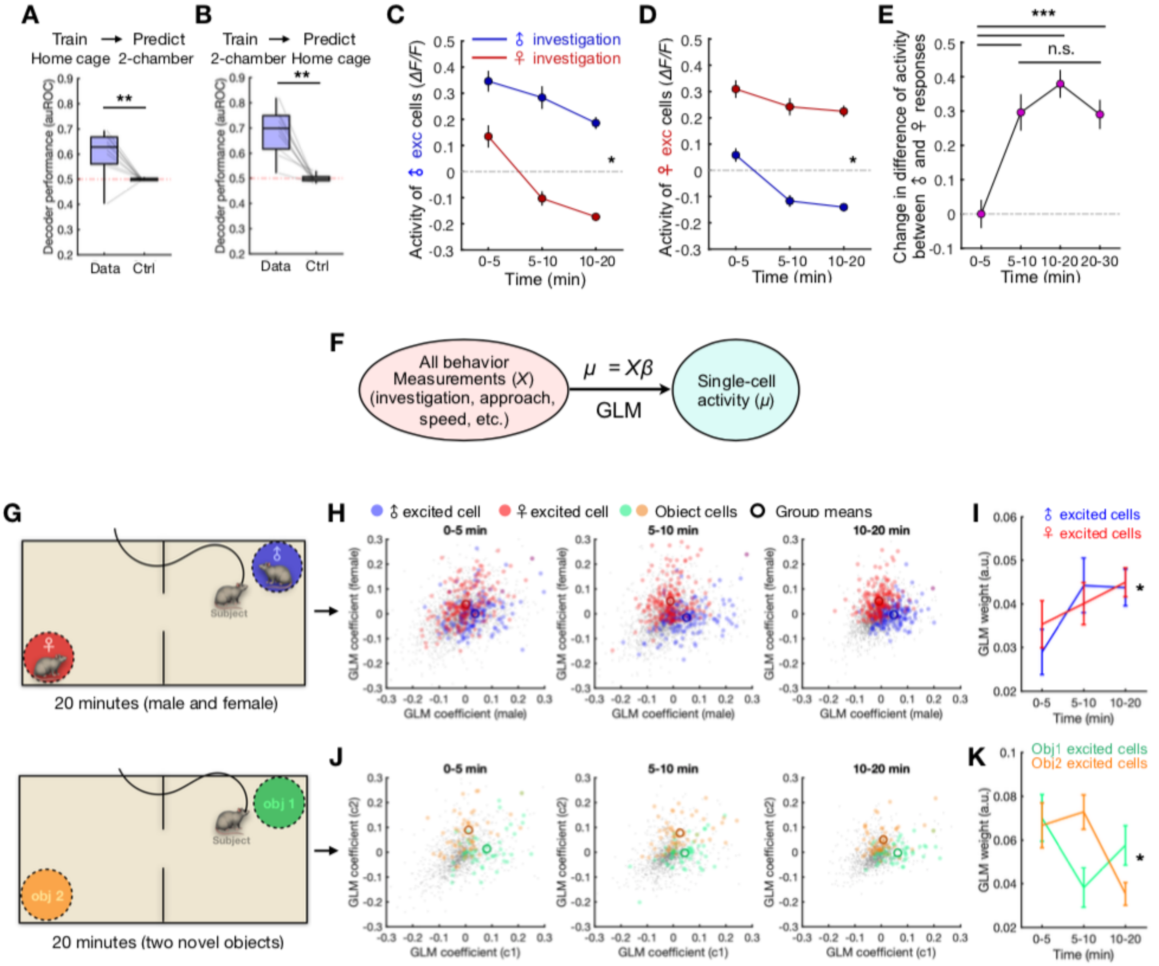

Figure 3.S3: Encoding of social information is strengthened over time. **(A)** Performance of Fisher's linear discriminant (FLD) decoders trained on calcium data from the home cage session to classify male vs. female investigation using population activity in the two-chamber session. Performance is compared with null models constructed using time-permuted calcium traces (p = 0.0098). **(B)** Performance of decoders constructed using data from the two-chamber session to predict the sex identity of interaction events in the home cage, compared with performance of null models as in (A) (p = 0.002). **(C-D)** Mean activity of male-excited or female-excited neurons during male and female investigation events over different time epochs during the two-chamber experiment (mean ± SEM, p = 0.021 (C), p = 0.048 (D), time/sex interaction). **(E)** Change in differential activity of male- and female-responsive cells evoked by male vs. female investigation during different time epochs in the two-chamber (mean ± SEM). Stimulus-evoked activity in each epoch is normalized to overall population activity. **(F)** Generalized linear models are constructed to model single-neuron activity using behavior variables including male and female investigation

(see Methods). Model weights are analyzed to measure cell tuning to different variables over time. **(G)** Schematic showing two-chamber sessions with either male and female stimuli (top) or novel objects (bottom). **(H)** Weights from generalized linear models (GLM) fit to model single-neuron calcium activity as a function of task parameters in the two-chamber assay (see Methods). Plots show weights fit to male and female investigation for male- (blue) and female-excited (red) cells during different epochs of the session. **(I)** GLM weights fit to tuned stimuli in male- and female-responsive cells over different epochs of the experiment (p = 0.032, time factor). **(J)** GLM weights fit to model single neuron activity as in (H) from two-chamber sessions with no social stimuli. Colored dots show weights fit to investigation of each of the novel objects for cells that are significantly active during novel object investigation. **(K)** GLM weights fit to model object-excited cells during investigation of tuned stimulus. Weights do not increase over time, as is seen with male- and female- responsive cells. (mean ± SEM, p = 0.047, time factor). In (A-B), Wilcoxon signed-rank test, n = 10 animals; (C-D) Two-way ANOVA, n = 125 cells (C), n = 166 cells (D); (E) One-way ANOVA followed by Tukey's range test, n = 291 cells; (I, K) Two-way ANOVA. ***p < 0.001, **p < 0.01, *p < 0.05, n.s. = not significant.

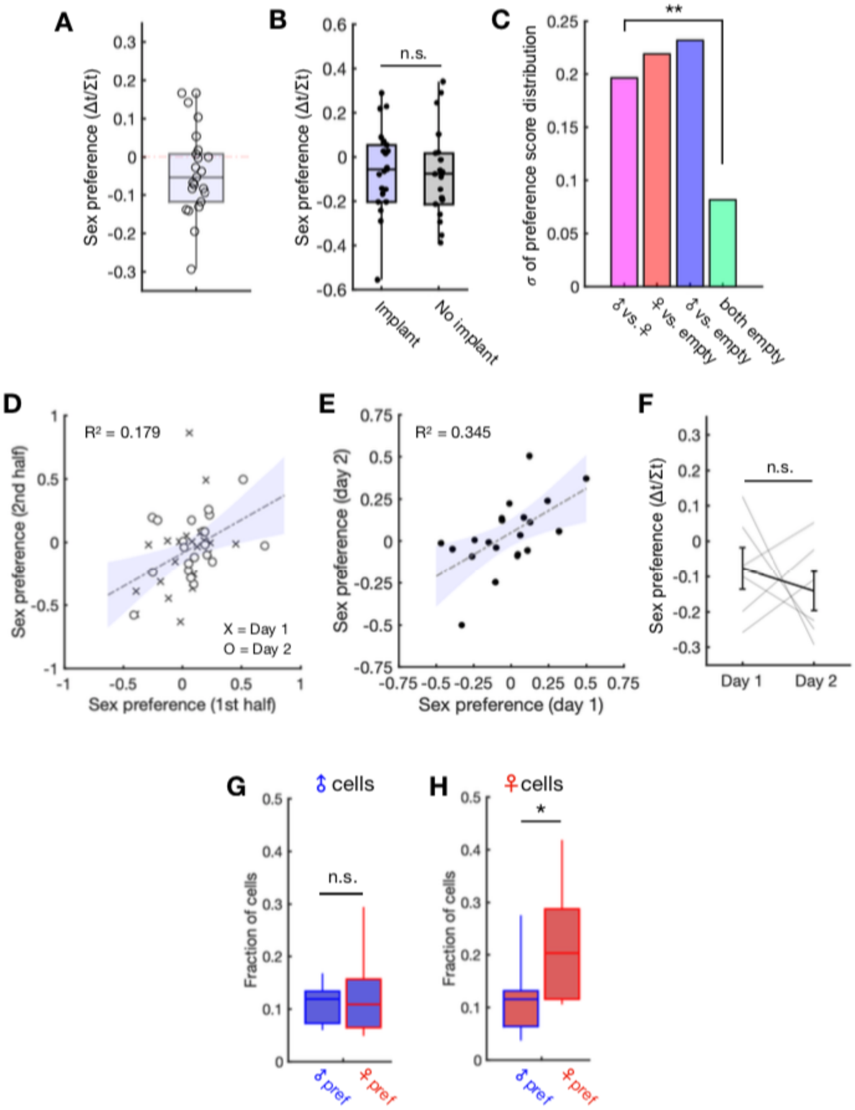

Figure 3.S4: Analysis of sex-preference behavior in the two-chamber assay. **(A)** Distribution of sex preference scores in the home cage assay [(*T*♂-*T*♀)/(*T*♂+*T*♀)]. **(B)** Distribution of sex preferences scores in the two-chamber assay for animals that were implanted and wearing a microendoscope and animals that were not (Mann- Whitney U test, p = 0.5629). **(C)** Variability (standard deviation) in preference scores for animals in different two-chamber conditions. The natural variability in sex preference during the male/female two-chamber (n = 21 animals) is similar to the variability in social preference elicited by male (n = 13 animals) or female (n = 13 animals) conspecifics compared to an empty cup, but higher than when social stimuli are not present (n = 15 animals) (p = 0.0016). **(D)** Linear correlation between sex preference scores calculated over two halves of the two-chamber. Data is pooled across two days of testing (R^2 = 0.179, p = 0.0052). **(E)** Linear correlation between sex preference scores across two days of testing with novel stimulus animals (95% confidence interval, R^2 = 0.345, p = 0.0051). **(F)** Sex preference scores for animals in a two-chamber assay before and after sexual experience with female conspecifics (Mann-Whitney U test, p = 0.69). (**G-H)** Fractions of male (G) and female (H) cells in male- and female-preferring animals (top and bottom 30%) computed using ROC analysis after equalizing representation of male and female bouts (G, p = 0.5; H, p = 0.0036). In (B), Mann-Whitney U test, n = 21 animals per condition; (C) two-sample F-test; (D-E) Linear regression, n = 42 sessions (D), n = 21 animals (E); (F) Wilcoxon signed-rank test, n = 6 animals; (G-H) One-sided Mann-Whitney U test, n = 7 animals per group. **p < 0.01, *p < 0.05, n.s. = not significant.

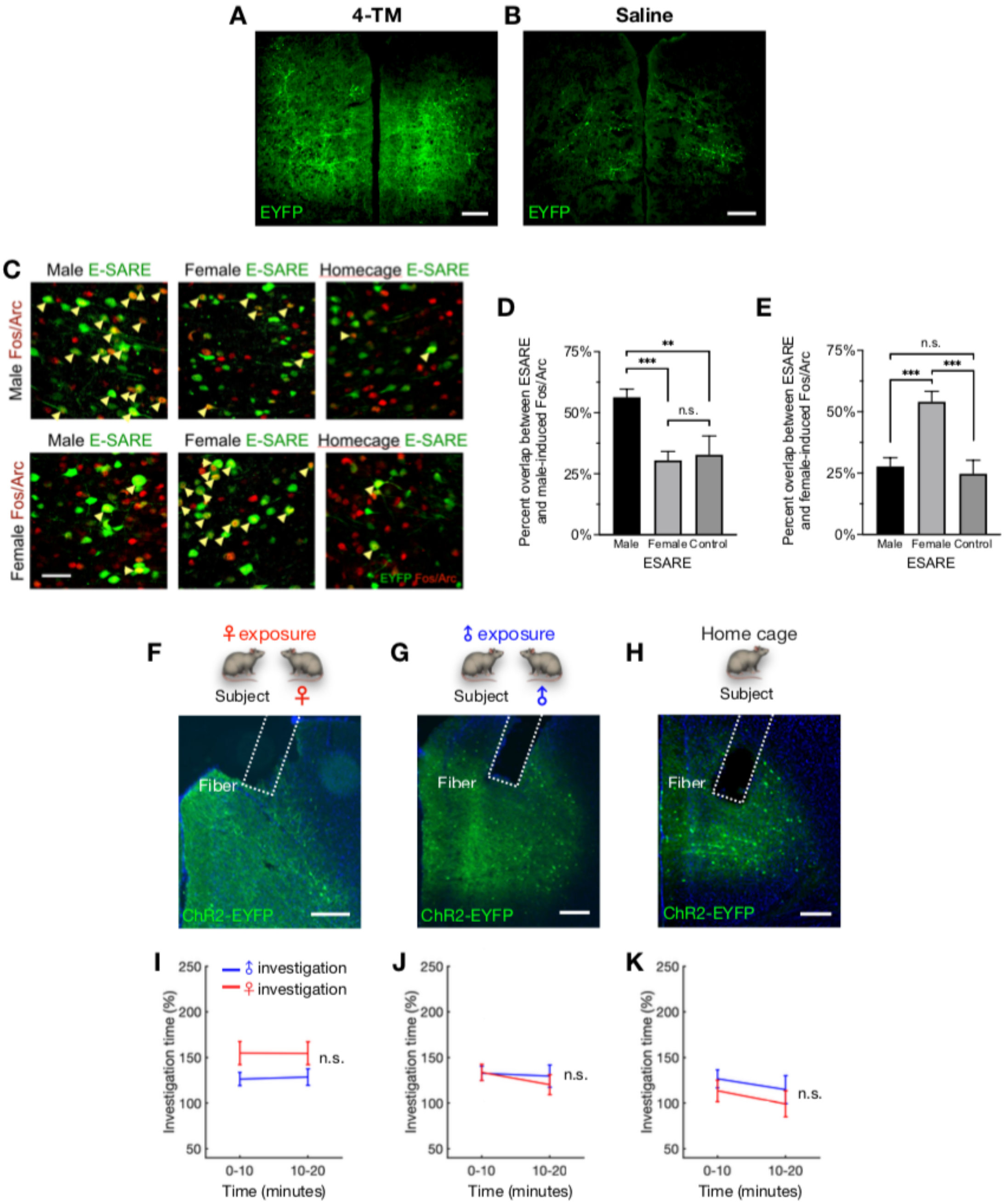

Figure 3.S5: Histological analysis of activity-dependent labeling and optogenetics controls. **(A)** Expression of ChR2-YFP in dmPFC neurons in an animal that was exposed to male conspecifics and received injection of tamoxifen. **(B)** Expression of ChR2-YFP in an animal that was exposed to male conspecifics and received a saline injection. **(C)** Representative fluorescent images showing overlap (yellow) between cells expressing Fos/Arc (red) after exposure to male (top row) or female (bottom row) conspecifics, and cells tagged by E-SARE (green) weeks prior when exposed to male (left) conspecifics, female (middle) conspecifics, or home cage (right). E-SARE- Cre labels 298 cells/mm^2 for male exposure and 323 cells/mm^2 for female exposure, and there is no significant difference between them (p = 0.46). **(D-E)** Quantification of percentage overlap of male-induced (D) or female-induced (E) Fos/Arc$^+$ cells with E-SARE labeled cells (mean ± SEM, p < 0.0001, E-SARE x Fos/Arc interaction). **(F)** Fiber placement above dmPFC

and expression of ChR2 in female-induced cells**.** **(G)** Fiber placement and expression of ChR2 in male-induced cells. **(H)** Fiber placement and expression of ChR2 in non-social dmPFC neurons. **(I)** Time spent investigating male and female in the two-chamber assay for the cohort of female-exposed animals in separate sham sessions where the optic fiber was attached but no light was delivered (mean ± SEM, p = 0.82, time/sex interaction). **(J)** Time spent investigating male and female for the male-exposed cohort in control sessions with fiber attachment and no light (mean ± SEM, p = 0.47, time x sex interaction). **(K)** Time spent investigating male and female for the home cage control cohort in sessions with fiber attachment and no light (mean ± SEM, p = 0.93 time, x sex interaction). In (D-E) two-way ANOVA followed by Sidak's multiple comparison test, n ≥ 10 sections from four independently injected hemispheres per condition (for all six conditions); (I-K), two-way repeated measures ANOVA, n = 8 (I), 14 (J), and 9 (K) animals. ***p < 0.001, **p < 0.01, n.s. = not significant. Scale bar = 200 μm (A, B, F, G, H), or 50 μm (C).

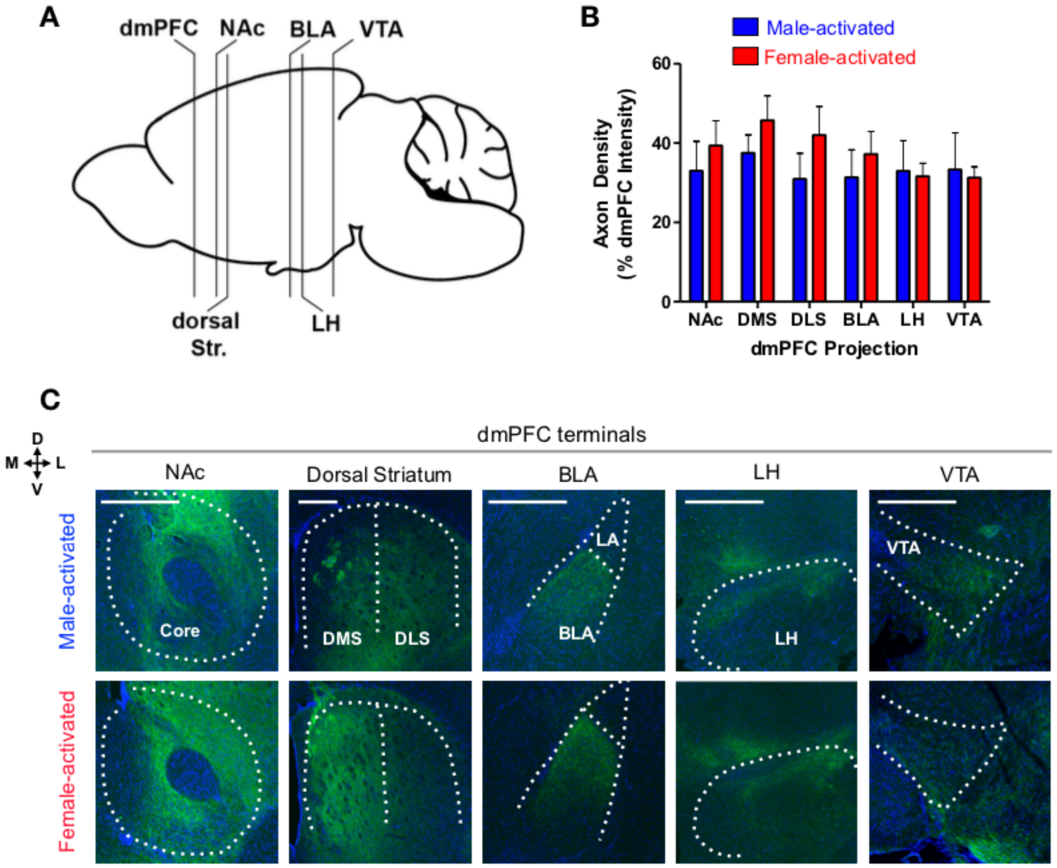

Figure 3.S6: Axonal projections of male- and female-activated neurons. **(A)** Sagittal view showing positions of brain regions imaged to quantify fluorescence intensity of axonal projections expressing ChR2-EYFP. **(B)** Quantification of fluorescent pixels per ROI for each condition (p = 0.20, region x cell interaction). The axon density (expressed as % dmPFC Intensity) represents the fraction of fluorescent pixels normalized to dmPFC intensity within each mouse. **(C)** Representative fluorescent images containing images of axon terminals in nucleus accumbens (NAc), dorsal striatum (dorsomedial: DMS and dorsolateral: DLS), basolateral amygdala (BLA), lateral hypothalamus (LH), and ventral tegmental area (VTA) from male-activated (top row) or female-activated (bottom row) mice. Scale bar = 500 μm.

3.7: References

1. Li, Y. & Dulac, C. Neural coding of sex-specific social information in the mouse brain. *Current Opinion in Neurobiology* **53**, 120–130 (2018).

2. Chen, P. & Hong, W. Neural Circuit Mechanisms of Social Behavior. *Neuron* **98**, 16–30 (2018).

3. Remedios, R. *et al.* Social behaviour shapes hypothalamic neural ensemble representations of conspecific sex. *Nature* **550**, 388–392 (2017).

4. Li, Y. *et al.* Neuronal Representation of Social Information in the Medial Amygdala of Awake Behaving Mice. *Cell* **171**, 1176-1190.e17 (2017).

5. Kim, D. W. *et al.* Multimodal Analysis of Cell Types in a Hypothalamic Node Controlling Social Behavior. *Cell* **179**, 713-728.e17 (2019).

6. Moffitt, J. R. *et al.* Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science (80-.).* **362**, (2018).

7. Kingsbury, L. *et al.* Correlated Neural Activity and Encoding of Behavior across Brains of Socially Interacting Animals. *Cell* **178**, 429-446.e16 (2019).

8. Liang, B. *et al.* Distinct and Dynamic ON and OFF Neural Ensembles in the Prefrontal Cortex Code Social Exploration. *Neuron* (2018). doi:10.1016/j.neuron.2018.08.043

9. Murugan, M. *et al.* Combined Social and Spatial Coding in a Descending Projection from the Prefrontal Cortex. *Cell* **171**, 1663-1677.e16 (2017).

10. Nakajima, M., Görlich, A. & Heintz, N. Oxytocin modulates female sociosexual behavior through a specific class of prefrontal cortical interneurons. *Cell* **159**, 295–305 (2014).

11. Wang, F. *et al.* Bidirectional Control of Social Hierarchy by Synaptic Efficacy in Medial Prefrontal Cortex. *Science (80-.).* **334**, (2011).

12. Nathanson, J. L., Yanagawa, Y., Obata, K. & Callaway, E. M. Preferential labeling of inhibitory and excitatory cortical neurons by endogenous tropism of adeno-associated virus and lentivirus vectors. *Neuroscience* **161**, 441–450 (2009).

13. Zhang, W. *et al.* BDNF rescues prefrontal dysfunction elicited by pyramidal neuron-specific DTNBP1 deletion in vivo. *J. Mol. Cell Biol.* **9**, 117 (2017).

14. Kim, C. K. *et al.* Molecular and Circuit-Dynamical Identification of Top-Down Neural Mechanisms for Restraint of Reward Seeking. *Cell* **170**, 1013-1027.e14 (2017).

15. McQuin, C. *et al.* CellProfiler 3.0: Next-generation image processing for biology. *PLoS Biol.* **16**, (2018).

16. Mathis, A. *et al.* DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nat. Neurosci.* **21**, 1281–1289 (2018).

17. Yao, S., Bergan, J., Lanjuin, A. & Dulac, C. Oxytocin signaling in the medial amygdala is required for sex discrimination of social cues. *Elife* **6**, (2017).

18. Beny-Shefer, Y. *et al.* Nucleus Accumbens Dopamine Signaling Regulates Sexual Preference for Females in Male Mice. *Cell Rep.* **21**, 3079–3088 (2017).

19. Pnevmatikakis, E. A. & Giovannucci, A. NoRMCorre: An online algorithm for piecewise rigid motion correction of calcium imaging data. *J. Neurosci. Methods* **291**, 83–94 (2017).

20. Mukamel, E. A., Nimmerjahn, A. & Schnitzer, M. J. Automated analysis of cellular signals from large-scale calcium imaging data. *Neuron* **63**, 747–60 (2009).

21. Cunningham, J. P. & Yu, B. M. Dimensionality reduction for large-scale neural recordings. *Nat. Neurosci.* **17**, 1500–1509 (2014).

22. Ghosh, K. K. *et al.* Miniaturized integration of a fluorescence microscope. *Nat. Methods* **8**, 871–8 (2011).

23. Chen, T. W. *et al.* Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* **499**, 295–300 (2013).

24. Luo, L., Callaway, E. M. & Svoboda, K. Genetic dissection of neural circuits. *Neuron* **57**, 634–60 (2008).

25. Luo, L., Callaway, E. M. & Svoboda, K. Genetic Dissection of Neural Circuits: A Decade of Progress. *Neuron* **98**, 256–281 (2018).

26. Lerner, T. N., Ye, L. & Deisseroth, K. Communication in Neural Circuits: Tools, Opportunities, and Challenges. *Cell* **164**, 1136–1150 (2016).

27. Wu, Y. E., Pan, L., Zuo, Y., Li, X. & Hong, W. Detecting Activated Cell Populations Using Single-Cell RNA-Seq. *Neuron* **96**, 313-329.e6 (2017).

28. Kawashima, T. *et al.* Functional labeling of neurons and their projections using the synthetic activity-dependent promoter E-SARE. *Nat. Methods* **10**, 889–895 (2013).

29. Franklin, T. B. *et al.* Prefrontal cortical control of a brainstem social behavior circuit. *Nat. Neurosci.* **20**, 260–270 (2017).

30. Levy, D. R. *et al.* Dynamics of social representation in the mouse prefrontal cortex. *Nat. Neurosci.* (2019). doi:10.1038/s41593-019-0531-z

31. McHenry, J. A. *et al.* Hormonal gain control of a medial preoptic area social reward circuit. *Nat. Neurosci.* **20**, 449–458 (2017).

32. Jennings, J. H. *et al.* Interacting neural ensembles in orbitofrontal cortex for social and feeding behaviour. *Nature* **565**, 645–649 (2019).

33. Ye, L. *et al.* Wiring and Molecular Features of Prefrontal Ensembles Representing Distinct Experiences. *Cell* **165**, 1776–1788 (2016).

34. Ramirez, S. *et al.* Activating positive memory engrams suppresses depression-like

behaviour. *Nature* **522**, 335–339 (2015).

35. Marshel, J. H. *et al.* Cortical layer-specific critical dynamics triggering perception. *Science (80-.).* **365**, (2019).

CHAPTER FOUR

Correlated Neural Activity and Encoding of Behavior Across Brains of Socially Interacting Individuals

4.1: Abstract

Social interactions involve complex decision-making tasks that are shaped by dynamic, mutual feedback between participants. An open question is whether and how emergent properties may arise across brains of socially interacting individuals to influence social decisions. By simultaneously performing microendoscopic calcium imaging in pairs of socially interacting mice, we find that animals exhibit interbrain correlations of neural activity in the prefrontal cortex that are dependent on ongoing social interaction. Activity synchrony arises from two neuronal populations that separately encode one's own behaviors and those of the social partner. Strikingly, interbrain correlations predict future social interactions as well as dominance relationships in a competitive context. Together, our study provides conclusive evidence for interbrain synchrony in rodents, uncovers how synchronization arises from activity at the single-cell level, and presents a role for interbrain neural activity coupling as a property of multi-animal systems in coordinating and sustaining social interactions between individuals.

4.2: Introduction

Social interactions involve some of the most complex decision-making tasks that animals must navigate to secure their survival and reproductive success [1], as individuals must integrate internal state with real time decisions of their social partners in a context-dependent manner. In interacting dyads, individuals thus become entrained as they attend to, predict, and react to each other's decisions (**Figure 4.S1A**) [2,3]. To date, social neuroscience has mostly focused on behavior in individual animals to interrogate the neural computations underlying social decision-making. But a full understanding of the social brain requires a broader picture that reflects the dynamic nature of social interactions, as well as the emergent neural properties that arise from multiple individuals as a single integrated system [1,4–6].

In recent years, much effort has been made to explore how neural systems coordinate across individuals engaged in social interaction. Simultaneous recordings from multiple human subjects using non-invasive techniques (e.g. functional magnetic resonance imaging [fMRI] and electroencephalography [EEG]), have revealed striking patterns of interbrain neural activity coupling during social engagement [7–10]. Despite these remarkable findings, little is concretely known about how interbrain synchrony arises from social interactions. Moreover, it remains unclear how synchrony emerges from individual neurons and neuronal populations, in part due to the limited spatial resolution of recording techniques in humans which cannot resolve single cell activity. It is also unclear whether brain synchrony is unique to primates, or whether it is a general phenomenon present in other social species.

Competitive interactions are common among social species, and play an important role in shaping social status hierarchies [11] which influence the long-term health of individuals [12–14]. Navigation of social interactions depends on circuitry in the medial prefrontal cortex (mPFC), which is implicated in the representation of social status [15–17] and shapes social and motivational states [18,19]. However, while previous work has shown that mPFC neurons are active during social interaction [20,21], it has not been clear how prefrontal ensembles encode behavioral decisions during real-time social engagements, such as social competition. Moreover, it is, to our knowledge, entirely unknown whether functional brain coupling arises during social interaction in rodents.

Here, we used microendoscopic calcium imaging to record from thousands of neurons in the dorsomedial prefrontal cortex (dmPFC) of pairs of mice engaged in social interactions. Our study provides conclusive evidence for interbrain activity correlations in interacting mice as well as a cellular level neural basis underlying this phenomenon, and identifies a critical role for interbrain synchrony in coordinating and facilitating social interaction.

4.3: Materials and methods

4.3.1: Experimental model and subject details

All experiments were carried out in accordance with the NIH guidelines and approved by the UCLA institutional animal care and use committee (IACUC). All subject mice were male C57BL6/J mice ordered from Jackson Laboratories at 8-10 weeks of age and 25-30 g of weight. Mice were maintained in a 12 h:12h light/dark cycle (lighted hours: 10:00 pm – 10:00 am) with food and water *ad libitum*. All mice were individually housed for three weeks prior to imaging and behavior experiments. All experiments were performed during the dark cycle of the animals.

4.3.2: Viral injections and GRIN lens implantations

For all surgical procedures, mice were anaesthetized with 1.0 to 2.0% isoflurane. We bilaterally injected 300 nl (on each side) of AAV1.Syn.GCaMP6f.WPRE.SV40 virus (titer: 4.65×10^{13} GC per ml, Penn Vector Core) at 30 nl min^{-1} into the dorsomedial prefrontal cortex (dmPFC; also prelimbic cortex, PL) using the stereotactic coordinates (AP: +2.0 mm, ML: ± 0.3mm, DV: -1.8mm to bregma skull surface). 30 minutes after injection, a 1.9mm diameter circular craniotomy was centered at the coordinates (AP: +2.0 mm, ML: 0.0 mm), and the GRIN lens (Edmund Optics; 1.8mm) was implanted above the injection site at a depth of -1.6mm ventral to the bregma skull surface and secured to the skull using super glue and dental cement. Mice were given one subcutaneous injection of Ketoprofen (4mg/kg) on the same day of surgery and Ibuprofen in drinking water (30mg/kg) starting on surgery day for 4 days. Mice were individually housed after surgery for two weeks. Then, the microscope together with a plastic baseplate were placed on top of the lens. We adjusted the position of the microscope until the cells and blood vessels appeared sharp in the focal plane and secured this position using dental cement. Left and right dmPFC were counterbalanced when choosing the field of view. The subjects included two mice that received a unilateral viral injection and were implanted with a 1 mm GRIN lens (Edmund Optics) above the right dmPFC. All mice were handled and habituated for at least 4 days before experiments. We did not observe any alterations in self-directed or social behavior in implanted animals.

4.3.3: Histology

Three weeks after imaging experiments, mice were transcardially perfused with 4% paraformaldehyde (PFA), followed by 24 h post-fixation in the same solution. 60-μm coronal sections were obtained using a cryostat. Finally, sections were stained with DAPI (1:5,000 dilution) and mounted on slides. Images were acquired using a Nikon A1 confocal microscope to confirm the position of lens implantation and GCaMP6f expression.

4.3.4: Behavior assays

4.3.4.1: Free social interaction in the open arena

Two novel male mice were simultaneously placed in an open arena (32 x 20 cm) which allows for free social interaction. During each imaging session (10-15 minutes), calcium fluorescence videos from both animals and their behavior were simultaneously recorded using microendoscopes and a video camera, respectively. The microendoscopes were connected to a digital acquisition device (DAQ) through a flexible, ultra-light coaxial cable. The long cable length prevented cables from becoming tangled during interaction between animals, ensuring that the social interaction was not affected by the presence of cables. For each pair, the social interaction assay was followed by a 10-minute separation assay (without removing the microscopes) where a solid, opaque board was inserted at the midline of the arena to prevent subjects from engaging in social interaction. All animals were habituated to being exposed to the open arena individually and to wearing the miniature microscope for at least 4 days before experimentation. We imaged from 10 pairs of animals that naturally displayed a high level of mutual social interaction (>15% of total time) and 9 pairs of animals that displayed a low level of mutual social interaction (<15% of total time). Here, mutual social interaction is defined as moments when both animals engaged in social behavior. A total of 8 implanted animals were used. Pairs that naturally displayed high levels of social interaction were used in further analyses of neural dynamics in Figures 4.1-4.2 and 4.S2, except in Figure 4.S2F. When we recorded from pairs of animals that naturally displayed lower levels of social interactions, relatively lower interbrain correlation was observed (Figure 4.S2F), consistent with the notion that interbrain synchrony depends on ongoing social interaction.

4.3.4.2: Competitive social interaction in the tube test

Animals were placed in a closed acrylic tube (length 60 cm; circumference 2.5 cm) with a 1.1 cm channel cut lengthwise down the center to allow movement of the head-mounted microscope. The microendoscopes were connected to a DAQ through a flexible, ultra-light coaxial cable. During each imaging session (12-15 minutes), subjects faced a novel male conspecific and were permitted to freely engage the opponent mouse by approaching, pushing, or retreating. Tube tests have previously been implemented using shorter tubes with individual trials lasting only seconds [16]. Here, the longer tube (60 cm) allowed us to perform longer sessions in order to permit each animal ample time to exhibit its full range of volitional behavior, and to respond dynamically to its opponent over the course of the encounter. Each session was broken into 2-5 trials, and the same pair of animals were manually reset to their respective end of the tube prior to each trial. All animals were habituated to engagement with a (different) novel male conspecific while wearing the miniature microscope for at least 4 days before behavior experiments. 18 pairs of mice were imaged using a total of 13 animals, and all pairs were used in further analyses of neural dynamics.

For simultaneous recording with and without social contact, 10-minute imaging sessions were performed in 13 pairs of mice using 6 animals (from the same cohort that were used in the other tube test experiments), immediately followed (without removing the microscope) by another 10-minute session after introduction of a translucent plastic separator in the center of the tube. Animals were free to move at will but were not in physical contact with one another.

4.3.5: Analysis of animal behavior

For both the open arena and the tube test experiments, behavior videos were recorded with a video camera at 20 frames per second (fps) and manually annotated frame by frame to identify onset and offset times for behavior of both animals. Behavior annotations were converted into a binary vector for each type of behavior that denotes precisely when animals are engaged in behavior ("1" indicates presence of a given behavior, and "0" indicates absence of that behavior). Epochs when animals engaged in no observable behavior or movement were considered to be "rest" epochs. During the "rest" epochs, the animal could observe the interacting partner, but was not actively behaving.

For the open arena experiments, a total of 15 social and non-social behaviors were annotated. Social behaviors included attacking, approaching, chasing, escaping, sniffing, social-grooming, defending, and mounting. Non-social behaviors included running, self-grooming, digging, exploration, rearing, climbing, and nesting. The level of social interaction for each pair was measured using the percentage of total time that both animals were engaged in social behaviors. 19 pairs of animals were used for basic behavior analyses shown in Figures 4.1B-D.

For the tube test experiments, the positions of both animals were tracked automatically using a supervised learning algorithm. For position tracking, we employed YOLOv2 (You Only Look Once), a convolutional neural network (CNN) framework optimized for high accuracy object detection [22]. We trained the CNN to detect and report bounding boxes around mice in each frame based on hundreds of example images. Accuracy for the automated tracking algorithm was confirmed by comparing the detected mouse positions with ground truth assessments in random samples of movie frames (>99% accuracy, Figure 4.S3A). For this analysis, individual scorers were blind to the identities of pairs and mice. Position vectors denoting the coordinates of each mouse were extracted and normalized to the length of the tube to obtain the relative tube position of each animal on a range from 0 (the starting end) to 1 (the opponent's end).

In order to quantitatively assess the relative dominance levels of each animal within each pair, we calculated their average position in the tube over the entire session. Previous reports have associated push and retreat behavior, as well as winning in the tube test, with overall social dominance status among male mice [17]. Because positional changes in the tube test correspond to gains or losses of territory that result from approach, push, or retreat behavior, each animal's average tube position can be considered as a measure of its overall dominance level within the pair. We confirmed that animals defined in this way (one dominant and subordinate in each pair) had significantly different average tube positions (Figure 4.3F). To assess whether dominant and subordinate animals exhibited different levels of push, retreat, and approach behavior, we compared dominant and subordinate animals in pairs that displayed large differences in dominance (having a tube position difference greater than 20% of the length of the tube) (Figures 4.3K-M). Indeed, pairs with large differences in tube position exhibited significantly different levels of push, retreat, and approach behavior, suggesting that the tube position metric corresponds to meaningful differences in behavioral repertoire that are consistent with previous studies.

4.3.6: Extraction of calcium signals

4.3.6.1: Motion-correction and preprocessing

During behavior experiments, calcium fluorescence videos from both animals were simultaneously recorded using customized miniature microscopes (UCLA miniscope) at 30Hz through custom-written data acquisition software. Raw videos from each imaging session were first processed using a MATLAB implementation of the NoRMCorre algorithm to correct for motion-induced artifacts across frames [23]. In order to normalize image frames prior to cell sorting, (*F-F0)*/*F0* (*ΔF*/*F*) was applied to each frame, where *F0* was the de-trended mean image from the entire movie. *ΔF*/*F* normalized videos were de-noised using an FFT spatial band-pass filter through a custom-written script in ImageJ (U.S. National Institutes of Health), and spatially down-sampled by a factor of 2 prior to ROI identification and cell sorting.

4.3.6.2: Segmentation and ROI identification

In order to identify putative cell bodies for extraction of neural signals, we employed an automated ROI detection algorithm that uses principal (PCA) and independent component analysis (ICA) to extract spatial filters based on spatiotemporal correlations among pixels [24]. Independent components were manually inspected to remove components that did not represent cell bodies, and binary thresholding was applied to remove contributions from pixels outside the bounds of putative neurons. Spatial filters were then applied to the *ΔF/F* movie to extract the calcium traces. All traces from recorded cells were manually inspected to ensure quality signals. Specifically, putative neurons that had abnormally shaped cell bodies (abnormally large or small), or that had calcium transients with low signal-to-noise ratio (< 2 standard deviations above the mean) were excluded from further analysis. Less than 5% of all putative neurons were removed based on these criteria. This approach ensured that the cells we included in our analyses had signal that reflected real neural activity and was robust enough for downstream analyses.

For open arena experiments, a total of 7535 (mean ± SEM = 198 ± 5) single neurons were analyzed. For tube test experiments, a total of 6728 (mean ± SEM = 187 ± 10) single neurons were analyzed. Here, a single neuron refers to one calcium trace extracted from an ROI, identified as described above, from one recording session.

4.3.7: Analysis of single cell responses during behavior

Prior to downstream analysis, all *ΔF/F* calcium traces were z-scored and are presented throughout in units of standard deviation (s.d.) unless otherwise specified. Responses of single neurons during behavior events (push, retreat, and approach) were quantified using an ROC (receiver operating characteristic) analysis, a commonly used approach that has previously been applied to calcium imaging data to characterize neural responses during social investigation (e.g. Li et al., 2017). Upon application of a binary threshold to the *ΔF/F* signal and comparison with a binary event vector denoting behavior bouts, behavior event detection based on neural activity can be measured using the true positive rate (TPR) and the false positive rate (FPR) over all time-points. Plotting the TPR against the FPR over a range of binary thresholds, spanning the minimum and maximum values of the neural signal, yields an ROC curve that describes how well the neural signal detects behavior events at each threshold. We used the area under the ROC curve (auROC) as a metric for how strongly neurons are modulated by each behavior. For each neuron/behavior category (for both subject and opponent behaviors), the observed auROC was compared to a null distribution of 1,000 auROC values generated from constructing ROC curves over randomly permuted calcium signals (that is, traces that were circularly permuted using a random time shift). A neuron was considered significantly responsive ($\alpha = 0.05$) if its auROC value exceeded the 95th percentile of the random distribution (auROC < 2.5th percentile for suppressed responses, auROC > 97.5th percentile for excited responses). Throughout, “neutral cells” refer to neurons that were not identified as responsive during subject or opponent behaviors.

While the significance of the auROC values for single cells can be analytically determined by performing a Mann-Whitney U test, the test statistic from the U test carries a caveat of being highly influenced by group sample sizes. Because of the kinetics of the calcium fluorescence signals, treating individual frames (sampled here at 30Hz) as independent samples for a U test would inappropriately inflate the power of the statistical test. Instead, we chose to use the permutation-based resampling method described above in order to test for statistical significance, as this approach is not sensitive to this particular sampling issue.

For comparison of response characteristics across subject and opponent cells (Figures 4.S8B-C), the response strength for each neuron and each behavior was calculated as the average z-scored *ΔF/F* activity during all behavior epochs of a given type. Response probability for each neuron and each behavior was calculated as the percentage of behavior events with average neural activity that exceeded 110% of the local baseline (increased by more than 10% above baseline), taken over the 10 seconds preceding behavior onset.

In order to ensure that opponent cell responses to opponent behaviors were not contaminated by activity associated with overlapping subject behavior, we analyzed the mean activity of opponent cells during isolated subject and opponent behavior bouts (Figures 4.7H-J). For this analysis, all events that overlapped across subject and opponent (within 2 seconds) were removed. We confirmed that subject animals did not display observable behavior, and did not exhibit changes in movement along the tube, during opponent behaviors used for this analysis (Figure 4.3F-G).

For analysis of cells responding during opponent behavior in the open arena assay (Figures 4.S8D-E), ROC analysis was performed using binary behavior vectors denoting all pooled social behaviors from the opponent that do not overlap with subject behavior and rest. Observed auROC values were compared with null distributions based on randomly permuted calcium traces (as described above, $\alpha = 0.05$). The mean activity of open arena opponent cells was computed over non-overlapping subject and opponent behavior, or baseline epochs.

Mean activity of opponent cells in the tube test was found to be significantly higher during social contact than after introduction of a separator to abolish contact (Figure 4.S8H), suggesting dependence on social context and interaction with another individual for opponent cell firing.

4.3.8: Analysis of population dynamics during behavior

4.3.8.1: Principal component analysis

To visualize population responses during social behavior, we applied principal component analysis (PCA) to obtain components that capture the covariance of the neural population during behavior events [26]. After binning neural traces into 1-second bins, trial-averaged responses were computed over a time window of 40 seconds (20 seconds prior to 20 second after event onset) for each neuron/behavior event, and concatenated across event types (e.g., approach, push, and retreat). Responses for each neuron were formed into a matrix which was used to perform PCA. Population vectors were then averaged over individual behavior bouts and projected onto the first

2 principal components for visualization (Figure 4.5K). For comparison of population responses to different behavior types (Figure 4.5L), we calculated the pairwise Euclidean distances between PC-projected population vectors (using the first 3 principal components) within or across different behaviors.

4.3.8.2: Mahalanobis distance

In order to visualize population response dynamics during behavior, we used the Mahalanobis distance, which provides a measurement of the separation between two population vectors while accounting for the covariance structure of the underlying distribution. This provides a way to quantify the strength of specific population response patterns, as opposed to simply measuring the average response of all neurons (Figure 4.S6A; see Li et al., 2017; Remedios et al., 2017). Average population vectors were constructed over frames from different behavior categories or over all baseline frames. The Mahalanobis distance between two vectors is computed as:

$$D_{\mathrm{MAH}}(x_1, x_2) = \sqrt{(x_1 - x_2)^{\mathrm{T}} S^{-1} (x_1 - x_2)}$$

where x_k is the mean population vector over all frames for event type k, and S is the covariance matrix computed over all baseline frames. For population response time-courses (Figure 4.5I), the Mahalanobis distance was measured between individual frame population vectors from a given class k and the average population vector over all baseline frames.

4.3.9: Behavior decoding based on population activity

In order to measure population-level encoding of social behaviors among dmPFC neurons, we constructed statistical models to predict behavior events based on population activity. For classification of individual behaviors, we used binary Fisher's linear discriminant (FLD) classifiers, and to distinguish between behavior types, we used a multi-class (3-way) Fisher's discriminant.

For all classifiers, training sets were constructed using population vectors during behavior bouts and negative training data was sampled from baseline (rest) frames. In order to measure the performance of FLD models, we split the data into training and tests sets and performed cross-validation. For each cross-validation fold, the test set represented 10% of the data drawn from 10 uniformly distributed 1% segments, and the remaining 90% training set was used to construct the model. For each fold, model performance was measured using the area under the ROC curve (auROC) for test data projected onto the Fisher discriminant. Overall model performance for each animal/session was calculated as the average over 50 folds where the training and test sets were randomly redrawn. Models were compared with null models constructed using training data with randomly shuffled class labels. Sessions with fewer than 5 bouts of the modeled behavior were not considered for this analysis. For frame-by-frame classification and visualization of the FLD projection (Figure 4.5M), frames were sampled uniformly every second over the entire session and used to construct training data to fit models. Population activity over the session was then projected onto the discriminant, and class predictions for each frame were evaluated.

For multi-class decoding of push, retreat, and approach behavior (Figure 4.5O), 3-way FLD models were constructed from population data using behavior vectors to define class labels, and cross-validation was performed as described above. Predictions were determined by taking the minimum Euclidean distance between test points and the mean of each class' training set after projection onto the first 2 FLD components. Performance for each fold was measured using the average accuracy for each class (weighted by the number of examples in the test set), and overall model performance was taken as the average over 50 folds (as described above). Models were compared with null models constructed using training data with randomly shuffled class labels.

For discrimination of subject vs. opponent behaviors (Figures 4.7K-L), behavior bouts within each animal were pooled together. Behavior frames that overlapped (concurrent subject and opponent behaviors) were removed from the analysis, and the remainder were used to construct training and test sets using the same cross-validation method as described above. Dimension reduction was first performed on the training data using partial least squares regression (PLS), and FLD components were computed from the training data after projection onto the first 10 PLS dimensions. For visualization (Figure 4.7K), population vectors from the test set from one example session/fold were projected onto the first two FLD components. For each model, ROC analysis was performed to quantify discriminability of subject and opponent population responses, and auROC values were averaged over the holdout partitions for each session. Overall model performance was quantified using the average of auROC values over all sessions (Figure 4.7L), and was compared with null models constructed using training data with randomly shuffled class labels.

4.3.10: Generalized linear models of single-neuron and population activity

4.3.10.1: Modeling neural activity across brains

In order to gain deeper insight into correlations of dmPFC neurons across animals in the tube test (Figure 4.6), we constructed Gaussian-residual generalized linear models (GLM) to express the mean activity of all neurons in one animal as a function of individual activities of neurons in the opponent. After binning calcium data from both animals into 1-second bins, GLMs were fit as:

$$\mu = X\beta + \varphi$$

where μ is the predicted mean activity in animal A, X is the matrix containing all normalized (to maximum) calcium traces from animal B, β is a vector of coefficients fit to each neuron in X, and φ is an error term. In order to validate the predictive power of GLMs, we performed 10-fold cross validation by withholding 10% of the data, sampled uniformly in 1% segments, from model fitting.

Full predicted activity traces were constructed by concatenating test predictions from each fold, and the overall performance of the model was evaluated using the Pearson's correlation coefficient (PCC) between the predicted activity μ and ground truth. Model performance was compared to the performance of null models constructed using randomly permuted calcium data—97.2% of the mean activity models individually exceeded chance levels (the 95th percentile of the null distribution). Cross-validated R^2 was also used as an alternative performance metric to confirm model significance and validated 97% of the mean activity models. Coefficients β from full models were z-scored before being pooled with those from other models (Figure 4.6G). For models of subpopulation activity (Figures 4.6H-M), the response variable μ was the mean activity of the top 15 behavior-excited neurons based on their rank-ordered auROC values for a given behavior type, and z-scored coefficients were averaged within each session according to cell identity before comparison across sessions/groups.

4.3.10.2: Modeling neural activity using behavior behaviors

To analyze the contributions of subject and opponent behaviors to the activity of individual neurons, we constructed GLMs using the behaviors and positions of both animals. Single-neuron GLMs were fit using a Poisson model with a log link function:

$$\ln(\mu) = X\beta + \varphi$$

where μ is calcium activity from one cell and X is a matrix of behavior and position vectors. The use of a log link function for single neuron models was based on the assumption that a Poisson distribution best characterizes the calcium data used to fit the model, as has been made in previous studies [28]. Binary behavior vectors were smoothed with an exponential decay function (τ = 3 seconds). Position vectors for each animal were projected onto four Gaussian functions centered at four positions (P_1, P_2, P_3, P_4) that uniformly tiled the length of the tube. In total, 14 variables were used to model activity: 6 behavior vectors (corresponding to push, approach, and retreat for both animals) and 8 position vectors (corresponding to the four tube positions for both animals). Model performance was quantified following 10-fold cross validation using the Pearson's correlation coefficient (PCC) of predicted and observed activity, and was compared to a distribution of null models fit using randomly permuted calcium data. Models were only considered significant and used for downstream analysis if their performance exceeded the 95th percentile of the null distribution. Significance testing for individual coefficients (Figures 4.8B-C) was based on a likelihood ratio test ($\alpha = 0.05$) which compares model performance with the associated variable against a null model without it. For comparisons of coefficients between dominant and subordinate animals, coefficients were z-scored and averaged within each animal/session. Results of coefficient analyses shown in Figures 4.8C-D were also consistent with analyses performed with models identified using R^2 as a performance metric (Figures 4.S8K-L).

4.3.10.3: GLM model comparison

In order to examine whether interbrain correlations observed in the open arena and tube test experiments exceeded modulations that could be only explained by observable behavior variables, we compared the performance of mean activity GLMs fit using both animal's behavior ("behavior-only" Model 1) with the performance of models that also included mean activity from the opponent animal as an additional explanatory variable ("interbrain" Model 2) (Figure 4.2J; Runyan et al., 2017). For these analyses, GLMs were Gaussian residual models, behavior vectors were exponentially smoothed (τ = 3 seconds), and behavior vectors and calcium activity were binned into 1 second bins prior to model fitting. Model performance was measured using cross-validated PCC with 10-fold cross-validation, as described above. We measured the change in model performance upon inclusion of opponent activity as *(*Model 2 – Model 1)/Model 1 ("GLM performance difference" in Figures 4.2K and 4.S4G). Performance indexes were compared with those of models constructed using randomly-permuted opponent activity (behavior variables were not permuted).

4.3.11: Analysis of interbrain neural activity correlations

4.3.11.1: Correlation of neural activity across brains

Because previous hyperscanning studies have investigated correlations of aggregate, region-level activity patterns, we used the mean activity of all z-scored *ΔF/F* traces in each dmPFC population (mean *ΔF/F*, effectively their summed activity normalized by the number of recorded neurons) as a measure of overall neural activity. For both open arena and tube test experiments, interbrain correlations across mouse dyads were calculated using the Pearson's correlation coefficient (PCC) of the overall neural activity across the entire session. To fairly compare interbrain correlations across sessions with different durations (Figures 4.2F and 4.4J), we cropped traces to the duration of the shortest session (10 min and 0 sec for the open arena; 11 min and 16 sec for the tube test). Interbrain correlations for each pair were compared to the 95th percentile of random permutation null distributions (Figures 4.S4A-B). In order to confirm that changes in interbrain correlation when animals were separated were not due to changes in the autocorrelation of each signal, we also compared phase-randomized signals before and after separation in both the open arena (Figure 4.S2E) and tube test experiments (Figure 4.S4E). Phase-randomized surrogate signals (Figure 4.S4D) were computed by independently randomizing the phase of each Fourier component, which disrupts the temporal structure of the signal but preserves its mean, variance, and autocorrelation. For comparison of overall correlations with dominance relationships (Figure 4.8J), interbrain correlations were measured over the first 5 minutes of each session to ensure a high degree of social interaction during each epoch.

4.3.11.2: Cross-correlation of neural activity across brains

In order to gain more insight into the timescale at which interbrain correlations occur, we performed a cross-correlation analysis using the neural activity from interacting animals in both the open arena and the tube test. We calculated the correlation between *ΔF/F* activity traces with different time shifts, ranging from −2 minutes to +2 minutes, and plotted the correlation as a function of time lag (Figures 4.1M and 4.4N). For interacting animals in both experiments, the peak of the average cross-correlation occurred precisely at 0.0 seconds lag. For both experiments, we also compared the correlation at the peak with the correlation at baseline, assessed using ±60 or ±30 seconds lag (based on the cross-correlation functions shown in Figures 4.1M and 4.4N). Cross-correlations were compared with those of phase-randomized signals (described above) to confirm that structure in the cross-correlation is not due to autocorrelations in each calcium trace (Figures 4.1N and 4.4O).

4.3.11.3: Interbrain correlations among subsets of neurons

To determine the contributions of subject-encoding and opponent-encoding neurons to interbrain correlations, we calculated correlations across animals after removing different types of cells from each neural population based on functional identity (e.g., behavior-excited or behavior-suppressed). While removal of behavior-excited cells resulted in a decrease in interbrain correlations, removal of behavior-suppressed cells did not (Figures 4.6A, 4.S7A-B, and 4.7M, 4.S8F-G). Neutral cells were neurons that were not identified as either subject-encoding or opponent-encoding by the ROC analysis. For subpopulation analyses in Figure 4.7N and Figures 4.8I, 4.8K, and 4.8O, subsets of 25 cells from each animal were used to calculate interbrain activity correlations in order to control for differences in correlation that could result from unequal population sizes. Subsets of the top behavior-encoding were selected (with the largest auROC values) for modulation by subject or opponent behavior. Neutral cells were defined as described above, and were sorted (in ascending order) and selected by |auROC$_{sub}$ – 0.5| + |auROC$_{opp}$ –

0.5|, where $\text{auROC}_{\text{sub}}$ and $\text{auROC}_{\text{opp}}$ are the auROC values calculated from neural responses to pooled subject and opponent behavior, respectively. To assess the relative contributions of subject and opponent cells to interbrain correlations, we also removed fixed numbers (to ensure a fair comparison between subject and opponent cells) of subject, opponent, or neutral cells (ranging from 1 to 25) from each animal and computed interbrain correlations over the down-sampled populations (Figures 4.7O-P).

4.3.11.4: Relationship between interbrain correlations and behavior interaction

In order to examine whether interbrain correlations could predict behavior interactions, we compared the degree of correlation prior to behavior in one animal to the probability of behavior response from the interacting partner (Figures 4.8G-I). For each behavior event (pooled across behavior categories) in each tube test session, the PCC of interbrain activity across the two animals was taken over the 30 seconds prior to behavior onset. All behavior events with any behavior from the interacting partner starting in the 15 seconds prior to behavior onset were removed from the analysis to ensure that preceding correlations were not contaminated by preceding behavior bouts. For each range of PCC (e.g., 0.1 – 0.2), the probability of behavior response in the reacting animal was calculated by summing all behavior events from the reacting animal over 3 seconds following the onset of its opponent's behavior for all epochs associated with that PCC range, and then dividing by total the number of epochs.

4.3.11.5: Interbrain correlations during matched behavior epochs across animal pairs

In order to address whether interbrain correlations could be accounted for simply by concurrent behaviors, we compared correlations of mean activity across animals during single epochs (30 seconds) of concurrent behavior (e.g. interacting animals A vs. B), with behavior-matched epochs across pairs that did not interact (e.g. non-interacting animals A vs. C) (Figures 4.2G and 4.4K). Specifically, we identified all epochs in which two interacting animals displayed behavior that have concurrent onset times (within 3 seconds), and computed interbrain activity correlations over these epochs (A vs. B). Behavior epochs in one animal were then matched with behavior epochs in another non-interacting animal from a separate session (A vs. C), such that the behavior types and onsets were identical to those in the epoch from the interacting pair (A vs. B). Other types of behaviors immediately before and after the temporally aligned behavior were also matched, such that overall behavior transitions, as well as the onsets of the aligned behaviors, were the same. The associated interbrain correlations were then compared. For the analysis shown in Figure 4.S4F, a single behavior bout in one animal was matched and aligned with an equivalent behavior bout from a separate non-interacting animal, and if multiple behavior bouts of the same type occurred within short intervals (1 s), they were considered as one bout. No other behavior bouts occurred during the epoch. For these analyses, lower PCC values are expected as interbrain correlation is lower in shorter temporal windows (Figure 4.S2C and 4.S4C).

4.3.12: Quantification and statistical analysis

All analyses for this study were conducted using custom routines in MATLAB (Mathworks), and are described in the respective Method Details, Results, and Figure Legends. All bar plots with error bars represent mean ± SEM; all box and whisker plots represent the median, interquartile range (box), and 5th to 95th percentile (whiskers) of the underlying distribution, unless otherwise specified. For all statistical tests throughout, normality of the data and equal variance of groups were not assumed, and non-parametric (Wilcoxon rank-sum and signed-rank) tests were used for unpaired and paired group comparisons, respectively. Statistical significance was defined with $\alpha < 0.05$ using two-tailed tests. For comparisons of proportions of binary-valued variables, Fisher's exact test was used. For comparisons of behavior bout length and cell pairwise distance distributions, two-sample Kolmogorov-Smirnov tests were used. Resampling methods based on temporally-permuted calcium traces were used to assess significance of auROC values for behavioral modulation of neural signals and performance of GLM models. Statistical significance of FLD classifiers was assessed by comparison with null models constructed using training data with shuffled class labels. The sizes of mouse groups were not pre-specified and approximated those of previous work.

4.4: Results

4.4.1: Correlated neural activity across animals during free social interaction

During natural social encounters, animals exhibit a wide range of behavior that engage them in complex, often reciprocal interactions. To study neural dynamics across brains of socially interacting mice, we first examined naturally occurring behaviors during social interactions in an open arena, where two novel animals were permitted to freely interact (**Figure 4.1A**). We recorded the interaction using a video camera, and annotated behaviors of both animals frame-by-frame (**Figure 4.1B**). Across all sessions, we identified 15 types of behaviors that included both social and non-social behavior. While animals spent about 43% of the time engaged in observable behavior (**Figure 4.1C**), the majority of this (~66%) was social behavior directed toward the interacting partner (**Figures 4.1D, 4.S2A**). Thus, the open arena provides an unconstrained context where animals freely engage in highly diverse and naturalistic social interactions.

To investigate neural dynamics during the social interaction, we employed microendoscopic calcium imaging to simultaneously monitor activity from hundreds of dmPFC neurons in both individuals. To gain optical access to neurons below the cortical surface, we implanted a gradient refractive index (GRIN) lens above the dmPFC following injection of an AAV expressing the fluorescent calcium indicator GCaMP6f (**Figure 4.1E**). Lens placement and GCaMP6f expression were confirmed histologically (**Figures 4.1F-G**). Calcium fluorescence videos were processed using independent component analysis to identify putative cell bodies, which were used to extract calcium traces from single cells, expressed throughout as relative change in fluorescence (*ΔF/F*) (**Figures 4.1H-I**). We analyzed a total of 7535 dmPFC neurons in

19 pairs of animals engaged in open arena social interaction. Overall neural activity varied across different types of behaviors (**Figure 4.S2B**), suggesting that activity in the dmPFC is differentially modulated by social behavior.

To explore how dmPFC neural dynamics were related across individuals, we computed the mean activity of neurons in each animal as aggregate signals that reflect the overall activity of the population (**Figure 4.1J**), and quantified correlations of activity (Pearson's correlation coefficient, PCC) across dyads in each session. Strikingly, dmPFC populations displayed highly correlated activity across animals, which far exceeded chance levels (**Figures 4.1K-L, 4.S2C**). To examine the timescale of interbrain correlations, we measured the cross-correlation of dmPFC activity across animals (**Figure 4.1M**); these showed a clear peak at 0.0 s, indicating precise synchrony of interbrain activity. This interbrain correlation was not due to autocorrelations in each signal, as the cross-correlation structure was abolished when traces were phase-randomized (**Figure 4.1N**). Together, these results establish that animals engaged in free social interaction exhibit highly correlated dmPFC activity.

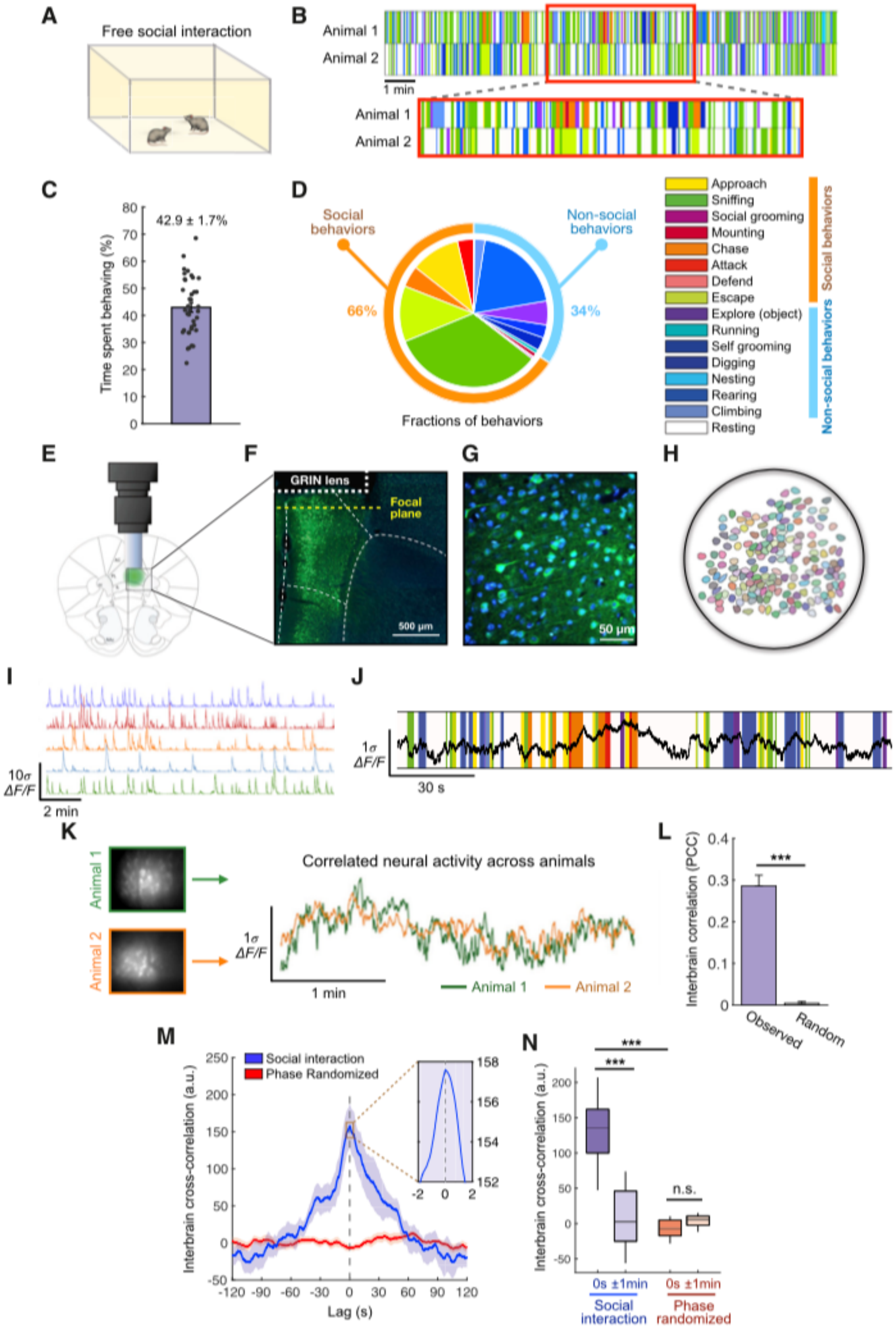

Figure 4.1: Correlated neural activity across brains of interacting animals during free social interaction. **(A)** Illustration of social interactions in the open arena. **(B)** Behavior raster plot of two animals interacting in the open arena. **(C)** Percentage of time animals engage in behavior in the open arena. Each dot represents one animal from one session. **(D)** Distribution of behaviors mice display in the open arena interaction. **(E)** Schematic of head-mounted microscope and GRIN lens implantation above dmPFC. **(F)** Example image of injection site showing expression of GCaMP6f in dmPFC**.** **(G)** Example image showing viral expression in dmPFC cell bodies. Green, GCaMP6f; blue, DAPI. **(H)** Example imaging field of view with individual cell bodies. **(I)** Example calcium

traces recorded from one session. **(J)** Example trace showing overall dmPFC activity (mean activity of all cells) in one animal during social interaction overlaid with behavior annotations. **(K)** Example calcium traces showing overall dmPFC activity from two animals engaged in social interaction. **(L)** Interbrain correlations of overall dmPFC activity in animals, compared with those of temporally permuted traces. **(M)** Cross-correlation of dmPFC activity traces from interacting animals compared with that of phase-randomized traces. **(N)** Quantification of cross-correlations shown in (M) at 0 s or ± 60 s.

4.4.2: Interbrain activity correlations depend on ongoing social interaction

Animals in a social environment are naturally inclined to engage with one another, but they occasionally exhibit periods of coordinated rest in which they are both quiescent (**Figure 4.S2D**). To address whether interbrain correlations could be simply explained by coordinated behavior/rest periods, we removed epochs in which both animals did not exhibit observable behavior and compared interbrain correlations during these epochs with those of full sessions. Activity after discounting periods of coordinated rest was as correlated as activity during full sessions (**Figure 4.2A**), suggesting that bouts of concurrent rest cannot account for activity correlations.

Although animals do not tend to exhibit the same behaviors at the same time (**Figure 4.S2D**), interacting animals do sometimes behave concurrently. To determine whether overall concurrent behavior could explain interbrain synchrony, we compared interbrain correlations during epochs with low vs. high levels of concurrent behavior (measured by correlation of overall behavior, **Figure 4.2B**). Again, interbrain correlations during these epochs were not different and were equally disrupted upon phase-randomization of activity traces.

To explore the relationship between interbrain synchrony and social interaction, we next compared the degree of interbrain correlation during social vs. non-social behavior. Correlations were significantly higher during social behavior (**Figure 4.2C**), suggesting dependence on social interaction. However, because animals are in the same environment, there is a possibility that correlated activity reflects shared sensory inputs such as ambient noise or lighting rather than social engagement. To rule this out, we separated the animals within the same physical environment using a barrier (**Figure 4.2D**). Abolishing social interaction significantly reduced interbrain correlations among dmPFC neurons (**Figures 4.2E-F, 4.S2E**), suggesting that correlated activity is not due to shared sensory input, but actually depends on ongoing interaction between the pair. Indeed, when we recorded from pairs of animals that naturally displayed low levels of social interaction, a lower degree of correlation was observed (**Figure 4.S2F**).

Given this observation, another hypothesis is that interbrain correlations reflect generic activity associated with social interaction, such as motivational state, regardless of whether animals are directly engaged. To rule this out, we examined neural activity across pairs of animals that each engaged in social interactions, but with separate animals and not with each other (**Figure 4.2G**). Activity correlations across animals from different sessions were significantly lower than those across interacting pairs (**Figure 4.2H**), confirming that directed engagement between two animals is necessary for interbrain coupling.

Moreover, it is possible that interbrain correlations could be purely explained by activity associated with individual coordinated behavior bouts at finer timescales. To address this, we computed correlations during epochs with coordinated behavior bouts, and compared them with correlations during behavior-matched epochs in non-interacting animals (**Figure 4.2I**). Activity from behavior-matched epochs across non-interacting pairs did not exhibit correlations; only those in socially interacting animals showed interbrain coupling (**Figures 4.2I**). This suggests that interbrain synchrony cannot be simply explained by overall concurrent behavior or individual coordinated behavior bouts, but depends upon the context of a direct, ongoing social interaction. For example, the same type of behavior may be associated with different patterns of activity depending on social context (e.g. interactions over a longer timescale or specific social relationships).

Lastly, to further understand the relationship between dmPFC activity and behavior, we modeled activity in each animal as a function of behavior and activity recorded from the interacting partner. We constructed generalized linear models (GLM) to model dmPFC activity from behaviors exhibited by both animals (**Figure 4.2J**, Model 1) and compared it to a second model fit using the overall activity from the interacting partner as an additional variable (Model 2). We reasoned that, if neural activity in one animal did not contain information relevant to activity in the interacting partner beyond what is explained by individual behaviors, models that included partner activity (Model 2) would not perform better than "behavior-only" models (Model 1). In fact, Model 2 performed significantly better than Model 1 (**Figure 4.2K**), suggesting that activity in one animal contains additional information about activity in the other that cannot be fully explained by moment-to-moment behavior. This is consistent with the notion that interbrain coupling depends on the larger context of an ongoing interaction.

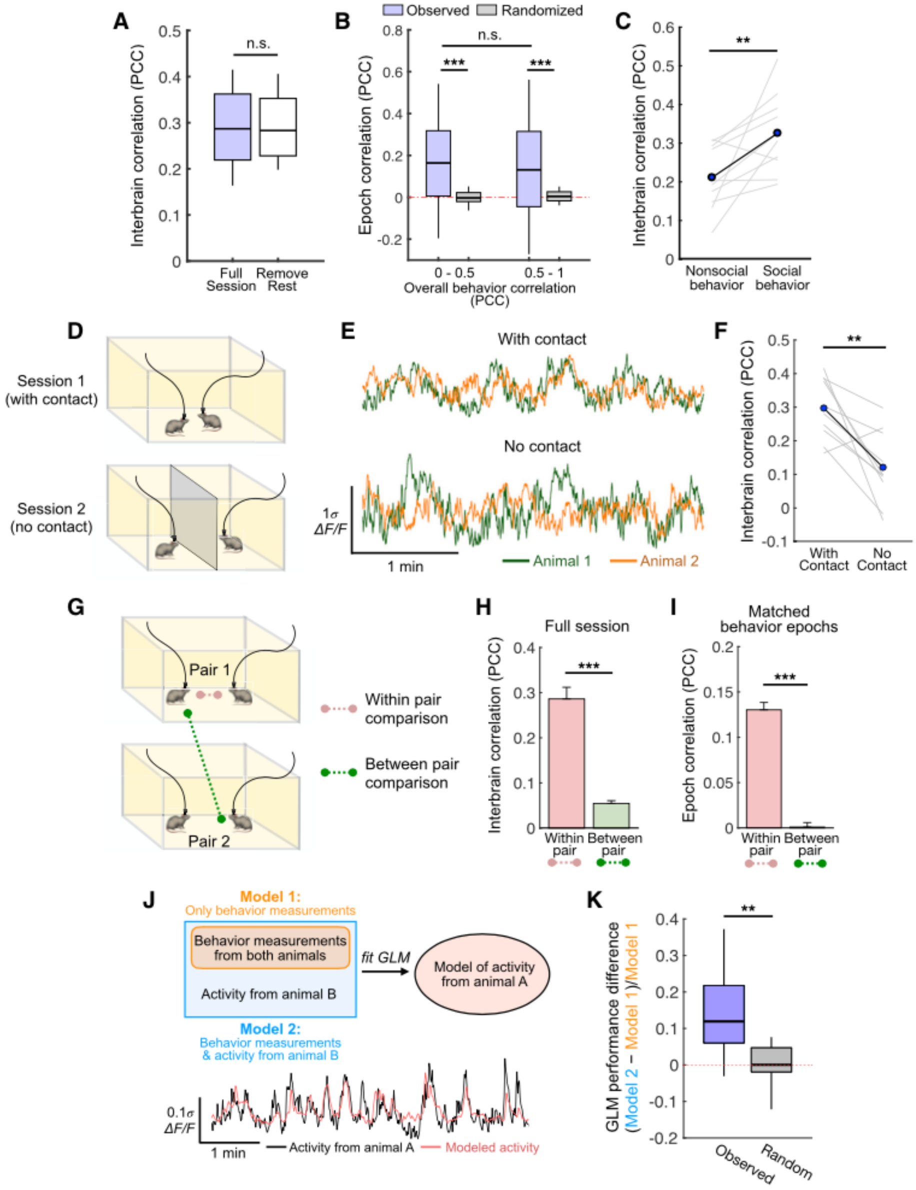

Figure 4.2: Interbrain correlations depend on ongoing social interaction. **(A)** Interbrain correlations of dyads during full open arena sessions or correlations after removing epochs of concurrent rest, defined as when both animals display no observable behavior. **(B)** Interbrain activity correlations during single epochs (1 min) with low or high behavior correlation (the PCC of binary vectors measuring the presence of any behavior), compared with correlations of phase-randomized signals. **(C)** Interbrain correlations of epochs when one or both animals engaged in social versus non-social behavior. **(D)** Schematic of the open arena interaction with social contact or with separation of animals with a barrier. Head-mounted microscopes were connected via an ultra-light cable that is long and flexible enough to prevent tangling during the course of social interactions. **(E)** Example calcium traces of overall dmPFC activity (mean activity) in a dyad with or without social contact. **(F)** Interbrain correlations in pairs with or without social contact. **(G)** Schematic showing comparisons of correlations across pairs engaged in social interaction (within

pair) and across animals that each interacted with a different animal (between pair). **(H)** Comparison between interbrain correlations across interacting or non-interacting pairs. **(I)** Interbrain correlations during single epochs (30 s) with concurrent behavior bouts in interacting pairs or those over behavior-matched epochs in non-interacting animals.

4.4.3: Behavioral dynamics during a competitive social encounter

To explore whether interbrain coupling was present in other contexts, such as competitive interaction, we adopted a social dominance assay (the tube test) that allowed us to examine competitive behavior and dominance relationships across dyads [30,31] (**Figure 4.3A**). In the tube test, mice are placed facing each other in a one-dimensional tube and allowed to push each other or retreat from conflict. Winning in the tube test (by pushing the other animal out of the tube) has previously been used to operationalize dominance behavior, as it correlates with other social status behavior in mice [17]. Compared to the open arena, the tube test also offers an advantage of narrowing the animals' decisions to a set of well-defined behaviors, enabling a precise interrogation of the relationship between interbrain synchrony and single cell encoding of behavioral decisions.

To analyze behavioral dynamics during the tube test, we recorded the interaction using a video camera, and developed an automated tracking algorithm using a convolutional neural network [22] to track the positions of both animals (**Figures 4.3B-C**), which we validated by unbiased visual assessment (>99% accuracy**; Figure 4.S3A**). We also manually annotated videos frame-by-frame to identify the onset and duration of behaviors in both animals. We observed that animals displayed three distinct types of behavior in the tube test: *approach*, a forward approach toward the opponent; *push*, a forceful push against the opponent sometimes resulting in forward movement; and *retreat*, a backward retreat away from the opponent. This parcellation, together with the position tracking, allowed us to examine how competitive interactions lead to gains or losses in territory for each animal.

On average, animals spent 23% of the time engaged in observable social interactions (**Figure 4.3D**), the majority of which (71%) was push behavior (**Figure 4.3E**). Although not all behavioral decisions lead to positional changes between the pair, position changes represent gains or losses of territory that result from competitive interaction. That is, each animal's position can be considered as a function of its individual decisions to approach, push, or retreat from conflict, thus characterizing its overall level of relative dominance. Within each pair, we identified the more dominant animal as the one who gained more territory on average, and confirmed that dominant and subordinate animals exhibited large differences in tube position (**Figure 4.3F**).

In any complex social engagement, reciprocal interaction is a common feature. Indeed, dyads behaved reciprocally in a fraction of the total time (**Figure 4.3G**), indicating that their behavioral decisions depended on one another. To examine how animals reacted to each other, we analyzed how the probability of each behavior in one animal changed following opponent behavior (**Figures 4.3H-J**). Overall, push behavior was followed by a probabilistic increase in retreat behavior in opponents, indicating that, while not all pushes result in opponent reactions, push and retreat behavior are sometimes linked (**Figure 4.3H**). There was also an increase in approach behavior following opponent retreats (**Figure 4.3I**), suggesting that animals were generally motivated to engage with their opponent.

Because dominants and subordinates exhibit similar levels of behavior overall (**Figure 4.S3B**), we reasoned that differences in tube position likely reflect differences in the distributions of displayed behavior. Indeed, dominants pushed more, retreated less, and approached more than subordinates (**Figures 4.3K-M**). We found no differences between the per-bout durations of behaviors displayed by dominants and subordinates (**Figures 4.S3C-E**), suggesting that differences in dominance (i.e. territory gained) depend mostly on the frequency of different social decisions.

Differences in overall dominance may also depend on how animals react to behavior from their opponent. To explore this, we constructed time-courses of animals' change in retreat probability following opponent push behavior. While both dominants and subordinates showed a probabilistic increase in retreats following opponent pushes, subordinates were more likely to retreat reactively (**Figures 4.3N-O**). Collectively, this analysis shows that outcomes of dominance encounters between mice depend not only on different behavioral choices in each animal, but also on how each animal responds to its opponent.

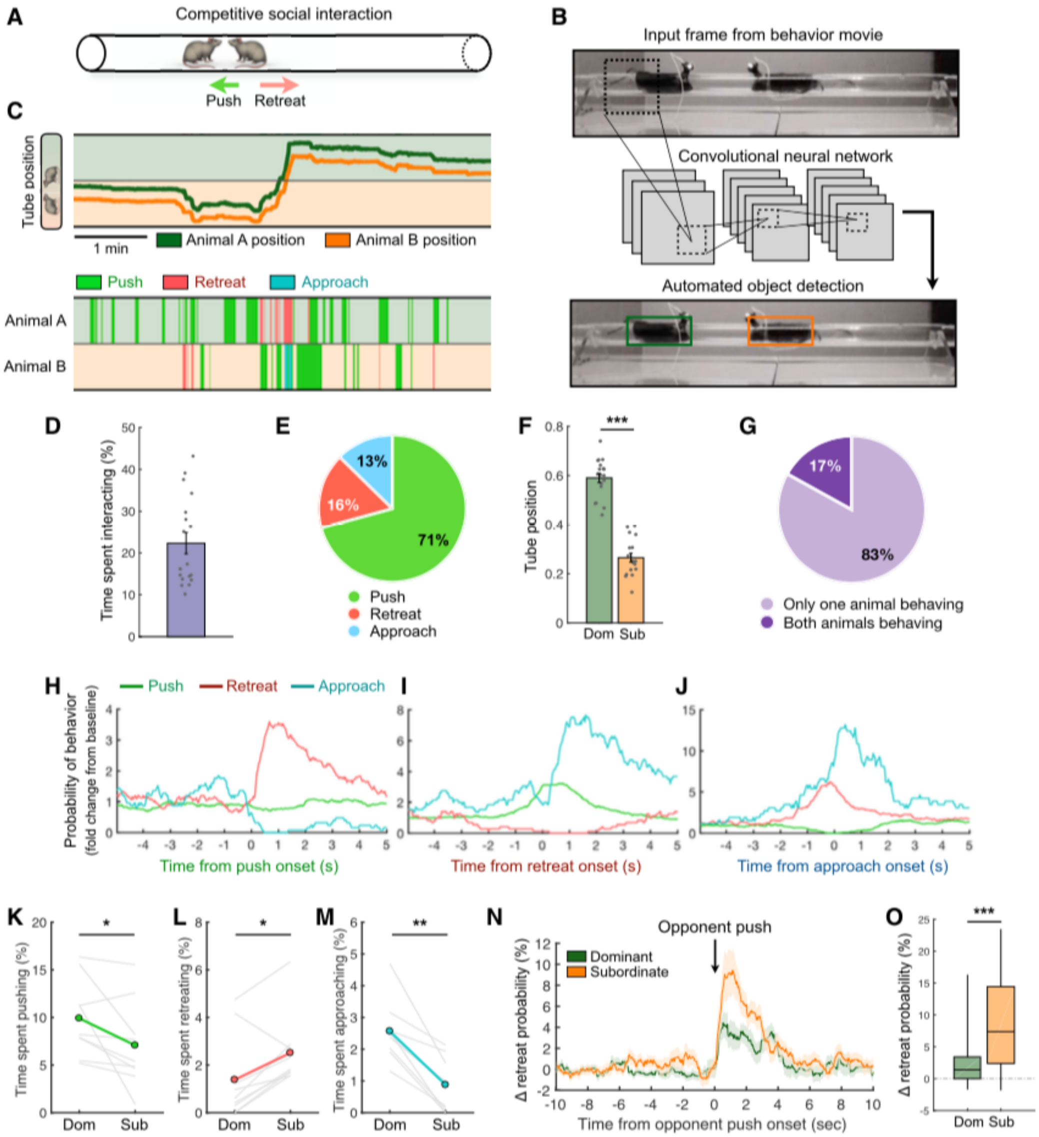

Figure 4.3: Dynamics of social behaviors during competitive interaction. **(A)** Cartoon of mice engaged in the tube test. **(B)** Illustration of the neural network used for automated tracking of mice. **(C)** Behavior annotations and position trajectories of a pair of mice in the tube test. **(D)** Total percentage of time animal pairs (either animal) engaged in social interaction. **(E)** Distribution of

time animals displayed different behaviors. **(F)** Average tube positions in dominant or subordinate animals. **(G)** Fraction of interaction time when only one or both animals are behaving. **(H–J)** Change in probability of opponent animal behavior with respect to subject animal push (H), retreat (I), or approach (J). **(K–M)** Percentage of time spent pushing (K), retreating (L), and approaching (M) in dominants or subordinates in pairs that displayed a large difference in dominance. **(N)** Change in relative probability of dominant or subordinate retreat behavior following opponent push. **(O)** Probability of retreat in dominants or subordinates 1 s following opponent push. ***$p < 0.001$, **$p < 0.01$, $p > 0.05$, n.s. (D, F, and N) Mean ± SEM.

4.4.4: Animals display interbrain correlations during social competition

To determine whether mice engaged in social competition also display interbrain coupling, we simultaneously imaged dmPFC activity using microendoscopes in animal dyads during the tube test (**Figure 4.4A**). As in the open arena, overall dmPFC activity was highly correlated across interacting animals in the tube test, far exceeding chance levels (**Figures 4.4B-C, 4.S4A-C**).

We first ruled out the possibility that correlated activity in this context is due simply to concurrent behavior or rest: neural activity correlations were consistently higher than correlations of overall behavior (**Figure 4.4D**), removing coordinated rest epochs did not reduce neural correlations (**Figure 4.4E**), correlations remained high when only one animal was behaving (**Figure 4.4F**), and activity correlations were higher than chance even during epochs with a lower level of concurrent behavior (**Figure 4.4G**). These suggest that interbrain coupling is not simply due to concurrent behavior or rest.

To confirm, as in the open arena, that interbrain coupling is not due to similar sensory inputs from a shared environment, we separated animals inside the tube so that both could freely move but could not interact (**Figure 4.4H**). Activity correlations were significantly reduced after separation (**Figures 4.4I-J, 4.S4D-E**), indicating that brain coupling in a competitive context also depends on ongoing interaction. In addition, comparisons of activity correlations in interacting vs. non-interacting pairs (**Figure 4.4K**) revealed that social engagement in the same encounter is necessary for correlated activity (**Figure 4.4L**). In support of this, while behavior epochs in interacting pairs had correlated activity, behavior-matched epochs in non-interacting pairs did not (**Figures 4.4M, 4.S4F**).

As in the open arena, dmPFC activity from interacting animals also exhibited peak cross-correlation at 0.0 s (**Figure 4.4N**), indicating that interbrain activity is precisely synchronized. However, the cross-correlation was disrupted upon phase randomization, and not significantly higher at zero time lag than at a lag of 30 s (**Figure 4.4O**), indicating a strong reduction in interbrain correlation.

Collectively, these results demonstrate that mice engaged in a competitive social encounter reliably display correlated activity across dmPFC neurons that depends on ongoing interactions in a larger social context and cannot be simply explained by overall concurrent behavior or individual coordinated behavior.

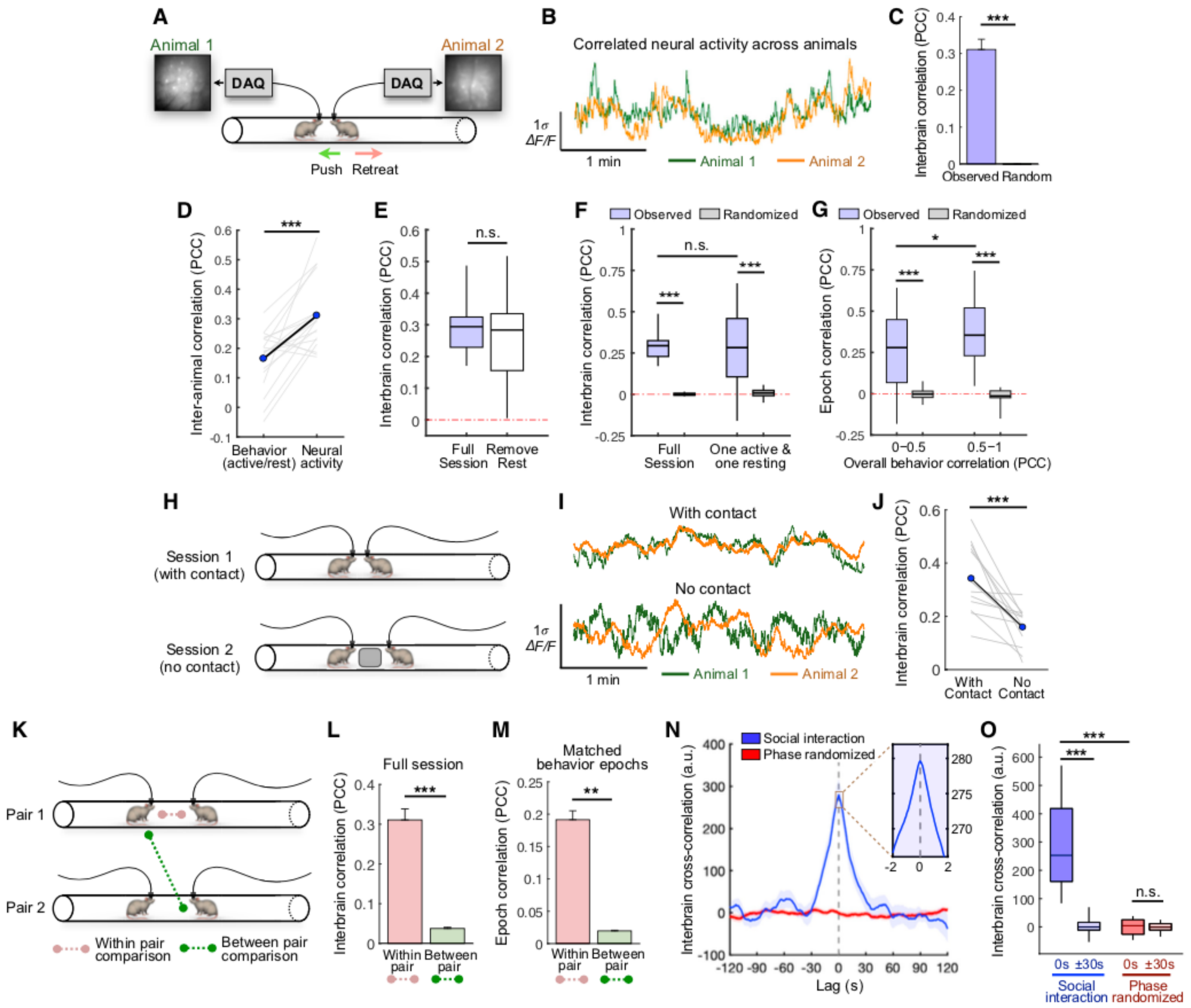

Figure 4.4: Correlated neural activity across animals during competitive social interaction. **(A)** Cartoon showing simultaneous imaging of two mice during the tube test. **(B)** Example traces of overall dmPFC activity (mean of all neurons) from two animals during the tube test. **(C)** Interbrain correlations in interacting pairs or correlations of randomly permuted traces. **(D)** Comparison of correlations of behavior (PCC of binary event vectors) across animals versus correlations of dmPFC activity. **(E)** Interbrain correlations during the tube test or after removing concurrent rest epochs when both animals display no observable behavior. **(F)** Interbrain correlations during tube test sessions or during epochs (R1 min) when one animal is behaving while the other is resting (displaying no observable behavior), compared with phase randomized controls. **(G)** Interbrain correlations during single epochs (1 min) of low or high overall behavior correlation (PCC of binary vectors measuring the presence of any behavior), compared with those of phase-randomized traces. **(H)** Schematic showing introduction of a separator in the tube test to abolish social contact. **(I)** Example traces showing dmPFC activity across two animals with or without social contact. **(J)** Interbrain correlations with or without social contact. **(K)** Schematic showing pairs engaged in social interaction (within pair) or pairs that each interact with a different animal (between pair). **(L)** Interbrain correlations across interacting or non-interacting animals. **(M)** Interbrain correlations during single epochs (30 s) with concurrent behavior bouts in interacting pairs or during behavior-matched epochs in non-interacting animals. **(N)** Cross-

correlation of dmPFC activity from pairs of mice in the tube test and that of phase-randomized controls. **(O)** Quantification of cross-correlations shown in (N) at 0 s versus ± 30 s. ***p < 0.001, **p < 0.01, p > 0.05, n.s. (C and L–N) Mean ± SEM.

4.4.5 dmPFC neurons encode distinct social behaviors during competitive interaction

Overall activity patterns of a brain region arise from individual cells, but a cellular-level basis for interbrain synchrony remains elusive. To explore how activities in single cells contribute to synchronous activity across animals, we first examined whether dmPFC neurons encode distinct social behaviors. dmPFC neurons as a whole exhibited time-locked excitation during push, retreat, and approach behavior (**Figure 4.5A**). However, this raises the question of whether behavioral decisions are associated with uniform activation of the dmPFC, or are encoded uniquely by distinct subsets of dmPFC neurons.

To address this, we examined whether single cells responded during specific behaviors. Using a receiver operating characteristic (ROC) analysis (**Figures 4.5B, 4.S5A**), we identified subsets of neurons that were excited or suppressed during push, approach, or retreat behavior (**Figures 4.5C-F, 4.S5B**). Of all recorded neurons, 29% encoded social behaviors (**Figure 4.5D**), and among these, ~76% showed selective tuning to specific behaviors. Cells that were not identified as behavior-encoding (hereafter referred to as “neutral cells”) were just as active, overall, as behavior cells (**Figure 4.S5C**), indicating that behavior encoding was due to specific time-locked responses. Interestingly, while behavior cells included both excited and suppressed ones, the majority were excited (**Figure 4.5E**). Overall, we found no differences in the spatial distributions of behavior cells compared with neutral cells (**Figures 4.5G-H**), indicating that behavior cells are spatially intermixed. These results demonstrate that a substantial fraction of dmPFC neurons selectively encode social behaviors in the tube test.

Information can be more robustly encoded at the population level than among single, highly tuned cells [32]. We next investigated whether neurons in the dmPFC formed stable activation patterns encoding social behaviors that could be read out at the population level. We examined how population response dynamics differed between types of behaviors using the Mahalanobis distance between behavior-evoked responses and baseline activity (**Figures 4.5I, 4.S6A**). Again, we found that all behaviors elicited time-locked responses. Interestingly, push and approach elicited stronger response patterns than retreat (**Figure 4.5J**), consistent with the idea that distinct behaviors are encoded differentially rather than as an aggregate of ensemble activity. To analyze the separability of population dynamics during behavior, we visualized population responses using principal component analysis (PCA); this revealed a clear separation of activity clusters based on behavior type (**Figures 4.5K, 4.S6B-D**). Further, the distance between different behaviors was significantly larger than within-behavior distances (**Figure 4.5L**), indicating that the separation of responses is not due to trial variability, but reflects unique patterns of activation that distinguish social behaviors.

Finally, to explore the robustness of behavior representations, we constructed decoders using Fisher's discriminant to predict the occurrence of behavior events based on population activity. Each behavior could be predicted by decoders (**Figures 4.5M, 4.S6E-G**), which significantly outperformed models constructed using randomized training data (**Figure 4.5N**). Moreover, multi-class decoders trained to predict specific behaviors among push, retreat, and approach achieved significantly higher performance than chance (**Figure 4.5O**), again indicating that neural representations are distinct and stable. Taken together, these results show that dmPFC neurons encode social behaviors at both the single-cell and population levels.

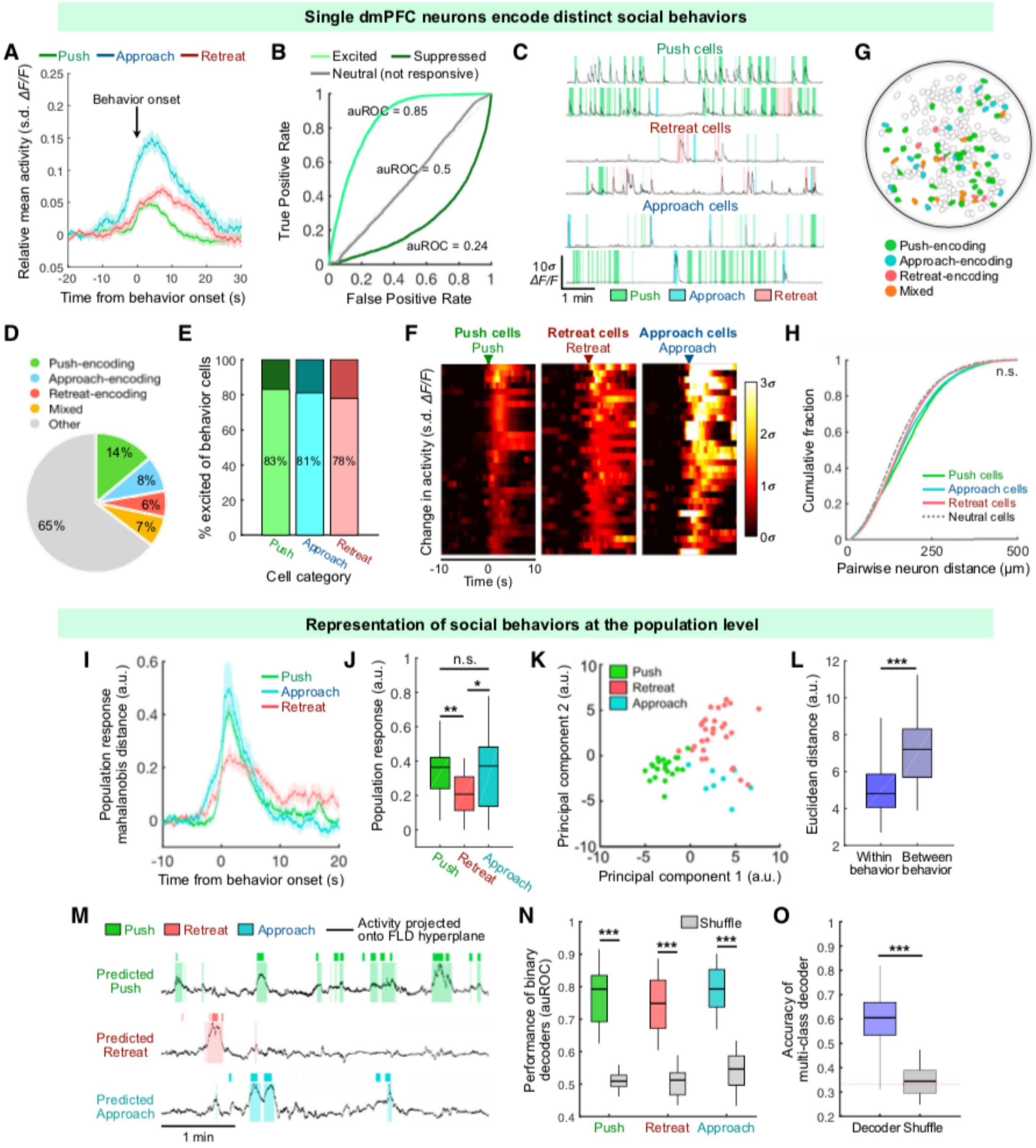

Figure 4.5: dmPFC neurons encode social behaviors during competitive interaction. **(A)** Mean trial-averaged response of dmPFC neurons (normalized to the 15 s preceding behavior) centered at onset of social behaviors. **(B)** ROC curves from example neurons for push behavior. **(C)** Examples of single cells that selectively encode different behaviors. **(D)** Distribution of behavior-encoding neurons. **(E)** Distribution of excited and suppressed cells within each behavior category. **(F)** Trial-averaged responses of example behavior cells. **(G)** Example field of view showing spatial distribution of behavior cells. **(H)** Cumulative fraction of pairwise distances among different subsets of behavior cells, compared with neutral cells (Kolmogorov-Smirnov test). **(I)** Population responses during behavior events (Mahalanobis distance between trial population vectors and baseline activity), averaged across sessions. **(J)** Population responses (as in I) during different behaviors over 3 s following behavior onset. **(K)** Principal component (PC) separation of

behavior-evoked population responses from one session; each dot is the mean response from one behavior bout. **(L)** Euclidean distance between PC-projected population vectors within or between behavior types, averaged within each session. **(M)** FLD decoders trained to predict different behaviors from rest using population activity. Plots: projections of population activity onto the linear discriminant; dark patches: annotated behavior; light patches: frame-by-frame predictions of example classifiers. **(N)** Performance of FLD decoders exemplified in (M), compared with models constructed using shuffled class labels. **(O)** Performance of 3-way multi-class FLD decoders trained to distinguish between push, approach, and retreat behavior. Red line: expected chance level in the three-way decoder. ***p < 0.001, **p < 0.01, p > 0.05, n.s. (A and I) Mean ± SEM. (A, C, and F) D*F*/*F* calcium traces are presented in units of SD.

4.4.6: Interbrain activity correlations depend on cells encoding social behavior

To determine how interbrain coupling depends on activity in individual cells, we next examined whether interbrain correlations arise from uniform dmPFC activation or specific subsets of cells (e.g. behavior cells). Removal of behavior cells resulted in a marked reduction in the activity correlation across animals (**Figure 4.S7A**), and this was driven specifically by behavior-excited cells, as removal of behavior-suppressed cells did not affect interbrain correlations (**Figures 4.6A, 4.S7B**). Moreover, interbrain correlations were equally disrupted upon removal of behavior cells in only one animal, indicating that brain coupling requires encoding of social information in both animals simultaneously. In contrast, removing neutral cells did not reduce activity correlations. This was not due to neutral cells being unresponsive, as their overall activity was as high as that of behavior cells (**Figure 4.S5C**). Instead, this suggests that correlated brain activity depends on subsets of cells encoding social information, rather than uniformly distributed neural dynamics.

Following this, we next examined correlations between specific subpopulations of behavior-encoding cells. Indeed, certain categories of behavior cells exhibited elevated interbrain correlations (**Figures 4.6B-D**). In particular, push-vs-retreat subpopulations were more highly correlated across animals than were neutral cells, consistent with our observation that these behaviors are sometimes coupled (**Figure 4.3H**). Interestingly, the synchronization of push and retreat cells was unidirectional across dyads, such that only push cells in dominants, but not in subordinates, were more correlated with retreat cells in the opponent. This suggests that interbrain correlations not only depend on specific subsets of cells, but that neurons encoding specific behavior interactions contribute preferentially to brain coupling.

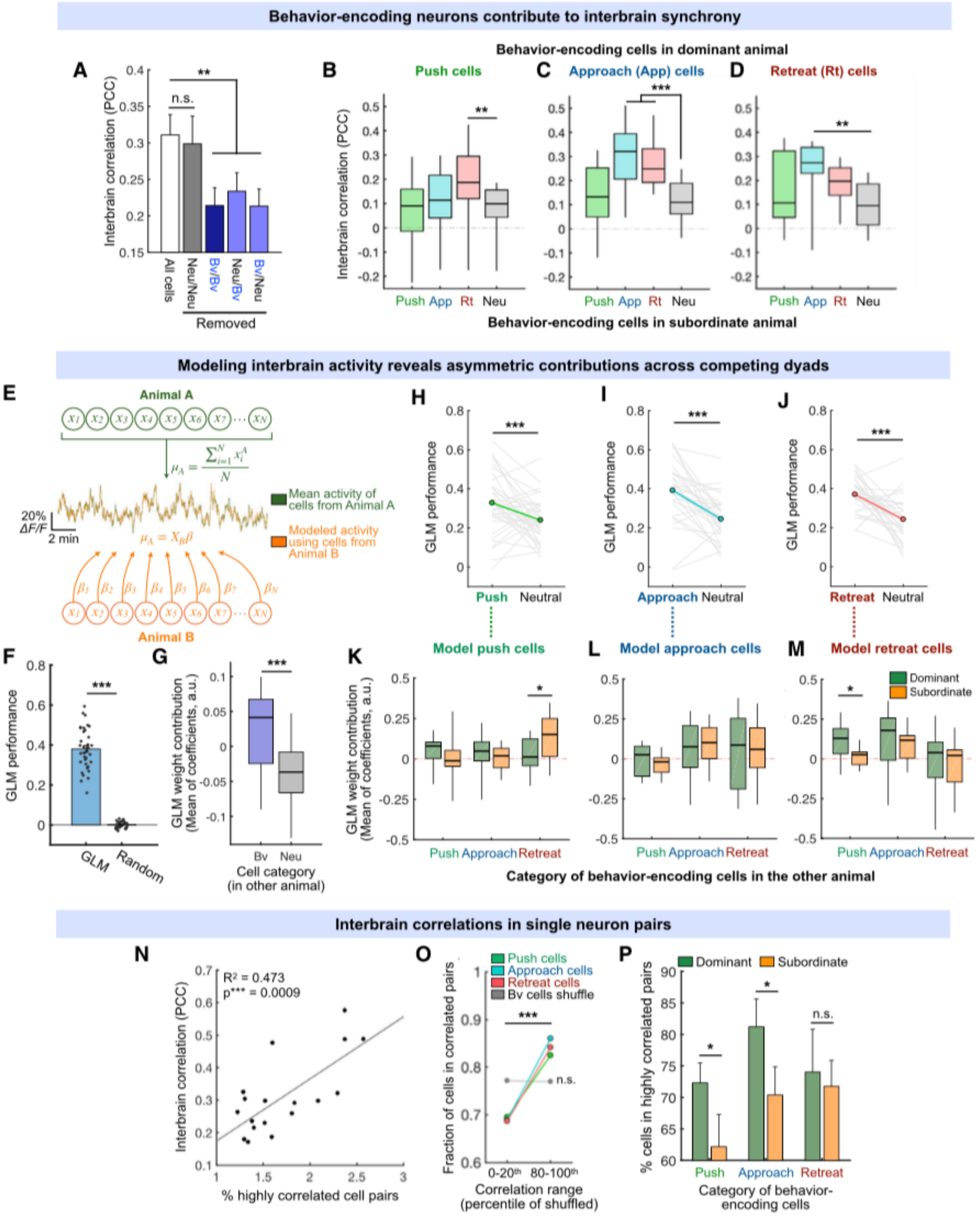

Figure 4.6: Interbrain correlations depend on neurons encoding one's own social behavior. **(A)** Interbrain activity correlations after removal of behavior-excited (Bv) or neutral (Neu) cells from both animals. **(B–D)** Interbrain correlations between the mean activity of subsets of push- (B), approach- (C), and retreat- (D) excited cells (the top 15 cells based on area under the ROC curve [auROC] values). **(E)** Schematic of models of interbrain activity across animals. The mean activity of all neurons in one animal (top) is modeled as a function of single-cell activities in the interacting partner (bottom) using a GLM. **(F)** Performance (cross-validated PCC) of GLMs to predict activity in one animal using single-cell activities from the other, compared with that of models using randomly permuted controls. **(G)** Weight contributions of behavior (Bv) and neutral (Neu) cells in GLMs of overall activity in (F), computed as the average of *Z*-scored coefficients fit to Bv or Neu cells in each model. **(H–J)** Performance of GLMs modeling the mean activity of subsets of push (H), approach (I), and retreat (J) cells, as in (B)–(D), compared with that of GLMs modeling the mean of neutral cells. **(K–M)** Weight contributions of behavior (push-, approach-, or retreat-excited) cells fit to models of push (K), approach (L), and retreat (M) cells in (H)-(J), computed as the average of z-scored coefficients for each cell type. **(N)** Correlation between the percentage of highly correlated single-cell pairs (>99th percentile of random distribution, see Figures 4.S7E and 4.S7F) and the interbrain activity correlation across pairs. **(O)** Fraction of behavior cells that belong to an interbrain cell pair with low (bottom 20% of random distribution) or high (top 20% of random distribution) correlations. **(P)** Fraction of behavior cells in highly correlated (>99th percentile of random distribution) cell pairs in dominants and subordinates. ***p < 0.001, **p < 0.01, p > 0.05, n.s. (A, F, and P) Mean ± SEM.

4.4.7: Interbrain activity correlations arise from single cell dynamics

To gain more insight into how interbrain correlations emerge from single dmPFC neurons, we constructed GLMs to express the overall dmPFC activity in one animal as a function of single cells in the interacting opponent (**Figure 4.6E**). These GLMs performed significantly better than chance (**Figure 4.6F**), suggesting that a weighted combination of individual cell activities in one animal could provide a good model of overall activity in the opponent. Moreover, behavior cells had significantly higher weight contributions in the models than neutral cells (**Figure 4.6G**), consistent with our results that interbrain correlations depend on behavior cells.

We next constructed GLMs using single cells from one animal to model subsets of behavior cells in the other (**Figure 4.S7C**), and found that these models performed significantly better than models of neutral cells (**Figures 4.6H-J, 4.S7D**). Examination of subpopulation models in dominants and subordinates revealed further asymmetries that mirrored unidirectional behavior interactions displayed by the dyads (**Figures 4.6K-M**): while the push-encoding population in dominants was best explained by subordinate retreat cells, the retreat-encoding population in subordinates was better modeled by dominant push cells. This further suggests that interbrain correlations in dmPFC arise from unique subpopulations in each animal that preserve individual differences in behavior.

Lastly, we investigated whether interbrain correlations were related to correlations between single pairs of cells across animals. Interacting animals contained more highly correlated cell pairs than expected by chance (**Figures 4.S7E-F**), and the fraction of highly correlated cell pairs in each dyad was itself correlated with the degree of overall brain coupling between them (**Figure 4.6N**), supporting the notion that correlated activity at the population level arises from subsets of single cells. Moreover, behavior cells were enriched among more highly correlated cell pairs (**Figure 4.6O**). In particular, in dominants, a larger fraction of push and approach cells were highly correlated with cells in subordinates, possibly reflecting a greater influence of behaviors of dominants on opponent responses (**Figure 4.6P**).

Taken as a whole, these results show that interbrain correlations in the dmPFC arise from specific subsets of cells encoding distinct behaviors in both animals, and reflect ensemble correlations that extend to the single-cell level.

4.4.8: Interbrain correlations depend on cells encoding behaviors of the social partner

The observation that interbrain coupling depends on subsets of behavior-encoding neurons raises the possibility that correlated activity could be completely explained by activity in these cells. However, our findings that (1) the degree of activity correlation consistently exceeds behavior correlations (**Figure 4.4D**), (2) activity correlations cannot be explained simply by concurrent or coordinated behavior bouts (**Figure 4.4M**), and (3) activity correlations persist when only one animal is behaving (**Figure 4.4F**), raise the alternative possibility that other information in the circuit also contributes to interbrain coupling. In particular, one hypothesis is that some correlated activity arises from subsets of dmPFC neurons that encode the behavior of the interacting partner.

To examine this hypothesis, we first asked whether any dmPFC neurons contained information about opponent behavior. Using ROC analysis, we identified a fraction of dmPFC neurons that responded specifically during opponent behavior, but not during subject behavior (**Figures 4.7A-B**), which constituted 8% of all recorded cells. On the other hand, 21% responded only during subject, but not opponent, behavior. We hereafter referred to neurons that only encoded opponent behavior as “opponent cells” and neurons that only encoded subject behavior as “subject cells.” Of the cells that were only responsive to opponent behavior (**Figure 4.S8A**), the majority (93%) responded selectively to single categories of behavior (**Figure 4.7C**), with response characteristics that were comparable to those of subject cells (**Figures 4.S8B-C**). Subject and opponent cells were spatially intermixed within the population (**Figures 4.7D-E**). Interestingly, we also identified a comparable fraction of cells that encoded behavior of the interacting partner during free social interaction in the open area (**Figures 4.S8D-E**), suggesting that behavior of social partners is encoded across multiple social contexts.

Opponent cells showed responses to specific opponent behaviors, but did not appear to respond during the subject’s own behavior (**Figures 4.7F-G**). To confirm that these cells were selectively active during opponent behavior, we compared their mean activity during opponent push, retreat, or approach with activity during subject behaviors (**Figures 4.7H-J**). Opponent cell activity during opponent behaviors (when the subject is not behaving or moving; **Figures 4.S3F-G**) was significantly higher than baseline, while activity during subject behavior was not, confirming that opponent cells selectively encode opponent behavior.

To further explore the population encoding of opponent behavior, we constructed decoders to classify the identities of subject vs. opponent behaviors (**Figure 4.7K**) and found that discrimination was significantly higher than chance levels (**Figure 4.7L**), indicating that neural responses during subject and opponent behavior form distinct population-level representations.

To test whether opponent cells also contribute to brain coupling, we next examined the effect of removing subsets of opponent cells on interbrain correlations. As with removal of subject cells (**Figure 4.6A**), removal of opponent cells, even in only one animal, markedly decreased correlated activity (**Figure 4.S8F**), an effect that was driven specifically by opponent-excited cells (**Figures 4.7M, 4.S8G**). Conversely, examining interbrain correlations only among subject and opponent cells, we found that they displayed even higher correlations than the whole population, and that replacing these with neutral cells in either animal abolished interbrain correlations (**Figure 4.7N**). Interestingly, we also observed that removing opponent cells had a stronger effect (~63% more) on reducing interbrain correlations than subject cells, suggesting that they contribute relatively more, cell for cell, to synchronized activity (**Figures 4.7O-P**).

Taken together, these results indicate that correlated brain activity depends not only on subject cells encoding one's own behavior, but also on a separate subset of neurons in each animal that encode the behavior of the interacting partner (**Figure 4.7Q**). As each brain represents a common behavior repertoire consisting of both animals' behavior, overall neural activity becomes synchronized across dyads. This offers an explanation for why interbrain synchrony cannot be fully explained by coordinated rest or concurrent behavior, and why it can be observed even when only one animal behaves.

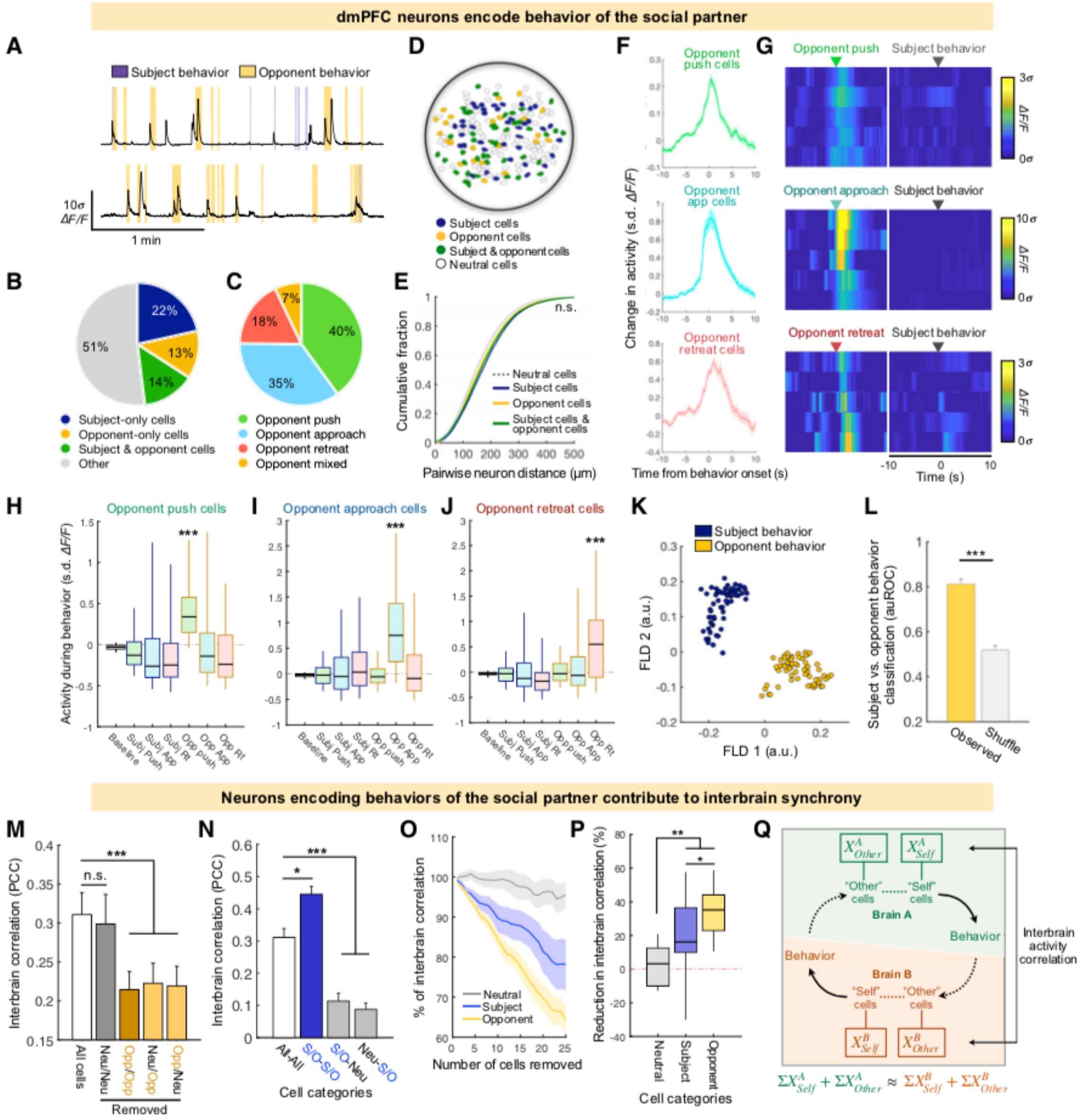

Figure 4.7: Neurons encoding behavior of the social partner contribute to interbrain correlations. **(A)** Example traces from dmPFC neurons that respond during opponent behavior. **(B)** Fraction of neurons that are significantly responsive during subject, opponent, or both types of behavior based on ROC analysis. **(C)** Distribution of opponent-encoding neurons that selectively respond during specific behaviors. **(D)** Example field of view showing the spatial distribution of subject and opponent cells. **(E)** Cumulative fraction of pairwise distances among different subsets of cells, compared with neutral cells (Kolmogorov-Smirnov test). **(F)** Trial-averaged responses of behavior-selective opponent cells. **(G)** Trial-averaged responses of example opponent push-, approach-, and retreat-excited neurons during opponent or subject behavior. **(H–J)** Mean activity of opponent push- (H), approach- (I), and retreat (J)-excited cells during each type of subject or opponent behavior. Behavior bouts that overlapped across subject and opponent were excluded to ensure that activity during opponent behavior was not

contaminated by subject behavior. During opponent behaviors used for this analysis, the subject animal did not exhibit any behavior or positional change (see Figures 4.S3F and 4.S3G). **(K)** Population responses during subject and opponent behavior (from a cross-validation test set) projected onto the first two FLD dimensions. **(L)** Performance of FLD decoders to distinguish between subject and opponent behavior based on population activity. **(M)** Interbrain activity correlations after removal of opponent-excited (Opp) or neutral (Neu) cells from both animals. **(N)** Interbrain activity correlations between subsets of subject and opponent (S/O) or neutral (Neu) cells (the top 25 cells based on rank-ordered auROC values). **(O)** Interbrain correlation upon removal of different numbers of subject, opponent, or neutral cells from each animal. **(P)** Reduction in interbrain correlation after removing 25 subject, opponent, or neutral cells from each animal, as in (O). **(Q)** Schematic showing that interbrain correlations arise from the collective contributions of neurons encoding subject and opponent behavior in both animals. As these neurons in each brain represent a common behavior repertoire (i.e., behavior of both animals), overall neural activity becomes synchronized across dyads. ***p < 0.001, **p < 0.01, p > 0.05, n.s. (F and L–O) Mean ± SEM. (A, F, and G) D*F*/*F* calcium traces are presented in units of SD.

4.4.9: Dominant animals exert a greater influence on interbrain correlations than subordinates

Next, to explore whether cells in dominants and subordinates encode subject and opponent information differently, we constructed GLMs to model the activity of each neuron as a function of the behaviors of both animals and their positions in the tube (**Figure 4.8A, 4.S8I**). Overall, ~30% of all cells in both dominants and subordinates were well-modeled (**Figure 4.S8J**), and the majority of these were significantly fit by only subject behavior, opponent behavior, or a combination of both (**Figure 4.8B**). Moreover, a subset of dmPFC neurons were only explained by opponent – but not subject – behavior, and a substantial fraction were fit with significant coefficients to specific opponent behaviors (**Figures 4.8C, 4.S8K**), again indicating that activity in some dmPFC neurons is selectively modulated by opponent behavior.

Intriguingly, models of cells in dominants placed higher weight on the subject's own behavior, whereas opponent behaviors had a stronger weight contribution to cells in subordinates (**Figures 4.8D, 4.S8L**). This indicates that while cells in dominants respond more to subject behaviors compared to cells in subordinates, cells in subordinates respond more to opponent behaviors compared to cells in dominants. This possibly reflects stronger engagement of attention in subordinates toward dominant animal behavior.

These observations led us to hypothesize that dmPFC neurons might exhibit stronger interbrain correlations when dominants behave compared to subordinates. To test this, we examined interbrain correlations during epochs when one animal, but not the interacting partner, was behaving. Strikingly, activity correlations were higher during dominant than during subordinate behavior (**Figure 4.8E**), suggesting that interbrain correlations are driven more strongly by dominant animals (**Figure 4.8F**).

4.4.10: Interbrain correlations predict social interactions and dominance relationships across dyads

The observation that dominant animals more strongly drive brain coupling suggests a more direct relationship between interbrain correlations and social interaction. To explore this more deeply, we first asked whether interbrain correlations could predict behavior interactions. We constructed time-courses of the probability of behavioral response in one animal as a function of time following partner behavior (**Figure 4.8G**). Decisions in one animal preceded by highly correlated activity were more likely followed by a behavioral reaction from the opponent. Moreover, the probability of behavioral response following partner behavior was positively correlated with the degree of synchrony preceding the interaction (**Figure 4.8H**), suggesting that correlated activity not only arises during social interaction but actually predicts future interactions. As expected, correlations among subsets of subject and opponent cells in each animal also predict future interactions (**Figure 4.8I**). However, this relationship was abolished when considering correlations with neutral cells (**Figure 4.8I**), again highlighting the dependency of activity synchrony on neurons encoding social information.

Given that the overall dominance relationship between animals is a consequence of individual social interactions, we hypothesized that the degree of activity correlation across a dyad, which predict their interactions, may reflect their difference in overall dominance levels. Using average tube positions of animals as a dominance metric (i.e., territory gained), we compared interbrain correlations across dyads with their difference in relative dominance. Strikingly, we observed a significant positive correlation across all pairs (**Figure 4.8J**). In particular, subsets of neurons encoding social behaviors of self and others significantly predicted differences in dominance behavior, while replacement with neutral cells in either animal abolished this relationship (**Figure 4.8K**).

Since brain coupling *predicted* future social interactions, we also asked whether correlations during only the initial phase of the encounter could predict dominance outcomes. Interestingly, the degree of interbrain correlation in just the first two minutes of each session predicted differences in dominance across the whole session (**Figure 4.8L**). Again, this relationship depended critically on behavior cells in both animals (**Figures 4.8M-O**). Despite this, the degree of overall behavior correlation in the first 2 minutes was unrelated to differences in dominance (**Figure 4.S8M**), suggesting that activity correlations may be a better predictor of dominance outcomes than behavior itself. Taken as a whole, these results demonstrate that activity correlations predict social interaction on timescales ranging from seconds to minutes, suggesting a functional role for brain coupling as an emergent property of a multi-animal system in coordinating social interactions and facilitating the development of social relationships (**Figures 4.S1B, 4.8P**).

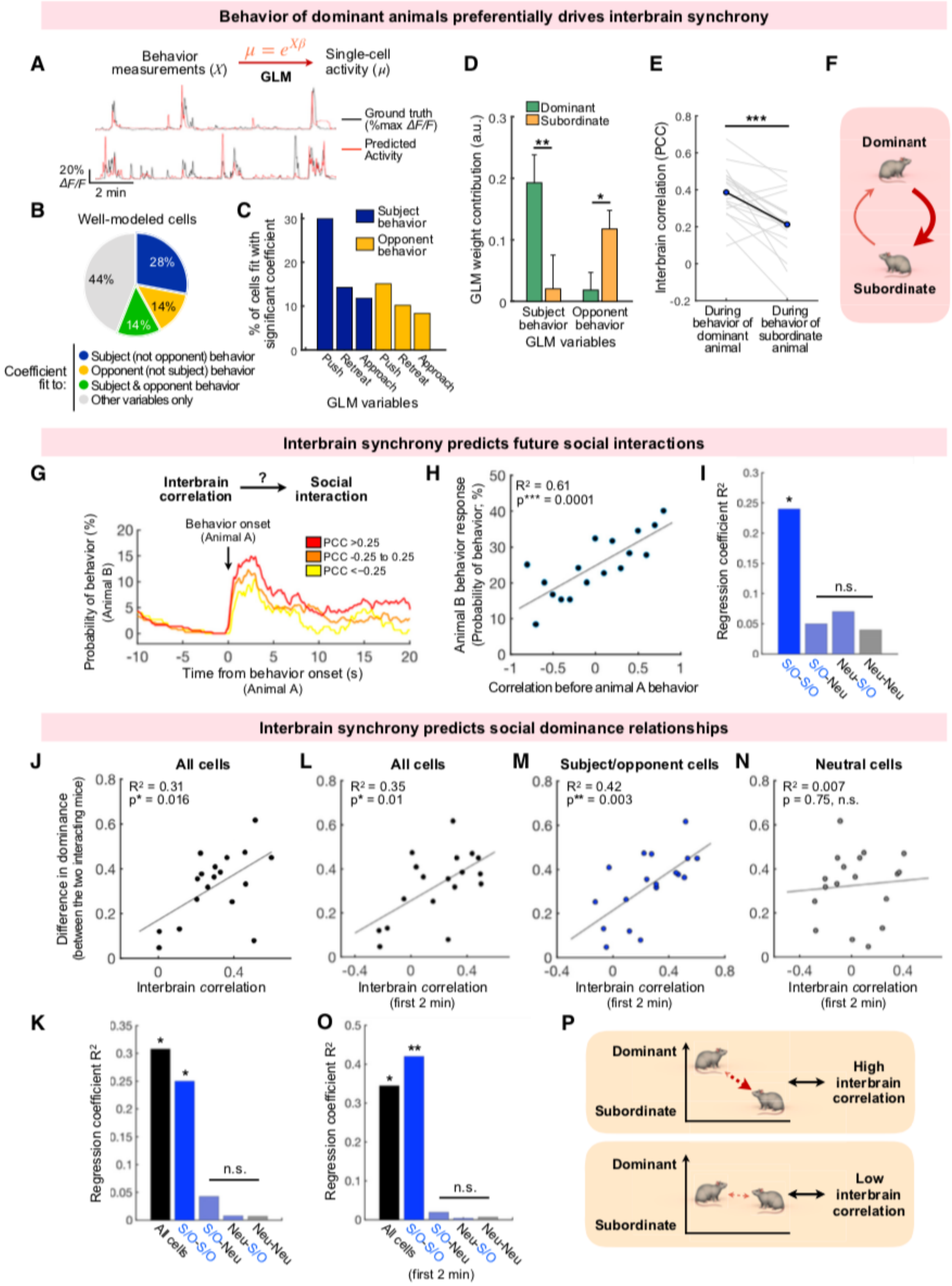

Figure 4.8: Interbrain correlations predict future interaction and dominance relationships. **(A)** Examples of neurons with activity modeled by GLMs using positions and behavior of both animals. **(B)** Distribution of single-neuron GLMs with statistically significant ($p < 0.05$) coefficients fit to subject (but not opponent) behavior, opponent (but not subject) behavior, subject and opponent behavior, or other variables only. **(C)** Distribution of cells with significant coefficients for

specific subject and opponent behaviors. **(D)** Weight contributions (the average of *Z*-scored coefficients) in single-neuron GLMs for subject and opponent behavior in dominants and subordinates. **(E)** Interbrain correlations during behaviors of dominants versus subordinates. **(F)** Schematic showing greater influence on interbrain synchrony by dominant animals. **(G)** Time courses showing the probability of behavior in one animal as a function of time following behavior onset in the interacting partner, color coded based on the interbrain correlation over the preceding 30 s. **(H)** Correlation between the interbrain activity PCC preceding behavior in one animal and the response probability of the interacting partner. **(I)** Regression coefficients (R^2) for the linear relationship shown in (H) using subsets of neurons (S/O, subject and opponent cells; Neu, neutral cells). **(J)** Correlation between the interbrain activity PCC across pairs and the differences in their mean tube position. **(K)** Regression coefficients (R^2) for the linear relationship shown in (J) using subsets of neurons. **(L–N)** Correlation between interbrain activity PCC during the first 2 min of interaction and overall difference in tube position over the session using all cells (L), only subject- and opponent-encoding cells (M), or only neutral cells (N). **(O)** Regression coefficients (R^2) for the linear relationships between interbrain activity correlations during the first 2 min of interaction and dominance difference using subsets of neurons. **(P)** Schematic showing that interbrain coupling is higher when one animal is significantly more dominant than its opponent, and lower when two animals have similar levels of dominance. ***$p < 0.001$, **$p < 0.01$, $p > 0.05$, n.s. (D) Mean ± SEM.

4.5: Discussion

4.5.1: Interbrain correlated neural activity during social interaction

Previous research on interbrain synchrony has illuminated the capacity for neural circuits to coordinate across individuals during social engagement [10,33]. However, it has been largely unclear how region-wide interbrain correlations arise from activity patterns at the circuit or single-cell levels. Using simultaneous large-scale recordings in interacting animal dyads, we provide conclusive evidence that mice exhibit interbrain correlations of neural activity in the dmPFC that arise from ongoing social interaction. We observed correlated activity in an unconstrained environment, as well as during dominance competitions in the tube test, suggesting that social brain coupling is a general phenomenon present in multiple contexts. Importantly, interbrain correlations could not be simply explained by activity associated with concurrent or coordinated behavior. Rather, the coupling of brain activity likely reflects specific types of meaningful engagement, as well as attentional entrainment across pairs of animals embedded in a larger social context. As interbrain coupling has only previously been observed in humans and non-human primates, this finding strongly suggests generality and conservation of the phenomenon across a wide range of animal species.

Importantly, rather than reflecting uniform changes in the firing patterns of cell populations, we find that activity synchrony depends specifically on subsets of neurons that separately encode behaviors of the subject animal and those of the interacting partner. These cells allow each brain to represent a common repertoire of behavior (behavior of both interacting animals), such that activity across separate brains becomes synchronized. The existence of opponent-encoding cells in part explains why interbrain synchrony is not simply accounted for by coordinated rest and concurrent behavior, and highlights the complexity of mechanisms underlying synchrony that invite deeper investigation at the circuit level.

4.5.2: Encoding of one's own and the social partner's behavior in dmPFC neurons

In many social species, including humans, social interactions between individuals are shaped by status relationships and dominance competitions [11]. Recent work has begun to investigate the neural mechanisms underlying the expression of dominance behavior [34,35]. In our study, we identified a substantial fraction of neurons in the dmPFC that encode distinct social dominance behaviors during a competitive encounter. These single-cell responses collectively formed stable representations of push, retreat, and approach behavior, suggesting a role for dmPFC neurons in regulating multiple, sometimes opposing, behavioral strategies.

In addition to coordinating one's own behavior, social interactions also require animals to anticipate and react to the decisions of their social partners. However, it is not well understood how neural systems represent observed behavior. Studies in humans and non-human primates report that prefrontal, motor, and parietal regions can respond to actions displayed by other individuals [36–39]. Yet many of these studies were done in the context of passive and unidirectional behavioral observation. It is largely unclear how representations of self and others' behavior arise during dynamic interactions where animals must simultaneously observe and respond within seconds. We find that a fraction of dmPFC neurons in mice encode specific behaviors of the interacting partner, and collectively form a neural response pattern that distinguishes opponent and subject behavior. The presence of these neurons in the rodent dmPFC suggests conservation of function across diverse species and sets the groundwork for deeper investigation using a genetically tractable animal model.

We also explored whether encoding of observed behavior is shaped by dominance status. Interestingly, while subject behavior was more strongly encoded in dominants than in subordinates, opponent behavior was more robustly encoded in subordinates than in dominants, suggesting an asymmetry in the computational structure of the dmPFC circuit based on social status. Moreover, synchrony was consistently higher during dominant animals' behavior than during subordinate animals' behavior. These suggest that during competitive interactions, subordinates may be more attentive to dominants. Indeed, in primates, subordinates pay more attention to the actions and gazes of dominant individuals [40,41]. Our results suggest that, in rodents, this feature of directed social attention could be instantiated in the activity of dmPFC neurons.

Animals also have the capacity to encode other information about conspecifics, such as their physical location or emotional state [42–44]. How these processes are related to the encoding of volitional behavior of others is unclear and remains an exciting topic for future study.

4.5.3: Interbrain correlations predict social interactions and dominance relationships

Beyond providing a neural basis for how interbrain synchrony arises from individual cells, our study also functionally links it to the coordination of social interactions—stronger interbrain correlations across dyads predict future social interaction. While interbrain coupling originates from activities in individual brains, it represents a state of multi-individual systems that operates at the level of the system itself and is not accessible to each brain to directly influence one's own decisions. Instead, this state reflects one or several underlying neural processes within each brain that operate to shape animal behavior. Given the role of opponent-encoding neurons in interbrain synchrony, correlated activity may in part reflect attentional engagement between animals, effectively coupling their decisions and increasing their behavioral reciprocity. As interbrain coupling both arises from and predicts dyadic behavior, the behavioral interaction and its interbrain neural correlate may form a bidirectional feedback loop that serves to facilitate and sustain ongoing interaction (**Figure 4.S1B**).

In addition to our observation that dominants drive stronger responses from subordinates, we also found that the degree of interbrain correlation across each pair predicted dominance relationships, whereas correlations between their behavior could not. This echoes previous reports in humans that brain coupling can predict leader-follower relationships, even before leadership roles are manifested [45–47]. Our results suggest that synchrony across individuals with unequal status relationships depends on circuitry that encodes actions of social partners, and in such contexts, may reflect the directed engagement of "followers" toward more dominant individuals leading an interaction.

Collectively, our results shed new light on the neural basis and functional role of interbrain synchrony in coordinating social interactions. More importantly, they set the groundwork for a more incisive investigation of the emergent neural properties of multi-individual systems, which may yet reveal a richer and deeper understanding of the social brain as it is embedded in a truly social world.

4.6: Supplemental data

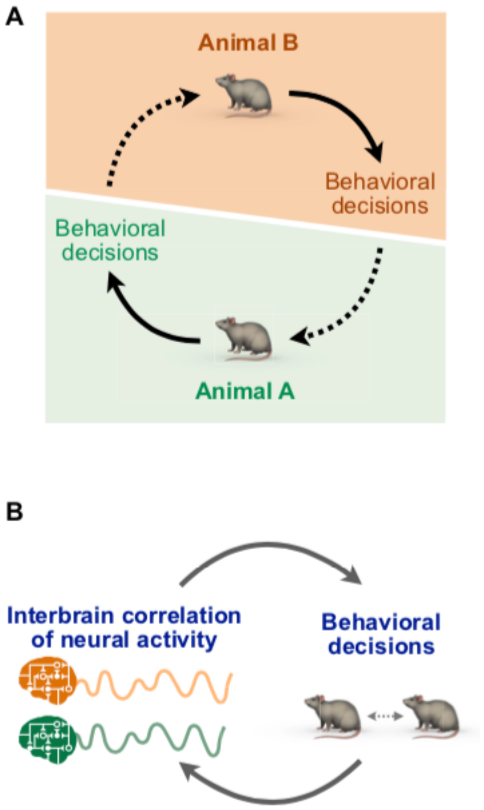

Figure 4.S1: Social behavior and interbrain coupling in interacting animals. **(A)** Schematic showing social behavioral decisions of animals engaged in dyadic social interaction. **(B)** Feedback loop between interbrain synchrony and social interactions. The coupling of activity between interacting animals facilitates and sustains ongoing social interaction.

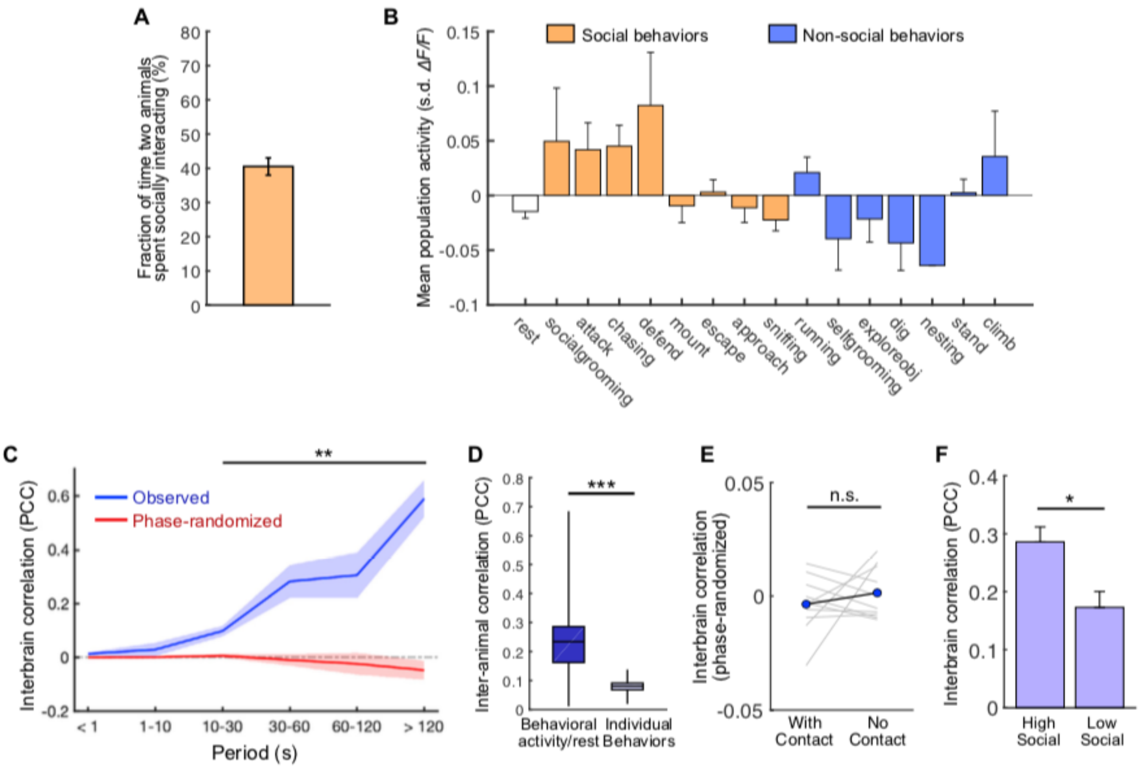

Figure 4.S2: Analysis of behavior and interbrain correlations in the open arena. **(A)** Total time two animals spent interacting in the open arena (which includes time when a single animal or both animals are engaged in social behavior). **(B)** Mean dmPFC activity during different types of social (orange) and non-social (blue) behaviors across all animals engaged in open arena interactions (mean ± SEM). **(C)** Correlations of dmPFC activity (blue) and phase-randomized traces (red) across animal pairs at different timescales. Mean activity traces were decomposed into different frequency bands using a Fourier transform. Interbrain correlations are stronger at slower timescales, consistent with the notion that correlations depend on a larger context of continuous, ongoing interaction on a scale of seconds to minutes. **(D)** Correlation of behavioral activity and rest across animals interacting in the open arena (all types of behavior pooled, left) and correlation of specific types of behaviors across animals (right) (p*** < 0.001). This suggests that, across animals, behavior activity and rest are somewhat correlated (left), whereas individual behaviors are not correlated (right). **(E)** Correlations of phase-randomized activity traces across animals in the open arena with or without social contact ($p > 0.05$ – not significant). **(F)** Comparison of interbrain correlations among animal pairs that naturally displayed high or low levels of mutual social interaction. Pairs with a higher degree of social interaction showed higher interbrain synchrony, consistent with the notion that synchrony depends on ongoing interaction.

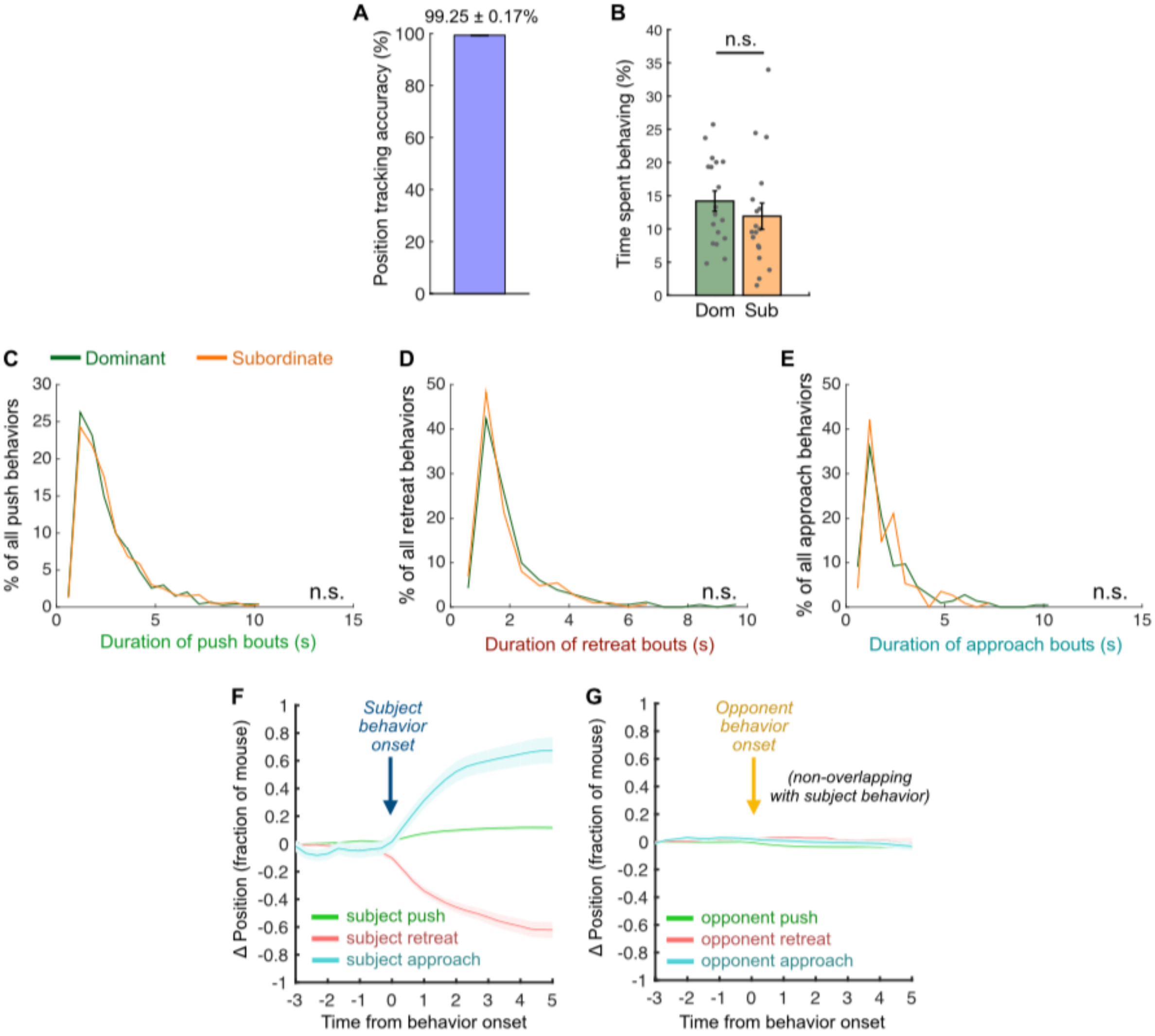

Figure 4.S3: Automated tracking and analysis of animal behavior during the tube test. **(A)** Performance of the convolutional neural network to automatically track the locations of interacting mice in behavior movies, measured by the accuracy of the algorithm to properly identify both mice and correctly determine their positions in a subset of randomly drawn frames, compared with ground truth assessment determined by an unbiased individual (mean ± SEM). **(B)** Total percentage of time spent behaving among dominant and subordinate animals across all pairs. For each pair, the dominant animal is the one with the greater mean tube position (mean ± SEM, $p > 0.05$; not significant). **(C-E)** Distribution of per-bout behavior durations for push (C), retreat (D), and approach (E) behavior in dominant or subordinate animals (Kolmogorov-Smirnov test, $p > 0.05$; not significant). **(F)** Average change in position of mice during subject push, retreat, or approach behavior (mean ± SEM). **(G)** Average change in position of mice during opponent push, retreat, or approach behavior (mean ± SEM; behavior bouts when subject and opponent behavior overlapped were removed from analysis).

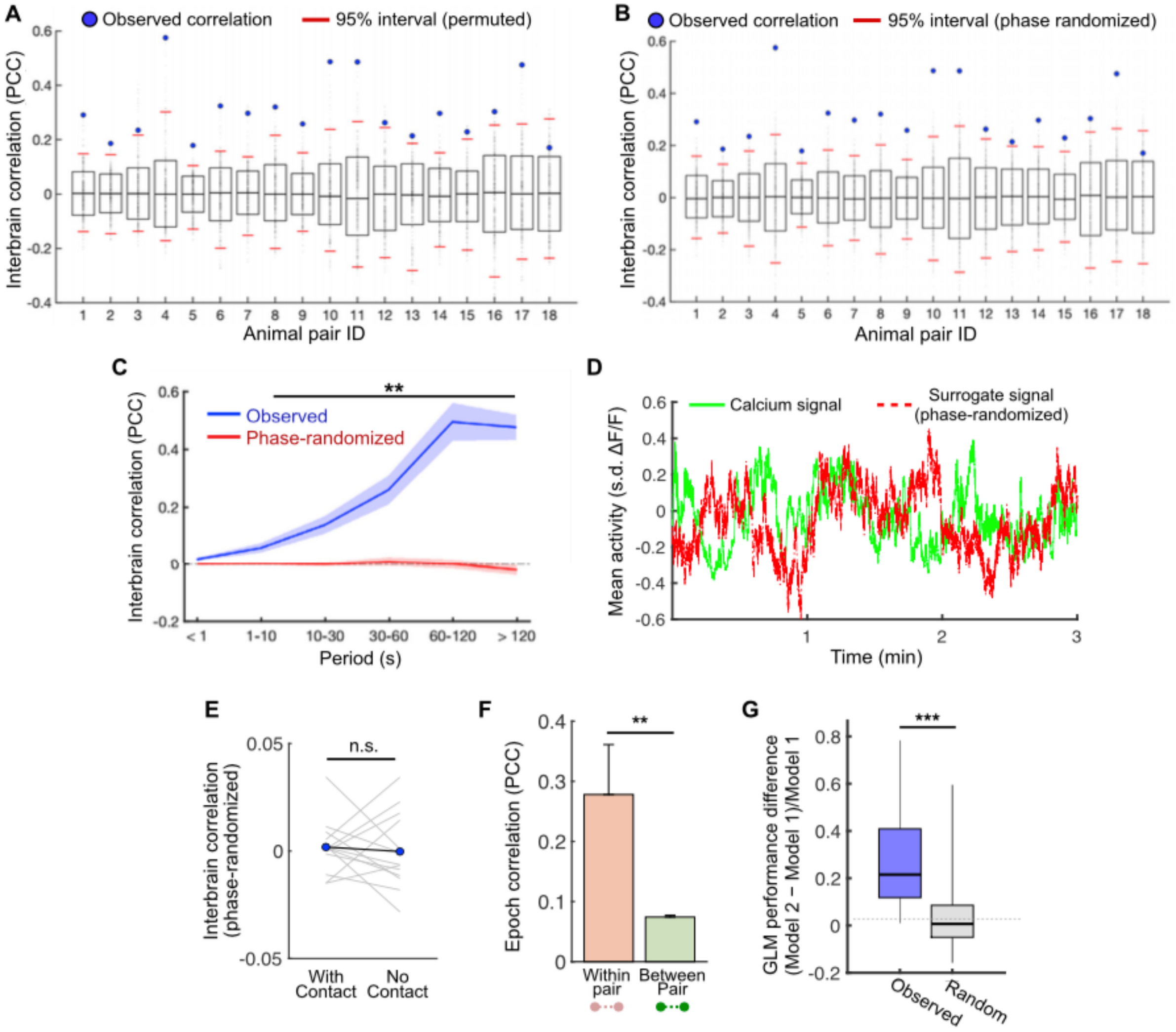

Figure 4.S4: Analysis of interbrain correlations in the tube test. **(A, B)** For each animal pair, the observed interbrain correlation (PCC; blue dots) shown against a null distribution of PCCs. Boxes indicate mean ± standard deviation of the null distributions; red lines indicate 95% intervals (2.5th and 97.5th percentile). (A) Null distributions are generated from temporally permuted traces. (B) Null distributions are generated from phase-randomized traces. **(C)** Correlations of dmPFC activity (blue) and phase-randomized traces (red) across pairs at different timescales using Fourier decomposition of signals into different frequency bands. Interbrain correlations are stronger at slower timescales, consistent with the notion that correlations depend on a larger context of continuous, ongoing interaction on a scale of seconds to minutes. **(D)** Example trace of the average activity of all dmPFC neurons in one animal (green), and a surrogate phase-randomized signal (red) with disrupted temporal structure but identical mean, variance, and autocorrelation as the original trace. **(E)** Interbrain correlations of phase-randomized traces from tube test experiments with or without social contact, as in Figure 4J ($p > 0.05$; not significant). **(F)** Comparison of interbrain correlations during epochs with concurrent isolated behavior bouts in interacting pairs in the tube test (left), and behavior-matched epochs from non-interacting pairs (right) (mean ± SEM). **(G)** The difference in performance of GLM models schematized in Figure 2J for animals engaged in the tube test, compared with that using phase-randomized activity from

the interacting partner. The GLM performance difference quantifies the relative difference in model performance when activity from the interacting partner is included as a variable in addition to behavior variables (p** < 0.01).

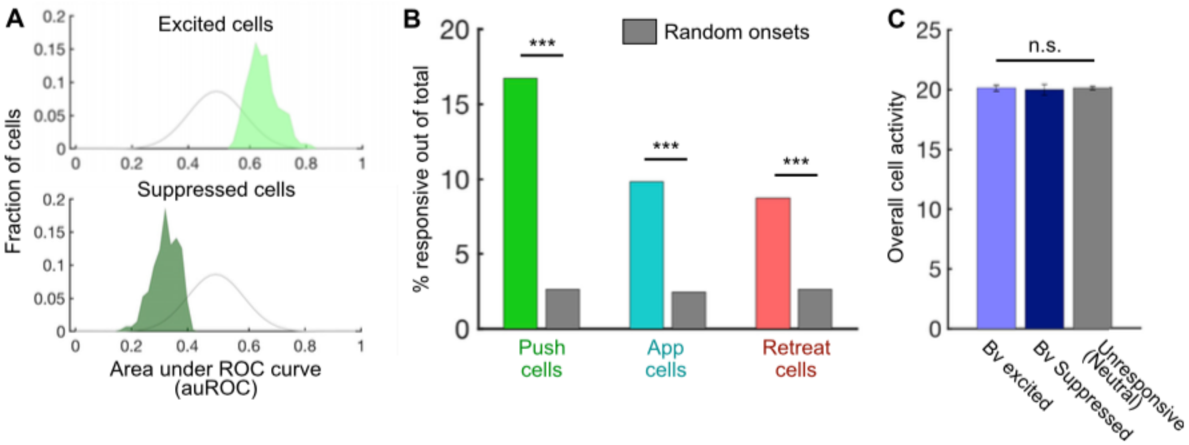

Figure 4.S5: Activity and spatial intermixing of behavior cells in dmPFC. **(A)** Distributions of auROC (area under the ROC curve) values for cells that are excited (top) or suppressed (bottom) during behavior. Significantly responsive cells were determined using permutation testing. Gray curve indicates the distribution of auROC values from neutral cells that do not respond during behavior. **(B)** Comparison between the percentage of behavior-excited cells identified over all tube test sessions and the percentage expected by chance. Chance levels were determined by comparing auROC values of temporally permuted calcium traces against random null distributions (p*** < 1.0e-10, Fisher's exact test). **(C)** Average cell activities for behavior-excited, behavior-suppressed, and neutral (behavior-unresponsive) cells. For each neuron, overall activity is measured as the percentage of time the calcium trace is above 10% of its maximum value (p > 0.05; not significant).

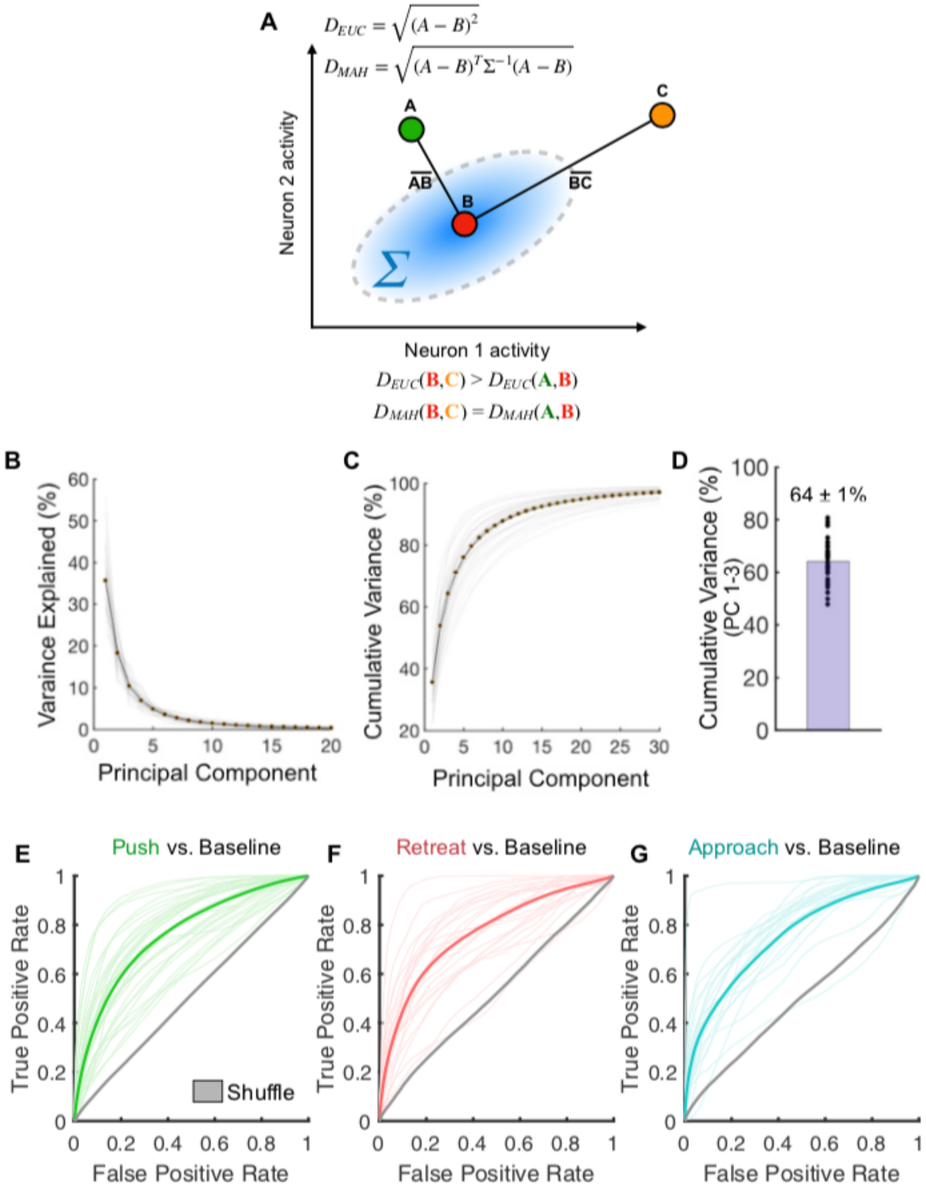

Figure 4.S6: Separation of population responses encoding distinct social behaviors. **(A)** Cartoon illustration of the Mahalanobis distance and Euclidean distance between pairs of points on a 2D plane. The Mahalanobis distance considers the shape of the underlying distribution of data by scaling dimensions based on their covariance (the correlational structure of the neural population). Although point C is further from B than A is in Euclidean terms, A and C are equidistant from B using the Mahalanobis distance. **(B)** Percentage of the total variance of trial-averaged population activity during behavior in tube test sessions that is captured by principal components (gray curves); average over all sessions shown with black curve. **(C)** The cumulative variance of trial-averaged population activity captured by principal components as a function of the number of components (gray curves); average over all sessions shown with black curve. **(D)** Average variance in population activity captured by the first three principal components, as shown in (C) (mean ± SEM). **(E-G)** ROC curves quantifying the performance of FLD decoders to predict push (E), retreat (F), and approach (G) behavior based on population activity. Thin color lines:

performance for each session; dark color lines: average ROC curves taken over all sessions; gray lines: the average of chance decoders constructed using training data with randomly shuffled class labels.

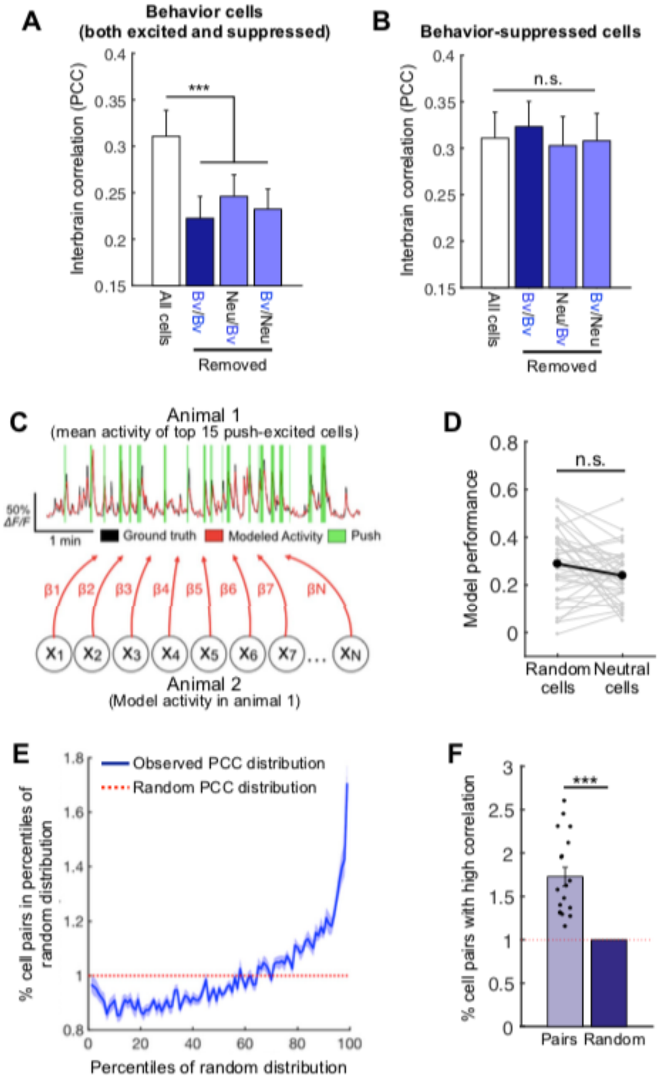

Figure 4.S7: Single behavior cell contributions to interbrain correlations. **(A)** Interbrain activity correlations after removal of behavior cells (Bv; including both excited and suppressed cells) or neutral cells (Neu) from both animals (mean ± SEM). **(B)** Interbrain activity correlations after removal of behavior-suppressed (Bv) or neutral (Neu) cells from both animals (mean ± SEM). **(C)** Schematic of GLM fit to model mean activity of subsets of behavior-excited cells as shown in Figure 6H using single neuron activities from the interacting partner. Red line: modeled activity of the top 15 push cells from one animal/session. Black line: ground truth activity of the same group of cells. **(D)** Comparison between the performance of GLMs constructed to model the mean of subsets (15 cells) of randomly selected or neutral cells ($p > 0.05$; not significant). **(E)** Distribution of PCC of all single neuron pairs across interacting animals in the tube test (blue). Each bin represents one percentile of the random distribution (chance level of 1%, red) of correlations generated from calculating PCCs over temporally permuted calcium traces (mean ± SEM). This indicates that pairs of single cells across interacting animals exhibit a higher level of correlation than expected by chance. **(F)** The percentage of single cell interbrain correlations that exceed the

99th percentile of null distributions generated from randomly permuted calcium traces, as in (E). Percentage of highly correlated cell pairs is compared with the chance level of 1% (mean ± SEM, p*** < 10^{5}).

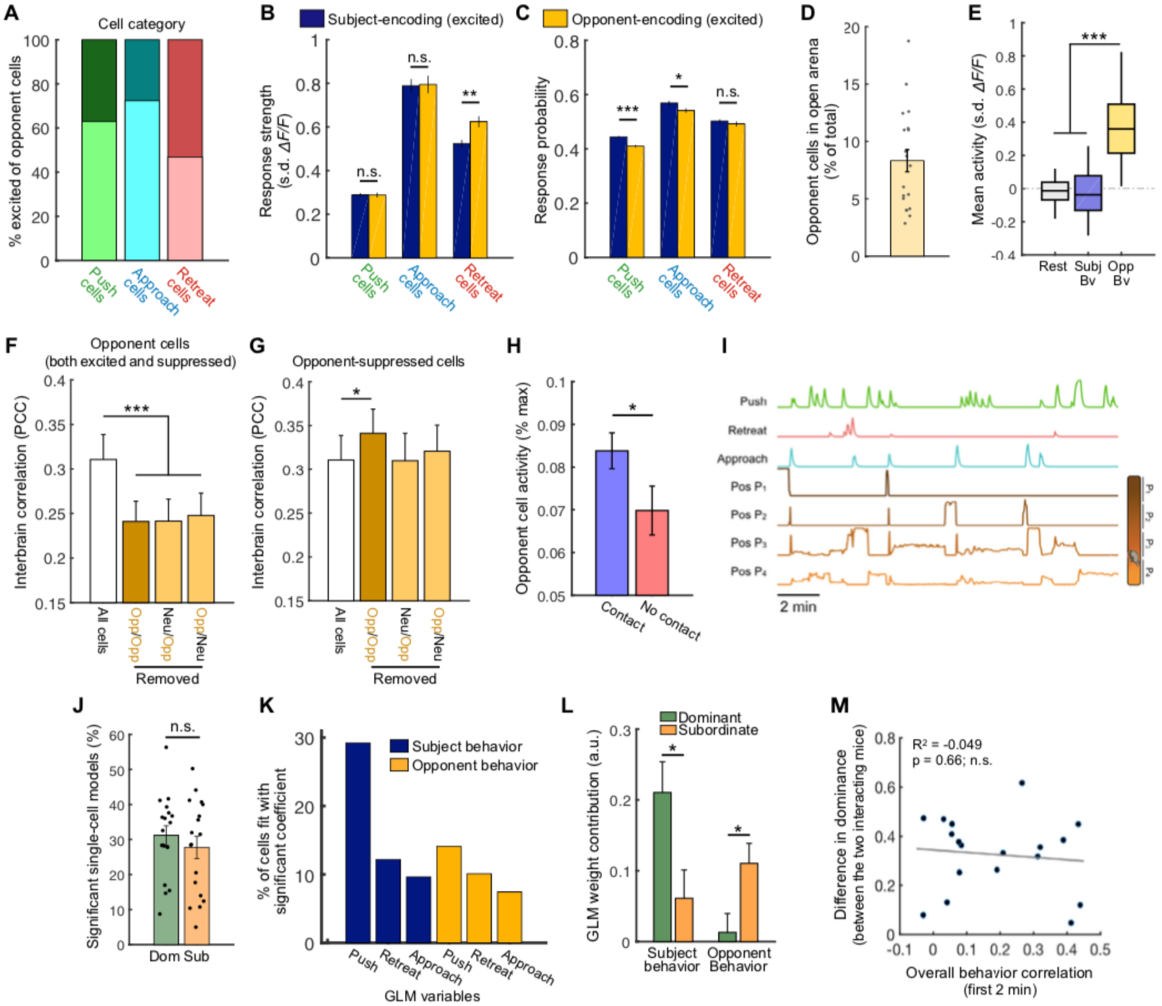

Figure 4.S8: Analysis of behavior cell properties and single neuron models. **(A)** Distribution of excited (light color) and suppressed (dark color) opponent cells within each behavior category. **(B, C)** Response strength (B) and response probability (C) of subject behavior-excited and opponent behavior-excited cells for different behavior categories. The response strength for each cell is calculated as the mean activity over all behavior epochs. The response probability is calculated as the percentage of behavior events with neural activity exceeding 110% of the local baseline. **(D)** Percentage of neurons recorded during open arena interactions that respond selectively during opponent social behavior. Opponent cells in open arena interactions were identified using ROC analysis based on opponent behavior (not overlapping with subject behavior) and rest epochs. **(E)** Mean activity of opponent cells during subject behavior, opponent behavior, or rest (when neither animal is behaving) in open area interactions. **(F, G)** Interbrain activity correlations after removal of opponent cells (Opp) or neutral cells (Neu) from both animals. Opponent cells includes both excited and suppressed cells (F) or only suppressed cells (G). **(H)** Activity (percent of the max activity value) of all behavior-excited opponent cells during the tube test with or without social contact. **(I)** Illustration of the variables used to fit single neuron GLM

models. Behavior vectors denoting social behavior of each animal are exponentially smoothed, and position coordinates for each animal are decomposed into four positions that tile the length of the tube. **(J)** Percentage of single neurons in each tube test session that are modeled well (exceed chance levels based on cross-validation) by a GLM fit to the behavior and positions of both animals. **(K)** Percentage of cells fit with significant coefficients for individual subject and opponent behaviors, as in Figure 8C. Here, single-neuron GLMs were identified using cross-validated R^2 as an alternative performance metric. **(L)** Contributions of coefficients in single neuron GLMs for subject and opponent behavior in dominant and subordinate animals. Weight contribution was calculated as the average of normalized coefficients over all cells in each animal, as in Figure 8D. Here, single-neuron GLMs were identified using cross-validated R^2 as an alternative performance metric. **(M)** Relationship between overall behavior correlation across pairs during the first 2 min of interaction and their overall dominance difference over the session. Overall behavior correlations were measured by the correlation of the presence of behaviors of any types, which reflects the level of overall concurrent behavior. ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$, $p > 0.05$, n.s. (B–D, F–H, J, L) Mean ± SEM.

4.7: References

1. Chen, P. & Hong, W. Neural Circuit Mechanisms of Social Behavior. *Neuron* **98**, 16–30 (2018).

2. Sanfey, A. G. Social Decision-Making: Insights from Game Theory and Neuroscience. *Science (80-.).* **318**, 598–602 (2007).

3. Rilling, J. K. & Sanfey, A. G. The Neuroscience of Social Decision-Making. *Annu. Rev. Psychol.* **62**, 23–48 (2011).

4. Adolphs, R. Conceptual Challenges and Directions for Social Neuroscience. *Neuron* **65**, 752–767 (2010).

5. Ochsner, K. N. & Lieberman, M. D. The emergence of social cognitive neuroscience. *Am. Psychol.* **56**, 717–34 (2001).

6. Schilbach, L. *et al.* Toward a second-person neuroscience. *Behav. Brain Sci.* **36**, 393–414 (2013).

7. Babiloni, F. *et al.* Hypermethods for EEG hyperscanning. in *2006 International Conference of the IEEE Engineering in Medicine and Biology Society* 3666–3669 (IEEE, 2006). doi:10.1109/IEMBS.2006.260754

8. King-Casas, B. *et al.* Getting to Know You: Reputation and Trust in a Two-Person Economic Exchange. *Science (80-.).* **308**, 78–83 (2005).

9. Liu, T. & Pelowski, M. A new research trend in social neuroscience: Towards an interactive-brain neuroscience. *PsyCh J.* **3**, 177–188 (2014).

10. Montague, P. R. *et al.* Hyperscanning: simultaneous fMRI during linked social interactions. *Neuroimage* **16**, 1159–64 (2002).

11. Williamson, C. M., Lee, W. & Curley, J. P. Temporal dynamics of social hierarchy formation and maintenance in male mice. *Anim. Behav.* **115**, 259–272 (2016).

12. Cooper, M. A., Clinard, C. T. & Morrison, K. E. Neurobiological mechanisms supporting experience-dependent resistance to social stress. *Neuroscience* **291**, 1–14 (2015).

13. Sapolsky, R. M. Social Status and Health in Humans and Other Animals. *Annu. Rev. Anthropol.* **33**, 393–418 (2004).

14. Sapolsky, R. M. The Influence of Social Hierarchy on Primate Health. *Science (80-.).* **308**, 648–652 (2005).

15. Utevsky, A. V & Platt, M. L. Status and the brain. *PLoS Biol.* **12**, e1001941 (2014).

16. Zhou, T. *et al.* History of winning remodels thalamo-PFC circuit to reinforce social dominance. *Science (80-.).* **357**, 162–168 (2017).

17. Wang, F. *et al.* Bidirectional Control of Social Hierarchy by Synaptic Efficacy in Medial Prefrontal Cortex. *Science (80-.).* **334**, (2011).

18. Franklin, T. B. *et al.* Prefrontal cortical control of a brainstem social behavior circuit. *Nat. Neurosci.* **20**, 260–270 (2017).

19. Warden, M. R. *et al.* A prefrontal cortex–brainstem neuronal projection that controls response to behavioural challenge. *Nature* **492**, 428 (2012).

20. Murugan, M. *et al.* Combined Social and Spatial Coding in a Descending Projection from the Prefrontal Cortex. *Cell* **171**, 1663-1677.e16 (2017).

21. Liang, B. *et al.* Distinct and Dynamic ON and OFF Neural Ensembles in the Prefrontal Cortex Code Social Exploration. *Neuron* (2018). doi:10.1016/j.neuron.2018.08.043

22. Redmon, J. & Farhadi, A. YOLO9000: Better, Faster, Stronger. (2016).

23. Pnevmatikakis, E. A. & Giovannucci, A. NoRMCorre: An online algorithm for piecewise rigid motion correction of calcium imaging data. *J. Neurosci. Methods* **291**, 83–94 (2017).

24. Mukamel, E. A., Nimmerjahn, A. & Schnitzer, M. J. Automated analysis of cellular signals

from large-scale calcium imaging data. *Neuron* **63**, 747–60 (2009).

25. Li, Y. *et al.* Neuronal Representation of Social Information in the Medial Amygdala of Awake Behaving Mice. *Cell* **171**, 1176-1190.e17 (2017).

26. Cunningham, J. P. & Yu, B. M. Dimensionality reduction for large-scale neural recordings. *Nat. Neurosci.* **17**, 1500–1509 (2014).

27. Remedios, R. *et al.* Social behaviour shapes hypothalamic neural ensemble representations of conspecific sex. *Nature* **550**, 388–392 (2017).

28. Driscoll, L. N., Pettit, N. L., Minderer, M., Chettih, S. N. & Harvey, C. D. Dynamic Reorganization of Neuronal Activity Patterns in Parietal Cortex. *Cell* **170**, 986-999.e16 (2017).

29. Runyan, C. A., Piasini, E., Panzeri, S. & Harvey, C. D. Distinct timescales of population coding across cortex. *Nature* **548**, 92–96 (2017).

30. Drews, C. The Concept and Definition of Dominance in Animal Behaviour. *Behaviour* **125**, 283–313 (1993).

31. Wang, F. *et al.* The mouse that roared: neural mechanisms of social hierarchy. *Trends Neurosci.* **37**, 674–82 (2014).

32. Pouget, A., Dayan, P. & Zemel, R. Information processing with population codes. *Nat. Rev. Neurosci.* **1**, 125–132 (2000).

33. Liu, T. & Pelowski, M. A new research trend in social neuroscience: Towards an interactive-brain neuroscience. *PsyCh J.* **3**, 177–188 (2014).

34. Stagkourakis, S. *et al.* A neural network for intermale aggression to establish social hierarchy. *Nat. Neurosci.* **21**, 834–842 (2018).

35. Zhou, T., Sandi, C. & Hu, H. Advances in understanding neural mechanisms of social

dominance. *Curr. Opin. Neurobiol.* **49**, 99–107 (2018).

36. Ogawa, K. & Inui, T. Neural representation of observed actions in the parietal and premotor cortex. *Neuroimage* **56**, 728–735 (2011).

37. Hardwick, R. M., Caspers, S., Eickhoff, S. B. & Swinnen, S. P. Neural correlates of action: Comparing meta-analyses of imagery, observation, and execution. *Neurosci. Biobehav. Rev.* **94**, 31–44 (2018).

38. Rozzi, S. & Fogassi, L. Neural Coding for Action Execution and Action Observation in the Prefrontal Cortex and Its Role in the Organization of Socially Driven Behavior. *Front. Neurosci.* **11**, 492 (2017).

39. Tseng, P.-H., Rajangam, S., Lehew, G., Lebedev, M. A. & Nicolelis, M. A. L. Interbrain cortical synchronization encodes multiple aspects of social interactions in monkey pairs. *Sci. Rep.* **8**, 4699 (2018).

40. Deaner, R. O., Khera, A. V & Platt, M. L. Monkeys pay per view: adaptive valuation of social images by rhesus macaques. *Curr. Biol.* **15**, 543–8 (2005).

41. Klein, J. T., Shepherd, S. V & Platt, M. L. Social attention and the brain. *Curr. Biol.* **19**, R958-62 (2009).

42. Danjo, T., Toyoizumi, T. & Fujisawa, S. Spatial representations of self and other in the hippocampus. *Science* **359**, 213–218 (2018).

43. Panksepp, J. & Panksepp, J. B. Toward a cross-species understanding of empathy. *Trends Neurosci.* **36**, 489–496 (2013).

44. Allsop, S. A. & Wichmann, R. Corticoamygdala Transfer of Socially Derived Information Gates Observational Learning. *Cell* **173**, (2018).

45. Jiang, J. *et al.* Leader emergence through interpersonal neural synchronization. *Proc. Natl. Acad. Sci.* **112**, 4274–4279 (2015).

46. Konvalinka, I. *et al.* Frontal alpha oscillations distinguish leaders from followers: Multivariate decoding of mutually interacting brains. *Neuroimage* **94**, 79–88 (2014).

47. Sänger, J., Müller, V. & Lindenberger, U. Directionality in hyperbrain networks discriminates between leaders and followers in guitar duets. *Front. Hum. Neurosci.* **7**, 234 (2013).

CHAPTER FIVE

Conclusions and Future Directions

5.1: Conclusions:

I have presented two studies in this dissertation, each focused on investigating the neural processes in the brain that shape social perception, behavior, and interaction between individuals. The first study, described in chapter 3, examined the process of social decision-making in the brain by exploring how a complex social variable – sex identity – is encoded in cortical neurons, and then linking these neural representations of sex to behavior through correlational analyses and optogenetic manipulations [1]. One important conclusion from this study was that neurons in the prefrontal cortex robustly encode sex as a variable and can discriminate the sex identity of conspecifics (**Figure 3.1**). Previous work in the field had largely focused on the role of subcortical circuits in processing sensory cues and extracting social variables (see section 1.3.3.2) [2–5]. Although some recent studies have pointed to a role for prefrontal cortical circuits in processing social information [6,7], the encoding of specific social variables such as sex had not been explored. In addition, our study also established a causal role for native representations of sex in the control of animal behavior using optogenetic manipulations (**Figure 3.6**). Previous studies had identified an association between the encoding of sex in the brain and social behavior, implying that internal representations of social variables play a causal role in the transformation of sensory inputs into behavioral changes [3,4]. Although this is compelling interpretation of observational data, the idea that native representations of sex affect animal behavior was a hypothesis that had not been tested. Our study tested this hypothesis explicitly using activity-dependent expression of channelrhodopsin, allowing us to directly manipulate the neurons that encode male or female cues. These findings close an important knowledge gap in the field, and lay groundwork for more incisive experiments that can uncover the mechanics of how neurons encoding sex modulate expressed behavior.

In chapter 4, I presented work that investigated the synchronization of neural activity across the brains of animals engaged in social interaction. Over the last 20 years, a rich literature has emerged in human social neuroscience that has examined the relationship between brain activity across interacting people, and in many cases, has linked properties of these signals (such as time-series correlation and coherence in specific frequency domains) to behavioral, cognitive, and relational variables [8–11]. These interbrain dynamics are of interest because they may capture the relationship between internal processes within each agent that are informative about the interaction (such as attention) but may not be explicitly observable at the behavioral level (**Figure 1.2**) [12]. In principle, such signals may be useful as biomarkers for certain types of interaction, and they may even help to distinguish deficits in interaction that are present in individuals with developmental or psychiatric disorders [13,14]. Despite several years of fascinating discoveries using this approach, two fundamental questions have remained unaddressed: 1) *What is the biological reason that signals recorded from physically separate neural systems are correlated?* 2) *Why does the correlation or coherence between these signals contain information about behavioral, cognitive, or relational variables?* Our study moved understanding in both of these knowledge gaps forward. By recording activity of hundreds of individual neurons and using these data to derive a "bulk" regional signal, we were able to then decompose the synchronized signals into their biological components and ask specific questions about which components give rise to correlated dynamics. In the prefrontal cortex of mice, we found that interbrain correlations emerge because of the coding properties of individual neurons – in this case, because single neurons encode the behavior of the subject animal and its social partner (**Figure 4.7**) [15]. We also linked the degree of interbrain correlation to social rank relationships between pairs, echoing previous reports of synchronization across leaders and followers in a group (**Figure 4.8**) [16,17]. While not yet tested, it appears that this activity-behavior relationship may arise because of asymmetric coding properties of prefrontal cortex neurons across dominant and subordinate animals (**Figure 4.8**). In general, our study models a method for analyzing how interbrain dynamics emerge from underlying neural components, and it demonstrates how the use of animal models can provide deeper insight into the biology underlying this phenomenon. It also suggests a more general theoretical framework for thinking about interbrain dynamics across species, contexts, and brain regions (elaborated in a review article entitled "A Multi-Brain Framework for Social Interaction" [12]) – namely, that *interbrain dynamics and their relationship with behavior arise because of the specific coding properties of the underlying neural components*. This framework may be used to guide future research in both animals and humans toward more informed hypotheses and deeper mechanistic understanding.

5.2: Future directions:

The study presented in chapter 3 leaves open several important directions of inquiry. On the mechanistic side, immediate directions include using anatomical tracing techniques to investigate the inputs to male- and female-encoding mPFC neurons and to understand how their sex-specific responses are shaped [18,19]. An important open question on this front is whether encoding of sex in mPFC is shaped by one dominant sensory input (for example, pheromonal information routed from MeA or BNST), or is integrated across multiple sensory features. Clarification of how sex representations are formed in the brain may uncover more general principles of how complex social features, like social status and familiarity, are extracted from basic sensory inputs. Another important question surrounds how male- and female-encoding mPFC neurons affect behavior. The optogenetics experiments presented in the study indicate that these neurons are causally effective, but they do not indicate precisely how their projections to other circuits in the brain shape behavioral output. Initial circuit tracing experiments showed that these neurons project widely to limbic and subcortical centers, including the amygdala, lateral hypothalamus, ventral tegmental area, nucleus accumbens, and striatum (**Figure 3.S6**). Do any one of these projection targets specifically mediate the effects on behavior, or are the effects redundantly distributed across multiple downstream circuits? Optogenetics manipulations targeted to specific mPFC projection neurons may help to elucidate this [18,19].

On the more conceptual side, one important question in social neuroscience centers around the specificity of social processing and social functions in the brain [20]. Are regions of the brain that appear to process social information in some sense designed for social functions, as the "social brain" hypotheses suggests [21,22]? Or do social interactions simply engage a specific set of primitive cognitive and emotional processes in a unique way that seems to *suggest* social-specific functioning [23]? While the issue may ultimately be semantic, it is possible and important to clarify how basic cognitive and emotional processes overlap with processing of social information. Building on the results presented in chapter 3, future experiments could use longitudinal imaging to examine how neural populations that represent social-sensory variables (like sex identity) are engaged in different types of cognitive processes. One extreme hypothesis is that neurons encoding social information are relatively inactive during non-social cognitive tasks (and vice versa), suggesting that cortical processing of social and non-social information occur in orthogonal subspaces of the population activity state space. The alternative is that these processes are overlapping, such that neurons encoding social cues are also engaged in non-social cognitive tasks. If this is true, it would be informative to examine whether there is any structure in the response profiles of individual neurons across distinct types of tasks. This could help to refine further hypotheses about how the role of mPFC neurons in a particular task or context is shaped by its input architecture, by internal state variables, or other factors.

The investigations of interbrain synchronization presented in chapter 4 also open several interesting lines of inquiry. On the mechanistic side, it would be informative to trace more deeply the response properties of mPFC neurons from inputs, and to explore whether social information encoded in mPFC neurons (which gives rise to the correlational signal across animals) is enriched in specific subpopulations or projection pathways. While our study demonstrated that neurons encoding social behavior play an important role in shaping correlated dynamics, other internal processes may also be important [12,24,25]. For example, the prefrontal cortex is known to play a role in the control of attention – interbrain correlations in the mPFC may also reflect the alignment of attention across interacting animals, possibly at a slower timescale than the encoding of individual behavioral choices. This and related hypotheses could begin to be tested by recording neuromodulatory signals that are thought to be related to attentional processes (such as noradrenergic signaling) [26–28], or by recording specifically from subpopulations of neurons involved in attention (such as parvalbumin-positive interneurons) [29,30].

As discussed in the previous section, this study also presents a more general theory that links interbrain neural dynamics and their behavioral correlates to the coding properties of underlying neural components [12]. While this theory is based on observations in the specific experiments that we performed, at this point it is only speculative, and remains to be tested thoroughly. One important future direction is to examine interbrain dynamics across multiple brain regions and in different social contexts, and to analyze how the heterogeneity in interbrain dynamics is related to the coding properties of individual neurons or neural populations. If our framework is correct, one should be able to explain regional and contextual variability in interbrain dynamics based on the coding structure of the underlying neural components. In principle, the same approach could be applied in the setting of human neuroimaging experiments using fMRI or EEG – agreement across studies performed in different species will be important to validate or revise the conceptual framework.

Finally, the experiments performed in this study were focused on dyadic interaction, but this multi-individual level of analysis is not limited to pairs. Some studies in humans have begun to examine group-level neural dynamics (in groups ranging from 3 to 20 and more), suggesting that this approach may reveal interesting neural correlates of complex group variables such as group learning [31–33]. Although it is practically challenging in any case, simultaneous recording of more than two individuals is probably easier to implement in animals than it is in human subjects. Recent advances in wireless recording technology, including wireless microendoscopes, open the possibility of studying interbrain dynamics across multiple of animals during naturalistic group interaction. Such experiments will expand our understanding of how interbrain dynamics emerge from processes in individual brains and will generate new hypotheses about how these signals may be used to understand more about group interaction and collective behavior.

5.3: References

1. Kingsbury, L. *et al.* Cortical Representations of Conspecific Sex Shape Social Behavior. *Neuron* (2020). doi:10.1016/j.neuron.2020.06.020
2. Li, Y. & Dulac, C. Neural coding of sex-specific social information in the mouse brain. *Current Opinion in Neurobiology* **53**, 120–130 (2018).
3. Li, Y. *et al.* Neuronal Representation of Social Information in the Medial Amygdala of Awake Behaving Mice. *Cell* **171**, 1176-1190.e17 (2017).
4. Remedios, R. *et al.* Social behaviour shapes hypothalamic neural ensemble representations of conspecific sex. *Nature* **550**, 388–392 (2017).
5. Bayless, D. W. *et al.* Limbic Neurons Shape Sex Recognition and Social Behavior in Sexually Naive Males. *Cell* **176**, 1190-1205.e20 (2019).
6. Liang, B. *et al.* Distinct and Dynamic ON and OFF Neural Ensembles in the Prefrontal Cortex Code Social Exploration. *Neuron* (2018). doi:10.1016/j.neuron.2018.08.043
7. Levy, D. R. *et al.* Dynamics of social representation in the mouse prefrontal cortex. *Nat. Neurosci.* (2019). doi:10.1038/s41593-019-0531-z
8. Montague, P. R. *et al.* Hyperscanning: simultaneous fMRI during linked social interactions. *Neuroimage* **16**, 1159–64 (2002).
9. King-Casas, B. *et al.* Getting to Know You: Reputation and Trust in a Two-Person Economic Exchange. *Science (80-.).* **308**, 78–83 (2005).
10. Babiloni, F. *et al.* Hypermethods for EEG hyperscanning. in *2006 International Conference of the IEEE Engineering in Medicine and Biology Society* 3666–3669 (IEEE, 2006). doi:10.1109/IEMBS.2006.260754
11. Schilbach, L. *et al.* Toward a second-person neuroscience. *Behav. Brain Sci.* **36**, 393–414 (2013).
12. Kingsbury, L. & Hong, W. A Multi-Brain Framework for Social Interaction. *Trends Neurosci.* (2020). doi:10.1016/j.tins.2020.06.008

13. Tanabe, H. C. *et al.* Hard to 'tune in': Neural mechanisms of live face-to-face interaction with high-functioning autistic spectrum disorder. *Front. Hum. Neurosci.* **6**, 268 (2012).

14. Wang, Q. *et al.* Autism Symptoms Modulate Interpersonal Neural Synchronization in Children with Autism Spectrum Disorder in Cooperative Interactions. *Brain Topogr.* **33**, 112–122 (2020).

15. Kingsbury, L. *et al.* Correlated Neural Activity and Encoding of Behavior across Brains of Socially Interacting Animals. *Cell* **178**, 429-446.e16 (2019).

16. Konvalinka, I. *et al.* Frontal alpha oscillations distinguish leaders from followers: Multivariate decoding of mutually interacting brains. *Neuroimage* **94**, 79–88 (2014).

17. Jiang, J. *et al.* Leader emergence through interpersonal neural synchronization. *Proc. Natl. Acad. Sci.* **112**, 4274–4279 (2015).

18. Luo, L., Callaway, E. M. & Svoboda, K. Genetic Dissection of Neural Circuits. *Neuron* **57**, 634–660 (2008).

19. Luo, L., Callaway, E. M. & Svoboda, K. Genetic Dissection of Neural Circuits: A Decade of Progress. *Neuron* **98**, 256–281 (2018).

20. Chen, P. & Hong, W. Neural Circuit Mechanisms of Social Behavior. *Neuron* **98**, 16–30 (2018).

21. Adolphs, R. The Social Brain: Neural Basis of Social Knowledge. *Annu. Rev. Psychol.* **60**, 693–716 (2009).

22. Spunt, R. P. & Adolphs, R. A new look at domain specificity: Insights from social neuroscience. *Nature Reviews Neuroscience* **18**, 559–567 (2017).

23. Adolphs, R. Investigating the cognitive neuroscience of social behavior. *Neuropsychologia* **41**, 119–126 (2003).

24. Stephens, G. J., Silbert, L. J. & Hasson, U. Speaker-listener neural coupling underlies successful communication. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 14425–14430 (2010).

25. Koike, T. *et al.* Neural substrates of shared attention as social memory: A hyperscanning

functional magnetic resonance imaging study. *Neuroimage* **125**, 401–412 (2016).

26. Riga, D. *et al.* Optogenetic dissection of medial prefrontal cortex circuitry. *Front. Syst. Neurosci.* **8**, 230 (2014).

27. Schmitt, L. I. *et al.* Thalamic amplification of cortical connectivity sustains attentional control. *Nature* **545**, 219–223 (2017).

28. Yizhar, O. Optogenetic Insights into Social Behavior Function. *Biol. Psychiatry* **71**, 1075–1080 (2012).

29. Kim, H., Ährlund-Richter, S., Wang, X., Deisseroth, K. & Carlén, M. Prefrontal Parvalbumin Neurons in Control of Attention. *Cell* **164**, 208–218 (2016).

30. Yizhar, O. *et al.* Neocortical excitation/inhibition balance in information processing and social dysfunction. *Nature* **477**, 171–178 (2011).

31. Dikker, S. *et al.* Brain-to-Brain Synchrony Tracks Real-World Dynamic Group Interactions in the Classroom. *Curr. Biol.* **27**, 1375–1380 (2017).

32. Yang, J., Zhang, H., Ni, J., De Dreu, C. K. W. & Ma, Y. Within-group synchronization in the prefrontal cortex associates with intergroup conflict. *Nat. Neurosci.* **23**, 754–760 (2020).

33. Dai, B. *et al.* Neural mechanisms for selectively tuning in to the target speaker in a naturalistic noisy situation. *Nat. Commun.* **9**, 1–12 (2018).